DERIVING CLUSTER ANALYTIC DISTANCE FUNCTIONS FROM GAUSSIAN MIXTURE MODELS

Michael E. Tipping

Microsoft Research, St George House, 1 Guildhall Street, Cambridge, CB2 3NH, U.K.

ABSTRACT

The reliable detection of clusters in datasets of non-trivial dimensionality is notoriously difficult. Clustering algorithms are generally driven by some distance function (usually Euclidean) defined over pairs of examples, which implicitly treats distances within and between clusters alike. In this paper, a more effective distance measure is proposed, derived from an *a priori* estimated Gaussian mixture model. Examples illustrate how the proposed approach can effectively de-emphasise within-cluster structure, and thus implicitly magnify the separation between regions of high data density.

INTRODUCTION

Finding genuine clusters within multivariate data is a challenging task. While there is inevitable ambiguity over what precisely constitutes a cluster[1], we are generally searching for distinguished modes of the data density. There are a variety of established cluster analytic approaches, ranging from traditional hierarchical linkage methods to probabilistic mixture models. Hierarchical methods are relatively heuristic and are often observed to perform poorly [4], while it is difficult to verify the correct number of clusters in a mixture model.[2]

One approach to cluster detection is through visualisation, where a two (or three) dimensional representation of the data is generated for human study. Clusters can often be readily distinguished by the eye from such plots, but the challenge is to generate projections which retain the relevant detail, and which do not obscure genuine clusters or introduce artifacts. Interactive visualisation methods have been developed with cluster analysis in mind, ranging from 'projection pursuit' techniques [3] to the more recent application of hierarchical mixtures of latent variable models [2].

In this paper, we propose the use of a Gaussian mixture model as a pre-processing stage for visual cluster detection. Specifically, we will exploit the model to derive a distance function over the data space, which we will utilise to visualise the examples (although the utility of the technique is not confined exclusively to visualisation). A key feature of the approach is that employing the 'correct' number of mixture components is not critical.

To motivate the problem we are trying to tackle, consider noisy data distributed over two parallel square planes in three dimensions. One of the simplest, and most commonly used, methods for obtaining a two-dimensional representation of data is principal component analysis (PCA). In this case however, the planes are sufficiently close such that the linear PCA projection shown in Fig. 1 effectively collapses the two planes together.

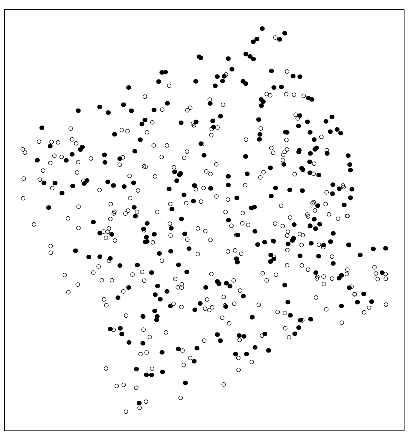

Figure 1: *Principal component projection of 200 data points located on two parallel planes, data from each being denoted by filled and open circles respectively.*

The projection of Fig. 1 fails to reveal the interesting structure of the data as the variance within each plane is treated identically to variance between the planes. Intuitively, we would like to reduce the influence of the within-plane variance in favour of that between the planes, which implies basing our clustering al-

[1] It has been said that the definition of a cluster is "that which is output by a clustering algorithm".

[2] We might use cross-validation to optimise the number of mixture components, but this will only be in the *predictive* sense — i.e. the number of components which gives the best likelihood on the validation set. This needn't correspond to the 'true' number of clusters in the data, as neatly demonstrated by Smyth and Wolpert [7]. Bayesian methods behave no differently in this respect either.

gorithms on some non-Euclidean, and perhaps non-global, distance function. To this end, we can see that this example dataset might be effectively modelled with a two-component Gaussian mixture, where the covariance structure of the components conveys much of the necessary information that we require. The primary aim of this paper is therefore to elucidate a mechanism for effectively exploiting this information to derive a superior distance measure for clustering purposes.

