Mixture Models for Clustering and Dimension Reduction
Verbeek, J.J.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
In many current information processing systems the measurements are very high dimensional: each measurement comprises many variables. Examples include:

- Systems which use a camera as a sensor, delivering images consisting of several thousands or millions of pixels.

- Systems which process text documents, frequently represented as a vector indicating for each of several thousands of words in a dictionary how often it occurs in the document.

- Systems which analyze sound by measuring the energy in several hundreds of frequency bands every few milliseconds.

By measuring many different variables a very rich representation is obtained, that can be used for a wide variety of tasks. For example, images can be used to recognize the different digits of a postal code written on an envelope, or to recognize a person from an image of his face.

Information processing tasks where high dimensional data sets are common include classification, regression and data visualization. In classification, the goal is to predict the class of a new measurement on the basis of supervised examples. For example, we may want to predict which digit is displayed in an image: does the image belong to class 0, 1, ..., 8 or 9? The supervised examples then consist of images of digits together with a label indicating which digit is displayed. The supervised examples are used to estimate a function that can be used to map new images from an unknown class to a prediction of the class label. Regression is similar to classification, but the goal is to predict some continuous quantity from the measurement rather than a discrete class label. Visualization techniques are used to produce an insightful graphical display of the data to assist a user in interpreting the data. For example, the results of image database queries can be presented using a two dimensional visualization that displays similar images near to each other. In this manner a user quickly gets an idea of what kind of images were found.
**Introduction**

Figure 1.1: The points are given by a linear projection of the 1600 pixel values of images of a face. The images trace out a curve along which the pose of the face gradually changes. For some points the corresponding image is plotted next to it.

Degrees of freedom and clusters. The high dimensional data can be thought of as a set of points in a high dimensional data space; the value of a measured variable gives the coordinate on a corresponding dimension of this space. In many applications where such high dimensional data is processed, the data is not distributed evenly over the data space but exhibits some structure.

Often the dimensionality of the data space is much larger than the number of degrees of freedom of the process from which the measurements are obtained. The degrees of freedom are the manners in which the process that is monitored can vary. Since the measurements are in a sense a function of the degrees of freedom, the number of degrees of freedom also limits the number of ways in which the measurements can vary. For example, consider $40 \times 40 = 1600$ pixel gray-scale images of a face that looks in different directions, as depicted in Fig. 1.1. We may consider each image as a 1600 dimensional measurement, one dimension for the gray value of each pixel. However, there is only one degree of freedom since the face changes only by turning from left to right. If the direction in which the face looks changes, the values of many of the 1600 pixels will change. However, these changes will be highly correlated. Apart from measuring many correlated variables, irrelevant variables may be measured as well: variables that are not dependent on the degrees of freedom. If the sensor is noisy, the value of an irrelevant variable does vary over the different measurements.
If small changes in the degrees of freedom also lead to small changes in the measured variables, then the value of the measured variables is a smooth function of the degrees of freedom. An illustration is given in Fig. 1.1, where we plotted images of a face as points in a two dimensional space, each coordinate is given by a linear combination of the original pixel values. We see that the points are roughly distributed around a smooth curve. By plotting for several points the corresponding image next to it, we see that indeed the direction in which the face looks changes gradually along the curve.

If the number of degrees of freedom is larger than one, then the data lies on or near surfaces—or in general manifolds—rather than curves. The degrees of freedom are referred to as latent variables underlying the data, and the number of degrees of freedom is referred to as the intrinsic dimensionality of the data.

In addition, the data may exhibit a clustered structure: the data forms several disconnected clouds of points in the data space. For example, consider a collection of text documents which contains articles about soccer and articles about religion. Suppose that the documents are represented as suggested above, and that very common words like 'a', 'the', 'have', 'is', etc., have been removed from the dictionary. It may be expected that articles about the same topic have a lot of words in common. On the other hand, two articles about different topics are expected to use rather different words. Thus, the documents of the two topics form two separated clusters in the data space. There are also situations where the data exhibits both a clustered structure and the data within each cluster is distributed near a low dimensional manifold.

**The curse of dimensionality.** It is well known that for classification and regression the amount of supervised examples required to reliably estimate a function with a given accuracy grows exponentially with the dimensionality of the data. Therefore, already for several tens of dimensions the number of required supervised examples becomes impractically large. This effect was termed ‘the curse of dimensionality’ by Bellman (Bellman, 1961). To illustrate this, consider a rather simplistic approach to classification and regression where we partition each input dimension into \( m \) cells. For \( D \) input dimensions, the total number of cells is then \( m^D \), and to estimate the function output of each cell we would need at least \( m^D \) supervised examples.

