Regression Fixed Point Clusters: Motivation, Consistency and Simulations

Christian Hennig - Fachbereich Mathematik -
Universität Hamburg

February 20, 2000

Abstract

In this paper, the theoretical foundation of Least Squares-Fixed Point Clusters for clusterwise linear regression is given in full detail, as well as a discussion of the computation and application of the approach and a comparison with other cluster analysis methods based on stochastic models. Fixed Point Clustering is based on iteratively reweighted estimation with zero weight for all outliers. A Fixed Point Cluster is defined as a data subset that is exactly the set of non-outliers w.r.t. its own parameter estimators. Consistency results are given for certain mixture models of interest in cluster analysis. Convergence of a fixed point algorithm is shown and the implementation is discussed in detail. Simulations and the application to a real dataset show that Fixed Point Clustering has advantages over maximum likelihood methods to detect well separated homogeneous subpopulations in the presence of deviations from the usual assumptions of model based cluster analysis.

1 Introduction

Cluster analysis is related to the concept of outliers. If a part of a dataset forms a well separated cluster, this means that the other points of the dataset appear outlying with respect to the cluster. It may be interpreted synonymously that the cluster is homogeneous and that it does not contain any outlier. The idea of Fixed Point Clusters (FPCs) is to formalize a cluster as a data subset that does not contain any outlier and with respect to which all other data points are outliers. It is rooted in robust statistics as explained in Section 2.

The concept is applied to clusterwise linear regression in this paper, that is, a relation

$$y = x'\beta + u,\ E(u) = 0,$$

between dependent variable y and independent variable $x \in I\!R^p \times \{1\}$ (β_{p+1} denoting the intercept parameter) should be adequate for a single cluster. Figure 1.1 shows data from the Old Faithful Geyser in the Yellowstone National Park, collected in August 1985. The duration of an eruption of the geyser is modeled here as dependent on the waiting time since the previous eruption. Besides other features, which are discussed in more detail in Section 9, one can recognize roughly two groups of linear dependence

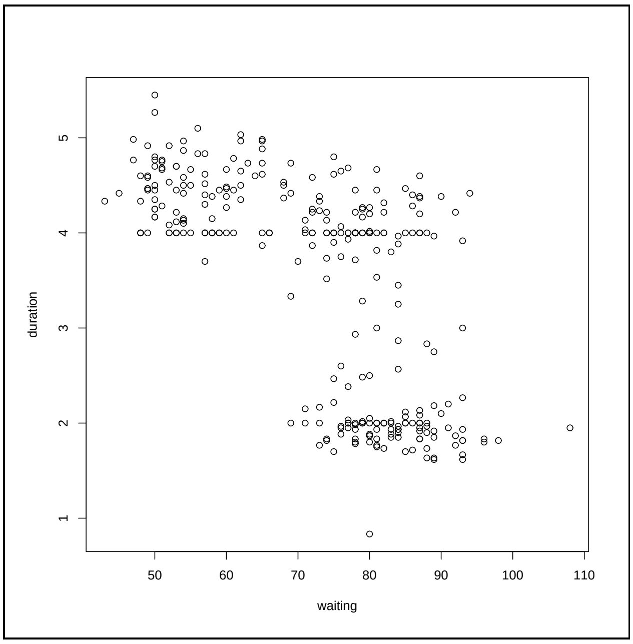

Figure 1.1: Old Faithful Geyser data.

between "waiting" and "duration", corresponding to the eruptions with lower and higher duration, the latter group with a moderately decreasing tendency for increasing waiting times. The data was taken from Azzalini and Bowman (1990). The aim of clusterwise linear regression is to find such kind of heterogeneity. Further applications of clusterwise linear regression appear in e.g. in biology (Hosmer 1974) and market segmentation (DeSarbo and Cron 1988).

For the sake of simplicity I discuss the one dimensional location clustering problem (i.e., linear regression without slope, $p = 0$) to motivate the FPC idea in Section 2. Least Squares-FPCs for the linear regression setup are defined in Section 3. Section 4 discusses generalizations of the approach. Section 3.2 introduces a convergent algorithm to find FPCs. Conditions for the consistency of Least Squares FPCs for theoretical FPCs are given in Section 5. Theoretical FPCs are calculated for certain mixture distributions and the consistency conditions are checked in Section 6. Section 7 discusses a reasonable implementation of the method along with the choice of all required constants.

The literature on clusterwise linear regression concentrates mainly on least squares and maximum likelihood methods for mixture and partition models (with exception of the paper of Morgenthaler, 1990, see Section 2). An overview is given by Hennig (1999). FPC analysis is compared to the maximum likelihood method of DeSarbo and Cron (1988) by means of simulations in Section 8 and by application to the Old Faithful data in Section 9. Linear regression clusters with normal distributed independent variable may be estimated as well by methods for normal mixture distributions. The procedure of DasGupta and Raftery (1998) concentrates on linearly shaped clusters and allows for noise modeled by a Poisson process mixture component. It was included in the comparisons as well.

After a short conclusion has been given in Section 10, all lemmas and theorems are proven in Section 11.

Here is some notation. "$\|x\|$" denotes the Euklidean norm of $x \in I\!R^k$. $\mathbf{I}_p$ denotes the $p \times p$-unit matrix. For $\epsilon > 0$, $B_\epsilon(x)$ is the closed ϵ-ball around x w.r.t. the Euklidean metric. For a given probability distribution P, $k \in I\!N \cup \{\infty\}$, P^k denotes the k-fold independent product. P^∞ is used as parent distribution for i.i.d. random variables $x_1, \ldots, x_n$ with $\mathcal{L}(x_1) = P$, which means that x_1 is distributed according to P. P_k denotes the empirical distribution according to $(x_1, \ldots, x_k)$ where $\mathcal{L}(x_1, \ldots, x_k) = P^k$. I write Pf for $\int f dP$. $1[(x, y) \in B]$ denotes the indicator function of the set B.

2 Clusters, outliers, M-estimators and fixed points

The link between outlier identification, robust statistics and cluster analysis is mentioned first by Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 46), to my knowledge. Robust statistics often deals with the location of a large homogeneous "main part" of the data in presence of outliers, which may be produced by mechanisms different from the rest, and which should not largely affect the estimation of the main part. Cluster analysis more generally aims to locate any homogeneous part of the data. The recognition of such a part should not be strongly affected by changes in distant parts of the data. This demand is violated by many CA methods, in particular by partitioning methods such as k-means (see Garcia-Escudero and Gordaliza, 1999). If there is a clear separation between main part and outliers, the main part can be regarded as the largest cluster, and robust statistics may serve to find it. But it can also point to the other ones, as explained in the following.

Imagine a one-dimensional dataset $(x_1, \ldots, x_n)$, $n = 30$, with 20 observations from $\mathcal{N}(0, 1)$ (avoiding the extreme tail areas), 5 observations from $\mathcal{N}(10, 1)$, and 5 observations from $\mathcal{N}(30, 1)$, i.e., three strongly separated clusters. M-estimators T_ρ of location (see e.g. Huber 1981) are defined by

$$\sum_{i=1}^{n} \rho\left(\frac{x_i - T_\rho}{s}\right) \overset{!}{=} \min \tag{2.1}$$

with suitable chosen loss function ρ and scale $s > 0$, or alternatively by

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - T_\rho}{s}\right) \overset{!}{=} 0, \tag{2.2}$$

where $\psi = \rho'$ (possibly piecewise). A solution of (2.2) is a fixed point of

$$f(t) := \frac{\sum_{i=1}^{n} w((x_i - t)/s) x_i}{\sum_{i=1}^{n} w((x_i - t)/s)}, \quad w(y) := \frac{\psi(y)}{y}. \tag{2.3}$$

That is, T_ρ is a weighted mean, where the weights depend on T_ρ itself. It may be obtained by the ordinary fixed point algorithm under certain conditions (Huber, 1981, p. 146; in linear regression such algorithms are sometimes called "iteratively reweighted least squares", see e.g. Morgenthaler, 1990). $w((x_i - t)/s)$ gives the weight of x_i at the computation of t and may be interpreted as a measure of centrality (outlyingness, respectively) of the point x_i with respect to t.

For example, the median corresponds to $\rho(x) = x * 1[x > 0] - x * 1[x < 0]$ regardless of s. As many robust location estimators, it will appear close to 0 for the data above, but positively biased (if interpreted as estimator for the data from $\mathcal{N}(0,1)$)) because of the asymmetrical contamination in positive direction.

The bias may be avoided by the so called "redescending M-estimators", which are M-estimators with $\rho(y)$ constant for large absolute values of y, and therefore $\psi(y) = w(y) = 0$. Such points do not have any weight at the computation of T_ρ, as desired for outliers. If s is chosen small enough, such an estimator estimates the center of $\mathcal{N}(0,1)$ unaffected by any point from the smaller populations. Furthermore it remains a solution of (2.3) under addition or deletion of outliers in the sense of this definition, i.e., of points with $w((x-t_0)/s) = 0$. But a solution of (2.2), (2.3), respectively, is usually not unique for redescending M-estimators. If s is chosen such that $w((x-t)/s) = 0$ holds for $|x-t| > 4$, say, there will be solutions estimating the centers of $\mathcal{N}(10,1)$ and $\mathcal{N}(30,1)$ as well, since the "window" of points with positive weight w around the center of each of the three clusters will only contain points from the same cluster. This leads to the thought that the solutions of (2.3) for redescending M-estimators might be used to locate an unknown number of clusters stably in the presence of outliers.

The main problem is the choice of s. In robust statistics one often uses a preliminary robust estimate of scale, for example the MAD. But such an estimate depends on at least half of the points. That is, if the largest cluster contains fewer than half of the points, s depends on points of at least two clusters and gets too large for a single cluster. Furthermore the clusters may have differing scales. If $\mathcal{N}(30,1)$ would be replaced by $\mathcal{N}(30,6)$, a weight window adjusted to variance 1 may capture only few points of this component, while working with variance 6 may destroy the separation between the other two populations.

The idea of FPC analysis is to define the location (regression parameters, respectively) and scale estimators jointly via a fixed point condition using only the corresponding non-outliers, so that both parameters are adapted to the local cluster. Such parameter estimators can no longer be described as minima of some global criterion like (2.1), since there is no natural quality ordering among them. The weights will be chosen so that they can only take the values 0 (outlier) and 1 (non-outlier). That is, a solution of (2.3) is characterized as corresponding to a subpopulation (defined by the weights for all points) that is exactly the set of non-outliers w.r.t. its own parameter estimators. A generalization to continuous choices of w, leading to fuzzy clusterings, is possible.

The resulting estimators fall in the class of simultaneous M-estimators of location and scale as defined by Huber (1980, p. 136), but the theory given there does exclude redescending ψ-functions. Morgenthaler (1990) to my knowledge was the first author to investigate the use of redescending M-estimators for locating different subpopulations. He discussed the choice of s in a linear regression setup based on the MAD of residuals of the LS-estimator as well as using a decreasing sequence of values for s, but he did not treat clusters with differing scales.

Alternative suggestions for the use of robust techniques in cluster analysis - not generalized to linear regression clusters up to now - were made by Davies (1988) and Cuesta-Albertos, Gordaliza and Matran (1997).

3 Fixed Point Clusters in Linear Regression

3.1 Definition for datasets

Let $\mathbf{Z} := \mathbf{Z}_n := (\mathbf{X}, y) := ((x_1', y_1), \ldots, (x_n', y_n))'$, where $x_i \in I\!R^p \times \{1\}$, $y_i \in I\!R$, $i = 1, \ldots, n$, be a regression dataset. For a given indicator (weight) vector $w \in \{0,1\}^n$ let $\mathbf{Z}(w) = (\mathbf{X}(w), y(w))$ be the dataset consisting only of the points (x_i', y_i) with $w_i = 1$. $n(w)$ is the number of points indicated by w. For FPC analysis in the regression setup, particular weight vectors are of interest. They indicate the points lying close (in terms of a variance parameter σ^2) to the regression hyperplane defined by a parameter β:

$$w_{\mathbf{Z},\beta,\sigma^2} := \left(1[(y_i - x_i'\beta)^2 \leq c\sigma^2]\right)_{i=1,\ldots,n}.$$

An FPC is defined as a data subset defined by some weight vector w indicating the non-outliers w.r.t. to the LS-estimator $\hat\beta(\mathbf{Z}(w))$ weighted by w itself. Outlyingness is measured by means of the weighted error variance estimator $\hat\sigma(\mathbf{Z}(w))$, i.e., parameter estimators satisfying a fixed point condition analogously to (2.3). Some tuning constant $c > 1$ has to be chosen to define the tolerance of the outlier classification. The choice of c is discussed in Section 7.1.

Definition 3.1 *An indicator vector* $w_{\mathbf{Z},\beta,\sigma^2} \in \{0,1\}^n$ *is called* **Least Squares-Fixed Point Cluster Vector** *(LS-FPCV) w.r.t* $\mathbf{Z}$ *(and the indicated points form an LS-FPC), iff* $(\beta, \sigma^2) \in I\!R^{p+1} \times I\!R_0^+$ *is a fixed point of*

$$f_{\mathbf{Z}}: \ (\beta, \sigma^2) \mapsto \left(\hat\beta\left[\mathbf{Z}(w_{\mathbf{Z},\beta,\sigma^2})\right], \hat\sigma^2\left[\mathbf{Z}(w_{\mathbf{Z},\beta,\sigma^2})\right]\right),$$

$$\textit{where } \hat\beta(\mathbf{Z}(w)) := (\mathbf{X}(w)'\mathbf{X}(w))^{-1}\mathbf{X}(w)'y(w),$$

$$\hat\sigma^2(\mathbf{Z}(w)) := \tfrac{1}{n(w)-p-1}\sum_{i=1}^n w_i\left(y_i - x_i'\hat\beta(\mathbf{Z}(w))\right)^2.$$

In case of the non-existence of $(\mathbf{X}(w)'\mathbf{X}(w))^{-1}$, $f_{\mathbf{Z}}(\beta, \sigma^2) := (\beta, \infty)$.

For example, consider the points indicated by triangles in Figure 3.1. They are indicated by the weight vector $w = w_{\mathbf{Z},\beta,\sigma^2}$ where β corresponds to the solid line and the dotted lines show $x'\beta \pm \sqrt{c\sigma^2}$. They form an LS-FPC for $c = 6.635$, since one finds $(\hat\beta(\mathbf{Z}(w)), \hat\sigma^2(\mathbf{Z}(w)) = (\beta, \sigma^2)$.

Consider on the other hand the squares with values of “duration” between 2 and 4. If the LS-regression line is estimated for this data subset, their error variance is so large that some of the circles and some of the triangles would get inside the corresponding strip. This would make the error variance of the resulting data subset even larger and it would also change the regression line, so that the fixed point condition is not fulfilled and this data subset is not separated enough from the rest to form an LS-FPC. The full result for the Geyser data is discussed in Section 9.

Note that FPCs may intersect or include each other. In particular, all subsets $\mathbf{Z}(w_{\mathbf{Z},\beta,\sigma^2})$ with $\sigma^2 = 0$ and non-collinear covariate points form LS-FPCs. It will be

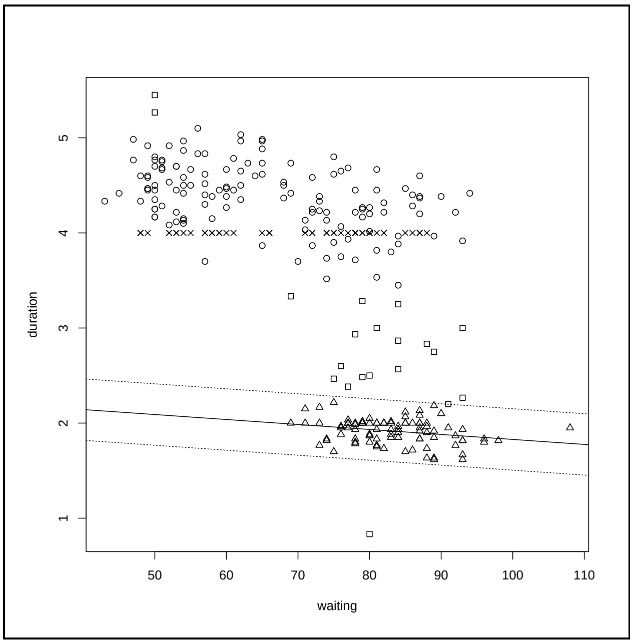

Figure 3.1: Old Faithful Geyser data with LS-FPCs, $c = 6.635$. The points indicated by crosses form an FPC as well as the triangles. A further FPC consists of the circles together with the crosses.

discussed later how to avoid trivial meaningless LS-FPCs in the output of the procedure.

Since the FPC-property of a subset does only depend on the points inside the strip defined by its parameter estimators, the deletion of any of the points outside the three shown FPCs (i.e., the points denoted by squares), or the addition of such points, would not change the FPC-property of any of these clusters.

3.2 A fixed point algorithm for LS-Fixed Point Cluster Vectors

It is practically impossible to check the FPC-property of every subset of a dataset, except if it is very small. But LS-FPCVs can be found by means of the usual fixed point algorithm. Its convergence is shown as Theorem 3.2. The use of the algorithm for an implementation of FPC analysis is described in Section 7. The convergence result will be needed for the consistency theory of Section 5 as well.

Fixed point algorithm (FPA): Choose $w^0 \in \{0,1\}^n$ with $n(w^0) > p+1,\ k = 0$.

Step 1: Compute $\hat\beta(\mathbf{Z}(w^k)), \hat\sigma^2(\mathbf{Z}(w^k))$.

Step 2: $w_i^{k+1} = w_{\mathbf{Z}i}(w^k) = 1((y_i - x_i'\hat\beta(\mathbf{Z}(w^k)))^2 \le c\sigma^2(\mathbf{Z}(w^k))),\ i = 1,\ldots,n.$

Step 3: End if $w^k = w^{k+1}$, else $k = k+1$, step 1.

Theorem 3.2 (Convergence) *Let $c > 1$. If $(\mathbf{X}(w)'\mathbf{X}(w))^{-1}$ exists for all $w \in \{0,1\}^n$ with $n(w) > p+1$, then for some $k < \infty$: $w^k = w_{\mathbf{Z}}(w^k)$, i.e., the FPA converges in finitely many steps.*

3.3 Definition for distributions

In order to investigate the statistical properties of LS-FPC analysis, I define a distribution version of LS-FPCs. Let P denote a distribution on $\mathbb{R}^p \times \{1\} \times \mathbb{R}$, i.e., a distribution for regression data points (x', y) as above. LS-FPCs of a distribution should consist of all points of appropriate strips around regression hyperplanes where the distribution is "regression cluster-shaped". They are indicated by weight functions of the form

$$w_{\beta,\sigma^2}(x,y) := 1[(y - x'\beta)^2 \le c\sigma^2].$$

For some measurable indicator function w let P_w denote the conditional distribution of P under $\{w = 1\}$, i.e., the restriction of P to the points indicated by w.

Theoretical LS-FPCs of distributions are defined by replacement of the regression and scale estimators by their corresponding functionals in Definition 3.1.