We begin, in the next section, by considering two candidate approaches, based first on linear discriminant analysis and second on recently developed Fisher kernel methodology [5]. The disadvantages of these two methods are briefly outlined, before a new distance function is detailed. This measure is exploited as a basis to obtain visualisations, and examples of application to real and synthetic data sets are given.

OBTAINING DISTANCE FUNCTIONS FROM MIXTURE MODELS

Consider a data set $\{\mathbf{x}_n\}_{n=1}^N$ in d dimensions, to which has been fitted a Gaussian mixture with K components with mixing proportions, means and covariance matrices given by π_k, $\boldsymbol{\mu}_k$ and $\mathbf{C}_k$ respectively. We seek to exploit the information within the mixture model to derive an effective visualisation in two dimensions. As primarily we are motivated by the desire to detect clusters, we begin by assuming that the number of mixture components is reflective of genuine structure in the dataset, and consider how therefore to emphasise the distance between them.

Linear discriminant analysis

One straightforward approach would be to perform a linear discriminant analysis on the clusters, where a $d \times 2$ projection matrix $\mathbf{U}$ is sought which minimises within-component variance while maximising the scatter between component means in the space of projections $\mathbf{U}^{\mathrm{T}}\mathbf{x}_n$. This can be achieved by maximising a measure such as trace $\left[(\mathbf{U}^{\mathrm{T}}\mathbf{C}_W\mathbf{U})^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{C}_B\mathbf{U})\right]$, where

$$\mathbf{C}_W = \sum_{k=1}^{M} \pi_k \mathbf{C}_k,$$

$$\mathbf{C}_B = \sum_{k} \pi_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^{\mathrm{T}},$$

and $\bar{\boldsymbol{\mu}} = \sum_k \boldsymbol{\mu}_k / K$. It can be shown that the trace criterion is optimised when the columns of $\mathbf{U}$ are the leading eigenvectors of $\mathbf{C}_W^{-1}\mathbf{C}_B$.

This approach is effective for the earlier parallel planes example, but it has two primary disadvantages. First, it relies on the mixture model comprising the 'correct' number of components, and second, its linearity limits its flexibility when applied to more complex datasets.

As suggested in the introduction, exploitation of the mixture model structure might be achieved by defining an appropriate non-Euclidean distance function over the example set. In the above case, the squared distance between examples $\mathbf{x}_i$ and $\mathbf{x}_j$ in the projection is given by

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{U}\mathbf{U}^{\mathrm{T}})(\mathbf{x}_i - \mathbf{x}_j).$$

The projection thus implicitly represents an optimal (in the distance-preserving sense) embedding of the original data given this simplistic model-derived metric.

The 'Fisher kernel' method

Another potential approach to extracting a more complex distance measure from a probability model is offered by the *Fisher kernel* approach of Jaakkola and Haussler [5]. The authors' consider how a generative model of the data can be utilised to derive a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ to act as an inner product between examples which can be plugged into, say, a support vector classifier. Although the ideas were developed for discrimination, the authors note that the principle might also be applied to clustering.

Consider a generative model $p(\mathbf{x}|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$. The 'Fisher kernel' is defined as $\mathbf{s}(\mathbf{x}_i)^{\mathrm{T}}\mathbf{F}^{-1}\mathbf{s}(\mathbf{x}_j)$, where $\mathbf{s}(\mathbf{x})$ is the vector of partial derivatives of the log-likelihood of $\mathbf{x}$ (known as the 'Fisher score'), i.e. $s_\alpha(\mathbf{x}) = \partial \log p(\mathbf{x}|\boldsymbol{\theta})/\partial\theta_\alpha$. The matrix $\mathbf{F}$ is the Fisher information matrix, $E\left[\mathbf{s}(\mathbf{x})\mathbf{s}(\mathbf{x})^{\mathrm{T}}\right]$.

