Clustering and Outlier Identification: Fixed Point Cluster Analysis

C. Hennig
Institut für Mathematische Stochastik,
Universität Hamburg, D-20146 Hamburg, Germany

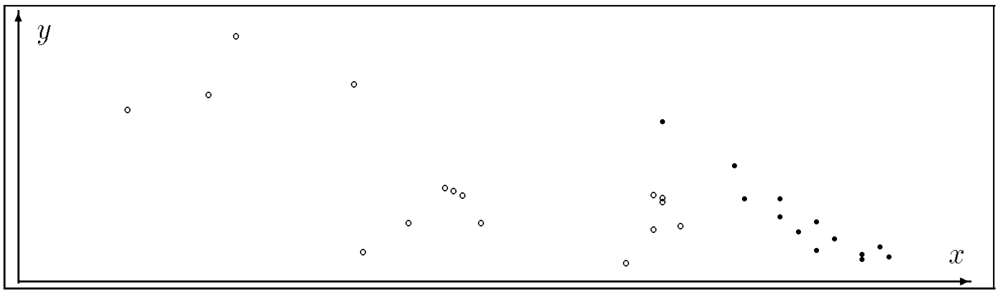
Abstract: Fixed Point Cluster Analysis (FPCA) is introduced in this paper. FPCA is a new method for non-hierarchical cluster analysis. It is related to outlier identification. Its aim is to find groups of points generated by a common stochastic model without assuming a global model for the whole dataset. FPCA allows for points not belonging to any cluster, for the existence of clusters with a different shape, and for overlapping clusters. FPCA is applicated to the clustering of $p-$dimensional metrical data, 0-1-vectors, and linear regression data.

Keywords: Stochastical clustering, overlapping clusters, mixture model, outlier identification, linear regression

1. The cluster concept of Fixed Point Clusters

Sometimes, a dataset does not consist of a partition into some "natural clusters", but nevertheless there seem to be clusters, perhaps of different types. The following dataset is taken from Hand et al. (1994), p. 58, and gives the rates of child mortality ($y-$axis) and adult female literacy ($x-$axis) of 29 countries.

Figure 1: *UNICEF-data*



There seems to be some "grouping" in the data. A linear regression model could be adequate for all or a large part of the filled points, but hardly for considerable parts of the rest.

The aim of Fixed Point Clustering is to find "natural clusters" in a stochastic sense. For a given dataset, a "cluster" in a stochastic sense is a subset whose points are described adequately by a common distribution. This distribution comes from a parametric family of "cluster reference distributions" which model the form of the clusters of interest. For example, it could be Normal or other suitable unimodal distributions, while bimodal distributions heuristically mostly generate two clusters. Frequently, such stochastic cluster analysis problems are treated by means of a mixture model

$$P = \sum_{i=1}^{s} \epsilon_i P_i, \quad \sum_{i=1}^{s} \epsilon_i = 1, \quad \epsilon_i > 0, \ i = 1, \ldots, s, \tag{1}$$

where the $P_i, i = 1, \ldots, s$, are cluster reference distributions with different parameters and the $\epsilon_i, i = 1, \ldots, s$, are the proportions of the s clusters. While the estimation of s is difficult, the parameters of the P_i can often be estimated consistently by Maximum Likelihood estimators if s is known.

In difference to that, FPCA is a local cluster concept, i.e. the cluster property of a data subset depends only on the particular set and its relation to the whole dataset, but not on a global model. FPCA bases on "contamination models" of the form

$$P = (1 - \epsilon) P_0 + \epsilon P^*, \ 0 < \epsilon \leq 1, \tag{2}$$

where P_0 is a cluster reference distribution. P^* can be arbitrary. In particular, it can be a mixture of other cluster reference distributions so that this model includes mixture distributions of the form (1). Additionally, P^* has to be "well seperated" from P_0 in order to guarantee that the points generated by P_0 form a cluster with respect to the whole dataset. An ideal FPC should consist of the points generated by P_0. Contamination models are often used in robust statistics and outlier identification, and a cluster could be seen as a data subset which belongs together, i.e. it contains no outliers, and which is separated from the remaining data, i.e. all other data points are outliers w.r.t. the cluster. FPCA provides a formalization of this imagination.

