Validating visual clusters in large datasets

fixed point clusters of spectral features

Working Paper

Author(s):
Hennig, Christian; Christlieb, Norbert

Publication date:
2002

Permanent link:
https://doi.org/10.3929/ethz-a-004297813

Originally published in:
Research Report / Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) 101

Validating visual clusters in large datasets: fixed point clusters of spectral features

by

Christian Hennig [1] and Norbert Christlieb [2]

Research Report No. 101
January 2002

Seminar für Statistik
Eidgenössische Technische Hochschule (ETH)
CH-8092 Zürich
Switzerland

[1] Seminar for Statistics, ETH Zentrum, Zurich, Switzerland and University of Hamburg, Germany
[2] Hamburg Observatory at University of Hamburg, Germany

Validating visual clusters in large datasets: fixed point clusters of spectral features

Christian Hennig [‡] and Norbert Christlieb [§]

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

January 2002

Abstract

Finding clusters in large datasets is a difficult task. Almost all computationally feasible methods are related to k-means and need a clear partition structure of the data, while most such datasets contain masking outliers and other deviations from the usual models of partitioning cluster analysis. It is possible to look for clusters informally using graphic tools like the grand tour, but the meaning and the validity of such patterns is unclear. In this paper, a three-step-approach is suggested: In the first step data visualization methods like the grand tour are used to find cluster candidate subsets of the data. In the second step, reproducible clusters are generated from them by means of fixed point clustering, a method to find a single cluster at a time based on the Mahalanobis distance. In the third step, the validity of the clusters is assessed by use of classification plots. The approach is applied to an astronomical dataset of spectra from the Hamburg/ESO survey.

Key words: Discriminant coordinates, outliers, contamination model, sky surveys, stellar populations

1 Introduction

Finding clusters in large datasets is a difficult task, because such datasets typically do not match the implicit and explicit underlying assumptions of the conventional methods of cluster analysis. Often they contain outliers that are not easily detectable because of masking phenomena. Moreover, only a minority of the points of such a dataset may be described adequately as clustered, the clusters may have different shapes, which cannot always be modeled by the same family of parametric distributions, and there are often non-elliptical and asymmetrical parts of the data.

[‡] Seminar for Statistics, ETH Zentrum, Zurich, Switzerland and University of Hamburg, Germany

[§] Hamburg Observatory at University of Hamburg, Germany

Figure 1: Above: Principal components of spectra data (all variables standardized to median $= 0$, $1.4826 \cdot \mathrm{MAD} = 1$) with FPC 1 (gray; consisting of 47978 points, which are all concentrated in one tiny point near the origin). Below: Scatterplot of Calcium-break variables. Points on the diagonal line form FPC 2, see Section **??**.

To illustrate these issues, we present a dataset of 74159 points of 16 features of spectra from the Hamburg/ESO survey. The dataset is explained in Section **??**. Figure **??** makes some of the features of the dataset visible. It shows two two-dimensional projections of these data, namely the principal component plot and a scatterplot of two of the 16 variables. Both plots suggest that the dataset as a whole can neither be adequately described by some single elliptical distribution, nor by a mixture of few of them. Note that more than half of the data are concentrated in the tiny gray point near the origin of the principal components plot (above; the gray "area" is referred to as FPC 1, as explained further in Section **??**) so that the principal components plot is entirely dominated by fewer than 10% of the points which appear outlying from the core of the data forming "strings" parallel to the first two principal components. The points along the horizontal string clearly show a skew distribution. In the scatterplot shown below it can be seen that there is a data subset arranged on a straight line which intersects the majority of the data only slightly. The rest of the data looks skew in both variables. Further projections of the dataset are shown in Section **??**.

Most cluster analysis methods that can be applied to datasets of this size assume more or less implicitly that the dataset can be partitioned completely into clusters which can all be modeled by the same parametric family. For example, k-means is a maximum likelihood estimator for a fixed partition model of k Gaussian distributions with the same covariance matrix, which must be a constant multiplied by the unit matrix. There are some approaches which account explicitly for outliers, e.g. maximum likelihood mixture fitting based on t-distributions (**?**, Ch. 7) as implemented in the software EMMIX, Gaussian mixture modeling with additional noise component (**?**) as implemented in EMCLUST, and trimmed k-means (**?**), but datasets of the given size may cause computational problems. For example, the result of a single EM-algorithm run as used by EMMIX and EMCLUST may depend strongly on the starting value (**?**, Sec. 2.12). The choice of starting values by means of an initial hierarchical clustering, which is the default for both packages, needs too much memory for large datasets, and a single run of the EM-algorithm takes so much time that it becomes unfeasible to try multiple starting values, especially if the number of clusters is treated as unknown. However, there is some research about reasonable starting values for large datasets, see **?**.

An alternative approach is visual clustering based on interactive software for graphical data analysis (**?**). This may reveal many interesting and nonstandard patterns in the data, which cannot be found by the existing formal methods. However, it remains unclear if these cluster patterns are statistically relevant in some sense, which points belong to them exactly, and how the patterns could be reproduced by a repetition of the analysis of the same dataset. Especially the latter issues are difficult in large datasets, because with ten thousands of points on the screen it is not possible to demarcate clusters exactly, and the inspection of high dimensional data needs so many two-dimensional projections that it becomes improbable that exactly the same projections lead again to the discovery of clusters during a repetition of the graphical analysis.