This Fisher kernel induces a distance function in the data space given by:

$$d_{ij}^2 = \left[\mathbf{s}(\mathbf{x}_i) - \mathbf{s}(\mathbf{x}_j)\right]^{\mathrm{T}}\mathbf{F}^{-1}\left[\mathbf{s}(\mathbf{x}_i) - \mathbf{s}(\mathbf{x}_j)\right].$$

As an example, consider the simple case of a univariate Gaussian distribution with parameters $\boldsymbol{\theta} = \{\mu, \sigma^2\}$. Here, d_{ij}^2 is given by:

$$\frac{(x_i - x_j)^2}{\sigma^2} + \left[\frac{(x_i-\mu)^2}{2\sigma^2} - \frac{(x_j-\mu)^2}{2\sigma^2}\right]^2. \quad (1)$$

Advantageous features of this distance function are that it is applicable to all types of generative models (i.e. not just to Gaussian mixtures), and it is invariant to invertible transformations of the model parameters.

The Fisher distance function was applied to clustering using Gaussian mixture models, but proved ineffective. We may understand why this is so by considering the case of a two-component model of two well-separated clusters of univariate data. For two examples x_i and x_j located in different clusters, the derivatives of the log-likelihood at x_i with respect to the parameters of the Gaussian modelling x_j will be negligible, and vice-versa. (This can be seen by inspection of the standard parameter derivatives for a Gaussian mixture, e.g. as given in [1].) Thus, the Fisher score vectors $\mathbf{s}(x_i)$ and $\mathbf{s}(x_j)$ will be orthogonal, and the Fisher information matrix will be diagonal (block diagonal in the multivariate case). It follows that the distance function will comprise terms in x_i^2 and x_j^2, but not cross-product terms of the form $x_i x_j$. This is undesirable, as it implies that the Fisher distance must be independent of correlations between the examples, and even well separated data does not cluster with respect to the Fisher distance. In the following section we consider the form that a more appropriate metric might take.

A model-weighted distance

Consider data modelled by a single multivariate Gaussian with covariance matrix $\mathbf{C}$. The 'natural' measure of distance based on this model between examples $\mathbf{x}_i$ and $\mathbf{x}_j$ should intuitively be given by the squared Mahalanobis distance:

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{x}_j). \qquad (2)$$

This is a distance measured in terms of a rotated and scaled coordinate system whose axes are perpendicular to lines of constant probability density. Furthermore, given maximum-likelihood estimation of $\mathbf{C}$, d_{ij}^2 is invariant to non-singular linear transformation of the variables. This feature is important in cluster analysis, since the results should not depend on the units of measurement of a given coordinate.

We cannot apply (2) to the mixture case, as we cannot identify a unique $\mathbf{C}$. To extend the concept of distance as applied above, the data space is interpreted as a Riemannian manifold with metric $\mathbf{G}(\mathbf{x})$ which is defined to be:

$$\mathbf{G}(\mathbf{x}) = \sum_{k=1}^{K} p(k|\mathbf{x}) \mathbf{C}_k^{-1}. \qquad (3)$$

The metric at $\mathbf{x}$ is thus the average of the individual component metrics (inverse covariances), weighted by the posterior probability of each component at $\mathbf{x}$. This definition is consistent with (2), but is nevertheless inevitably heuristic in character. From this, the distance along the straight line from $\mathbf{x}_i$ to $\mathbf{x}_j$ is given by:

$$d_{ij} = \int_0^1 \left[\nabla \mathbf{x}(t)^{\mathrm{T}} \mathbf{G}(t) \nabla \mathbf{x}(t) \right]^{1/2} dt, \qquad (4)$$

where the line is parameterised as $\mathbf{x} = \mathbf{x}_i + t(\mathbf{x}_j - \mathbf{x}_i)$ and $\nabla \mathbf{x}(t) = (\partial x_1/\partial t, \ldots, \partial x_d/\partial t)$. Unfortunately, this integral is analytically intractable and, since we must calculate it $O(N^2)$ times, it is appropriate to seek some workable (non-numeric) approximation. We make two simplifications, by first taking the square root outside the integral, and then averaging the inverse covariances with respect to $p(\mathbf{x}|k)\pi_k$ rather than the posteriors $p(k|\mathbf{x})$, giving:

