Probabilistic Principal Component Analysis

Michael E. Tipping; Christopher M. Bishop

Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 61, No. 3 (1999), 611-622.

Stable URL:
http://links.jstor.org/sici?sici=1369-7412%281999%2961%3A3%3C611%3APPCA%3E2.0.CO%3B2-E

Journal of the Royal Statistical Society. Series B (Statistical Methodology) is currently published by Royal Statistical Society.

J. R. Statist. Soc. B (1999)
61, *Part* 3, *pp.* 611–622

Probabilistic principal component analysis

Michael E. Tipping and Christopher M. Bishop

Microsoft Research, Cambridge, UK

[Received April 1997. Final revision October 1998]

Summary. Principal component analysis (PCA) is a ubiquitous technique for data analysis and processing, but one which is not based on a probability model. We demonstrate how the principal axes of a set of observed data vectors may be determined through maximum likelihood estimation of parameters in a latent variable model that is closely related to factor analysis. We consider the properties of the associated likelihood function, giving an EM algorithm for estimating the principal subspace iteratively, and discuss, with illustrative examples, the advantages conveyed by this probabilistic approach to PCA.

Keywords: Density estimation; EM algorithm; Gaussian mixtures; Maximum likelihood; Principal component analysis; Probability model

1. Introduction

Principal component analysis (PCA) (Jolliffe, 1986) is a well-established technique for dimensionality reduction, and a chapter on the subject may be found in numerous texts on multivariate analysis. Examples of its many applications include data compression, image processing, visualization, exploratory data analysis, pattern recognition and time series prediction.

The most common derivation of PCA is in terms of a standardized linear projection which maximizes the variance in the projected space (Hotelling, 1933). For a set of observed d-dimensional data vectors $\{\mathbf{t}_n\}$, $n \in \{1, \ldots, N\}$, the q *principal axes* $\mathbf{w}_j$, $j \in \{1, \ldots, q\}$, are those orthonormal axes onto which the retained variance under projection is maximal. It can be shown that the vectors $\mathbf{w}_j$ are given by the q dominant eigenvectors (i.e. those with the largest associated eigenvalues λ_j) of the sample covariance matrix $\mathbf{S} = \Sigma_n (\mathbf{t}_n - \bar{\mathbf{t}})(\mathbf{t}_n - \bar{\mathbf{t}})^{\mathrm{T}}/N$, where $\bar{\mathbf{t}}$ is the data sample mean, such that $\mathbf{S}\mathbf{w}_j = \lambda_j \mathbf{w}_j$. The q principal components of the observed vector $\mathbf{t}_n$ are given by the vector $\mathbf{x}_n = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \bar{\mathbf{t}})$, where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q)$. The variables x_j are then uncorrelated such that the covariance matrix $\Sigma_n \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}}/N$ is diagonal with elements λ_j.

A complementary property of PCA, and that most closely related to the original discussions of Pearson (1901), is that, of all orthogonal linear projections $\mathbf{x}_n = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \bar{\mathbf{t}})$, the principal component projection minimizes the squared reconstruction error $\Sigma_n \|\mathbf{t}_n - \hat{\mathbf{t}}_n\|^2$, where the optimal linear reconstruction of $\mathbf{t}_n$ is given by $\hat{\mathbf{t}}_n = \mathbf{W}\mathbf{x}_n + \bar{\mathbf{t}}$.

However, a notable feature of these definitions of PCA (and one remarked on in many texts) is the absence of an associated probabilistic model for the observed data. The objective of this paper is therefore to address this limitation by demonstrating that PCA may indeed be derived within a density estimation framework.

Address for correspondence: Michael E. Tipping, Microsoft Research, St George House, 1 Guildhall Street, Cambridge, CB2 3NH, UK.
E-mail: mtipping@microsoft.com

 1369–7412/99/61611

We obtain a probabilistic formulation of PCA from a Gaussian latent variable model which is closely related to statistical factor analysis. This model is outlined in Section 2, where we discuss the existing precedence for our approach in the literature. Within the framework that we propose, detailed in Section 3, the principal axes emerge as maximum likelihood parameter estimates which may be computed by the usual eigendecomposition of the sample covariance matrix and subsequently incorporated in the model. Alternatively, the latent variable formulation leads naturally to an iterative, and computationally efficient, expectation–maximization (EM) algorithm for effecting PCA.

Such a probabilistic formulation is intuitively appealing, as the definition of a likelihood measure enables a comparison with other probabilistic techniques, while facilitating statistical testing and permitting the application of Bayesian methods. However, a further motivation is that probabilistic PCA conveys additional practical advantages as follows.

(a) The probability model offers the potential to extend the scope of conventional PCA. For example, we illustrate in Section 4 how multiple PCA models may usefully be combined as a probabilistic mixture and how PCA projections may be obtained when some data values are missing.
(b) As well as its application to dimensionality reduction, probabilistic PCA can be utilized as a constrained Gaussian density model. The benefit of so doing is that maximum likelihood estimates for the parameters associated with the covariance matrix can be efficiently computed from the data principal components. Potential applications include classification and novelty detection, and we again give an example in Section 4.

We conclude with a discussion in Section 5, while mathematical details concerning key results are left to Appendixes A and B.