The aim of this paper is to provide a strategy for datasets of the size of our spectra data, which connects the more intuitive visual clustering with a formal method to generate well defined, reproducible clusters.

The strategy consists of three elements:

(1) Cluster candidate sets can be found and selected by interactive software for graphical data analysis. Here the grand tour (**?**) as implemented in the software XGOBI (**?**) is used. This step is discussed in Section **??**.

(2) The candidate data subsets can be used as starting configurations for an algorithm which finds a fixed point cluster (FPC). FPC analysis (**?**) is a method to find a data subset which does not contain any outlier, and with respect to which all other points of the dataset are outliers. Such a subset can be considered as cluster. The definition of an outlier is based on a parametric model for a single cluster, here a multivariate Gaussian distribution. That is, like for the other formal clustering methods discussed above, there is an underlying parametric model for FPC analysis. For combinatorial reasons it is not possible to find all FPCs of a high dimensional dataset. But there are two advantages of FPC analysis: Firstly, while a cluster is modeled as elliptical, there is no stochastic assumption on the rest of the data which does not belong to the cluster, except of being well separated from it in some sense. In particular, a single cluster candidate can be used as a starting configuration, which is not possible for partition or mixture based methods. Secondly, for a given cluster candidate subset, the algorithm is fast and easy even for large datasets. This step is discussed in Section **??**.

(3) Found fixed point clusters can be visualized by use of plots of optimal discriminating projections (**??**). The projection directions may be of use for the interpretation. This step is discussed in Section **??**.

2 Finding cluster candidates

The easiest method to find cluster candidates is to inspect all scatterplots of two variables. However, not all clusters can be discovered by considering only the interactions between two variables. The visual inspection of complex datasets raises the problem how more than two dimensions can be visualized on the computer screen. This can be done by inspection of a moving sequence of two-dimensional projections. The easiest form of such a sequence is the three dimensional rotation. Note that the idea behind the rotation is not only to show many projections. These projections form a smooth sequence, and this is crucial for the intuition, because we do not only observe the location of the points in the two-dimensional subspace, but also their movement. The movement of the three dimensional rotation is one-dimensional, and together with the two-dimensional projections we perceive three dimensions. The grand tour (**?**) is a generalization of the rotation, where the local movement is defined by a two-dimensional vector, so that we get the impression of four dimensions quickly (**?**). **?** defines a sequence of projections of p-dimensional data into the $\mathbb{R}^2$ such, that

- the sequence of projections is asymptotically dense in the space R of all possible projections,

- all the different possible projections are approximated roughly as fast as possible (this requires the definition of a topology on R),

- the distribution of projections in R is asymptotically uniform,
- the sequence should appear continuously,
- the sequence should appear as “straight” as possible (which is in conflict to the fast approximation of all areas of the dataset).

The implementation of XGOBI is based on a smooth interpolation of a sequence of randomly generated planes. The reader is referred to **?** for details.

For the task of finding cluster candidates, which can be used as starting configurations for FPC analysis, data subsets in plots must be “brushed”, i.e., interactively marked with various colors and symbols. Files of membership vectors of the brushed groups must be generated. This is possible with XGOBI.

Though the grand tour is designed to cross the neighborhoods of all two-dimensional projections as fast as possible, the inspection of large datasets with lots of dimensions is very time-consuming. So it is always advisable to reduce the dimensionality. If the dimension is moderate, it is possible to carry out the grand tour until all significant characteristics are repeatedly seen. Such a tour through six of the 16 variables of the spectra dataset took about 40 minutes, including brushing of cluster candidates.

XGOBI provides a lot of methods and features that are useful for visual clustering. Selected variables can be removed, added or held fixed manually, which can be done according to a priori information about the meaning of the variables.

There are various possibilities to scale the data. They can be centered about the mean or the median, and the screen can show the whole range of the data or only a central part of adjustable size. The whole range of the data is of interest to find extreme cluster candidates, while it is advantageous to consider only a central part about the median to assess the majority of the data in detail, because otherwise the projections may be dominated by gross outliers. Examples can be seen in the Figures 1 and 5. In the recent version of XGOBI the plotted range of all variables can be chosen as k times the value range. The constant k is the same for all variables and can be specified by the user. In older versions it had been possible to choose the plotted range of the values in terms of the standard deviation and the median absolute deviation (MAD), which had served our aims better because the value range is often determined by extreme outliers, such that $k = 0.01$ or even smaller would be necessary to inspect the central part of the values of such variables, but this would exclude a majority of the values of other variables from the plot, which are less affected by outliers. Scaling based on standard deviation and MAD is already documented in **?**. We communicated the problem to the designers of XGOBI and hope that the option will appear again in the next version.

As will be discussed in Section **??**, the cluster candidates obtained from the visual inspection should be as homogeneous as possible to be used as starting configurations for FPC analysis. Therefore they should be followed through many projections. If they look heterogeneous in some of the plots, they should be split and outliers should be excluded. For the same reason another feature of XGOBI can be useful, namely the connection of selected points of brushed subsets with lines to assess the shape of the subsets through various plots.