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \mathbf{G} (\mathbf{x}_i - \mathbf{x}_j), \qquad (5)$$

where we now must compute

$$\mathbf{G} = \frac{\sum_{k=1}^{K} \mathbf{C}_k^{-1} \pi_k \int_{\mathbf{x}_i}^{\mathbf{x}_j} p(\mathbf{x}|k)\, d\mathbf{x}}{\sum_{k=1}^{K} \pi_k \int_{\mathbf{x}_i}^{\mathbf{x}_j} p(\mathbf{x}|k)\, d\mathbf{x}}. \qquad (6)$$

The term $\int_{\mathbf{x}_i}^{\mathbf{x}_j} p(\mathbf{x}|k)\, d\mathbf{x}$ represents the unidimensional integral of the probability density of component k along the straight line from $\mathbf{x}_i$ to $\mathbf{x}_j$ and is given by:

$$\int_{\mathbf{x}_i}^{\mathbf{x}_j} p(\mathbf{x}|k)\, d\mathbf{x} = \sqrt{\frac{\pi b^2}{2}} \exp\{-Z/2\} \times \left[\operatorname{erf}\left(\frac{1-a}{\sqrt{2b^2}} \right) - \operatorname{erf}\left(\frac{-a}{\sqrt{2b^2}} \right) \right], \qquad (7)$$

where $\operatorname{erf}(x) = 2/\sqrt{\pi} \int_0^x e^{-t^2} dt$, the error function, and

$$b^2 = \left(\mathbf{v}^{\mathrm{T}} \mathbf{C}_k^{-1} \mathbf{v} \right)^{-1}, \qquad (8)$$

$$a = b^2 \mathbf{v}^{\mathrm{T}} \mathbf{C}_k^{-1} \mathbf{u}, \qquad (9)$$

$$Z = \mathbf{u}^{\mathrm{T}} \mathbf{C}_k^{-1} \mathbf{u} - b^2 \left(\mathbf{v}^{\mathrm{T}} \mathbf{C}_k^{-1} \mathbf{u} \right)^2, \qquad (10)$$

with $\mathbf{v} = \mathbf{x}_i - \mathbf{x}_j$ and $\mathbf{u} = \boldsymbol{\mu}_k - \mathbf{x}_j$.

Equation (5) thus assumes a constant metric along the path between the data points, which is given by (6) as the probabilistically-weighted average of the inverse covariances of each mixture component.

The model-weighted distance function (5) has the following properties:

1. It is invariant to arbitrary nonsingular linear transformations of the variables (and therefore the choice of measurement units).

2. For the case of a single Gaussian, or in regions of space where only one component density $p(\mathbf{x}|k)$ is non-zero along $\mathbf{x}_i$ to $\mathbf{x}_j$, then d_{ij}^2 is as given by (2) in the single Gaussian case.

3. The measure is relatively robust to the number of component Gaussians in the mixture model. If in the parallel planes example earlier, one plane was fitted with

more than a single Gaussian, each Gaussian in that plane would have a similar inverse covariance matrix, and the weighted average would be similar to that of a single Gaussian fitted to the plane.

4. The measure is not robust to poor local likelihood maxima. If the density model is unrepresentative of the data structure (e.g. two Gaussians each straddling both planes in the above example), then poor performance is inevitable.

VISUALISATION OF DISTANCE RELATIONSHIPS

Although we can exploit the model-weighted distance measure in a general clustering algorithm, in this paper we concentrate on two-dimensional visualisation of data based on an optimal embedding of their inter-point distances. There are a number of ways in which a geometric representation of points can be derived from a distance matrix, but in the examples which follow, embeddings generated by the *Sammon mapping* (see [9]) will be used. Vectors $\mathbf{y}_i$ in the visualisation space are adjusted so as to minimise the criterion $\sum_{i,j}(\|\mathbf{y}_i - \mathbf{y}_j\| - d_{ij})^2$, with d_{ij} as defined by (5).

