LETTER Communicated by Todd Leen

Mixtures of Probabilistic Principal Component Analyzers

Michael E. Tipping
Christopher M. Bishop
Microsoft Research, St. George House, Cambridge CB2 3NH, U.K.

Principal component analysis (PCA) is one of the most popular techniques for processing, compressing, and visualizing 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. Therefore, previous attempts to formulate mixture models for PCA have been ad hoc to some extent. In this article, 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 analyzers, whose parameters can be determined using an expectation-maximization algorithm. We discuss the advantages of this model in the context of clustering, density modeling, 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 proved 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 $\{\mathbf{t}_n\}$, $n \in \{1 \ldots N\}$, the q principal axes $\mathbf{w}_j$, $j \in \{1, \ldots, q\}$, are those orthonormal axes onto which the retained variance under projection is maximal. It can be shown that the vectors $\mathbf{w}_j$ are given by the q dominant eigenvectors (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 $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q)$, 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 $\mathbf{S}$. 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, Dayan, and Revow (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, Myers, and Gluck (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 defines only a *linear* projection of the data, the scope of its application is necessarily somewhat limited. This has naturally motivated various developments of *nonlinear* PCA in an effort to retain a greater proportion of the variance using fewer components. Examples include principal curves (Hastie & Stuetzle, 1989; Tibshirani, 1992), multilayer autoassociative neural networks (Kramer, 1991), the kernel-function approach of Webb (1996), and the generative topographic mapping (GTM) of Bishop, Svensén, and Williams (1998). An alternative paradigm to such global nonlinear approaches is to model nonlinear structure with a collection, or mixture, of local linear submodels. This philosophy is an attractive one, motivating, for example, the mixture-of-experts technique for regression (Jordan & Jacobs, 1994).

A number of implementations of "mixtures of PCA" have been proposed in the literature, each defining a different algorithm or a variation. 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 esti-

mation 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 used as the criterion for cluster assignments. This conveys the advantage that a common cost measure is used in both stages. However, even recently proposed models that adopt this cost measure still define different algorithms (Hinton et al., 1997; Kambhatla & Leen, 1997), while a variety of alternative approaches for combining local PCA models have also been proposed (Broomhead, Indik, Newell, & Rand, 1991; Bregler & Omohundro, 1995; Hinton, Revow, & Dayan, 1995; Dony & Haykin, 1995). None of these algorithms defines a probability density.

One difficulty in implementation is that when using "hard" clustering in the partitioning phase (Kambhatla & Leen, 1997), the overall cost function is inevitably nondifferentiable. Hinton et al. (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 analyzer, the responsibility for its generation is shared among all of the analyzers. 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 analyzer for reconstructing data point $\mathbf{t}_n$ be given by $r_{nj} = \exp(-E_j^2/2\sigma^2)/\{\sum_{j'} \exp(-E_{j'}^2/2\sigma^2)\}$, 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 analyzers. 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 modeling. We demonstrate the operation of the algorithm on a simple toy problem and compare its performance with that of an explicit reconstruction-based nonprobabilistic modeling method on both synthetic and real-world data sets.

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 overconstraining 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 provided in the appendixes.

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 $\mathbf{x}$:

$$\mathbf{t} = \mathbf{y}(\mathbf{x}; \mathbf{w}) + \boldsymbol{\epsilon}, \tag{2.1}$$

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 2.1 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 equation 2.1.

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 $\mathbf{x}$:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}. \tag{2.2}$$

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 nonzero 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 $\mathbf{C} = \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}$. (As a result of this parameterization, $\mathbf{C}$ is invariant under postmultiplication of $\mathbf{W}$ by an orthogonal matrix, equivalent to a rotation of the $\mathbf{x}$ coordinate 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. There is no closed-form analytic solution for $\mathbf{W}$ and $\boldsymbol{\Psi}$, 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 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, p. 361), 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 eigendecomposition of $\mathbf{S}$, without resort to iteration (Anderson, 1963).

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 characterization 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 still emerges 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 in appendix A) we extend this earlier

work to give a full characterization of the properties of the model we term probabilistic PCA. Specifically, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, 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 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, equation 2.2 implies a probability distribution over $\mathbf{t}$-space for a given $\mathbf{x}$ of the form

$$p(\mathbf{t}|\mathbf{x}) = (2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{W}\mathbf{x} - \boldsymbol{\mu}\|^2\right\}. \tag{3.1}$$

With a gaussian prior over the latent variables defined by

$$p(\mathbf{x}) = (2\pi)^{-q/2} \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}\right\}, \tag{3.2}$$

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

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x}, \tag{3.3}$$

$$= (2\pi)^{-d/2}|\mathbf{C}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{t} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{t} - \boldsymbol{\mu})\right\}, \tag{3.4}$$

where the model covariance is

$$\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}}. \tag{3.5}$$

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

$$\begin{aligned} p(\mathbf{x}|\mathbf{t}) = {} & (2\pi)^{-q/2}|\sigma^{-2}\mathbf{M}|^{1/2} \\ & \times \exp\left[-\frac{1}{2}\left\{\mathbf{x} - \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu})\right\}^{\mathrm{T}}(\sigma^{-2}\mathbf{M})\right. \\ & \left.\left\{\mathbf{x} - \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu})\right\}\right], \end{aligned} \tag{3.6}$$

where the posterior covariance matrix is given by

$$\sigma^2\mathbf{M}^{-1} = \sigma^2(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}. \tag{3.7}$$

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

$$\mathcal{L} = \sum_{n=1}^{N} \ln \left\{ p(\mathbf{t}_n) \right\},$$
$$= -\frac{N}{2} \left\{ d \ln(2\pi) + \ln |\mathbf{C}| + \mathrm{tr} \left(\mathbf{C}^{-1}\mathbf{S} \right) \right\}, \tag{3.8}$$

where

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}, \tag{3.9}$$

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

3.2 Properties of the Maximum Likelihood Estimators. The maximum likelihood estimate of the parameter $\boldsymbol{\mu}$ is given by the mean of the data:

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{t}_n. \tag{3.10}$$

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

3.2.1 The Weight Matrix $\mathbf{W}$. The log-likelihood (see equation 3.8) is maximized when the columns of $\mathbf{W}$ span the principal subspace of the data. To show this we consider the derivative of equation 3.8 with respect to $\mathbf{W}$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}). \tag{3.11}$$

In appendix A it is shown that with $\mathbf{C}$ given by equation 3.5, the only nonzero stationary points of equation 3.11 occur for

$$\mathbf{W} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{3.12}$$

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 $\boldsymbol{\Lambda}_q$, 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 $\boldsymbol{\Lambda}_q$ 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 equation 3.12, the latent variable model defined by equation 2.2 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

$$\sigma^2_{\mathrm{ML}} = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j, \tag{3.13}$$

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 eigendecomposition of $\mathbf{S}$ (we consider an alternative, iterative approach shortly), after which σ^2_{ML} is found from equation 3.13 followed by $\mathbf{W}_{\mathrm{ML}}$ from equation 3.12. 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 2.2 represents a mapping from the lower-dimensional latent space to the data space. So in PPCA, the probabilistic analog 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 3.6, 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 equation 3.6. 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 equation 3.6 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 counterintuitively given equation 2.2, $\mathbf{W}_{\mathrm{ML}}\langle \mathbf{x}_n \rangle + \boldsymbol{\mu}$ 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 toward 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)

$$\hat{\mathbf{t}}_n = \mathbf{W}_{\mathrm{ML}} \left(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}} \mathbf{W}_{\mathrm{ML}} \right)^{-1} \mathbf{M} \langle \mathbf{x}_n \rangle + \boldsymbol{\mu}, \tag{3.14}$$

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 & Thayer 1982), we can obtain a principal component projection by maximizing the likelihood function (see equation 3.8). 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 eigendecomposition of $\mathbf{S}$, and subsequently incorporated in the probability model using equations 3.12 and 3.13 to realize 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 eigendecomposition, 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 behavior 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 nonsingular linear transformation of the data variables, so that $\mathbf{t} \rightarrow \mathbf{At}$. Using equation 3.10, we see that under such a transformation, the maximum likelihood solution for the mean will be transformed as $\boldsymbol{\mu}_{\mathrm{ML}} \rightarrow \mathbf{A}\boldsymbol{\mu}_{\mathrm{ML}}$. From equation 3.9, it then follows that the covariance matrix will transform as $\mathbf{S} \rightarrow \mathbf{ASA}^{\mathrm{T}}$.

The log-likelihood for the latent variable model, from equation 3.8, is

given by

$$\mathcal{L}(\mathbf{W}, \mathbf{\Psi}) = -\frac{N}{2}\Big\{d\ln(2\pi) + \ln|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \mathbf{\Psi}| + \mathrm{tr}\Big[(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \mathbf{\Psi})^{-1}\mathbf{S}\Big]\Big\}, \tag{3.15}$$

where $\mathbf{\Psi}$ is a general noise covariance matrix. Thus, using equation 3.15, we see that under the transformation $\mathbf{t} \rightarrow \mathbf{At}$, the log-likelihood will transform as

$$\mathcal{L}(\mathbf{W}, \mathbf{\Psi}) \rightarrow \mathcal{L}(\mathbf{A}^{-1}\mathbf{W}, \mathbf{A}^{-1}\mathbf{\Psi}\mathbf{A}^{-\mathrm{T}}) - N\ln|\mathbf{A}|, \tag{3.16}$$

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 $\mathbf{AW}_{\mathrm{ML}}$ 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 be proportional to the unit matrix only 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 nonprobabilistic PCA since a rotation of the coordinates induces a corresponding rotation of the principal axes. Thus we see that factor analysis is covariant under componentwise rescaling, while PPCA and PCA are covariant under rotations, as illustrated in Figure 1.

4 Mixtures of Probabilistic Principal Component Analyzers

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 & 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

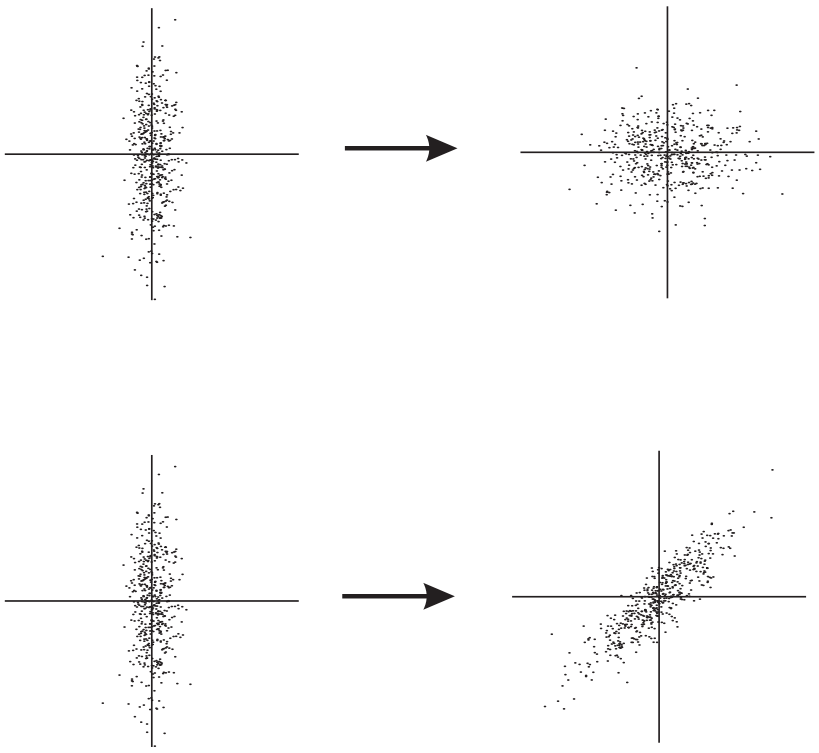

Figure 1: Factor analysis is covariant under a componentwise rescaling of the data variables (top plots), while PCA and probabilistic PCA are covariant under rotations of the data space coordinates (bottom plots).

model is:

$$\mathcal{L} = \sum_{n=1}^{N} \ln \left\{ p(\mathbf{t}_n) \right\}, \tag{4.1}$$

$$= \sum_{n=1}^{N} \ln \left\{ \sum_{i=1}^{M} \pi_i p(\mathbf{t}_n | i) \right\}, \tag{4.2}$$

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 & Tipping, 1998), and 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 2.2 as in the single model case, taking care to use 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

$$R_{ni} = \frac{p(\mathbf{t}_n|i)\pi_i}{p(\mathbf{t}_n)}, \tag{4.3}$$

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

$$\widetilde{\pi}_i = \frac{1}{N}\sum_{n=1}^{N} R_{ni}, \tag{4.4}$$

$$\widetilde{\boldsymbol{\mu}}_i = \frac{\sum_{n=1}^{N} R_{ni}\mathbf{t}_n}{\sum_{n=1}^{N} R_{ni}}. \tag{4.5}$$

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:

$$\mathbf{S}_i = \frac{1}{\widetilde{\pi}_i N}\sum_{n=1}^{N} R_{ni}(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)^{\mathrm{T}}, \tag{4.6}$$

by standard eigendecomposition in exactly the same manner as for a single PPCA model. However, as noted 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 4.3, 4.4, and 4.5 in sequence followed by computation of $\mathbf{W}_i$ and σ_i^2, from either equation 4.6 using equations 2.12 and 2.13 or using the iterative updates in appendix C, is guaranteed to find a local maximum of the log-likelihood in equation 4.2. 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 modeling 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 used in a typical algorithm.

