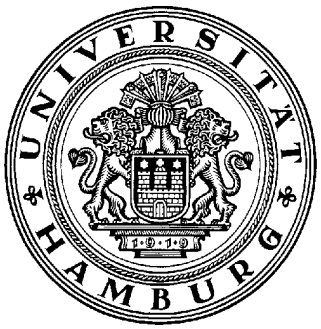


A A

Fixed Point Clusters,

Clusterwise Linear Regression

Christian Hennig

Preprint–No. 98-05 März 1998

INSTITUT FÜR MATHEMATISCHE STOCHASTIK

BUNDESSTR. 55, D – 20146 HAMBURG

Clustering and Outlier Identification: Fixed Point Cluster Analysis

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

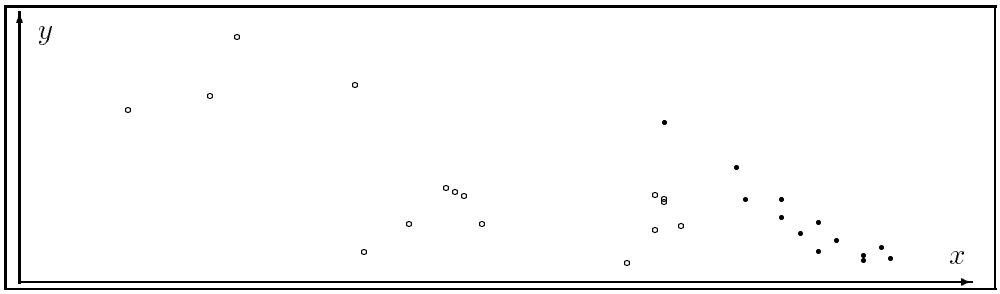
Abstract: Fixed Point Cluster Analysis (FPCA) is introduced in this paper. FPCA is a new method for non-hierarchical cluster analysis. It is related to outlier identification. Its aim is to find groups of points generated by a common stochastic model without assuming a global model for the whole dataset. 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:

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

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

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

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.

MODELS AND METHODS FOR CLUSTERWISE LINEAR REGRESSION[1]

C. Hennig

Institut für Mathematische Stochastik,
Universität Hamburg, Bundesstr. 55, D-20146 Hamburg, Germany

Abstract: Three models for linear regression clustering are given, and corresponding methods for classification and parameter estimation are developed and discussed: The mixture model with fixed regressors (ML-estimation), the fixed partition model with fixed regressors (ML-estimation), and the mixture model with random regressors (Fixed Point Clustering). The number of clusters is treated as unknown. The approaches are compared via an application to Fisher's Iris data. By the way, a broadly ignored feature of these data is discovered.

1 Introduction

Cluster analysis problems based on stochastic models can be divided into two classes:

1. A cluster is considered as a subset of the data points, which can be modeled adequately by a distribution from a class of *cluster reference distributions* (c.r.d.). These distributions are chosen to reflect the meaning of *homogeneity* with respect to the certain data analysis problem. Therefore c.r.d. are often unimodal. If the class of c.r.d. is parametric, then one is interested in classification of the data points and parameter estimation within each cluster.

2. A cluster is considered as an area of high density of the distribution of the whole dataset. No distributional assumption is made for the single clusters.

Clusterwise linear regression is a problem of the first kind since the points of each cluster are considered to be generated according to some linear regression relation, i.e. one imagines a separate model for each cluster. The class of c.r.d. for the regression clustering problem contains distributions of the following kind: Consider a dataset $\mathbf{Z} = (x_i', y_i)_{i \in I}$, $x_i \in \{1\} \times I\!R^p, y_i \in I\!R$, I being some index set.

$$\mathcal{L}(y_i|x_i) = F_{(x_i,\beta,\sigma^2)}, \text{ defined by}$$
$$y_i = x_i'\beta + u_i, \qquad \mathcal{L}(u_i) = \mathcal{N}_{(0,\sigma^2)},$$
$$(\beta, \sigma^2) \in I\!R^{p+1} \times I\!R_0^+.$$