This concept can cope with more complex data situations:

- If a dataset contains outliers, they need not to be included in any cluster.
- Sometimes a dataset contains clusters of various shapes. For example, there are growth curves in biology, where linear growth is followed by logarithmic growth.
- Clusters can overlap.
- No assumption on the number of clusters is needed.

2. Definition of Fixed Point Clusters

An FPC is a subset of the data, which does not contain any outlier. All other points of the dataset have to be outliers with respect to the FPC.

More formally: Let M be the space of the data points and $\mathbf{Z}_0 \in M^n = (z_1, \ldots, z_n)'$ be the dataset. For $m \in I\!N$ let

$$I_m : \; M^m \mapsto \{0,1\}^M \quad (\mathbf{Z} \in M^m \text{ is mapped on a mapping.}) \tag{3}$$

an "outlier identifier" with the interpretation that a point $z \in M$ is outlier with respect to $\mathbf{Z} \in M^m$, if $I_m[\mathbf{Z}](z) = 1$. The construction of reasonable outlier identifiers will be discussed in later sections. For an indicator vector $g = (g_1, \ldots, g_n)' \in \{0,1\}^n$, let $\mathbf{Z}_0(g)$ be the matrix of the data points z_i, for which $g_i = 1$, and $n(g)$ be the number of these points.
Definition. *With outlier identifiers $I_1, \ldots, I_n$ define*

$$f_{\mathbf{Z}_0} : \; \{0,1\}^n \mapsto \{0,1\}^n, \; g \mapsto (1 - I_{n(g)}[\mathbf{Z}_0(g)](z_i))_{i=1,\ldots,n}, \tag{4}$$

i.e. $f_{\mathbf{Z}_0}$ maps g on the indicator vector of the $\mathbf{Z}_0(g)-$non-outliers. Then $\mathbf{Z}_0(g)$ is an FPC if g is a fixed point of $f_{\mathbf{Z}_0}$. Thus, the non-outliers w.r.t. $\mathbf{Z}_0(g)$ have to be precisely the points of $\mathbf{Z}_0(g)$.

This definition can be related to stochastic modeling as follows: Let M be a space with $\sigma-$Algebra $\mathcal{B}$. Let $\mathcal{P}$ be the space of distributions on $(M, \mathcal{B})$, and $\mathcal{P}_0 \subset \mathcal{P}$ the set of cluster reference distributions.
Definition. *(Davies and Gather 1993) $A(P)$ is called $\alpha-$outlier region w.r.t. P under*

$$A : \; \mathcal{P} \mapsto \mathcal{B} \text{ with } \forall P \in \mathcal{P}_0 : \; P(A(P)) \leq \alpha. \tag{5}$$

$0 < \alpha < 1$ should be small so that the points in $A(P)$ "lie really outside". Now, the set of points $A_m(\mathbf{Z}) := \{z : \; I_m[\mathbf{Z}](z) = 1\}$ is an estimator of $A(P)$, if $\mathbf{Z}$ is a dataset of i.i.d. points distributed by P. If $\mathbf{Z} = \mathbf{Z}_0(g)$ forms an FPC w.r.t. $\mathbf{Z}_0$, $A_m(\mathbf{Z})^c$, i.e. the set of $\mathbf{Z}$-non-outliers, is an estimator (whose quality is to be assessed) for an FPC-set of a distribution:
Definition. *For a given distribution $P \in \mathcal{P}$ and $B \in \mathcal{B}$ let $P_B := P(\bullet|B)$ the distribution P restricted to the points of B. Let A be an outlier region. Define*

$$f_P : \; \mathcal{B} \mapsto \mathcal{B}, \quad B \mapsto A(P_B)^c. \tag{6}$$

B is an FPC-set w.r.t. P if B is a fixed point of f_P.