The cluster candidates used for FPC analysis should rather contain too few points than too much. The FPC algorithm of Section **??** is usually able to find a whole homogeneous subpopulation if the starting configuration contains only a portion of it (in large datasets, usually even 10–20% suffice), but it may be seriously affected by gross outliers and heterogeneity in the starting configuration, see Section **??**.

Furthermore, XGOBI allows the grand tour to be guided in order to optimize certain projection pursuit indices. However, this feature did not work with the spectra data, because the guided tour of XGOBI initially performs a sphering of the data based on principal components. This led to an error, presumably caused by extreme outliers which made almost all eigenvalues of the principal components appear to be approximately zero.

Further techniques for visual clustering are discussed in **?** and may also be applied with advantage for the generation of starting configurations for FPC analysis.

3 Mahalanobis fixed point clusters

3.1 *Definition*

The basic idea of FPC analysis is that a cluster can be formalized as a data subset, which is homogeneous in the sense that it does not contain any outlier, and which is well separated from the rest of the data meaning that all other points are outliers with respect to the cluster. That is, the FPC concept is a local cluster concept: It does not assume a cluster structure or some parametric model for the whole dataset. It is based only on the cluster candidate itself and its relation to its surroundings.

In order to define FPCs we need a definition of an outlier with respect to a data subset. The definition should be based only on a parametric model for the non-outliers (reference model), but not for the outliers. That is, if we take the Gaussian family as reference model, the whole dataset is treated as if it came from a contamination mixture

$$(1-\epsilon)N_p(a, \mathbf{\Sigma}) + \epsilon P^*, \quad 0 \leq \epsilon < 1,$$

where p is the number of variables, $N_p(a, \mathbf{\Sigma})$ denotes the p-dimensional Gaussian distribution with mean vector a and covariance matrix $\mathbf{\Sigma}$, and P^* is assumed to generate points well separated from the core area of $N_p(a, \mathbf{\Sigma})$. The principle to define the outliers is taken from **?**. They define α-outliers as points which lie in a region with low density such that the probability of the so-called outlier region is α under the reference distribution. α has to be small in order to match the impression of outlyingness. For the $N_p(a, \mathbf{\Sigma})$-distribution, the α-outlier region is

$$\{x: \ (x-a)^t \mathbf{\Sigma}^{-1}(x-a) > \chi^2_{p;1-\alpha}\},$$

$\chi^2_{p;1-\alpha}$ denoting the $1-\alpha$-quantile of the χ^2-distribution with p degrees of freedom. Because a and $\mathbf{\Sigma}$ are not known, they have to be estimated. The easiest way to do this is to use the sample mean and the maximum likelihood covariance matrix.

We write the dataset as a matrix $\mathbf{X} = (x_1, \ldots, x_n) \in (\mathbb{R}^p)^n$. Data subsets (such as cluster candidates obtained from visual clustering) are represented by an indicator vector $g \in \{0,1\}^n$. Let $\mathbf{X}(g)$ be the matrix with only the points x_i, for which $g_i = 1$, and

Figure 2: Four overlapping FPCs of artificial data with means M and borders of outlier regions.

$n(g) = \sum_{i=1}^{n} g_i$. Let $m(g) = \frac{1}{n(g)} \sum_{g_i=1} x_i$ the mean vector and $\mathbf{S}(g) = \frac{1}{n(g)} \sum_{g_i=1} (x - m(g))(x - m(g))'$ the ML covariance matrix estimator for the points indicated by g.

The set of outliers from $\mathbf{X} = (x_1, \dots, x_n)$ with respect to a data subset $\mathbf{X}(g)$ is

$$\{x : \ (x - m(g))'\mathbf{S}(g)^{-1}(x - m(g)) > \chi^2_{p;1-\alpha}\}.$$

That is, a point is defined as an outlier w.r.t $\mathbf{X}(g)$, if its Mahalanobis distance to the points of $\mathbf{X}(g)$ is large.

An FPC is defined as a data subset which is exactly the set of non-outliers w.r.t. itself:

Definition 1 *A data subset $\mathbf{X}(g)$ of $\mathbf{X}$ is called Mahalanobis fixed point cluster of level α, if for $i = 1, \dots, n$:*

$$g = \left(I\left[(x_i - m(g))'\mathbf{S}(g)^{-1}(x_i - m(g)) \leq \chi^2_{p;1-\alpha} \right] \right)_{i=1,\dots,n}, \qquad (1)$$

where I denotes the indicator function.

3.2 *Computation*

For combinatorial reasons it is impossible to check (**??**) for all g. But FPCs can be found by a fixed point algorithm defined by

$$g^{k+1} = \left(I\left[(x_i - m(g^k))'\mathbf{S}(g^k)^{-1}(x_i - m(g^k)) \leq \chi^2_{p;1-\alpha} \right] \right)_{i=1,\dots,n}, \qquad (2)$$

given reasonable starting configurations g^0, which are to be found during the grand tour. Experience shows that this algorithm is fast and easy enough to work with large datasets.

