Dimensionality Reduction of Clustered Data Sets

Guido Sanguinetti

Abstract—We present a novel probabilistic latent variable model to perform linear dimensionality reduction on data sets which contain clusters. We prove that the maximum likelihood solution of the model is an unsupervised generalization of linear discriminant analysis. This provides a completely new approach to one of the most established and widely used classification algorithms. The performance of the model is then demonstrated on a number of real and artificial data sets.

Index Terms—Dimensionality reduction, clustering, discriminant analysis, probabilistic algorithms.

1 Introduction

Dimensionality reduction techniques form an important chapter in statistical machine learning. Linear methods such as principal component analysis (PCA) and multidimensional scaling are classical statistical tools and more sophisticated techniques such as latent variable models and Independent Component Analysis (ICA) have attracted much interest in the past two decades. While a lot of current research focuses on nonlinear dimensionality reduction techniques, linear methods are still widely used in a number of applications, from computer vision to bioinformatics.

An important advance in the understanding of dimensionality reduction was provided by Tipping and Bishop's probabilistic interpretation of PCA (PPCA) [11]. The starting point for PPCA is in defining a factor-analysis style latent variable model

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

Here, $\mathbf{y}$ is a D-dimensional vector, W is a tall, thin matrix with D rows and q columns (with $q < D$), $\mathbf{x}$ is a q-dimensional latent variable, $\boldsymbol{\mu}$ is a constant (mean) vector, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I)$ is an error term. Placing a spherical unitary normal prior on the latent variable $\mathbf{x}$, Tipping and Bishop proved that, having observed i.i.d. data $\mathbf{y}_i$, at maximum likelihood the columns of W spanned the principal subspace of the data. Therefore, the generative model (1) could be said to provide a probabilistic equivalent of PCA.

This result opened the way for a number of applications and generalizations: It allowed a principled treatment of missing data and mixtures of PCA models [12], noise propagation through the model [9] and, through different choices of priors, allowed the linear dimensionality reduction framework to be extended to cases where the data could not be modeled as Gaussian [6].

While PCA is indubitably one of the most widely used statistical tools, it is important to understand its limitations. Since PCA is based entirely on matching the data covariance, it may not be appropriate in cases where the major interest lies in the structure of the data, rather than its covariance.

In many cases, there is some prior knowledge available about the structure of the data; often, it is known a priori that the data may have a clustered structure. For example, in computer vision one is often presented with images which consist of different parts (background and several objects) without explicit class labels of which pixel/area correspond to which part of the image. Similarly, one might consider using genome-wide gene expression measurement to discriminate between different conditions that are difficult to diagnose (a famous example is [8] which classified two different types of leukaemia which have very similar phenotypes). In these cases, dimensionality reduction based solely on the data covariance may not lead to good visualizations of the data set.

The problem is even more significant if we view dimensionality reduction as a feature extracting technique. Feature extraction techniques based on the covariance of the data (such as PCA or factor analysis) do not necessarily give meaningful features when the data set contains clusters. The most important features in this case are clearly the ones that discriminate between clusters, which, in general, do not coincide with the directions of greatest variation in the whole data set.

While unsupervised techniques for dimensionality reduction of data sets with clusters are relatively understudied, the supervised case is addressed by linear discriminant analysis (LDA). It was Fisher in a pioneering work [4] who first provided the solution to the problem of identifying the optimal projection in order to separate two classes. Under the assumption that the class conditional probabilities are Gaussian with identical covariance, he proved that the optimal one-dimensional projection to separate the data maximizes the *Rayleigh coefficient*

$$J(\mathbf{e}) = \frac{d^2}{\sigma_1^2 + \sigma_2^2}. \tag{2}$$

Here, $\mathbf{e}$ is the unit vector that defines the projection, d is the projected distance of the (empirical) class centers, and σ_i^2 is the projected empirical variance of class i. The generalization to the multiclass case is straightforward (see, e.g., [1]): defining the intraclass covariance matrix S_W (i.e., the sum of the empirical covariance matrices of the various classes) and the interclass covariance matrix S_B (i.e., the sum of the exterior products of the vectors connecting the class means), the optimal projection onto a q-dimensional subspace is given by selecting the q generalized eigenvectors with largest eigenvalues for the matrices S_B and S_W. Since the interclass covariance matrix has a rank equal to the number of classes K minus 1, it follows that the maximal value for q is $K - 1$.

In this contribution, we propose a latent variable model to perform dimensional reduction of clustered data sets in an unsupervised way. The maximum likelihood solution for the model fulfills an optimality criterion which is a generalization of Fisher's criterion, and reduces to maximizing Rayleigh's coefficient in the limit when cluster assignments become certain.