Definition 3.3 *An indicator function w_{β,σ^2}, $(\beta, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_0^+$ is called* **Least Squares-Fixed Point Cluster Indicator** *(LS-FPCI) w.r.t P, iff (β, σ^2) is a fixed point of*

$$f_P : \ (\beta,\sigma^2) \mapsto \left(\tilde\beta\left[P_{w_{\beta,\sigma^2}}\right], \tilde\sigma^2\left[P_{w_{\beta,\sigma^2}}\right]\right),$$
$$\textit{where } \tilde\beta(P_w) := \arg\min_\beta P_w(y - x'\beta)^2,$$
$$\tilde\sigma^2(P_w) := P_w(y - x'\tilde\beta(P_w))^2.$$

If $\arg\min_\beta P_w(y - x'\beta)^2$ is not defined uniquely, $f_P(\beta,\sigma^2) := (\beta, \infty)$.

$$\text{(This implies } Pw_{\beta,\sigma^2} > 0 \text{ for all LS-FPCIs.)} \qquad (3.1)$$

Under suitable conditions, LS-FPCVs turn out to be consistent estimators for LS-FPCIs in Section 5. That is, LS-FPCVs can be viewed as reasonable estimators of clusters of distributions, if the LS-FPCIs are such reasonable clusters. Some examples are discussed in Section 6.

The components of the functions $f_{\mathbf{Z}}$, f_P respectively, are written as follows from now on: $\beta_{\mathbf{Z}}(\beta,\sigma^2) := \hat\beta\left[\mathbf{Z}(w_{\mathbf{Z},\beta,\sigma^2})\right]$, $\sigma^2_{\mathbf{Z}}(\beta,\sigma^2) := \hat\sigma^2\left[\mathbf{Z}(w_{\mathbf{Z},\beta,\sigma^2})\right]$, $\beta_P(\beta,\sigma^2) := \tilde\beta\left[P_{w_{\beta,\sigma^2}}\right]$, $\sigma^2_P(\beta,\sigma^2) := \hat\sigma^2\left[P_{w_{\beta,\sigma^2}}\right]$.

Remark 3.4 *The regression equivariance properties of the LS- and variance estimator carry over to FPCVs and FPCIs, i.e., w_{β,σ^2} is FPCI w.r.t. P iff $w_{\beta,\sigma^2} \circ D = w_{(\mathbf{A}^{-1})'(a\beta+b),a^2\sigma^2}$ is FPCI w.r.t. P^D under linear transformations of the form*

$$D : \ \mathbb{R}^{p+2} \mapsto \mathbb{R}^{p+2}, \ (x,y) \mapsto (\mathbf{A}x, ay + x'b),$$

$\mathbf{A} \in I\!R^{(p+1)^2}$ invertible with last column $(0,\dots,0,1)$, $a \in I\!R \setminus \{0\}$, $b \in I\!R^{p+1}$. This holds analogously for FPCVs. The proof is straightforward, see Hennig(1997), Remarks 8.5 and 8.7.

4 Fixed Point Clusters - General

Here is a rougher description of FPCs: Consider a subset of the dataset. Decide for all points of the data subset, whether they are close to the subset (represented by its regression and scale parameter estimator) or lie out. If the non-outlying points are exactly the points of the subset, the subset forms an FPC. That is, the FPC property defines *homogeneity* (no outlier included) and *separateness* (all others are outliers) of a cluster in terms of outlier identification.

This description may be generalized to arbitrary clustering problems. Only an outlier identification rule is needed, that divides the whole dataset into outliers and non-outliers w.r.t. any given subset. The subsets, which do not contain any outlier, and w.r.t. which the whole rest of the data consists of outliers, are the FPCs. Applications to clustering problems apart from clusterwise linear regression are sketched out in Hennig (1998).

Appropriate outlier identifiers can be found as follows: Davies and Gather (1993) emphasized that a definition of the term "outlier" should rely on the idea of an underlying distribution of the homogeneous part of the data. They define "outlier regions" (ORs) as atypical regions of such "reference distributions". For example, in the linear regression case the class of distributions of the type $P_{\beta,\sigma^2,G}$ can be considered as the class of reference distribution for homogeneous data, where $P_{\beta,\sigma^2,G}$ is defined as the common distribution of (x, y) according to

$$y = x'\beta + u, \quad \mathcal{L}(u) = \mathcal{N}(0,\sigma^2), \quad \mathcal{L}(x) = G, \tag{4.1}$$

i.e., a model with random covariates, where

$$x \text{ and } u \text{ stochastically independent, } G\|x\|^2 < \infty, \quad (Gxx')^{-1} \text{ exists.} \tag{4.2}$$

Then,

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

$c := c(\alpha)$ being the $(1-\alpha)-$quantile of the χ^2_1-distribution, defines an α-OR in the sense of Davies and Gather, i.e., $A(\alpha, P_{\beta,\sigma^2,G}) = \alpha$ so that the points in the area of low density of the error distribution are defined as outliers. For example, $c(0.01) = 6.635$. In the definition of LS-FPCs, the parameters β and σ^2 are simply replaced by estimators.

That is, an OR is estimated on the basis of the data subset under consideration. It is treated as a set of non-outliers coming from a member of the family of reference distributions, and the whole dataset is treated as generated by a distribution of the form

$$(1-\epsilon)P_0 + \epsilon P^*, \quad 0 \le \epsilon < 1, \tag{4.3}$$

where P_0 is a reference distribution for homogeneous data, and P^* is arbitrary, but should be concentrated on $A(\alpha, P_0)$ with appropriate α. Models of the form (4.3) are called "contamination models". They are often used in robust statistics (e.g. Huber, 1981). Mixture models of the form

$$\sum_{i=1}^{k} \epsilon_i P_i, \quad \sum_{i=1}^{k} \epsilon_i = 1, \tag{4.4}$$

where $P_i,\ i = 1, \ldots, k$ are cluster reference distributions with distinct parameters, are more familiar in cluster analysis (e.g. DeSarbo and Cron, 1988). They are of the contamination type (4.3) as well, but they assume a particular structure for P^*, while FPC analysis needs P^* to be more clearly separated from the cluster generating distribution P_0, i.e., $P_{\beta,\sigma^2,G}$ in this paper. This is illustrated from a theoretical viewpoint in Section 6, while the simulation study of Section 8.2 shows the benefit of allowing a less restrictive P^* than mixture based CA methods.

From the viewpoint of robust outlier identification, it is questionable to estimate an OR by use of non-robust estimators like the LS-regression estimator. If a dataset (or a subset) contains outliers, they will affect such estimators. Davies and Gather (1993) discuss alternative outlier identifiers for the case $p = 0$ and show the superiority of identifiers based on robust estimators for the problem of finding large outliers in the presence of multiple outliers. Boscher (1992) suggests alternative outlier identifiers for the linear regression case. FPCs may be defined by the use of more general estimators of ORs. The most obvious idea is the replacement of regression and scale parameters $\hat{\beta}$ and $\hat{\sigma}^2$ of Definition 3.1, the corresponding functionals of Definition 3.3, respectively, by more robust alternatives.

I concentrate on the LS-version here for reasons of computational and theoretical simplicity. Its non-robustness may do less damage for the purposes of cluster analysis, since the aim is to find outlier-free data subsets, and there is no robustness problem for the data subsets which are *in fact* homogeneous and well separated. Recall from Section 2 that Definition 3.1 defines a *redescending* M-estimator as opposed to an LS-estimator for the whole dataset.

For heterogeneous data subsets, however, the estimated OR may get very large, so that there is usually an additional FPC corresponding to (almost) the whole dataset, even if the latter consists of some clearly separated clusters.

5 Consistency of LS-Fixed Point Cluster Vectors

The LS-FPCIs of the models are the "theoretical clusters" to be estimated by the LS-FPCVs. FPC analysis is intended to be a reasonable tool to analyze data from contamination models (4.3) where the component $P_0 = P_{\beta_1,\sigma_1,G}$ is well separated from P^*. Therefore it is desirable that the parameters of the LS-FPCVs are consistent for the parameters (β_1, σ_1^2) in some sense. The number of the FPCIs of distributions can vary as well as the number of FPCVs of datasets, even that of data drawn i.i.d. from the same distribution with $n \to \infty$. Here are aspects of the consistency of FPCs:

1. Do LS-FPCVs estimate LS-FPCIs consistently?

   (a) If P has an LS-FPCI, there should be a sequence of LS-FPCVs consistent for it (Theorem 5.3).

   (b) For large enough n, all LS-FPCVs should appear close to some LS-FPCI of P with large probability. (Corollary 5.2. There is no result relating the *number* of LS-FPCVs to that of LS-FPCIs.)

2. Do LS-FPCIs adequately reflect the structure of distributions of the contamination type (4.3)?

   (a) The contamination model should have an LS-FPCI belonging to $P_{\beta_1,\sigma_1,G}$, if it is well separated from P^* (Theorem 6.1, Corollary 6.2, Examples 6.4-6.8).

   (b) P^* may contain further parts of the type $P_{\beta,\sigma^2,G}$. Therefore it is not reasonable to expect that the LS-FPCI mentioned above would be the only one. But P should not have LS-FPCIs in areas where it does not give rise to any clustering of the data (Theorem 6.1, Examples 6.4-6.8).

   (c) If the LS-FPCIs correspond to well separated components of the type $P_{\beta,\sigma^2,G}$, they should fulfill the assumptions of the consistency results (Lemma 6.3, Examples 6.4-6.8).

Throughout this section P denotes a distribution on $I\!R^p \times \{1\} \times I\!R$, where $\mathcal{L}(\mathbf{Z}_n) = P^n, n \in I\!N$.

The basic result for the asymptotic existence of LS-FPCVs close to the LS-FPCIs, the non-existence elsewhere, respectively, is the uniform consistency of $f_{\mathbf{Z}_n}(\beta, \sigma^2)$ for $f_P(\beta, \sigma^2)$ for all (β, σ^2) belonging to some suitable set.

Let C be some compact subset of $I\!R^{p+1} \times I\!R_0^+$. Define

$$V(C) := \bigcup_{(\beta,\sigma^2)\in C} \{(y - x'\beta)^2 \leq c\sigma^2\}$$

as the union of all (x, y) belonging to one of the w_{β,σ^2}-stripes for $(\beta, \sigma^2) \in C$. $V(C)$ is closed and hence measurable as proven by Hennig (1997), Remark 13.10. Consistency of $f_{\mathbf{Z}_n}$ for f_P within C requires the following assumptions:

$$\forall(\beta, \sigma^2) \in C: \; P\{(y - x'\beta)^2 = c\sigma^2\} = 0, \tag{5.1}$$

$$\forall z \in I\!R^{p+1} \setminus \{0\}: \; P(\{x'z = 0\} \cap V(C)) = 0 \tag{5.2}$$

$$Py^2 1[(x, y) \in V(C)] < \infty, \; P\|x\|^2 1[(x, y) \in V(C)] < \infty, \tag{5.3}$$

$$\inf_{(\beta,\sigma^2)\in C} Pw_{\beta,\sigma^2} > 0. \tag{5.4}$$

The assumptions (5.1) and (5.2) are fulfilled if P is Lebesgue-dominated. Finiteness of Py^2 and $P\|x\|^2$ suffices for (5.3). (5.1) and the moment conditions (5.3) are needed to ensure the continuity of f_P. (5.2) prevents the covariate matrix from getting collinear. The assumption (5.4) together with (5.1) forces C to be bounded away from $\sigma^2 = 0$. Since C is compact, this suffices for (5.4) to hold if P has a non-vanishing Lebesgue-density. (5.4) is necessary since FPC analysis deals with arbitrary small subsets of the data, and increasing n does not prevent the occurrence of very small data subsets such that their local estimators of regression and error variance lie far from their theoretical values.

Lemma 5.1 (Uniform consistency of $f_{\mathbf{Z}_n}$ on C) *If (5.1)-(5.4) hold for some compact $C \subset I\!R^{p+1} \times I\!R_0^+$, then for all $\kappa > 0$*

$$P^\infty\{\exists n_0 > p+1\ \forall n > n_0, (\beta, \sigma^2) \in C: \ \left\| f_{\mathbf{Z}_n}(\beta, \sigma^2) - f_P(\beta, \sigma^2) \right\| < \kappa\} = 1$$

This means that for such C, which can be arbitrary large as long as it is compact and bounded away from $\sigma^2 = 0$, LS-FPCVs occur eventually only outside of C or where $f_P(\beta, \sigma^2)$ is very close to (β, σ^2):

Corollary 5.2 (Non-existence of LS-FPCVs) *Let $\kappa > 0$. Let C fulfill the assumptions of Lemma 5.1. Then for large enough n there exists P^∞-a.s. no LS-FPCV $w_{\mathbf{Z}_n,\beta,\sigma^2}$ with $(\beta, \sigma^2) \in C$ and $\|f_P(\beta, \sigma^2) - (\beta, \sigma^2)\| \geq \kappa$.*

The corollary follows directly from Lemma 5.1.

It is necessary to investigate f_P to assess the statistical meaning of this statement. The examples of Section 6 show how f_P may look like.

This will help to understand the meaning of the following assumption as well, which is additionally required to show the existence of consistent sequences of LS-FPCVs for LS-FPCIs. Suppose

$$\exists \text{ LS-FPCI } w_{\beta_0,\sigma_0^2} \text{ w.r.t. } P,\ \sigma_0^2 > 0. \tag{5.5}$$

(If there exists such LS-FPCI with $\sigma_0^2 = 0$, then $Pw_{\beta_0,0} > 0$ and for large enough n there are P^∞-a.s. enough points (x, y) with $(y - x'\beta_0)^2 = 0$, so that $w_{\mathbf{Z}_n,\beta_0,0}$ is an LS-FPCV. That is, in this case there exists a consistent sequence of LS-FPCVs.)

It will be assumed that $\exists \epsilon_0 > 0,\ 1 > \alpha \geq 0:$

$$\forall 0 < \epsilon \leq \epsilon_0: \ (\beta, \sigma^2) \in B_\epsilon(\beta_0, \sigma_0^2) \Rightarrow f_P(\beta, \sigma^2) \in B_{\alpha\epsilon}(\beta_0, \sigma_0^2). \tag{5.6}$$

This assumption is needed to force $f_{\mathbf{Z_n}}(\beta, \sigma^2)$, where (β, σ^2) is close to (β_0, σ_0^2), into shrinking neighborhoods of (β_0, σ_0^2). Let $C := B_{\epsilon_0}(\beta_0, \sigma_0^2)$. (5.6) follows immediately, if

$$\begin{gathered} f_P(C) \subseteq C,\ 1 > \alpha \geq 0: \ \forall (\beta_1, \sigma_1^2), (\beta_2, \sigma_2^2) \in C: \\ \|f_P(\beta_1, \sigma_1^2) - f_P(\beta_2, \sigma_2^2)\| \leq \alpha \|(\beta_1, \sigma_1^2) - (\beta_2, \sigma_2^2)\|, \end{gathered} \tag{5.7}$$

i.e., contractivity of f_P within C as needed for Banach's Fixed Point Theorem that guarantees the existence of a fixed point within C (but only for f_P, not for the non-continuous $f_{\mathbf{Z}_n}$). See Section 6 for a discussion of cases where this is fulfilled.

Theorem 5.3 (Consistency) *Assume (5.5), (5.6) and (5.1)-(5.3) for* $C = B_{\epsilon_0}(\beta_0, \sigma_0^2)$. *Then,*

$$P^\infty\{\forall n > p+1 \ \exists w_{\mathbf{Z}_n,\beta_n,\sigma_n^2} \text{ LS-FPCV w.r.t. } \mathbf{Z}_n : \lim_{n\to\infty}(\beta_n, \sigma_n^2) = (\beta_0, \sigma_0^2)\} = 1$$

Remark 5.4 *The conditions (5.6) and (5.7) in the given form are not invariant under data transformations of the form D of Remark 3.4. But the convergence statements of Lemma 5.1 and Theorem 5.3 remain fulfilled: With $D(x,y) = (\mathbf{A}x, ay + x'b)$ let P^D denote the distribution of $D(x,y)$ under $\mathcal{L}(x,y) = P$. Assume $a \neq 0$ and $\mathbf{A}^{-1}$ as existent. Let $\mathbf{B} := \begin{pmatrix} a(\mathbf{A}^{-1})' & 0 \\ 0 & a^2 \end{pmatrix}$. $\|z\|_{(\mathbf{B}^{-1})'\mathbf{B}^{-1}} := z`(\mathbf{B}^{-1})'\mathbf{B}^{-1}z$ defines a norm on $I\!R^{p+2}$. If the Euklidean norm in the conditions (5.6) and (5.7) (including the definition of B_ϵ) is replaced by $\| \bullet \|_{(\mathbf{B}^{-1})'\mathbf{B}^{-1}}$, then the conditions hold for f_{P^D} in an ϵ_0-neighborhood of $\mathbf{B}\binom{\beta_0}{\sigma_0^2}$, iff they hold for f_P in the original form. Since all norms on $I\!R^{p+2}$ metrize the same topology, the proofs of Lemma 5.1 and Theorem 5.3 can be adapted easily to the norm $\| \bullet \|_{(\mathbf{B}^{-1})'\mathbf{B}^{-1}}$.*

Remark 5.5 *The FPCs are defined by a particular outlier region here. But the main ideas to prove consistency might be used for some other definitions of FPCs as well:*

1. *Get convergence of the FPA by finding a statistics that is strictly decreased by the algorithm.*
2. *Check uniform consistency of $f_{\mathbf{Z}_n}$ for f_P.*
3. *Assume contractivity of f_P in some neighborhood of the fixed point (β_0, σ_0^2) to be reached consistently.*
4. *The FPA does not leave small neighborhoods of (β_0, σ_0^2) for large enough n because of 2. and 3. and there must be a fixed point of $f_{\mathbf{Z}_n}$ because of 1.*

6 LS-Fixed Point Cluster Indicators of some contamination and mixture models

This section starts with some theoretical results under relatively strong conditions. First, the existence and uniqueness of an LS-FPCI in the case $\epsilon = 0$ is shown. Corollary 6.2 and Lemma 6.3 (giving conditions for an LS-FPCI to fulfill the assumptions of Theorem 5.3) allow $\epsilon > 0$, but require P^* to give mass 0 to some neighborhood of $\{y = x'\beta_1\}$. This does clearly not hold for mixtures of more than one regression with normal distributed errors.

The theory is supplemented by numerical evaluations of f_P for some mixtures of one-dimensional normal distributions (i.e., $p = 0$). The results show that reasonable LS-FPCIs again exist, if the mixture components are separated well enough. They provide an illustration of the practical meaning of Corollary 5.2 and assumption (5.7).

In the case $\epsilon = 0$, $P = P_{\beta_1,\sigma_1^2,G}$ is a homogeneous linear regression distribution. Consequently there is only one LS-FPCI:

Theorem 6.1 (Homogeneous normal regression) *Let $c > 3$. $w_{\beta_1,k\sigma_1^2}$ is the unique LS-FPCI w.r.t. $P = P_{\beta_1,\sigma_1^2,G}$ where k is the unique zero of*

$$h(k) := 1 - k - \frac{2\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck}) - \Phi(-\sqrt{ck})}.$$

Theorem 6.1 shows that, given an LS-FPCV w w.r.t. some data set $\mathbf{Z}$, one can interpret $(\beta_{\mathbf{Z}}(w), \frac{\sigma_{\mathbf{Z}}^2(w)}{k})$ as a Fisher consistent estimator of the parameters (β_1, σ_1^2) of some homogeneous linear regression distribution, of some linear regression part of some contamination mixture respectively, as in Corollary 6.2. k depends on the pre-chosen c only, e.g. $c = 10$ yields $k = 0.9815$, $c = 6.635$ yields $k = 0.9001$. Essentially, $\sigma_{\mathbf{Z}}^2$ does not estimate the variance of the normal error distribution, but the variance of the truncated normal distribution of the non-outliers w.r.t. $\beta_{\mathbf{Z}}$ and itself.