EXAMPLES OF APPLICATION

Synthetic Gaussian data

First we illustrate the application of the method to some three-dimensional synthetic data generated from a three component mixture model. Each generative Gaussian has the same covariance matrix, and is relatively flat (has small variance) in one dimension. One of these clusters is well separated from the other two, which lie 'on top of' each other. This data set is thus a slightly more complex version of the parallel planes case earlier, and Fig. 2 shows a two-dimensional 'side' elevation of 300 data points.

This 'toy' data set has been studied previously in the context of hierarchical interactive visualisation [2], and was deliberately contrived such that a standard PCA projection (Fig. 3) would fail to separate the clusters (as, indeed, would a Sammon mapping based on the Euclidean distances).

A three-component Gaussian mixture model (illustrated also in Fig. 2) was fitted to the data, and, using the model-weighted distance function, the embedding of Fig. 4 was obtained. In this case, the three clusters are as separated in the visualisation as can be expected, given the close proximity of two of them. Of course, in this example we have chosen the correct number of components in our model, so in Fig. 5 an equivalent plot is given for a model with 6 Gaussian components. There is some distortion compared to Fig. 4, but the presence of three clusters is still clear.

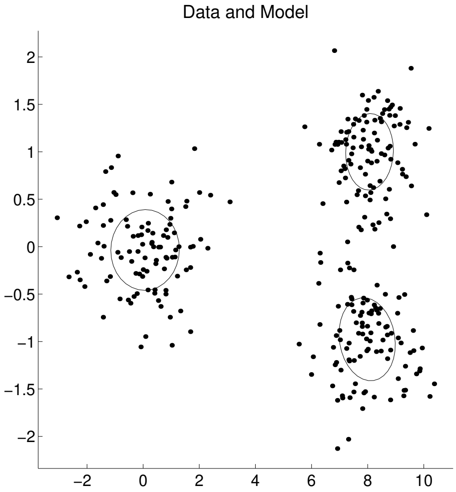

Figure 2: *Side elevation of the Gaussian mixture data illustrating the close proximity of two of the clusters. The contours illustrate projected lines of unit Mahalanobis distance from the fitted (not generative) model. Note the significant difference in scales of the two axes.*

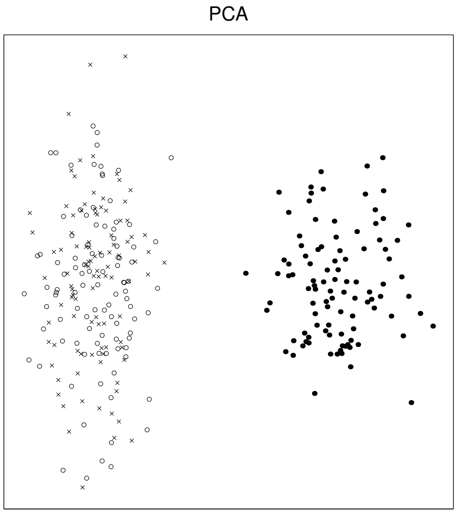

Figure 3: *Principal component projection of the three clusters, each denoted by a separate symbol. The two adjacent clusters are superimposed as the data has been generated such that the principal axes of variance are orthogonal to the line separating those clusters.*

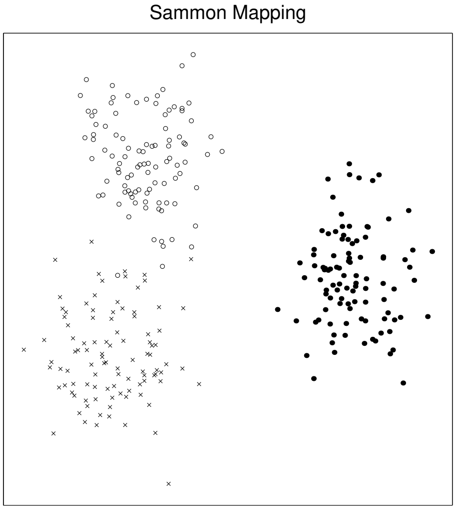

Figure 4: *Sammon mapping based on the model-weighted distance matrix for a 3-component mixture model.*