The rest of the paper is organized as follows: In the next section, we describe the latent variable model and provide an expectation-maximization (EM) algorithm for finding the maximum likelihood solution. We then prove that, under certain assumptions, the likelihood of the model is a monotonic function of Rayleigh's coefficient, so that LDA can be retrieved as a limiting case of our model. We then present experimental results on several data sets, both real and artificial. The results support the correctness of our theoretical derivation, giving good performance both for visualization and for classification. Finally, we discuss the advantages and limitations of our model, as well as possible generalizations and relationships with existing models.

2 Latent Variable Model

Let our data set be composed of N D-dimensional vectors $\mathbf{y}_i$, $i = 1, \ldots, N$. We have prior knowledge that the data set contains K clusters, and we wish to obtain a $q < D$-dimensionality reduction of the data which is optimal with respect to the clustered structure of the data. The latent variable model formulation we present is inspired by the extreme component analysis (XCA) model of [14]. We explain the data using two sets of latent variables which are

• *The author is with the Department of Computer Science, University of Sheffield, 211 Portobello Street, Sheffield, S1 4DP, UK. E-mail: guido@dcs.shef.ac.uk.*

Manuscript received 1 Mar. 2007; revised 17 July 2007; accepted 15 Oct. 2007; published online 22 Oct. 2007.
Recommended for acceptance by S. Chaudhuri.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-03-0137.
Digital Object Identifier no. 10.1109/TPAMI.2007.70819.

forced to map to orthogonal subspaces in the data; the model then takes the form

$$\mathbf{y} = V\mathbf{x} + \boldsymbol{\mu} + W\mathbf{z}. \tag{3}$$

In this equation, V is a $D \times q$ matrix whose columns span the subspace we wish to project on. XCA constrains V to have orthonormal columns; we will see later, though, that other constraints are more appropriate in our case, so for the time being V will simply have full column rank. $\boldsymbol{\mu}$ is a D-dimensional mean vector, and W is a $D \times (D-q)$ matrix whose columns span the orthogonal complement of the subspace spanned by the columns of V, i.e., $W^T V = 0$. Notice a key difference from the PPCA model of (1) is the absence of a noise term; this is common to our model and the XCA model.

The assumption that the data has a clustered structure is mirrored in our choice of prior distributions for the latent variables $\mathbf{x}$ and $\mathbf{z}$,

$$\begin{aligned} &\mathbf{z} \sim \mathcal{N}\big(\mathbf{0}, I_{D-q}\big), \\ &\mathbf{x} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}\big(\mathbf{m}_k, \sigma^2 I_q\big), \\ &\sum_{k=1}^{K} \pi_k = 1. \end{aligned} \tag{4}$$

Thus, we have a mixture of Gaussians generative distribution in the directions determined by the columns of V and a spherical Gaussian covariance structure in the orthogonal directions.

The motivation for these definitions is as follows: Intuitively, the mixture of Gaussians part will tend to optimize the fit to the clusters, which (given the spherical covariances) will be obtained when the subspace spanned by V interpolates the centers of the cluster in the optimal (least squares) way. Meanwhile, the general Gaussian covariance in the orthogonal directions will seek to maximize the total variance in the directions orthogonal to V, hence minimizing the sum of the projected variances. Therefore, the projection on the subspace spanned by V will seek an optimal compromise between separating the cluster centers and obtaining tight clusters.

From the definitions of the model and the priors, it is clear that we can set the mean vector $\boldsymbol{\mu}$ to zero and consider centred data sets without loss of generality. We will systematically do this in the following.

2.1 Likelihood

Given the model (3) and the prior distributions on the latent variables (4), it is straightforward to marginalize the latent variables and obtain a likelihood for the model. Using the orthogonality of V and W, we readily obtain a log-likelihood for the model (see the Appendix for a derivation, which can be found at http://computer.org/tpami/archives.htm)

$$\begin{aligned} &\mathcal{L}\big(\mathbf{y}_j, V, W, \mathbf{m}_k, \sigma^2, \pi\big) = -\frac{N}{2}\log|C| \\ &-\frac{1}{2}\mathrm{tr}\big(C^{-1}S\big) + \frac{N}{2}\log\Big|\big(\sigma^2 V^T V\big)^{-1}\Big| + \\ &\sum_{j=1}^{N}\log\sum_{k=1}^{K}\pi_k \exp\left[-\frac{\big(\mathbf{y_j} - V\mathbf{m}_k\big)^T \hat{C}\big(\mathbf{y_j} - V\mathbf{m}_k\big)}{2\sigma^2}\right]. \end{aligned} \tag{5}$$

Here, $S = \frac{1}{N}\sum_{j=1}^{N}\mathbf{y}_j\mathbf{y}_j^T$ is the empirical covariance of the (centered) data, $C = WW^T$, $\hat{C} = V\big(V^TV\big)^{-2}V^T$ and we have omitted a number of constants. With a slight notational abuse, we denote $C^{-1} = W\big(WW^T\big)^{-2}W^T$ (using the pseudoinverse of W) and $|C|$ as the product of the nonzero eigenvalues of C.

