A A

Fixed Point Clusters and their Relation to Stochastic Models

Christian Hennig

Preprint–No. 96-03 März 1996

INSTITUT FÜR MATHEMATISCHE STOCHASTIK

BUNDESSTR. 55, D – 20146 HAMBURG

Fixed Point Clusters and Their Relation to Stochastic Models

C. Hennig[1]

[1] Institut für Mathematische Stochastik,
Universität Hamburg, Bundeststr. 55, D-20146 Hamburg, Germany

Summary: The idea of Fixed Point Clusters (FPC) will be introduced in this paper. An FPC is a set of data points which is a fixed point of a procedure consisting of a statistical parameter estimation and a classification of data in "near" and "outlying" with respect to the estimator. The relation of Fixed Point Clustering to parameter estimation in stochastic models will be discussed. FPCs use stochastic models as orientation rather than as basis. They are mainly a tool for explorative data analysis and model building. FPCs are a useful alternative to other clustering approaches at least in clusterwise Linear Regression.

1. Fixed Point Clusters - The General Case

The idea of Fixed Point Clusters is as follows: Given a dataset from $(\mathbb{R}^d)^n$, we call a subset of the data points a "cluster" if one can define an area, where these data points appear highly concentrated compared to the points which lie in the neighborhood outside this area. How can we find such areas? We take a subset of the data points and classify every point in the whole dataset with respect to the criterion, if it lies "near" or "far" to the points of our subset. In other words, we carry out an outlier classification. Here this means to compute an area which represents the points of the subset as a whole and to look if each data point of the dataset lies inside or outside this area. We call the subset of data points an FPC if it reproduces itself (i.e. it is a fixed point) with respect to this procedure. Thus, all points of an FPC are classified to lie "near" to themselves, all other points are classified to lie outside. "Themselves" here means a structure which represents the "general tendency" of the points in the subset.

The crucial point is how to assign such a structure to a set of data points. If we imagine a stochastic model as an adequate description of the points belonging to a common cluster, it is natural to estimate the parameters of the model and compute an area where points generated by the model would lie with high probability. In many stochastic models this can be done by computing those data points which have small deviation from the predicted value by the model. This will be formalized using a distance between a data point and a parameter estimator.

Definition: Let $(z_1, \ldots, z_n)$, $\quad z_i \in \mathbb{R}^d, i = 1, \ldots, n$ denote the dataset, $g \in \{0,1\}^n$ be an indicator vector where $g_i = 1$ means that z_i belongs to the set indicated by g.

Let $\hat{\theta}$ be some parameter estimator, i.e. $\hat{\theta}$ maps from the space of datasets

into some parameter space Θ. $\hat{\theta}(g)$ denotes $\hat{\theta}$ applied to the points indicated by g. $D : I\!R^d \times \Theta \mapsto [0, \infty)$ denotes a distance between a data point and a parameter estimator. Let $c > 0$ be a tuning constant.
Then g is a **Fixed Point Cluster Vector** (FPCV), iff g is fixed point of

$$f : \{0,1\}^n \mapsto \{0,1\}^n, \tag{1}$$

$$f_i(g) := 1\{D(z_i, \hat{\theta}(g)) < c\}, \quad i = 1, \ldots, n. \tag{2}$$

c will be chosen such that

$$P_{\hat{\theta}(g)}\{D(z, \hat{\theta}(g)) < c\} \text{ is large.} \tag{3}$$

Example: In the one-dimensional location setup, one could choose the sample mean or median as $\hat{\theta}$ and $D(x,t) := |x - t|$. The choice of c then depends on the expected variance of the clusters. If one wants to find clusters with different and unknown variances, $\hat{\theta} = (\hat{t}, \hat{s}^2)$ is a more reasonable choice, where $\hat{t}$ is some location estimator and $\hat{s}^2$ is some scale estimator. Then one could choose $D(x, (t, s^2)) := \frac{(x-t)^2}{s^2}$.
Computation: The natural approach to compute a fixed point is the usual fixed point algorithm:

Step 1: Choose $g^0 \in \{0,1\}^n$.