The Theorem leads easily to the existence of a suitable LS-FPCI in the contamination model with $\epsilon > 0$, if there is no overlap between the LS-FPCI of the component $P_{\beta_1,\sigma_1^2,G}$ and P^*:

Corollary 6.2 *$w_{\beta_1,k\sigma_1^2}$ is LS-FPCI w.r.t. P defined by (4.3) with $P_0 = P_{\beta_1,\sigma_1^2,G}$, if*

$$P^* w_{\beta_1,k\sigma_1^2} = 0. \tag{6.1}$$

(Proven as Theorem 13.1 of Hennig (1997).)

The uniqueness of the LS-FPCI is lost in this case. This is reasonable since P^* may generate clusters elsewhere. (6.1) means that P^* has to generate outliers w.r.t. $P_{w_{\beta_1,k\sigma_1^2}}$ with probability 1. Note that (6.1) becomes weaker with smaller c. If one allows some overlap of P^* and $P_{w_{\beta_1,k\sigma_1^2}}$, proofs get rather complicated (Hennig (1997), but consider the examples at the end of this section). If the overlap or ϵ would be small enough and P^* would be continuous, continuity considerations lead again to the existence of some LS-FPCI w belonging to $P_{\beta_1,\sigma_1^2,G}$. Since the form of P^* is not specified, FPC analysis cannot distinguish between non-outliers w.r.t. this w generated by $P_{\beta_1,\sigma_1^2,G}$ and those generated by P^*. Therefore $(\tilde{\beta}(P_w), \tilde{\sigma}^2(P_w))$ would not be equal to $(\beta_1, k\sigma_1^2)$, but lie in some neighborhood.

Now conditions will be given, which ensure that the LS-FPCI of Corollary 6.2 fulfills the assumptions of the consistency theorem.

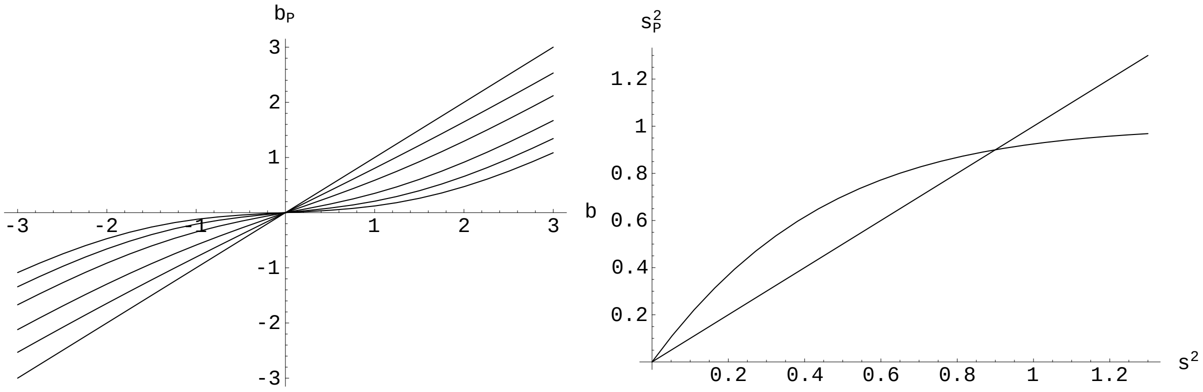

Figure 6.1: $P = \mathcal{N}_{0,1}$: a)x-axis: β, y-axis: β_P - b) x-axis: σ^2, y-axis: σ_P^2

Lemma 6.3 (Contamination mixture with normal regression) *Let* $c > 3$, $P = (1-\epsilon)P_{\beta_1,\sigma_1^2,G} + \epsilon P^*$, $1 > \epsilon \geq 0$, *where*

$$G\|x\|^3 < \infty,\ \forall a \neq 0:\ G\{a'x = 0\} = 0, \tag{6.2}$$

$$\exists \epsilon_1 > 0:\ P^*\left(V(B_{\epsilon_1}(\beta_1, k\sigma_1^2)\right) = 0, \tag{6.3}$$

where $k > 0$ *is defined as in Theorem 6.1 and fulfills furthermore*

$$k > 1 - \frac{2}{c-1}. \tag{6.4}$$

Then the assumptions of Theorem 5.3 are fulfilled with $\beta_0 = \beta_1$, $\sigma_0^2 = k\sigma_1^2$.

Remark: If $\epsilon P^*\{(y - x'\beta_1)^2 \leq c\sigma_1^2\} > 0$ but sufficiently small, condition (6.3) could be presumably replaced by assuming $P^* y^2 1[(y - x'\beta)^2 \leq \sigma^2]$ and $P^*\|x^2\|1[(y - x'\beta)^2 \leq \sigma^2]$ to be continuously differentiable w.r.t. (β, σ^2) for $(\beta, \sigma^2) \in B_{\epsilon_1}(\beta_1, k\sigma_1^2)$. Then (β_0, σ_0^2) would appear in some neighborhood of $(\beta_1, k\sigma_1^2)$.

(6.4) can be verified numerically for given c and holds for all values applied in this paper.

In the case $p = 0$, the function f_P is easy to evaluate and visualize numerically for normal mixtures[1]. Here are some examples. $c = 6.635$ was used, except if indicated.

Example 6.4 $P = \mathcal{N}_{0,1}$. *That is,* $\beta_1 = 0, \sigma_1^2 = 1$ *in Theorem 6.1. Figure 6.1a shows* $\beta_P(\beta, \sigma^2)$ *as a function of* β *for* $\sigma^2 = 0.1, 0.25, 0.5, 0.75, 1$. *(I always added the identity curve to make the fixed points visible.) Regardless of* σ^2, *the only fixed point of* β_P *is 0. Note that assumption (5.7) would follow for a one-dimensional function* f *if its increase in a neighborhood of a fixed point would be smaller than 1 (contractivity). This holds obviously for* β_P.

Figure 6.1b shows $\sigma_P^2(0, \sigma^2)$ *as a function of* σ^2. *There are two fixed points: 0 and 0.9001. Step 1 of Theorem 6.1 shows unique existence of a fixed point for* $\sigma^2 > 0$. $w_{0,0}$

[1] For $p > 0$, suitable parallel regression hyperplanes result in the same values

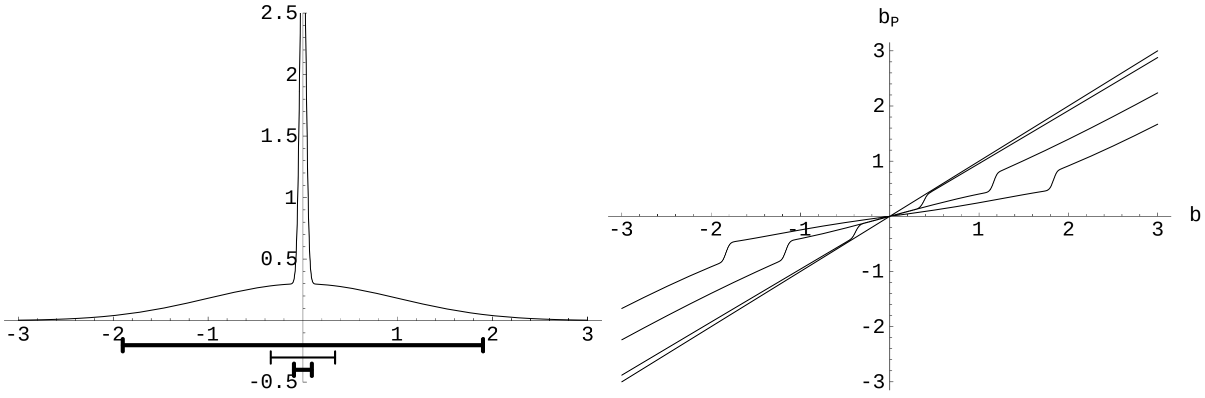

Figure 6.2: $P = \frac{3}{4}\mathcal{N}_{0,1} + \frac{1}{4}\mathcal{N}_{0,0.001}$: a) p.d.f. with LS-FPCs - b) x-axis: β, y-axis: β_P

cannot be an LS-FPCI unless $Pw_{0,0} > 0$.

Given some $\kappa > 0$, Corollary 5.2 almost surely excludes the existence of LS-FPCVs where $\|f_P(\beta,\sigma^2) - (\beta,\sigma^2)\| \geq \kappa$ for large enough n, (β,σ^2) from any compact set C. Eventually, all LS-FPCVs will fall in one of two classes:

- *They appear in a shrinking neighborhood of $(\beta,\sigma^2) = (0, 0.9815)$,*
- *or they contain only very few points since $\sigma^2 \approx 0$ or $|\beta|$ very large or both. (If σ^2 would be very large but not $|\beta|$, $\{g_{\beta,\sigma^2} = 1\}$ would contain almost all points. $\sigma^2_{\mathbf{Z}_n}(\beta,\sigma^2)$ would approximate σ_P of Figure 6.1b and would be smaller than a large σ^2.)*

The implementation of FPC analysis given in Section 7 excludes too small LS-FPCVs. In this example, such small FPCVs do not provide interesting information about the data, while (β_1,σ_1^2) can be consistently estimated by $(\beta_{\mathbf{Z}_n}, \sigma^2_{\mathbf{Z}_n}/0.9001)$.

Observe that σ_P^2 is contractive in the neighborhood of $\sigma^2 = 0.9815$, but not in the neighborhood of 0. A fixed point algorithm applied to f_P with starting values (β,σ^2) from some neighborhood of $(0,0)$ would never converge to $(0,0)$, since the increase of σ_P^2 is larger than 1 in a neighborhood of 0. Collatz (1966) calls such fixed points "repulsive". As mentioned before, there are LS-FPCVs $w_{\mathbf{Z}_n,\beta,\sigma^2}$ with $\sigma = 0$ even for large n, but if n is large enough, and β is not too far from 0, $\sigma_{\mathbf{Z}_n}(\beta,\sigma^2)$ comes close to $\sigma_P(0,\sigma^2)$ with high probability and therefore gets larger than σ^2. This means that $\hat\sigma^2(\mathbf{Z}_n(w^k))$ usually not decreases further during the FPA, if $\hat\sigma^2(\mathbf{Z}_n(w^1))$ has been already small.

Remark 6.5 *I have restricted the considerations about contractivity of f_P to the one-dimensional functions $\beta_P(\bullet,\sigma^2)$ for fixed σ^2 and $\sigma_P^2(\beta,\bullet)$ for fixed β. As shown in Lemma 6.3, f_P is contractive around $(0, 0.9001)$ for $P = \mathcal{N}(0,1)$, and it cannot be contractive if this is not even fulfilled for the one-dimensional projections. But for the further examples I only presume the contractivity of f_P in case of the contractivity of both $\beta_P(\bullet,\sigma^2)$ and $\sigma_P^2(\beta,\bullet)$ on the basis of numerical inspection and smoothness considerations.*

Example 6.6 $P = \frac{3}{4}\mathcal{N}_{0,1} + \frac{1}{4}\mathcal{N}_{0,0.001}$. *A cluster of "inliers" is added to the standard normal distribution here (Figure 6.2a. The LS-FPCs are drawn below the x-axis; they*

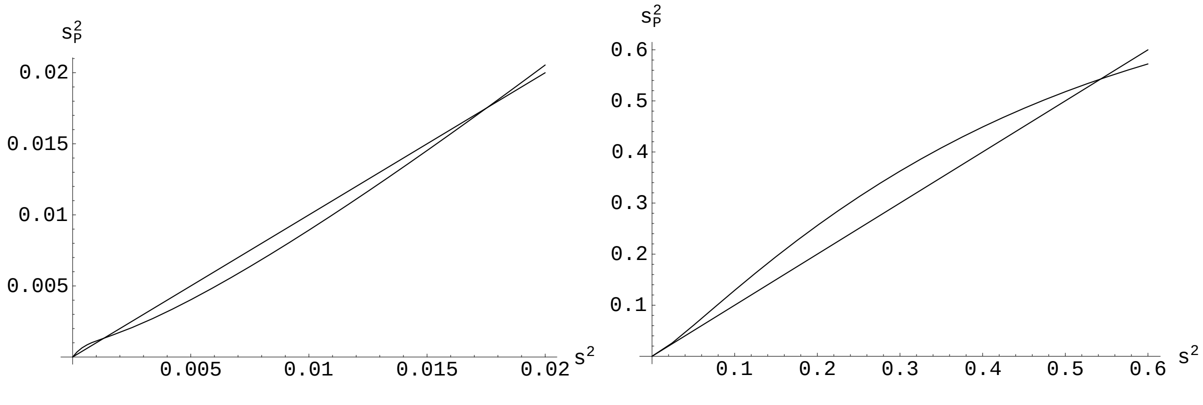

Figure 6.3: $P = \frac{3}{4}\mathcal{N}_{0,1} + \frac{1}{4}\mathcal{N}_{0,0.001}$: a), b) x-axis: σ^2, y-axis: σ_P^2

are indicated fat, if they fulfill (5.7)). Figure 6.2b shows again $\beta_P(\beta, \sigma^2)$ as a function of β for fixed $\sigma^2 = 0.02, 0.2, 0.5$. For $\sigma^2 = 0.001$ the curve would have the same shape but could not be distinguished optically from the identity curve. Again the only fixed point of β_P is 0, regardless of σ^2, and again β_P is contractive in a neighborhood of 0. There can be no LS-FPCI with $\beta_P \neq 0$, and there is no such mixture component as well.

Figure 6.3a and b show $\sigma_P^2(0, \sigma^2)$ as a function of different domains of σ^2. There are fixed points at 0 and about 0.0013, 0.0174 and 0.544, while there are only two components in the mixture[2]*. Again $w_{0,0}$ cannot be an LS-FPCI. The discussion of the previous example concerning $\sigma^2 \approx 0$ holds again*[3]*. The fixed point 0.0174 of σ_P^2 is repulsive for $\beta = 0$. This means that if n is so large that $\sigma_{\mathbf{Z}_n}$ provides a good approximation of σ_P, the FPA will move $\hat{\sigma}^2(\mathbf{Z}_n(g))$ away from 0.0174, so that an LS-FPCV is found only seldom in this area, but this cannot be excluded theoretically.*

σ_P is contractive around $(0, 0.0013)$ and $(0, 0.544)$, so that almost surely there exist consistent sequences of LS-FPCVs for these two LS-FPCIs (under the reservation of Remark 6.5). These are not the parameters of the mixture components. The reason is that FPC analysis does not enforce a partition of the data. The second mixture component $\mathcal{N}_{0,0.001}$ is only separated from those points of $\mathcal{N}_{0,1}$ that lie far enough from 0. But points from the first component $\mathcal{N}_{0,1}$, truncated to the non-outlier region of $\mathcal{N}_{0,0.001}$, have nevertheless a larger variance than points from $\mathcal{N}_{0,0.001}$. And there is no data analytic information to separate the first one from the second one.

However, there are only few points from the first component in this area and so the LS-FPCI with $(\beta, \sigma^2) \approx (0, 0.0013)$ corresponds well to the points from $\mathcal{N}_{0,0.001}$.

Contrarily, $(0, 0.544)$ would be a bad choice as an estimator of the parameters of the first component. Since almost all points from $\mathcal{N}_{0,0.001}$ are non-outliers with respect to $\mathcal{N}_{0,1}$, the standard normal component of this mixture is poorly separated from the rest (and the concept of "separateness of mixture components" as used in this paper turns out to be asymmetrical). The interpretation is that the relation between the width of the non-outlier region and variance of the truncated distribution does not seem "normal" to the FPC analysis procedure if $(\beta, \sigma^2) \approx (0, 1)$, but for $(0, 0.544)$, because of the "inliers" from the second mixture component[4]*.*

[2] For $\sigma^2 > 0.6$, there is no further fixed point.

[3] Note, however, that "≈ 0" means "remarkably smaller than 0.001" in this example.

[4] If the proportion of $\mathcal{N}(0, 0.001)$ would be increased to $\frac{1}{2}$, only one LS-FPCI would remain, namely

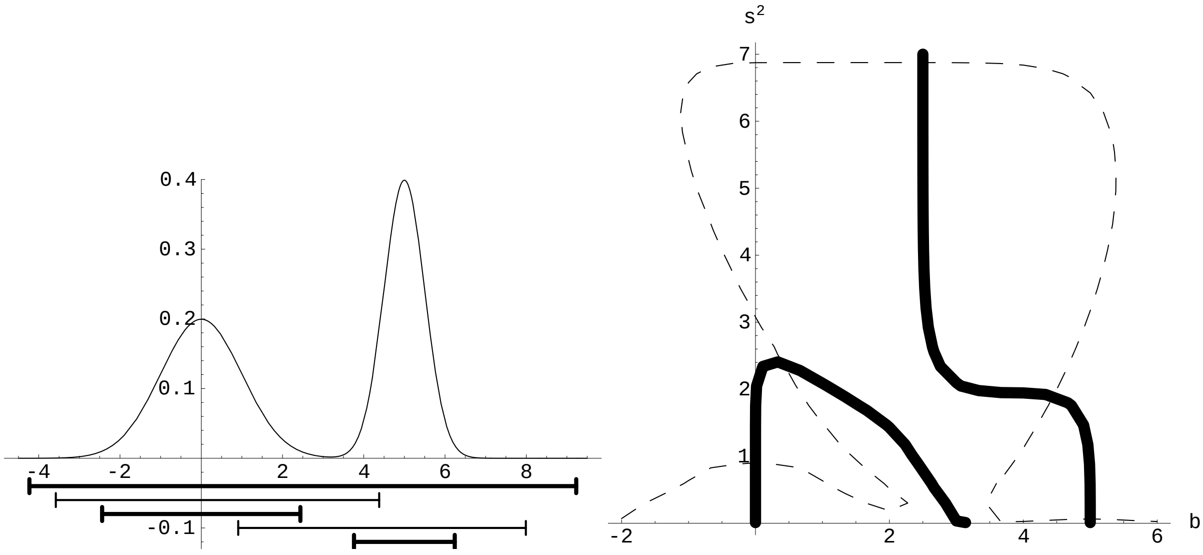

Figure 6.4: $P = \frac{1}{2}\mathcal{N}_{0,1} + \frac{1}{2}\mathcal{N}_{5,0.25}$: a) p.d.f. with FPCs - b)x-axis: β, y-axis: σ^2. Fat line: $\beta_P(\beta, \sigma^2) \overset{!}{=} \beta$, dotted line: $\sigma_P^2(\beta, \sigma^2) \overset{!}{=} \sigma^2$

If one would consider FPC analysis to be an estimation procedure for parameters of mixtures, it should be ruled out by this example. The example shows that this is not what the method essentially does. The components of mixture models cannot always be interpreted as "clusters". A mixture of two normals with only moderately different scales and means can hardly be distinguished from a homogenous population. Surprisingly large differences of the parameters are necessary for two normal distributions to be "separated", i.e.,to produce a large amount of outliers with respect to the other one (Wellmann and Gather 1999).