With help of this definition, the quality of the results of an FPCA with respect to distributions P of the form (2) can be investigated by comparing the FPC-sets B of P (the distributions P_B, respectively) to P_0, e.g. the high density areas of P_0.

The computation of all FPCs of a dataset requires the evaluation of (4) for every subset of data. This is impossible, except for very small datasets. Instead, one can apply the usual fixed point algorithm

$$g^{i+1} = f(g^i) \tag{7}$$

to various starting vectors g^0.

Now, the application of this concept to various clustering problems will be demonstrated.

3. Multidimensional Normal clusters

The estimation of Normal mixture parameters is discussed broadly in the literature, see e.g. Titterington et al. (1985). In this setup, a cluster of points from $I\!R^p$ is considered as a group of points generated by a common $\mathcal{N}_p(\mu, \mathbf{\Sigma})$-distribution. The definition of FPCs is given by the definition of the cluster reference distributions and the corresponding outlier identifier. Here:

$$\mathcal{P}_0 := \{\mathcal{N}_p(\mu, \mathbf{\Sigma}) : \ \mu \in I\!R^p, \mathbf{\Sigma} \text{ pos. semidef. } p \times p\}. \tag{8}$$

An easy $\alpha-$outlier region for the distributions of $\mathcal{P}_0$ has the form

$$A(\mathcal{N}_p(\mu, \mathbf{\Sigma})) := \{x \in I\!R^p : (x-\mu)'\mathbf{\Sigma}^{-1}(x-\mu) > \chi^2_{p;1-\alpha}\}. \tag{9}$$

Such outlier regions can be estimated by replacing μ and $\mathbf{\Sigma}$ by estimators $\hat{\mu}_m$ and $\hat{\mathbf{\Sigma}}_m$. This leads to the following outlier identifier: For $\mathbf{X} \in (I\!R^p)^m$ and $x \in I\!R^p$:

$$I_m[\mathbf{X}](x) := 1\left((x - \hat{\mu}_m(\mathbf{X}))'\hat{\mathbf{\Sigma}}_m(\mathbf{X})^{-1}(x - \hat{\mu}_m(\mathbf{X})) > \chi^2_{p;1-\alpha}\right). \tag{10}$$

Chosing $\hat{\mu}_m$ and $\hat{\mathbf{\Sigma}}_m$ as mean vector and sample covariance matrix leads to an outlier identification by means of the Mahalanobis distance. This approach is computationally simple and is discussed e.g. in Rousseeuw and Leroy (1988). They show that a better identification procedure can be obtained if $\hat{\mu}_m$ and $\hat{\mathbf{\Sigma}}_m$ are replaced by more robust estimates.

4. Clustering 0-1-vectors

Let $\mathbf{Z} \in (\{0,1\}^k)^n$ a dataset of $k-$dimensional 0-1-vectors. Then data points can be considered as belonging to a common cluster if they seem to be generated by the same independent product of Bernoulli distributions. That is,

$$P[k, p_1, \ldots, p_k] := \bigotimes_{j=1}^{k} B[1, p_j], \quad \mathcal{P}_0 := \left\{P[k, p_1, \ldots, p_k] : (p_1, \ldots, p_k) \in [0,1]^k\right\},$$

where $B[k,p]$ denotes the Binomial distribution with parameters k,p. An α-outlier region for this kind of data can be defined by counting the components for which a data point does not belong to the expected majority in the cluster:

$$\begin{aligned} &\forall P[k,p_1,\ldots,p_k] \in \mathcal{P}_0 : \\ A(P[k,p_1,\ldots,p_k]) &:= \left\{ z \in \{0,1\}^k : v[p_1,\ldots,p_k](z) \geq c(\alpha) \right\}, \\ &\text{where } c(\alpha) := \min \left\{ c : 1 - B[k,\tfrac{1}{2}](c) \leq \alpha \right\}, \\ v[p_1,\ldots,p_k](z) &:= \sum_{j=1}^{k} \left[z_j 1\left(p_j \leq \frac{1}{2}\right) + (1-z_j) 1\left(p_j > \frac{1}{2}\right)\right]. \end{aligned} \tag{11}$$

