Mixtures of Probabilistic Principal Component Analysers

Michael E. Tipping and Christopher M. Bishop
Microsoft Research, Cambridge, U.K.

Published as:	"Mixtures of probabilistic principal component analysers", Neural Computation 11(2), pp 443–482. MIT Press.
Year of publication:	1999
This version typeset:	June 26, 2006
Available from:	`http://www.miketipping.com/papers.htm`
Correspondence:	`mail@miketipping.com`

Abstract Principal component analysis (PCA) is one of the most popular techniques for processing, compressing and visualising data, although its effectiveness is limited by its global linearity. While nonlinear variants of PCA have been proposed, an alternative paradigm is to capture data complexity by a combination of local linear PCA projections. However, conventional PCA does not correspond to a probability density, and so there is no unique way to combine PCA models. Previous attempts to formulate mixture models for PCA have therefore to some extent been *ad hoc*. In this paper, PCA is formulated within a maximum-likelihood framework, based on a specific form of Gaussian latent variable model. This leads to a well-defined mixture model for probabilistic principal component analysers, whose parameters can be determined using an EM algorithm. We discuss the advantages of this model in the context of clustering, density modelling and local dimensionality reduction, and we demonstrate its application to image compression and handwritten digit recognition.

1 Introduction

Principal component analysis (PCA) (Jolliffe 1986) has proven to be an exceedingly popular technique for dimensionality reduction and is discussed at length in most texts on multivariate analysis. Its many application areas include data compression, image analysis, visualization, pattern recognition, regression and time series prediction.

The most common definition of PCA, due to Hotelling (1933), is that, for a set of observed d-dimensional data vectors $\{\mathbf{t}_n\}$, $n \in \{1 \dots N\}$, the q *principal axes* $\mathbf{w}_j$, $j \in \{1 \dots q\}$, 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) of the sample covariance matrix $\mathbf{S} = \sum_n (\mathbf{t}_n - \bar{\mathbf{t}})(\mathbf{t}_n - \bar{\mathbf{t}})^{\mathrm{T}}/N$ such that $\mathbf{S}\mathbf{w}_j = \lambda_j \mathbf{w}_j$ and where $\bar{\mathbf{t}}$ is the sample mean. The vector $\mathbf{x}_n = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \bar{\mathbf{t}})$, where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_q)$, is thus a q-dimensional reduced representation of the observed vector $\mathbf{t}_n$.

A complementary property of PCA, and that most closely related to the original discussions of Pearson (1901), is that the projection onto the *principal subspace* minimizes the squared reconstruction error $\sum \|\mathbf{t}_n - \hat{\mathbf{t}}_n\|^2$. The optimal linear reconstruction of $\mathbf{t}_n$ is given by $\hat{\mathbf{t}}_n = \mathbf{W}\mathbf{x}_n + \bar{\mathbf{t}}$, where $\mathbf{x}_n = \mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \bar{\mathbf{t}})$, and the orthogonal columns of $\mathbf{W}$ span the space of the leading q eigenvectors

of **S**. In this context, the principal component projection is often known as the *Karhunen-Loève transform*.

One limiting disadvantage of these definitions of PCA is the absence of an associated probability density or generative model. Deriving PCA from the perspective of density estimation would offer a number of important advantages including the following:

- The corresponding likelihood would permit comparison with other density-estimation techniques and facilitate statistical testing.
- Bayesian inference methods could be applied (e.g. for model comparison) by combining the likelihood with a prior.
- In classification, PCA could be used to model class-conditional densities, thereby allowing the posterior probabilities of class membership to be computed. This contrasts with the alternative application of PCA for classification of Oja (1983) and Hinton *et al.* (1997).
- The value of the probability density function could be used as a measure of the 'degree of novelty' of a new data point, an alternative approach to that of Japkowicz *et al.* (1995) and Petsche *et al.* (1996) in autoencoder-based PCA.
- The probability model would offer a methodology for obtaining a principal component projection when data values are missing.
- The single PCA model could be extended to a *mixture* of such models.

This final advantage is particularly significant. Because PCA only defines a *linear* projection of the data, the scope of its application is necessarily somewhat limited. This has naturally motivated various developments of *nonlinear* principal component analysis in an effort to retain a greater proportion of the variance using fewer components. Examples include principal curves (Hastie and Stuetzle 1989; Tibshirani 1992), multi-layer auto-associative neural networks (Kramer 1991), the kernel-function approach of Webb (1996) and the generative topographic mapping, or GTM, of Bishop, Svensén, and Williams (1998). However, an alternative paradigm to such *global nonlinear* approaches is to model nonlinear structure with a collection, or mixture, of *local linear* sub-models. This philosophy is an attractive one, motivating, for example, the 'mixture of experts' technique for regression (Jordan and Jacobs 1994).

A number of implementations of 'mixtures of PCA' have been proposed in the literature, each of which defines a different algorithm, or a variation thereupon. The variety of proposed approaches is a consequence of ambiguity in the formulation of the overall model. Current methods for local PCA generally necessitate a two-stage procedure: a partitioning of the data space followed by estimation of the principal subspace within each partition. Standard Euclidean distance-based clustering may be performed in the partitioning phase, but more appropriately, the *reconstruction error* may be utilised as the criterion for cluster assignments. This conveys the advantage that a common cost measure is utilised in both stages. However, even recently proposed models which adopt this cost measure still define different algorithms (Hinton, Dayan, and Revow 1997; Kambhatla and Leen 1997), while a variety of alternative approaches for combining local PCA models have also been proposed (Broomhead *et al.* 1991; Bregler and Omohundro 1995; Hinton *et al.* 1995; Dony and Haykin 1995). None of these algorithms defines a probability density.

One difficulty in implementation is that when utilising 'hard' clustering in the partitioning phase (Kambhatla and Leen 1997), the overall cost function is inevitably non-differentiable. Hinton, Dayan, and Revow (1997) finesse this problem by considering the partition assignments as 'missing data' in an expectation-maximization (EM) framework, and thereby propose a 'soft' algorithm where instead of any given data point being assigned exclusively to one principal component analyser, the 'responsibility' for its 'generation' is shared amongst all of the analysers. The authors concede that the absence of a probability model for PCA is a limitation to their approach and

propose that the responsibility of the jth analyser for reconstructing data point $\mathbf{t}_n$ be given by $r_{nj} = \exp(-E_j^2/2\sigma^2) / \left\{\sum_{j'} \exp(-E_{j'}^2/2\sigma^2)\right\}$, where E_j is the corresponding reconstruction cost. This allows the model to be determined by the maximization of a pseudo-likelihood function, and an explicit two-stage algorithm is unnecessary. Unfortunately, this also requires the introduction of a variance parameter σ^2 whose value is somewhat arbitrary, and again, no probability density is defined.

Our key result is to derive a probabilistic model for PCA. From this a mixture of local PCA models follows as a natural extension in which all of the model parameters may be estimated through the maximization of a single likelihood function. Not only does this lead to a clearly defined and unique algorithm, but it also conveys the advantage of a probability density function for the final model with all the associated benefits as outlined above.

In Section 2, we describe the concept of latent variable models. We then introduce *probabilistic principal component analysis* (PPCA) in Section 3, showing how the principal subspace of a set of data vectors can be obtained within a maximum-likelihood framework. Next we extend this result to mixture models in Section 4, and outline an efficient EM algorithm for estimating all of the model parameters in a mixture of probabilistic principal component analysers. The partitioning of the data and the estimation of local principal axes are automatically linked. Furthermore, the algorithm implicitly incorporates a soft clustering similar to that implemented by Hinton *et al.* (1997), in which the parameter σ^2 appears naturally within the model. Indeed, σ^2 has a simple interpretation and is determined by the same EM procedure used to update the other model parameters.

The proposed PPCA mixture model has a wide applicability, and we discuss its advantages from two distinct perspectives. First, in Section 5, we consider PPCA for dimensionality reduction and data compression in local linear modelling. We demonstrate the operation of the algorithm on a simple toy problem, and compare its performance with that of an explicit reconstruction-based non-probabilistic modelling method on both synthetic and real-world datasets.

A second perspective is that of general Gaussian mixtures. The PPCA mixture model offers a way to control the number of parameters when estimating covariance structures in high dimensions, while not over-constraining the model flexibility. We demonstrate this property in Section 6, and apply the approach to the classification of images of handwritten digits.

Proofs of key results and algorithmic details are left to the appendices.

2 Latent Variable Models and PCA

2.1 Latent Variable Models

A latent variable model seeks to relate a d-dimensional observed data vector $\mathbf{t}$ to a corresponding q-dimensional vector of latent variables $\mathbf{x}$:

$$\mathbf{t} = \mathbf{y}(\mathbf{x}; \mathbf{w}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y}(\cdot;\cdot)$ is a function of the latent variables $\mathbf{x}$ with parameters $\mathbf{w}$, and $\boldsymbol{\epsilon}$ is an $\mathbf{x}$-independent noise process. Generally, $q < d$ such that the latent variables offer a more parsimonious description of the data. By defining a prior distribution over $\mathbf{x}$, together with the distribution of $\boldsymbol{\epsilon}$, equation (1) induces a corresponding distribution in the data space, and the model parameters may then be determined by maximum-likelihood techniques. Such a model may also be termed 'generative', as data vectors $\mathbf{t}$ may be generated by sampling from the $\mathbf{x}$ and $\boldsymbol{\epsilon}$ distributions and applying (1).

2.2 Factor Analysis

Perhaps the most common example of a latent variable model is that of statistical *factor analysis* (Bartholomew 1987), in which the mapping $\mathbf{y}(\mathbf{x};\mathbf{w})$ is a linear function of $\mathbf{x}$:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}. \tag{2}$$

Conventionally, the latent variables are defined to be independent and Gaussian with unit variance, so $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The noise model is also Gaussian such that $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi}$ diagonal, and the $(d \times q)$ parameter matrix $\mathbf{W}$ contains the *factor loadings*. The parameter $\boldsymbol{\mu}$ permits the data model to have non-zero mean. Given this formulation, the observation vectors are also normally distributed $\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$, where the model covariance is $\mathbf{C} = \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}$. (Note that as a result of this parameterisation $\mathbf{C}$ is invariant under post-multiplication of $\mathbf{W}$ by an orthogonal matrix, equivalent to a rotation of the $\mathbf{x}$ co-ordinate system.) The key motivation for this model is that, because of the diagonality of $\boldsymbol{\Psi}$, the observed variables $\mathbf{t}$ are *conditionally independent* given the latent variables, or factors, $\mathbf{x}$. The intention is that the dependencies between the data variables $\mathbf{t}$ are explained by a smaller number of latent variables $\mathbf{x}$, while $\boldsymbol{\epsilon}$ represents variance unique to each observation variable. This is in contrast to conventional PCA, which effectively treats both variance and covariance identically. It should also be noted that there is no closed-form analytic solution for $\mathbf{W}$ and $\boldsymbol{\Psi}$, and so their values must be determined by iterative procedures.

2.3 Links from Factor Analysis to PCA

In factor analysis the subspace defined by the columns of $\mathbf{W}$ will generally *not* correspond to the principal subspace of the data. Nevertheless, certain links between the two methods have previously been noted. For instance, it has been observed that the factor loadings and the principal axes are quite similar in situations where the estimates of the elements of $\boldsymbol{\Psi}$ turn out to be approximately equal (e.g. Rao 1955). Indeed, this is an implied result of the fact that if $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$ and an isotropic, rather than diagonal, noise model is assumed, then PCA emerges if the $d - q$ smallest eigenvalues of the sample covariance matrix $\mathbf{S}$ are exactly equal. This "homoscedastic residuals model" is considered by Basilevsky (1994, p361), for the case where the model covariance is identical to its data sample counterpart. Given this restriction, the factor loadings $\mathbf{W}$ and noise variance σ^2 are identifiable (assuming correct choice of q) and can be determined analytically through eigen-decomposition of $\mathbf{S}$, without resort to iteration (Anderson 1963).

However, this established link with PCA requires that the $d - q$ minor eigenvalues of the sample covariance matrix be equal (or more trivially, be negligible) and thus implies that the covariance model must be *exact*. Not only is this assumption rarely justified in practice, but when exploiting PCA for dimensionality reduction, we do not require an exact characterisation of the covariance structure in the minor subspace, as this information is effectively 'discarded'. In truth, what is of real interest in the homoscedastic residuals model is the form of the maximum-likelihood solution when the model covariance is *not* identical to its data sample counterpart.

Importantly, we show in the following section that PCA does still emerge in the case of an approximate model. In fact, this link between factor analysis and PCA had been partially explored in the early factor analysis literature by Lawley (1953) and Anderson and Rubin (1956). Those authors showed that the maximum-likelihood solution in the approximate case was related to the eigenvectors of the sample covariance matrix, but did not show that these were the *principal* eigenvectors but instead made this additional assumption. In the next section (and Appendix A) we extend this earlier work to give a full characterisation of the properties of the model we term "probabilistic PCA". Specifically, with $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$, the columns of the maximum-likelihood estimator $\mathbf{W}_{\mathrm{ML}}$ are shown to span the principal subspace of the data even when $\mathbf{C} \neq \mathbf{S}$.