Step 2: $g^{j+1} := f(g^j)$.

Step 3: $g^j = f(g^j)$ or $j := j + 1$, step 2 again.

In general, convergence of this algorithm is not clear. It can be forced by replacing step 2 by

Step 2b: $g_i^{j+1} := f_i(g^j)$ for $g_i^j = k$, $\quad g_i^{j+1} := g_i^j$ else,

$k = 0$ or $k = 1$ fixed. Then, on the other hand, the result is not necessarily an FPCV. If step 2b is used alternating with $k = 0$ and $k = 1$, then the algorithm has been proven to converge towards an FPCV at least in the Regression setup (Hennig (1995)).
In order to find all FPCVs contained in a dataset, one theoretically has to check the fixed point property for every subset of the data. This is impossible in practice. Thus, one has to start the algorithm many times with randomly chosen starting sets. Depending on the concrete clustering problem, it has to be investigated if there is a high probability to get every "relevant" cluster by this method.
However, it is always possible to check if a particular set of data points forms an FPC. One can also find a "familiar" cluster by using it as starting point for the algorithm.

2. Application to Clusterwise Linear Regression

In clusterwise Linear Regression, data points should be treated as "similar", if they follow the same linear relation between an independent variable $x \in I\!R^p$ and a dependent variable $y \in I\!R$. There are different possibilities to model such a situation. In the *Finite Mixture Model* the independent variables and the cluster membership of each point are assumed to be randomly generated:

Finite Mixture Model:

$$H(x, y) := \sum_{j=1}^{s} \pi_j H_j(x, y), \quad \pi_j > 0 \forall j = 1, \ldots, s, \quad \sum_{j=1}^{s} \pi_j = 1, \tag{4}$$

where the distribution functions H_j are defined by $x \sim G_j$ with some c.d.f. G_j on $I\!R^p$,

$$y = x'\beta_j + e, \qquad e \sim F_j, \quad j = 1, \ldots, s, \tag{5}$$

where $\beta_j \in I\!R^p$, $\quad E_{F_j}(e) = 0$, $\quad e$ stochastically independent of x. To specify the estimators and the constant for the definition of Fixed Point Clusters, let $F_j = \Phi(0, \sigma_j^2)$ the c.d.f. of the Normal distribution with mean 0 and variance σ_j^2. An intercept parameter can be included by setting the $p+1$th component of each x_i equal to 1.

Remark: Although the theoretical results on Regression FPCs use the Finite Mixture Model, it would be also possible to work with fixed independent variables or fixed but unknown cluster membership of each points ("*Fixed Partition Model*").

Cluster analysis in such a situation is difficult. There are no adequate distances between two points, because two points can always be fitted perfectly by a suitable Regression parameter β. The components of the model can be estimated by a Maximum Likelihood approach (DeSarbo and Cron (1988)), but this approach is not robust against model deviations (e.g. the presence of outliers), the number of clusters has to be assumed as known (DeSarbo and Cron suggest the use of an information criterion) and the model is not generally identifiable which implies that some parameters cannot be estimated consistently (Hennig (1996)).
Fixed Point Clusters are useful in this case, because they do not force a partition of the dataset. Thus, they can exclude outliers and can make various suggestions if the parameters are not uniquely defined because of identifiability problems.

Definition: To define Fixed Point Clusters in the Regression setup, let $z_i := (x_i, y_i)$, $i = 1, \ldots, n$ a single observation, $X := (x_1, \ldots, x_n)'$, $y := (y_1, \ldots, y_n)'$,

$$\hat{\theta} := (\hat{\beta}, \hat{\sigma}^2), \text{ where} \tag{6}$$

$$\hat{\beta}(X, y) := (X'X)^{-1}X'y, \quad \hat{\sigma}^2(X, y) := \frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - x_i'\hat{\beta}(X, y))^2, \tag{7}$$