5.1 Illustration for Synthetic Data. For a demonstration of the mixture of PPCA algorithm, we generated a synthetic data set comprising 500 data points sampled uniformly over the surface of a hemisphere, with additive gaussian noise. Figure 2a shows this data.

A mixture of 12 probabilistic principal component analyzers 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 emphasize 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 2b. 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 use 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 reestimated 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

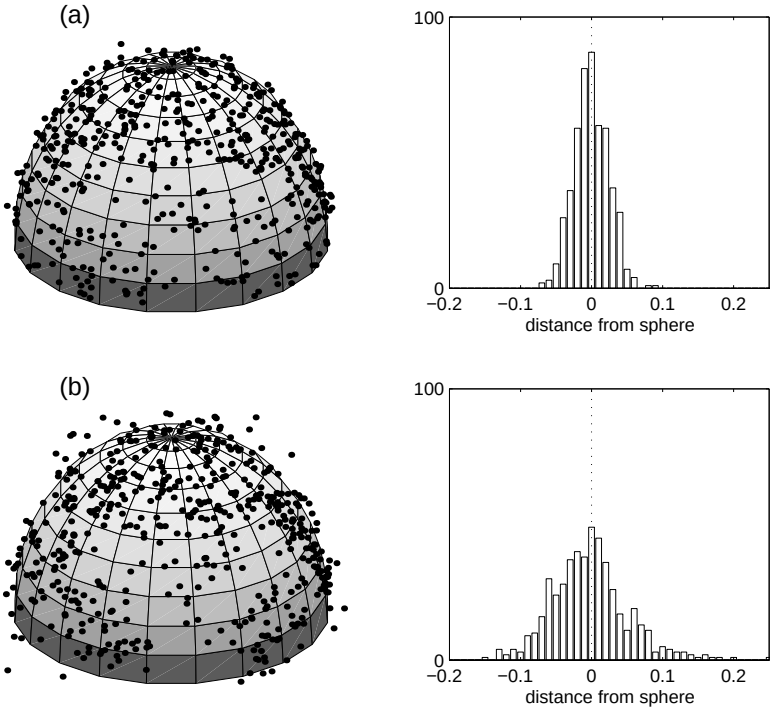

Figure 2: Modeling noisy data on a hemisphere. (a) On the left, the synthetic data; on the right, a histogram of the Euclidean distances of each data point to the sphere. (b) Data generated from the fitted PPCA mixture model with the synthetic data on the left and the histogram on the right.

data point $\mathbf{t}_n$:

$$p(\mathbf{t}_n|i) = (2\pi)^{-d/2}|\mathbf{C}_i|^{-1/2}\exp\left\{-E_{ni}^2/2\right\}, \tag{5.1}$$

where

$$E_{ni}^2 = (\mathbf{t}_n - \boldsymbol{\mu}_i)^{\mathrm{T}}\mathbf{C}_i^{-1}(\mathbf{t}_n - \boldsymbol{\mu}_i), \tag{5.2}$$

$$\mathbf{C}_i = \sigma_i^2\mathbf{I} + \mathbf{W}_i\mathbf{W}_i^{\mathrm{T}}. \tag{5.3}$$

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):

$$\mathbf{W} = \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{5.4}$$

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 parameterized, without loss of generality, in terms of $(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}$, where $\mathbf{K}_q = \text{diag}(k_1, k_2, \ldots, k_q)$ is a $q \times q$ diagonal matrix. Then

$$E_n^2 = (\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}} \left\{ \sigma^2\mathbf{I} + \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})\mathbf{U}_q^{\mathrm{T}} \right\}^{-1} (\mathbf{t}_n - \boldsymbol{\mu}). \tag{5.5}$$

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 equation 5.5 may then be written as

$$E_n^2 = \mathbf{z}_n^{\mathrm{T}}\mathbf{U}^{\mathrm{T}} \left\{ \sigma^2\mathbf{I} + \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})\mathbf{U}_q^{\mathrm{T}} \right\}^{-1} \mathbf{U}\mathbf{z}_n, \tag{5.6}$$

$$= \mathbf{z}_n^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{z}_n, \tag{5.7}$$

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

$$E_n^2 = \mathbf{z}_{\text{in}}^{\mathrm{T}}\mathbf{K}_q^{-1}\mathbf{z}_{\text{in}} + \frac{\mathbf{z}_{\text{out}}^{\mathrm{T}}\mathbf{z}_{\text{out}}}{\sigma^2}, \tag{5.8}$$

$$= E_{\text{in}}^2 + E_{\text{rec}}^2/\sigma^2, \tag{5.9}$$

where we have partitioned the elements of $\mathbf{z}$ into $\mathbf{z}_{\text{in}}$, the projection of $\mathbf{t}_n - \boldsymbol{\mu}$ onto the subspace spanned by $\mathbf{W}$, and $\mathbf{z}_{\text{out}}$, 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. 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 $\sigma_i^2 \to 0$, $R_{ni} \propto \pi_i \exp\left(-E_{\text{rec}}^2/2\right)$ 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 that lies in the subspace of a relatively distant component A, which may be reconstructed with zero error yet lies closer to the mean of a second component B. The effect of the noise variance σ_B^2 in equation 5.9 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 localized clusters, but with the 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 localized partitions (as noted by Kambhatla, 1995, who also considers the relationship

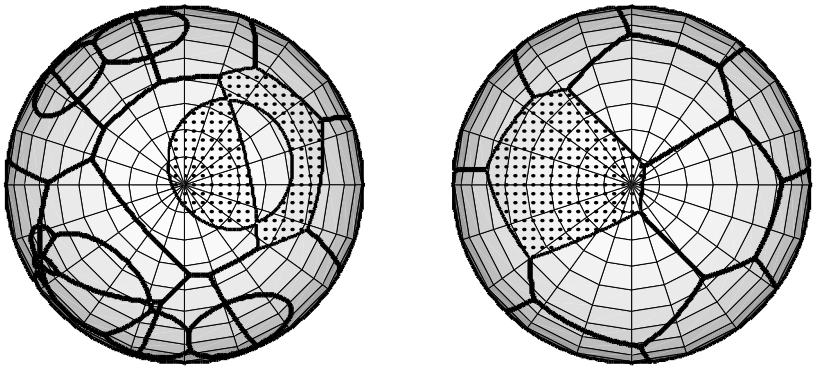

Figure 3: Comparison of the partitioning of the hemisphere effected by a VQPCA-based model (left) and a PPCA mixture model (right). The illustrated boundaries delineate regions of the hemisphere that are best reconstructed by a particular local PCA model. One such region is shown shaded to emphasize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

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), defined as follows:

1. Select initial cluster centers $\boldsymbol{\mu}_i$ at random from points in the data set, and assign all data points to the nearest (in terms of Euclidean distance) cluster center.
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 that best reconstructs them, setting each $\boldsymbol{\mu}_i$ to the mean of those data points assigned to cluster i.
4. Repeat from step 2 until the cluster allocations are constant.

In Figure 3, data points have been sampled over the hemisphere, without noise, and allocated to the cluster that 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

Table 1: Data Sets Used for Comparison of Clustering Criteria.

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 electroencephalogram time-series signal

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 nonlocal, 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.

Figure 4 summarizes 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 counterintuitively superior, despite the fact that the partitioning of the data space was based only partially on reconstruction error. This behavior is, we hypothesize, a result of the particular structure of that data set. The oil data are known to comprise a number of disjoint, but locally smooth, two-dimensional cluster structures (see Bishop & Tipping, 1998, for a visualization).

For the oil data set, we observed that many of the models found by the VQPCA algorithm exhibit partitions that are not only often nonconnected (similar to those shown for the hemisphere in Figure 3) but 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 reconstruc-

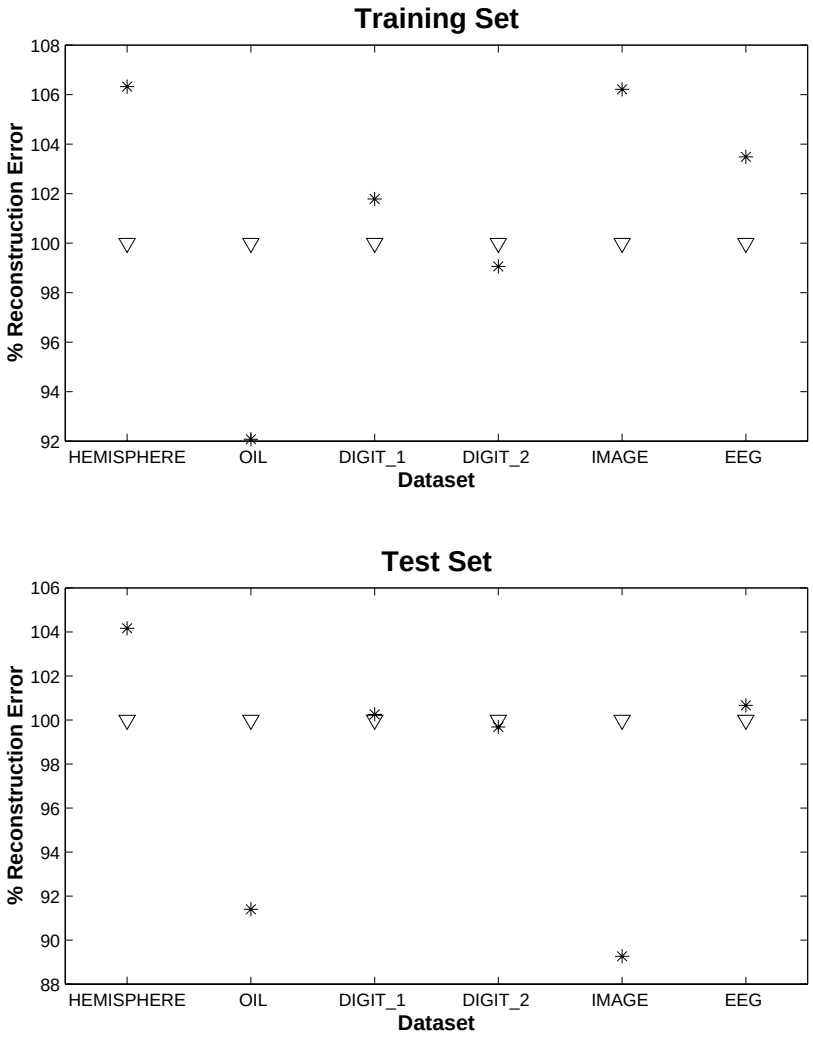

Figure 4: Reconstruction errors for reconstruction-based local PCA (VQPCA) and the PPCA mixture. Errors for the latter ($*$) have been shown relative to the former (∇), and are averaged over 100 runs with random initial configurations.

tion error cost function. The PPCA mixture algorithm does not find such suboptimal 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 apply only to the case of the oil data set. For the hemisphere, digit 1, image, and electroencephalogram (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 overfitting. As would be expected, compared with the training set, errors on the test set increase for both algorithms (although, because the errors have been normalized to allow comparisons between data sets, this is not 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 data set, 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 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 neighboring 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 nonoverlapping blocks, giving a total data set of 4050 64-dimensional vectors. Half of these data, corresponding to the left half of the picture, were used as training data. The right half was reserved for testing; 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 7 shows the reconstruction of the original image when modeled by a mixture of probabilistic principal component analyzers. The model parameters were estimated using only the left half of the image. In this example, 12

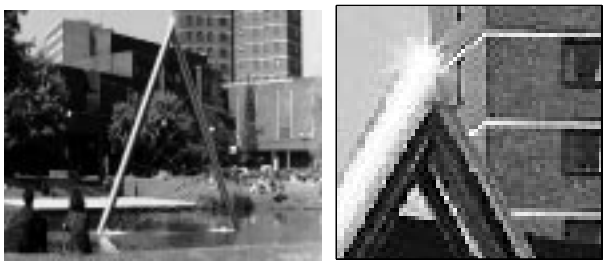

Figure 5: (Left) The original image. (Right) Detail.