3 Probabilistic PCA

3.1 The Probability Model

For the case of isotropic noise $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$, equation (2) implies a probability distribution over $\mathbf{t}$-space for a given $\mathbf{x}$ of the form

$$p(\mathbf{t}|\mathbf{x}) = (2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{W}\mathbf{x} - \boldsymbol{\mu}\|^2\right\}. \tag{3}$$

With a Gaussian prior over the latent variables defined by

$$p(\mathbf{x}) = (2\pi)^{-q/2} \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}\right\}, \tag{4}$$

we obtain the marginal distribution of $\mathbf{t}$ in the form

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x}, \tag{5}$$

$$= (2\pi)^{-d/2}|\mathbf{C}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{t} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{t} - \boldsymbol{\mu})\right\}, \tag{6}$$

where the model covariance is

$$\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}}. \tag{7}$$

Using Bayes' rule, the *posterior* distribution of the latent variables $\mathbf{x}$ given the observed $\mathbf{t}$ may be calculated:

$$\begin{aligned} p(\mathbf{x}|\mathbf{t}) \quad = \quad & (2\pi)^{-q/2}|\sigma^{-2}\mathbf{M}|^{1/2} \times \\ & \exp\left[-\frac{1}{2}\left\{\mathbf{x} - \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu})\right\}^{\mathrm{T}}(\sigma^{-2}\mathbf{M})\left\{\mathbf{x} - \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu})\right\}\right], \end{aligned} \tag{8}$$

where the posterior covariance matrix is given by

$$\sigma^2\mathbf{M}^{-1} = \sigma^2(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}. \tag{9}$$

Note that $\mathbf{M}$ is $q \times q$ while $\mathbf{C}$ is $d \times d$.

The log-likelihood of observing the data under this model is

$$\begin{aligned} \mathcal{L} &= \sum_{n=1}^{N} \ln\left\{p(\mathbf{t}_n)\right\}, \\ &= -\frac{N}{2}\left\{d\ln(2\pi) + \ln|\mathbf{C}| + \mathrm{tr}\left(\mathbf{C}^{-1}\mathbf{S}\right)\right\}, \end{aligned} \tag{10}$$

where

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}, \tag{11}$$

is the sample covariance matrix of the observed $\{\mathbf{t}_n\}$.

3.2 Properties of the Maximum-Likelihood Estimators

It is easily seen that the maximum-likelihood estimate of the parameter $\boldsymbol{\mu}$ is given by the mean of the data:

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{t}_n. \tag{12}$$

We now consider the maximum–likelihood estimators for the parameters $\mathbf{W}$ and σ^2.

3.2.1 The Weight Matrix W

The log-likelihood (10) is maximized when the columns of $\mathbf{W}$ span the principal subspace of the data. To show this we consider the derivative of (10) with respect to $\mathbf{W}$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}). \tag{13}$$

In Appendix A it is shown that, with $\mathbf{C}$ given by (7), the only non-zero stationary points of (13) occur for:

$$\mathbf{W} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{14}$$

where the q column vectors in the $d \times q$ matrix $\mathbf{U}_q$ are eigenvectors of $\mathbf{S}$, with corresponding eigenvalues in the $q \times q$ diagonal matrix $\mathbf{\Lambda}_q$, and $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal rotation matrix. Furthermore, it is also shown that the stationary point corresponding to the *global maximum* of the likelihood occurs when $\mathbf{U}_q$ comprises the *principal* eigenvectors of $\mathbf{S}$, and thus $\mathbf{\Lambda}_q$ contains the corresponding eigenvalues $\lambda_1, \ldots, \lambda_q$, where the eigenvalues of $\mathbf{S}$ are indexed in order of decreasing magnitude. All other combinations of eigenvectors represent saddle-points of the likelihood surface. Thus, from (14), the latent variable model defined by equation (2) effects a mapping from the latent space into the *principal subspace* of the observed data.

3.2.2 The Noise Variance σ^2

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

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q}\sum_{j=q+1}^{d} \lambda_j, \tag{15}$$

where $\lambda_{q+1}, \ldots, \lambda_d$ are the smallest eigenvalues of $\mathbf{S}$ and so σ^2_{ML} has a clear interpretation as the average variance 'lost' per discarded dimension.

3.3 Dimensionality Reduction and Optimal Reconstruction

To implement probabilistic PCA, we would generally first compute the usual eigen-decomposition of $\mathbf{S}$ (we consider an alternative, iterative, approach shortly), after which σ^2_{ML} is found from (15) followed by $\mathbf{W}_{\mathrm{ML}}$ from (14). This is then sufficient to define the associated density model for PCA, allowing the advantages listed in Section 1 to be exploited.

In conventional PCA, the reduced-dimensionality transformation of a data point $\mathbf{t}_n$ is given by $\mathbf{x}_n = \mathbf{U}_q^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$ and its reconstruction by $\hat{\mathbf{t}}_n = \mathbf{U}_q\mathbf{x}_n + \boldsymbol{\mu}$. This may be similarly achieved within the PPCA formulation. However, we note that in the probabilistic framework, the generative model defined by equation (2) represents a mapping *from* the lower-dimensional latent space *to* the data space. So, in PPCA, the probabilistic analogue of the dimensionality reduction process of conventional PCA would be to invert the conditional distribution $p(\mathbf{t}|\mathbf{x})$ using Bayes' rule, in equation (8), to give $p(\mathbf{x}|\mathbf{t})$. In this case, each data point $\mathbf{t}_n$ is represented in the latent space not by a single vector, but by the Gaussian *posterior distribution* defined by (8). As an alternative to the standard PCA projection then, a convenient summary of this distribution and representation of $\mathbf{t}_n$ would be the *posterior mean* $\langle \mathbf{x}_n \rangle = \mathbf{M}^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$, a quantity that also arises naturally in (and is computed in) the EM implementation of PPCA considered in Section 3.4. Note, also from (8), that the *covariance* of the posterior distribution is given by $\sigma^2\mathbf{M}^{-1}$ and is therefore constant for all data points.

However, perhaps counter-intuitively given equation (2), $\mathbf{W}_{\mathrm{ML}}\langle \mathbf{x}_n \rangle + \boldsymbol{\mu}$ is *not* the optimal linear reconstruction of $\mathbf{t}_n$. This may be seen from the fact that, for $\sigma^2 > 0$, $\mathbf{W}_{\mathrm{ML}}\langle \mathbf{x}_n \rangle + \boldsymbol{\mu}$ is not an

orthogonal projection of $\mathbf{t}_n$, as a consequence of the Gaussian prior over $\mathbf{x}$ causing the posterior mean projection to become skewed towards the origin. If we consider the limit as $\sigma^2 \to 0$, the projection $\mathbf{W}_{\mathrm{ML}}\langle\mathbf{x}_n\rangle = \mathbf{W}_{\mathrm{ML}}(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}})^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$ does become orthogonal and is equivalent to conventional PCA, but then the density model is singular and thus undefined.

Taking this limit is not necessary however, since the optimal least-squares linear reconstruction of the data from the posterior mean vectors $\langle\mathbf{x}_n\rangle$ may be obtained from (see Appendix B):

$$\hat{\mathbf{t}}_n = \mathbf{W}_{\mathrm{ML}}\left(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}\right)^{-1}\mathbf{M}\langle\mathbf{x}_n\rangle + \boldsymbol{\mu}, \tag{16}$$

with identical reconstruction error to conventional PCA.

For reasons of probabilistic elegance therefore, we might choose to exploit the posterior mean vectors $\langle\mathbf{x}_n\rangle$ as the reduced-dimensionality representation of the data, although there is no material benefit in so doing. Indeed, we note that in addition to the conventional PCA representation $\mathbf{U}_q^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$, the vectors $\hat{\mathbf{x}}_n = \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$ could equally be used without loss of information, and reconstructed using $\hat{\mathbf{t}}_n = \mathbf{W}_{\mathrm{ML}}\left(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}\right)^{-1}\hat{\mathbf{x}}_n + \boldsymbol{\mu}$.

3.4 An EM Algorithm For PPCA

By a simple extension of the EM formulation for parameter estimation in the standard linear factor analysis model (Rubin and Thayer 1982), we can obtain a principal component projection by maximizing the likelihood function (10). We are not suggesting that such an approach necessarily be adopted for probabilistic PCA — normally the principal axes would be estimated in the conventional manner, via eigen-decomposition of $\mathbf{S}$, and subsequently incorporated in the probability model using equations (14) and (15) to realise the advantages outlined in the introduction. However, as discussed in Appendix A.5, there may be an advantage in the EM approach for large d since the presented algorithm, although iterative, requires neither computation of the $d \times d$ covariance matrix, which is $O(Nd^2)$, nor its explicit eigen-decomposition, which is $O(d^3)$. We derive the EM algorithm and consider its properties from the computational perspective in Appendix A.5.

3.5 Factor Analysis Revisited

The probabilistic PCA algorithm was obtained by introducing a constraint into the noise matrix of the factor analysis latent variable model. This apparently minor modification leads to significant differences in the behaviour of the two methods. In particular, we now show that the covariance properties of the PPCA model are identical to those of conventional PCA, and are quite different from those of standard factor analysis.

Consider a non-singular linear transformation of the data variables, so that $\mathbf{t} \to \mathbf{A}\mathbf{t}$. Using (12) we see that under such a transformation the maximum likelihood solution for the mean will be transformed as $\boldsymbol{\mu}_{\mathrm{ML}} \to \mathbf{A}\boldsymbol{\mu}_{\mathrm{ML}}$. From (11) it then follows that the covariance matrix will transform as $\mathbf{S} \to \mathbf{A}\mathbf{S}\mathbf{A}^{\mathrm{T}}$.

The log-likelihood for the latent variable model, from (10), is given by

$$\mathcal{L}(\mathbf{W}, \boldsymbol{\Psi}) = -\frac{N}{2}\left\{d\ln(2\pi) + \ln|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}| + \ \mathrm{tr}\ \left[(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi})^{-1}\mathbf{S}\right]\right\} \tag{17}$$

where $\boldsymbol{\Psi}$ is a general noise covariance matrix. Thus, using (17), we see that under the transformation $\mathbf{t} \to \mathbf{A}\mathbf{t}$ the log likelihood will transform as

$$\mathcal{L}(\mathbf{W}, \boldsymbol{\Psi}) \to \mathcal{L}(\mathbf{A}^{-1}\mathbf{W}, \mathbf{A}^{-1}\boldsymbol{\Psi}\mathbf{A}^{-\mathrm{T}}) - N\ln|\mathbf{A}| \tag{18}$$

where $\mathbf{A}^{-\mathrm{T}} \equiv (\mathbf{A}^{-1})^{\mathrm{T}}$. Thus if $\mathbf{W}_{\mathrm{ML}}$ and $\mathbf{\Psi}_{\mathrm{ML}}$ are maximum likelihood solutions for the original data, then $\mathbf{A}\mathbf{W}_{\mathrm{ML}}$ and $\mathbf{A}\mathbf{\Psi}_{\mathrm{ML}}\mathbf{A}^{\mathrm{T}}$ will be maximum likelihood solutions for the transformed data set.

In general the form of the solution will not be preserved under such a transformation. However, we can consider two special cases. First, suppose $\mathbf{\Psi}$ is a diagonal matrix, corresponding to the case of factor analysis. Then $\mathbf{\Psi}$ will remain diagonal provided $\mathbf{A}$ is also a diagonal matrix. This says that factor analysis is *covariant* under component-wise rescaling of the data variables: the scale factors simply become absorbed into rescaling of the noise variances, and the rows of $\mathbf{W}$ are rescaled by the same factors. Second, consider the case $\mathbf{\Psi} = \sigma^2\mathbf{I}$, corresponding to PPCA. Then the transformed noise covariance $\sigma^2\mathbf{A}\mathbf{A}^{\mathrm{T}}$ will only be proportional to the unit matrix if $\mathbf{A}^{\mathrm{T}} = \mathbf{A}^{-1}$, in other words if $\mathbf{A}$ is an orthogonal matrix. Transformation of the data vectors by multiplication with an orthogonal matrix corresponds to a rotation of the coordinate system. This same covariance property is shared by standard non-probabilistic PCA since a rotation of the coordinates induces a corresponding rotation of the principal axes. Thus we see that factor analysis is covariant under component-wise rescaling, while PPCA and PCA are covariant under rotations, as illustrated in Figure 1.

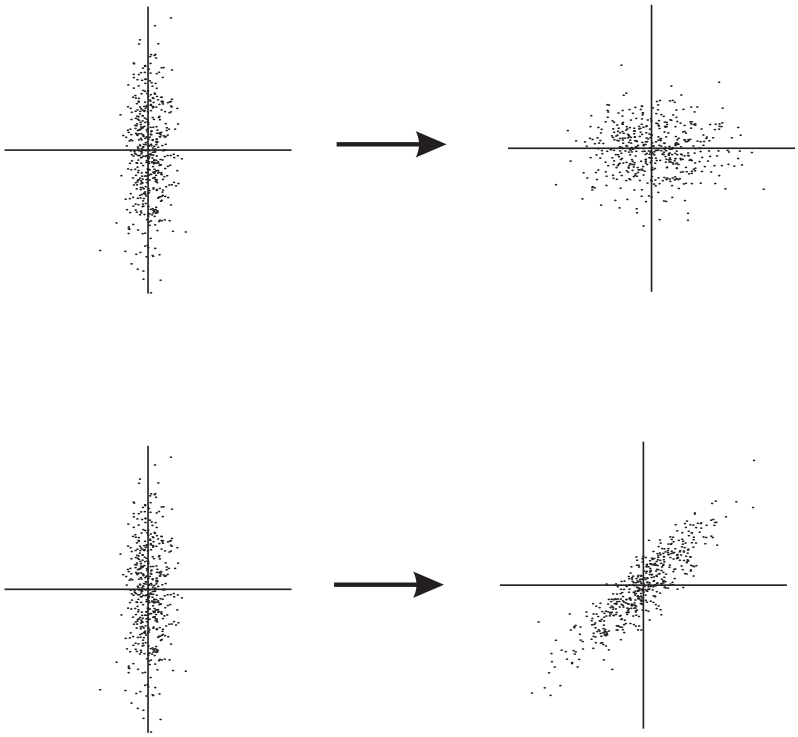

Figure 1: Factor analysis is covariant under a component-wise rescaling of the data variables (top plots) while PCA and probabilistic PCA are covariant under rotations of the data space coordinates (bottom plots).