FPC analysis provides a proposal how to define the "clusters" of a normal mixture. According to the concept presented here, clusters are interpreted in terms of internal homogeneity and external separateness. The normal distribution serves to define "homogeneity" and "separateness". Internal homogeneity is measured by the relation between the width of the non-outlier region and variance of the truncated distribution. P truncated to $[-\sqrt{6.635 * 0.544}, \sqrt{6.635 * 0.544}]$ *is "more normal" in this sense than P restricted to the non-outlier region of* $\mathcal{N}_{0,1}$.

Example 6.7 $P = \frac{1}{2}\mathcal{N}_{0,1} + \frac{1}{2}\mathcal{N}_{5,0.25}$. *This is an example with two normal mixture components that generate clearly separated clusters (Figure 6.4a). The parameters of the LS-FPCIs should be close to the parameters of the mixture components. Figure 6.4b) shows the pairs* (β, σ^2), *for which* $\beta_P(\beta, \sigma^2) \overset{!}{=} \beta$ *(fat line),* $\sigma_P^2(\beta, \sigma^2) \overset{!}{=} \sigma^2$ *respectively (dotted line). The intersections of the fat and the dotted line mark the fixed points of* f_P. *Apart from* $\sigma^2 = 0$ *there are five of them: approximately* $(0, 0.9), (0.4, 2.39), (5, 0.23)$, $(4.45, 1.89)$, *and* $(2.5, 6.87)$. *Figure 6.5a shows* β_P *as a function of* β *for* $\sigma^2 = 0.25$ *(dotted),* $\sigma^2 = 1$ *(thick),* $\sigma^2 = 6.8$ *(very fat). In the latter case,* β_P *is nearly constant around* $\beta = 2.5$ *since* σ^2 *is so large that* $\{w_{\beta,\sigma^2} = 1\}$ *has almost probability 1.* β_P *is clearly*

that corresponding to $\mathcal{N}(0, 0.001)$

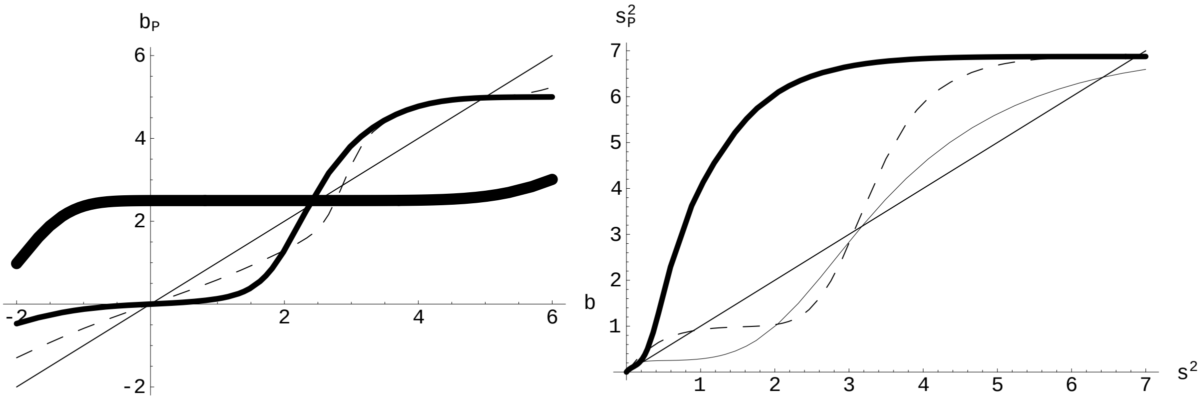

Figure 6.5: $P = \frac{1}{2}\mathcal{N}_{0,1} + \frac{1}{2}\mathcal{N}_{5,0.25}$: a) x-axis: β, y-axis: β_P - b) x-axis: σ^2, y-axis: σ^2_P

contractive here around 2.5 for $\sigma^2 = 6.8$*, as well as around 0 and 5 for the smaller values of* σ^2*. Figure 6.5b shows* σ^2_P *as a function of* σ^2 *for* $\beta = 0$ *(dotted),* $\beta = 5$ *(thick), and* $\beta = 2.5$ *(very fat). It turns out that* σ^2_P *is only contractive in the neighborhoods of the fixed points* $(\beta, \sigma^2) = (0, 0.9)$, $(5, 0.23)$ *corresponding to the two mixture components, and* $(2.5, 6.87)$ *corresponding to the non-outlier region of the whole dataset. FPC analysis does not take care if the distribution inside the non-outlier region of a cluster has a normal shape*[5]*. This is not the case for the LS-FPCI corresponding to the parameters* $(2.5, 6.87)$*. If an outlier identification is computed on the basis of points from distinct mixture components, the non-outlier region gets large. The corresponding LS-FPCI is only homogeneous in the sense that the points inside would "belong together" in some manner compared to any added gross outliers*[6]*.*

Observe further that in the examples up to now the contractivity assumption (5.7) held for the "more meaningful" LS-FPCIs in terms of mixture components, while other fixed points were repulsive.

Example 6.8 *The last example illustrates the meaning of the term "clearly separated". Figure 6.6a shows the p.d.f. of* $P = \frac{1}{2}\mathcal{N}_{0,1} + \frac{1}{2}\mathcal{N}_{d,0.25}$, $d = 2.5$*. The density is bimodal, but the amount of mass that cannot be clearly assigned to one or the other mixture component is not negligible. Apart from* $\sigma^2 = 0$ *there is only one further fixed point of* f_P *as can be seen from the plot of* $\beta_P(\beta, \sigma^2) \overset{!}{=} \beta$ *(fat line),* $\sigma^2_P(\beta, \sigma^2) \overset{!}{=} \sigma^2$ *respectively (dotted line) (Figure 6.6b). This fixed point corresponds to the non-outlier region of the whole dataset, as discussed in the previous example. If the distance d between the two mixture components increases, the fat line gets closer to the dotted line. At about* $d = 2.9$, f_P *starts to have a fixed point at about* $(d, 0.25)$ *(Figure 6.7a). At this stage, the second component is separated well enough from the first one, but not the first one from the second one. At about* $d = 3.7$*, an LS-FPCI for the first component appears (Figure 6.7b). The example illustrates that the fixed point concept of "clustering" requires more separation than the identification of "clusters" with mixture components or neighborhoods*

[5] Hennig (2000) discusses the distinction between more and less meaningful FPCIs on the basis of the Kolmogorov-distance to an LS-FPCI of a homogeneous normal population.

[6] This effect does not vanish if LS-functional/expectation and variance are replaced by robust functionals as suggested in Section 4

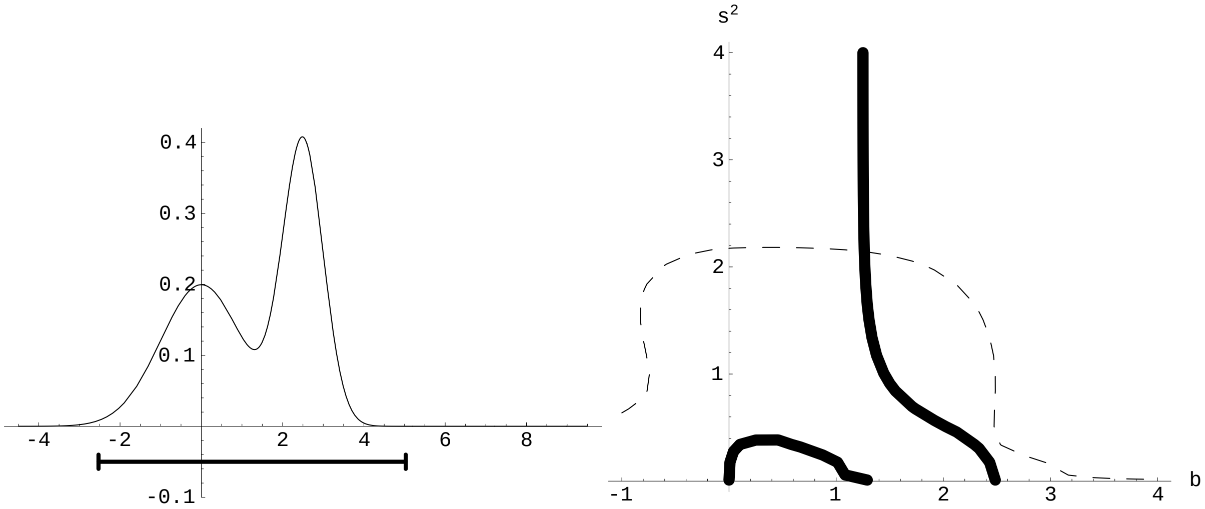

Figure 6.6: $P = \frac{1}{2}\mathcal{N}_{0,1} + \frac{1}{2}\mathcal{N}_{2.5,0.25}$: a) p.d.f. with LS-FPC - b)x-axis: β, y-axis: σ^2. Fat line: $\beta_P(\beta,\sigma^2) \overset{!}{=} \beta$, dotted line: $\sigma^2_P(\beta,\sigma^2) \overset{!}{=} \sigma^2$

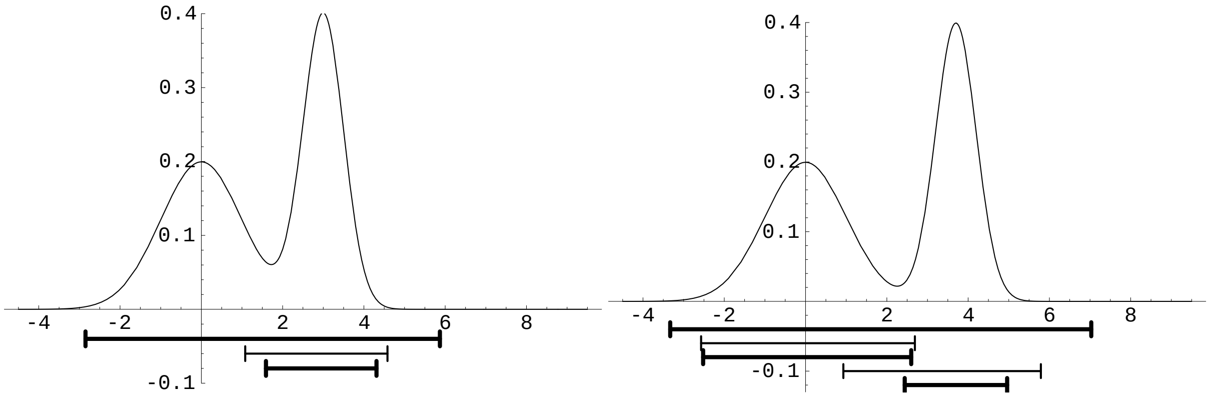

Figure 6.7: a) p.d.f of $P = \frac{1}{2}\mathcal{N}_{0,1} + \frac{1}{2}\mathcal{N}_{3,0.25}$. b) p.d.f of $P = \frac{1}{2}\mathcal{N}_{0,1} + \frac{1}{2}\mathcal{N}_{3.7,0.25}$. Both with LS-FPCs.

of modes. The degree of separateness needed for FPCs can be modified by the use of the tuning constant c. The larger c, the larger the region of non-outliers and the more separation between clusters is needed. For $c = 10$, there is only one LS-FPCI for d smaller than 3.7, where the LS-FPCI corresponding to $\mathcal{N}(d, 0.25)$ appears. For $c = 5$, this mixture component keeps its LS-FPCI down to $d = 2.4$.

Again it turns out that FPC analysis leads to essentially different findings than mixture estimation or mode seeking. It is subject to the applicator to decide if it corresponds more to his or her image of a cluster.

7 Implementation of the procedure

As mentioned before, a complete search for all LS-FPCVs of a given dataset is impossible. In this section an implementation of a procedure is described to find all "interesting" FPCVs with high probability. The term "interesting" is discussed below. The implementation involves the choice of some tuning constants, which is discussed in detail. Find a short description of the implementation in Section 7.4. The basic procedure is simple:

1. Choose the number of algorithm runs $i_{n,p}$ and the tuning constant c.

2. Repeat $i_{n,p}$ times: Generate a subset indicator w^0 with $n(w^0) = p + 2$ randomly[7] and apply the FPA until convergence. Store all found FPCVs, count the number of times that each FPCV has been found.

The choice of c and $i_{n,p}$ is discussed in the Sections 7.1 and 7.3.

Applying the basic procedure, one may observe that the number of found FPCs is often larger than one would like to interpret, unless n is very large or $i_{n,p}$ is so small that the result of the analysis depends strongly on chance. There are some reasons for that:

- Often there is more than one version of the same visible cluster in the data, since FPC analysis allows an arbitrarily large overlap between the clusters. Consider a single point that is neither a clear outlier, nor clearly consistent with the other points of a given FPC. If such a point is added to or removed from an FPC, the FPC property might remain fulfilled. It does not decide necessarily between these two versions of the cluster. Contrarily, partitioning methods base such a decision on assumptions about the nature of the rest of the data. Such assumptions are avoided in the FPC setup.

- Often there appear small subsets of a dataset that can be fitted almost exactly by a regression hyperplane. Since their separation from the remaining points is measured by their very small error variance, they can meet the FPCV definition. For example, all data subsets lying exactly on a regression line (including all subsets of $p+1$ points) lead to FPCVs with variance 0, as mentioned in Section 3.1. If $i_{n,p}$ is large enough, some algorithm runs lead to very small FPCVs by chance.

- If the dataset is sparse in $I\!R^{p+1}$, i.e., if n is small or p is large, it is not very difficult for subsets to be separated well enough from the rest of the data and thus to be FPCVs. This is in agreement with Rousseeuws (1994) words that "*my interpretation of the "curse of dimensionality" is that several structures can exist simultaneously in the same dataset.*" Partitioning methods and methods based on mixture models suffer from lots of local optima of the criterion function in this situation (usually hidden from the applicator), FPC analysis yields lots of FPCs, as can be seen in the simulations of Section 8.1.

- The consistency theory does not exclude the occurrence of FPCVs in the neighborhood of repulsive fixed points, which turned out to be of low interpretative value in the Examples 6.6-6.8. However, if n is large, then one can expect that the FPA converges to such FPCVs very seldom because of the repulsiveness of the corresponding fixed points w.r.t. f_P, which is well approximated eventually.

[7] The choice of $n(w^0)$ is justified in Section 7.3.

All these FPCs might give relevant information about the dataset, but the researcher might not want to work through more than 20 or even more than 200 FPCVs to find the most important features of the data. Therefore an applicable procedure needs a third step, which will be described precisely in the Sections 7.2 and 7.4:

3. Reduce the number of FPCs to be interpreted by

- defining groups of similar found FPCVs (*clusters of clusters*),
- discarding all FPCVs of groups that were found too seldom,
- choosing a representative FPCV for each of the remaining groups.

7.1 The tuning constant c

The tuning constant c defines the size of the distance that a point must have from the center of a normal distribution to call it an "outlier". The larger c, the more separated data subsets have to be to get FPCs. The choice of c determines the *definition* of the structures FPC analysis looks for, and not the *quality* of their estimation. As for the p-value in testing, there is no "optimal" or "correct" choice of c. The researcher has to decide about the degree of separation of clusters he is interested in.

Some statisticians do not like tuning constants because of the lack of "objectivity" of subjective choices. But it seems to me that such choices are required for all good statistics since careful model choices are always subjective decisions. Tuning constants are a way to make the subjectivity explicit.

Here are some theoretical considerations. Example 6.8 shows the degree of separation of clusters required by some choices of c asymptotically. For example, $c = 10$ and even $c = 6.635$ need a very strong separation to find FPCs corresponding to the existing mixture components. From a stochastic viewpoint, "clusters" are often associated with modes or mixture components, and this may lead to the belief that c must be chosen much smaller. But there are some arguments against that.

c is defined to be the $(1-\alpha)$-quantile of the χ^2-distribution in the definition of $A(\alpha, P_{\beta,\sigma^2,G})$. For $\alpha = 0.01, c = 6.635$ is obtained, i.e., 99% of normal distributed data (linear regression data with normal errors, respectively) are defined to be non-outliers. Davies and Gather (1993) define the tuning constants of their outlier identifiers in such a way that the probability exceeds 0.95 that none of n points from a homogeneous population falls into the corresponding outlier region. For $c = 6.635$, the corresponding n is 4. For $n > 68$, the probability for at least one regularly generated outlier is larger than 0.5. $c = 10$ corresponds to $\alpha = 0.00157$. This choice leads to a probability of approximately 0.95 that none of 33 i.i.d. observations from $P_0 \in \mathcal{P}_0$ falls into $A(\alpha, P_0)$.

For the clustering problem, which is not treated by Davies and Gather, it could be analogously desired that all points of a homogeneous population should belong to the corresponding FPC. This would be the case if none of them would fall into the *estimated* outlier region based on LS-estimator and estimated residual variance, which corresponds to $A(\alpha, P_{\beta,\sigma^2,G})$ only asymptotically. But for $p > 0$ the distribution of the residual of a point depends on its regressor value and cannot be utilized directly for the choice of c. A table for the case $p = 0$, $n \leq 20$ can be found in Barnett and Lewis (1994, Table XIIIb, p.485). For 20 i.i.d. normal distributed observations and a probability of 0.95 for absence of estimated outliers, $c = 7.344$ must be taken, compared to $c = 9.12$ using the method of Davies and Gather. For large n, both methods lead to very large (and asymptotically equal) choices of c.

For large n it is clearly very restrictive to demand that *no single* observation from a homogeneous population should be classified as an outlier. But one may consider the variance of the members of the corresponding LS-FPCI to measure its similarity with a non-truncated normal distribution. For $c = 6.635$, it is about 90% of the variance of the underlying normal distribution, for $c = 10$ it gets more than 98%, see the remark following the proof of Theorem 6.1.

According to the arguments given up to now, one would choose c increasing with n. For the aims of cluster analysis, however, the experiences from the simulations (see Sections 8.1 and 8.2) point in the opposite direction. A smaller c leads to a weaker requirement of separation and therefore to a larger number of FPCs. For small n and especially for large p there appear lots of meaningless FPCs. It would be reasonable to choose c such that the expected number of found FPCs for a homogeneous normal population does not clearly exceed 1. This leads to a very large c for small n and large p, while there seem to be no problems with $c = 6.635$ for e.g. $p = 1,\ n = 200$.

It is difficult to establish a theoretical foundation for the relation between $n,\ p$, and c. One may specify an asymptotic value for c, $c = 10$ or $c = 6.635$, say, and choose c_n according to the Table 8.2 in Section 8.1 as long as such c_n is not smaller than c. This way, $c_n = c$ for large enough n, and the results from the Section 5 remain valid for the concrete procedure [8].