[1]To appear in *Proceedings of GfKl-98, Dresden.*

The first component of β denotes the intercept. The $u_i, i \in I$ are considered to be stochastically independent. The x_i are called *regressors* in the following. They can be fixed or random with $\mathcal{L}(x_i) = G$ from some class of distributions $\mathcal{G}$. In the latter case the regressors are assumed to be i.i.d. and independent of $(u_i)_{i\in I}$. F_{G,β,σ^2} then denotes the joint distribution of (x_i, y_i). In our setup, all parameters are considered as unknown.

The models will be divided into fixed and random regressor models, and into mixture and fixed partition models. Mixture models treat the cluster membership of a point as random, fixed partition models contain parameters for the cluster membership of each point. A fixed partition model with random regressors will not be given because this does not lead to an easy clustering method. The purpose of the model based approach presented here is not to describe the mechanism generating the data, but to find an adequate description of the data themselves. Thus, all models can be applied to the same data. In particular, the question is ignored if the regressors were *really* fixed or random.

The literature on clusterwise linear regression either treats the mixture model with fixed regressors (e.g. Quandt and Ramsey (1978), for general p and number of clusters DeSarbo and Cron (1988)) or discusses algorithms for a least squares solution (e.g. Bock (1969), Spaeth (1979)) which is related to the fixed regressors fixed partition model presented here in the case of equal error variances for each cluster.

2 Fixed regressors, mixture model

Let I be an index set, usually $I = \{1, \ldots, n\}$. With a given regressor design $(x_i)_{i\in I} \in (\{1\} \times I\!R^p)^I$ the **fixed regressors mixture model** (FRM) is defined by

$$\mathcal{L}((y_i)_{i\in I}) = \bigotimes_{i\in I} \sum_{j=1}^{s} \epsilon_j F_{(x_i,\beta_j,\sigma_j)},$$

$$\sum_{j=1}^{s} \epsilon_j = 1, \ \epsilon_j > 0, \ j = 1, \ldots, s.$$

That is, s denotes the number of clusters and ϵ_j denotes the proportion of the cluster j. The log-likelihood function

$$\ln L_n(s, (\epsilon_j, \beta_j, \sigma_j^2)_{j=1,\ldots,s}, \mathbf{Z}) =$$
$$\sum_{i\in I} \ln \left(\sum_{j=1}^{s} \epsilon_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left[\frac{-(y_i - \beta_j' x_i)^2}{2\sigma_j^2} \right] \right)$$

can be locally maximized for given s via the EM-algorithm described in DeSarbo and Cron (1988). This works only subject to $\sigma_j^2 > c \ \forall j$ with some lower bound $c > 0$ (e.g. $c = 0.001$) since otherwise $\ln L_n$ would be

unbounded. After having performed the algorithm, point i can be classified to cluster $\hat{\gamma}(i) \in \{1, \ldots, s\}$ according to

$$\hat{\gamma}(i) = \arg\max_j \hat{\epsilon}_{ij}, \quad \hat{\epsilon}_{ij} = \frac{\hat{\epsilon}_j \varphi_{(x_i'\beta_j, \hat{\sigma}_j^2)}(y_i)}{\sum_{l=1}^{s} \hat{\epsilon}_l \varphi_{(x_i'\beta_l, \hat{\sigma}_l^2)}(y_i)}.$$

$\hat{\epsilon}_{ij}$ denotes the estimated a posteriori probability for point i to be generated by mixture component j.

DeSarbo and Cron (1988) suggest Akaike's Information Criterion (AIC) for the estimation of s:

$$\hat{s} := \arg\max_s \ln \hat{L}_n(s) - k(s), \; k(s) = (p+3)s - 1.$$

$k(s)$ denotes the number of free parameters to estimate for the cluster number s and $\ln \hat{L}_n(s)$ is the estimated maximum log-likelihood. Their simulations do not treat the performance of this proposal, and Hennig (1997b) showed the tendency of the AIC to overestimate a small number of clusters. Schwarz' Criterion (SC) gives smaller estimates of s for $n > e^2$ and seems to work better:

$$\hat{s} := \arg\max_s \ln \hat{L}_n(s) - \frac{\ln n}{2} k(s).$$

The discussion of the Iris data example in section 5 illustrates this performance.