2. Latent variable models, factor analysis and principal component analysis

2.1. Factor analysis

A latent variable model seeks to relate a d-dimensional observation vector $\mathbf{t}$ to a corresponding q-dimensional vector of latent (or unobserved) variables $\mathbf{x}$. Perhaps the most common such model is *factor analysis* (Bartholomew, 1987; Basilevsky, 1994) where the relationship is linear:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}. \tag{1}$$

The $d \times q$ matrix $\mathbf{W}$ relates the two sets of variables, while the parameter vector $\boldsymbol{\mu}$ permits the model to have non-zero mean. The motivation is that, with $q < d$, the latent variables will offer a more parsimonious explanation of the dependences between the observations. Conventionally, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the latent variables are defined to be independent and Gaussian with unit variance. By additionally specifying the error, or noise, model to be likewise Gaussian $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, equation (1) induces a corresponding Gaussian distribution for the observations $\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi})$. The model parameters may thus be determined by maximum likelihood, although, because there is no closed form analytic solution for $\mathbf{W}$ and $\boldsymbol{\Psi}$, their values must be obtained via an iterative procedure.

The motivation, and indeed key assumption, for the factor analysis model is that, by constraining the error covariance $\boldsymbol{\Psi}$ to be a diagonal matrix whose elements ψ_i are usually estimated from the data, the observed variables t_i are *conditionally independent* given the values of the latent variables $\mathbf{x}$. These latent variables are thus intended to explain the

correlations between observation variables while ϵ_i represents variability unique to a particular t_i. This is where factor analysis fundamentally differs from standard PCA, which effectively treats covariance and variance identically.

2.2. *Links from factor analysis to principal component analysis*

Because of the distinction made between variance and covariance in the standard factor analysis model, the subspace defined by the maximum likelihood estimates of the columns of $\mathbf{W}$ will generally *not* correspond to the principal subspace of the observed data. However, certain links between the two methods have been previously established, and such connections centre on the special case of an *isotropic* error model, where the residual variances $\psi_i = \sigma^2$ are constrained to be equal.

This approach was adopted in the early Young–Whittle factor analysis model (Young, 1940; Whittle, 1952), where in addition the residual variance σ^2 was presumed known (i.e. the model likelihood was a function of $\mathbf{W}$ alone). In that case, maximum likelihood is equivalent to a least squares criterion, and a principal component solution emerges in a straightforward manner.

The methodology employed by Young and Whittle differed from that conventionally adopted, since the factors $\mathbf{x}$ were considered as parameters to be estimated rather than as random variables. However, a stochastic treatment of $\mathbf{x}$ recovers a similar result, given that the $d - q$ smallest eigenvalues of the sample covariance $\mathbf{S}$ are equal to σ^2. In that case, it is simple to show that the observation covariance model $\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$ can be made exact (assuming the correct choice of q), and both $\mathbf{W}$ and σ^2 may then be determined analytically through eigendecomposition of $\mathbf{S}$, without resort to iteration (Anderson (1963) and Basilevsky (1994), pages 361–363).

However, it is restrictive (and rarely justified in practice) to assume that either σ^2 is known or that the model of the second-order statistics of the data is exact. Indeed, in the presence of additive observation noise, an exact covariance model is generally undesirable. This is particularly true in the practical application of PCA, where we often do not require an exact characterization of the covariance structure in the minor subspace, since this information is effectively 'discarded' in the dimensionality reduction process.

In the remainder of this paper we therefore focus on the case of most interest and consider the nature of the maximum likelihood estimators for $\mathbf{W}$ and σ^2 in the realistic case where the model covariance proposed is not equal to its sample counterpart, and where σ^2 must be estimated from the data (and so enters the likelihood function). This case has indeed been investigated, and related to PCA, in the early factor analysis literature by Lawley (1953) and by Anderson and Rubin (1956), although this work does not appear widely known. They showed that stationary points of the likelihood function occur when $\mathbf{W}$ is a matrix whose columns are scaled eigenvectors of the sample covariance matrix $\mathbf{S}$, and σ^2 is the average variance in the discarded dimensions (we give details shortly). These derivations, however, fall short of showing that the *principal* eigenvectors represent the *global* maximum of the likelihood.

In the next section we re-establish this link between PCA and factor analysis, while also extending the earlier derivation to show (in Appendix A) that the maximum likelihood estimators $\mathbf{W}_{\mathrm{ML}}$ and σ^2_{ML} for the isotropic error model do correspond to *principal* component analysis. We give a full characterization of the properties of the likelihood function for what we choose to term 'probabilistic PCA' (PPCA). In addition, we give an iterative EM algorithm for estimating the parameters of interest with potential computational benefits. Finally, to

motivate this work and to underline how the definition of the probability model can be advantageously exploited in practice, we offer some examples of the practical application of PPCA in Section 4.

3. Probabilistic principal component analysis

3.1. The probability model

The use of the isotropic Gaussian noise model $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ for $\boldsymbol{\epsilon}$ in conjunction with equation (1) implies that the $\mathbf{x}$ conditional probability distribution over $\mathbf{t}$-space is given by

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

With the marginal distribution over the latent variables also Gaussian and conventionally defined by $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the marginal distribution for the observed data $\mathbf{t}$ is readily obtained by integrating out the latent variables and is likewise Gaussian:

$$\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}), \tag{3}$$

where the observation covariance model is specified by $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$. The corresponding log-likelihood is then

$$\mathcal{L} = -\frac{N}{2}\{d\ \ln(2\pi) + \ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})\}, \tag{4}$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}. \tag{5}$$

The maximum likelihood estimator for $\boldsymbol{\mu}$ is given by the mean of the data, in which case $\mathbf{S}$ is the sample covariance matrix of the observations $\{\mathbf{t}_n\}$. Estimates for $\mathbf{W}$ and σ^2 may be obtained by iterative maximization of $\mathcal{L}$, e.g. by using the EM algorithm given in Appendix B, which is based on the algorithm for standard factor analysis of Rubin and Thayer (1982). However, in contrast with factor analysis, maximum likelihood estimators for $\mathbf{W}$ and σ^2 may be obtained explicitly, as we see shortly.