We notice immediately that, as W is unconstrained (except for being orthogonal to V and full column rank), at maximum likelihood WW^T will match *exactly* the covariance of the data in the directions perpendicular to V. Therefore, the second term in (5) will just be a constant $\frac{D-q}{2}$ irrespective of the other parameters and can be dropped from the likelihood. Also, we can use the orthogonality between V and W to rewrite $|C|$ in terms of V alone. This can be done by observing that, if P and $P_\perp$ represent two mutually orthogonal projections that sum to the identity, then the determinant of a symmetric matrix A can be expressed as $|A| = \big|PAP^T\big|\big|P_\perp AP_\perp^T\big|$. Introducing the matrix $E = V\big(V^TV\big)^{-\frac{1}{2}}$, which defines the orthonormal projection on the subspace spanned by V, we obtain that

$$|C| = \frac{|S|}{|E^TSE|}.$$

Substituting these results into the log-likelihood (5) and omitting terms which do not depend on the parameters, we obtain

$$\begin{aligned} &\mathcal{L}\big(\mathbf{y}_j, V, \mathbf{m}_k, \sigma^2, \pi,\big) = \\ &\frac{N}{2}\log\big|E^TSE\big| + \frac{N}{2}\log\Big|\big(\sigma^2 V^TV\big)^{-1}\Big| + \\ &\sum_{j=1}^{N}\log\sum_{k=1}^{K}\pi_k \exp\left[-\frac{\big(\mathbf{y_j} - V\mathbf{m}_k\big)^T \hat{C}\big(\mathbf{y_j} - V\mathbf{m}_k\big)}{2\sigma^2}\right]. \end{aligned} \tag{6}$$

This expression is now simply a mixture of Gaussian likelihood in the projected space spanned by the columns of V, plus a term that depends only on the projection E. This suggests a simple iterative strategy to maximise the likelihood.

2.2 E-M Algorithm

We can optimize the likelihood (6) using an E-M algorithm [5]. This is an iterative procedure that is proven to converge to a (possibly local) maximum of the likelihood. We introduce unobserved binary class membership vectors $\mathbf{c} \in \Re^K$; we can then define the *responsibilities*

$$\gamma_{jk} = \frac{\pi_k \exp\Big[-\frac{1}{2\sigma^2}\big(\mathbf{y_j} - V\mathbf{m}_k\big)^T \hat{C}\big(\mathbf{y_j} - V\mathbf{m}_k\big)\Big]}{\sum_{k=1}^{K}\pi_k \exp\Big[-\frac{1}{2\sigma^2}\big(\mathbf{y_j} - V\mathbf{m}_k\big)^T \hat{C}\big(\mathbf{y_j} - V\mathbf{m}_k\big)\Big]}.$$

We then obtain a lower bound on the log-likelihood in the form

$$\mathcal{Q}\big(\mathbf{y}, V, \mathbf{m}_k, \sigma^2, \pi\big) = \sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_{jk}\log\big\{\pi_k p\big(\mathbf{y}_j|\mathbf{c}_j\big)\big\} + \frac{N}{2}\log\big|E^TSE\big|, \tag{7}$$

where

$$p\big(\mathbf{y}_j|\mathbf{c}_j\big) = \frac{\exp\left[-\frac{\big(\mathbf{y_j} - V\mathbf{m}_k\big)^T \hat{C}\big(\mathbf{y_j} - V\mathbf{m}_k\big)}{2\sigma^2}\right]}{\sqrt{2\pi|\sigma^2 V^TV|}}$$

is the mixture component corresponding to class k (i.e., $\mathbf{c_j}$ is zero except for position k). Equation (7) can be interpreted as the expectation of the joint likelihood taken under the posterior probability of class membership.

Optimizing the bound (7), we obtain explicit values for the parameters. For the latent means, these are

$$\mathbf{m}_k = \big(V^TV\big)^{-1}V^T\boldsymbol{\mu}_k = \big(V^TV\big)^{-1}V^T\frac{\sum_{j=1}^{N}\gamma_{jk}\mathbf{y}_j}{\sum_{j=1}^{N}\gamma_{jk}}. \tag{8}$$