The set of classification or regression functions that is considered for a specific task plays a crucial role here. If the set of functions is very small, then, regardless of the data dimensionality, chances are high that already with relatively few supervised examples we can determine which function gives the best predictions for future data. As the set of functions becomes larger, more supervised examples are needed to identify the best function among the many candidates. When processing high dimensional data, the absence of prior knowledge to reduce the set of potentially useful functions often leads to a very large set of functions that is considered.

High dimensional data is also problematic for data visualization, since the number of
variables that can be displayed is inherently limited to two or three. With colors and temporal effects, i.e. showing a movie rather than a still plot, it is possible to somewhat increase the number of variables that can be displayed. However, the total number of variables that can be displayed remains limited to about six.

**Finding the relevant information.** To avoid the curse of dimensionality and to visualize high dimensional data, there is a need for methods which are able to detect clusters and to identify the manifolds on which the data resides. The data can then be represented in terms of cluster membership and/or low dimensional coordinates on these manifolds. In a sense, such methods find the relevant information in an overwhelming amount of variables. In Chapter 2 we give a review of different types of clustering and dimension reduction methods that have been proposed.

In most classification and regression applications it is time and/or money consuming to generate supervised examples: for a set of measurements the class label or regression variable has to be determined manually. Therefore the number of supervised examples is often limited. Unsupervised examples—lacking the class label or regression variable—on the other hand, are often much easier to acquire. For example, when the goal is to classify email messages as ‘spam’ or ‘not-spam’, it is easy to acquire thousands or more email messages but quite cumbersome to determine the class label for all these messages. Therefore, it is attractive to apply clustering or dimension reduction techniques on a relatively large set of unsupervised data to recover a compact representation. A classification or regression function can then be estimated from a low dimensional description of a smaller set of supervised data and/or from the supervised data in each cluster separately. In this thesis we mainly focus on such unsupervised methods for clustering and dimension reduction.

Among the best known clustering approaches are the k-means algorithm and the EM algorithm to estimate Gaussian mixture densities. Both algorithms iterate two steps, and both algorithms are guaranteed to converge and to increase some performance measure after each iteration. The main drawback of these algorithms is that the resulting clustering depends strongly on the initialization of the algorithm and is not guaranteed to maximize the performance criterion. In Section 3.1 we introduce mixture densities and the EM algorithm. Then in Section 3.2 and Section 3.3 we consider how to overcome the drawbacks of the k-means algorithm and the EM algorithm for Gaussian mixtures. In Section 3.4 we consider how the EM algorithm can be accelerated when applied to data sets containing many points.

Dimension reduction methods can be divided into methods which are limited to finding linear manifolds in the data space (lines, planes, etc.), and methods which also find non-linear manifolds. The self-organizing map, introduced by Kohonen in the early 1980’s, is a non-linear dimension reduction method that has been used in many practical applications. Nevertheless, there are several problematic issues with self-organizing
maps. First, Kohonen's parameter estimation algorithm is not guaranteed to converge. Second, the self-organizing map was originally designed for dimension reduction of real valued data. Extensions for data that are not specified as real numbers have been proposed, but were rather ad-hoc and not derived from a general principle. The same holds when there are data points with missing values; i.e. data points for which the value of some variables is unknown. In Chapter 4 we present a dimension reduction technique based on mixture densities, similar to Kohonen's self-organizing map, that resolves the mentioned problems.

In some applications it is not only desirable to have a mapping from the high dimensional data space to a low dimensional representation, but also a mapping from the low dimensional representation to the original high dimensional data space. Such a mapping allows us to reconstruct the high dimensional data from the low dimensional representation. In the example of facial images given above, such a mapping would allow us to generate the different facial images by specifying a single number: their location along the curve. Linear dimension reduction methods, such as principal component analysis, are widely used for data compression and reconstruction. Although linear methods can be implemented very efficiently and provide a smooth two-way mapping, they are limited to map to and from linear manifolds. Most non-linear dimension reduction methods do not provide a smooth two-way mapping or lack an efficient algorithm for parameter estimation.

Several methods exist that combine clustering and linear dimension reduction. Such methods have been applied to data which lies on or near a non-linear manifold in the data space, i.e. like the images in Fig. 1.1 which lie along a non-linear curve. The data is clustered in such a manner that the data within each cluster lies close to a linear manifold; the curve is approximated by several linear segments. These methods find a separate low dimensional representation for the data in each cluster by projecting the data onto the linear subspace associated with the cluster. Thus, there is no single — global — low dimensional representation for the complete data set. In Chapter 5 we explore a probabilistic approach to combine the several linear low dimensional representations into a global non-linear one. This approach provides a smooth two-way mapping and parameters can be estimated with an EM-like algorithm.

In Chapter 6, we summarize our conclusions from the preceding chapters and discuss directions for further research.