Later, we shall make use of the conditional distribution of the latent variables $\mathbf{x}$ given the observed $\mathbf{t}$, which may be calculated by using Bayes's rule and is again Gaussian:

$$\mathbf{x}|\mathbf{t} \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}), \tag{6}$$

where we have defined $\mathbf{M} = \mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I}$. Note that $\mathbf{M}$ is of size $q \times q$ whereas $\mathbf{C}$ is $d \times d$.

3.2. Properties of the maximum likelihood estimators

In Appendix A it is shown that, with $\mathbf{C}$ given by $\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$, the likelihood (4) is maximized when

$$\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{7}$$

where the q column vectors in the $d \times q$ matrix $\mathbf{U}_q$ are the principal eigenvectors of $\mathbf{S}$, with corresponding eigenvalues $\lambda_1, \ldots, \lambda_q$ in the $q \times q$ diagonal matrix $\boldsymbol{\Lambda}_q$, and $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal rotation matrix. Other combinations of eigenvectors (i.e. non-principal ones) correspond to saddlepoints of the likelihood function. Thus, from equation (7), the

latent variable model defined by equation (1) effects a mapping from the latent space into the *principal subspace* of the observed data.

It may also be shown that for $\mathbf{W} = \mathbf{W}_{\mathrm{ML}}$ the maximum likelihood estimator for σ^2 is given by

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j, \tag{8}$$

which has a clear interpretation as the variance 'lost' in the projection, averaged over the lost dimensions.

In practice, to find the most likely model given $\mathbf{S}$, we would first estimate σ^2_{ML} from equation (8), and then $\mathbf{W}_{\mathrm{ML}}$ from equation (7), where for simplicity we would effectively ignore $\mathbf{R}$ (i.e. choose $\mathbf{R} = \mathbf{I}$). Alternatively, we might employ the EM algorithm detailed in Appendix B, where $\mathbf{R}$ at convergence can be considered arbitrary.

3.3. *Factor analysis revisited*

Although the above estimators result from the application of a simple constraint to the standard factor analysis model, we note that an important distinction resulting from the use of the isotropic noise covariance $\sigma^2\mathbf{I}$ is that PPCA is covariant under rotation of the original data axes, as is standard PCA, whereas factor analysis is covariant under componentwise rescaling. Another point of contrast is that in factor analysis neither of the factors found by a two-factor model is necessarily the same as that found by a single-factor model. In PPCA, we see above that the principal axes may be found incrementally.

3.4. *Dimensionality reduction*

The general motivation for PCA is to transform the data into some reduced dimensionality representation, and with some minor algebraic manipulation of $\mathbf{W}_{\mathrm{ML}}$ we may indeed obtain the standard projection onto the principal axes if desired. However, it is more natural from a probabilistic perspective to consider the dimensionality reduction process in terms of the distribution of the latent variables, conditioned on the observation. From expression (6), this distribution may be conveniently summarized by its *mean*:

$$\langle \mathbf{x}_n | \mathbf{t}_n \rangle = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}_{\mathrm{ML}}(\mathbf{t}_n - \boldsymbol{\mu}). \tag{9}$$

(Note, also from expression (6), that the corresponding conditional covariance is given by $\sigma^2_{\mathrm{ML}}\mathbf{M}^{-1}$ and is thus independent of n.) It can be seen that, when $\sigma^2 \to 0$, $\mathbf{M}^{-1} \to (\mathbf{W}^{\mathrm{T}}_{\mathrm{ML}}\mathbf{W}_{\mathrm{ML}})^{-1}$ and equation (9) then represents an orthogonal projection into latent space and so standard PCA is recovered. However, the density model then becomes singular, and thus undefined. In practice, with $\sigma^2 > 0$ as determined by equation (8), the latent projection becomes skewed towards the origin as a result of the Gaussian marginal distribution for $\mathbf{x}$. Because of this, the reconstruction $\mathbf{W}_{\mathrm{ML}}\langle \mathbf{x}_n | \mathbf{t}_n \rangle + \boldsymbol{\mu}$ is *not* an orthogonal projection of $\mathbf{t}_n$ and is therefore not optimal (in the squared reconstruction error sense). Nevertheless, optimal reconstruction of the observed data from the conditional latent mean may still be obtained, in the case of $\sigma^2 > 0$, and is given by $\mathbf{W}_{\mathrm{ML}}(\mathbf{W}^{\mathrm{T}}_{\mathrm{ML}}\mathbf{W}_{\mathrm{ML}})^{-1}\mathbf{M}\langle \mathbf{x}_n | \mathbf{t}_n \rangle + \boldsymbol{\mu}$.

4. Examples

Here we give three examples of how PPCA can be exploited in practice. We first consider the

visualization of data sets with missing values and then extend this single projection model to the mixture case, before finally giving an example of how the covariance parameterization that is implicit in PPCA offers an effective mechanism for restricting the number of degrees of freedom in a Gaussian model.