4 Mixtures of Probabilistic Principal Component Analysers

The association of a probability model with PCA offers the tempting prospect of being able to model complex data structures with a combination of local PCA models through the mechanism

of a mixture of probabilistic principal component analysers (Tipping and Bishop 1997). This formulation would permit all of the model parameters to be determined from maximum-likelihood, where both the appropriate partitioning of the data and the determination of the respective principal axes occur automatically as the likelihood is maximized. The log-likelihood of observing the data set for such a mixture model is:

$$\mathcal{L} = \sum_{n=1}^{N} \ln \{p(\mathbf{t}_n)\}, \tag{19}$$

$$= \sum_{n=1}^{N} \ln \left\{ \sum_{i=1}^{M} \pi_i p(\mathbf{t}_n|i) \right\}, \tag{20}$$

where $p(\mathbf{t}|i)$ is a single PPCA model and π_i is the corresponding mixing proportion, with $\pi_i \geq 0$ and $\sum \pi_i = 1$. Note that a separate mean vector $\boldsymbol{\mu}_i$ is now associated with each of the M mixture components, along with the parameters $\mathbf{W}_i$ and σ_i^2. A related model has recently been exploited for data visualization (Bishop and Tipping 1998), while a similar approach, based on the standard factor analysis diagonal ($\boldsymbol{\Psi}$) noise model, has been employed for handwritten digit recognition (Hinton *et al.* 1997), although it does not implement PCA.

The corresponding generative model for the mixture case now requires the random choice of a mixture component according to the proportions π_i, followed by sampling from the $\mathbf{x}$ and $\boldsymbol{\epsilon}$ distributions and applying equation (2) as in the single model case, taking care to utilise the appropriate parameters $\boldsymbol{\mu}_i$, $\mathbf{W}_i$ and σ_i^2. Furthermore, for a given data point $\mathbf{t}$, there is now a posterior distribution associated with each latent space, the mean of which for space i is given by $(\sigma_i^2\mathbf{I} + \mathbf{W}_i^{\mathrm{T}}\mathbf{W}_i)^{-1}\mathbf{W}_i^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu}_i)$.

We can develop an iterative EM algorithm for optimization of all of the model parameters π_i, $\boldsymbol{\mu}_i$, $\mathbf{W}_i$ and σ_i^2. If $R_{ni} = p(i|\mathbf{t}_n)$ is the posterior *responsibility* of mixture i for generating data point $\mathbf{t}_n$, given by

$$R_{ni} = \frac{p(\mathbf{t}_n|i)\pi_i}{p(\mathbf{t}_n)}, \tag{21}$$

then in Appendix C it is shown that we obtain the following parameter updates:

$$\widetilde{\pi}_i = \frac{1}{N} \sum_{n=1}^{N} R_{ni}, \tag{22}$$

$$\widetilde{\boldsymbol{\mu}}_i = \frac{\sum_{n=1}^{N} R_{ni}\mathbf{t}_n}{\sum_{n=1}^{N} R_{ni}}. \tag{23}$$

Thus the updates for $\widetilde{\pi}_i$ and $\widetilde{\boldsymbol{\mu}}_i$ correspond exactly to those of a standard Gaussian mixture formulation (e.g. see Bishop 1995). Furthermore, in Appendix C, it is also shown that the combination of the E- and M-steps leads to the intuitive result that the axes $\mathbf{W}_i$ and the noise variance σ_i^2 are determined from the *local responsibility-weighted* covariance matrix:

$$\mathbf{S}_i = \frac{1}{\widetilde{\pi}_i N} \sum_{n=1}^{N} R_{ni}(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)^{\mathrm{T}}, \tag{24}$$

by standard eigen-decomposition in exactly the same manner as for a single PPCA model. However, as noted earlier in Section 3.4 (and also in Appendix A.5), for larger values of data dimensionality d, computational advantages can be obtained if $\mathbf{W}_i$ and σ_i^2 are updated iteratively according to an EM schedule. This is discussed for the mixture model in Appendix C.

Iteration of equations (21), (22) and (23) in sequence followed by computation of $\mathbf{W}_i$ and σ_i^2, either from equation (24) using (14) and (15) or using the iterative updates in Appendix C, is guaranteed to find a local maximum of the log-likelihood (19). At convergence of the algorithm each weight matrix $\mathbf{W}_i$ spans the principal subspace of its respective $\mathbf{S}_i$.

In the next section we consider applications of this PPCA mixture model, beginning with data compression and reconstruction tasks. We then consider general density modelling in Section 6.

5 Local Linear Dimensionality Reduction

In this section we begin by giving an illustration of the application of the PPCA mixture algorithm to a synthetic data set. More realistic examples are then considered, with an emphasis on cases in which a principal component approach is motivated by the objective of deriving a reduced-dimensionality representation of the data which can be reconstructed with minimum error. We will therefore contrast the clustering mechanism in the PPCA mixture model with that of a hard clustering approach based explicitly on reconstruction error as utilised in a typical algorithm.

5.1 Illustration for Synthetic Data

For a demonstration of the mixture of PPCA algorithm, we generated a synthetic dataset comprising 500 data points sampled uniformly over the surface of a hemisphere, with additive Gaussian noise. Figure 2(a) shows this data.

A mixture of 12 probabilistic principal component analysers was then fitted to the data using the EM algorithm outlined in the previous section, with latent space dimensionality $q = 2$. Because of the probabilistic formalism, a generative model of the data is defined and we emphasise this by plotting a second set of 500 data points, obtained by sampling from the fitted generative model. These data points are shown in Figure 2(b). Histograms of the distances of all the data points from the hemisphere are also given to indicate more clearly the accuracy of the model in capturing the structure of the underlying generator.

5.2 Clustering Mechanisms

Generating a local PCA model of the form illustrated above is often prompted by the ultimate goal of accurate data reconstruction. Indeed, this has motivated Kambhatla and Leen (1997) and Hinton *et al.* (1997) to utilise squared reconstruction error as the clustering criterion in the partitioning phase. Dony and Haykin (1995) adopt a similar approach to image compression, although their model has no set of independent 'mean' parameters $\boldsymbol{\mu}_i$. Using the reconstruction criterion, a data point is assigned to the component that reconstructs it with lowest error, and the principal axes are then re-estimated within each cluster. For the mixture of PPCA model, however, data points are assigned to mixture components (in a soft fashion) according to the responsibility R_{ni} of the mixture component for its generation. Since, $R_{ni} = p(\mathbf{t}_n|i)\pi_i/p(\mathbf{t}_n)$ and $p(\mathbf{t}_n)$ is constant for all components, $R_{ni} \propto p(\mathbf{t}_n|i)$ and we may gain further insight into the clustering by considering the probability density associated with component i at data point $\mathbf{t}_n$:

$$p(\mathbf{t}_n|i) = (2\pi)^{-d/2}|\mathbf{C}_i|^{-1/2} \exp\left\{-E_{ni}^2/2\right\}, \tag{25}$$

where

$$E_{ni}^2 = (\mathbf{t}_n - \boldsymbol{\mu}_i)^{\mathrm{T}}\mathbf{C}^{-1}{}_i(\mathbf{t}_n - \boldsymbol{\mu}_i), \tag{26}$$

$$\mathbf{C}_i = \sigma_i^2\mathbf{I} + \mathbf{W}_i\mathbf{W}_i^{\mathrm{T}}. \tag{27}$$

It is helpful to express the matrix $\mathbf{W}_i$ in terms of its singular value decomposition (and although we are considering an individual mixture component i, the i subscript will be omitted for notational clarity):

$$\mathbf{W} = \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{28}$$

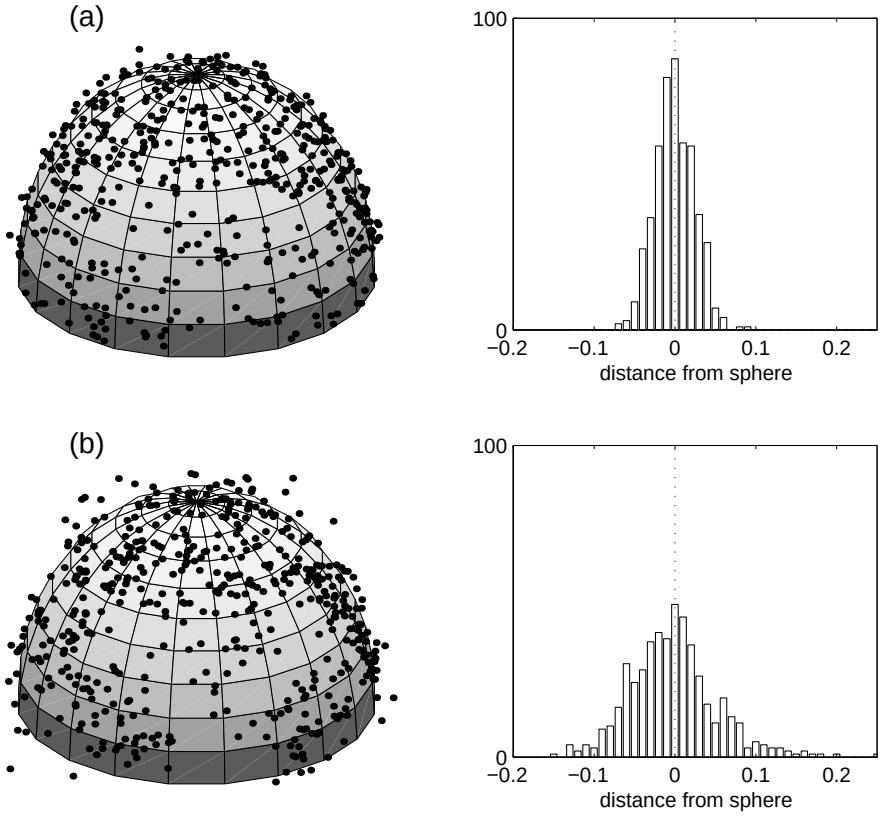

Figure 2: Modelling noisy data on a hemisphere. (a) On the left, the synthetic data, on the right, a histogram of the Euclidean distances of each data point to the sphere. (b) Similarly for data generated from the fitted PPCA mixture model.

where $\mathbf{U}_q$ is a $d \times q$ matrix of orthonormal column vectors and $\mathbf{R}$ is an arbitrary $q \times q$ orthogonal matrix. The singular values are parameterised, without loss of generality, in terms of $(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}$, where $\mathbf{K}_q = \mathrm{diag}(k_1, k_2, \ldots, k_q)$ is a $q \times q$ diagonal matrix. Then

$$E_n^2 = (\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}} \left\{\sigma^2\mathbf{I} + \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})\mathbf{U}_q^{\mathrm{T}}\right\}^{-1} (\mathbf{t}_n - \boldsymbol{\mu}). \tag{29}$$

The data point $\mathbf{t}_n$ may also be expressed in terms of the basis of vectors $\mathbf{U} = (\mathbf{U}_q, \mathbf{U}_{d-q})$, where $\mathbf{U}_{d-q}$ comprises $(d-q)$ vectors perpendicular to $\mathbf{U}_q$ which complete an orthonormal set. In this basis, we define $\mathbf{z}_n = \mathbf{U}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$ and so $\mathbf{t}_n - \boldsymbol{\mu} = \mathbf{U}\mathbf{z}_n$, from which (29) may then be written as

$$E_n^2 = \mathbf{z}_n^{\mathrm{T}}\mathbf{U}^{\mathrm{T}} \left\{\sigma^2\mathbf{I} + \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})\mathbf{U}_q^{\mathrm{T}}\right\}^{-1} \mathbf{U}\mathbf{z}_n, \tag{30}$$

$$= \mathbf{z}_n^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{z}_n, \tag{31}$$

where $\mathbf{D} = \mathrm{diag}(k_1, k_2, \ldots, k_q, \sigma^2, \ldots, \sigma^2)$ is a $d \times d$ diagonal matrix. Thus:

$$E_n^2 = \mathbf{z}_{\mathrm{in}}^{\mathrm{T}}\mathbf{K}^{-1}{}_q\mathbf{z}_{\mathrm{in}} + \frac{\mathbf{z}_{\mathrm{out}}^{\mathrm{T}}\mathbf{z}_{\mathrm{out}}}{\sigma^2}, \tag{32}$$

$$= E_{\mathrm{in}}^2 + E_{\mathrm{rec}}^2/\sigma^2, \tag{33}$$

where we have partitioned the elements of $\mathbf{z}$ into $\mathbf{z}_{\mathrm{in}}$, the projection of $\mathbf{t}_n - \boldsymbol{\mu}$ onto the subspace spanned by $\mathbf{W}$, and $\mathbf{z}_{\mathrm{out}}$, the projection onto the corresponding perpendicular subspace. Thus E_{rec}^2 is the squared reconstruction error and E_{in}^2 may be interpreted as an 'in-subspace' error term. Note that at the maximum-likelihood solution, $\mathbf{U}_q$ is the matrix of eigenvectors of the local covariance matrix and $\mathbf{K}_q = \boldsymbol{\Lambda}_q$.