Some alternative proposals for parameter estimation within this model were made (e.g. Quandt and Ramsey (1978)), but they lead to greater numerical difficulties and were investigated only for $s = 2, p = 1$.

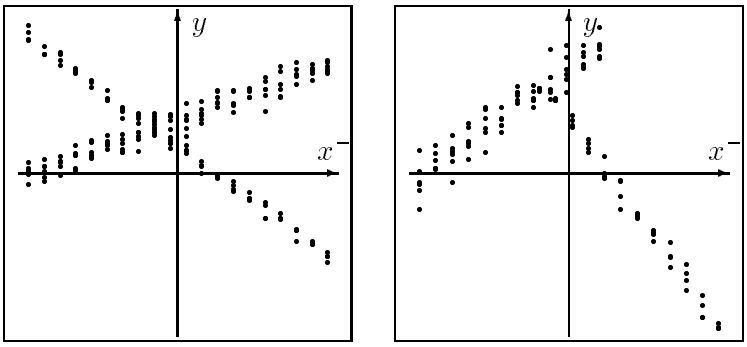


Figure 1: Assignment independence - assignment dependence

The implicit assumption of **assignment independence** is a disadvantage of the FRM. That is, the clusters keep the same proportions $\epsilon_j, j = 1, \ldots, s$ for every fixed regressor x_i (see figure 1). The probability of a point (x_i, y_i) to be generated by cluster j is independent of x and i. This is not generally true. For example in a **change point** setup, the cluster membership is considered as *determined* by x or i. Methods concerning this particular assumption can be found e.g. in Krishnaiah and Miao (1988). Also for the Iris data in section 5, assignment independence seems not to be fulfilled.

3 Fixed regressors, fixed partition model

In the fixed partition approach, the cluster membership of each point i is indicated by a parameter $\gamma(i)$. Thus, general kinds of assignment dependency can be modelled. The **fixed regressors fixed partition model** (FRFP) is given by

$$\mathcal{L}((y_i)_{i\in I}) = \bigotimes_{i\in I} F_{\left(x_i,\beta_{\gamma(i)},\sigma^2_{\gamma(i)}\right)},$$
$$\gamma:\ I \mapsto \{1,\ldots,s\},$$

$(x_i)_{i\in I} \in (\mathbb{R}^{p+1})^I$ again given fixed. Under known s, ML-estimation is also possible within this model. The log-likelihood function is given by

$$\ln L_n(s,\gamma,(\beta_j,\sigma_j^2)_{j=1,\ldots,s},\mathbf{Z}) =$$
$$-\tfrac{1}{2}\sum_{j=1}^{s}\sum_{\gamma(i)=j}\left(\ln 2\pi + \ln\sigma_j^2 + \frac{-(y_i-\beta_j' x_i)^2}{2\sigma_j^2}\right). \qquad (1)$$

For given $(\hat\beta_j,\hat\sigma_j^2)_{j=1,\ldots,s}$, (1) is minimized according to

$$\hat\gamma(i) = \arg\min_j\left(\ln\hat\sigma_j^2 + \frac{y_i - x_i'\hat\beta_j}{\hat\sigma_j^2}\right). \qquad (2)$$

For given $\hat\gamma$, (1) is the sum of the usual log-likelihood functions for homogenous linear regressions within each cluster. Therefore, it is minimized by the LS-estimator $\hat\beta_j$ from the points (x_i,y_i) with $\hat\gamma(i)=j$ and

$$\hat\sigma_j^2 := \frac{\sum_{\hat\gamma(i)=j}(y_i-\hat\beta_j' x_i)^2}{\hat n_j},\quad \hat n_j := \sum_{i=1}^{n} 1(\hat\gamma(i)=j),\ j=1,\ldots,s. \qquad (3)$$