4.1. Missing data

PPCA offers a natural approach to the estimation of the principal axes in cases where some, or indeed all, of the data vectors $\mathbf{t}_n = (t_{n1}, t_{n2}, \ldots, t_{nd})$ exhibit one or more missing (at random) values. Drawing on the standard methodology for maximizing the likelihood of a Gaussian model in the presence of missing values (Little and Rubin, 1987) and the EM algorithm for PPCA given in Appendix B, we may derive an iterative algorithm for maximum likelihood estimation of the principal axes, where both the latent variables $\{\mathbf{x}_n\}$ and the missing observations $\{t_{nj}\}$ make up the 'complete' data. Fig. 1(a) shows a projection of 38 examples from the 18-dimensional *Tobamovirus* data utilized by Ripley (1996), p. 291, to illustrate standard PCA. Of interest in the plot is the evidence of three subgroupings and the atypicality of example 11. We simulated missing data by randomly removing each value in the data set with probability 20%. Fig. 1(b) shows an equivalent PPCA projection obtained by using an EM algorithm, where the conditional means have also been averaged over the conditional distribution of missing, given observed, values. The salient features of the projection remain clear, despite the fact that all the data vectors suffered from at least one missing value.

4.2. Mixtures of probabilistic principal component analysis models

Because PCA defines a single linear projection and is thus a relatively simplistic technique, there has been significant recent interest in obtaining more complex projection methods by combining *multiple* PCA models, notably for image compression (Dony and Haykin, 1995) and visualization (Bishop and Tipping, 1998). Such a complex model is readily implemented as a *mixture* of such PPCA models. By means of a simple illustration, Fig. 2 shows *three* PCA projections of the virus data obtained from a three-component mixture model, optimized by

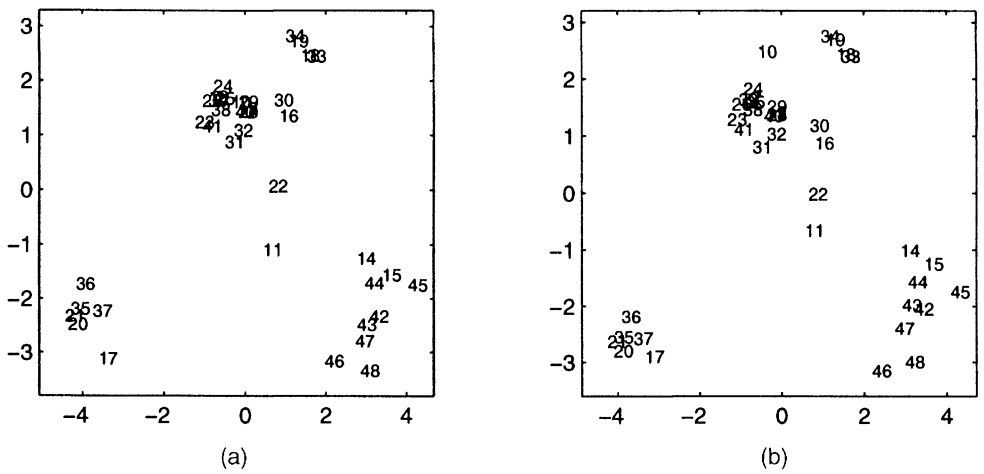

Fig. 1. Projections of the *Tobamovirus* data by using (a) PCA on the full data set and (b) PPCA with 136 missing values

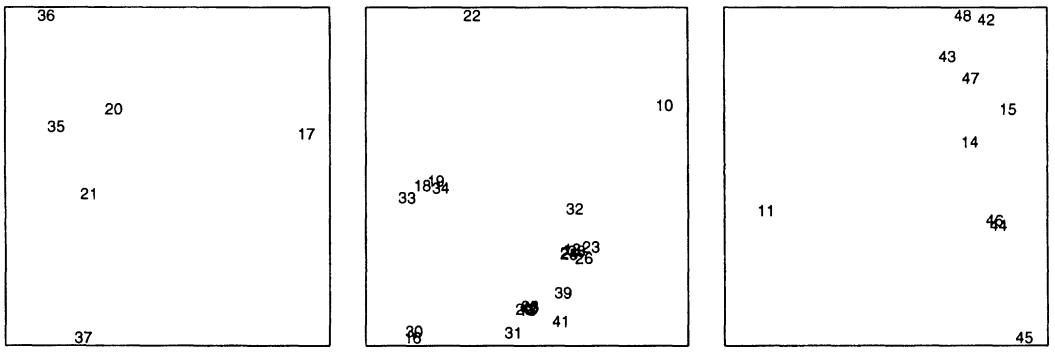

Fig. 2. Projections of the *Tobamovirus* data obtained from a three-component PPCA mixture model: the locations of these three projection planes can be superimposed on the single principal component projection plot (Fig. 1(a)) to aid the interpretation of the data structure further

using an EM algorithm again derived by combining standard methods (Titterington *et al.*, 1985) with the algorithm given in Appendix B. In theory, the projection of every data point would appear in each plot, corresponding to three sets of principal axes associated with each component in the mixture, but in practice examples need not be shown in the plot if the corresponding component model has negligible conditional probability of having generated them. This effectively implements a simultaneous automated clustering and visualization of data, which is much more powerful than simply subsetting the data by eye and performing individual PCAs. Multiple plots such as these offer the potential to reveal more complex structure than a single PCA projection alone.