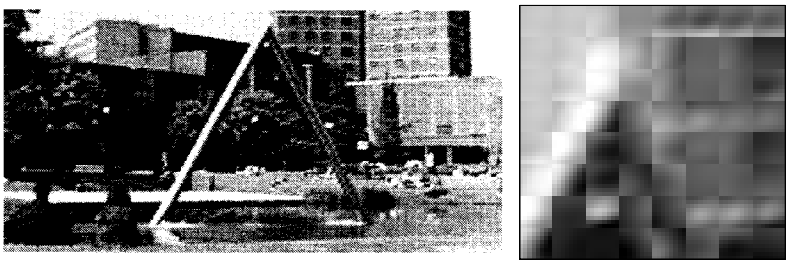

Figure 6: The PCA reconstructed image, at 0.5 bit per pixel. (Left) The original image. (Right) Detail.

components were used, of dimensionality 4; after the model likelihood had been maximized, the image coding was performed in a "hard" fashion, by allocating data to the component with the lowest reconstruction error. The resulting coded image was uniformly quantized, 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 to 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 normalized test error was 7.1×10^{-2}; for the mixture model, it was 5.7×10^{-2}. The VQPCA algorithm gave a test error of 6.2×10^{-2}.

6 Density Modeling

A popular approach to semiparametric density estimation is the gaussian mixture model (Titterington, Smith, & 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

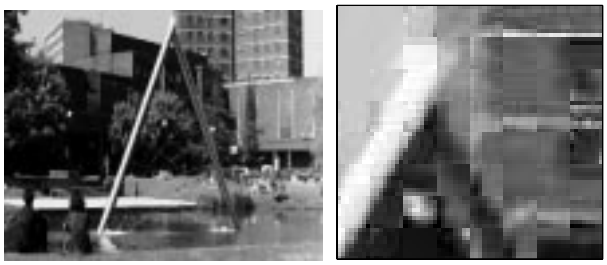

Figure 7: The mixture of PPCA reconstructed image, using the same bit rate as Figure 6. (Left) The original image. (Right) Detail.

approach 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 & Tresp, 1996.) Two common constraints are to restrict the covariance to be isotropic or to be diagonal. The isotropic model is highly constrained as it assigns only 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 among the variables.

A mixture of PPCA models, where the covariance of each gaussian is parameterized 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 were then modeled by both a mixture of PPCA models and a mixture of diagonal covariance gaussians, using eight mixture components. In the mixture of PPCA case, $q = 1$ for each component, and so there are four variance parameters per component compared with three for the diagonal model. The results are visualized in Figure 8, which illustrates both side and end projections of the data.

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

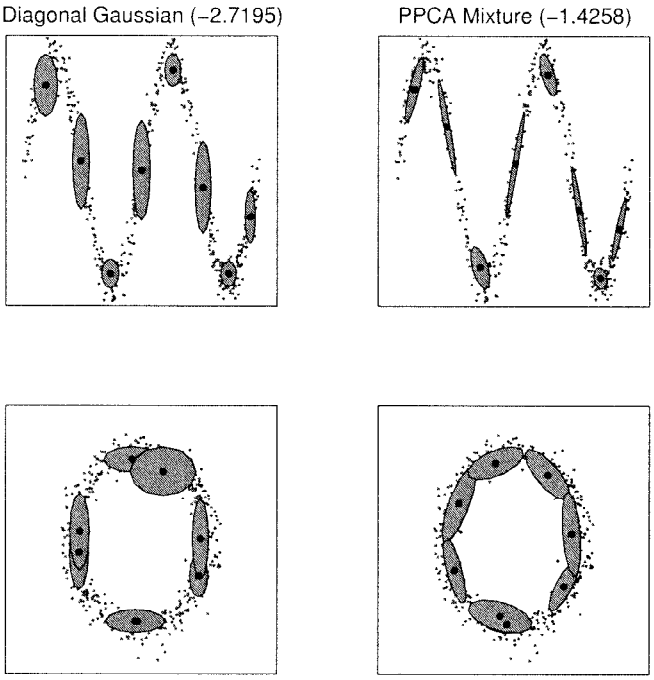

Figure 8: Comparison of an eight-component diagonal variance gaussian mixture model with a mixture of PPCA model. The upper two plots give a view perpendicular to the major axis of the spiral; the lower two plots show the end elevation. The covariance structure of each mixture component is shown by projection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

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 parameterization is that the means are also implicitly constrained because 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 parentheses in the figure. Such a result would be expected given that the PPCA 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 inappropriate when modeling a smooth

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.

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

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 modeling a three-dimensional spiral data set of 100 data points (the number of data points was reduced to emphasize the overfitting) 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 for both 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 parameterization 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 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 8×8 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*

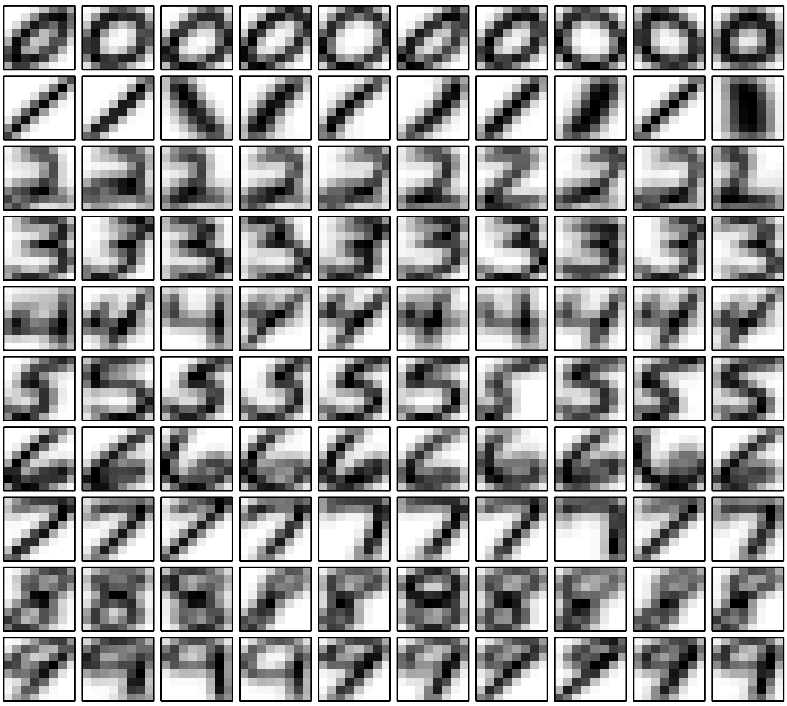

Figure 9: Mean vectors μ_i, illustrated as gray-scale digits, for each of the 10 digit models. The model for a given digit is a mixture of 10 PPCA models, one centered at each of the pixel vectors shown on the corresponding row. Note how different components can capture different styles of digit.

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 using the same choice of parameter values ($M = 10$ and $q = 10$). To help visualize 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 localized 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 used 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%. (The availability of posteriors can be advantageous in other applications, where they may be used in various forms of follow-on processing.)

7 Conclusions

Modeling 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 article, 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 that 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 data sets ('oil') considered in the article, 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 data sets 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 analyzers enables data to be modeled in high dimensions with relatively few free parameters, while 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 parameterization was demonstrated by performance on a handwritten digit recognition task.

Appendix A: Maximum Likelihood PCA

A.1 The Stationary Points of the Log-Likelihood. The gradient of the log-likelihood (see equation 3.8) with respect to $\mathbf{W}$ may be obtained from standard matrix differentiation results (e.g., see Krzanowski & Marriott, 1994, p. 133):

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

At the stationary points

$$\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}, \tag{A.2}$$

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 A.2:

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

$$\begin{aligned} &\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}} = \mathbf{S}, \\ &\quad \Rightarrow \ \mathbf{W} = \mathbf{U}(\mathbf{\Lambda} - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \end{aligned} \tag{A.3}$$

   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:

$$\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}, \tag{A.4}$$

where $\mathbf{U}$ is a $d \times q$ matrix of orthonormal column vectors, $\mathbf{L} = \mathrm{diag}(l_1, l_2, \ldots, l_q)$ is the $q \times q$ diagonal matrix of singular values, and $\mathbf{V}$ is a $q \times q$ orthogonal matrix. Now,

$$\begin{aligned} \mathbf{C}^{-1}\mathbf{W} &= (\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1}\mathbf{W}, \\ &= \mathbf{W}(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}, \\ &= \mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^{\mathrm{T}}. \end{aligned} \tag{A.5}$$

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

$$\begin{aligned} & \mathbf{S}\mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^{\mathrm{T}} = \mathbf{U}\mathbf{L}\mathbf{V}^{\mathrm{T}}, \\ \Rightarrow \quad & \mathbf{S}\mathbf{U}\mathbf{L} = \mathbf{U}(\sigma^2\mathbf{I} + \mathbf{L}^2)\mathbf{L}. \end{aligned} \tag{A.6}$$

For $l_j \neq 0$, equation A.6 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

$$l_j = (\lambda_j - \sigma^2)^{1/2}. \tag{A.7}$$

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

$$\mathbf{W} = \mathbf{U}_q(\mathbf{K}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{A.8}$$

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:

$$k_j = \begin{cases} \lambda_j, & \text{the corresponding eigenvalue to } \mathbf{u}_j, \text{ or,} \\ \sigma^2, & \end{cases} \tag{A.9}$$

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 that maximize the

likelihood, the expression for $\mathbf{W}$ in equation A.8 is substituted into the log-likelihood function (see equation 3.8) to give

$$\mathcal{L} = -\frac{N}{2}\left\{ d\ln(2\pi) + \sum_{j=1}^{q'} \ln(\lambda_j) + \frac{1}{\sigma^2}\sum_{j=q'+1}^{d} \lambda_j + (d-q')\ln\sigma^2 + q' \right\}, \tag{A.10}$$

where q' is the number of nonzero 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 equation A.10 with respect to σ^2 gives

$$\sigma^2 = \frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j, \tag{A.11}$$

and so

$$\mathcal{L} = -\frac{N}{2}\left\{ \sum_{j=1}^{q'} \ln(\lambda_j) + (d-q')\ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j\right) + d\ln(2\pi) + d \right\}. \tag{A.12}$$

Note that equation A.11 implies that $\sigma^2 > 0$ if $\mathrm{rank}(\mathbf{S}) > q$ as stated earlier. We wish to find the maximum of equation A.12 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

$$E = \ln\left(\frac{1}{d-q'}\sum_{j=q'+1}^{d} \lambda_j\right) - \frac{1}{d-q'}\sum_{j=q'+1}^{d} \ln(\lambda_j), \tag{A.13}$$

which conveniently depends on only the discarded values and is nonnegative (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:

$$\frac{\partial E}{\partial \lambda_k} = \frac{1}{\sum_{j=q'+1}^{d} \lambda_j} - \frac{1}{(d-q')\lambda_k}. \tag{A.14}$$

From equation A.14, 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 discarded 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 that 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 that to minimize E, the discarded eigenvalues $\lambda_{q'+1}, \ldots, \lambda_d$ must be chosen to be adjacent among 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 A.7, 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 equation A.11, 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 equation A.13 that 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 (nonprincipal) 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 equation A.8, 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 $\mathbf{W} = \widehat{\mathbf{W}} + \epsilon\mathbf{PR}$, 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 $\mathbf{W} = \widehat{\mathbf{W}} + \epsilon\mathbf{PR}$, then from

equation A.1,

$$\begin{aligned}\mathbf{CG} &= \mathbf{SC}^{-1}\mathbf{W} - \mathbf{W},\\ &= \mathbf{SW}(\sigma^2\mathbf{I} + \mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1} - \mathbf{W},\\ &= \mathbf{SW}(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}} + \epsilon^2\mathbf{R}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{PR})^{-1} - \mathbf{W}, \qquad \text{(A.15)}\end{aligned}$$