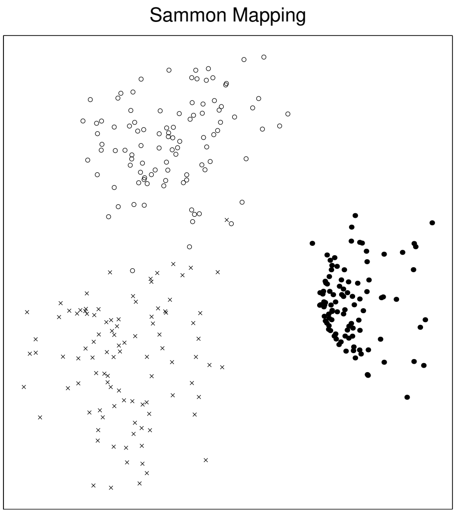

Figure 5: *Sammon mapping for 6-components.*

Oil flow data

For a more complex, higher-dimensionality, example, we consider data arising from the non-invasive monitoring of multi-phase flow of oil, water and gas in a pipeline (again studied in [2]). The data is derived from 12 noisy measurements of gamma ray attenuation through the pipeline for varying fractions of flow constituents. Locally, there are only two degrees of freedom in the measurements (since the fractions must sum to one), but the structure of the data is considerably complicated by the fact that the mixture can adopt a number of geometric flow configurations. Since we know that the data is locally two-dimensional, we choose to model it with a mixture of *probabilistic principal component analysers* [8]. This is a mixture of Gaussians where each component has a covariance matrix constrained by the parameterisation $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2\mathbf{I}$, where $\mathbf{W}$ is a $d \times q$ matrix. This restriction implies that the variance can be freely adjusted only in q orthogonal (but adaptive) directions, while the variance in each dimension orthogonal to this 'pancake' is forced to be identical (σ^2). The name of the model derives from the fact that the maximum likelihood estimate for $\mathbf{W}$ spans the local principal subspace. The motivation for this model is that the number of free parameters which must be estimated (and which prohibitively scales with d^2 for a full covariance matrix) can be significantly constrained, while local correlations in the data can still be captured. Here, based on the knowledge of the data, we choose $q = 2$. Fig. 6 shows a PCA projection of 1000 example data points, while Fig. 7 shows the Sammon mapping based on the weighted model distance matrix for the mixture.

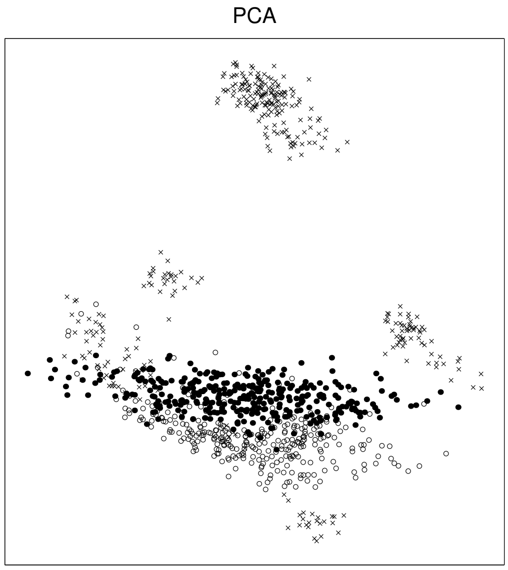

Figure 6: *PCA of the oil pipeline dataset. Examples corresponding to laminar flow are labelled '×', annular by '∘' and homogeneous by '•'.*

Knowledge of the generating process indicates that the data set comprises a number of distinct and connected clusters, representing interactions of the different geometric configurations with the gamma beam lines. A prolonged interactive investigation which used the configuration labels as a guide, illustrated in [2], showed there to be 8 distinct laminar flow clusters, 4 connected planes representing annular flow, and