4.3. Controlling the degrees of freedom

An alternative perspective on PPCA is that it can be applied simply as a covariance model of data, where the covariance $\mathbf{C}$ is defined in terms of the auxiliary parameters $\mathbf{W}$ and σ^2. This is particularly relevant for larger values of data dimensionality d and moderately sized data sets, where it is usually inappropriate to fit a full covariance model, as this implies the estimation of $d(d+1)/2$ free parameters. In such cases, constraints are often placed on the covariance matrix, that it be, for example, diagonal (with d parameters) or proportional to the identity matrix (one parameter). The covariance model in PPCA comprises $dq + 1 - q(q-1)/2$ free parameters and thus permits control of the model complexity through the choice of q. (Here, we stress that we are considering the predictive power of the model, rather than the explanatory sense in which we might interpret q in traditional factor analysis.)

We illustrate this in Table 1, which shows the estimated prediction error (in this case, the negative log-likelihood per example) for various Gaussian models fitted to the *Tobamovirus* data. The dimensionality of the data is large, at 18, compared with the number of examples, 38, and so more complex models easily overfit. However, a PPCA density model with latent space dimension $q = 2$ gives the lowest error. More practically, in other problems we could apply PPCA to the modelling of class conditional densities, and select a value (or values) of q which optimized the classification accuracy. Because the maximum likelihood estimators for $\mathbf{W}$ and σ^2, and thus $\mathbf{C}$, can be found explicitly by eigendecomposition of the sample covariance matrix, the search for an appropriate complexity of model can be performed explicitly and relatively cheaply.

Table 1. Complexity and bootstrap estimate of the prediction error for various Gaussian models of the *Tobamovirus* data†

Covariance model	*q (equivalent)*	*Number of parameters*	*Prediction error*
Isotropic	(0)	1	18.6
Diagonal	(—)	18	19.6
PPCA	1	19	16.8
	2	36	14.8
	3	52	15.6
Full	(17)	171	3193.5

†The isotropic and full covariance models are equivalent to special cases of PPCA, with $q = 0$ and $q = d - 1$ respectively.

5. Discussion

We have reiterated and extended the earlier work of Lawley (1953) and Anderson and Rubin (1956) and shown how PCA may be viewed as a maximum likelihood procedure based on a probability density model of the observed data. This probability model is Gaussian, and the model covariance is determined simply by applying equations (7) and (8), requiring only the computation of the eigenvectors and eigenvalues of the sample covariance matrix. However, in addition to this explicit formulation, we have also given an EM algorithm for finding the principal axes by iteratively maximizing the likelihood function, and this approach may be more efficient for larger values of data dimensionality as discussed in Appendix B.

Examples given in Section 4 demonstrated the utility of the probabilistic formalism, where we performed PCA on a data set with missing values, generalized the single model to the mixture case and demonstrated the capacity to constrain the number of free parameters in a Gaussian density model. Indeed, we have exploited these possibilities in practice to obtain more powerful algorithms for the visualization of data and more efficient methods for image compression.

Finally, we note that factor analysis is generally applied to elucidate an *explanation* of data, and, whereas PPCA is closely related to factor analysis in its formulation, the above examples reflect that our motivation for its application has in general *not* been explanatory. Rather, we have considered PPCA as a mechanism for probabilistic dimension reduction, or as a variable complexity predictive density model.

Acknowledgements

The authors would like to thank the referees for their valuable comments and suggestions. Michael Tipping was originally supported at Aston University by Engineering and Physical Sciences Research Council contract GR/K51808, 'Neural networks for visualisation of high dimensional data', and both authors would like to thank the Isaac Newton Institute, Cambridge, for its hospitality during the preparation of this work.

Appendix A: Maximum likelihood principal component analysis

A.1. The stationary points of the log-likelihood

The gradient of log-likelihood (4) with respect to $\mathbf{W}$ may be obtained from standard matrix differentiation results (for example see Krzanowski and Marriott (1994), p. 133):

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}). \tag{10}$$

At the stationary points

$$\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}, \tag{11}$$

assuming that $\mathbf{C}^{-1}$ exists, which we shall see requires that $q < \text{rank}(\mathbf{S})$, so this assumption implies no loss of practicality.

There are three possible classes of solutions to equation (11). The first is, trivially, $\mathbf{W} = \mathbf{0}$, which will be seen to be a minimum of the log-likelihood. Second is the case $\mathbf{C} = \mathbf{S}$, where the covariance model is exact and the $d - q$ smallest eigenvalues of $\mathbf{S}$ are identical as discussed in Section 2.2. Then, $\mathbf{W}$ is identifiable since $\mathbf{W}\mathbf{W}^{\mathrm{T}} = \mathbf{S} - \sigma^2\mathbf{I}$ has a known solution at $\mathbf{W} = \mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$, where $\mathbf{U}$ is a square matrix whose columns are the eigenvectors of $\mathbf{S}$, with $\mathbf{\Lambda}$ the corresponding diagonal matrix of eigenvalues, and $\mathbf{R}$ is an arbitrary orthogonal (i.e. rotation) matrix.