As $\sigma_i^2 \to 0$, $R_{ni} \propto \pi_i \exp\left(-E_{rec}^2/2\right)$ and, for equal prior probabilities, cluster assignments are equivalent to a soft reconstruction-based clustering. However, for $\sigma_A^2, \sigma_B^2 > 0$, consider a data point which lies in the subspace of a relatively distant component A, which may be reconstructed with zero error, yet which lies more closely to the mean of a second component B. The effect of the noise variance σ_B^2 in (33) is to moderate the contribution of E_{rec}^2 for component B. As a result, the data point may be assigned to the nearer component B even though the reconstruction error is considerably greater, given that it is sufficiently distant from the mean of A such that E_{in}^2 for A is large.

It should be expected, then, that mixture of PPCA clustering would result in more localised clusters, but with final reconstruction error inferior to that of a clustering model based explicitly on a reconstruction criterion. Conversely, it should also be clear that clustering the data according to the proximity to the subspace alone will not necessarily result in localised partitions (as noted by Kambhatla (1995), who also considers the relationship of such an algorithm to a probabilistic model). That this is so is simply illustrated in Figure 3, in which a collection of 12 conventional PCA models have been fitted to the hemisphere data, according to the 'VQPCA' (Vector-Quantization PCA) algorithm of Kambhatla and Leen (1997) which is defined as:

1. Select initial cluster centres $\boldsymbol{\mu}_i$ at random from points in the dataset, and assign all data points to the nearest (in terms of Euclidean distance) cluster centre.
2. Set the $\mathbf{W}_i$ vectors to the first two principal axes of the covariance matrix of cluster i.
3. Assign data points to the cluster which best reconstructs them, setting each $\boldsymbol{\mu}_i$ to the mean of those data points assigned to cluster i.
4. Repeat from 2 until the cluster allocations are constant.

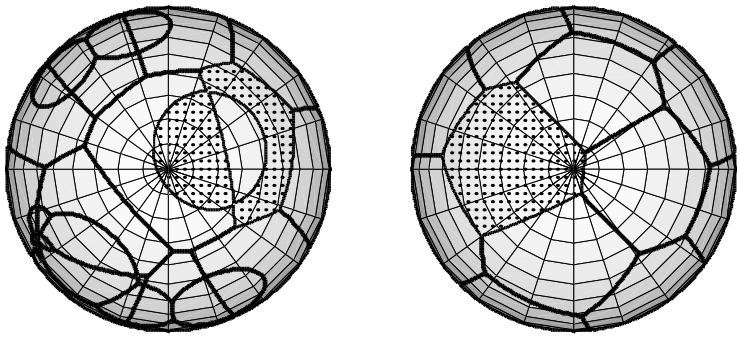

Figure 3: Comparison of the partitioning of the hemisphere effected by a VQPCA-based model (left) and a PPCA mixture model (right). The illustrated boundaries delineate regions of the hemisphere that are best reconstructed by a particular local PCA model. One such region is shown shaded to emphasize that clustering according to reconstruction error results in a non-localised partitioning. In the VQPCA case, the circular effects occur when principal component planes intersect beneath the surface of the hemisphere.

In Figure 3, data points have been sampled over the hemisphere, without noise, and allocated to the cluster which best reconstructs them. The left plot shows the partitioning associated with the best (i.e. lowest reconstruction error) model obtained from 100 runs of the VQPCA algorithm. The right plot shows a similar partitioning for the best (i.e. greatest likelihood) PPCA mixture model using the same number of components, again from 100 runs. Note that the boundaries illustrated in this latter plot were obtained using assignments based on reconstruction error for the final model, in identical fashion to the VQPCA case, and not on probabilistic responsibility. We see that the partitions formed when clustering according to reconstruction error alone can

be non-local, as exemplified by the shaded component. This phenomenon is rather contrary to the philosophy of *local* dimensionality reduction and is an indirect consequence of the fact that reconstruction-based local PCA does not model the data in a probabilistic sense.

However, we might expect that algorithms such as VQPCA should offer better performance in terms of the reconstruction error of the final solution, having been designed explicitly to optimize that measure. In order to test this, we compared the VQPCA algorithm with the PPCA mixture model on six data sets, detailed in Table 1.

Data Set	N	d	M	q	**Description**
Hemisphere	500	3	12	2	Synthetic data used above.
Oil	500	12	12	2	Diagnostic measurements from oil pipeline flows.
Digit_1	500	64	10	10	8×8 gray-scale images of handwritten digit '1'.
Digit_2	500	64	10	10	8×8 gray-scale images of handwritten digit '2'.
Image	500	64	8	4	8×8 gray-scale blocks from a photographic image.
EEG	300	30	8	5	Delay vectors from an EEG time series signal.

Table 1: Datasets used for comparison of clustering criteria.

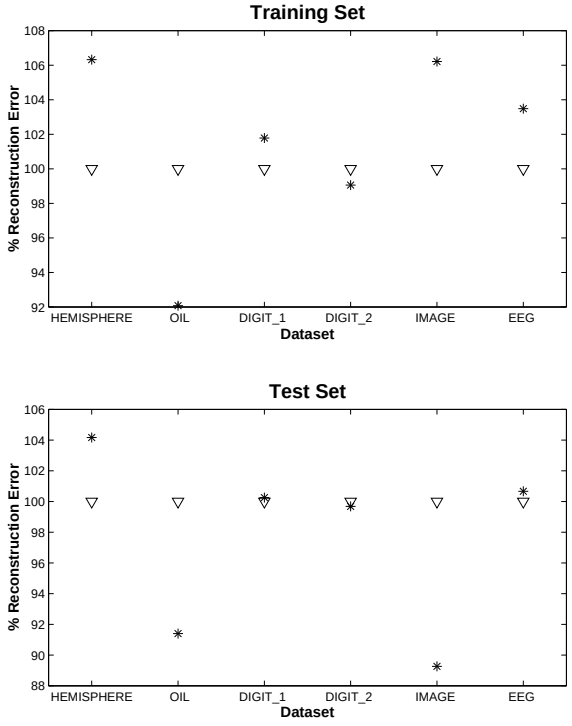

Figure 4: Reconstruction errors for reconstruction-based local PCA (VQPCA) and the PPCA mixture. Errors for the latter ($*$) have been shown relative to the former ($\bigtriangledown$), and are averaged over 100 runs with random initial configurations.

Figure 4 summarises the reconstruction error of the respective models and in general, VQPCA performs better as expected. However, we also note two interesting aspects of the results.

First, in the case of the 'oil' data, the final reconstruction error of the PPCA model on both training and test sets is counter-intuitively superior, despite the fact that the partitioning of the data space was based only partially on reconstruction error. This behaviour is, we hypothesize, a result of the particular structure of that dataset. The oil data is known to comprise a number of disjoint, but locally smooth, two-dimensional cluster structures (see Bishop and Tipping 1998 for a visualization thereof).

For the oil dataset, we observed that many of the models found by the VQPCA algorithm exhibit partitions that are not only often non-connected (similar to those shown for the hemisphere in Figure 3) but which may also span more than one of the disjoint cluster structures. The evidence of Figure 4 suggests that these models represent poor local minima of the reconstruction error cost function. The PPCA mixture algorithm does not find such sub-optimal solutions, which would have low likelihood due to the locality implied by the density model. The experiment indicates that by avoiding these poor solutions, the PPCA mixture model is able to find solutions with lower reconstruction error (on average) than VQPCA.

These observations only apply to the case of the oil dataset. For the hemisphere, digit '1', image and EEG training sets, the data manifolds are less disjoint and the explicit reconstruction-based algorithm, VQPCA, is superior. For the digit '2' case, the two algorithms appear approximately equivalent.

A second aspect of Figure 4 is the suggestion that the PPCA mixture model algorithm may be less sensitive to over-fitting. As would be expected, compared with the training set, errors on the test set increase for both algorithms (although, because the errors have been normalised to allow comparisons between datasets, this isn't shown in Figure 4). However, with the exception of the case of the digit '2' data set, for the PPCA mixture model this increase is proportionately smaller than for VQPCA. This effect is most dramatic for the image data set, where PPCA is much superior on the test set. For that dataset, the test examples were derived from a separate portion of the image (see below), and as such, the test set statistics can be expected to differ more significantly from the respective training set than for the other examples.

A likely explanation for this is that, because of the 'soft' clustering of the PPCA mixture model, there is an inherent 'smoothing' effect occurring when estimating the local sets of principal axes. Each set of axes is determined from its corresponding local responsibility-weighted covariance matrix which in general will be influenced by many data points, not just the subset that would be associated with the cluster in a 'hard' implementation. Because of this, the parameters in the $\mathbf{W}_i$ matrix in cluster i are also constrained by data points in neighbouring clusters ($j \neq i$) to some extent. This notion is discussed in the context of regression by Jordan and Jacobs (1994) as motivation for their 'mixture of experts' model, where the authors note how soft-partitioning can reduce variance (in terms of the bias-variance decomposition). Although it is difficult to draw firm conclusions from this limited set of experiments, the evidence of Figure 4 does point to the presence of such an effect.

5.3 Application: Image Compression

As a practical example, we consider an application of the PPCA mixture model to block transform image coding. Figure 5 shows the original image. This 720×360 pixel image was segmented into 8×8 non-overlapping blocks, giving a total dataset of 4050 64-dimensional vectors. Half of this data, corresponding to the left half of the picture, was used as training data. The right half was reserved for testing, and a magnified portion of the test image is also shown in Figure 5. A reconstruction of the entire image based on the first four principal components of a *single* PCA model determined from the block-transformed left half of the image is shown in Figure 6.

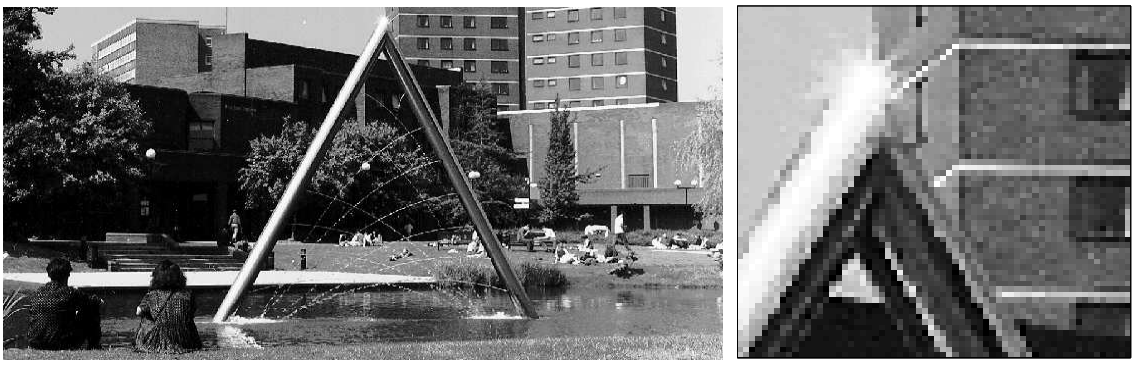

Figure 5: The original image (left), and detail therein (right).

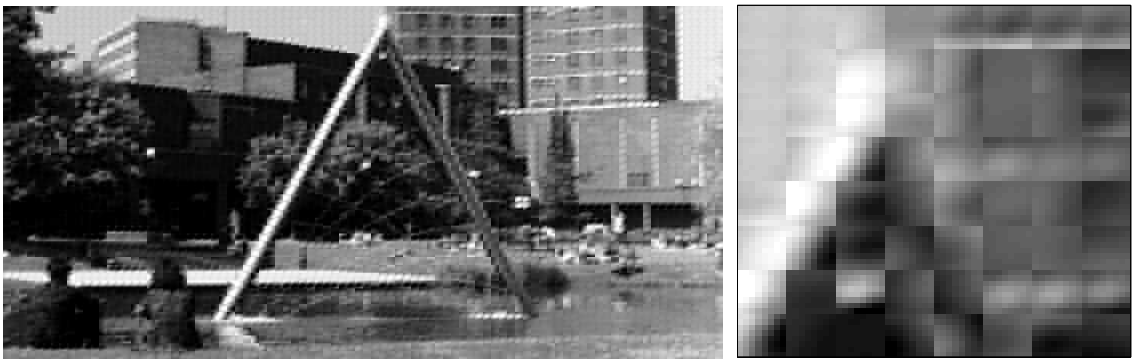

Figure 6: The PCA reconstructed image, at 0.5 bits-per-pixel.

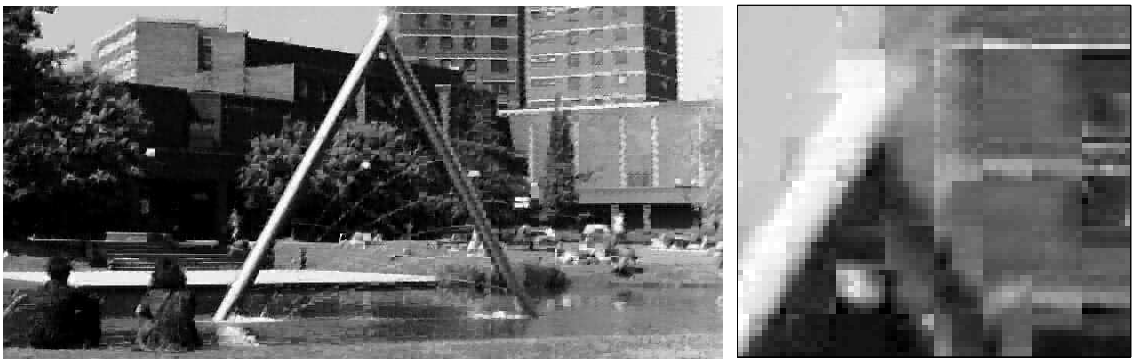

Figure 7: The mixture of PPCA reconstructed image, using the same bit-rate as Figure 6.