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

$$\begin{aligned}\mathbf{CG} &= \mathbf{S}(\widehat{\mathbf{W}} + \epsilon\mathbf{PR})(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}})^{-1} - (\widehat{\mathbf{W}} + \epsilon\mathbf{PR}),\\ &= \epsilon\mathbf{SPR}(\sigma^2\mathbf{I} + \widehat{\mathbf{W}}^{\mathrm{T}}\widehat{\mathbf{W}})^{-1} - \epsilon\mathbf{PR}, \qquad \text{(A.16)}\end{aligned}$$

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

$$\begin{aligned}&\mathbf{CG} = \epsilon\mathbf{SPR}(\mathbf{R}^{\mathrm{T}}\mathbf{K}_q^{-1}\mathbf{R}) - \epsilon\mathbf{PR},\\ &\Rightarrow \mathbf{G} = \epsilon\mathbf{C}^{-1}\mathbf{P}(\mathbf{\Lambda}\mathbf{K}_q^{-1} - \mathbf{I})\mathbf{R}, \qquad \text{(A.17)}\end{aligned}$$

where $\mathbf{\Lambda}$ is a $d \times d$ matrix of zeros, except for the Wth 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

$$\begin{aligned}\mathrm{sign}\left\{\mathrm{tr}\left(\mathbf{G}^{\mathrm{T}}\mathbf{PR}\right)\right\} &= \mathrm{sign}\left\{\epsilon\,\mathrm{tr}\left[\mathbf{R}^{\mathrm{T}}(\mathbf{\Lambda}\mathbf{K}_q^{-1} - \mathbf{I})\mathbf{P}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{PR}\right]\right\},\\ &= \mathrm{sign}\left\{(\lambda_P/k_W - 1)\mathbf{u}_P^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{u}_P\right\},\\ &= \mathrm{sign}\left\{\lambda_P/k_W - 1\right\}, \qquad \text{(A.18)}\end{aligned}$$

since $\mathbf{C}^{-1}$ is positive definite and where k_W is the Wth 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 equation A.18 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 equation A.11, 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, 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 A.7 is zero, and $\mathbf{W}$ effectively comprises only 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 2.2 by applying standard least-squares techniques. However, for a given $\mathbf{t}_n$, we do not know the value of $\mathbf{x}_n$ that 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 that 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, & Rubin, 1977).

The complete-data log-likelihood is given by:

$$\mathcal{L}_C = \sum_{n=1}^{N} \ln \left\{ p(\mathbf{t}_n, \mathbf{x}_n) \right\}, \tag{A.19}$$

where, in PPCA, from equations 3.1 and 3.4,

$$p(\mathbf{t}_n, \mathbf{x}_n) = (2\pi\sigma^2)^{-d/2} \exp\left\{ -\frac{\|\mathbf{t}_n - \mathbf{W}\mathbf{x}_n - \boldsymbol{\mu}\|^2}{2\sigma^2} \right\} (2\pi)^{-q/2} \times \exp\left\{ -\frac{1}{2}\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n \right\}. \tag{A.20}$$

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

$$\langle \mathcal{L}_C \rangle = -\sum_{n=1}^{N} \left\{ \frac{d}{2}\ln\sigma^2 + \frac{1}{2}\mathrm{tr}\left(\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle \right) + \frac{1}{2\sigma^2}\|\mathbf{t}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2}\langle \mathbf{x}_n \rangle^{\mathrm{T}} \mathbf{W}^{\mathrm{T}} (\mathbf{t}_n - \boldsymbol{\mu}) + \frac{1}{2\sigma^2}\mathrm{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle \right) \right\}, \tag{A.21}$$

where we have omitted terms independent of the model parameters and

$$\langle \mathbf{x}_n \rangle = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}), \tag{A.22}$$

$$\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle = \sigma^2 \mathbf{M}^{-1} + \langle \mathbf{x}_n \rangle \langle \mathbf{x}_n \rangle^{\mathrm{T}}, \tag{A.23}$$

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 equation A.22 is simply the posterior mean from equation 3.6. Equation A.23 follows from this in conjunction with the posterior covariance of equation 3.7.

In the M-step, $\langle \mathcal{L}_C \rangle$ is maximized with respect to $\mathbf{W}$ and σ^2 by differentiating equation A.21 and setting the derivatives to zero. This gives:

$$\widetilde{\mathbf{W}} = \left[\sum_n (\mathbf{t}_n - \boldsymbol{\mu}) \langle \mathbf{x}_n^{\mathrm{T}} \rangle \right] \left[\sum_n \langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle \right]^{-1} \tag{A.24}$$

$$\widetilde{\sigma}^2 = \frac{1}{Nd} \sum_{n=1}^{N} \left\{ \|\mathbf{t}_n - \boldsymbol{\mu}\|^2 - 2\langle \mathbf{x}_n^{\mathrm{T}} \rangle \widetilde{\mathbf{W}}^{\mathrm{T}} (\mathbf{t}_n - \boldsymbol{\mu}) + \mathrm{tr}\left(\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle \widetilde{\mathbf{W}}^{\mathrm{T}} \widetilde{\mathbf{W}} \right) \right\} \tag{A.25}$$

To maximize the likelihood then, the sufficient statistics of the posterior distributions are calculated from the E-step equations A.22 and A.23, followed by the maximizing M-step equations (A.24 and A.25). 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 A.24 and A.25 by substituting for $\langle \mathbf{x}_n \rangle$ and $\langle \mathbf{x}_n \mathbf{x}_n^{\mathrm{T}} \rangle$ from A.22 and A.23. Taking care not to confuse new and old parameters, some further manipulation leads to both the E-step and M-step's being combined and rewritten as:

$$\widetilde{\mathbf{W}} = \mathbf{S}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W})^{-1}, \text{ and} \tag{A.26}$$

$$\widetilde{\sigma}^2 = \frac{1}{d} \mathrm{tr}\left(\mathbf{S} - \mathbf{S}\mathbf{W}\mathbf{M}^{-1}\widetilde{\mathbf{W}}^{\mathrm{T}} \right), \tag{A.27}$$

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

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}}. \tag{A.28}$$

Note that the first instance of $\mathbf{W}$ in equation A.27 is the *old* value of the weights, while the second instance $\widetilde{\mathbf{W}}$ is the *new* value calculated from equation A.26. Equations A.26, A.27, and A.28 indicate that the data enter 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 $\mathbf{S}$, 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 equations A.24 and A.25 indicates that complexity is only $O(Ndq)$, and is reflected in equations A.26 and A.27 by the fact that $\mathbf{S}$ appears only 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 $\sum_n (\mathbf{t}_n - \boldsymbol{\mu}) \left\{ (\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{W} \right\}$, as that form is more efficient than $\left\{ \sum_n (\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathrm{T}} \right\} \mathbf{W}$, 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 (see equation 3.8), then at convergence, $\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$. 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 constraint $\mathbf{R} = \mathbf{I}$ is imposed and the axes may be found).

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

$$\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}} \mathbf{W}_{\mathrm{ML}} = \mathbf{R}^{\mathrm{T}} (\boldsymbol{\Lambda}_q - \sigma^2 \mathbf{I}) \mathbf{R} \tag{A.29}$$

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 $\boldsymbol{\Lambda}_q$ may be found by inverting the rotation followed by normalization of $\mathbf{W}_{\mathrm{ML}}$. That the rotational ambiguity may be resolved in PPCA is a consequence of the scaling of the eigenvectors by $(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}$ 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 $O(q^3)$.

Appendix 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,

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

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

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

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

$$\mathbf{B} = \mathbf{S}\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W})^{-1}\mathbf{M}, \tag{B.3}$$

where we have substituted for $\langle \mathbf{x}_n \rangle$ from equation A.22. 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 $\mathbf{W} = \mathbf{W}_{\text{ML}} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}$, and the optimal reconstructing matrix becomes

$$\mathbf{B}_{\text{ML}} = \mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{M}, \tag{B.4}$$

and so

$$\hat{\mathbf{t}}_n = \mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{M}\langle \mathbf{x}_n \rangle, \tag{B.5}$$

$$= \mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{t}_n, \tag{B.6}$$

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 B.5 given the parameters $\mathbf{W}$ and σ^2. Substituting for $\mathbf{B}$ in equation B.2 gives $E_{\text{rec}}^2 = (d-q)\sigma^2$, and the noise term σ^2 thus represents the expected squared reconstruction error per "lost" dimension.

Appendix C: EM for Mixtures of Probabilistic PCA

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

takes the form

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

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 that 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 include variables z_{ni} labeling 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

$$\mathcal{L}_{\mathrm{C}} = \sum_{n=1}^{N} \sum_{i=1}^{M} z_{ni} \ln\{\pi_i p(\mathbf{t}_n, \mathbf{x}_{ni})\}. \tag{C.2}$$

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 equation 4.3 and similarly evaluate the expectations $\langle \mathbf{x}_{ni} \rangle$ and $\langle \mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}} \rangle$:

$$\langle \mathbf{x}_{ni} \rangle = \mathbf{M}_i^{-1}\mathbf{W}_i^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu}_i), \tag{C.3}$$

$$\langle \mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}} \rangle = \sigma_i^2\mathbf{M}_i^{-1} + \langle \mathbf{x}_{ni} \rangle \langle \mathbf{x}_{ni} \rangle^{\mathrm{T}}, \tag{C.4}$$

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

$$\begin{aligned} \langle \mathcal{L}_{\mathrm{C}} \rangle = \sum_{n=1}^{N} \sum_{i=1}^{M} R_{ni} \Bigg\{ & \ln \pi_i - \frac{d}{2} \ln \sigma_i^2 - \frac{1}{2} \mathrm{tr} \left(\langle \mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}} \rangle \right) \\ & - \frac{1}{2\sigma_i^2} \|\mathbf{t}_{ni} - \boldsymbol{\mu}_i\|^2 + \frac{1}{\sigma_i^2} \langle \mathbf{x}_{ni} \rangle^{\mathrm{T}} \mathbf{W}_i^{\mathrm{T}} (\mathbf{t}_n - \boldsymbol{\mu}_i) \\ & - \frac{1}{2\sigma_i^2} \mathrm{tr} \left(\mathbf{W}_i^{\mathrm{T}}\mathbf{W}_i \langle \mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}} \rangle \right) \Bigg\}, \end{aligned} \tag{C.5}$$

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 equation C.5 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

$$\langle \mathcal{L}_{\mathrm{C}} \rangle + \lambda \left(\sum_{i=1}^{M} \pi_i - 1 \right). \tag{C.6}$$

Together with the results of maximizing equation C.5 with respect to the remaining parameters, this gives the following M-step equations:

$$\widetilde{\pi}_i = \frac{1}{N} \sum_n R_{ni} \tag{C.7}$$

$$\widetilde{\boldsymbol{\mu}}_i = \frac{\sum_n R_{ni} (\mathbf{t}_{ni} - \widetilde{\mathbf{W}}_i \langle \mathbf{x}_{ni} \rangle)}{\sum_n R_{ni}} \tag{C.8}$$

$$\widetilde{\mathbf{W}}_i = \left[\sum_n R_{ni} (\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i) \langle \mathbf{x}_{ni} \rangle^{\mathrm{T}} \right] \left[\sum_n R_{ni} \langle \mathbf{x}_{ni} \mathbf{x}_{ni}^{\mathrm{T}} \rangle \right]^{-1} \tag{C.9}$$

$$\widetilde{\sigma}_i^2 = \frac{1}{d \sum_n R_{ni}} \left\{ \sum_n R_{ni} \| \mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i \|^2 - 2 \sum_n R_{ni} \langle \mathbf{x}_{ni} \rangle^{\mathrm{T}} \widetilde{\mathbf{W}}_i^{\mathrm{T}} (\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i) + \sum_n R_{ni} \mathrm{tr} \left(\langle \mathbf{x}_{ni} \mathbf{x}_{ni}^{\mathrm{T}} \rangle \widetilde{\mathbf{W}}_i^{\mathrm{T}} \widetilde{\mathbf{W}}_i \right) \right\} \tag{C.10}$$

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 equations C.8 and C.9, 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

$$\mathcal{L} = \sum_{n=1}^{N} \ln \left\{ \sum_{i=1}^{M} \pi_i p(\mathbf{t}_n | i) \right\}. \tag{C.11}$$

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

$$\widehat{\mathcal{L}}_C = \sum_{n=1}^{N} \sum_{i=1}^{M} R_{ni} \ln \left\{ \pi_i p(\mathbf{t}_n | i) \right\}, \tag{C.12}$$

where R_{ni} represent the posterior probabilities (corresponding to the expected values of z_{ni}) and are given by equation 4.2. Maximization of equation C.12 with respect to π_i, again using a Lagrange multiplier, gives the