Theorem 2 *The fixed point algorithm defined by (**??**) reaches an FPC vector g^m (i.e., $\mathbf{X}(g^m)$ is an FPC) after a finite number of steps, if $\chi^2_{p;1-\alpha} > p$ and unless $\mathbf{S}(g^k)$ is singular for some $k \leq m$.*

The proof is given in the appendix.

The assumption $\chi^2_{p;1-\alpha} > p$ is fulfilled for $\alpha \leq 0.25$ for arbitrary p. The algorithm can cope as well with singular $\mathbf{S}(g^k)$ in practice by use of the Moore-Penrose inverse. In this case, all points that do not lie on the lower dimensional hyperplane spanned by $\mathbf{X}(g)$ have to be defined as outliers.

3.3 *Discussion*

The fact that the definition of FPCs is based on a Gaussian model does not mean that the points of an FPC always look Gaussian. For non-homogeneous data subsets, the region of non-outliers can be very large, and so there may be large FPCs which clearly do not stem from a single Gaussian distribution, see Figure **??**. This problem does not arise if the starting configuration is homogeneous, because in this case the usual mean and covariance estimators work well. Otherwise they might be affected by outliers or irregularities *inside* the candidate subsets, while FPC analysis is robust against such problems outside their non-outlier regions. It must be noted that the ML estimators are not advisable from the viewpoint of robust outlier identification (**?**). FPCs may be defined where they are replaced by more robust estimators. But it is not clear if Theorem **??** remains valid for such estimators, and the computation may be very time consuming for large datasets, because the estimators have to be computed for each cluster candidate and each algorithm step. Furthermore, **?** showed that resistant outlier identification is possible by application of the usual estimators for Gaussian distributions to data subsets. Therefore we restrict ourselves to the easiest estimators in this paper, aware of the necessity to keep the candidate subsets as homogeneous as possible.

Note further the ability of FPCs to include and overlap each other. They do neither form a partition nor a hierarchy. The inclusion property can be useful sometimes, see e.g. Figure **??**, where it is not adequate to exclude the points of the inner left FPC from the outer left cluster.

It is possible to define other kinds of FPCs analogously for other clustering problems from a given outlier definition, see **?**. FPCs are not adequately interpreted as estimators of mixture components. Instead, they estimate “theoretical FPCs” (**?**), i.e., they are based on an alternative definition of a cluster. Publications on consistency properties are in preparation.

The tuning constant α has to be chosen by the statistician. The interpretation of α is that the smaller α is, the more separated a data subset must be from the rest of the dataset to form an FPC. A large value of α, say, $\alpha = 0.1$ or $\alpha = 0.05$, classifies lots of points from a Gaussian population as outliers. On the other hand, if there is a Gaussian cluster among others, points from neighboring clusters do not affect its FPC property, if they lie in the corresponding outlier region. If α is smaller, the outlier region is more distant from the center, and therefore the other clusters need to be better separated. Illustrations are given in **?** and Section **??**.

4 Discriminant projections

Discriminant projections are useful to get a clear impression on the separateness of the found clusters, and they may sometimes be of use for their interpretation. Two kinds of discriminant projections are used in this work.

4.1 *Discriminant coordinates*

The method of discriminant coordinates is the most common approach for the projection of high-dimensional data with a given grouping to a lower-dimensional subspace. The term “discriminant coordinates” is used according to **?**, they are also known under the name “canonical variates”. The approach goes back at least to **?**, who developed discriminant coordinates as a generalization of Fisher's linear discriminant function to more than two groups.

The definition is as follows: Let $x_{i1}, \ldots, x_{in_i}$ the p-dimensional points of group $i = 1, \ldots, s,\ n = \sum_{i=1}^{s} n_i$, i.e., it is assumed that the data is partitioned. Let

$$m_i = \tfrac{i}{n_i} \textstyle\sum_{j=1}^{n_i} x_{ij},\ m = \tfrac{1}{n} \sum_{i=1}^{s} \sum_{j=1}^{n_i} x_{ij},\ \mathbf{C}_i = \sum_{j=1}^{n_i} (x_{ij} - m_i)(x_{ij} - m_i)',$$
$$\mathbf{W} = \tfrac{1}{n-s} \textstyle\sum_{i=1}^{s} \mathbf{C}_i,\ \mathbf{B} = \tfrac{1}{s-1} \sum_{i=1}^{s} n_i (m_i - m)(m_i - m)',$$

where $\mathbf{W}$ denotes the pooled within clusters-covariance matrix and $\mathbf{B}$ denotes the between clusters-covariance matrix. The first discriminant coordinate is the linear combination of the original variables, along which the ratio between projected between-clusters-variance and within-clusters-variance of the data is maximized, i.e. a vector c_1 maximizing

$$F_c = \frac{c'\mathbf{B}c}{c'\mathbf{W}c}.$$

The second discriminant coordinate maximizes F_c subject to orthogonality with respect to $\mathbf{W}$, i.e., $c_2'\mathbf{W}c_1 = 0$. Further projection directions are defined by analogy. The discriminant coordinates turn out to be the orthonormal eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$, c_1 corresponding to the largest eigenvalue and so on, see **?**. $s-1$ coordinates suffice to show all the separation information of s groups contained in $\mathbf{B}$, because the principle of discriminant coordinates is to show the differences between the mean vectors of the groups with respect to a metric defined by $\mathbf{W}$, which is an estimator of the covariance matrix of a single group under the assumption that it is the same for all groups, and s means can be embedded in an $s - 1$-dimensional subspace.