Figure 7 shows the reconstruction of the original image when modelled by a mixture of probabilistic principal component analysers. The model parameters were estimated using only the left half of the image. In this example, 12 components were used, of dimensionality 4, and after the model likelihood had been maximized, the image coding was performed in a 'hard' fashion — i.e. by allocating data to the component with lowest reconstruction error. The resulting coded image was uniformly quantised, with bits allocated equally to each transform variable, before reconstruction in order to give a final bit-rate of 0.5 bits-per-pixel (and thus compression of 16:1) in both Figures 6 and 7. In the latter case, the cost of encoding the mixture component label was included. For the simple principal subspace reconstruction, the normalised test error was 7.1×10^{-2}, while for the mixture model, it was 5.7×10^{-2}. The VQPCA algorithm gave a test error of 6.2×10^{-2}.

6 Density Modelling

A popular approach to semi-parametric density estimation is the Gaussian mixture model (Titterington, Smith, and Makov 1985). However, such models suffer from the limitation that if each Gaussian component is described by a full covariance matrix, then there are $d(d+1)/2$ independent covariance parameters to be estimated for each mixture component. Clearly, as the dimensionality of the data space increases, the number of data points required to specify those parameters reliably will become prohibitive. An alternative approach, then, is to reduce the number of parameters by placing a constraint on the form of the covariance matrix. (Another would be to introduce priors over the parameters of the full covariance matrix, as implemented by Ormoneit and Tresp (1996).) Two common constraints are either to restrict the covariance to be isotropic or to be diagonal. The isotropic model is highly constrained as it only assigns a single parameter to describe the entire covariance structure in the full d dimensions. The diagonal model is more flexible, with d parameters, but the principal axes of the elliptical Gaussians must be aligned with the data axes and thus each individual mixture component is unable to capture correlations amongst the variables.

A mixture of PPCA models, where the covariance of each Gaussian is parameterised by the relation $\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}}$, comprises $dq + 1 - q(q-1)/2$ free parameters.[1] (Note that the $q(q-1)/2$ term takes account of the number of parameters needed to specify the arbitrary rotation $\mathbf{R}$.) It thus permits the number of parameters to be controlled by the choice of q. When $q = 0$, the model is equivalent to an isotropic Gaussian. With $q = d - 1$, the general covariance Gaussian is recovered.

6.1 A Synthetic Example: Noisy Spiral Data

The utility of the PPCA mixture approach may be demonstrated with the following simple example. A 500 point data set was generated along a three-dimensional spiral configuration with added Gaussian noise. The data was then modelled by both a mixture of PPCA models and a mixture of diagonal covariance Gaussians, using 8 mixture components. In the mixture of PPCA case, $q = 1$ for each component, and so there are 4 variance parameters per component compared with 3 for the diagonal model. The results are visualised in Figure 8, which illustrates both 'side' and 'end' projections of the data.

The orientation of the ellipses in the diagonal model can be seen not to coincide with the local data structure, which is a result of the axial alignment constraint. A further consequence of the diagonal parameterisation is that the means are also implicitly constrained as they tend to lie where the tangent to the spiral is parallel to either axis of the end elevation. This qualitative superiority of the PPCA approach is underlined quantitatively by the log-likelihood per data point given in brackets in the figure. Such a result would of course be expected given that the PPCA

[1] An alternative would be a mixture of factor analysers, implemented by Hinton *et al.* (1997), although that comprises more parameters.

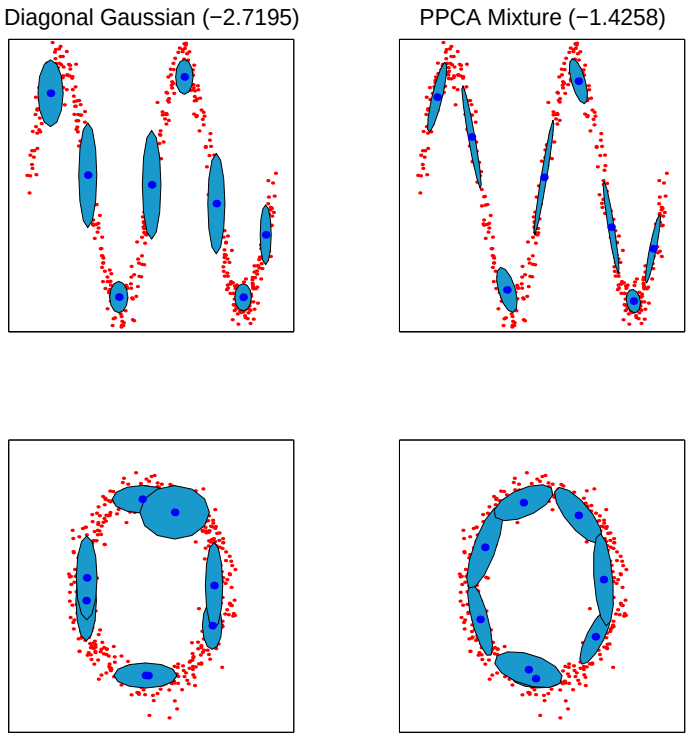

Figure 8: Comparison of an 8-component diagonal variance Gaussian mixture model with a mixture of PPCA model. The upper two plots give a view perpendicular to the major axis of the spiral, while the lower two plots show the end elevation. The covariance structure of each mixture component is shown by projection of a unit Mahalanobis distance ellipse and the log-likelihood per data-point is given in brackets above the figures.

model has an extra parameter in each mixture component, but similar results are observed if the spiral is embedded in a space of much higher dimensionality where the extra parameter in PPCA is proportionately less relevant.

It should be intuitive that the axial alignment constraint of the diagonal model is, in general, particularly *in*appropriate when modelling a smooth and continuous lower dimensional manifold in higher dimensions, regardless of the intrinsic dimensionality. Even with $q = 1$, the PPCA approach is able to track the spiral manifold successfully.

Finally, we demonstrate the importance of the use of an appropriate number of parameters by modelling a three-dimensional spiral data set of 100 data points (the number of data points was reduced to emphasise the over-fitting) as above with isotropic, diagonal and full covariance Gaussian mixture models, along with a PPCA mixture model. For each model, the log-likelihood per data point both for the training data set, and an unseen test set of 1000 data points, is given in Table 2.

As would be expected in this case of limited data, the full covariance model exhibits the best likelihood on the training set, but test set performance is worse than for the PPCA mixture. For this simple example, there is only one intermediate PPCA parameterisation with $q = 1$ ($q = 0$ and $q = 2$ are equivalent to the isotropic and full covariance cases respectively). In realistic applications, where the dimensionality d will be considerably larger, the PPCA model offers the choice of a range of q, and an appropriate value can be determined using standard techniques for

	Isotropic	Diagonal	Full	PPCA
Training	−3.14	−2.74	−1.47	−1.65
Test	−3.68	−3.43	−3.09	−2.37

Table 2: Log-likelihood per data point measured on training and test sets for Gaussian mixture models with eight components and a 100-point training set.

model selection. Finally, note that these advantages are not limited to mixture models, but may equally be exploited for the case of a single Gaussian distribution.

6.2 Application: Handwritten Digit Recognition

One potential application for high-dimensionality density models is handwritten digit recognition. Examples of gray-scale pixel images of a given digit will generally lie on a lower-dimensional smooth continuous manifold, the geometry of which is determined by properties of the digit such as rotation, scaling and thickness of stroke. One approach to the classification of such digits (although not necessarily the best) is to build a model of each digit separately, and classify unseen digits according to the model to which they are most 'similar'.

Hinton *et al.* (1997) gave an excellent discussion of the handwritten digit problem, and applied a mixture of PCA approach, using soft reconstruction-based clustering, to the classification of scaled and smoothed 8-by-8 gray-scale images taken from the CEDAR U.S. Postal Service database (Hull 1994). The models were constructed using an 11,000-digit subset of the '*br*' data set (which was further split into training and validation sets), and the '*bs*' test set was classified according to which model best reconstructed each digit (in the squared-error sense). We repeated the experiment with the same data using the PPCA mixture approach utilising the same choice of parameter values ($M = 10$ and $q = 10$). To help visualise the final model, the means of each component $\boldsymbol{\mu}_i$ are illustrated in digit form in figure 9.

The digits were again classified, using the same method of classification, and the best model on the validation set misclassified 4.64% of the digits in the test set. Hinton *et al.* (1997) reported an error of 4.91%, and we would expect the improvement to be a result partly of the localised clustering of the PPCA model, but also the use of individually-estimated values of σ_i^2 for each component, rather than a single, arbitrarily-chosen, global value.

One of the advantages of the PPCA methodology is that the definition of the density model permits the posterior probabilities of class membership to be computed for each digit and utilised for subsequent classification, rather than using reconstruction error as above. Classification according to the largest posterior probability for the $M = 10$ and $q = 10$ model resulted in an increase in error, and it was necessary to invest significant effort to optimize the parameters M and q for each model to provide comparable performance. Using this approach, our best classifier on the validation set misclassified 4.61% of the test set. An additional benefit of the use of posterior probabilities is that it is possible to reject a proportion of the test samples about which the classifier is most 'unsure', and thus hopefully improve the classification performance. Using this approach to reject 5% of the test examples resulted in a misclassification rate of 2.50%. (Note that the availability of posteriors can be advantageous in other applications, where they may be utilised in various forms of follow-on processing.)

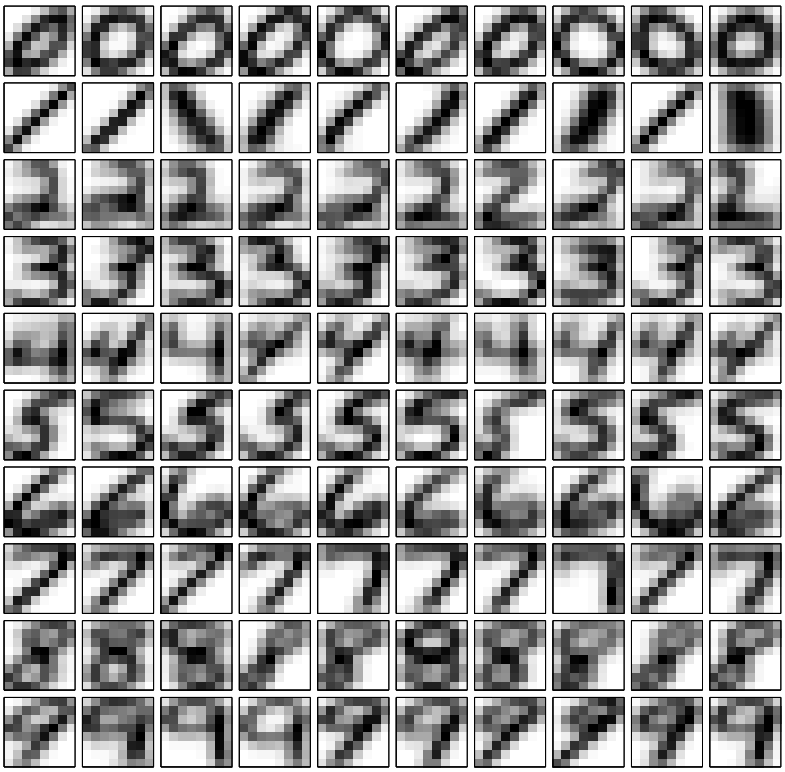

Figure 9: The mean vectors $\boldsymbol{\mu}_i$, illustrated as gray-scale digits, for each of the ten digit models. The model for a given digit is a mixture of ten PPCA models, one centred at each of the pixel vectors shown on the corresponding row. Note how different components can capture different styles of digit.

7 Conclusions

Modelling complexity in data by a combination of simple linear models is an attractive paradigm offering both computational and algorithmic advantages along with increased ease of interpretability. In this paper we have exploited the definition of a probabilistic model for PCA in order to combine local PCA models within the framework of a probabilistic mixture in which all the parameters are determined from maximum-likelihood using an EM algorithm. In addition to the clearly-defined nature of the resulting algorithm, the primary advantage of this approach is the definition of an observation density model.

A possible disadvantage of the probabilistic approach to combining local PCA models is that, by optimizing a likelihood function, the PPCA mixture model does not directly minimize squared reconstruction error. For applications where this is the salient criterion, algorithms which explicitly minimize reconstruction error should be expected to be superior. Experiments indeed showed this to be generally the case, but two important caveats must be considered before any firm conclusions can be drawn concerning the suitability of a given model. First, and rather surprisingly, for one of the datasets ('oil') considered in the paper the final PPCA mixture model was actually superior in the sense of squared reconstruction error, even on the training set. It was demonstrated that algorithms incorporating reconstruction-based clustering do not necessarily generate local clusters and it was reasoned that for datasets comprising a number of disjoint data structures, this phenomenon may lead to poor local minima. Such minima are not found by the PPCA density model approach. A second consideration is that there was also evidence that the smoothing implied by the soft clustering inherent in the PPCA mixture model helps to reduce overfitting, particularly

in the case of the image compression experiment where the statistics of the test data set differed from the training data much more so than for other examples. In that instance the reconstruction test error for the PPCA model was, on average, more than 10% lower.

In terms of a Gaussian mixture model, the mixture of probabilistic principal component analysers enables data to be modelled in high dimensions with relatively few free parameters, while at the same time not imposing a generally inappropriate constraint on the covariance structure. The number of free parameters may be controlled through the choice of latent space dimension q, allowing an interpolation in model complexity from isotropic to full covariance structures. The efficacy of this parameterisation was demonstrated by performance on a handwritten digit recognition task.