M-step equation (4.4). Similarly, maximization of equation C.12 with respect to $\boldsymbol{\mu}_i$ gives equation 4.5. 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 generalized 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, 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 twofold. First, the new values $\widetilde{\boldsymbol{\mu}}_i$ calculated in the first stage are used to compute the sufficient statistics of the posterior distribution of $\mathbf{x}_{ni}$ in the second stage using equations C.3 and C.4. 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 equation C.5 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 equation C.5, we see that the expected complete-data log-likelihood now takes the form

$$\begin{aligned}\langle \mathcal{L}_{\mathrm{C}} \rangle = \sum_{n=1}^{N}\sum_{i=1}^{M} R_{ni}\Bigg\{ \ln \widetilde{\pi}_i - \frac{d}{2}\ln \sigma_i^2 - \frac{1}{2}\mathrm{tr}\left(\langle \mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}} \rangle\right) \\ - \frac{1}{2\sigma_i^2}\|\mathbf{t}_{ni} - \widetilde{\boldsymbol{\mu}}_i\|^2 + \frac{1}{\sigma_i^2}\langle \mathbf{x}_{ni}^{\mathrm{T}} \rangle \mathbf{W}_i^{\mathrm{T}}(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i) \\ - \frac{1}{2\sigma_i^2}\mathrm{tr}\left(\mathbf{W}_i^{\mathrm{T}}\mathbf{W}_i \langle \mathbf{x}_{ni}\mathbf{x}_{ni}^{\mathrm{T}} \rangle\right)\Bigg\}. \end{aligned} \tag{C.13}$$

Now when we maximize equation C.13 with respect to $\mathbf{W}_i$ and σ_i^2 (keeping $\widetilde{\boldsymbol{\mu}}_i$ fixed), we obtain the much simplified M-step equations:

$$\widetilde{\mathbf{W}}_i = \mathbf{S}_i\mathbf{W}_i(\sigma_i^2\mathbf{I} + \mathbf{M}_i^{-1}\mathbf{W}_i^{\mathrm{T}}\mathbf{S}_i\mathbf{W}_i)^{-1}, \tag{C.14}$$

$$\widetilde{\sigma}_i^2 = \frac{1}{d}\mathrm{tr}\left(\mathbf{S}_i - \mathbf{S}_i\mathbf{W}_i\mathbf{M}_i^{-1}\widetilde{\mathbf{W}}_i^{\mathrm{T}}\right), \tag{C.15}$$

where

$$\mathbf{S}_i = \frac{1}{\widetilde{\pi}_i N}\sum_{n=1}^{N} R_{ni}(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)^{\mathrm{T}}. \tag{C.16}$$

Iteration of equations 4.3 through 4.5 followed by equations C.14 and C.15 in sequence is guaranteed to find a local maximum of the likelihood (see equation 4.1).

Comparison of equations C.14 and C.15 with equations A.26 and A.27 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 submodel is then performing a local PCA, where each data point is weighted by the responsibility of that submodel 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 (see equations C.14 and C.15) since $\mathbf{W}_i$ and σ_i^2 may be determined by eigendecomposition 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 C.14 and C.15 can be used, taking care to evaluate $\mathbf{S}_i\mathbf{W}_i$ efficiently as $\sum_n R_{ni}(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)\left\{(\mathbf{t}_n - \widetilde{\boldsymbol{\mu}}_i)^{\mathrm{T}}\mathbf{W}_i\right\}$. In the mixture case, unlike for the single model, $\mathbf{S}_i$ must be recomputed 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 $\mathbf{M}^{-1}$.

Acknowledgments

This work was supported by EPSRC contract GR/K51808: Neural Networks for Visualization of High Dimensional Data, at Aston University. 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., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 5, pp. 111–150). Berkeley: University of California, Berkeley.

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: Clarendon Press.

Bishop, C. M., Svensén, M., & Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation, 10*(1), 215–234.

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

Bregler, C., & Omohundro, S. M. (1995). Nonlinear image interpolation using manifold learning. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems, 7* (pp. 973–980). Cambridge, MA: MIT Press.

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B39*(1), 1–38.

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

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

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

Hinton, G. E., Revow, M., & Dayan, P. (1995). Recognizing handwritten digits using mixtures of linear models. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems, 7* (pp. 1015–1022). Cambridge, MA: 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., Myers, C., & Gluck, M. (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., & Jacobs, R. A. (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*. Unpublished doctoral dissertation, Oregon Graduate Institute, Center for Spoken Language Understanding.

Kambhatla, N., & Leen, T. K. (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., & Marriott, F. H. C. (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*. Nordisk Psykologi Monograph Series (pp. 35–42). Uppsala: Almqvist and Wiksell.

Oja, E. (1983). *Subspace methods of pattern recognition*. New York: Wiley.

Ormoneit, D., & Tresp, V. (1996). Improved gaussian mixture density estimates using Bayesian penalty terms and network averaging. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems, 8* (pp. 542–548). Cambridge, MA: MIT Press.

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

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

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

Rubin, D. B., & Thayer, D. T. (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., & Bishop, C. M. (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., Smith, A. F. M., & Makov, U. E. (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.

Received June 19, 1997; accepted May 19, 1998.

This article has been cited by:

1. Rafael Do Espírito Santo, Fabio Henrique Pereira, Edson Amaro Júnior. 2014. Image Compression Based on Generalized Principal Components Analysis and Simulated Annealing. *International Journal of Cognitive Informatics and Natural Intelligence* **6**:2, 41-67. [CrossRef]
2. Giuliano Galimberti, Gabriele Soffritti. 2013. Using conditional independence for parsimonious model-based Gaussian clustering. *Statistics and Computing* **23**:5, 625-638. [CrossRef]
3. Pierrick Bruneau, Marc Gelgon, Fabien Picarougne. 2013. A low-cost variational-Bayes technique for merging mixtures of probabilistic principal component analyzers. *Information Fusion* **14**:3, 268-280. [CrossRef]
4. Chenping Hou, Feiping Nie, Yuanyuan Jiao, Changshui Zhang, Yi Wu. 2013. Learning a subspace for clustering via pattern shrinking. *Information Processing & Management* **49**:4, 871-883. [CrossRef]
5. Qingchao Jiang, Xuefeng Yan. 2013. Probabilistic monitoring of chemical processes using adaptively weighted factor analysis and its application. *Chemical Engineering Research and Design* . [CrossRef]
6. Jukka J. Remes, Ahmed Abou Elseoud, Esa Ollila, Marianne Haapea, Tuomo Starck, Juha Nikkinen, Osmo Tervonen, Olli Silven. 2013. On applicability of PCA, voxel-wise variance normalization and dimensionality assumptions for sliding temporal window sICA in resting-state fMRI. *Magnetic Resonance Imaging* . [CrossRef]
7. Irene Vrbik, Paul D. McNicholas. 2013. Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics & Data Analysis* . [CrossRef]
8. Mohssen Mohammed, Al-Sakib PathanSignature Generation Algorithms for Polymorphic Worms 169-260. [CrossRef]
9. Kim Vincs, Kim Barbour. 2013. Snapshots of complexity: using motion capture and principal component analysis to reconceptualise dance. *Digital Creativity* 1-17. [CrossRef]
10. Anastasios Bellas, Charles Bouveyron, Marie Cottrell, Jérôme Lacaille. 2013. Model-based clustering of high-dimensional data streams with online mixture of probabilistic PCA. *Advances in Data Analysis and Classification* . [CrossRef]
11. Iván Machón-González, Hilario López-García, Jesús Rodríguez-Iglesias, Elena Marañón-Maison, Leonor Castrillón-Peláez, Yolanda Fernández-Nava. 2013. Comparing feed-forward versus neural gas as estimators: application to coke wastewater treatment. *Environmental Technology* **34**:9, 1131-1140. [CrossRef]
12. Yi-Qing Wang, Jean-Michel Morel. 2013. SURE Guided Gaussian Mixture Image Denoising. *SIAM Journal on Imaging Sciences* **6**:2, 999-1034. [CrossRef]
13. Timur Pekhovsky, Aleksandr Sizov. 2013. Comparison between Supervised and Unsupervised Learning of Probabilistic Linear Discriminant Analysis Mixture Models for Speaker Verification. *Pattern Recognition Letters* . [CrossRef]
14. M. Fauvel, J. Chanussot, J.A. Benediktsson, A. Villa. 2013. Parsimonious Mahalanobis kernel for the classification of high dimensional data. *Pattern Recognition* **46**:3, 845-854. [CrossRef]
15. Sanjeena Subedi, Antonio Punzo, Salvatore Ingrassia, Paul D. McNicholas. 2013. Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification* **7**:1, 5-40. [CrossRef]
16. Julien Jacques, Cristian Preda. 2013. Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing* . [CrossRef]
17. E. Zhang, J.D. Chazot, J. Antoni, M. Hamdi. 2013. Bayesian characterization of Young's modulus of viscoelastic materials in laminated structures. *Journal of Sound and Vibration* . [CrossRef]
18. Ognjen Arandjelović, Roberto Cipolla. 2013. Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant. *Pattern Recognition* **46**:1, 9-23. [CrossRef]
19. Fuxi Shi, Dan Zhang, Jun Chen, Hamid Reza Karimi. 2013. Missing Value Estimation for Microarray Data by Bayesian Principal Component Analysis and Iterative Local Least Squares. *Mathematical Problems in Engineering* **2013**, 1-5. [CrossRef]
20. Charles Bouveyron, Camille Brunet-Saumard. 2012. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis* . [CrossRef]
21. Teng Zhang, Arthur Szlam, Yi Wang, Gilad Lerman. 2012. Hybrid Linear Modeling via Local Best-Fit Flats. *International Journal of Computer Vision* **100**:3, 217-240. [CrossRef]
22. Matteo Falasconi, Matteo Pardo, Giorgio SberveglieriMethods and Graphical Tools for Exploratory Data Analysis of Artificial Olfaction Experiments 317-339. [CrossRef]
23. Dušan Kocur, Jana RovňákováShort-Range Tracking of Moving Targets by a Handheld UWB Radar System 207-225. [CrossRef]