Similarly, we can easily obtain an update for the mixing coefficients

$$\pi_k = \frac{N_k}{N} = \frac{\sum_{j=1}^{N}\gamma_{jk}}{N}. \tag{9}$$

Substituting (8) into (7) we obtain, ignoring terms not depending on V and σ^2,

$$\begin{aligned}\mathcal{Q}(V,\sigma^2) &= \frac{N}{2}\log\left|\left(V^TV\right)^{-1}\right| \\ &- \frac{Nq}{2}\log(\sigma^2) + \frac{N}{2}\log\left|E^TSE\right| \\ &- \frac{N}{2\sigma^2}\operatorname{trace}\left[\left(V^TV\right)^{-1}V^TS_KV\left(V^TV\right)^{-1}\right],\end{aligned} \tag{10}$$

where

$$S_K = \frac{1}{N}\sum_{k=1}^{K}\left\{\sum_{j=1}^{N}\gamma_{jk}\left(\mathbf{y}_j-\boldsymbol{\mu}_k\right)\left(\mathbf{y}_j-\boldsymbol{\mu}_k\right)^T\right\},$$

where $\boldsymbol{\mu}_k$ was defined in (8). Optimization with regard to σ^2 is then straightforward and leads to

$$\sigma^2 = \frac{1}{q}\operatorname{trace}\left[\left(V^TV\right)^{-1}V^TS_KV\left(V^TV\right)^{-1}\right]. \tag{11}$$

Substituting (11) into (10) we can now obtain an expression that depends on V alone. Introducing the matrix $\hat{E} = V\left(V^TV\right)^{-1} = E\left(V^TV\right)^{-\frac{1}{2}}$ and using elementary properties of logarithms and determinants, (10) simplifies to

$$\mathcal{Q}(V) = -\log\left[\frac{1}{q}\operatorname{trace}\left(\hat{E}^TS_K\hat{E}\right)\right] + \frac{1}{q}\log\left|\hat{E}^TS\hat{E}\right|. \tag{12}$$

Notice the similarity of (12) with (20) in [11]. Maximization of (12) with respect to $\hat{E}$ reduces to finding the q generalized eigenvectors with largest generalized eigenvalues of the matrices S and S_K. Care must be excercised, however, in the implementation of this step: The usual constraint in generalized eigenvalue problems is that the columns of $\hat{E}$ be orthonormal with respect to the matrix S_K, but the matrix S_K will change after each update of the responsibilities. To avoid this problem, we will impose orthonormality with respect to S, which is fixed, and look for the eigenvectors with the smallest eigenvalues.

Equation (12) is noteworthy. The matrix S_K is a soft-assignment, unsupervised analogue of the matrix S_W, the sum of intraclass variances appearing in the multiclass formulation of LDA. The matrix S can be decomposed as $S_K + S_T$ (see, e.g., [1]), where S_T is the generalization of the interclass covariance S_B to the unsupervised case. Therefore, the intuition motivating the model was indeed correct and the maximum likelihood estimates of the model parameters return the natural unsupervised generalization of LDA.

3 Relationship with Linear Discriminant Analysis

In this section, we will explicitly show how, in the limiting case of large separation between the clusters, the likelihood of the model becomes a monotonic function of the objective function of LDA. For clarity's sake, we will limit our discussion to the two classes, one latent dimension case ($K = 2, q = 1$); the general case is not significantly different. The projection is then defined by a single vector $\mathbf{e}$; since the length of the vector $\mathbf{e}$ is clearly irrelevant, we will choose here $\mathbf{e}^T\mathbf{e} = 1$ to simplify the computations, even if this differs (by a constant) from our previous choices.

If the clusters are well separated, assuming they are balanced in the data, the likelihood of (6) can be simplified neglecting the probability that points in one cluster were generated by the other cluster, obtaining

$$\mathcal{L} = N\log\sigma^{-2} + \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{N_k} -\frac{1}{2\sigma^2}\left[\left(\mathbf{y_j}-\mathbf{e}m_k\right)^T\mathbf{e}\right]^2 + N\log\left|\mathbf{e}^TS\mathbf{e}\right|, \tag{13}$$

where we replaced $\pi_i = \frac{1}{2}$, N_i is the number of points assigned cluster i and m_k are the latent means as in (4). Using the fact that $\sum_{j=1}^{N}\mathbf{y}_j = 0$, we define

$$\mathbf{y}_j = \boldsymbol{\mu_i} + \mathbf{v}_j, \tag{14}$$

where $\boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{y}_j$. The intracluster covariances are then $S_i = \frac{1}{N_i}\sum\mathbf{v}_j\mathbf{v}_j^T$. Since there are two balanced clusters, we have $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = \boldsymbol{\mu}$ and $m_1 = -m_2 = m$. Using (14), we obtain that

$$S = \frac{1}{2}\left(S_1 + S_2 + 2\boldsymbol{\mu}\boldsymbol{\mu}^T\right).$$

Defining $\sigma_i^2 = \mathbf{e}^TS_i\mathbf{e}$ the projected variance of cluster i and $d = 2\boldsymbol{\mu}^T\mathbf{e}$ the projected distance separating the clusters, it is easy to compute the maximum likelihood estimates for m and σ^2, given, respectively, by