Acknowledgements: This work was supported by EPSRC contract GR/K51808: *Neural Networks for Visualization of High Dimensional Data*. We thank Michael Revow for supplying the handwritten digit data in its processed form.

References

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics 34*, 122–148.

Anderson, T. W. and H. Rubin (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume V, U. Cal, Berkeley, pp. 111–150.

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

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

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bishop, C. M., M. Svensén, and C. K. I. Williams (1998). GTM: the Generative Topographic Mapping. *Neural Computation 10*(1), 215–234.

Bishop, C. M. and M. E. Tipping (1998). A hierarchical latent variable model for data visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*(3), 281–293.

Bregler, C. and S. M. Omohundro (1995). Nonlinear image interpolation using manifold learning. In G. Tesauro, D. S. Touretzky, and T. K. Leen (Eds.), *Advances in Neural Information Processing Systems 7*, pp. 973–980. Cambridge, Mass: MIT Press.

Broomhead, D. S., R. Indik, A. C. Newell, and D. A. Rand (1991). Local adaptive Galerkin bases for large-dimensional dynamical systems. *Nonlinearity 4*(1), 159–197.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B 39*(1), 1–38.

Dony, R. D. and S. Haykin (1995). Optimally adaptive transform coding. *IEEE Transactions on Image Processing 4*(10), 1358–1370.

Hastie, T. and W. Stuetzle (1989). Principal curves. *Journal of the American Statistical Association 84*, 502–516.

Hinton, G. E., P. Dayan, and M. Revow (1997). Modelling the manifolds of images of handwritten digits. *IEEE Transactions on Neural Networks 8*(1), 65–74.

Hinton, G. E., M. Revow, and P. Dayan (1995). Recognizing handwritten digits using mixtures of linear models. In G. Tesauro, D. S. Touretzky, and T. K. Leen (Eds.), *Advances in Neural Information Processing Systems 7*, pp. 1015–1022. Cambridge, Mass: MIT Press.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*, 417–441.

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence 16*, 550–554.

Japkowicz, N., C. Myers, and M. Gluck (1995). A novelty detection approach to classification. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence*, pp. 518–523.

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

Jordan, M. I. and R. A. Jacobs (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation 6*(2), 181–214.

Kambhatla, N. (1995). *Local Models and Gaussian Mixture Models for Statistical Data Processing*. Ph. D. thesis, Oregon Graduate Institute, Center for Spoken Language Understanding.

Kambhatla, N. and T. K. Leen (1997). Dimension reduction by local principal component analysis. *Neural Computation 9*(7), 1493–1516.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal 37*(2), 233–243.

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

Lawley, D. N. (1953). A modified method of estimation in factor analysis and some large sample results. In *Uppsala Symposium on Psychological Factor Analysis*, Number 3 in Nordisk Psykologi Monograph Series, pp. 35–42. Uppsala: Almqvist and Wiksell.

Oja, E. (1983). *Subspace Methods of Pattern Recognition*. New York: John Wiley.

Ormoneit, D. and V. Tresp (1996). Improved Gaussian mixture density estimates using Bayesian penalty terms and network averaging. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8*, pp. 542–548. Cambridge: MIT Press.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series 2*, 559–572.

Petsche, T., A. Marcantonio, C. Darken, S. J. Hanson, G. M. Kuhn, and I. Santoso (1996). A neural network autoassociator for induction motor failure prediction. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8*, pp. 924–930. Cambridge: MIT Press.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika 20*, 93–111.

Rubin, D. B. and D. T. Thayer (1982). EM algorithms for ML factor analysis. *Psychometrika 47*(1), 69–76.

Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computing 2*, 183–190.

Tipping, M. E. and C. M. Bishop (1997). Mixtures of principal component analysers. In *Proceedings of the IEE Fifth International Conference on Artificial Neural Networks*, Cambridge, pp. 13–18. London: IEE.

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

Webb, A. R. (1996). An approach to nonlinear principal components–analysis using radially symmetrical kernel functions. *Statistics and Computing 6*(2), 159–168.

A Maximum-Likelihood PCA

A.1 The Stationary Points of the Log-Likelihood

The gradient of the log-likelihood (10) with respect to $\mathbf{W}$ may be obtained from standard matrix differentiation results (e.g. see Krzanowski and Marriott 1994, p. 133):

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}). \tag{34}$$

At the stationary points:

$$\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}, \tag{35}$$

assuming that $\sigma^2 > 0$, and thus that $\mathbf{C}^{-1}$ exists. This is a necessary and sufficient condition for the density model to remain nonsingular, and we will restrict ourselves to such cases. It will be seen shortly that $\sigma^2 > 0$ if $q < \text{rank}(\mathbf{S})$, so this assumption implies no loss of practicality.

There are three possible classes of solutions to equation (35):

1. $\mathbf{W} = \mathbf{0}$. This is shown later to be a minimum of the log-likelihood.
2. $\mathbf{C} = \mathbf{S}$, where the covariance model is exact, such as is discussed by Basilevsky (1994, pp 361–363) and considered in Section 2.3. In this unrealistic case of an exact covariance model, where the $d-q$ smallest eigenvalues of $\mathbf{S}$ are identical and equal to σ^2, $\mathbf{W}$ is identifiable since

$$\begin{aligned} \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}} &= \mathbf{S}, \\ \Rightarrow\ \mathbf{W} &= \mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \end{aligned} \tag{36}$$

   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.
3. $\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}$, with $\mathbf{W} \neq \mathbf{0}$ and $\mathbf{C} \neq \mathbf{S}$.

We are interested in case 3 where $\mathbf{C} \neq \mathbf{S}$ and the model covariance need not be equal to the sample covariance. First, we express the weight matrix $\mathbf{W}$ in terms of its singular value decomposition:

$$\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}, \tag{37}$$

where $\mathbf{U}$ 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. Now,

$$\begin{aligned} \mathbf{C}^{-1}\mathbf{W} &= (\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}, \\ &= \mathbf{W}(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}, \\ &= \mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^{\mathrm{T}}. \end{aligned} \tag{38}$$

Then at the stationary points, $\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}$ implies that

$$\begin{aligned} \mathbf{S}\mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^{\mathrm{T}} &= \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}, \\ \Rightarrow \qquad \mathbf{S}\mathbf{U}\mathbf{L} &= \mathbf{U}(\sigma^2\mathbf{I} + \mathbf{L}^2)\mathbf{L}. \end{aligned} \tag{39}$$

For $l_j \neq 0$, equation (39) implies that if $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q)$, then the corresponding column vector $\mathbf{u}_j$ must be an eigenvector of $\mathbf{S}$, with eigenvalue λ_j such that $\sigma^2 + l_j^2 = \lambda_j$, and so

$$l_j = (\lambda_j - \sigma^2)^{1/2}. \tag{40}$$

For $l_j = 0$, $\mathbf{u}_j$ is arbitrary (and if all l_j are zero, then we recover case 1). All potential solutions for $\mathbf{W}$ may thus be written as

$$\mathbf{W} = \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{41}$$

where $\mathbf{U}_q$ is a $d \times q$ matrix comprising q column eigenvectors of $\mathbf{S}$, and $\mathbf{K}_q$ is a $q \times q$ diagonal matrix with elements:

$$k_j = \begin{cases} \lambda_j, & \text{the corresponding eigenvalue to } \mathbf{u}_j, \text{ or,} \\ \sigma^2, & \end{cases} \tag{42}$$

where the latter case may be seen to be equivalent to $l_j = 0$. Again, $\mathbf{R}$ is an arbitrary orthogonal matrix, equivalent to a rotation in the principal subspace.

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 (41) is substituted into the log-likelihood function (10) to give

$$\mathcal{L} = -\frac{N}{2}\left\{ d\ln(2\pi) + \sum_{j=1}^{q'} \ln(\lambda_j) + \frac{1}{\sigma^2}\sum_{j=q'+1}^{d} \lambda_j + (d-q')\ln\sigma^2 + q' \right\}, \tag{43}$$

where q' is the number of non-zero l_j, $\{\lambda_1, \ldots, \lambda_{q'}\}$ are the eigenvalues corresponding to those 'retained' in $\mathbf{W}$, and $\{\lambda_{q'+1}, \ldots, \lambda_d\}$ are those 'discarded'. Maximizing (43) with respect to σ^2 gives

$$\sigma^2 = \frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j, \tag{44}$$

and so

$$\mathcal{L} = -\frac{N}{2}\left\{ \sum_{j=1}^{q'} \ln(\lambda_j) + (d-q')\ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j\right) + d\ln(2\pi) + d \right\}. \tag{45}$$

Note that (44) implies that $\sigma^2 > 0$ if $\text{rank}(\mathbf{S}) > q$ as stated earlier. We wish to find the maximum of (45) with respect to the choice of eigenvectors/eigenvalues to retain in $\mathbf{W}$, $j \in \{1, \ldots, q'\}$, and those to discard, $j \in \{q'+1, \ldots, d\}$. By exploiting the constancy of the sum of all eigenvalues with respect to this choice, the condition for maximization of the likelihood can be expressed equivalently as minimization of the quantity

$$E = \ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j\right) - \frac{1}{d-q'}\sum_{j=q'+1}^{d} \ln(\lambda_j), \tag{46}$$

which conveniently depends only on the discarded values and is non-negative (Jensen's inequality).

We consider minimization of E by first assuming that $d-q'$ discarded eigenvalues have been chosen arbitrarily, and, by differentiation, consider how a single such value λ_k affects the value of E:

$$\frac{\partial E}{\partial \lambda_k} = \frac{1}{\sum_{j=q'+1}^{d} \lambda_j} - \frac{1}{(d-q')\lambda_k}. \tag{47}$$

From (47), it can be seen that $E(\lambda_k)$ is convex and has a single minimum when λ_k is equal to the mean of the discarded eigenvalues (including itself). The eigenvalue λ_k can only take discrete values, but if we consider exchanging λ_k for some retained eigenvalue λ_j, $j \in \{1 \ldots q'\}$, then if λ_j lies between λ_k and the current mean retained eigenvalue, swapping λ_j and λ_k must decrease E. If we consider that the eigenvalues of $\mathbf{S}$ are ordered, for any combination of discarded eigenvalues which includes a 'gap' occupied by a retained eigenvalue, there will always be a sequence of adjacent

eigenvalues with a lower value of E. It follows then that to minimize E, the discarded eigenvalues $\lambda_{q'+1}, \ldots, \lambda_d$ must be chosen to be adjacent amongst the ordered eigenvalues of $\mathbf{S}$.

This alone is not sufficient to show that the *smallest* eigenvalues must be discarded in order to maximize the likelihood. However, a further constraint is available from equation (40), since $l_j = (\lambda_j - \sigma^2)^{1/2}$ implies that there can be no real solution to the stationary equations of the log-likelihood if any retained eigenvalue $\lambda_j < \sigma^2$. Since, from (44), σ^2 is the average of the discarded eigenvalues, this condition would be violated if the *smallest* eigenvalue were not discarded. Now, combined with the previous result, this indicates that E must be minimized when $\lambda_{q'+1}, \ldots, \lambda_d$ are the smallest $d - q'$ eigenvalues and so $\mathcal{L}$ is maximized when $\lambda_1, \ldots, \lambda_q$ are the principal eigenvalues of $\mathbf{S}$.

It should also be noted that the log-likelihood $\mathcal{L}$ is maximized, with respect to q', when there are fewest terms in the sum in (46) which occurs when $q' = q$ and therefore no l_j is zero. Furthermore, $\mathcal{L}$ is minimized when $\mathbf{W} = \mathbf{0}$, which is equivalent to the case of $q' = 0$.

A.3 The Nature of Other Stationary Points

If stationary points represented by minor (non-principal) eigenvector solutions are stable maxima of the likelihood, then local maximization (via an EM algorithm for example) is not guaranteed to find the principal eigenvectors. We may show, however, that minor eigenvector solutions are in fact saddle points on the likelihood surface.

Consider a stationary point of the log-likelihood, given by (41), at $\widehat{\mathbf{W}} = \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$, where $\mathbf{U}_q$ may contain q arbitrary eigenvectors of $\mathbf{S}$ and $\mathbf{K}_q$ contains either the corresponding eigenvalue or σ^2. We examine the nature of this stationary point by considering a small perturbation of the form $\mathbf{W} = \widehat{\mathbf{W}} + \epsilon\mathbf{P}\mathbf{R}$, where ϵ is an arbitrarily small positive constant and $\mathbf{P}$ is a $d \times q$ matrix of zeroes *except* for column W which contains a 'discarded' eigenvector $\mathbf{u}_P$ not contained in $\mathbf{U}_q$. By considering each potential eigenvector $\mathbf{u}_P$ individually applied to each column W of $\widehat{\mathbf{W}}$, we may elucidate the nature of the stationary point by evaluating the inner product of the perturbation with the gradient at $\mathbf{W}$ (where we treat the parameter matrix $\mathbf{W}$ or its derivative as a single column vector). If this inner product is negative for all possible perturbations, then the stationary point will be stable and represent a (local) maximum.