24. Hwa Jeon Song, Yunkeun Lee, Hyung Soon Kim. 2012. Probabilistic Bilinear Transformation Space-Based Joint Maximum A Posteriori Adaptation. *ETRI Journal* **34**:5, 783-786. [CrossRef]
25. Masashi Sugiyama, Taiji Suzuki, Takafumi Kanamori. 2012. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics* **64**:5, 1009-1044. [CrossRef]
26. Leandro A.F. Fernandes, Manuel M. Oliveira. 2012. A general framework for subspace detection in unordered multidimensional data. *Pattern Recognition* **45**:9, 3566-3579. [CrossRef]
27. Arpan Banerjee, Ajay S. Pillai, Justin R. Sperling, Jason F. Smith, Barry Horwitz. 2012. Temporal microstructure of cortical networks (TMCN) underlying task-related differences. *NeuroImage* **62**:3, 1643-1657. [CrossRef]
28. Charles Bouveyron, Camille Brunet. 2012. Theoretical and practical considerations on the convergence properties of the Fisher-EM algorithm. *Journal of Multivariate Analysis* **109**, 29-41. [CrossRef]
29. Bo He, Xianhui Yang, Tao Chen, Jie Zhang. 2012. Reconstruction-based multivariate contribution analysis for fault isolation: A branch and bound approach. *Journal of Process Control* **22**:7, 1228-1236. [CrossRef]
30. Charles Bouveyron, Camille Brunet. 2012. Probabilistic Fisher discriminant analysis: A robust and flexible alternative to Fisher discriminant analysis. *Neurocomputing* **90**, 12-22. [CrossRef]
31. Shikui Tu, Lei Xu. 2012. A theoretical investigation of several model selection criteria for dimensionality reduction. *Pattern Recognition Letters* **33**:9, 1117-1126. [CrossRef]
32. László Dobos, János Abonyi. 2012. On-line detection of homogeneous operation ranges by dynamic principal component analysis based time-series segmentation. *Chemical Engineering Science* **75**, 96-105. [CrossRef]
33. Chenkun Qi, Han-Xiong Li, Shaoyuan Li, Xianchao Zhao, Feng Gao. 2012. Probabilistic PCA-Based Spatiotemporal Multimodeling for Nonlinear Distributed Parameter Processes. *Industrial & Engineering Chemistry Research* 120501132748008. [CrossRef]
34. Wan-Lun Wang, Tsung-I Lin. 2012. An efficient ECM algorithm for maximum likelihood estimation in mixtures of t-factor analyzers. *Computational Statistics* . [CrossRef]
35. Thiago Feital, Uwe Kruger, Julio Dutra, José Carlos Pinto, Enrique Luis Lima. 2012. Modeling and performance monitoring of multivariate multimodal processes. *AIChE Journal* n/a-n/a. [CrossRef]
36. Shivam Tripathi, Rao S. Govindaraju. 2011. Appraisal of Statistical Predictability under Uncertain Inputs: SST to Rainfall. *Journal of Hydrologic Engineering* **16**:12, 970-983. [CrossRef]
37. Paul D. McNicholas, Sanjeena Subedi. 2011. Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference* . [CrossRef]
38. Jochen Rau. 2011. Assessing thermalization and estimating the Hamiltonian with output data only. *Physical Review A* **84**:5. . [CrossRef]
39. Zhiqiang Ge, Furong Gao, Zhihuan Song. 2011. Two-dimensional Bayesian monitoring method for nonlinear multimode processes. *Chemical Engineering Science* **66**:21, 5173-5183. [CrossRef]
40. Tan Dat Nguyen, Surendra Ranganath. 2011. Facial expressions in American sign language: Tracking and recognition. *Pattern Recognition* . [CrossRef]
41. Charles Bouveyron, Julien Jacques. 2011. Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification* . [CrossRef]
42. Yue-Fei Guo, Xiaodong Lin, Zhou Teng, Xiangyang Xue, Jianping Fan. 2011. A covariance-free iterative algorithm for distributed principal component analysis on vertically partitioned data. *Pattern Recognition* . [CrossRef]
43. Charles Bouveyron, Gilles Celeux, Stéphane Girard. 2011. Intrinsic Dimension Estimation by Maximum Likelihood in Isotropic Probabilistic PCA. *Pattern Recognition Letters* . [CrossRef]
44. Mohamed Deriche. 2011. Multichannel feature selection using information maximization with application to multichannel EEG signals. *Journal of Communications Technology and Electronics* **56**:7, 838-846. [CrossRef]
45. Penghui Wang, Lei Shi, Lan Du, Hongwei Liu, Lei Xu, Zheng Bao. 2011. Radar HRRP statistical recognition with temporal factor analysis by automatic Bayesian Ying-Yang harmony learning. *Frontiers of Electrical and Electronic Engineering in China* **6**:2, 300-317. [CrossRef]
46. Shikui Tu, Lei Xu. 2011. An investigation of several typical model selection criteria for detecting the number of signals. *Frontiers of Electrical and Electronic Engineering in China* **6**:2, 245-255. [CrossRef]

47. MANOJIT CHATTOPADHYAY, PRANAB K. DAN, SITANATH MAZUMDAR. 2011. PRINCIPAL COMPONENT ANALYSIS AND SELF-ORGANIZING MAP FOR VISUAL CLUSTERING OF MACHINE-PART CELL FORMATION IN CELLULAR MANUFACTURING SYSTEM. *Systems Research Forum* **05**:01, 25-51. [CrossRef]
48. Shikui Tu, Lei Xu. 2011. Parameterizations make different model selections: Empirical findings from factor analysis. *Frontiers of Electrical and Electronic Engineering in China* **6**:2, 256-274. [CrossRef]
49. Ezequiel López-Rubio. 2011. Stochastic approximation learning for mixtures of multivariate elliptical distributions. *Neurocomputing* . [CrossRef]
50. John Ashburner, Stefan Klöppel. 2011. Multivariate models of inter-subject anatomical variability. *NeuroImage* **56**:2, 422-439. [CrossRef]
51. Geoffrey J. McLachlan, Jangsun Baek, Suren I. RathnayakeMixtures of Factor Analysers for the Analysis of High-Dimensional Data 189-212. [CrossRef]
52. Charles Bouveyron, Camille Brunet. 2011. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing* . [CrossRef]
53. Ahmed Fawzi Otoom, Hatice Gunes, Oscar Perez Concha, Massimo Piccardi. 2011. MLiT: mixtures of Gaussians under linear transformations. *Pattern Analysis and Applications* . [CrossRef]
54. Gyemin Lee, William Finn, Clayton Scott. 2011. Statistical file matching of flow cytometry data. *Journal of Biomedical Informatics* . [CrossRef]
55. Xu Lei. 2011. Codimensional matrix pairing perspective of BYY harmony learning: hierarchy of bilinear systems, joint decomposition of data-covariance, and applications of network biology. *Frontiers of Electrical and Electronic Engineering in China* **6**:1, 86-119. [CrossRef]
56. L. Hu, Z.G. Zhang, Y.S. Hung, K.D.K. Luk, G.D. Iannetti, Y. Hu. 2011. Single-trial detection of somatosensory evoked potentials by probabilistic independent component analysis and wavelet filtering. *Clinical Neurophysiology* . [CrossRef]
57. H. M. Mashaly, N. A. Masood, Abdalla S. A. Mohamed. 2011. Classification of Papulo-Squamous skin diseases using image analysis. *Skin Research and Technology* no-no. [CrossRef]
58. Bibliography 289-320. [CrossRef]
59. Zhiqiang Ge, Furong Gao, Zhihuan Song. 2011. Mixture probabilistic PCR model for soft sensing of multimode processes. *Chemometrics and Intelligent Laboratory Systems* **105**:1, 91-105. [CrossRef]
60. Gareth A. Tribello, Jéro^me Cuny, Hagai Eshet, Michele Parrinello. 2011. Exploring the free energy surfaces of clusters using reconnaissance metadynamics. *The Journal of Chemical Physics* **135**:11, 114109. [CrossRef]
61. Keith Worden, Wieslaw J. Staszewski, James J. Hensman. 2011. Natural computing for mechanical systems research: A tutorial overview. *Mechanical Systems and Signal Processing* **25**:1, 4-111. [CrossRef]
62. Fiora Pirri. 2011. The well-designed logical robot: Learning and experience from observations to the Situation Calculus. *Artificial Intelligence* **175**:1, 378-415. [CrossRef]
63. Minhua Chen, Jorge Silva, John Paisley, Chunping Wang, David Dunson, Lawrence Carin. 2010. Compressive Sensing on Manifolds Using a Nonparametric Mixture of Factor Analyzers: Algorithm and Performance Bounds. *IEEE Transactions on Signal Processing* **58**:12, 6140-6155. [CrossRef]
64. Sotirios Chatzis. 2010. A method for training finite mixture models under a fuzzy clustering principle. *Fuzzy Sets and Systems* **161**:23, 3000-3013. [CrossRef]
65. Zhiqiang Ge, Zhihuan Song. 2010. Mixture Bayesian regularization method of PPCA for multimode process monitoring. *AIChE Journal* **56**:11, 2838-2849. [CrossRef]
66. P. D. McNicholas, T. B. Murphy. 2010. Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics* **26**:21, 2705-2712. [CrossRef]
67. TAREM AHMED, RUMANA RAHMANSurvey of Anomaly Detection Algorithms 65-89. [CrossRef]
68. G. A. Tribello, M. Ceriotti, M. Parrinello. 2010. A self-learning algorithm for biased molecular dynamics. *Proceedings of the National Academy of Sciences* **107**:41, 17509-17514. [CrossRef]
69. Luca Scrucca. 2010. Dimension reduction for model-based clustering. *Statistics and Computing* **20**:4, 471-484. [CrossRef]
70. Ezequiel López-Rubio. 2010. Probabilistic Self-Organizing Maps for Continuous Data. *IEEE Transactions on Neural Networks* **21**:10, 1543-1554. [CrossRef]
71. Alexandre Carvalho, Georgios Skoulakis. 2010. Time Series Mixtures of Generalized t Experts: ML Estimation and an Application to Stock Return Density Forecasting. *Econometric Reviews* **29**:5, 642-687. [CrossRef]

72. Narihisa Matsumoto, Shotaro Akaho, Yasuko Sugase-Miyamoto, Masato Okada. 2010. Visualization of multi-neuron activity by simultaneous optimization of clustering and dimension reduction. *Neural Networks* **23**:6, 743-751. [CrossRef]
73. Shipeng Yu, Jinbo Bi, Jieping Ye. 2010. Matrix-variate and higher-order probabilistic projections. *Data Mining and Knowledge Discovery* . [CrossRef]
74. Shankar R. Rao, Allen Y. Yang, S. Shankar Sastry, Yi Ma. 2010. Robust Algebraic Segmentation of Mixed Rigid-Body and Planar Motions from Two Views. *International Journal of Computer Vision* **88**:3, 425-446. [CrossRef]
75. Robin C. Gilbert, Shivakumar Raman, Theodore B. Trafalis. 2010. Mathematical framework for form inspection. *The International Journal of Advanced Manufacturing Technology* . [CrossRef]
76. Scott C. Markley, David J. Miller. 2010. Joint Parsimonious Modeling and Model Order Selection for Multivariate Gaussian Mixtures. *IEEE Journal of Selected Topics in Signal Processing* **4**:3, 548-559. [CrossRef]
77. Kentaro Katahira, Narihisa Matsumoto, Yasuko Sugase-Miyamoto, Kazuo Okanoya, Masato Okada. 2010. Doubly sparse factor models for unifying feature transformation and feature selection. *Journal of Physics: Conference Series* **233**, 012021. [CrossRef]
78. Zhiqiang Ge, Zhihuan Song. 2010. Maximum-likelihood mixture factor analysis model and its application for process monitoring. *Chemometrics and Intelligent Laboratory Systems* **102**:1, 53-61. [CrossRef]
79. Banchar Arnonkijpanich, Alexander Hasenfuss, Barbara Hammer. 2010. Local matrix learning in clustering and applications for manifold visualization#. *Neural Networks* **23**:4, 476-486. [CrossRef]
80. Yisong Chen, Horace H.S. Ip, Sheng Li, Guoping Wang. 2010. Discovering hidden knowledge in data classification via multivariate analysis. *Expert Systems* **27**:2, 90-100. [CrossRef]
81. Jeffrey L. Andrews, Paul D. McNicholas. 2010. Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing* . [CrossRef]
82. Jo-Anne Ting, Aaron D'Souza, Sethu Vijayakumar, Stefan Schaal. 2010. Efficient Learning and Feature Selection in High-Dimensional Regression. *Neural Computation* **22**:4, 831-886. [Abstract] [Full Text] [PDF] [PDF Plus]
83. M NURULHAQUEMOLLAH. 2010. Robust extraction of local structures by the minimum $\beta\beta$-divergence method. *Neural Networks* **23**:2, 226-238. [CrossRef]
84. K.S. Gurumoorthy, A. Rajwade, A. Banerjee, A. Rangarajan. 2010. A Method for Compact Image Representation Using Sparse Matrix and Tensor Projections Onto Exemplar Orthonormal Bases. *IEEE Transactions on Image Processing* **19**:2, 322-334. [CrossRef]
85. Yuxiang Shan, Jia Liu. 2010. Robust speaker recognition in cross-channel condition based on Gaussian mixture model. *Multimedia Tools and Applications* . [CrossRef]
86. Julien Jacques, Charles Bouveyron, Stéphane Girard, Olivier Devos, Ludovic Duponchel, Cyril Ruckebusch. 2010. Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data. *Journal of Chemometrics* n/a-n/a. [CrossRef]
87. Paul D. McNicholas, T. Brendan Murphy. 2010. Model-based clustering of longitudinal data. *Canadian Journal of Statistics* n/a-n/a. [CrossRef]
88. Gang Qian, Feng GuoArticulated Human Motion Tracking 181-221. [CrossRef]
89. Guangliang Chen, Gilad Lerman. 2009. Foundations of a Multi-way Spectral Clustering Framework for Hybrid Linear Modeling. *Foundations of Computational Mathematics* **9**:5, 517-558. [CrossRef]
90. Ezequiel López-Rubio, Juan Miguel Ortiz-de-Lazcano-Lobato. 2009. Automatic Model Selection by Cross-Validation for Probabilistic PCA. *Neural Processing Letters* **30**:2, 113-132. [CrossRef]
91. EZEQUIEL LÓPEZ-RUBIO. 2009. ROBUST LOCATION AND SPREAD MEASURES FOR NONPARAMETRIC PROBABILITY DENSITY FUNCTION ESTIMATION. *International Journal of Neural Systems* **19**:05, 345-357. [CrossRef]
92. Ezequiel Lopez-Rubio, Juan Miguel Ortiz-de-Lazcano-Lobato, Domingo Lopez-Rodriguez. 2009. Probabilistic PCA Self-Organizing Maps. *IEEE Transactions on Neural Networks* **20**:9, 1474-1489. [CrossRef]
93. S.P. Chatzis, D.I. Kosmopoulos, T.A. Varvarigou. 2009. Robust Sequential Data Modeling Using an Outlier Tolerant Hidden Markov Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**:9, 1657-1669. [CrossRef]
94. B. Goossens, A. Pizurica, W. Philips. 2009. Image Denoising Using Mixtures of Projected Gaussian Scale Mixtures. *IEEE Transactions on Image Processing* **18**:8, 1689-1702. [CrossRef]
95. Peyman Adibi, Reza Safabakhsh. 2009. Information Maximization in a Linear Manifold Topographic Map. *Neural Processing Letters* **29**:3, 155-178. [CrossRef]