However, the 'interesting' solutions represent the third case, where $\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}$, but $\mathbf{W} \neq \mathbf{0}$ and $\mathbf{C} \neq \mathbf{S}$. To find these we first express the parameter matrix $\mathbf{W}$ in terms of its singular value decomposition:

$$\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}, \tag{12}$$

where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q)$ is a $d \times q$ matrix of orthonormal column vectors, $\mathbf{L} = \text{diag}(l_1, l_2, \ldots, l_q)$ is the $q \times q$ diagonal matrix of singular values and $\mathbf{V}$ is a $q \times q$ orthogonal matrix. Then, substituting this decomposition into equation (11) gives after some manipulation

$$\mathbf{S}\mathbf{U}\mathbf{L} = \mathbf{U}(\sigma^2\mathbf{I} + \mathbf{L}^2)\mathbf{L}. \tag{13}$$

For $l_j \neq 0$, equation (13) implies that $\mathbf{S}\mathbf{u}_j = (\sigma^2 + l_j^2)\mathbf{u}_j$ for each vector $\mathbf{u}_j$. Therefore, each column of $\mathbf{U}$ must be an eigenvector of $\mathbf{S}$, with corresponding eigenvalue $\lambda_j = \sigma^2 + l_j^2$, and so

$$l_j = (\lambda_j - \sigma^2)^{1/2}. \tag{14}$$

For $l_j = 0$, $\mathbf{u}_j$ is arbitrary. All potential solutions for $\mathbf{W}$ may thus be written as

$$\mathbf{W} = \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{15}$$

where $\mathbf{U}_q$ is a $d \times q$ matrix whose q columns $\mathbf{u}_j$ are eigenvectors of $\mathbf{S}$, $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal matrix and $\mathbf{K}_q$ is a $q \times q$ diagonal matrix with elements

$$k_j = \begin{cases} \lambda_j, & \text{the corresponding eigenvalue to } \mathbf{u}_j, \text{ or} \\ \sigma^2, & \end{cases} \tag{16}$$

where the latter case may be seen to be equivalent to $l_j = 0$.

A.2. The global maximum of the likelihood

The matrix $\mathbf{U}_q$ may contain any of the eigenvectors of $\mathbf{S}$, so to identify those which maximize the likelihood the expression for $\mathbf{W}$ in equation (15) is substituted into the log-likelihood function (4) to give

$$\mathcal{L} = -\frac{N}{2}\left\{ d\ \ln(2\pi) + \sum_{j=1}^{q'} \ln(\lambda_j) + \frac{1}{\sigma^2}\sum_{j=q'+1}^{d} \lambda_j + (d - q')\ \ln(\sigma^2) + q' \right\}, \tag{17}$$

where q' is the number of non-zero l_j, $\lambda_1, \ldots, \lambda_{q'}$ are the eigenvalues corresponding to the eigenvectors 'retained' in $\mathbf{W}$ and $\lambda_{q'+1}, \ldots, \lambda_d$ are those 'discarded'. Maximizing the log-likelihood (17) with respect to σ^2 gives

$$\sigma^2 = \frac{1}{d - q'}\sum_{j=q'+1}^{d} \lambda_j, \tag{18}$$

and so

$$\mathcal{L} = -\frac{N}{2}\left\{\sum_{j=1}^{q'} \ln(\lambda_j) + (d-q')\ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d}\lambda_j\right) + d\ln(2\pi) + d\right\}. \tag{19}$$

Equation (18) implies that $\sigma^2 > 0$ if $\text{rank}(\mathbf{S}) > q$ as stated earlier. We wish to find the maximum of the log-likelihood (19) with respect to the choice of eigenvectors or eigenvalues to retain in $\mathbf{W}$ and those to discard. By exploiting the constancy of the sum of all eigenvalues, the condition for the maximization of the likelihood can be expressed equivalently as the minimization of the quantity

$$E = \ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d}\lambda_j\right) - \frac{1}{d-q'}\sum_{j=q'+1}^{d}\ln(\lambda_j), \tag{20}$$

which only depends on the discarded values and is non-negative (Jensen's inequality). Interestingly, the minimization of E leads only to the requirement that the discarded λ_j be adjacent within the spectrum of the ordered eigenvalues of $\mathbf{S}$. However, equation (14) also requires that $\lambda_j > \sigma^2$, $\forall j \in \{1, \ldots, q'\}$, so from equation (18) we can deduce from this that the *smallest* eigenvalue *must be discarded.* This is now sufficient to show that E must then be minimized when $\lambda_{q'+1}, \ldots, \lambda_d$ are the *smallest* $d-q'$ eigenvalues and so the likelihood $\mathcal{L}$ is maximized when $\lambda_1, \ldots, \lambda_{q'}$ are the largest eigenvalues of $\mathbf{S}$.

It should also be noted that $\mathcal{L}$ is maximized, with respect to q', when there are fewest terms in the sums in equation (20) which occurs when $q' = q$ and therefore no l_j is 0. Furthermore, $\mathcal{L}$ is *minimized* when $\mathbf{W} = \mathbf{0}$, which may be seen to be equivalent to the case of $q' = 0$.

A.3. The nature of other stationary points

If stationary points represented by minor eigenvector solutions are stable maxima, then local maximization (via an EM algorithm for example) is not guaranteed to converge to the global maximum comprising the principal eigenvectors. We may show, however, that minor eigenvector solutions are saddlepoints on the likelihood surface.

Consider a stationary point of the gradient equation (10) at $\hat{\mathbf{W}} = \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}$, where $\mathbf{U}_q$ may contain q arbitrary eigenvectors of $\mathbf{S}$, and $\mathbf{K}_q$, as defined in equations (16), contains either the corresponding eigenvalue or σ^2. (For clarity, the rotation $\mathbf{R}$ is ignored here, but it can easily be incorporated in the following analysis.) Then consider a small perturbation to a column vector $\hat{\mathbf{w}}_i$ in $\hat{\mathbf{W}}$ of the form $\epsilon\mathbf{u}_j$, where ϵ is an arbitrarily small positive constant and $\mathbf{u}_j$ is a discarded eigenvector.