$$D(x,y,\beta,\sigma^2) := \frac{(y-x'\beta)^2}{\sigma^2}. \tag{8}$$

To choose c, the distribution of D would be useful under the hypothesis, that β and σ^2 would be the parameter estimators for the mixture component containing (x,y). This distribution is unknown since denominator and nominator are dependent, because (x,y) was used to compute the estimators. As an approximation, a high quantile of the χ^2-distribution is a reasonable choice for c.
An FPC is "relevant" in this situation (concerning the discussion on the algorithm in the previous section), if it corresponds to a component of the mixture distribution. Simulations show, that if the proportions π_i of the components are not too small and the dimension p of the independent variables is not too large, one can get all relevant FPCVs with a moderate number of algorithm runs. The result of such a Fixed Point Cluster analysis will be illustrated as follows.

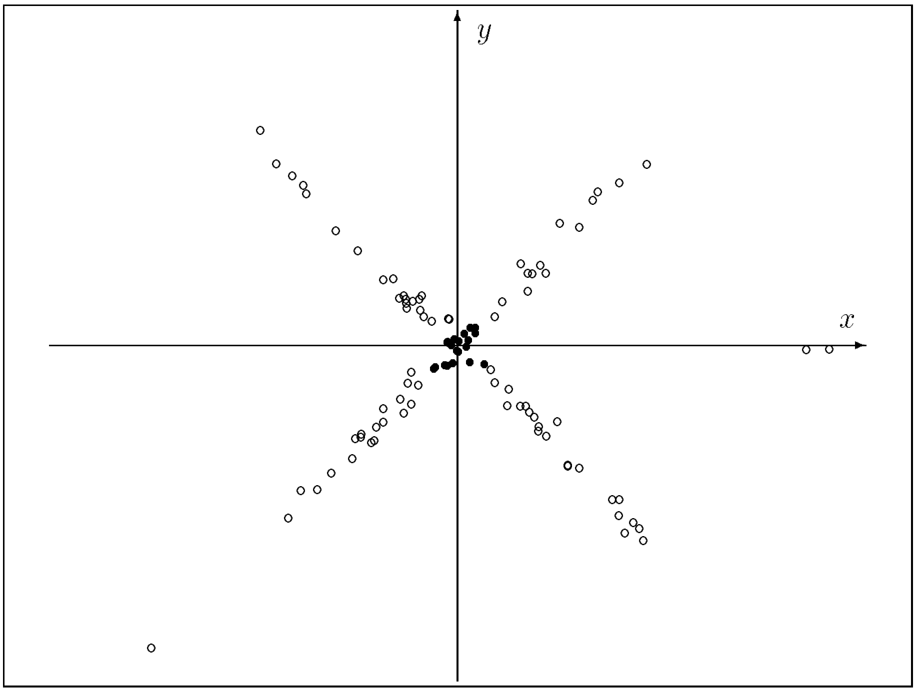

Figure 1: Artificial dataset

Example: Figure 1 shows an artificial generated dataset, where the points 1-50 were generated according to

$$x \sim \Phi(0,1), \quad y = x + e, \quad e \sim \Phi(0, 0.01), \tag{9}$$

the points 51-98 were generated according to

$$x \sim \Phi(0,1), \quad y = -x + e, \quad e \sim \Phi(0, 0.01), \tag{10}$$

and the points 99 and 100 are outliers. $c = 10$ and 150 runs of the algorithm yielded 8 clusters of size larger than 6. A minimum cluster size has to be specified, since one often gets many very small and useless FPCs. Especially all sets of points which can be fitted perfectly are trivially FPCs.
One cluster contains all points, a second contains all points except the point with lowest y-value. These FPCs can be left out from the interpretation of the results, since FPCs, which contain almost all points, usually exist. But they are only useful for outlier analysis if almost the whole dataset is homogenous. The existence of additional clusters, which were found more than once by the algorithm, indicates that this is not the case.
The remaining six clusters contain either the points 1-50 or the points 51-98. None of them contain the outliers. The two FPCs found most often from this six contain 53 and 64 points, respectively. Their parameter estimators (including intercept component) are

$$\hat{\beta}_1 = (0.998, -0.001), \quad \hat{\sigma}_1^2 = 0.013, \tag{11}$$

$$\hat{\beta}_2 = (-1.003, -0.007), \quad \hat{\sigma}_2^2 = 0.028. \tag{12}$$

The points represented by bullets lie in the intersection of these FPCs. They bias $\hat{\sigma}_2^2$, since they were almost all generated by the first component. The other clusters only differ from these two in what they contain from the "intersection" in the center. This can be used in the analysis of how clear the single points can be assigned to the components.