That is, $\ln\hat L_n$ is monotonely increased if the steps (2) and (3) are carried out alternately. This algorithm leads to a local maximum in finitely many steps since there are only finitely many choices for $\hat\gamma$. In my experience, this is noticeably the fastest algorithm discussed in this paper. Under $\sigma_1^2 = \ldots = \sigma_s^2$, the procedure is equivalent to the least squares algorithm of Bock (1969).

It can be shown that FRFP-ML leads to inconsistent parameter estimators (see Marriott (1975) for the location case). This does not matter in practice if the clusters are well separated, but causes serious problems otherwise. Like FRM-ML, FRFP-ML needs some lower bound on the error variance parameters since otherwise $\ln\hat L_n$ would be unbounded.

The approaches for the estimation of s discussed in section 2 are not reasonable here because the number of parameters $\gamma(i)$ increases with n and their

value range increases with s. The following modification of the SC worked very well in simulations (Hennig (1997b)):

$$\hat{s} := \arg\max_s \ln \hat{L}_n(s) - \frac{\ln n}{2} k(s) - 0.7sn, \quad k(s) := (p+2)s, \tag{4}$$

$k(s)$ denoting the number of regression- and scale parameters.

4 Random regressors, mixture model

Random regressors have the advantage that the observations can be treated as i.i.d. The **random regressors mixture model** (RRM) has the following form: $(x_i, y_i) \in \{1\} \times I\!R^p \times I\!R,\ i \in I$, are distributed i.i.d. according to

$$\mathcal{L}(x,y) = \sum_{j=1}^{s} \epsilon_j F_{(G_j,\beta_j,\sigma_j^2)},$$

$$\sum_{j=1}^{s} \epsilon_j = 1,\ G_1, \ldots, G_s \in \mathcal{G},$$

that is, $\mathcal{L}(x) = G_j$ within cluster j. Suitable choices for $G_j,\ j = 1, \ldots, s$, enable us to model every kind of assignment (in-)dependence. Usually the G_j are not of interest, but unknown. For performing ML-estimation, there need to be a parametric specification of $\mathcal{G}$. This will not be discussed here. A more general approach is presented instead. The RRM is a special case of the **contamination model** (CM) (choose $F^* = \sum_{j=2}^{s} \frac{\epsilon_j}{\epsilon} F_{(G_j,\beta_j,\sigma_j^2)},\ (G,\beta,\sigma^2) = (G_1,\beta_1,\sigma_1^2)$ below):

$$\mathcal{L}(x,y) = (1-\epsilon)F_{(G,\beta,\sigma^2)} + \epsilon F^*,\ 0 \leq \epsilon < 1,\ G \in \mathcal{G}. \tag{5}$$

There is some basic difference between the CM and the former models. The parameters (G,β,σ^2) are clearly not unique in (5) since they can correspond to (G_j,β_j,σ_j^2) of the RRM for each j. Further, if F^* is not assumed to be of a mixture type, the CM allows for outliers, i.e. points in the data, which do not belong to any regression population. In robust statistics, the CM with $\epsilon < \frac{1}{2}$ is a standard tool to describe the occurence of outliers.

A method to analize the CM should find possible choices for (β,σ^2) (G is treated as nuisance) and therefore needs no specification of some number of clusters.

This goal can be achieved by means of **Fixed Point Clustering** (Hennig (1997a,b)). The idea of this approach is that a data subset, which contains no outliers, can be viewn as homogenous. If at the same time all other points of the dataset are outliers with respect to the subset, then the subset is separated from the rest and can be considered as a cluster.

For an indicator vector $g \in \{0,1\}^n$ define $\mathbf{Z}(g) := (x_i', y_i)_{g_i=1}$.