For $\hat{\mathbf{W}}$ to represent a stable solution, the dot product of the likelihood gradient at $\hat{\mathbf{w}}_i + \epsilon\mathbf{u}_j$ and the perturbation must be negative. This dot product may be straightforwardly computed and, ignoring terms in ϵ^2, is given by

$$\epsilon N(\lambda_j/k_i - 1)\mathbf{u}_j^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{u}_j, \tag{21}$$

where k_i is the value in $\mathbf{K}_q$ corresponding to $\hat{\mathbf{w}}_i$ and λ_j is the eigenvalue corresponding to the perturbation $\mathbf{u}_j$. Since $\mathbf{C}^{-1}$ is positive definite, $\mathbf{u}_j^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{u}_j > 0$ and so the sign of the gradient is determined by $\lambda_j/k_i - 1$. When $k_i = \lambda_i$, this term is negative if $\lambda_i > \lambda_j$, in which case the maximum is stable. If $\lambda_i < \lambda_j$ then $\hat{\mathbf{W}}$ must be a saddlepoint. If $k_i = \sigma^2$, the stationary point can never be stable since, from equation (18), σ^2 is the average of $d-q'$ eigenvalues, and so $\lambda_j > \sigma^2$ for at least one of those eigenvalues, *except* when all those eigenvalues are identical. Such a case is considered in the next section.

From expression (21), by considering all possible perturbations $\mathbf{u}_j$ to all possible column vectors $\hat{\mathbf{w}}_i$ of $\hat{\mathbf{W}}$, it can be seen that the only stable maximum occurs when $\hat{\mathbf{W}}$ comprises the q principal eigenvectors.

A.4. Equality of eigenvalues

Equality of any of the q principal eigenvalues does not affect the analysis presented. However, a consideration should be given to the instance when all the $d-q$ minor (discarded) eigenvalue(s) are equal and identical with one or more of the smallest principal (retained) eigenvalue(s). (In practice, particularly in the case of sampled covariance matrices, this exact $\mathbf{C} = \mathbf{S}$ case is unlikely.)

Consider the example of extracting two components from data with a covariance matrix having eigenvalues 2, 1 and 1. In this case, the second principal axis is not uniquely defined within the minor subspace. The spherical noise distribution defined by σ^2, in addition to explaining the residual variance, can also optimally explain the second principal component. Because $\lambda_2 = \sigma^2$, l_2 in equation (14) is 0,

and $\mathbf{W}$ effectively only comprises a single vector. The combination of this single vector and the noise distribution still represents the maximum of the likelihood.

Appendix B: An EM algorithm for probabilistic principal component analysis

In the EM approach to maximizing the likelihood for PPCA, we consider the latent variables $\{\mathbf{x}_n\}$ to be 'missing' data and the 'complete' data to comprise the observations together with these latent variables. The corresponding complete-data log-likelihood is then

$$\mathcal{L}_C = \sum_{n=1}^{N} \ln\{p(\mathbf{t}_n, \mathbf{x}_n)\}, \tag{22}$$

where, in PPCA, from the definitions in Section 3.1,

$$p(\mathbf{t}_n, \mathbf{x}_n) = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|\mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \boldsymbol{\mu}\|^2}{2\sigma^2}\right)(2\pi)^{-q/2} \exp\left(-\frac{\|\mathbf{x}_n\|^2}{2}\right). \tag{23}$$

In the E-step, we take the expectation of $\mathcal{L}_C$ with respect to the distributions $p(\mathbf{x}_n|\mathbf{t}_n, \mathbf{W}, \sigma^2)$:

$$\langle\mathcal{L}_C\rangle = -\sum_{n=1}^{N}\left\{\frac{d}{2}\ln(\sigma^2) + \frac{1}{2}\operatorname{tr}(\langle\mathbf{x}_n\mathbf{x}_n^T\rangle) + \frac{1}{2\sigma^2}(\mathbf{t}_n - \boldsymbol{\mu})^T(\mathbf{t}_n - \boldsymbol{\mu}) - \frac{1}{\sigma^2}\langle\mathbf{x}_n\rangle^T\mathbf{W}^T(\mathbf{t}_n - \boldsymbol{\mu}) + \frac{1}{2\sigma^2}\operatorname{tr}(\mathbf{W}^T\mathbf{W}\langle\mathbf{x}_n\mathbf{x}_n^T\rangle)\right\}, \tag{24}$$

where we have omitted terms that are independent of the model parameters and

$$\langle\mathbf{x}_n\rangle = \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{t}_n - \boldsymbol{\mu}), \tag{25}$$

$$\langle\mathbf{x}_n\mathbf{x}_n^T\rangle = \sigma^2\mathbf{M}^{-1} + \langle\mathbf{x}_n\rangle\langle\mathbf{x}_n\rangle^T, \tag{26}$$

in which $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}$ as before. Note that these statistics are computed using the current (fixed) values of the parameters and follow from distribution (6) earlier.