$$m = d, \quad \sigma^2 = \alpha\left(\sigma_1^2 + \sigma_2^2\right)$$

with α a constant. Notice that these can also be obtained from (8)-(11) by taking the limit when the responsibilities are 0 or 1 (hard assignments). Ignoring constants, we obtain that the likelihood (13) can be rewritten as

$$\mathcal{L} = \log\left(\frac{\sigma_1^2+\sigma_2^2+2d^2}{\sigma_1^2+\sigma_2^2}\right) = \log[1+2J(\mathbf{e})],$$

where $J(\mathbf{e})$ is the Rayleigh coefficient of (2). Therefore, the maximum likelihood solution is obtained at the maximum of the Rayleigh coefficient, showing that the model reduces to Fisher's discriminant in these conditions.

4 Experimental Results

In this section, we examine the behavior of our model on a number of data sets. First, to clarify the way the model works, we demonstrate it on a toy data set similar to the one used in [1] to explain Fisher's Discriminant Analysis. We then perform experiments on three more challenging data sets. Two of them are taken from the Machine Learning Repository at UCI[1] while the third one is the USPS data set of handwritten digits.[2] Matlab code to recreate all the results shown, as well as the data sets used, can be freely downloaded from http://www.dcs.shef.ac.uk/~guido/software.html.

Before analyzing the results, it is important to stress the limitations and possible pitfalls of the model. First of all, a natural question when dealing with parameter estimations is whether the likelihood is unimodal. Obviously, whenever mixture models are involved, there is a trivial multimodality due to the permutations of the components. It is clear that, for data sets where the clusters are well separated, this is the only source of ambiguity in our model. However, the situation becomes more delicate when the clusters partially overlap; in this case, the likelihood can acquire multiple modes and plateaus, leading to potentially hard optimization problems.

Most importantly, we have proven in the previous section that the model can be viewed as a probabilistic, unsupervised version of LDA. As such, it shares all the limitations of LDA, aggravated by the use of incomplete information (absence of class labels). Therefore, the classification performance can be expected to be poor in cases when the classes are overlapped or nonconvex, or when the covariance structure of different clusters varies greatly. An example of a case where these assumptions are not met, and our model fails to produce a good visualization, is given in the Appendix, which can be found at http://computer.org/tpami/archives.htm.

As a toy example, we first consider a simple artificial data set consisting of two highly elongated clusters in two dimensions (Fig. 1a). Each cluster consists of 100 points and the covariance of each cluster is 36 times greater in the horizontal direction than in the vertical direction. This is a classical example when PCA would fail miserably; projecting onto the first principal component (shown in the figure by the dashed-dotted line) would lead to almost completely overlapping clusters. To obtain a reasonable

1. www.ics.uci.edu/~mlearn/MLRepository.html.
2. http://www.gaussianprocess.org/gpml/data/.

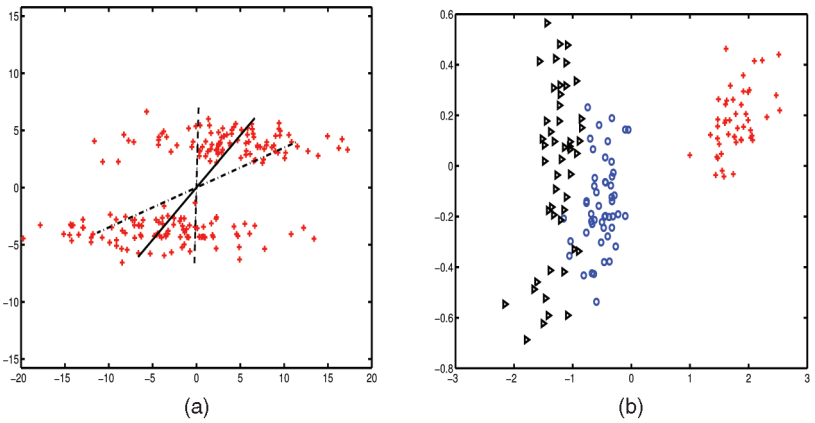

Fig. 1. Experimental results. (a) Toy data: The solid dotted line represents the first principal direction of the data; the solid line is the initialization (obtained using k-means followed by Fisher's discriminant); the dashed and dotted line gives the maximum likelihood estimate of the model, which coincides with (supervised) Fisher's discriminant. (b) Iris data set: The three classes are shown as triangles, crosses, and circles.