Here v counts the "minority components" of z and $c(\alpha)$ is chosen such that (5) holds for all $P \in \mathcal{P}_0$, because $B[k,\frac{1}{2}]$ is the distribution of $v[\frac{1}{2},\ldots,\frac{1}{2}](z)$. This is stochastically largest among the $v[p_1,\ldots,p_k](z)$, where $\mathcal{L}(z) \in \mathcal{P}_0$.

Again, a proper outlier identifier can be defined by replacing $p_1,\ldots,p_k$ by estimators in the definition of v, e.g.

$$\hat{p}_{jm}(\mathbf{X}) := \frac{1}{m}\sum_{i=1}^{m} x_{ij}, \text{ where } \mathbf{X} := (x_1,\ldots,x_m) \in (\{0,1\}^k)^m. \tag{12}$$

5. Linear regression clusters

For analyzing the UNICEF-Data, the female literacy can be considered as explanatory variable for the child mortality. While there is no clear structure for the whole dataset, the relation for certain subsets of the data seems to be approximatively linear. Cluster reference distributions for linear regression can be defined as follows:

$$\mathcal{P}_0 := \{P_{\beta,\sigma^2,G} : \ \beta \in I\!R^{p+1}, \sigma^2 \in I\!R^+, G \in \mathcal{G}\} \tag{13}$$

where $P_{\beta,\sigma^2,G} \in I\!B^{p+1}$ is the distribution of $(x,y) \in I\!R^{p+1} \times I\!R$ defined by

$$y = \beta' x + u, \quad u \sim \mathcal{N}_{0,\sigma^2}, \tag{14}$$

$$(x_1,\ldots,x_p)' \sim G \text{ independently of } u, \ x_{p+1} = 1. \tag{15}$$

$\mathcal{G}$ is some suitable space of distributions on $(I\!R^p, I\!B^p)$, and a simple α-outlier region can be constructed:

$$A(P_{\beta,\sigma^2,G}) := \{(x,y) \in I\!R^p \times \{1\} \times I\!R : (y - x'\beta)^2 > c(\alpha)\sigma^2\}, \tag{16}$$

where $c(\alpha)$ is the $1-\alpha$-quantile of the χ^2-distribution.

If β and σ^2 are replaced by estimators $\hat{\beta}_m$ and $\hat{\sigma}^2_m$, the indicator function of A is again a reasonable outlier identifier. The best choices for $\hat{\beta}_m$ and $\hat{\sigma}^2_m$

would be robust estimators, but they impose formal and numerical problems. If one uses Least Squares and

$$\hat{\sigma}^2_m(\mathbf{X}) := \tfrac{1}{m-1} \textstyle\sum_{i=1}^n (y_i - x_i'\hat{\beta}_m(\mathbf{X}))^2, \tag{17}$$
$$\text{where } \mathbf{X} := ((x_1, y_1), \ldots, (x_m, y_m)) \in (I\!R^p \times \{1\} \times I\!R)^m,$$

then the subset marked by full points in figure 1 turns out as an FPC with respect to the UNICEF-data. This subset contains all south american countries, Oman and Czechoslovakia, the only european country in the sample.

6. Conclusion

FPCA needs less restrictive assumptions than usual cluster analysis methods based on stochastic models. It enables the analysis of more complex data situations. The idea applies to various problems of cluster analysis. Theoretical background is presented in Hennig (1997), especially for the linear regression setup.

References

Davies, P. L. and Gather, U. (1993). The identification of multiple outliers, *Journal of the American Statistical Association* 88, 782-801.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994). *A Handbook of Small Datasets*, Chapman & Hill, London.

Hennig, C. (1997). *Datenanalyse mit Modellen für Cluster linearer Regression*, dissertation, Universität Hamburg.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.