In the M-step, $\langle\mathcal{L}_C\rangle$ is maximized with respect to $\mathbf{W}$ and σ^2 giving new parameter estimates

$$\tilde{\mathbf{W}} = \left\{\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})\langle\mathbf{x}_n\rangle^T\right\}\left(\sum_{n=1}^{N}\langle\mathbf{x}_n\mathbf{x}_n^T\rangle\right)^{-1}, \tag{27}$$

$$\tilde{\sigma}^2 = \frac{1}{Nd}\sum_{n=1}^{N}\{\|\mathbf{t}_n - \boldsymbol{\mu}\|^2 - 2\langle\mathbf{x}_n\rangle^T\tilde{\mathbf{W}}^T(\mathbf{t}_n - \boldsymbol{\mu}) + \operatorname{tr}(\langle\mathbf{x}_n\mathbf{x}_n^T\rangle\tilde{\mathbf{W}}^T\tilde{\mathbf{W}})\}. \tag{28}$$

To maximize the likelihood then, the sufficient statistics of the conditional distributions are calculated from equations (25) and (26), after which revised estimates for the parameters are obtained from equations (27) and (28). These four equations are iterated in sequence until the algorithm is judged to have converged.

We may gain considerable insight into the operation of the EM algorithm by substituting for $\langle\mathbf{x}_n\rangle$ and $\langle\mathbf{x}_n\mathbf{x}_n^T\rangle$ from equations (25) and (26) into equations (27) and (28). Some further manipulation leads to both the E-step and the M-step being combined and rewritten as

$$\tilde{\mathbf{W}} = \mathbf{S}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^T\mathbf{S}\mathbf{W})^{-1}, \tag{29}$$

$$\tilde{\sigma}^2 = \frac{1}{d}\operatorname{tr}(\mathbf{S} - \mathbf{S}\mathbf{W}\mathbf{M}^{-1}\tilde{\mathbf{W}}^T), \tag{30}$$

where $\mathbf{S}$ is again given by

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^T. \tag{31}$$

The first instance of $\mathbf{W}$ in equation (30) is the *old* value of the parameter matrix, whereas the second instance $\widetilde{\mathbf{W}}$ is the *new* value calculated from equation (29). Equations (29)–(31) indicate that the data enter the EM formulation only through the covariance matrix $\mathbf{S}$, as would be expected.

Although it is algebraically convenient to express the EM algorithm in terms of $\mathbf{S}$, care should be exercised in the implementation. When $q \ll d$, considerable computational savings might be obtained by not explicitly evaluating $\mathbf{S}$, even though this need only be done once at initialization. The computation of $\mathbf{S}$ requires $O(Nd^2)$ operations, but an inspection of equations (27) and (28) indicates that the complexity is only $O(Ndq)$. This is reflected by the fact that equations (29) and (30) only require terms of the form $\mathbf{SW}$ and $\mathrm{tr}(\mathbf{S})$. For the former, computing $\mathbf{SW}$ as $\Sigma_n \mathbf{x}_n(\mathbf{x}_n^{\mathrm{T}}\mathbf{W})$ is $O(Ndq)$ and so more efficient than $(\Sigma_n \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}})\mathbf{W}$, which is equivalent to finding $\mathbf{S}$ explicitly. The trade-off between the cost of initially computing $\mathbf{S}$ directly and that of computing $\mathbf{SW}$ more cheaply at each iteration will clearly depend on the number of iterations needed to obtain the accuracy of solution required and the ratio of d to q.

A final point to note is that, at convergence, although the columns of $\mathbf{W}_{\mathrm{ML}}$ will span the principal subspace, they need not be orthogonal since

$$\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}} = \mathbf{R}^{\mathrm{T}}(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})\mathbf{R}, \tag{32}$$

which is not diagonal for $\mathbf{R} \neq \mathbf{I}$. In common with factor analysis, and indeed some other iterative PCA algorithms, there is an element of rotational ambiguity. However, if required, the true principal axes may be determined by noting that equation (32) represents an eigenvector decomposition of $\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}$, where the transposed rotation matrix $\mathbf{R}^{\mathrm{T}}$ is simply the matrix whose columns are the eigenvectors of the $q \times q$ matrix $\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}$ (and therefore also of $\mathbf{M}$).

References

Anderson, T. W. (1963) Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, **34**, 122–148.

Anderson, T. W. and Rubin, H. (1956) Statistical inference in factor analysis. In *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability* (ed. J. Neyman), vol. V, pp. 111–150. Berkeley: University of California Press.

Bartholomew, D. J. (1987) *Latent Variable Models and Factor Analysis*. London: Griffin.

Basilevsky, A. (1994) *Statistical Factor Analysis and Related Methods*. New York: Wiley.

Bishop, C. M. and Tipping, M. E. (1998) A hierarchical latent variable model for data visualization. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 281–293.

Dony, R. D. and Haykin, S. (1995) Optimally adaptive transform coding. *IEEE Trans. Image Process.*, **4**, 1358–1370.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, **24**, 417–441.

Jolliffe, I. T. (1986) *Principal Component Analysis*. New York: Springer.

Krzanowski, W. J. and Marriott, F. H. C. (1994) *Multivariate Analysis*, part 2, *Classification, Covariance Structures and Repeated Measurements*. London: Arnold.

Lawley, D. N. (1953) A modified method of estimation in factor analysis and some large sample results. *Nord. Psykol. Monogr. Ser.*, **3**, 35–42.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. Chichester: Wiley.

Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dub. Phil. Mag. J. Sci.*, 6th ser., **2**, 559–572.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rubin, D. B. and Thayer, D. T. (1982) EM algorithms for ML factor analysis. *Psychometrika*, **47**, 69–76.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *The Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Whittle, P. (1952) On principal components and least square methods of factor analysis. *Skand. Akt.*, **36**, 223–239.

Young, G. (1940) Maximum likelihood estimation and factor analysis. *Psychometrika*, **6**, 49–53.