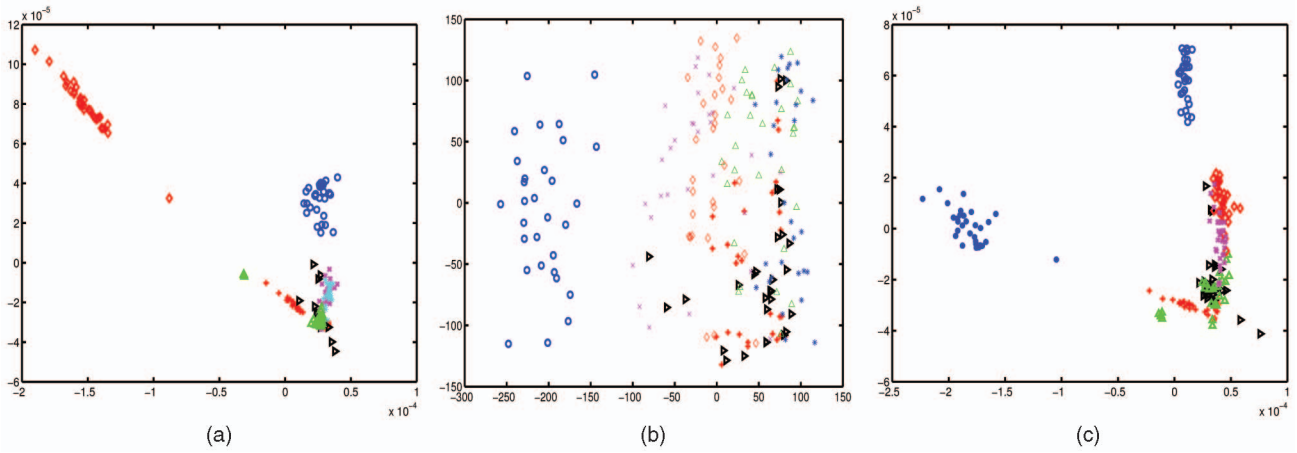

Fig. 2. Visualization results for the image data set: (a) projection given by our model, (b) projection given by PCA, and (c) projection obtained by supervised LDA.

initialization, we ran k-means (initialized at random data points) followed by LDA (using the cluster assignments as class labels). This procedure was followed in all the experiments.

The initial direction of the projection vector is shown in Fig. 1a by a solid line. As can be seen, this is a better projection than the one obtained by PCA, but fails to be optimal. The reason for this is that k-means implicitly assumes isotropic clusters; in the presence of highly elongated clusters (as is the case shown), it will break the clusters, hence causing the application of LDA to find a suboptimal projection. The results of our model are shown by the dashed and dotted line and coincide exactly with the projection obtained by LDA using the correct class labels. This is obviously not surprising since the clusters are well separated and have equal covariance matrix, precisely the conditions under which our model (and LDA) provide good results.

We then turn to the highly used *Iris* data set, initially used by Fisher in his groundbreaking paper [4], and still widely used as a benchmark (for a recent usage see [7]). This data set consists of 150 measurements of four different attributes of three different species of *Iris*, *I. versicolor*, *I. setosa*, and *I. virginica*. Each of the three classes contains 50 data points. One of the classes is well separated, the other two are partially overlapping. The results of applying our model to project in two dimensions are shown in Fig. 1b. The visualization gives one compact cluster for the separable class, the other two classes are partially overlapped. To assess the quality of the visualization, we used a One Nearest Neighbor (1NN) technique, which showed that approximately 96 percent of the points are nearest to a point within the same class. While this percentage is clearly quite high, it is very similar to the one obtained by PCA, and indeed to the one obtained by supervised LDA.

A more challenging data set is the *Image* data set, created in the nineties by the computer vision group at University of Massachusetts. This consists of 210 data points from seven different types of images: brick-face, sky, foliage, cement, window, path, and grass. Each data point consists of 19 different features extracted from the image using standard techniques. We removed from the data seven features that were obviously non-Gaussian distributed (such as edge detectors and pixel counts), and performed dimensionality reduction from 12 dimensions to two.

Fig. 2 shows the results of 1) our model, 2) compared with PCA, 3) and with supervised LDA. The first thing to notice is that PCA provides a very poor visualization of the data; one class is reasonably separated, but the other six are completely overlapped. The visualization provided by our model clearly shows the presence of more clusters; in particular, besides the circle class, the diamond class and the crosses class are quite well separated. The visualization obtained with our model is similar to the one obtained with supervised LDA, besides a permutation of classes. In order to give a more quantitative appreciation of the quality of the visualization, we again applied a 1NN classifier to the three projections. This confirmed that the projection given by PCA indeed does not respect the clustered structure; only 49.5 percent of the projected points are nearest to a point of the same class. The situation is dramatically better in our model, with the percentage of points closest to points in the same class rising to 74.3 percent. This comes very close indeed to the 75.7 percent obtained by LDA (which obviously employed the class information). It may be worthwhile reporting that the same test