3. The Cluster Concept of Fixed Point Clusters

So far, FPCs have been defined and illustrated. The following section summarizes some ideas of the meaning of the concept "cluster", defined as above. Particular attention will be paid to the relation between FPCs and stochastic modelling. The notion of FPCs is inspired by stochastic models, but it is not a usual model based procedure. Here models offer a possibility to construct structure in the perception of a dataset, but they are not assumed to be true, not even approximately true.

3.1 Fixed Point Clusters are Probabilistic

The main role of stochastic models for FPCs is that the points belonging to a single FPC are treated as if coming from a common model. In the fixed point iteration, all of them and no other are used to compute the parameter estimators, none of them but all others are classified as outliers with respect to their parameter estimators. The common model for a single FPC has to be "homogenous" in some sense, i.e. it must be considered as "cluster-generating". For example, a Normal distribution can generate a "cluster" in contrast to a bimodal distribution with well seperated areas of high density. For the general idea it does not matter where the remaining points come from, if some points are classified to belong to an FPC (though it does matter for the quality of the parameter estimators). Thus, the models H for which FPCs are supposed to be adequate only need to have a single component H_0 generated by a distribution from the class which defines the estimators:

$$H := \epsilon H_0 + (1-\epsilon)H^*, \quad 0 < \epsilon \leq 1. \tag{13}$$

This formalization is familiar in Robust Statistics, if the data is treated as a combination of "good data" and outliers. ("*ϵ-Contamination Neighborhood*" of H_0, usually $\epsilon < \frac{1}{2}$. The latter is not necessary for FPCs.) The points modelled by the arbitrary H^* need not even be considered as coming from any stochastic model. The Finite Mixture Regression Model is only the special case, in which FPCs would be adequate for every data point.

Remark: Unfortunately this cannot be used to investigate the adequacy of a certain model for the points of a particular FPC, since the algorithm treats all points as generated by the model until they are classified as outliers. In the given example the whole dataset forms an FPC and only in comparison with the other FPCs in the dataset it becomes clear that a homogenous model is not reasonable for it. If a Linear Regression model is not adequate for any part of the data, there nevertheless will be FPCs.

According to that, stochastic models should be used to assess the quality of the method. A first approach to do this for the Linear Regression case is to formulate the method as a functional on the space of probability distributions of the domain of the data points. The values of the functional are sets of indicator functions (or their corresponding parameters), each corresponding to an FPC. Then the key question is: What is the result for the distributions of (13) ?
It is proven (Hennig (1995)) that for a homogenous population ($\epsilon = 1$) the functional is Fisher consistent, i.e. there exists an unique FPC with the same parameters as the model H_0 (σ^2 standardized with a proper constant). If the population is not homogenous, H^* usually will generate some points which are not recognized as outliers with respect to H_0. This causes a bias on the parameter estimates of the FPC corresponding to H_0. The existence of such an FPC and bounds for the bias can be proven if H_0 and H^* are well separated.
Consequently Fixed Point Clustering is at least a "probabilistic" concept.

3.2 Fixed Point Clusters are not Model Based

A procedure is considered as "model based" in this context, if it assumes the whole dataset to be generated by a member of some class of stochastic distributions and, additionally, a decision problem concerning some distribution parameters is solved under this assumption.
Some differences between FPCs and model based methods:

- The data is not assumed to be generated by a common stochastic model.
  - For a single FPC only a part of the data is considered: The data modelled by H_0 in the contamination model (13).