7.2 A similarity measure for clusters and representative FPCVs

The number of found FPCVs can be reduced without loss of too much relevant information, if very similar FPCVs are interpreted as corresponding to the same "pattern" of the data. This can be done formally by defining groups (clusters) of FPCVs and by declaring only one "representative" FPCV of each group as "interesting". It is reasonable to define "similarity" between FPCVs by means of the number of common data points and not by means of their regression and error variance parameters to avoid sensitivity to the scaling of the parameters. A similarity measure between the indicator vectors v and w of subsets of a dataset is defined by relating the number of points of the intersection of the subsets to the sum of their sizes:

$$s_*(v,w) := \frac{2|\{i:\ v_i w_i = 1\}|}{|\{i:\ v_i = 1\}| + |\{i:\ w_i = 1\}|}, \tag{7.1}$$

so that $0 \leq s_*(v,w) \leq 1$, where $s_*(v,w) = 0$ iff v and w have no point in common, and $s_*(v,w) = 1$ iff $v = w$. To define a partition of the FPCVs, one can specify $0 < s_{cut} < 1$ so that v, w are interpreted as "similar" if $s_*(v,w) \geq s_{cut}$. The Single Linkage clusters of index s_{cut} are defined as the connectivity components of the graph with the FPCVs as vertices and edges between all pairs v, w where $s_*(v,w) \geq s_{cut}$. They can be computed by an algorithm described in the Section 7.4 as "Step 5". This seems to be the easiest method to get a reasonable partition based on similarities without assumptions about the

[8] If $c_n \searrow c$, but not constant for large enough n, the theory holds as well, but it is not obvious how to relate the interpretation of the results for n with $c_n > c$ to the examples of Section 6.

number of groups[9]. In the resulting partition, v and w are always joined if $s_*(v, w) \geq s_{cut}$. If $s_*(v, w) < s_{cut}$, v and w are sometimes joined, namely if there is another FPCV (or a "chain" of joined FPCVs) similar to both. The choice of a small s_{cut} can lead to heterogenous groups that hide valuable information about the dataset. A large s_{cut} may not reduce the number of FPCVs considerably. I suppose $s_{cut} = 0.85$, which means that two FPCs of 20 points each are considered as similar if they have at least 17 points in common. A subset of at least 16 points is considered as similar to a set of 20 points.

For each of the groups, a representative FPCV is chosen. The set of the representative FPCVs is designed to reveal the essential clusters of the data. Let i_w denote the number of findings of the FPCV w during $i_{n,p}$ algorithm runs. Section 7.3 treats the choice of $i_{n,p}$ and gives arguments for the following suggestions: For each of the groups, the FPCV w with the largest i_w can be considered as representative. In case of equal i_w, the smallest FPC should be chosen.

The results of Table 8.1 in Section 8.1 may be used to assess the effect of the Single Linkage reduction. The similarity s_* is used in Section 8.2 as well to measure the quality of the cluster recovery by the compared cluster analysis methods.

7.3 The number of algorithm runs, the size of FPCVs and the number of findings

As mentioned above, an FPC analysis may lead to a number n_c of FPCs that is too large to interpret. In this section it is attempted to choose the number $i_{n,p}$ of algorithm runs in such a way that

- all FPCs corresponding to relevant structure of the data are found with large probability and
- as few further FPCs as possible are found.

Furthermore, the result of an FPC analysis should be as stable as possible in spite of the dependence on random starting subsets.

Suppose that a dataset of size n contains an FPCV w of size $n(w)$, which is homogeneous and well separated from the rest of the data. To approximate roughly the probability p_w that this FPC is found by one run of the FPA, imagine that it is found by the FPA with a starting constellation of $n(w^0)$ points if and only if all these points are contained in the FPC. That is,

$$p_w \approx \frac{\binom{n(w)}{n(w^0)}}{\binom{n}{n(w^0)}} =: q_{n(w)}. \tag{7.2}$$

From my experience, p_w seldom happens to be much larger than $q_{n(w)}$. It often appears between $q_{n(w)}/2$ and $q_{n(w)}$, and sometimes it is smaller. The latter case means that the FPC has lots of subsets of size $n(w^0)$ leading to other FPCVs if used as starting

[9] In the literature there are some criticisms of Single Linkage clustering. Sometimes it leads to a "chaining effect" by joining vertices with small similarity by chains of vertices in between. This can happen here as well. However, the speed of computing and the clarity of the interpretation of the clusters in terms of s_* is in favor of Single Linkage clustering for the application discussed here.

configurations for the FPA. Then w could not be considered as stable; its parameters would not be supported by many of its subsets. This may indicate that

1. there are similar FPCVs supported by some of the subsets of w,

2. w is not in the neighborhood of an "attractive" (i.e., fulfilling (5.6)) FPCI,

3. the points of w do not form a homogeneous regression population (this is possible for FPCs; see the discussion of Example 6.7) so that small data subsets lead to considerably different parameter estimators and possibly to FPCVs that are subsets of w.

In the first case, the similar FPCVs are expected to fall into the same Single Linkage group of clusters (see Section 7.2). The approximation (7.2) may then be reasonable for the probability of finding an FPCV of this group by an algorithm run.

In the cases 2 and 3, w may be considered as "less interesting" than a more attractive fixed point of $f_{\mathbf{Z}_n}$. Thus it is reasonable to base the choice of $i_{n,p}$ on the approximation (7.2) to find the Single Linkage groups of FPCs *of main interest.*

$q_{n(w)}$ gets smaller for larger $n(w^0)$, and less than $p+2$-points lead to an error variance of 0. Therefore it is advisable to start with $n(w^0) = p+2$ points. If $n(w)/n$ is very small, $q_{n(w)}$ is very small as well, so that one cannot expect to find very small FPCVs by a feasible number of algorithm runs. Therefore a decision is necessary about the smallest size $n_{\min}$ of an FPCV that one wants to find with high probability.

$i_{n,p}$ can be chosen so that the approximated probability to find an FPCV with $n_{\min}$ points at least $i_{\min}$ times is at least 0.95:

$$i_{n,p} := \min\{i: \ \mathrm{QB}(i, q_{n_{\min}}; 0.05) < i_{\min}\}, \tag{7.3}$$

where $\mathrm{QB}(n,p;\alpha)$ denotes the α-quantile of the Binomial(n,p)-distribution. To make the computation fast by keeping $i_{n,p}$ small, $i_{min} = 1$ must be chosen. But the result of the analysis can be made more stable by choosing i_{min} larger:

In order to keep the result stable, it is reasonable to exclude FPCVs that appear only by chance. Two kinds of FPCVs are suspicious of being not reproducable by further applications of the whole FPC analysis procedure:

1. FPCs found too seldom: $i_w/i_{n,p}$ is a positively biased estimate for p_w, since FPCVs with $i_w = 0$ cannot be observed.

2. Too small FPCs: p_w cannot be considerably larger as $q_{n(w)}$ for small $n(w)$, and p_w is presumably overestimated by $i_w/i_{n,p}$, if a small FPCV is found too often.

If $n(w) < n_-$ defined by

$$n_- := \min\{\bar{n}: \ \mathrm{QB}(i_{n,p}, q_{\bar{n}}; 0.5) < i_{\min}\} \tag{7.4}$$

then an FPCV w will be reproduced with an estimated probability of at most 0.5. Thus it should be excluded.

If an FPCV is found seldom, there can be similar FPCVs corresponding to the same, possibly relevant, pattern of the data. But Single Linkage groups of FPCVs should be excluded as well, if they are found less than i_{min} times. This is the more stable the larger

i_{min} is. $i_{\min} = 3$ does not lead to too long computing times for the n and p values used in the simulations of the Section 8.2, but for $p > 2$ it could be advisable to take $i_{min} = 2$, see Section 8.1 for a comparison.

Since p_w measures the stability of an FPCV, it is reasonable to choose the FPCV with the largest i_w as the representative FPCV. In case of equality I propose the FPCV with smaller $n(w)$, since in general it is easier to find larger FPCVs, as explained above.

7.4 A description "ready to run"

Step 1 Choose c according to Table 8.2 (take the closest n), but not smaller than 6.635, $n_{\min} = \frac{n}{5}$, $i_{\min} = 3$, $s_{cut} = 0.85$. Of course, all these choices are subjective since they concern trade-offs between more information and better interpretability, more stability and lower computing time, respectively, as discussed in the previous sections.

Step 2 Compute $i_{n,p}$ according to (7.3), n_- according to (7.4). (If $i_{n,p}$ becomes too large, $i_{min} = 2$ does not change too much and may be the easiest way to save computing time, see Section 8.1).

Step 3 Repeat $i_{n,p}$ times: Generate a subset indicator w^0 with $n(w^0) = p+2$ by random and apply the FPA. Store all found FPCVs w with $n(w) \geq n_-$. Count the number of times that each FPCV has been found.

Step 4 Compute the similarities for each pair of FPCVs according to (7.1).

Step 5 Compute the Single Linkage clusters of index s_{cut} of FPCVs by the following algorithm:

Step a) Suppose that $1 < n_c$ FPCVs $w_1, \ldots, w_{n_c}$ were found. Let $j_F := 1$ (number of FPCV under consideration), $j_G := 1$ (number of group under consideration), $\text{group}(w_j) := F(j) := 0,\ j = 1, \ldots, n_c$ (F(j) indicates if w_j had been already "under consideration").

Step b) $\text{group}(w_{j_F}) := j_G,\ F(j_F) := 1$.

Step c) Find all w_j where $j \neq j_F$, $\text{group}(w_j) = 0$, and $s_*(w_{j_F}, w_j) \geq s_{cut}$. $\text{group}(w_j) := j_G$ for them all.

Step d) Let j_- be the smallest $j \neq j_F$ with $F(j) = 0$, $\text{group}(w_j) = j_G$, if there is any. Else $j_- := 0$.

Step e) If $j_- > 0$: $j_F := j_-$, $F(j_-) := 1$, step c). Else:

Step f) Take the smallest j with $\text{group}(w_j) = 0$. If there is any: $j_F := j$, $j_G := j_G + 1$, step b). Else end. j_G groups of FPCVs are constructed.

Step 6 For $j = 1, \ldots, j_G$:

$$i_j := \sum_{\text{group}(w)=j} i_w.$$

If $i_j \geq i_{\min}$, choose

$$w_j^* := \arg\max_{\text{group}(w)=j} \{i_w\} \text{ (In case of equality take } w \text{ with smallest } n(w)).$$

The w_j^*, $j = 1, \ldots, j_G$, with $i_j \geq i_{\min}$ are the representative FPCVs. Let n_r denote its number in the following.

The C-software needed about 40 seconds for a dataset from constellation "Square-p2" of Section 8.2 as well as for the `Geyser`-data on a Pentium-266. Because of the increasing number of algorithm runs, the time increases exponentially with p and seems to increase linearly with n. Therefore, $p > 4$ is not feasible at the moment.

8 Simulations

8.1 FPCs in homogeneous populations and choice of c

It was discussed previously that in a population consisting of only one homogeneous regression component asymptotically only one FPC is to be expected. But, as mentioned in Section 7.1, for small n and increasing p, the data gets too sparse, resulting in very large numbers of FPCs. The simulations of this section are to show the relation between the number of found FPCs, n, p and the tuning constant c. Here, all points $(x_1, \ldots, x_p, y) \in \mathbb{R}^{p+1}$ were generated according to $\mathcal{N}_{0,\mathbf{I}_{p+1}}$. The procedure of Section 7.4 was used with $i_{min} = 2$. (I repeated some trials with $i_{min} = 3$ resulting in moderately lower values of n_r and larger values of n_c. For example: $p = 1,\ c = 10,\ n = 50$ leads to $n_r(n_c) = 4.4(16.3)$ instead of 6.1(13.3) for $i_{min} = 2$. That is, $i_{min} = 3$ yields better results in terms of n_r.)

There were 10 simulations runs for each constellation. For larger p and small c not only $i_{n,p}$ got very large, but the large n_c also resulted in a large amount of computing time (and required memory) to handle the similarity matrix between the FPCs. Therefore it is generally advisable to have many points for larger p, or, at least, to use a large c. The number of found FPCs n_c was recorded as well as the number of their Single Linkage clusters found often enough (n_r). The simulation results are given in Table 8.1.

While it is obvious that the number of FPCs decreases with increasing n, increasing c and decreasing p, these relations do not seem to follow a uniform functional pattern. To give an orientation, Table 8.2 shows for each combination of n and p the smallest value of c from Table 8.1 such that the average n_r equals or is smaller as 1.5, that is, one may expect a homogeneous population to lead to only one representative FPC with a probability of at least 0.5. The entry "> 30" means that $c = 30$ still leads to considerably more than 1.5 representative FPCs, but n_r is small enough to perform an FPC analysis of such data and interprete the result exploratory. For $p \geq 2$ and too small n it does not seem reasonable to carry out an FPC analysis. (I doubt that any kind of linear regression cluster analysis would be reasonable in these cases.) If desired, n_r can also be made smaller by the choice of a larger i_{min} (to the price of a higher computing time, see above) or by choice of $n_{min} > \frac{n}{5}$, which may be reasonable for small n.

8.2 Comparison of methods

The performance of LS-FPC analysis was compared to two other procedures from the literature by means of a Monte Carlo simulation, namely

Maximum Likelihood Clusterwise Linear Regression (MLCLR) as explained by DeSarbo and Cron (1988). They assume a fixed sequence of regressor values

$p = 0$	$c = 6$	$c = 10$	$c = 15$	$c = 20$	$c = 30$
$n = 25$	4.4 (9.0)	2.4 (3.8)	1.8 (2.3)	1.9 (2.5)	1.4 (1.6)
$n = 50$	2.4 (4.6)	1.0 (1.4)	1.0 (1.1)	1.0 (1.0)	1.0 (1.0)
$n = 100$	1.3 (2.2)	1.1 (1.1)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
$n = 200$	1.0 (1.7)	1.0 (1.1)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
$p = 1$	$c = 6$	$c = 10$	$c = 15$	$c = 20$	$c = 30$
$n = 25$	28.4 (78.2)	19.7 (39.2)	13.7 (23.0)	11.0 (19.0)	6.7 (10.7)
$n = 50$	17.5 (52.0)	6.1 (13.3)	3.4 (6.0)	1.8 (2.6)	1.5 (1.9)
$n = 100$	5.5 (15.0)	1.3 (2.0)	1.0 (1.1)	1.0 (1.1)	1.0 (1.0)
$n = 200$	1.4 (2.7)	1.0 (1.2)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
$p = 2$	$c = 6$	$c = 10$	$c = 15$	$c = 20$	$c = 30$
$n = 25$	137.7 (703.0)	150.0 (366.6)	124.0 (223.8)	103.0 (164.6)	79.7 (106.6)
$n = 50$	90.5 (383.8)	42.6 (100.9)	19.4 (34.4)	10.5 (16.6)	5.1 (6.5)
$n = 100$	36.8 (142.5)	5.6 (15.7)	1.3 (2.3)	1.4 (1.6)	1.0 (1.1)
$n = 200$	4.2 (13.3)	1.0 (1.1)	1.0 (1.0)	1.0 (1.1)	1.0 (1.0)
$p = 3$	$c = 6$	$c = 10$	$c = 15$	$c = 20$	$c = 30$
$n = 50$	508.0 (1980.8)	480.8 (1636.5)	278.9 (684.9)	170.6 (363.7)	78.7 (143.0)
$n = 100$	254.8 (1171.4)	41.4 (122.9)	7.8 (14.3)	2.2 (3.6)	1.1 (1.6)
$n = 200$	31.7 (147.1)	1.0 (1.8)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)
$n = 300$	3.0 (10.3)	1.0 (1.5)	1.0 (1.0)	1.0 (1.0)	1.0 (1.0)

Table 8.1: Average number of representative FPCs (found FPCs) $n_r(n_c)$ for homogeneous data

$x_1, \ldots, x_n$ and model $y_1, \ldots, y_n$ as independently distributed according to

$$\mathcal{L}(y_i) = \sum_{j=1}^{s} \epsilon_j \mathcal{N}_{x_i'\beta_j, \sigma_j^2},$$

where $\epsilon_j > 0,\ j = 1, \ldots, s$, and $\sum_{j=1}^{s} \epsilon_j = 1$. The ϵ_j denote the proportions of the s mixture components. They compute Maximum Likelihood estimators for the parameters $(\epsilon_j, \beta_j, \sigma_j^2),\ j = 1, \ldots, s$ under fixed s using an EM-algorithm, which also provides estimators $\hat{\epsilon}_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, s$, for the probability that the point (x_i, y_i), conditional on its value, was generated by the mixture component j. The point (x_i, y_i) can then be classified as belonging to component $j(i) := \arg\max_j \{\hat{\epsilon}_{ij}\}$. Wedel and DeSarbo (1995) propose the Consistent Akaike's Information Criterion (CAIC) of Bozdogan (1987) to estimate the number of mixture components s. I applied this procedure with estimators from a random partition of the data points as starting values for the parameter estimators, an upper bound of 7 for s and a lower bound of 10^{-6} for the $\sigma_j^2,\ j = 1, \ldots, s$ (otherwise the likelihood function would be unbounded). The algorithm was terminated when the increase of the loglikelihood function fell below 10^{-7}.

Model Based Gaussian Clustering with Noise (MBGCN) as implemented in the

	$p=0$	$p=1$	$p=2$	$p=3$
$n=25$	30	>30	too large	
$n=50$	10	30	>30	too large
$n=100$	6	10	15	30
$n=200$	<6	6	10	30
$n=300$				10

Table 8.2: Smallest c from simulations with $n_r \leq 1.5$

software package `mclust` based on the work of Banfield and Raftery (1993). A current version is treated in DasGupta and Raftery (1998). They assume the points $z_i = (x_i, y_i)$, $i = 1, \ldots, n$, as i.i.d. distributed according to

$$\mathcal{L}(z_i) = \epsilon_0 \mathcal{U}_C + \sum_{j=1}^{s} \epsilon_j \mathcal{N}_{a_j, \mathbf{\Sigma}_j},$$

where $\mathcal{U}_C$ denotes the uniform distribution on some convex set C, $a_j \in I\!R^{p+1}$, $\mathbf{\Sigma}_j$ positive definite $(p+1) \times (p+1)$-covariance matrices for $j = 1, \ldots, s$, $\epsilon_j > 0$, $j = 0, \ldots, s$, and $\sum_{j=0}^{s} \epsilon_j = 1$. The x_i-values from $I\!R^p$ do not include a component for the regression intercept in this setup. Such a normal mixture model can also be applied to linear regression data, since a linear regression distribution $P_{\beta,\sigma^2,G}$ is a $p+1$-variate normal distribution if G is assumed to be a p-variate normal. In fact, DasGupta and Raftery (1998) propose their method for "highly linear" data. The mixture component $\mathcal{U}_C$ is designed to contain noise or outliers not belonging to any of the normal components. The covariance matrices $\mathbf{\Sigma}_j$ can be decomposed as $\mathbf{\Sigma}_j = \lambda_j \mathbf{D}_j \mathbf{A}_j \mathbf{D}_j$, where λ_j is the largest eigenvalue of $\mathbf{\Sigma}_j$, $\mathbf{D}_j$ is the matrix of eigenvectors, $\mathbf{A}_j = \mathrm{diag}(1, \alpha_{2j}, \ldots, \alpha_{(p+1)j})$. The software `mclust` computes Maximum Likelihood estimators using the EM algorithm for the parameters $\epsilon_0, (\epsilon_1, a_1, \mathbf{\Sigma}_1), \ldots, (\epsilon_s, \mathbf{\Sigma}_\mathbf{s})$ from starting values given by some hierarchical model based method from Banfield and Raftery (1993). The component memberships of the points can be estimated by analogy to the MLCLR procedure. The Bayesian Information Criterion BIC (Schwarz 1978) was used for the estimation of the number of components s.