So defining $\mathbf{G} = (\partial\mathcal{L}/\partial\mathbf{W})/N$ evaluated at $\mathbf{W} = \widehat{\mathbf{W}} + \epsilon\mathbf{P}\mathbf{R}$, then from (34),

$$\begin{aligned}\mathbf{C}\mathbf{G} &= \mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{W}, \\ &= \mathbf{S}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1} - \mathbf{W}, \\ &= \mathbf{S}\mathbf{W}(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}} + \epsilon^2\mathbf{R}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{P}\mathbf{R})^{-1} - \mathbf{W}, \end{aligned} \tag{48}$$

since $\mathbf{P}^{\mathrm{T}}\widehat{\mathbf{W}} = \mathbf{0}$. Ignoring the term in ϵ^2 then gives:

$$\begin{aligned}\mathbf{C}\mathbf{G} &= \mathbf{S}(\widehat{\mathbf{W}} + \epsilon\mathbf{P}\mathbf{R})(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}})^{-1} - (\widehat{\mathbf{W}} + \epsilon\mathbf{P}\mathbf{R}), \\ &= \epsilon\mathbf{S}\mathbf{P}\mathbf{R}(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}})^{-1} - \epsilon\mathbf{P}\mathbf{R}, \end{aligned} \tag{49}$$

since $\mathbf{S}\widehat{\mathbf{W}}(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}}) - \widehat{\mathbf{W}} = \mathbf{0}$ at the stationary point. Then substituting for $\widehat{\mathbf{W}}$ gives $\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}} = \mathbf{R}^{\mathrm{T}}\mathbf{K}_q\mathbf{R}$, and so

$$\begin{aligned}\mathbf{C}\mathbf{G} &= \epsilon\mathbf{S}\mathbf{P}\mathbf{R}(\mathbf{R}^{\mathrm{T}}\mathbf{K}_q^{-1}\mathbf{R}) - \epsilon\mathbf{P}\mathbf{R}, \\ \Rightarrow \mathbf{G} &= \epsilon\mathbf{C}^{-1}\mathbf{P}(\mathbf{\Lambda}\mathbf{K}_q^{-1} - \mathbf{I})\mathbf{R}, \end{aligned} \tag{50}$$

where $\mathbf{\Lambda}$ is a $d \times d$ matrix of zeros, except for the W^{th} diagonal element which contains the eigenvalue corresponding to $\mathbf{u}_P$, such that $(\mathbf{\Lambda})_{WW} = \lambda_P$. Then the sign of the inner product of

the gradient $\mathbf{G}$ and the perturbation $\epsilon\mathbf{PR}$ is given by

$$\begin{aligned}\text{sign}\left(\text{tr}\ (\mathbf{G}^{\mathrm{T}}\mathbf{PR})\right) &= \text{sign}\left(\epsilon\text{tr}\ \left[\mathbf{R}^{\mathrm{T}}(\mathbf{\Lambda}\mathbf{K}_q^{-1} - \mathbf{I})\mathbf{P}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{PR}\right]\right), \\ &= \text{sign}\left((\lambda_P/k_W - 1)\mathbf{u}_P^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{u}_P\right), \\ &= \text{sign}\left(\lambda_P/k_W - 1\right), \end{aligned} \tag{51}$$

since $\mathbf{C}^{-1}$ is positive definite and where k_W is the W^{th} diagonal element value in $\mathbf{K}_q$, and thus in the corresponding position to λ_P in $\mathbf{\Lambda}$. When $k_W = \lambda_W$, the expression given by (51) is negative (and the maximum a stable one) if $\lambda_P < \lambda_W$. For $\lambda_P > \lambda_W$, $\widehat{\mathbf{W}}$ must be a saddle point.

In the case that $k_W = \sigma^2$, the stationary point will generally not be stable since, from (44), σ^2 is the average of $d - q'$ eigenvalues, and so $\lambda_P > \sigma^2$ for at least one of those eigenvalues, *except* when all those eigenvalues are identical. Such a case is considered shortly.

From this, by considering all possible perturbations $\mathbf{P}$, it can be seen that the only stable maximum occurs when $\mathbf{W}$ comprises the q principal eigenvectors, for which $\lambda_P < \lambda_W$, $\forall P \neq W$.

A.4 Equality of Eigenvalues

Equality of any of the q principal eigenvalues does not affect the maximum likelihood estimates. However, if thinking in terms of conventional PCA, consideration should be given to the instance when all the $d - q$ minor (discarded) eigenvalue(s) are equal and identical to at least one retained eigenvalue. (In practice, particularly in the case of sample covariance matrices, this is unlikely.)

To illustrate, consider the example of extracting two components from data with a covariance matrix possessing eigenvalues λ_1, λ_2 and λ_2, and $\lambda_1 > \lambda_2$. In this case, the second principal axis is not uniquely defined within the minor subspace. The spherical noise distribution defined by $\sigma^2 = \lambda_2$, in addition to explaining the residual variance, can also optimally explain the second principal component. Because $\lambda_2 = \sigma^2$, l_2 in equation (40) is zero, 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, but no second eigenvector is defined.

A.5 An EM Algorithm For PPCA

In the EM approach to PPCA, we consider the latent variables $\{\mathbf{x}_n\}$ to be 'missing' data. If their values were known, estimation of $\mathbf{W}$ would be straightforward from equation (2) by applying standard least-squares techniques. However, for a given $\mathbf{t}_n$, we don't know the value of $\mathbf{x}_n$ which generated it, but we do know the joint distribution of the observed and latent variables, $p(\mathbf{t}, \mathbf{x})$, and we can calculate the *expectation* of the corresponding *complete-data* log-likelihood. In the E-step of the EM algorithm this expectation, calculated with respect to the posterior distribution of $\mathbf{x}_n$ given the observed $\mathbf{t}_n$, is computed. In the M-step, new parameter values $\widetilde{\mathbf{W}}$ and $\widetilde{\sigma}^2$ are determined which maximize the expected complete-data log-likelihood and this is guaranteed to increase the likelihood of interest, $\prod_n p(\mathbf{t}_n)$, unless it is already at a local maximum (Dempster, Laird, and Rubin 1977).

The complete-data log-likelihood is given by:

$$\mathcal{L}_C = \sum_{n=1}^{N} \ln\left\{p(\mathbf{t}_n, \mathbf{x}_n)\right\}, \tag{52}$$

where, in PPCA, from equations (3) and (4)

$$p(\mathbf{t}_n, \mathbf{x}_n) = (2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{\|\mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \boldsymbol{\mu}\|^2}{2\sigma^2}\right\} (2\pi)^{-q/2} \exp\left\{-\frac{1}{2}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n\right\}. \tag{53}$$

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

$$\begin{aligned}\langle \mathcal{L}_C \rangle = -\sum_{n=1}^{N}\Big\{ \frac{d}{2}\ln\sigma^2 + \frac{1}{2}\mathrm{tr}\,(\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle) + \frac{1}{2\sigma^2}\|\mathbf{t}_n - \boldsymbol{\mu}\|^2 \\ - \frac{1}{\sigma^2}\langle \mathbf{x}_n\rangle^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}) + \frac{1}{2\sigma^2}\mathrm{tr}\,(\mathbf{W}^{\mathrm{T}}\mathbf{W}\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle)\Big\},\end{aligned} \tag{54}$$

where we have omitted terms independent of the model parameters and

$$\langle \mathbf{x}_n\rangle = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}), \tag{55}$$

$$\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle = \sigma^2\mathbf{M}^{-1} + \langle \mathbf{x}_n\rangle\langle \mathbf{x}_n\rangle^{\mathrm{T}}, \tag{56}$$

with $\mathbf{M} = (\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})$. Note that these statistics are computed using the current (fixed) values of the parameters, and that (55) is simply the posterior mean from equation (8). Equation (56) follows from this in conjunction with the posterior covariance of equation (9).

In the M-step, $\langle \mathcal{L}_C \rangle$ is maximized with respect to $\mathbf{W}$ and σ^2 by differentiating equation (54) and setting the derivatives to zero. This gives:

$$\widetilde{\mathbf{W}} = \left[\sum_n (\mathbf{t}_n - \boldsymbol{\mu})\langle \mathbf{x}_n^{\mathrm{T}}\rangle\right]\left[\sum_n \langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle\right]^{-1} \tag{57}$$

$$\widetilde{\sigma}^2 = \frac{1}{Nd}\sum_{n=1}^{N}\left\{\|\mathbf{t}_n - \boldsymbol{\mu}\|^2 - 2\langle \mathbf{x}_n^{\mathrm{T}}\rangle\widetilde{\mathbf{W}}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}) + \mathrm{tr}\left(\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle\widetilde{\mathbf{W}}^{\mathrm{T}}\widetilde{\mathbf{W}}\right)\right\} \tag{58}$$

To maximize the likelihood then, the sufficient statistics of the posterior distributions are calculated from the E-step equations (55) and (56) followed by the maximizing M-step equations (57) and (58). These four equations are iterated in sequence until the algorithm is judged to have converged.

We may gain considerable insight into the operation of equations (57) and (58) by substituting for $\langle \mathbf{x}_n\rangle$ and $\langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}\rangle$ from (55) and (56). Taking care not to confuse 'new' and 'old' parameters, some further manipulation leads to both the E-step and M-step being combined and re-written as:

$$\widetilde{\mathbf{W}} = \mathbf{S}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W})^{-1}, \text{ and} \tag{59}$$

$$\widetilde{\sigma}^2 = \frac{1}{d}\mathrm{tr}\left(\mathbf{S} - \mathbf{S}\mathbf{W}\mathbf{M}^{-1}\widetilde{\mathbf{W}}^{\mathrm{T}}\right), \tag{60}$$

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

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}. \tag{61}$$

Note that the first instance of $\mathbf{W}$ in equation (60) above is the *old* value of the weights, while the second instance $\widetilde{\mathbf{W}}$ is the *new* value calculated from equation (59). Equations (59), (60) and (61) indicate that the data enters into the EM formulation only through its covariance matrix $\mathbf{S}$, as we would expect.

Although it is algebraically convenient to express the EM algorithm in terms of $\mathbf{S}$, note that care should be exercised in any implementation. When $q \ll d$, it is possible to obtain considerable computational savings by not explicitly evaluating the covariance matrix, computation of which is $O(Nd^2)$. This is because inspection of (57) and (58) indicates that complexity is only $O(Ndq)$, and is reflected in (59) and (60) by the fact that $\mathbf{S}$ only appears within the terms $\mathbf{SW}$ and tr $(\mathbf{S})$, which may be computed with $O(Ndq)$ and $O(Nd)$ complexity respectively. That is, $\mathbf{SW}$ should be computed as $\sum_n(\mathbf{t}_n - \boldsymbol{\mu})\{(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{W}\}$, as that form is more efficient than $\{\sum_n(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}\}\mathbf{W}$, which is equivalent to finding $\mathbf{S}$ explicitly. However, because $\mathbf{S}$ need only be computed once in

the single model case and the EM algorithm is iterative, potential efficiency gains depend on the number of iterations required to obtain the desired accuracy of solution, as well as the ratio of d to q. For example, in our implementation of the model using $q = 2$ for data visualization, we found that an iterative approach could be more efficient for $d > 20$.

A.6 Rotational Ambiguity

If $\mathbf{W}$ is determined by the above algorithm, or any other iterative method that maximizes the likelihood (10), then at convergence, $\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$. If it is desired to find the true principal axes $\mathbf{U}_q$ (and not just the principal subspace) then the arbitrary rotation matrix $\mathbf{R}$ presents difficulty. This rotational ambiguity also exists in factor analysis, as well as in certain iterative PCA algorithms where it is usually not possible to determine the actual principal axes if $\mathbf{R} \neq \mathbf{I}$ (although there are algorithms where the contraint $\mathbf{R} = \mathbf{I}$ is imposed and the axes may be found).

However, in probabilistic PCA, $\mathbf{R}$ may actually be found since

$$\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}} = \mathbf{R}^{\mathrm{T}}(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})\mathbf{R}, \tag{62}$$

implies that $\mathbf{R}^{\mathrm{T}}$ may be computed as the matrix of eigenvectors of the $q \times q$ matrix $\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}$. Hence, both $\mathbf{U}_q$ and $\mathbf{\Lambda}_q$ may be found by inverting the rotation followed by normalisation of $\mathbf{W}_{\mathrm{ML}}$. That the rotational ambiguity may be resolved in PPCA is a consequence of the scaling of the eigenvectors by $(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}$ prior to rotation by $\mathbf{R}$. Without this scaling, $\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}} = \mathbf{I}$, and the corresponding eigenvectors remain ambiguous. Also, note that while finding the eigenvectors of $\mathbf{S}$ directly requires $O(d^3)$ operations, to obtain them from $\mathbf{W}_{\mathrm{ML}}$ in this way requires only $O(q^3)$.

B Optimal Least-Squares Reconstruction

One of the motivations for adopting PCA in many applications, notably in data compression, is the property of optimal linear least-squares reconstruction. That is for all orthogonal projections $\mathbf{x} = \mathbf{A}^{\mathrm{T}}\mathbf{t}$ of the data, the least-squares reconstruction error

$$E_{\mathrm{rec}}^2 = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{t}_n - \mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{t}_n\|^2 \tag{63}$$

is minimized when the columns of $\mathbf{A}$ span the principal subspace of the data covariance matrix, and $\mathbf{B} = \mathbf{A}$. (For simplification, and without loss of generality, we assume here that the data has zero mean.)