- Stochastic models (including the contamination model) are not treated as "true" but as suggestions to give adequate images for at least parts of the data.

- FPCs are not a solution of a certain decision problem. They do not optimize a target criterion formalized in the context of a certain model.
- Overlap and inclusion of FPCs as well as points, which do not belong to any FPC, are possible. The result of a Fixed Point Cluster analysis gives neither a model nor a class of models for the whole dataset nor a probabilistic classification for single data points.

Consequently, FPCs offer a different point of view on the data and their relation to stochastic models than model based procedures. The question of adequecy of a stochastic model is treated as open. FPCs give various possibilities for interpretations of parts of the data but they do not force decisions between them.
For example, classification methods based on a Fixed Partition Model classify each point to a particular cluster. This is not necessarily adequate. In the data of figure 1 one would loose the information, that the points in the intersection are fitted good from both mixture components.
Another example: If there were an FPC in the data example which contained the outliers and some small part of the other points, the decision whether the outliers are "outlying" to the "good" data or whether they are a part of an own mixture component could not be made on the basis of the data alone. A model based procedure would base a decision on an uncertain assumption. Contrarily to that, FPCs are more an explorative tool. They make suggestions for considering points as belonging together, i.e. generated by a common model. This information can be used only to get an idea from the present data, but it can also be used to build a model for a later survey, to decide on significant extern factors defining a cluster membership or to validate the results of a model based clustering.

3.3 Modeling the Self Organization of Perception

There is another way of motivating FPCs. It is related to the idea of cognition in self organization theory. Heinz von Förster (1976) formalizes human cognitive actions as recursive operators on a domain of "observables", i.e. a cognitive action operates with its own output:

$$\mathrm{obs}_{i+1} = \mathrm{COORDINATION}(\mathrm{obs}_i). \tag{14}$$

Cognitive processes are treated as the recursive computation of a reality depending on the way the subject operates. This is illustrated by von Förster (1973)[1] as follows:

$$\textit{Cognition} \longrightarrow \textit{computations of}$$

[1]The German translation of both papers is included in von Förster (1993)

This idea can also be explained by biological arguments. Von Förster refers to Piaget (1975) and Maturana (1970).
Von Förster states that for an external observer the "objects" of a person cannot be distinguished from the fixed points[2] of the operation (14). This means: The observer constitutes an object as "recognized" by the person observed if the coordinations of actions of the person concerning the object are stable.
Fixed Point Clustering models such a process. The clusters (or patterns) of a dataset can be viewed as the "objects" to be recognized. The function f defines the action the algorithm carries out with the data points which play the role of the initial observables. An FPC is a stable point of this process.

References:

DESARBO, W. S., and CRON, W. L. (1988): A Maximum Likelihood Methodology for Clusterwise Linear Regression. *Journal of Classification 5*, 249-282.

HENNIG, C. (1995): Fixed Point Cluster Analysis for Regression Data: Computation and Weak Consistency, Preprint No. 95-2, Institut für Mathematsche Stochastik, Universität Hamburg.

HENNIG, C. (1996): Identifiability of Finite Linear Regression Mixtures, Preprint No. 96-6, Institut für Mathematsche Stochastik, Universität Hamburg.

MATURANA, H. R. (1970): Neurophysiology of Cognition. In: P.L. Garvon (ed.): *Cognition: A Multiple View*, Spartan Books, New York, 3-23.

PIAGET, J. (1975): L'Equilibration des structures cognitives, P.U.F., Paris.

VON FÖRSTER, H. (1973): On Constructing a Reality. In: W.F.E. Preiser (ed.): *Environmental Design Research*, vol. 2, Stroudberg, 35-46.

VON FÖRSTER, H. (1976): Objects: Tokens for Eigen-Behaviors. *ASC Cybernetic Forum 8*, 91-96.

VON FÖRSTER, H. (1993): Wissen und Gewissen (ed. by S.J. Schmidt), Suhrkamp, Frankfurt/Main.

[2]Von Förster uses the older term "Eigenvalue"