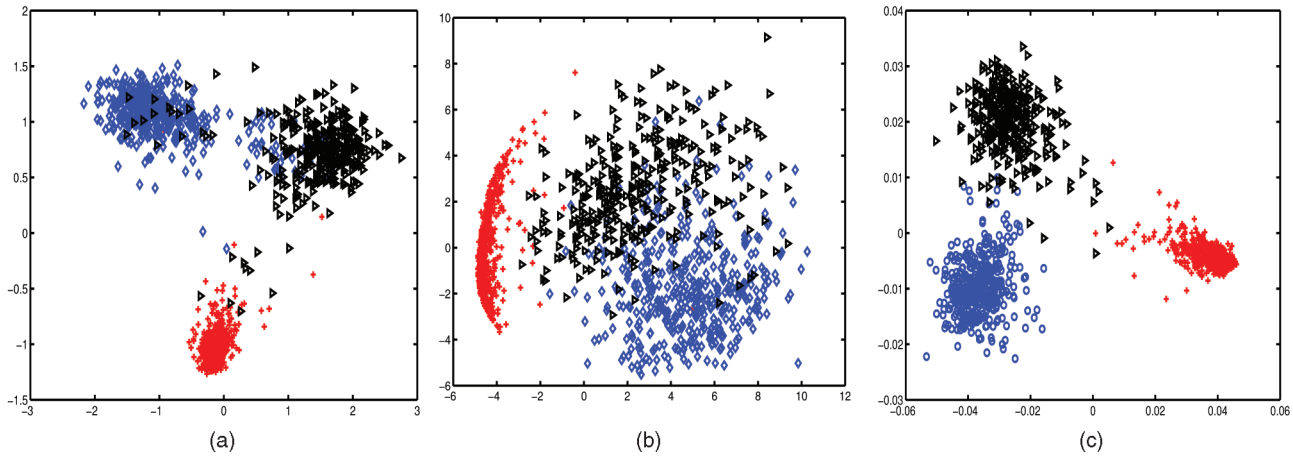

Fig. 3. Visualization results for the USPS digits data set: (a) Projection given by our model, (b) projection given by PCA, and (c) projection given by supervised LDA.

carried out on the whole 12-dimensional data set gives a result of 83.8, so there is a substantial level of overlap among some classes.

While the visualization results are very reasonable, the best separation among classes in LDA can be obtained projecting onto a $K-1$-dimensional space, in this case a six-dimensional space. We therefore ran our model in this case and considered the cluster assignments given naturally by the mixture responsibilities. The classification thus obtained had an accuracy of 64.7 percent. Most of the misclassifications were due to the algorithm grouping together the brick-face class with the cement class, which is quite plausible.

As a further example, we considered a subset of the USPS handwritten digits data set, which has long been used as benchmark for classification. This data set consists of 256-dimensional vectors representing gray level pixel intensities; we selected a subset of 1,396 points representing the digits 1, 3, and 8. The results are shown in Fig. 3; while there is some (unavoidable) overlap between the 3s and the 8s, the visualization given by our model in Fig. 3a is clearly superior to the visualization given by PCA Fig. 3b. In terms of 1NN classification, our model returns a 93 percent success rate against the 90 percent of PCA; however, the overall separation between classes is clearly much better. Interestingly, using the label information in supervised LDA on this data set (Fig. 3c) leads to a marked improvement in accuracy as measured by 1NN (99 percent), but does not lead to a significantly different visualization.

5 Discussion

In this paper, we have addressed the question of selecting an optimal linear projection of a high-dimensional data set, given prior knowledge of the presence of clusters in the data. The proposed solution is a generative model which is made up of two orthogonal latent variable models: One is an unconstrained, PPCA-like model [11], while the other has a mixture of Gaussians prior on the latent variables. The rationale behind this choice of priors is that the PPCA-like part will tend to maximize the variance of the data in the directions orthogonal to the projection, hence forcing the projected clusters to maximise the intercluster spread, while minimizing the intracluster variance. Indeed, we prove that the maximum likelihood estimator for the projection in our model satisfies an unsupervised, soft assignment generalization of the criterion for LDA. We also prove that when the projection is one dimensional and the clusters are well separated, the likelihood of the model reduces to a monotonic function of the Rayleigh coefficient used in Fisher's discriminant analysis.

Our model is related to several previously proposed models. It can be viewed as an adaptation of the XCA model of [14] to the case when the data set has a clustered structure; it is remarkable though that the generalized LDA criterion is obtained without imposing that the variance in the directions orthogonal to the projection be maximal. Bishop and Tipping [2] dealt with the problem of visualizing clustered data sets by proposing a hierarchical model based on a mixture of PPCAs [12]. While this approach addresses successfully the problem of analyzing the detailed structure of each cluster, it does not provide a single optimal projection for visualizing all the clusters simultaneously. Another approach where a dimensionality reduction is combined with a mixture model is [15]; however, the dimensionality reduction there is a random one, performed solely to avoid the curse of dimensionality. It would be interesting, however, to apply the estimation framework proposed in [15] in a case where the projection is not random, as an alternative to the EM framework proposed here.