The form of the covariance matrices may be restricted. DasGupta and Raftery propose to assume $\mathbf{A}_j = (1, \alpha)$, $\alpha < 1$ for all $j = 1, \ldots, s$ in their two-dimensional setup to get linearly shaped clusters. The simulations were carried out without restrictions ("MBGCN.vvv") as well as with assuming all $\mathbf{A}_j$ as equal but unknown("MBGCN.vev").

An initial estimation of the noise component is needed, which was generated by the software `NNclean` explained in Byers and Raftery (1998). This software requires the choice of a constant K for the number of nearest neighbors of a point involved in the calculations. I chose $K = 10$. $s \leq 7$ was again assumed. A lower bound for the covariance determinant (to bound the likelihood function) and a convergence criterion were used as implemented in `mclust`.

The used implementation of Least Squares-Fixed Point Clustering (LS-FPC) was described in Section 7.4.

Obviously the procedures differ with respect to their underlying models. MBGCN assumes normal regressor distributions. The MLCLR model does not contain an outlier component, and it assumes the probability of (x_i, y_i) to be generated by mixture component j to be ϵ_j regardless of x_i. I call this latter assumption "assignment independence". It will be illustrated by the discussion of the simulated data constellations. The assumptions of LS-FPC are most general, but it is no exact estimation procedure for normal mixture components, as shown in Section 6. That is, procedures are compared that clearly do not estimate the same features of the data. However, all the methods may be applied to the same data with similar interpretation of the results, since it is not obvious how to decide between the models for given real data. For example, the minefield datasets treated by DasGupta and Raftery (1998) do apparently not meet the normal assumption, and DeSarbo and Cron (1988) do not give arguments in favor of assignment independence or the absence of outliers for their marketing application. Therefore it is interesting to study the performance of their procedures in cases where the assumptions are not fulfilled, but where one may consider the methods as appropriate.

The simulations deal with the recognition of clusters generated by normal linear regression distributions $P_{\beta,\sigma^2,G}$ in mixture models. The clusters are strongly separated so that there exist LS-FPCIs estimated by LS-FPC matching the mixture components.

The results are to illustrate the behavior of the various methods. I have chosen four different constellations of n, p, the β, σ^2, G-parameters and noise for this paper. These and further simulations (Hennig 1997) indicate that the results depend strongly on all the parameter choices. An arbitrary "ranking" of the procedures could easily be illustrated by the choice of the appropriate constellation. I do not attempt to show that LS-FPC is generally better than the ML-methods, but at least there are some situations where it is superior. The constellations are:

Square-p2: $n = 48$, $p = 2$, all x_1-values were generated by $\mathcal{U}_{[0,1]}$, $x_2 := x_1^2$. For the first 24 points: $y = x_1 + u$. For the points 25-48: $y = 0.5x_2 + u$, $\mathcal{L}(u) = \mathcal{N}_{0,0.0001}$ for all points. See Figure 8.1. The assumptions of MLCLR are met since the data contain only linear regression clusters and the distribution of the independent variable does not vary between clusters. The assumptions of MBGCN are not met because of the non-normal regressor distribution.

Square-p1: The data sets of this constellation were generated as the data of the constellation `Square`, but the values of x_2 were not included, thus $n = 48$, $p = 1$. This is data with a linear and a nonlinear cluster. It meets only the assumptions of LS-FPC.

3+Noise: $n = 100$, $p = 1$. Points 1-40 were generated by $\mathcal{L}(x_1) = \mathcal{N}_{0,0.09}$, $y = x_1 + u$, points 41-60 were generated by $\mathcal{L}(x_1) = \mathcal{N}_{5,0.09}$, $y = -x_1 + 5 + u$, points 61-90 were generated by $\mathcal{L}(x_1) = \mathcal{N}_{2.5,0.25}$, $y = 1 + u$, $\mathcal{L}(u) = \mathcal{N}_{0,0.01}$ for points 1-90. The points 91-100 were generated by $\mathcal{U}_{[-1,6]\times[-1,2]}$. See the left side of Figure 8.2. The assumptions of MBGCN are met, approximately even that of MBGCN.vev. The assumptions of MLCLR are not met because of the noise and a strong violation of assignment independence: The domains of the regressors are nearly disjoint for the three clusters.

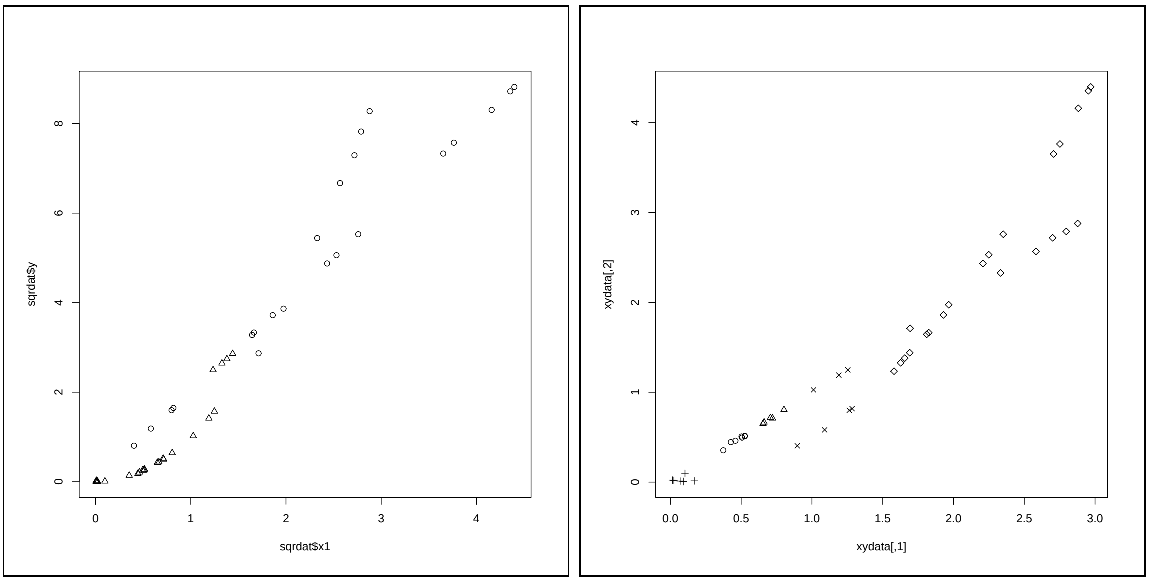

Figure 8.1: Data from the "Square"-constellations. Left: Triangles denote one of the FPCs not corresponding to a mixture component. Right: MBGCN.vev-partition.

2+Noise: $n = 73$, $p = 1$. Points 1-40 were generated by $x_1 = |x_*|$, $\mathcal{L}(x_*) = \mathcal{N}_{0,1}$, $y = x_1 + u$, $\mathcal{L}(u) = \mathcal{N}_{0,0.0004}$. Points 41-70 were generated by $\mathcal{L}(x) = \mathcal{N}_{3.5,1}$, $y = 2 + u$, $\mathcal{L}(u) = \mathcal{N}_{0,0.0025}$. Points 71-73 were generated by $x = 6$, $y = 2+u$, $\mathcal{L}(u) = \mathcal{N}_{0,4}$. See the right side of Figure 8.2. This is a constellation of two regression clusters and three outliers. Neither the assumptions of MLCLR, nor them of MBGCN are met since the regressor distribution of the first cluster is not normal, the noise is not uniform, and the regressor distributions of the two clusters differ strongly.

There were 200 simulation runs for each method with each constellation. The results are given in the Tables 8.3 and 8.4. The estimated number of clusters n_C (meaning the number n_r of found representative FPCs in the case of FPC analysis) was recorded as well as the maximum similarity s_* between an estimated cluster and the given clusters of the constellation. n_C does not include the noise component in case of MBGCN. The estimated number of LS-FPCs cannot be interpreted in the same manner as for the ML-methods. The number of FPCs can often be expected to be larger than the number of clusters found by the partition procedures since

- FPCs may intersect or include each other (in particular there is always almost an LS-FPC containing the whole dataset, similar to the largest LS-FPCIs of the Examples 6.7 and 6.8), and
- the number of found clusters of the other two methods was limited by seven.

The average maximum similarity (over the simulation runs) was used as a measure of how good the methods discovered the given clusters. s_* was discussed in Section 7.2. To interprete the simulation results, recall that a found cluster of similarity of smaller than, say, 0.7 or 0.75 to a desired cluster can hardly be interpreted as a good recovery of this cluster.

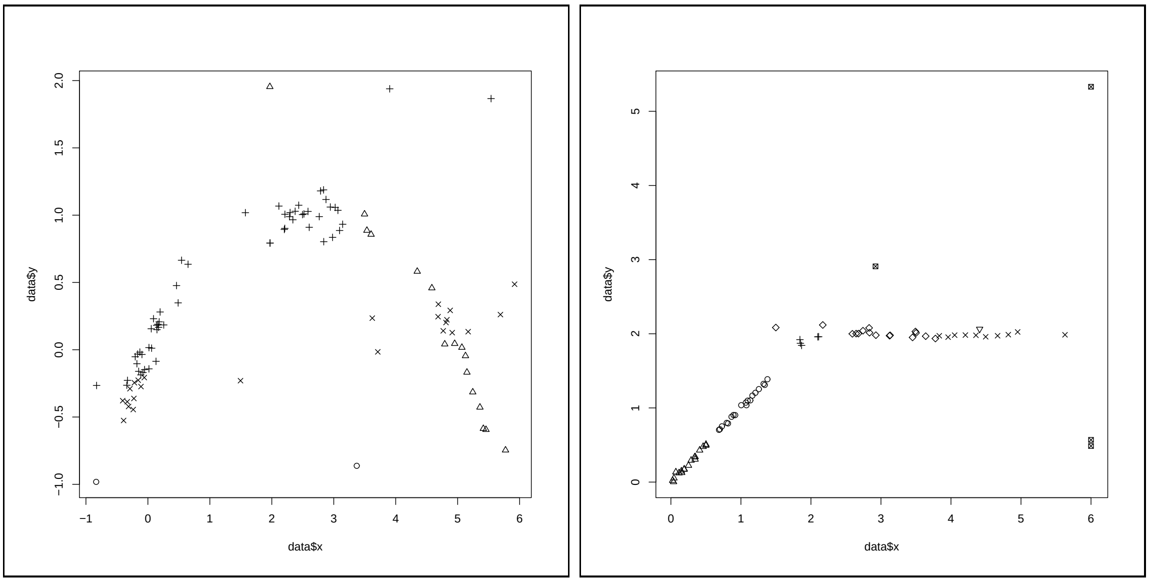

Figure 8.2: Left: Data from constellation “3+Noise” with MLCLR-partition. Right: Data from constellation “2+Noise” with MBGCN.vvv-partition.

Where the model assumptions of one of the ML methods were fulfilled, the corresponding method led to the best results, as one should expect: MLCLR was best for “Square-p2”, MBGCN was best for “3+Noise” (overestimating n_C, however). In both cases, LS-FPC yielded better results than the misspecified ML-method. MBGCN was completely confused by the non-normality of the regressor distribution of the “Square”-constellations, MLCLR suffered strongly from the strong assignment dependence in “3+Noise”. In the constellation “2+Noise”, where the model assumptions of both ML methods were violated, LS-FPC led to the best results. The solutions shown in the Figures 8.1 and 8.2 may illustrate the weaknesses of the three methods.

The performance of LS-FPC depends on the size of the cluster, as can be seen in “3+Noise”. A smaller FPC is more difficult to find, as explained in Section 7.3, while the partition methods often find “all or nothing”. Note the good result of MLCLR for the first cluster of “Square-p1”, where the rest of the data deviates extremely from the model assumption. Usually, the presence of points not belonging to any linear regression mixture component does not prevent the “correct” clusters from being found by MLCLR (except in the case of strong assignment dependence). The other points are simply divided to further clusters (sometimes lots of them). This behavior can be observed at “2+Noise” as well, where the recovery of the two clusters is good, but n_C gets too large.

The number n_r of representative FPCVs is reasonably low at three of the four constellations. After subtracting the usual FPC of the whole dataset, it can keep abreast of the ML methods as an estimator of the number of “cluster-shaped” mixture components. At “Square-p1”, it is clearly the best. However, the result of more than 11 Single Linkage groups of FPCVs found for the constellation “Square-p2” is a serious drawback, even though the groups corresponding to point 1-24, point 25-48, respectively, usually came out with a number of findings i_w on ranks between 2 and 4 among all groups of

Constellation	Square-p2			Square-p1		
Method	Pt. 1-24	Pt. 25-48	n_C	Pt. 1-24	Pt. 25-48	n_C
MBGCN.vvv	0.464	0.448	4.86	0.678	0.531	4.00
MBGCN.vev	0.393	0.402	5.85	0.597	0.577	4.63
MLCLR	0.983	0.979	2.10	0.973	0.652	3.93
LS-FPC ($c = 30$)	0.960	0.956	11.49	0.960	0.667	2.87
LS-FPC ($c = 10$)	0.922	0.914	22.78	0.875	0.674	4.60
LS-FPC ($c = 50$)	0.894	0.894	5.54	0.879	0.668	2.10

Table 8.3: Average maximum similarity s_* of found cluster and average number of found clusters n_C for constellations “Square-p1” and “Square-p2”

Constellation	3+Noise				2+Noise		
Method	Pt. 1-40	Pt. 41-60	Pt. 61-90	n_C	Pt. 1-40	Pt. 41-70	n_C
MBGCN.vvv	0.918	0.878	0.907	3.95	0.848	0.933	3.03
MBGCN.vev	0.934	0.909	0.930	4.02	0.772	0.883	3.77
MLCLR	0.694	0.714	0.529	3.04	0.934	0.914	3.68
LS-FPC ($c = 10$)	0.989	0.762	0.823	5.03	0.966	0.971	3.30

Table 8.4: Average maximum similarity s_* of nearest found cluster and average number of found clusters n_C for constellations “3+Noise” and “2+Noise”

FPCVs (behind the FPCV where the iteration started with the whole dataset). This is in agreement with the result of Table 8.1. The comparison with $c = 10$ and $c = 50$ shows, that $c = 30$ allows the best recovery of the clusters in terms of s_*. The smaller number of FPCVs for a larger c has to be paid by a worse estimation of the relevant patterns. The results for $c = 10$ show that a larger number of FPCVs does not automatically lead to FPCVs corresponding perfectly to the mixture components.

9 Application to the Old Faithful data

Data on the duration of eruptions and the waiting time between the eruptions of the Old Faithful Geyser in the Yellowstone National Park have been discussed in several publications on the basis of data from various time periods. A literature overview, as well as the data set analyzed here, can be found in Azzalini and Bowman (1990). These data were collected in August 1985. Measurements are in minutes. They are shown in Figure 1.1.

The duration of an eruption of the geyser is modeled here as dependent on the waiting time since the previous eruption. There seem to be at least two different groups of dependency, corresponding to the eruptions with lower and higher duration, the latter group with a moderately decreasing tendency for increasing waiting times.

Some authors (e.g. Cook and Weisberg, 1982) model the duration of an eruption as an independent covariate for the subsequent waiting time. Their approach does not reveal any differences between groups. There are no publications up to now that address clustering and dependency between successive events at the same time. Azzalini and Bowman (1990) analyze the data with time series models. They *assume* for their analysis that there are two different patterns of dependency, while I use the data set to illustrate how to *find* such kind of heterogeneity.

The data show some other features: There is a clear outlier with duration smaller than 1. The probability for a long eruption was clearly larger if the waiting time had been short, i.e., the assignment to the two groups is dependent on the independent variable. There are 53 points with duration= 4 exactly, and there are about 20 points with duration= 2. This is due to inexact observations during the night, which were coded as 2 (short eruption), 3 (medium length eruption, only once) and 4 (long eruption) by Azzalini and Bowman.

FPC analysis was applied according to the procedure described in Section 7.4, i.e., $c = 6.635$, $i_{n,p} = 809$. This resulted in eight FPCs. There were six Single Linkage groups of FPCs, and four of them were found three times or more. I concentrate on the interpretation of the four representative FPCs.

The whole dataset was found 521 times as an FPC. It has been discussed previously (Section 4, Examples 6.7, 6.8) that the computation of LS- and error scale estimator on the basis of points from heterogeneous populations leads usually to a very large region of non-outliers and therefore to an FPC corresponding to (almost) the whole dataset. According to the discussion of Section 7.3, this FPC can be expected to be found often. This is an artifact of the method and has to be known in order to interprete the results.

The other representative FPCs are more interesting. The Single Linkage group of the second one was found 217 times. It consists of the circles together with the crosses of Figure 3.1 of Section 3 and corresponds to the group with the longer durations. It excludes the points with the two largest durations as outliers as well as most of the points with medium duration of the eruption. Some more of these points would be included by the choice of $c = 10$, while maintaining roughly the same four representative FPCs. The Single Linkage group of the third representative FPC was found 31 times. It contains the points denoted by triangles in Figure 3.1 and corresponds to the group with the shorter durations, excluding the outlier with duration smaller than 1. The points with duration= 4, denoted by crosses, form the fourth representative FPC, which was found 8 times.

The second, third and fourth representative FPCs give a good description of the main features of the dataset. The possibility of overlapping FPCs is useful here, since an interpretation of the points with duration= 4 as an own cluster is reasonable ("group of inexactly observed long eruptions") *as well as* an interpretation of them as a part of the larger "long duration"-group. The points with duration= 2 form an FPC as well, but it was not found often enough during the iterations, since its number of points is too small (see Section 7.3 for an explanation).

I applied MBGCN.vvv and MLCLR to the dataset as well, as explained in Section 8.2. In both cases, there are multiple local maxima of the likelihood function for most numbers of clusters n_C, leading to strongly varying partitions, so that the results depend on the randomly chosen starting configurations. The best solution of five iterations for each $n_C \leq 7$, followed by the choice of an optimal number of clusters n_C, are shown in Figure 9.1. The procedures needed about the same computing time as the LS-FPC analysis.

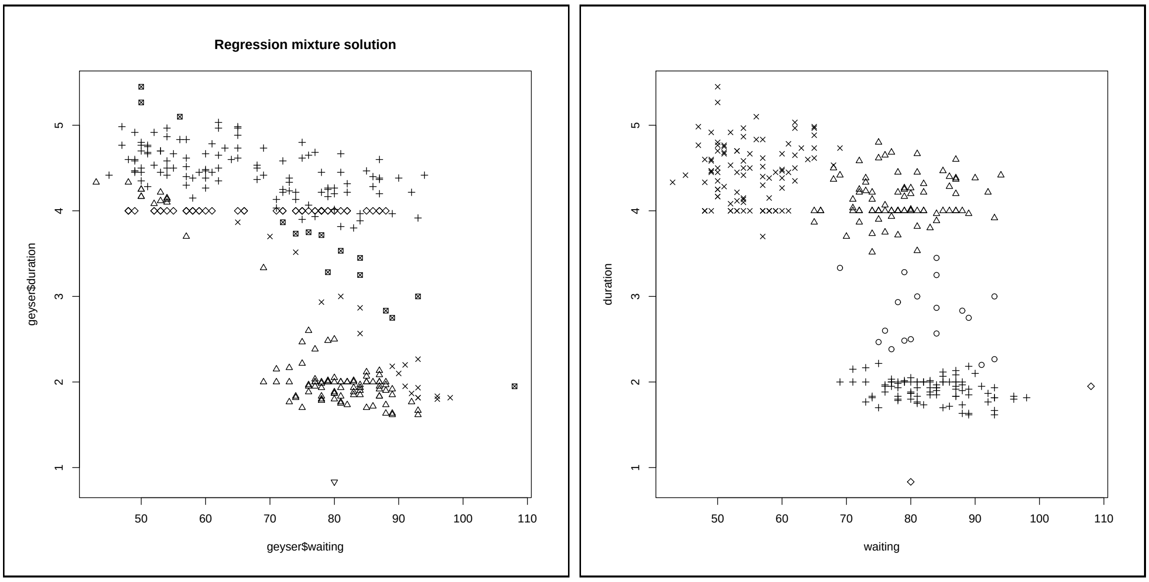

Figure 9.1: Old Faithful Geyser data with partition of MLCLR (left) and MBGCN.vvv (right, diamonds denoting estimated noise).