Definition: $\mathbf{Z}(g)$ is called *Fixed Point Cluster* (FPC) w.r.t. $\mathbf{Z}$, iff g is a fixed point of

$$f: \{0,1\}^n \mapsto \{0,1\}^n,$$
$$f_i(g) = 1\left[(y_i - x_i'\hat{\beta}(\mathbf{Z}(g)))^2 \leq c\hat{\sigma}^2(\mathbf{Z}(g))\right]$$

with some prechosen constant c (e.g. $c = 10$). $\hat{\beta}(\mathbf{Z}(g))$ and $\hat{\sigma}^2(\mathbf{Z}(g))$ are regression parameter and error variance estimators based only on the data subset $\mathbf{Z}(g)$, e.g. the ML-estimators from (3).

The function f is a kind of outlier identifier (0 for outliers) and therefore an FPC $\mathbf{Z}(g)$ is exactly the set of non-outliers in $\mathbf{Z}$ with respect to $\mathbf{Z}(g)$.
FPCs can be computed with the usual fixed point algorithm ($g^{j+1} = f(g^j)$) which converges in finitely many steps (Hennig (1997b)). In order to find all relevant clusters, the algorithm must be started many times with various starting vectors g. A complete search is numerically impossible. However, this also holds for the other two methods unless one is satisfied with a local maximum of unknown quality of the log-likelihood function.
The FPC methodology does not force a partition of the dataset. Non-disjoint FPCs and points are possible, which do not belong to any FPC. According to that, FPCs are rather an exploratory tool than a parameter estimation procedure in the case of a valid partition or mixture model.

5 Iris data example and comparison

Fisher's Iris data (Fisher (1936)) consists of four measurements of three species of Iris plants. The measurements are sepal width (SW), sepal length (SL), petal width (PW) and petal length (PL). The species are Iris setosa (empty circles in figure 2a), Iris virginica (filled circles) and Iris versicolor (empty squares). Each species is represented by 50 points. Originally, the classification of the plants was no regression problem. The dataset is used for illustratory purposes here. Only the variables SW and PW are considered. PW is modelled as dependent of SW. The distinction in "regressor" and "dependent variable" is artificial. The methods use no information about the real partition. By eye, the setosa plants are clearly seperated from the other two species, while virginica and versicolor overlap. A linear regression relation between SW and PW seems to be appropriate within each of the species.
Using the SC for estimating the number of clusters, FRM-ML-estimation finds the four clusters shown in figure 2b. Three clusters correspond to the three species. FRM-ML is the only method which provides a rough distinction between the virginica and versicolor plants. The fourth cluster (crosses in figure 2b) is some kind of "garbage cluster". It contains some points which are not fitted good enough by one of the other three regression equations. Note that the deviation from assignment independence of the