FPC analysis does not result in a partition of the data, but in single clusters, which may overlap and are not necessarily exhaustive. There are two ways to apply discriminant coordinates in our setup. They may be computed between disjoint FPCs, and the rest of the data may be projected into the resulting image, or a single FPC can be treated as one class and the rest of the data as a second. However, the complement of an FPC can by no means expected to be homogeneous and in particular it is not plausible that it should have the same covariance matrix. While projection on the first discriminant coordinate may show valuable information about the difference in means, we suggest to choose a second projection direction in order to show the difference in the covariance matrices.

4.2 *Bhattacharyya coordinates*

The separation between two distributions with means μ_1, μ_2 and covariance matrices Σ_1, Σ_2 can be measured by the following expression:

$$\begin{aligned}
D(\mu_1, \mu_2, \Sigma_1, \Sigma_2) &= D_1\left(\mu_1, \mu_2, \frac{\Sigma_1 + \Sigma_2}{2}\right) + D_2(\Sigma_1, \Sigma_2), \\
D_1(\mu_1, \mu_2, \Sigma) &= \frac{1}{8}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2), \\
D_2(\Sigma_1, \Sigma_2) &= \frac{1}{2} \log \frac{\det\left(\frac{\Sigma_1 + \Sigma_2}{2}\right)}{\sqrt{\det(\Sigma_1)\det(\Sigma_2)}}.
\end{aligned}$$

D is called "Bhattacharyya distance" according to **?**, p. 99, where it is derived as an upper bound for the Bayes misclassification probability between two Gaussian distributions with the given parameters. In the case of two groups in the data, it is appealing to choose projection directions d in order to maximize $D(d'm_1, d'm_2, d'\mathbf{W}_1 d, d'\mathbf{W}_2 d)$, $\mathbf{W}_i = \frac{1}{n_i}\mathbf{C}_i,\ i = 1, 2$. However, such directions are hard to compute. **?**, p. 455ff. suggests to consider the special cases $\Sigma_1 = \Sigma_2$ and $\mu_1 = \mu_2$. In the first case, the projected D_1 is to be maximized, which is almost equivalent to the computation of the first discriminant coordinate with the only difference that for calculating D_1 the covariance matrices of the two groups have to be pooled as $\mathbf{W}_D = \frac{1}{2}(\mathbf{W}_1 + \mathbf{W}_2)$, which corresponds to F_c only if $n_1 = n_2$. In the second case, $D_2(d'\mathbf{W}_1 d, d'\mathbf{W}_2 d)$ must be maximized, which is done by the eigenvector d_1 of $\mathbf{W}_1^{-1}\mathbf{W}_2$ where the corresponding eigenvalue λ_1 maximizes $\lambda + \frac{1}{\lambda}$ among the eigenvalues of $\mathbf{W}_1^{-1}\mathbf{W}_2$. We call d_1 the first Bhattacharyya coordinate, the further eigenvectors correspond again to further optimal projection directions under an orthogonality constraint, see **?** for details. Now it can be proceeded as follows:

(1) Compute the discriminant coordinate c_1, where $\mathbf{W}$ is replaced by $\mathbf{W}_D$[1].

(2) Project the data on the subspace orthogonal to c_1 with respect to $\mathbf{W}_D$, i.e. compute $y = (c_2', \ldots, c_p')'x$ for a data point x. Note that the whole difference in means between the two groups is projected on c_1. Thus, the group means are equal on the orthogonal subspace.

(3) Compute the Bhattacharyya coordinate d_1 for the projected data.

The resulting two-dimensional projection maximizes the difference in means along the first axis and the remaining difference of the projected covariance matrices along the second axis. Examples are shown in Section **??**. As an alternative, **?**, p. 458 suggests to take the two orthonormal eigenvectors of $\mathbf{W}_1^{-1}\mathbf{W}_2$ maximizing the projected D in the case that the difference in means is negligible compared to a more than one-dimensional difference in the covariance matrices.

Note that because of their internal scaling, Mahalanobis FPCs and the discriminant plots discussed here are invariant w.r.t. linear transformations of the whole dataset as opposed to principal components, k-means clustering and related methods.

5 Clusters of spectral features

The Hamburg/ESO survey (HES; **??**) was carried out with a special wide-angle telescope, a so-called Schmidt telescope. The 1 m Schmidt telescope of the European Southern Observatory (ESO) in Chile allows to image an area of 5 by 5 degrees of the sky, corresponding to 10× by 10× the diameter of the full moon. The main aim of the HES, which covers half of the sky visible from southern hemisphere, is to find new quasars. For the work discussed here, we used the database in a more exploratory spirit to look for unknown clusters and patterns in the collected spectra.

By mounting a prism in front of the telescope, the images of celestial objects are converted into spectra (see left panel in Fig. **??**). On one photographic plate, typically the spectra

[1] $\mathbf{W}$ could be used as well. But if FPCs are to be projected, then only one of the two groups is considered as homogeneous, namely the FPC, as opposed to the rest of the data. If the FPC does not contain the majority of the points, it gets more weight in the computation of $\mathbf{W}_D$ than in the computation of $\mathbf{W}$.