96. Sotirios Chatzis, Theodora Varvarigou. 2009. Factor Analysis Latent Subspace Modeling and Robust Fuzzy Clustering Using $t $-Distributions. *IEEE Transactions on Fuzzy Systems* **17**:3, 505-517. [CrossRef]
97. Tao Chen, Yue Sun. 2009. Probabilistic contribution analysis for statistical process monitoring: A missing variable approach. *Control Engineering Practice* **17**:4, 469-477. [CrossRef]
98. Heng Lian. 2009. Bayesian Nonlinear Principal Component Analysis Using Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**:4, 749-754. [CrossRef]
99. EZEQUIEL LÓPEZ-RUBIO, JUAN MIGUEL ORTIZ-DE-LAZCANO-LOBATO. 2009. DYNAMIC COMPETITIVE PROBABILISTIC PRINCIPAL COMPONENTS ANALYSIS. *International Journal of Neural Systems* **19**:02, 91-103. [CrossRef]
100. Xiang-Yun QING. 2009. Probabilistic Two-dimensional Principal Component Analysis. *Acta Automatica Sinica* **34**:3, 353-359. [CrossRef]
101. Guangliang Chen, Gilad Lerman. 2009. Spectral Curvature Clustering (SCC). *International Journal of Computer Vision* **81**:3, 317-330. [CrossRef]
102. Peng Qiu, Z.J. Wang, K. Liu, Z. Szabo. 2009. An Activity-Subspace Approach for Estimating the Integrated Input Function and Relative Distribution Volume in PET Parametric Imaging. *IEEE Transactions on Information Technology in Biomedicine* **13**:1, 25-36. [CrossRef]
103. Dechang Chen, Chang-Tien Lu, Yufeng Kou, Feng Chen. 2008. On Detecting Spatial Outliers. *GeoInformatica* **12**:4, 455-475. [CrossRef]
104. Jian-Hua Zhao, P.L.H. Yu. 2008. Fast ML Estimation for the Mixture of Factor Analyzers via an ECM Algorithm. *IEEE Transactions on Neural Networks* **19**:11, 1956-1961. [CrossRef]
105. Xianchao Xie, Shuicheng Yan, J.T. Kwok, T.S. Huang. 2008. Matrix-Variate Factor Analysis and Its Applications. *IEEE Transactions on Neural Networks* **19**:10, 1821-1826. [CrossRef]
106. Paul David McNicholas, Thomas Brendan Murphy. 2008. Parsimonious Gaussian mixture models. *Statistics and Computing* **18**:3, 285-296. [CrossRef]
107. Yi Fang, Myong K Jeong. 2008. Robust Probabilistic Multivariate Calibration Model. *Technometrics* **50**:3, 305-316. [CrossRef]
108. Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, Pierre Dumouchel. 2008. A Study of Interspeaker Variability in Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing* **16**:5, 980-988. [CrossRef]
109. Peng Wang, Qiang Ji. 2008. Robust Face Tracking via Collaboration of Generic and Specific Models. *IEEE Transactions on Image Processing* **17**:7, 1189-1199. [CrossRef]
110. José Camacho, Jesús Picó, Alberto Ferrer. 2008. Bilinear modelling of batch processes. Part I: theoretical discussion. *Journal of Chemometrics* **22**:5, 299-308. [CrossRef]
111. Lan Du, Hongwei Liu, Zheng Bao. 2008. Radar HRRP Statistical Recognition: Parametric Model and Model Selection. *IEEE Transactions on Signal Processing* **56**:5, 1931-1944. [CrossRef]
112. Yuan Cao, Rui Ming Liu, Jie Yang. 2008. Infrared Small Target Detection Using PPCA. *International Journal of Infrared and Millimeter Waves* **29**:4, 385-395. [CrossRef]
113. Sotirios P. Chatzis, Dimitrios I. Kosmopoulos, Theodora A. Varvarigou. 2008. Signal Modeling and Classification Using a Robust Latent Space Model Based on t Distributions. *IEEE Transactions on Signal Processing* **56**:3, 949-963. [CrossRef]
114. Renaud Maroy, Raphail Boisgard, Claude Comtat, Vincent Frouin, Pascal Cathier, Edouard Duchesnay, Fridiric Dolle, Peter E. Nielsen, Rigine Trebossen, Bertrand Tavitian. 2008. Segmentation of Rodent Whole-Body Dynamic PET Images: An Unsupervised Method Based on Voxel Dynamics. *IEEE Transactions on Medical Imaging* **27**:3, 342-354. [CrossRef]
115. Guido Sanguinetti. 2008. Dimensionality Reduction of Clustered Data Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**:3, 535-540. [CrossRef]
116. Yun Tang, Richard Rose. 2008. Rapid Speaker Adaptation Using Clustered Maximum-Likelihood Linear Basis With Sparse Training Data. *IEEE Transactions on Audio, Speech, and Language Processing* **16**:3, 607-616. [CrossRef]
117. Herminia García-Mozo, Rosa Perez-Badía, Carmen Galán. 2008. Aerobiological and meteorological factors' influence on olive (Olea europaea L.) crop yield in Castilla-La Mancha (Central Spain). *Aerobiologia* **24**:1, 13-18. [CrossRef]
118. W YINGHUI, N XIAOJUAN, Y CHUNXIA, W QIONGFANG. 2008. A Method of Illumination Compensation for Human Face Image Based On Quotient Image. *Information Sciences* . [CrossRef]

119. Akisato Kimura, Kunio Kashino, Takayuki Kurozumi, Hiroshi Murase. 2008. A Quick Search Method for Audio Signals Based on a Piecewise Linear Representation of Feature Trajectories. *IEEE Transactions on Audio, Speech, and Language Processing* **16**:2, 396-407. [CrossRef]
120. Qingfu Zhang, Aimin Zhou, Yaochu Jin. 2008. RM-MEDA: A Regularity Model-Based Multiobjective Estimation of Distribution Algorithm. *IEEE Transactions on Evolutionary Computation* **12**:1, 41-63. [CrossRef]
121. Boštjan Vesnicer, France Mihelič. 2008. The Likelihood Ratio Decision Criterion for Nuisance Attribute Projection in GMM Speaker Verification. *EURASIP Journal on Advances in Signal Processing* **2008**, 1-12. [CrossRef]
122. Feng Guo, Gang Qian. 2008. Monocular 3D Tracking of Articulated Human Motion in Silhouette and Pose Manifolds. *EURASIP Journal on Image and Video Processing* **2008**, 1-19. [CrossRef]
123. Haixian Wang, Sibao Chen, Zilan Hu, Bin Luo. 2007. Probabilistic two-dimensional principal component analysis and its mixture model for face recognition. *Neural Computing and Applications* . [CrossRef]
124. Feng Zhang, Timwah Luk. 2007. A Data Mining Algorithm for Monitoring PCB Assembly Quality. *IEEE Transactions on Electronics Packaging Manufacturing* **30**:4, 299-305. [CrossRef]
125. Schubert R. Carvalho, Ronan Boulic, Daniel Thalmann. 2007. Interactive low-dimensional human motion synthesis by combining motion models and PIK. *Computer Animation and Virtual Worlds* **18**:4-5, 493-503. [CrossRef]
126. T KIM, O ARANDJELOVIC, R CIPOLLA. 2007. Boosted manifold principal angles for image set-based recognition. *Pattern Recognition* **40**:9, 2475-2484. [CrossRef]
127. H GARCIAMOZO, M GOMEZCASERO, E DOMINGUEZ, C GALAN. 2007. Influence of pollen emission and weather-related factors on variations in holm-oak (Quercus ilex subsp. ballota) acorn production. *Environmental and Experimental Botany* **61**:1, 35-40. [CrossRef]
128. Yi Ma, Harm Derksen, Wei Hong, John Wright. 2007. Segmentation of Multivariate Mixed Data via Lossy Data Coding and Compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**:9, 1546-1562. [CrossRef]
129. Francois Destrempes, Max Mignotte, Jean-Francois Angers. 2007. Localization of Shapes Using Statistical Models and Stochastic Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**:9, 1603-1615. [CrossRef]
130. F. Flitti, Ch. Collet. 2007. Markovian regularization of latent-variable-models mixture for New multi-component image reduction/segmentation scheme. *Signal, Image and Video Processing* **1**:3, 191-201. [CrossRef]
131. Daniel Keysers, Thomas Deselaers, Christian Gollan, Hermann Ney. 2007. Deformation Models for Image Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**:8, 1422-1435. [CrossRef]
132. Long Quan, Jingdong Wang, Ping Tan, Lu Yuan. 2007. Image-Based Modeling by Joint Segmentation. *International Journal of Computer Vision* **75**:1, 135-150. [CrossRef]
133. Xin Miao, Peng Gong, Ruiliang Pu, Raymond I Carruthers, Jill S Heaton. 2007. Applying class-based feature extraction approaches for supervised classification of hyperspectral imagery. *Canadian Journal of Remote Sensing* **33**:3, 162-175. [CrossRef]
134. 2007. Full Issue in PDF / Numéro complet enform PDF. *Canadian Journal of Remote Sensing* **33**:3. . [CrossRef]
135. Torbjorn Vik, Fabrice Heitz, Pierre Charbonnier. 2007. Robust Pose Estimation and Recognition Using Non-Gaussian Modeling of Appearance Subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**:5, 901-905. [CrossRef]
136. Patrick Kenny, Gilles Boulianne, Pierre Ouellet, Pierre Dumouchel. 2007. Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech and Language Processing* **15**:4, 1435-1447. [CrossRef]
137. G. Heidemann, H. Bekel, I. Bax, H. Ritter. 2007. Interactive online learning. *Pattern Recognition and Image Analysis* **17**:1, 146-152. [CrossRef]
138. H HOFFMANN. 2007. Kernel PCA for novelty detection. *Pattern Recognition* **40**:3, 863-874. [CrossRef]
139. Carlotta Domeniconi, Dimitrios Gunopulos, Sheng Ma, Bojun Yan, Muna Al-Razgan, Dimitris Papadopoulos. 2007. Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery* **14**:1, 63-97. [CrossRef]
140. Bibliography **212**, 379-393. [CrossRef]
141. H HOFFMANN. 2007. Perception through visuomotor anticipation in a mobile robot#. *Neural Networks* **20**:1, 22-33. [CrossRef]
142. A VELLIDO. 2006. Missing data imputation through GTM as a mixture of tt-distributions. *Neural Networks* **19**:10, 1624-1635. [CrossRef]
143. V. Bochko, J. Parkkinen. 2006. A Spectral Color Analysis and Colorization Technique. *IEEE Computer Graphics and Applications* **26**:5, 74-82. [CrossRef]
144. J. Verbeek. 2006. Learning nonlinear image manifolds by global alignment of local linear models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**:8, 1236-1250. [CrossRef]

145. Tao Xiang, Shaogang Gong. 2006. Model Selection for Unsupervised Learning of Visual Context. *International Journal of Computer Vision* **69**:2, 181-201. [CrossRef]

146. C. Constantinopoulos, A. Likas. 2006. An Incremental Training Method for the Probabilistic RBF Network. *IEEE Transactions on Neural Networks* **17**:4, 966-974. [CrossRef]

147. R. Yoshida. 2006. ArrayCluster: an analytic tool for clustering, data visualization and module finder on gene expression profiles. *Bioinformatics* **22**:12, 1538-1539. [CrossRef]

148. H WANG, E HANCOCK. 2006. Correspondence matching using kernel principal components analysis and label consistency constraints. *Pattern Recognition* **39**:6, 1012-1025. [CrossRef]