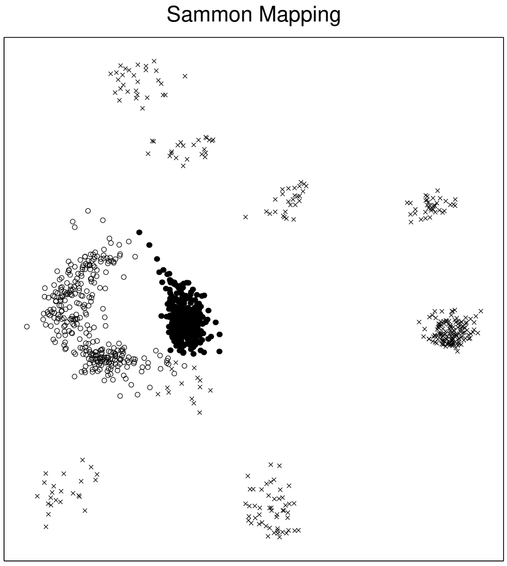

Figure 7: *Sammon mapping of the oil pipeline dataset based on the model distance function.*

a single homogeneous flow plane. In Fig. 7, 7 of the laminar clusters are clearly evident without the use of class labels. The single homogeneous, and connected annular, clusters can also be discerned. The final laminar cluster cannot be reliably detected without the aid of the labelling, but it should be noted that there are very few examples in that cluster.

Handwritten digits

For a more quantitative demonstration, we give an example of application to 8×8 grey-scale images of handwritten digits taken from the CEDAR database (e.g. see [8]). A 10-component probabilistic PCA mixture model (with $q = 16$) was fitted to a 1000-example dataset, comprising 100 examples each of digits '0' to '9'. Knowledge of the class labels was not used in the fitting process (although we chose to use 10 mixture components), and the model with greatest likelihood over 25 runs with random initialisations was selected.

In order to attempt to quantify how successfully the model-weighted distance measure is emphasising distances between clusters, for each example in the dataset, the *maximum* distance to another example in the same class was computed. The number of examples in other classes which were closer than this maximum was then calculated, as a measure of 'error'. This was performed for both the Euclidean and model-weighted distance functions. The comparison against the maximum within-class distance is a rather harsh criterion, but is the type of measure used in 'complete-linkage' clustering, for example. The results of this experiment are illustrated in Fig. 8 below. To summarise, 80.4% of the data was closer to a member of another class than to the furthest member of its own for the Euclidean distance, while for the model-weighted case, this was only 13.3%.

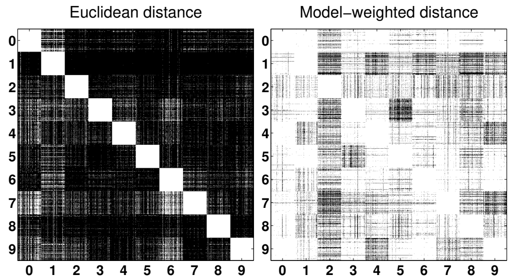

Figure 8: *Inter-digit distance comparison matrices (*1000×1000*). Each row represents a single example of a given class, and the columns associated with that row indicate (in black) out-of-class examples which are closer to the given example than the most distant in-class example. Note that the matrices will generally not be symmetric.*

REFERENCES

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[2] C. M. Bishop and M E. Tipping. A hierarchical latent variable model for data visualization. *IEEE Trans. PAMI*, 20:281–293, 1998.

[3] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comp.*, C-23:881–889, 1974.

[4] J. A. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2:63–76, 1985.

[5] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, 1999. MIT Press.

[6] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

[7] P. Smyth and D. Wolpert. Stacked density estimation. In *Advances in Neural Information Processing Systems 10*, pages 668–674, 1998. MIT Press.

[8] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11:443–482, 1999.

[9] M. E. Tipping and D. Lowe. Shadow targets: a novel algorithm for topographic projections by radial basis functions. In *IEE Fifth International Conference on Artificial Neural Networks*, pages 101–105, 1997. IEE.