Figure 3: Left: A small part of a Schmidt plate taken without prism (upper image), and the corresponding part of the HES plate, taken with prism (lower image). The wing-shaped object in middle height and slightly to the left in the upper image is a pair of interacting galaxies. Below this object, and slightly to the left, a spiral galaxy can be seen. Point-like objects are stars, or quasars. Right: HES example spectrum of an A-type star, exhibiting strong hydrogen absorption lines ("troughs" in the spectrum).

Figure 4: Left side: First two discriminant coordinates of 5-means solution (applied to standardized data). Right side: Discriminant projection plot of FPC 1 (gray) against rest of the data (black). About 200 gross outliers lie outside the range of the plot.

of 10000 objects are present. These objective-prism plates were digitized at Hamburger Sternwarte with a scanning machine. A reduction procedure yields spectra like shown in the right panel of Fig. **??**. The total HES data base consists of about 4 million digital spectra. As illustrated in Fig. **??**, their spectral resolution (typically 15 Å at Hγ) allows to detect numerous stellar absorption lines.

For each HES spectrum, consisting of 300 data points, a set of 16 spectral features (variables) is computed. The features 1–10 are strengths of stellar absorption lines. Feature 11 is the signal-to-noise ratio of the Calcium break. Feature 12 is its contrast to the continuum (smoothed spectrum). The features 13–16 contain color information: Features 13 and 14 are the first two principal components of the smoothed spectrum, and features 15 and 16 are two "half power-points" of the spectra computed in different spectral regions. For details we refer to **?**. The total feature set contains most, if not all, of the information present in the spectra. They allow e.g. for a rough spectral classification, and rough determination of stellar parameters.

The data set used in the work presented here consists of 74159 spectra, each represented by 16 data values. It includes 344 spectra of quasars selected by "classical" criteria (i.e., UV excess, and/or presence of emission lines and/or spectral brakes), and confirmed by follow-up observations; a learning sample of 2856 spectra of known class, and 70959 unclassified spectra from 10 HES plates.

We tried first to apply some conventional clustering algorithms. The only methods that

Figure 5: Discriminant projection plot of FPCs 3 and 4 (gray-scales). The rest of the data (black) is projected onto the same coordinates. Left side: All points. Right side: Area of the FPCs.

worked on our Sun UltraSparcIIi 333MHz were k-means and k-medoids (**?**). Both resulted in artificial partitions with approximately the same discriminant coordinates independent of k. The left side of Figure **??** shows the 5-means solution.

After some inspection with the grand tour we decided that the first ten variables of the data do not show interesting patterns and we excluded them from the cluster search. An exhaustive grand tour through 6 dimensions was carried out, and we ended up with ten cluster candidate subsets. These subsets were used to start the fixed point algorithm from Section **??** with $\alpha = 0.01$ and $\alpha = 0.05$. Discriminant projection plots were inspected to assess the clusters. Here are the main results with $\alpha = 0.05$. Four different patterns of the data manifested themselves as FPCs:

- The largest FPC 1 consists of 47978 points. Its discriminant projection (the rest of the data was treated as the second class) is shown on the right side of Figure **??**. There is no clear separation. As illustrated in Fig. **??**, the FPC can be interpreted as the core of the dataset, with respect to which (interpreted as Gaussian distributed) the other stars are more or less outlying.
- FPC 2 consists of 83 points, which can be seen on the right side of Figure **??**. They lie exactly on a straight line in the scatterplot of the two Calcium break variables (the line contains more than 83 points, but the others were classified as outlying with respect to FPC 2).
- FPC 3 contains 17362 points, and FPC 4 contains 14461 points. Both are subsets of FPC 1, i.e., usual stars from the core of the data. Because they are disjoint, a discriminant projection plot can be produced to separate them from each other. The rest of the data was projected into the plot, see Fig. **??**. The right side shows that the two clusters clearly form a significant pattern. The difficulty to find such clusters is illustrated on the left side of Fig. **??**: If all points are shown, the plot is dominated by gross outliers, so that the FPCs are concentrated in a tiny area in the core area of the dataset.

The results with $\alpha = 0.01$ were less interesting. The separation between FPCs 3 and 4 and the rest of the data is too weak for such a small α, and they disappeared. Because of the less strict definition of outliers, the FPCs corresponding to FPC 1 and 2 with $\alpha = 0.05$ contained more points, namely 60593 and 149 respectively.

The analysis with $\alpha = 0.05$ was repeated seven times to investigate if the FPCs can be reproduced from different subjectively marked cluster candidates from the grand tour. In fact, FPCs corresponding to all four clusters were found in all seven repetitions. Sometimes variants of the FPCs 1–3, differing in up to 40 points, were found. The original FPCs 1, 2 and 4 were found in all seven analyses, FPC 3 was found six times, while once only a variant appeared. No further clusters were found. The patterns corresponding to FPCs 1–4 were also reproduced by an analysis of stars of ten other objective-prism plates.