As mentioned in Section 8.2, both ML-methods need a lower bound on the error variance, the covariance determinant, respectively, because otherwise the likelihood function would be unbounded. Since there are data subsets with variance 0 along the "duration"-axis, the choice of these lower bounds affects crucially whether these subsets result in clusters of the ML-partitions or not. The lower bound of 10^{-6} of MLCLR was seemingly small enough to find the "duration= 4"-cluster. The "duration= 2"-cluster was not found, presumably because of unfortunate starting configurations. The problems here may be similar to those of the LS-FPC analysis with the search for too small clusters.

The MLCLR solution suffers strongly from the assignment dependence in the data. Therefore, the points with the shorter duration do not result as a cluster of the MLCLR-partition.

MBGCN.vvv did not find any of the subsets with singular covariance matrices. The point with the longest waiting time is excluded from the "short duration"-group. There is an ML-cluster of points with medium duration, even if it is not very well separated. The points with long duration are divided into two parts. This corresponds to the aim of the method, since the distribution of the "waiting"-values for the points with long duration seems more like a mixture of two normal distributions than like a homogeneous normal, and MBGCN.vvv looks for subsets that can be described by common 2-dimensional normal distributions, not necessarily by common regression lines (as well as MBGCN.vev).

For these data, the LS-FPC solution seems to be the best, in particular in the light of the discussion of Azzalini and Bowman (1990), which give geological evidence for the existence of two distinct patterns of eruptions, corresponding to the second and third representative FPC, while there are no arguments for breaking the "long duration"-group into two parts as in the MBGCN.vvv-solution. Important for these data are the abilities to deal with subsets of error variance 0 and with clusters including each other.

10 Conclusion

A new concept to define clusters was presented in this paper: An FPC is a data subset that does not contain any outlier, but all other points are outliers w.r.t. the FPC. The basic idea for FPC analysis is to compute iteratively reweighted estimators for which all outliers have zero weight, as for redescending M-estimators. FPC analysis was developed here for clusterwise linear regression, but it may be adapted to other clustering problems. FPCs are not necessarily exhaustive, they may intersect and include each other and they are locally defined, i.e., the FPC property of a data subset does not depend on very distant parts of the dataset.

The existence and consistent estimability of theoretical LS-FPCs of certain probability models of interest was investigated. FPC analysis is not meant as a procedure which is optimal with respect to particular reference models and target functions. It should be a data analytic tool which may be valuable under various deviations from the standard models of the model based CA. However, I tried to give it a solid stochastic foundation. The computation of FPCs and the choice of the required tuning constants was discussed in detail. The simulations and the application to the Old Faithful Geyser dataset showed the ability of the proposed method to deal with data sets that not exactly fulfill the usual mixture model assumptions.

`fixreg`, a C-software for LS-FPC analysis, can be obtained from

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

An `R/S-plus`-module is in preparation and will be available from the same web site.

11 Proofs

11.1 Proof of Theorem 3.2

Part 2 contains the main idea of the proof[10]: Step 2 of the FPA should get the error variance $\hat{\sigma}^2(\mathbf{Z}(w^k))$ small by adding points with squared residual $\leq c\hat{\sigma}^2(\mathbf{Z}(w^k))$ and excluding those with larger residuals. Since $c > 1$, the inclusion of new points does not necessarily decrease the error variance. Part 2 of the proof will show that it can be multiplied by an appropriate factor $\pi_{n(w^k)}$ such that the product $T = \pi_{n(w^k)}\hat{\sigma}^2(\mathbf{Z}(w^k))$ is always decreased. $\pi_{n(w^k)}$ is a product of $n(w^k)$ factors smaller than 1. That is, the FPA tries to unite many points ($\pi_{n(w^k)}$ small) with small $\hat{\sigma}^2(\mathbf{Z}(w^k))$. Part 1 shows $\hat{\beta}(\mathbf{Z}(w^k))$ to be always uniquely defined.

Some notation:

$$\beta_k := \hat{\beta}(\mathbf{Z}(w^k)), \qquad \sigma_k^2 := \hat{\sigma}^2(\mathbf{Z}(w^k)),$$
$$M_k := (y(w^k) - \mathbf{X}(w^k)'\beta_k)'(y(w^k) - \mathbf{X}(w^k)'\beta_k) = (n(w^k) - p - 1)\sigma_k^2,$$
$$\pi_m := \prod_{i=p+1}^{m-1}\left[1\left(\frac{c-1}{i-p} < 1\right)\left(1 - \frac{c-1}{i-p}\right) + 1\left(\frac{c-1}{i-p} \geq 1\right)\frac{1}{c}\right],$$
$$w^+ := \max(w^{k+1} - w^k, 0), \qquad w^- := \max(w^k - w^{k+1}, 0),$$

the maximum taken componentwise. w^+ and w^- indicate the data points that are added, removed respectively, by step 2 of the FPA. Assume $\sigma_k^2 > 0$ for all k up to part 3.

Part 1: Show $\forall k:\ n(w^k) > p{+}1$ by complete induction, which means that β_k is uniquely defined $\forall k$. Recall $n(w^0) > p{+}1$ and show for $m \geq 0 : n(w^m) > p{+}1 \Rightarrow n(w^{m+1}) > p{+}1$. By definition

$$\begin{aligned}&|\{i : w_i^m = 1 \wedge (y_i - x_i'\beta_m)^2 > c\sigma_m^2\}| = n(w^-) \Rightarrow \\ &\Rightarrow \sigma^2(w^m) \geq \tfrac{n(w^-)c\sigma_m^2}{n(w^m)-p-1} \Rightarrow n(w^-) \leq \tfrac{n(w^m)-p-1}{c}.\end{aligned} \tag{11.1}$$

Assuming $n(w^m) \geq p + c + 1$:

$$\begin{aligned}n(w^{m+1}) &\geq n(w^m) - n(w^-) \geq \left(1 - \tfrac{1}{c}\right)n(w^m) + \tfrac{p+1}{c} \geq \\ &\geq \left(1 - \tfrac{1}{c}\right)(p + c + 1) + \tfrac{p+1}{c} = p + c > p + 1.\end{aligned}$$

On the other hand $n(w^-) \in I\!N$ and with (11.1):

$$n(w^m) < p + c + 1 \Rightarrow 1 > n(w^-) = 0 \Rightarrow n(w^{m+1}) \geq n(w^m).$$

Part 2: Show that

$$T : \{0,1\}^n \mapsto [0,\infty) :\ w \mapsto \pi_{n(w)}\sigma_{\mathbf{Z}}^2(w) \tag{11.2}$$

is strictly decreased by step 2 of the FPA unless $n(w^+) = n(w^-) = 0$, i.e., $w^{k+1} = w_{\mathbf{Z}}(w^k) = w^k$. Therefore, no w^m with $w^{m+1} \neq w^m$ can be repeated during the FPA, and the FPA converges in a finite number of steps because there are only finitely many indicator vectors of length n.

[10] The main idea is taken from the proof of Theorem 9.2 in Hennig (1997), where the convergence of a modified FPA was proven.

Assume that $n(w^+) > 0$, or $n(w^-) > 0$ and show $T(w^{k+1}) - T(w^k) < 0$.

$$\begin{aligned} T(w^{k+1}) - T(w^k) &= \pi_{n(w^{k+1})}\sigma^2_{k+1} - \pi_{n(w^k)}\sigma^2_k = \\ &= \left(\tfrac{\pi_{n(w^{k+1})}(n(w^k)-p-1)}{n(w^{k+1})-p-1} - \pi_{n(w^k)}\right)\sigma^2_k + \tfrac{\pi_{n(w^{k+1})}}{n(w^{k+1})-p-1}(M_{k+1} - M_k). \end{aligned} \tag{11.3}$$

Yield

$$\begin{aligned} M_{k+1} &= \min_\beta (y(w^{k+1}) - \mathbf{X}(w^{k+1})\beta)'(y(w^{k+1}) - \mathbf{X}(w^{k+1})\beta) \leq \\ &\leq (y(w^{k+1}) - \mathbf{X}(w^{k+1})\beta_k)'(y(w^{k+1}) - \mathbf{X}(w^{k+1})\beta_k) \leq \\ &\leq M_k + n(w^+)c\sigma^2_k - n(w^-)c\sigma^2_k \end{aligned} \tag{11.4}$$

by definition of w^+ and w^-. If $n(w^-) > 0$, there is strict "<". Hence with (11.3):

$$T(w^{k+1}) - T(w^k) \leq \left(\frac{\pi_{n(w^{k+1})}(n(w^k) - p - 1 + [n(w^+) - n(w^-)]c)}{n(w^{k+1}) - p - 1} - \pi_{n(w^k)}\right)\sigma^2_k =: q. \tag{11.5}$$

Let $d := n(w^+) - n(w^-) = n(w^{k+1}) - n(w^k)$. If $d = 0$ then

$$1 = \frac{\pi_{n(w^k)}}{\pi_{n(w^{k+1})}} = \frac{n(w^k) - p - 1 + dc}{n(w^{k+1}) - p - 1}$$

and therefore $q = 0$ and $T(w^{k+1}) - T(w^k) < 0$ (there is strict inequality in (11.5) because $d = 0$ and $w^k \neq w^{k+1}$ imply $n(w^-) > 0$). Show further

$$d < 0 \Rightarrow \frac{\pi_{n(w^k)}}{\pi_{n(w^{k+1})}} \geq \frac{n(w^k) - p - 1 + dc}{n(w^{k+1}) - p - 1} \tag{11.6}$$

which implies $q \leq 0$ and again $T(w^{k+1}) - T(w^k) < 0$ by strict inequality in (11.5), and

$$d > 0 \Rightarrow \frac{\pi_{n(w^k)}}{\pi_{n(w^{k+1})}} > \frac{n(w^k) - p - 1 + dc}{n(w^{k+1}) - p - 1} \tag{11.7}$$

implying $q < 0$ and strict decrease of T.

Proof of (11.6): If $n(w^{k+1}) \leq p + c$, then

$$\frac{n(w^k) - p - 1 + dc}{n(w^{k+1}) - p - 1} = 1 + \frac{(c-1)d}{n(w^{k+1}) - p - 1} \leq 0 < \frac{\pi_{n(w^k)}}{\pi_{n(w^{k+1})}}.$$

On the other hand, with $n(w^k) > n(w^{k+1}) > p + c$, get

$$\frac{\pi_{n(w^k)}}{\pi_{n(w^{k+1})}} = \prod_{i=n(w^{k+1})}^{n(w^k)-1}\left(1 - \frac{c-1}{i-p}\right) \geq \left(1 - \frac{c-1}{n(w^{k+1}) - p - 1}\right)^{-d}.$$

Use $(1-b)^m \geq 1 - bm$ for $0 < b < 1, m \in I\!N$. Let $b = \frac{c-1}{n(w^{k+1})-p-1}$, $m = -d$. Then

$$\frac{\pi_{n(w^k)}}{\pi_{n(w^{k+1})}} \geq 1 + d\frac{c-1}{n(w^{k+1}) - p - 1} = \frac{n(w^k) - p - 1 + dc}{n(w^{k+1}) - p - 1}.$$

Proof of (11.7): By complete induction over $m > 0$ (see Hennig (1997) for details) get for $b > 0, m \in I\!N$:

$$1 - \tfrac{(c-1)m}{b+mc} > \left(1 - \tfrac{c-1}{b+m}\right)^{m_1} \left(\tfrac{1}{c}\right)^{m_2} \ \forall m_1, m_2 \in I\!N_0 : m_1 + m_2 = m \tag{11.8}$$

$$\text{assuming } c > 1 \text{ and } \tfrac{c-1}{b+m} < 1. \tag{11.9}$$

$$\text{In particular get } 1 - \tfrac{(c-1)m}{b+mc} = \tfrac{b+m}{b+mc} > \tfrac{1}{c} \geq \left(\tfrac{1}{c}\right)^m. \tag{11.10}$$

Start with $n(w^{k+1}) > p+c-1$. Since $n(w^{k+1}) > n(w^k) > p+1$, apply (11.8) with $m = d$, $m_1 = n(w^{k+1}) - \max(n(w^k), \lfloor p+c-1 \rfloor + 1)$, $b = n(w^k) - p - 1$. (11.9) is fulfilled since $\frac{c-1}{b+m} = \frac{c-1}{n(w^{k+1})-p-1} < 1$.

$$\frac{\pi_{n(w^{k+1})}}{\pi_{n(w^k)}} = \prod_{i=n(w^k)}^{n(w^{k+1})-1} \left(1\left(\frac{c-1}{i-p} < 1\right)\left(1 - \frac{c-1}{i-p}\right) + 1\left(\frac{c-1}{i-p} \geq 1\right)\frac{1}{c}\right) =$$

$$= \prod_{\max(n(w^k), \lfloor p+c-1 \rfloor+1) < i \leq n(w^{k+1})-1} \left(1 - \frac{c-1}{i-p}\right) \prod_{i=n(w^k)}^{\lfloor p+c-1 \rfloor} \frac{1}{c} \leq$$

$$\leq \left(1 - \tfrac{c-1}{n(w^{k+1})-p-1}\right)^{m_1} \left(\tfrac{1}{c}\right)^{d-m_1} < 1 - \tfrac{(c-1)d}{n(w^k)-p-1+dc} = \tfrac{n(w^{k+1})-p-1}{n(w^k)-p-1+dc},$$

i.e., (11.7). With $n(w^{k+1}) \leq p + c - 1$ get

$$\frac{\pi_{n(w^{k+1})}}{\pi_n(w^k)} = \left(\frac{1}{c}\right)^d < \frac{n(w^{k+1}) - p - 1}{n(w^k) - p - 1 + dc}$$

because of (11.10).

Part 3: $\sigma_k^2 = 0$ for some k. $\sigma_{k-1}^2 > 0 \Rightarrow n(w^k) > p + 1$ because of part 1. Further, $w_i^k = 1 \Rightarrow (y_i - x_i'\beta_k)^2 = 0$, $w_i^{k+1} = 1 \Leftrightarrow (y_i - x_i'\beta_k)^2 = 0$ and $n(w^{k+1}) \geq n(w^k) > p + 1$. Hence $\sigma_{k+1}^2 = \sigma_k^2 = 0$, $\beta_{k+1} = \beta_k$, $w^{k+2} = w^{k+1}$.

11.2 Some useful results for the following proofs

Assumptions: Let Q be some measure on $(I\!R^{p+2}, I\!B^{p+2})$, $M \subseteq I\!R^{p+1} \times I\!R^+$.

$$Qy^2 < \infty, \ Q\|x\|^2 < \infty, \ (Qxx')^{-1} \text{ exists,} \tag{11.11}$$

$$Q\{(y - x'\beta)^2 = c\sigma^2\} = 0 \ \forall (\beta, \sigma^2) \in M, \tag{11.12}$$

$$(Qxx'w_{\beta,\sigma^2}(x,y))^{-1} \text{ exists } \forall (\beta, \sigma^2) \in M, \tag{11.13}$$

Proposition 11.1 $\arg\min\limits_{\beta} Q(y - x'\beta)^2 = (Qxx')^{-1}Qxy$ *exists uniquely under (11.11).*

Proven as Proposition 11.1 of Hennig (1997).

Proposition 11.2 *Let* $l_1(a,b) := Qv(x,y)1[(y-x'b)^2 \leq ca^2]$ *where* $v : I\!R^{p+2} \mapsto I\!R^q$, $Q\|v(x,y)\| < \infty$. l_1 *is continuous on* M *under (11.12).*

$l_2(a_1, b_1, a_2, b_2) := Q1[(x'b_2)^2 > a_2^2]1[(y-x'b_1)^2 \leq ca_1^2]$ *is continuous in* $(a_1, b_1, a_2, b_2) \in M \times I\!R^{p+2}$ *under (11.12) and (11.13).*

f_Q *is continuous on* M *under (11.11), (11.12) and (11.13).*

Continuity of l_1 and f_Q were proven as Proposition 11.2/(11.9) of Hennig (1997). Continuity of l_2 follows by analogy to (11.9) of Hennig (1997).

11.3 Further preparations for the proof of Lemma 5.1

Here are useful parts of the "uniform convergence of the empirical process"-machinery, see e.g. van der Vaart and Wellner (1996), Pollard (1984):

Let $\mathcal{F}$ be a class of measurable functions $\mathbb{R}^k \mapsto \mathbb{R}$. For some real-valued function h, $|h|$ denotes the supremum norm. Some measurable function F with $|f| \leq F\ \forall f \in \mathcal{F}$ is called **Q-finite envelope** if Q is some measure with $QF < \infty$.

Definition 11.3 *For $\epsilon > 0$ and some measure Q on $(\mathbb{R}^d, \mathbb{B}^d)$, the* **covering number** *$N(\epsilon, \mathcal{F}, Q)$ is the minimum number of balls $\{g: \ Q|g-f| < \epsilon\}$ needed to cover $\mathcal{F}$. The centers f need not to belong to $\mathcal{F}$.*

Definition 11.4 *$\mathcal{F}$ is called* **permissible** *if it can be indexed by some T with Borel-σ-field $\mathcal{B}$ in such a way that $f(\bullet, \bullet)$ is $\mathbb{B}^k \otimes \mathcal{B}$-measurable and T is an analytic subset of some compact metric space. (Pollard 1984, Example 2.3.5. of van der Vaart and Wellner (1996))*

Theorem 11.5 *Let $\mathcal{F}$ be permissible with P-finite envelope F. If*

$$\forall \epsilon > 0: \ \log N(\epsilon, \mathcal{F}, P_n) = o_P(n), \tag{11.14}$$

$$\textit{then } \sup_{f \in \mathcal{F}} |P_n f - Pf| \to 0 \ P^\infty - a.s.$$

(Van der Vaart and Wellner (1996), Theorem 2.4.3)

Definition 11.6 *Let $\mathcal{C}$ be a collection of subsets of some set S. $\mathcal{C}$ is said to* **shatter** *some set $S_* \subset S$ if every subset of S_* can be formed as $S_* \cap C$, $C \in \mathcal{C}$. $\mathcal{C}$ is called a* **Vapnik-Chervonenkis(VC)-class** *with index $V(\mathcal{C}) \in \mathbb{N}$, if $\mathcal{C}$ shatters no subset of S with $V(\mathcal{C})$ elements.*

Proposition 11.7 *Subsets of VC-classes are trivially VC-classes. Classes of intersections of sets from VC-classes are VC-classes (van der Vaart and Wellner (1996), Lemma 2.6.17).*

Proposition 11.8 *$\mathcal{F}$ fulfills*

$$\forall \epsilon > 0: \ \sup_{Q \in \mathcal{Q}} \log N(\epsilon QF, \mathcal{F}, Q) < \infty,$$

$\mathcal{Q}$ being the set of all finite measures with $0 < QF < \infty$, and therefore (11.14), if the set of subgraphs $\{(x,t) \in \mathbb{R}^{k+1}: \ t < f(x)\}$, $f \in \mathcal{F}$, is a VC-class (van der Vaart and Wellner (1996), Theorem 2.6.7). Such $\mathcal{F}$ is itself called VC-class.