We can similarly obtain this property from our probabilistic formalism, without the need to determine the exact orthogonal projection $\mathbf{W}$, by finding the optimal reconstruction of the posterior mean vectors $\langle \mathbf{x}_n \rangle$. To do this we simply minimize

$$E_{\mathrm{rec}}^2 = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{t}_n - \mathbf{B}\langle \mathbf{x}_n \rangle\|^2, \tag{64}$$

over the reconstruction matrix $\mathbf{B}$, which is equivalent to a linear regression problem giving

$$\mathbf{B} = \mathbf{S}\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W})^{-1}\mathbf{M}, \tag{65}$$

where we have substituted for $\langle \mathbf{x}_n \rangle$ from (55). In general the resulting projection $\mathbf{B}\langle \mathbf{x}_n \rangle$ of $\mathbf{t}_n$ is not orthogonal, except in the maximum-likelihood case, where $\mathbf{W} = \mathbf{W}_{\mathrm{ML}} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$, and the optimal reconstructing matrix becomes

$$\mathbf{B}_{\mathrm{ML}} = \mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{M}, \tag{66}$$

and so

$$\hat{\mathbf{t}}_n = \mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{M}\langle\mathbf{x}_n\rangle, \tag{67}$$

$$= \mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{t}_n, \tag{68}$$

which is the expected orthogonal projection. The implication is thus that in the data compression context, at the maximum likelihood solution, the variables $\langle\mathbf{x}_n\rangle$ can be transmitted down the channel and the original data vectors optimally reconstructed using equation (67) given the parameters $\mathbf{W}$ and σ^2. Substituting for $\mathbf{B}$ in equation (64) gives $E^2_{\mathrm{rec}} = (d-q)\sigma^2$ and the noise term σ^2 thus represents the expected squared reconstruction error per 'lost' dimension.

C EM for Mixtures of Probabilistic PCA

In a mixture of probabilistic principal component analysers, we must fit a mixture of latent variable models in which the overall model distribution takes the form

$$p(\mathbf{t}) = \sum_{i=1}^{M} \pi_i p(\mathbf{t}|i), \tag{69}$$

where $p(\mathbf{t}|i)$ is a single probabilistic PCA model and π_i is the corresponding mixing proportion. The parameters for this mixture model can be determined by an extension of the EM algorithm. We begin by considering the standard form which the EM algorithm would take for this model and highlight a number of limitations. We then show that a two-stage form of EM leads to a more efficient algorithm.

We first note that in addition to a set of $\mathbf{x}_{ni}$ for each model i, the missing data includes variables z_{ni} labelling which model is responsible for generating each data point $\mathbf{t}_n$. At this point we can derive a standard EM algorithm by considering the corresponding complete-data log likelihood which takes the form

$$\mathcal{L}_{\mathrm{C}} = \sum_{n=1}^{N}\sum_{i=1}^{M} z_{ni} \ln\{\pi_i p(\mathbf{t}_n, \mathbf{x}_{ni})\}. \tag{70}$$

Starting with 'old' values for the parameters π_i, $\boldsymbol{\mu}_i$, $\mathbf{W}_i$ and σ_i^2 we first evaluate the posterior probabilities R_{ni} using (21) and similarly evaluate the expectations $\langle\mathbf{x}_{ni}\rangle$ and $\langle\mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}}\rangle$:

$$\langle\mathbf{x}_{ni}\rangle = \mathbf{M}^{-1}{}_i\mathbf{W}_i^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}_i), \tag{71}$$

$$\langle\mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}}\rangle = \sigma_i^2\mathbf{M}^{-1}{}_i + \langle\mathbf{x}_{ni}\rangle\langle\mathbf{x}_{ni}\rangle^{\mathrm{T}}, \tag{72}$$

with $\mathbf{M}_i = \sigma_i^2\mathbf{I} + \mathbf{W}_i^{\mathrm{T}}\mathbf{W}_i$.

Then we take the expectation of $\mathcal{L}_{\mathrm{C}}$ with respect to these posterior distributions to obtain

$$\begin{aligned}\langle\mathcal{L}_{\mathrm{C}}\rangle \;=\; & \sum_{n=1}^{N}\sum_{i=1}^{M} R_{ni}\left\{\ln\pi_i - \frac{d}{2}\ln\sigma_i^2 - \frac{1}{2}\mathrm{tr}\,(\langle\mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}}\rangle)\right. \\ & -\frac{1}{2\sigma_i^2}\|\mathbf{t}_{ni} - \boldsymbol{\mu}_i\|^2 + \frac{1}{\sigma_i^2}\langle\mathbf{x}_{ni}\rangle^{\mathrm{T}}\mathbf{W}_i^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}_i) \\ & \left.-\frac{1}{2\sigma_i^2}\mathrm{tr}\,(\mathbf{W}_i^{\mathrm{T}}\mathbf{W}_i\langle\mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}}\rangle)\right\},\end{aligned} \tag{73}$$

where $\langle\cdot\rangle$ denotes the expectation with respect to the posterior distributions of both $\mathbf{x}_{ni}$ and z_{ni} and terms independent of the model parameters have been omitted. The M-step then involves maximizing (73) with respect to π_i, $\boldsymbol{\mu}_i$, σ_i^2 and $\mathbf{W}_i$ to obtain 'new' values for these parameters.

The maximization with respect to π_i must take account of the constraint that $\sum_i \pi_i = 1$. This can be achieved with the use of a Lagrange multiplier λ (see Bishop 1995) and maximizing

$$\langle \mathcal{L}_{\mathrm{C}} \rangle + \lambda \left(\sum_{i=1}^{M} \pi_i - 1 \right). \tag{74}$$

Together with the results of maximizing (73) with respect to the remaining parameters, this gives the following M-step equations

$$\widetilde{\pi}_i = \frac{1}{N} \sum_n R_{ni} \tag{75}$$

$$\widetilde{\boldsymbol{\mu}}_i = \frac{\sum_n R_{ni} (\mathbf{t}_{ni} - \widetilde{\mathbf{W}}_i \langle \mathbf{x}_{ni} \rangle)}{\sum_n R_{ni}} \tag{76}$$

$$\widetilde{\mathbf{W}}_i = \left[\sum_n R_{ni} (\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i) \langle \mathbf{x}_{ni} \rangle^{\mathrm{T}} \right] \left[\sum_n R_{ni} \langle \mathbf{x}_{ni} \mathbf{x}_{ni}^{\mathrm{T}} \rangle \right]^{-1} \tag{77}$$

$$\begin{aligned} \widetilde{\sigma}_i^2 = \frac{1}{d \sum_n R_{ni}} \Bigg\{ & \sum_n R_{ni} \| \mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i \|^2 \\ & - 2 \sum_n R_{ni} \langle \mathbf{x}_{ni} \rangle^{\mathrm{T}} \widetilde{\mathbf{W}}_i^{\mathrm{T}} (\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i) + \sum_n R_{ni} \mathrm{tr} \left(\langle \mathbf{x}_{ni} \mathbf{x}_{ni}^{\mathrm{T}} \rangle \widetilde{\mathbf{W}}_i^{\mathrm{T}} \widetilde{\mathbf{W}}_i \right) \Bigg\} \end{aligned} \tag{78}$$

where the symbol $\widetilde{\ }$ denotes 'new' quantities that may be adjusted in the M-step. Note that the M-step equations for $\widetilde{\boldsymbol{\mu}}_i$ and $\widetilde{\mathbf{W}}_i$, given by (76) and (77), are coupled, and so further (albeit straightforward) manipulation is required to obtain explicit solutions.

In fact, simplification of the M-step equations, along with improved speed of convergence, is possible if we adopt a two-stage EM procedure as follows. The likelihood function we wish to maximize is given by

$$\mathcal{L} = \sum_{n=1}^{N} \ln \left\{ \sum_{i=1}^{M} \pi_i p(\mathbf{t}_n | i) \right\}. \tag{79}$$

Regarding the component labels z_{ni} as missing data, and ignoring the presence of the latent $\mathbf{x}$ variables for now, we can consider the corresponding expected complete-data log likelihood given by

$$\widehat{\mathcal{L}}_C = \sum_{n=1}^{N} \sum_{i=1}^{M} R_{ni} \ln \{ \pi_i p(\mathbf{t}_n | i) \} \tag{80}$$

where R_{ni} represent the posterior probabilities (corresponding to the expected values of z_{ni}) and are given by (21). Maximization of (80) with respect to π_i, again using a Lagrange multiplier, gives the M-step equation (22). Similarly, maximization of (80) with respect to $\boldsymbol{\mu}_i$ gives (23). This is the first stage of the combined EM procedure.

In order to update $\mathbf{W}_i$ and σ_i^2 we seek only to increase the value of $\widehat{\mathcal{L}}_C$ and not actually to maximize it. This corresponds to the generalised EM (or GEM) algorithm. We do this by considering $\widehat{\mathcal{L}}_C$ as our likelihood of interest and, introducing the missing $\mathbf{x}_{ni}$ variables, we perform one cycle of the EM algorithm, now with respect to the parameters $\mathbf{W}_i$ and σ_i^2. This second stage is guaranteed to increase $\widehat{\mathcal{L}}_C$, and therefore $\mathcal{L}$ as desired.

The advantages of this approach are two-fold. Firstly, the new values $\widetilde{\boldsymbol{\mu}}_i$ calculated in the first stage are utilised to compute the sufficient statistics of the posterior distribution of $\mathbf{x}_{ni}$ in the second stage using (71) and (72). By using updated values of $\boldsymbol{\mu}_i$ in computing these statistics, this leads to improved convergence speed.

A second advantage is that for the second stage of the EM algorithm, there is a considerable simplification of the M-step updates, since when (73) is expanded for $\langle \mathbf{x}_{ni} \rangle$ and $\langle \mathbf{x}_{ni} \mathbf{x}_{ni}^{\mathrm{T}} \rangle$, only

terms in $\widetilde{\boldsymbol{\mu}}_i$ (and not $\boldsymbol{\mu}_i$) appear. By inspection of (73) we see that the expected complete-data log likelihood now takes the form

$$\begin{aligned}\langle \mathcal{L}_{\mathrm{C}} \rangle &= \sum_{n=1}^{N}\sum_{i=1}^{M} R_{ni}\left\{ \ln \widetilde{\pi}_i - \frac{d}{2}\ln \sigma_i^2 - \frac{1}{2}\mathrm{tr}\,(\langle \mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}}\rangle) \right. \\ &\quad -\frac{1}{2\sigma_i^2}\|\mathbf{t}_{ni} - \widetilde{\boldsymbol{\mu}}_i\|^2 + \frac{1}{\sigma_i^2}\langle \mathbf{x}_{ni}^{\mathrm{T}}\rangle \mathbf{W}_i^{\mathrm{T}}(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i) \\ &\quad \left. -\frac{1}{2\sigma_i^2}\mathrm{tr}\,(\mathbf{W}_i^{\mathrm{T}}\mathbf{W}_i\langle \mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}}\rangle) \right\}. \end{aligned} \tag{81}$$

Now when we maximize (81) with respect to $\mathbf{W}_i$ and σ_i^2 (keeping $\widetilde{\boldsymbol{\mu}}_i$ fixed), we obtain the much simplified M-step equations:

$$\widetilde{\mathbf{W}}_i = \mathbf{S}_i\mathbf{W}_i(\sigma_i^2\mathbf{I} + \mathbf{M}^{-1}{}_i\mathbf{W}_i^{\mathrm{T}}\mathbf{S}_i\mathbf{W}_i)^{-1}, \tag{82}$$

$$\widetilde{\sigma}_i^2 = \frac{1}{d}\mathrm{tr}\left(\mathbf{S}_i - \mathbf{S}_i\mathbf{W}_i\mathbf{M}^{-1}{}_i\widetilde{\mathbf{W}}_i^{\mathrm{T}}\right), \tag{83}$$

where

$$\mathbf{S}_i = \frac{1}{\widetilde{\pi}_i N}\sum_{n=1}^{N} R_{ni}(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)^{\mathrm{T}}. \tag{84}$$

Iteration of equations (21)–(23) followed by equations (82) and (83) in sequence is guaranteed to find a local maximum of the likelihood (19).

Comparison of equations (82) and (83) with those of (59) and (60) earlier shows that the updates for the mixture case are identical to those of the single PPCA model, given that the *local responsibility-weighted* covariance matrix $\mathbf{S}_i$ is substituted for the global covariance matrix $\mathbf{S}$. Thus at stationary points, each weight matrix $\mathbf{W}_i$ contains the (scaled and rotated) eigenvectors of its respective $\mathbf{S}_i$, the local covariance matrix. Each sub-model is then performing a local PCA, where each data point is weighted by the responsibility of that sub-model for its generation, and a soft partitioning, similar to that introduced by Hinton *et al.* (1997), is automatically effected.

Given the established results for the single PPCA model, there is no need to use the iterative updates (82) and (83), since $\mathbf{W}_i$ and σ_i^2 may be determined by eigen-decomposition of $\mathbf{S}_i$, and the likelihood must still increase unless at a maximum. However, as discussed in Appendix A.5, the iterative EM scheme may offer computational advantages, particularly for $q \ll d$. In such a case, the iterative approach of equations (82) and (83) can be used, taking care to evaluate $\mathbf{S}_i\mathbf{W}_i$ efficiently as $\sum_n R_{ni}(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)\{(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)^{\mathrm{T}}\mathbf{W}_i\}$. Note that in the mixture case, unlike for the single model, $\mathbf{S}_i$ must be re-computed at each iteration of the EM algorithm, as the responsibilities R_{ni} will change.

As a final computational note, it might appear that the necessary calculation of $p(\mathbf{t}|i)$ would require inversion of the $d \times d$ matrix $\mathbf{C}$, an $O(d^3)$ operation. However, $(\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1} = \{\mathbf{I} - \mathbf{W}(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\}/\sigma^2$ and so $\mathbf{C}^{-1}$ may be computed using the already-calculated $q \times q$ matrix $\mathbf{M}^{-1}$.