Unfortunately, the clusters do not seem to have an astronomical meaning. FPC 2 stems from an error in the replacement of missing values. FPCs 3 and 4 are clearly connected to different objective-prism plates. We suspect that they are caused by variations of plate properties. Insofar the analysis has provided valuable information not about the celestial objects, but about the data processing of the survey.

6 Conclusion

We have shown that interesting cluster patterns in large datasets can be found by a three-step approach consisting of the visual inspection by means of the grand tour, iteration of Mahalanobis FPCs, and validation by discriminant projection plots. The approach combines intuitive and formal techniques. Patterns in complex data situations, which do not fulfill the assumptions of usual formal cluster analysis or mixture methods, can be found by use of the grand tour. FPC analysis serves to decide about the cluster candidates and to make them reproducible. Note that other cluster algorithms such as k-means, EMMIX and EMCLUST cannot be started from single cluster candidates, because they need the whole dataset to be partitioned. The discussed plots help to assess the homogeneity, separateness and meaning of the clusters. XGOBI can be obtained from

`http://lib.stat.cmu.edu/general/XGobi`

The FPC analysis was performed with the software FIXMAHAL written by the first author. It is available under

`http://www.math.uni-hamburg.de/home/hennig`

as well as the code to compute discriminant and Bhattacharyya coordinates with the statistical software R. All used software is freeware.

Appendix

PROOF. (Theorem 2) Define $c = \chi^2_{p;1-\alpha}$. Define

$$T(g^k) = f(n(g^k)) \det \mathbf{S}(g^k), \quad f(m) = \prod_{i=1}^{m} f_i, \quad f_i = \exp\left(\frac{p-c}{i-\frac{1}{2}}\right).$$

Below it is shown that T strictly decreases from step k to step $k+1$ unless $g^{k+1} = g^k$, i.e., g^k is an FPC vector. Find $g^{k+1} = g^k$ after a finite number of steps, because $T > 0$ and there are only finitely many indicator vectors in $\{0,1\}^n$.

Define $m_k = m(g^k)$, analogously $\mathbf{S}_k$, $\mathbf{X}_k$ and n_k for all k. Now consider a fixed k and show

$$\det \mathbf{S}_{k+1} \leq \exp\left(\frac{(n_{k+1} - n_k)(c-p)}{n_{k+1}}\right) \det \mathbf{S}_k \tag{3}$$

and

$$f(n_{k+1}) < \exp\left(\frac{(n_k - n_{k+1})(c-p)}{n_{k+1}}\right) f(n_k) \tag{4}$$

unless $n_k = n_{k+1}$, in which case either $g_k = g_{k+1}$ or "<" in (**??**) as proven below. T decreases by combination of (**??**) and (**??**).

Proof of (??) It is needed that

$$\sum_{i=1}^{n}(x_i - m)'\mathbf{S}^{-1}(x_i - m) = pn, \tag{5}$$

where m and $\mathbf{S}$ are the Gaussian ML-estimators of the mean and (non-singular) covariance matrix of p-dimensional data $x_1, \ldots, x_n$. To show that, use the decomposition $\mathbf{S} = \mathbf{O}'\mathbf{D}\mathbf{O}$, where the rows of $\mathbf{O}$ are orthonormal eigenvectors $o_1, \ldots, o_p$ of $\mathbf{S}$ and $\mathbf{D}$ is the diagonal matrix of the corresponding strictly positive eigenvalues $\lambda_1, \ldots, \lambda_p$. With that,

$$\sum_{i=1}^{n}(x_i - m)'\mathbf{S}^{-1}(x_i - m) = \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{((x_i - m)'o_j)^2}{\lambda_j}$$

and $\lambda_j = \frac{1}{n}\sum_{i=1}^{n}(x_i'o_j - m'o_j)^2$ because of $\mathbf{D} = \mathbf{O}\mathbf{S}\mathbf{O}'$, therefore (**??**).

$\mathbf{S}_{k+1}$ is an ML-estimator and minimizes $L(\mathbf{X}_{k+1}, m, \mathbf{S})$, where

$$L(\mathbf{X}_{k+1}, m, \mathbf{S}) = (\det \mathbf{S})^{n_{k+1}} \exp\left(\sum_{g_i^{k+1}=1}(x_i - m)'\mathbf{S}^{-1}(x_i - m)\right).$$

Because of (**??**),

$$\begin{aligned}
& (\det \mathbf{S}_{k+1})^{n_{k+1}} \exp(pn_{k+1}) && (6)\\
= \quad & L(\mathbf{X}_{k+1}, m_{k+1}, \mathbf{S}_{k+1}) \\
\leq \quad & (\det \mathbf{S}_k)^{n_{k+1}} \exp\left(\textstyle\sum_{g_i^{k+1}=1}(x_i - m_k)'\mathbf{S}_k^{-1}(x_i - m_k)\right) \\
\leq \quad & (\det \mathbf{S}_k)^{n_{k+1}} \exp\left(\textstyle\sum_{g_i^{k}=1}(x_i - m_k)'\mathbf{S}_k^{-1}(x_i - m_k) + (n_{k+1} - n_k)c\right) && (7)\\
= \quad & (\det \mathbf{S}_k)^{n_{k+1}} \exp\left(pn_k + (n_{k+1} - n_k)c\right),
\end{aligned}$$

proving (**??**). It may happen that the same number of points is included and excluded in a step of the algorithm, therefore $n_{k+1} = n_k$ while $g^{k+1} \neq g^k$. In this case the inequality becomes strict because of "<" in (**??**). Therefore $\det S_{k+1} < \det S_k$, and T decreases.