149. C. Wang, W. Wang. 2006. Links Between PPCA and Subspace Methods for Complete Gaussian Density Estimation. *IEEE Transactions on Neural Networks* **17**:3, 789-792. [CrossRef]

150. A. Torokhti, P. Howlett. 2006. Optimal transform formed by a combination of nonlinear operators: the case of data dimensionality reduction. *IEEE Transactions on Signal Processing* **54**:4, 1431-1444. [CrossRef]

151. J. Cho, J.C. Principe, D. Erdogmus, M.A. Motter. 2006. Modeling and Inverse Controller Design for an Unmanned Aerial Vehicle Based on the Self-Organizing Map. *IEEE Transactions on Neural Networks* **17**:2, 445-460. [CrossRef]

152. FENG ZHANG. 2006. NONLINEAR FEATURE EXTRACTION AND DIMENSION REDUCTION BY POLYGONAL PRINCIPAL CURVES. *International Journal of Pattern Recognition and Artificial Intelligence* **20**:01, 63-78. [CrossRef]

153. Alexandre X. Carvalho, Martin A. Tanner. 2006. Modeling nonlinearities with mixtures-of-experts of time series models. *International Journal of Mathematics and Mathematical Sciences* **2006**, 1-23. [CrossRef]

154. Wei Hong, John Wright, Kun Huang, Yi Ma. 2006. Multiscale Hybrid Linear Models for Lossy Image Representation. *IEEE Transactions on Image Processing* **15**:12, 3655-3671. [CrossRef]

155. Christian W. Hesse, Christopher J. James. 2006. On Semi-Blind Source Separation Using Spatial Constraints With Applications in EEG Analysis. *IEEE Transactions on Biomedical Engineering* **53**:12, 2525-2534. [CrossRef]

156. M GALES, S AIREY. 2006. Product of Gaussians for speech recognition. *Computer Speech & Language* **20**:1, 22-40. [CrossRef]

157. Zhimin Fan, Jie Zhou, Ying Wu. 2006. Multibody grouping by inference of multiple subspaces from high-dimensional data using oriented-frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**:1, 91-105. [CrossRef]

158. FENG ZHANG, ZHUJUN WENG. 2005. MIXTURE PRINCIPAL COMPONENT ANALYSIS MODEL FOR MULTIVARIATE PROCESSES MONITORING. *Journal of Advanced Manufacturing Systems* **04**:02, 151-166. [CrossRef]

159. 2005. Robust Real-time Intrusion Detection System. *Journal of Information Processing Systems* **1**:1, 9-13. [CrossRef]

160. M. Aladjem. 2005. Projection pursuit mixture density estimation. *IEEE Transactions on Signal Processing* **53**:11, 4376-4383. [CrossRef]

161. S. Benameur, M. Mignotte, F. Destrempes, J.A. DeGuise. 2005. Three-Dimensional Biplanar Reconstruction of Scoliotic Rib Cage Using the Estimation of a Mixture of Probabilistic Prior Models. *IEEE Transactions on Biomedical Engineering* **52**:10, 1713-1728. [CrossRef]

162. N. Balakrishnan, K. Hariharakrishnan, D. Schonfeld. 2005. A new image representation algorithm inspired by image submodality models, redundancy reduction, and learning in biological vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**:9, 1367-1378. [CrossRef]

163. Heiko Hoffmann, Wolfram Schenck, Ralf Möller. 2005. Learning visuomotor transformations for gaze-control and grasping. *Biological Cybernetics* **93**:2, 119-130. [CrossRef]

164. K. Honda, H. Ichihashi. 2005. Regularized linear fuzzy clustering and probabilistic PCA mixture models. *IEEE Transactions on Fuzzy Systems* **13**:4, 508-516. [CrossRef]

165. B.A. Draper, D.L. Elliott, J. Hayes, K. Baek. 2005. EM in High-Dimensional Spaces. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)* **35**:3, 571-577. [CrossRef]

166. C. F. Beckmann, M. DeLuca, J. T. Devlin, S. M. Smith. 2005. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**:1457, 1001-1013. [CrossRef]

167. P. Kenny, G. Boulianne, P. Dumouchel. 2005. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing* **13**:3, 345-354. [CrossRef]

168. Z HALBE, M ALADJEM. 2005. Model-based mixture discriminant analysis?an experimental study. *Pattern Recognition* **38**:3, 437-440. [CrossRef]

169. Alexandre X. Carvalho, Martin A. Tanner. 2005. Modeling nonlinear time series with local mixtures of generalized linear models. *Canadian Journal of Statistics* **33**:1, 97-113. [CrossRef]

170. Tae-Kyun Kim, J. Kittler. 2005. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**:3, 318-327. [CrossRef]
171. A.X. Carvalho, M.A. Tanner. 2005. Mixtures-of-Experts of Autoregressive Time Series: Asymptotic Normality and Model Specification. *IEEE Transactions on Neural Networks* **16**:1, 39-56. [CrossRef]
172. Janos Abonyi, Balazs Feil, Sandor Nemeth, Peter Arva. 2005. Modified Gath–Geva clustering for fuzzy segmentation of multivariate time-series. *Fuzzy Sets and Systems* **149**:1, 39-56. [CrossRef]
173. Bruce A. Draper, Kyungim Baek, Jeff Boody. 2004. Implementing the expert object recognition pathway. *Machine Vision and Applications* **16**:1, 27-32. [CrossRef]
174. Gunther Heidemann, Robert Rae, Holger Bekel, Ingo Bax, Helge Ritter. 2004. Integrating context-free and context-dependent attentional mechanisms for gestural object reference. *Machine Vision and Applications* **16**:1, 64-73. [CrossRef]
175. Rozenn Dahyot, Pierre Charbonnier, Fabrice Heitz. 2004. A Bayesian approach to object detection using probabilistic appearance-based models. *Pattern Analysis and Applications* **7**:3, 317-332. [CrossRef]
176. Ezequiel López-Rubio, Juan Miguel Ortiz-de-Lazcano-Lobato, José Muñoz-Pérez, José Antonio Gómez-Ruiz. 2004. Principal Components Analysis Competitive Learning. *Neural Computation* **16**:11, 2459-2481. [Abstract] [PDF] [PDF Plus]
177. X HU, L XU. 2004. A comparative investigation on subspace dimension determination. *Neural Networks* **17**:8-9, 1051-1059. [CrossRef]
178. P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel. 2004. Speaker Adaptation Using an Eigenphone Basis. *IEEE Transactions on Speech and Audio Processing* **12**:6, 579-589. [CrossRef]
179. Qiang Wang, Hai-Zhou Ai, Guang-You Xu. 2004. Learning-based tracking of complex non-rigid motion. *Journal of Computer Science and Technology* **19**:4, 489-500. [CrossRef]
180. G Heidemann. 2004. Combining spatial and colour information for content based image retrieval. *Computer Vision and Image Understanding* **94**:1-3, 234-270. [CrossRef]
181. C.F. Beckmann, S.M. Smith. 2004. Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. *IEEE Transactions on Medical Imaging* **23**:2, 137-152. [CrossRef]
182. D. Hoyle, M. Rattray. 2004. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physical Review E* **69**:2. . [CrossRef]
183. Hyun-Chul Kim, Daijin Kim, Sung Yang Bang, Sang-Youn Lee. 2004. Face recognition using the second-order mixture-of-eigenfaces method. *Pattern Recognition* **37**:2, 337-349. [CrossRef]
184. M. Davies, N. Mitianoudis. 2004. Simple mixture model for sparse overcomplete ICA. *IEE Proceedings - Vision, Image, and Signal Processing* **151**:1, 35. [CrossRef]
185. C. Archer, T.K. Leen. 2004. A Generalized Lloyd-Type Algorithm for Adaptive Transform Coder Design. *IEEE Transactions on Signal Processing* **52**:1, 255-264. [CrossRef]
186. M Kim. 2003. Face recognition using the embedded HMM with second-order block-specific observations. *Pattern Recognition* **36**:11, 2723-2735. [CrossRef]
187. J.G. Brankov, N.P. Galatsanos, Yongyi Yang, M.N. Wernick. 2003. Segmentation of Dynamic PET or fMRI Images Based on a Similarity Metric. *IEEE Transactions on Nuclear Science* **50**:5, 1410-1414. [CrossRef]
188. Shigeyuki Oba, Masa-aki Sato, Shin Ishii. 2003. Variational Bayes method for Mixture of Principal Component Analyzers. *Systems and Computers in Japan* **34**:11, 55-66. [CrossRef]
189. X Liu. 2003. Eigenspace updating for non-stationary process and its application to face recognition. *Pattern Recognition* **36**:9, 1945-1959. [CrossRef]
190. B Draper. 2003. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding* **91**:1-2, 115-137. [CrossRef]
191. H KIM. 2003. An efficient model order selection for PCA mixture model. *Pattern Recognition Letters* **24**:9-10, 1385-1393. [CrossRef]
192. 2003. Modified Kernel PCA Applied To Classification Problem. *The KIPS Transactions:PartB* **10B**:3, 243-248. [CrossRef]
193. J Wu. 2003. Efficient face candidates selector for face detection. *Pattern Recognition* **36**:5, 1175-1186. [CrossRef]
194. H Kim. 2003. Extensions of LDA by PCA mixture model and class-wise features. *Pattern Recognition* **36**:5, 1095-1105. [CrossRef]
195. Z Liu. 2003. Improved system for object detection and star/galaxy classification via local subspace analysis. *Neural Networks* **16**:3-4, 437-451. [CrossRef]
196. Jong-Hoon Ahn, Jong-Hoon Oh. 2003. A Constrained EM Algorithm for Principal Component Analysis. *Neural Computation* **15**:1, 57-65. [Abstract] [PDF] [PDF Plus]

197. L XU. 2002. BYY harmony learning, structural RPCL, and topological self-organizing on mixture models1. *Neural Networks* **15**:8-9, 1125-1151. [CrossRef]

198. B. Moghaddam. 2002. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**:6, 780-788. [CrossRef]

199. P. Tino, I. Nabney. 2002. Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**:5, 639-656. [CrossRef]

200. M.A.F. Figueiredo, A.K. Jain. 2002. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**:3, 381-396. [CrossRef]

201. J. Denzler, C.M. Brown. 2002. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**:2, 145-157. [CrossRef]

202. M.J.F. Gales. 2002. Maximum likelihood multiple subspace projections for hidden Markov models. *IEEE Transactions on Speech and Audio Processing* **10**:2, 37-47. [CrossRef]

203. M. Sadeghi, J. Kittler, K. Messer. 2002. Modelling and segmentation of lip area in face images. *IEE Proceedings - Vision, Image, and Signal Processing* **149**:3, 179. [CrossRef]

204. Wayne S. DeSarbo, Alexandru M. Degeratu, Michel Wedel, M. Kim Saxton. 2001. The Spatial Representation of Market Information. *Marketing Science* **20**:4, 426-441. [CrossRef]

205. Dong Kook Kim, Nam Soo Kim. 2001. Rapid speaker adaptation using probabilistic principal component analysis. *IEEE Signal Processing Letters* **8**:6, 180-183. [CrossRef]

206. Akio Utsugi, Toru Kumagai. 2001. Bayesian Analysis of Mixtures of Factor Analyzers. *Neural Computation* **13**:5, 993-1002. [Abstract] [PDF] [PDF Plus]

207. A.P. Torokhti, P.G. Howlett. 2001. Optimal fixed rank transform of the second degree. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing* **48**:3, 309-316. [CrossRef]

208. Peter Meinicke, Helge Ritter. 2001. Resolution-Based Complexity Control for Gaussian Mixture Models. *Neural Computation* **13**:2, 453-475. [Abstract] [PDF] [PDF Plus]

209. Dirk Husmeier. 2000. The Bayesian Evidence Scheme for Regularizing Probability-Density Estimating Neural Networks. *Neural Computation* **12**:11, 2685-2717. [Abstract] [PDF] [PDF Plus]

210. M.A. Carreira-Perpinan. 2000. Mode-finding for mixtures of Gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**:11, 1318-1323. [CrossRef]

211. R. Everson, S. Roberts. 2000. Inferring the eigenvalues of covariance matrices from limited, noisy data. *IEEE Transactions on Signal Processing* **48**:7, 2083-2091. [CrossRef]

212. Yue Wang, Lan Luo, M.T. Freedman, Sun-Yuan Kung. 2000. Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization. *IEEE Transactions on Neural Networks* **11**:3, 625-636. [CrossRef]