Perhaps the contribution that is closest to ours is in [7]. This paper considered the clustering problem using an Independent Component Analysis (ICA) model with one latent binary variable corrupted by Gaussian noise. The author then proved that the minimum of the negative entropy for the model returned an unsupervised analogue of Fisher's discriminant. This approach is somewhat complementary to ours in that it uses discrete rather than continuous variables, but the end result is remarkably close. In fact, in the one dimensional, two classes case, the objective function of the ICA model is the same as the likelihood (6). The major difference comes when considering latent spaces which are more than one dimensional or data sets with more than two clusters, which are generally the most important cases. Generalizing the ICA approach would enforce an independence constraint between the directions of the projection which is not plausible; this is not a problem in our approach.

The performance of our model depends critically on some assumptions. Being a natural unsupervised version of LDA, it tends to perform well under similar conditions. Specifically, there is a Gaussian generative assumption for each of the clusters, with a shared covariance structure. While removing the constraint of equal covariances can be done easily by introducing latent covariances (at a moderate computational cost), the assumption of normality is harder to remove. Also, since the algorithm is unsupervised, overlapped clusters may be difficult to detect, and lead to local optima in the likelihood. Notice, however, that the equal covariance assumption removes the unboundedness of the likelihood, a traditional weakness of mixture models.

There are several interesting directions for generalizing the present work. One possibility would be to extend it to nonlinear dimensional reductions techniques such as the GTM or the GPLVM [3], [10]. While this is in principle possible due to the shared probabilistic nature of these models, in practice it will require considerable further work. Another important direction would be to automate model selection issues (such as the number of clusters or number of latent dimensions) using Bayesian techniques such as Bayesian PCA [13]. While this is feasible, careful consideration should be given to issues of numerical efficiency. Finally, our model

provides an unsupervised, probabilistic extension of LDA; by combining it with a discriminative, supervised part, it should be relatively simple to obtain a semisupervised version of LDA, which could be useful in many applications.

ACKNOWLEDGMENTS

The author would like to thank Mahesan Niranjan and Mark Girolami for useful discussions and suggestions.

REFERENCES

[1] C.M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.
[2] C.M. Bishop and M.E. Tipping, "A Hierarchical Latent Variable Model for Data Visualisation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 3, pp. 281-293, Mar. 1998.
[3] C.M. Bishop, M. Svensen, and C.K.I. Williams, "GTM: The Generative Topographic Mapping," *Neural Computation,* vol. 10, no. 1, pp. 215-234, 1998.
[4] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics,* vol. 7, pp. 179-188, 1936.
[5] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc. B,* vol. 39, no. 1, pp. 1-38, 1977.
[6] M. Girolami and R. Breitling, "Biologically Valid Linear Factor Models of Gene Expression," *Bioinformatics,* vol. 20, no. 17, pp. 3021-3033, 2004.
[7] M. Girolami, "Latent Class and Trait Models for Data Classification and Visualisation," *Independent Component Analysis: Principles and Practice,* Cambridge Univ. Press, 2001.
[8] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science,* vol. 286, no. 5439, pp. 531-537, 1999.
[9] G. Sanguinetti, M. Milo, M. Rattray, and N.D. Lawrence, "Accounting for Probe-Level Noise in Principal Component Analysis of Microarray Data," *Bioinformatics,* vol. 21, no. 19, pp. 3748-3754, 2005.
[10] N.D. Lawrence, "Probabilistic Non-Linear Principal Component Analysis with Gaussian Process Latent Variable Models," *J. Machine Learning Research,* vol. 6, pp. 1783-1816, 2005.
[11] M. Tipping and C.M. Bishop, "Probabilistic Principal Component Analysis," *J. Royal Statistical Soc. B,* vol. 21, no. 3, pp. 611-622, 1999.
[12] M. Tipping and C.M. Bishop, "Mixtures of Probabilistic Principal Component Analyzers," *Neural Computation,* vol. 11, no. 2, pp. 443-482, 1999.
[13] C.M. Bishop, "Bayesian PCA," *Proc. Advances in Neural Information Processing Systems,* 1999.
[14] M. Welling, F. Agakov, and C.K.I. Williams, "Extreme Component Analysis," *Proc. Advances in Neural Information Processing Systems,* 2003.
[15] S. Dasgupta, "Learning Mixture of Gaussians," *Proc. 40th Ann. IEEE Symp. Foundations of Computer Science,* 1999.