Proof of (??) Remember that $p - c < 0$ is assumed. Case 1: $n_{k+1} > n_k$. Then,

$$\begin{aligned}
& \tfrac{f(n_{k+1})}{f(n_k)} \exp\left(\tfrac{(n_{k+1}-n_k)(c-p)}{n_{k+1}}\right) \\
= \quad & \textstyle\prod_{i=n_k+1}^{n_{k+1}} \exp\left(\tfrac{p-c}{i-\frac{1}{2}}\right) \exp\left(\tfrac{(n_{k+1}-n_k)(c-p)}{n_{k+1}}\right) \\
< \quad & \exp\left(\tfrac{(n_{k+1}-n_k)(p-c)}{n_{k+1}}\right) \exp\left(\tfrac{(n_{k+1}-n_k)(c-p)}{n_{k+1}}\right) = 1.
\end{aligned}$$

Case 2: $n_k > n_{k+1}$. Then, analogously,

$$\begin{aligned}
& \tfrac{f(n_k)}{f(n_{k+1})} \exp\left(\tfrac{(n_k-n_{k+1})(c-p)}{n_{k+1}}\right) \\
= \quad & \textstyle\prod_{i=n_{k+1}+1}^{n_k} \exp\left(\tfrac{p-c}{i-\frac{1}{2}}\right) \exp\left(\tfrac{(n_k-n_{k+1})(c-p)}{n_{k+1}}\right) \quad > 1.
\end{aligned}$$

References

D. Asimov, The grand tour: a tool for viewing multidimensional data. *SIAM Journal of Statistical Computing* **6** (1985) 128–143.

C. Becker and U. Gather, The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association* **94** (1999) 947–955.

A. Buja and D. Asimov, Grand Tour Methods: An Outline, in: D. M. Allen (Ed.): *Proceedings of the Seventh Symposium on The Interface* (Elsevier, Amsterdam, 1986) 63–67.

A. Buja, D. Cook, D. Asimov and C. Hurley, *Theory and Computational Methods for Dynamic Projections in High-Dimensional Data Visualization*, unpublished manuscript available from `http://www.research.att.com/~andreas/`.

N. Christlieb, L. Wisotzki, D. Reimers, D. Homeier, D. Koester and U. Heber, The stellar content of the Hamburg/ESO survey. I. Automated selection of DA white dwarfs, *Astronomy and Astrophysics* **366** (2001) 898–912.

D. A. Coleman and D. L. Woodruff, Cluster Analysis for Large Datasets: An Effective Algorithm for Maximizing the Mixture Likelihood. *Journal of Computational and Graphical Statistics* **9** (2000) 672–688.

J. A. Cuesta-Albertos, A. Gordaliza and C. Matran, Trimmed k-Means: An Attempt to Robustify Quantizers. *The Annals of Statistics* **25** (1997) 553–576.

C. Fraley and A. E. Raftery, How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis. *Computer Journal* **41** (1998) 578–588.

K. Fukunaga, *Introduction to Statistical Pattern Recognition (2nd Ed.)* (Academic Press, Boston, 1990).

R. Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations* (Wiley, New York, 1977).

C. Hennig, Clustering and outlier identification: Fixed Point Cluster analysis, in: Rizzi, A., Vichi, M., Bock, H.-H. (Eds.): *Advances in Data Science and Classification* (Springer, Berlin, 1998) 37–42.

C. Hennig, What clusters are generated by Normal mixtures? In: H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, M. Schader (Eds.): *Data Analysis, Classification and Related Methods* (Springer, Berlin, 2000) 53–58.

L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, New York, 1989).

A. S. Kosinski, A procedure for the detection of multivariate outliers. *Computational Statistics & Data Analysis* **29** (1999) 145–161.

G. McLachlan and D. Peel, *Finite Mixture Models* (Wiley, New York, 2000). Wiley.

C. R. Rao, *Advanced Statistical Methods in Biometric Research* (Wiley, New York, 1952).

D. F. Swayne, D. Cook and A. Buja, XGobi: Interactive Dynamic Data Visualization in the X Window System. *Journal of Computational and Graphical Statistics* **7** (1998) 113–130.

A. F. X. Wilhelm, E. J. Wegman and J. Symanzik, Visual clustering and classification: The Oronsay particle size data set revisited. *Computational Statistics* **14** (1999) 109–146.

L. Wisotzki, N. Christlieb, N. Bade, V. Beckmann, T. Köhler, C. Vanelle and D. Reimers, The Hamburg/ESO survey for bright QSOs. III. A large flux-limited sample of QSOs. *Astronomy and Astrophysics* **358** (2000) 77–87.

L. Wisotzki, T. Köhler, D. Groote and D. Reimers, The Hamburg/ESO survey for bright QSOs I. Survey design and candidate selection procedure, *Astronomy and Astrophysics* **115** (1996),227–233.