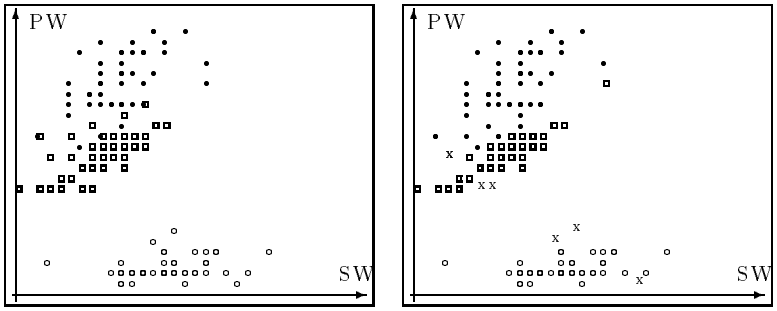


Figure 2: Iris data: a) original species - b) FRM-ML clusters with SC

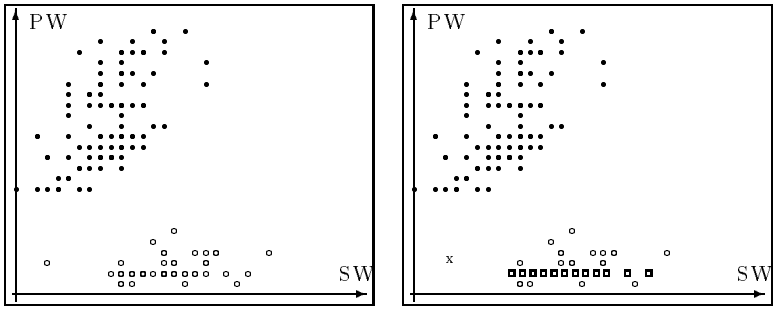


Figure 3: a) FRFP-ML clusters - b) Fixed Point Clusters

four cluster solution seems to be lower than that of the original partition of the species. The AIC for estimating the number of cluster leads to five clusters by removing further points from the three large clusters and building a second garbage cluster.

By application of (4), the number of clusters is estimated as 2. Figure 3a shows the ML-classification using the FRFP. It corresponds to the most natural[2] eye-fit. The well separated setosa plants form a cluster, the other two species are put together.

With 150 randomly chosen starting vectors, four FPCs are found. The first contains the whole dataset. This happens usually and is an artefact of the method, which one has to know to interpret the results adequately. The second and third cluster correspond to the setosa plants and the rest of the data, respectively. The point labelled by a cross falls in the intersection of both clusters and is therefore indicated as special. The fourth cluster is labelled by empty squares and consists of 29 points from the setosa cluster, which lie exactly on a line because of the rounding of the data. The other methods are not able to find this constellation because of the lower bounds on the error variances.

After having noticed this result, one realizes that there are other groups of points, which lie exactly on a line, and which are not found by the random

[2]It is not clear, what "most natural" means, but this is the impression of the author.

search of Fixed Point Clustering since they are too small. The fourth FPC contains more than half of the setosa species[3] and is therefore a remarkable feature of the Iris data.
The results from the Iris data highlight the special characteristics of the three methods. The simulation study of Hennig (1997b) leads to similar conclusions:

FRM-ML-estimation is the best procedure if assignment independence holds and if the clusters are not well separated. At the Iris data, it can discriminate between virginica and versicolor. The stress is on regression and error variance paramter estimation.

FRFP-ML-estimation is the best procedure under most kinds of assignment dependence to find well separated clusters if there is a clear partition of the dataset. At the Iris data, the procedure finds the visually clearest constellation. The stress is on point classification.

Fixed Point Clustering is the best procedure to find well separated clusters if outliers or identifiability problems (Hennig (1996)) exist. Its stress is on exploratory purposes. By means of Fixed Point Clustering, we discovered that a large part of the setosa cluster lies exactly on a line was made .

References

BOCK, H.H. (1969): The equivalence of two extremal problems and its application to the iterative classification of multivariate data. Lecture note, Mathematisches Forschungsinstitut Oberwolfach.

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

FISHER, R. A. (1936): The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7, 179-184.*

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

HENNIG, C. (1997a): Fixed Point Clusters and their Relation to Stochastic Models. In: Klar, R. and Opitz, O. (Eds.): *Classification and Knowledge Organization*, Springer, Berlin, 20-28.

HENNIG, C. (1997b): Datenanalyse mit Modellen für Cluster linearer Regression. Dissertation, Institut für Mathematsche Stochastik, Universität Hamburg.

KRISHNAIAH, P. R., MIAO, B. Q. (1988): Review about estimation of change points. In: Krishnaiah, P. R., Rao, P. C. (Eds.): *Handbook of Statistics, Vol. 7*, Elsevier, Amsterdam.

[3]One cannot see 29 squares because some of the points are identical.

MARRIOTT, F. H. C. (1975): Separating mixtures of normal distributions. *Biometrics 31, 767-769.*

QUANDT, R. E., RAMSEY, J. B. (1978): Estmating mixtures of Normal distributions and switching regressions. *Journal of the American Statistical Association 73, 730-752.*

SPAETH, H. (1979): Clusterwise linear regression. *Computing 22, 367-373.*