Proposition 11.9 *The set of indicator functions $\mathcal{I}(\mathcal{C})$ of some class of sets $\mathcal{C}$ is a VC-class iff $\mathcal{C}$ is a VC-class.*

Proof: $\mathcal{C}$ shatters some set $S_* = \{s_1, \ldots, s_n\}$ iff the class of subgraphs of $\mathcal{I}(\mathcal{C})$ shatters $\{(s_1, 0), \ldots, (s_n, 0)\}$.

Proposition 11.10 *For some positive, measurable function f and some measure Q on some S with $Qf < \infty$, let fQ be the measure with density f w.r.t. Q. Then, for sets of measurable functions $\mathcal{F}, \mathcal{G}$ with Q-finite envelopes F, G:*

$$\forall u > 0: \; N(u, \mathcal{F}\mathcal{G}, Q) \leq N\left(\frac{u}{2}, \mathcal{F}, GQ\right) N\left(\frac{u}{2}, \mathcal{G}, FQ\right),$$

$\mathcal{F}\mathcal{G}$ being the set of functions $fg, \; f \in \mathcal{F}, g \in \mathcal{G}$.

Proof: Let $\mathcal{F}_*$ with $|\mathcal{F}_*| = N(\frac{u}{2}, \mathcal{F}, GQ)$ some set of suitable functions such that the balls $\{g: \; GQ|g-f| < \frac{u}{2}\}, \; f \in \mathcal{F}_*$, cover $\mathcal{F}$, analogously $\mathcal{G}_*$. Let $h = fg, \; f \in \mathcal{F}, g \in \mathcal{G}$. Let $h_* := f_* g_*, \; f_* \in \mathcal{F}_*, g_* \in \mathcal{G}_*$ such that $GQ|f - f_*| \leq \frac{u}{2}, FQ|g - g_*| \leq \frac{u}{2}$. With that,

$$|fg - f_*g_*| \leq |fg - f_*g| + |f_*g_* - f_*g| \leq |f - f_*||g| + |f_*||g_* - g| \Rightarrow$$
$$\Rightarrow Q|h - h_*| \leq GQ|f - f_*| + FQ|g - g_*| \leq u \Rightarrow N(u, \mathcal{F}\mathcal{G}, Q) \leq |\mathcal{F}_*\mathcal{G}_*|.$$

Proposition 11.11 *If $\mathcal{F}_*$ is a finite dimensional vector space of real valued functions on S, $\mathcal{F}_*$ as well as the class of sets of the form $\{f \geq 0\}, \; f \in \mathcal{F}_*$ are VC-classes. "$\geq$" can be replaced by "$\leq$", "$>$", "$<$", respectively (van der Vaart and Wellner (1996), Lemma 2.6.15, 2.6.17, Pollard (1984), Lemma II.18).*

From Propositions 11.11 and 11.7 get that

Proposition 11.12 *The set $\{\{w_{\beta,\sigma^2} = 1\}: \; (\beta, \sigma^2) \in D\}$ is a VC-class for arbitrary $D \subseteq I\!R^{p+1} \times R_0^+$.*

Further propositions are needed for the proof of Lemma 5.1. For $(t, \beta, \sigma^2) \in I\!R^{p+1} \times I\!R^{p+1} \times I\!R_0^+$ define

$$f_{t,\beta,\sigma^2}(x,y) := (y - x't)^2 1[(y - x'\beta)^2 \leq c\sigma^2], \; (x,y) \in I\!R^p \times \{1\} \times I\!R.$$

Proposition 11.13 *Under the assumptions of Lemma 5.1, define $\sigma_*^2 := \max\{\sigma^2 : (\beta, \sigma^2) \in C\}$. $\forall \eta > 0 \; \exists d_{C,\eta} < \infty$ such that $\|t\| > d_{C,\eta} \Rightarrow \inf_{(\theta,s^2) \in C} Pf_{t,\theta,s^2} > c\sigma_*^2 + \eta$ and, P^∞-a.s. for large enough n, $\inf_{(\theta,s^2) \in C} P_n f_{t,\theta,s^2} > c\sigma_*^2 + \eta$.*

Proof: $S_p := \{x \in I\!R^p: \; \|x\| = 1\}$. Show that there exists $\tau > 0$ such that

$$a_P := \inf_{(\beta,\sigma^2) \in C, \; z \in S_{p+1}} P(L_{\beta,\sigma^2,\tau,z}) > 0, \quad L_{\beta,\sigma^2,\tau,z} := \{|x'z| > \tau\} \cap \{w_{\beta,\sigma^2} = 1\}. \quad (11.15)$$

Suppose that (11.15) does not hold. Then, because of the compactness of S_{p+1} and C, there is a sequence $(\beta_n, \sigma_n^2, \tau_n, z_n)_{n \in I\!N}$ with

$$(\beta_n, \sigma_n^2, \tau_n, z_n) \to (\beta_0, \sigma_0^2, 0, z_0), \; (\beta_0, \sigma_0^2) \in C, \; z_0 \in S_{p+1},$$

and $P(L_{\beta_n,\sigma_n^2,\tau_n,z_n}) \to 0$. Use Proposition 11.2 to get $P(L_{\beta_0,\sigma_0^2,0,z_0}) = 0$. But this contradicts (5.4) because of $P(\{|x'z_0| = 0\} \cap \{w_{\beta_0,\sigma_0^2} = 1\}) = 0$ by (5.2). Therefore (11.15).

Further get

$$a_{P_n} = \inf_{(\beta,\sigma^2) \in C, \; z \in S_{p+1}} P_n(L_{\beta,\sigma^2,\tau,z}) \to a_P > 0 \quad P^\infty - \text{a.s.}$$

since the sets $\{|x'z| > \tau\} = \{x'zz'x - \tau^2 > 0\}, z \in S_{p+1}$ form a VC-class by Propositions 11.11 and 11.7, and so do the intersections with the sets $\{w_{\beta,\sigma^2} = 1\}$, $(\beta, \sigma^2) \in C$ by Propositions 11.7 and 11.12. By Propositions 11.9 and 11.8, Theorem 11.5 can be applied to their indicator functions; permissibility is obvious. For $t \in I\!R^{p+1}$, $(\beta, \sigma^2) \in C$, $(x, y) \in L_{\beta,\sigma^2,\tau,\frac{t}{\|t\|}}$, get $|x't| = |x'\frac{t}{\|t\|}|\|t\| \geq \|t\|\tau$ and hence for $Q = P$ and P^∞-a.s. for sufficiently large n for $Q = P_n$:

$$Qf_{t,\beta,\sigma^2} = \int (y - x't)^2 1[(y - x'\beta)^2 \leq c\sigma^2] dQ \geq$$
$$\geq \int (y - x't)^2 1[L_{\beta,\sigma^2,\tau,\frac{t}{\|t\|}}] dQ \geq \int |t'x|(|t'x| - 2|y|) 1[L_{\beta,\sigma^2,\tau,\frac{t}{\|t\|}}] dQ \geq$$
$$\geq \|t\|\tau(\|t\|\tau a_Q - 2Q|y|),$$

which exceeds $c\sigma_*^2$ for sufficiently large $\|t\|$ since $a_Q > 0$ and $Q|y| < \infty$. Existence of $d_{C,\eta}$ follows.

Corollary 11.14 *Let $\eta > 0$, $(\beta, \sigma^2) \in C$. For $Q = P$ and P^∞-a.s. for $Q = P_n$, n large enough:*

$$\inf_{\|t\| \geq d_{C,\eta}} Qf_{t,\beta,\sigma^2} > \arg\min_t Qf_{t,\beta,\sigma^2} + \eta.$$

Proof: $Qf_{\beta,\beta,\sigma^2} \leq c\sigma_*^2$ by definition of f_{t,β,σ^2}.

Proposition 11.15 *Under the assumptions of Lemma 5.1,*

$$\forall \kappa > 0 : \inf_{(\beta,\sigma^2) \in C} \inf_{\|t - \beta_P(\beta,\sigma^2)\| \geq \kappa} \left(Pf_{t,\beta,\sigma^2} - Pf_{\beta_P(\beta,\sigma^2),\beta,\sigma^2} \right) > 0.$$

Proof: Suppose that the Proposition does not hold, i.e., there is some sequence $(t_m, \beta_m, \sigma_m^2)_{m \in I\!N}$ where $\|t_m - \beta_P(\beta_m, \sigma_m^2)\| > \kappa > 0$ and

$$\left| Pf_{t_m,\beta_m,\sigma_m^2} - Pf_{\beta_P(\beta_m,\sigma_m^2),\beta_m,\sigma_m^2} \right| \to 0.$$

$(t_m, \beta_m, \sigma_m^2)_{m \in I\!N}$ has a compact domain since $(\beta_m, \sigma_m^2) \in C$ and $\|t_m\| \leq d_{C,\kappa}$ by Corollary 11.14 for large m. Hence the sequence can be chosen convergent to some $(t_0, \beta_0, \sigma_0^2) \in \{\|t\| \leq d_{C,\eta}\} \times C$ where $|t_0 - \beta_P(\beta_0, \sigma_0^2)| \geq \kappa$. Pf_{t,β,σ^2} is continuous in (t, β, σ^2) under (5.1) and (5.3) by Proposition 11.2, thus

$$Pf_{t_m,\beta_m,\sigma_m^2} \to Pf_{t_0,\beta_0,\sigma_0^2} = Pf_{\beta_P(\beta_0,\sigma_0^2),\beta_0,\sigma_0^2}.$$

In contradiction to that, $\arg\min_t Pf_{t,\beta_0,\sigma_0^2}$ is uniquely defined because of (5.2) and (5.3) (Proposition 11.1). This proves the proposition.

11.4 Proof of Lemma 5.1

Define for some $\eta > 0$

$$\mathcal{F}_C := \{f_{t,\beta,\sigma^2} : \ \|t\| \leq d_{C,\eta}, \ (\beta, \sigma^2) \in C\}.$$

$\mathcal{F}_C$ is permissible by its parameterization. $\mathcal{F}_C$ has $P-$finite envelope $F_C(x, y) := y^2 + 2|y|\|x\|d_{C,\eta} + \|x\|^2 d_{C,\eta}^2$ and is VC-class with help of Propositions 11.12, 11.9, 11.11 and 11.10 (use $G \equiv 1, F = F_C$). Theorem 11.5 yields

$$\sup_{(\beta,\sigma^2) \in C, \|t\| \leq d_{C,\eta}} |P_n f_{t,\beta,\sigma^2} - Pf_{t,\beta,\sigma^2}| \to 0 \quad P^\infty - \text{a.s.} \tag{11.16}$$

By definition $\beta_{\mathbf{Z}_n}(w_{\beta,\sigma^2}) = \arg\min_t P_n f_{t,\beta,\sigma^2}$. For sufficiently large n, the arg min can be taken over $\{\|t\| \leq d_{C,\eta}\}$ by Corollary 11.14 and exists uniquely with probability 1 because of (5.2) and (5.4) . Thus $\|\beta_P(\beta,\sigma^2)\| \leq d_{C,\eta}\ \forall(\beta,\sigma^2) \in C$.

Now for arbitrary $\kappa > 0$:

$$\sup_{(\beta,\sigma^2)\in C} \left|P_n f_{\beta_P(\beta,\sigma^2),\beta,\sigma^2} - P f_{\beta_P(\beta,\sigma^2),\beta,\sigma^2}\right| \to 0\ P^\infty - \text{a.s.},$$

$$\sup_{(\beta,\sigma^2)\in C}\ \sup_{\{\|t-\beta_P(\beta,\sigma^2)\|>\kappa\}\cap\{\|t\|\leq d_{C,\eta}\}} \left|P_n f_{t,\beta,\sigma^2} - P f_{t,\beta,\sigma^2}\right| \to 0\ P^\infty - \text{a.s., thus}$$

$$P^\infty\left\{\exists n_0\ \forall n \geq n_0 :\ \sup_{(\beta,\sigma^2)\in C} \|\beta_{\mathbf{Z}_n}(\beta,\sigma^2) - \beta_P(\beta,\sigma^2)\| < \kappa\right\} = 1. \qquad (11.17)$$

Further, by definition,

$$\sigma^2_{\mathbf{Z}_n}(\beta,\sigma) = \frac{nP_n f_{\beta_{\mathbf{Z}_n}(\beta,\sigma^2),\beta,\sigma^2}}{(n-p-1)P_n\{w_{\beta,\sigma^2}=1\}},\ \sigma^2_P(\beta,\sigma^2) = \frac{P f_{\beta_P(\beta,\sigma^2),\beta,\sigma^2}}{P\{w_{\beta,\sigma^2}=1\}}.$$

Theorem 11.5 yields

$$\sup_{(\beta,\sigma^2)\in C} \left|P_n\{w_{\beta,\sigma^2}=1\} - P\{w_{\beta,\sigma^2}=1\}\right| \to 0\ P^\infty - \text{a.s., and}$$

$$\sup_{(\beta,\sigma^2)\in C} \left|P_n f_{\beta_{\mathbf{Z}_n}(\beta,\sigma^2),\beta,\sigma^2} - P f_{\beta_P(\beta,\sigma^2),\beta,\sigma^2}\right| \to 0\ P^\infty - \text{a.s.}$$

since, by (11.16),

$$\sup_{(\beta,\sigma^2)\in C} \left|P_n f_{\beta_{\mathbf{Z}_n}(\beta,\sigma^2),\beta,\sigma^2} - P f_{\beta_{\mathbf{Z}_n}(\beta,\sigma^2),\beta,\sigma^2}\right| \to 0\ P^\infty - \text{a.s.,}$$

$$\sup_{(\beta,\sigma^2)\in C} \left|P_n f_{\beta_P(\beta,\sigma^2),\beta,\sigma^2} - P f_{\beta_P(\beta,\sigma^2),\beta,\sigma^2}\right| \to 0\ P^\infty - \text{a.s.,}$$

$$P_n f_{\beta_P(\beta,\sigma^2),\beta,\sigma^2} \geq P_n f_{\beta_{\mathbf{Z}_n}(\beta,\sigma^2),\beta,\sigma^2}$$

$$P f_{\beta_{\mathbf{Z}_n}(\beta,\sigma^2),\beta,\sigma^2} \geq P f_{\beta_P(\beta,\sigma^2),\beta,\sigma^2}.$$

Observe

$$\sup_{(\beta,\sigma^2)\in C} \left|\frac{nP_n f_{\beta_{\mathbf{Z}_n}(\beta,\sigma^2),\beta,\sigma^2}}{(n-p-1)P_n\{w_{\beta,\sigma^2}=1\}} - \frac{P f_{\beta_P(\beta,\sigma^2),\beta,\sigma^2}}{P\{w_{\beta,\sigma^2}=1\}}\right| \to 0\ P^\infty - \text{a.s.,}$$

since the denominators are guaranteed to be P^∞-a.s. nonzero for large enough n by (5.4). This proves Lemma 5.1 together with (11.17).

11.5 Proof of Theorem 5.3

Because of Theorem 3.2 and (5.2), there exist LS-FPCV w_n w.r.t. $\mathbf{Z}_n$ for all $n \in I\!N_>$ with probability 1. For sufficiently large n, there exist P^∞-a.s. all $f_{\mathbf{Z}_n}(w_{\beta,\sigma^2})$, $(\beta,\sigma^2) \in B_{\epsilon_1}(\beta_0,\sigma_0^2)$ for some $\epsilon_1 > 0$ since

$$\exists \epsilon_1 > 0 :\ P\left(\bigcap_{(\beta,\sigma^2)\in B_{\epsilon_1}(\beta_0,\sigma_0^2)} \{w_{\beta,\sigma^2}=1\}\right) > 0, \qquad (11.18)$$

as shown at the end of the proof.

Choose $\epsilon, \kappa > 0$ small enough that $\alpha\epsilon + \kappa < \epsilon < \min(\epsilon_0, \epsilon_1)$. (5.4) follows from (11.18). Hence Lemma 5.1 can be applied. With help of (5.6) get

$$P^\infty\left\{\exists n_0 \in I\!N_> \; \forall n > n_0, (\beta, \sigma^2) \in B_\epsilon(\beta_0, \sigma_0^2) : \; f_{\mathbf{Z}_n}(\beta, \sigma^2) \in B_\epsilon(\beta_0, \sigma_0^2)\right\} = 1. \quad (11.19)$$

Thus, the fixed point algorithm started with some w_{β,σ^2}, $(\beta, \sigma^2) \in B_\epsilon(\beta_0, \sigma_0^2)$, stays inside of $B_\epsilon(\beta_0, \sigma_0^2)$ with probability 1. It converges by Theorem 3.2 and (5.2), and therefore $f_{\mathbf{Z}_n}$ has a fixed point almost surely for all $n \in I\!N_>$, and for $n > n_0$ it can be found in $B_\epsilon(\beta_0, \sigma_0^2)$. Let $(\eta_i)_{i \in I\!N}$ some sequence with $\eta_i \searrow 0$. Let $U :=$

$$\bigcap_{\eta_i, i \in I\!N} \left\{\exists n_0 \in I\!N_> \; \forall n > n_0 \; \exists w_{\beta_n, \sigma_n^2} \text{ LS-FPCV w.r.t. } \mathbf{Z}_n : \; \|f_{\mathbf{Z}_n}(\beta_n, \sigma_n^2) - (\beta_0, \sigma_0^2)\| < \eta_i\right\},$$

and observe $P^\infty(U) = 1$, which proves the theorem.

Proof of (11.18): Get by (5.5), (3.1) and (5.1)

$$P\{w_{\beta_0, \sigma_0^2} = 1\} = P\{(x, y) : \; (y - x'\beta_0)^2 < c\sigma_0^2\} > 0.$$

If $(\beta, \sigma^2) \in B_\epsilon(\beta_0, \sigma_0^2)$, $\{w_{\beta,\sigma^2} = 1\}$ contains all points (x, y) with $(y - x'\beta_0)^2 \leq c\sigma_0^2 - \kappa$, $\kappa > 0$, for which

$$\min\left(\frac{\kappa}{2c}, \sqrt{\frac{\kappa}{4\|x\|^2}}, \frac{\kappa}{8\sqrt{c}\sigma_0\|x\|}\right) \geq \epsilon, \quad (11.20)$$

since then $c\sigma^2 \geq c\sigma_0^2 - \frac{\kappa}{2}$ and

$$(y - x'\beta)^2 = (y - x'\beta_0 + x'(\beta_0 - \beta))^2 \leq (y - x'\beta_0)^2 + 2\sqrt{c}\sigma_0\|x\|\epsilon + \|x\|^2\epsilon^2 \leq (y - x'\beta_0)^2 + \frac{\kappa}{2}.$$

Hence with $\epsilon \searrow 0$,

$$\bigcap_{(\beta,\sigma^2) \in B_\epsilon(\beta_0, \sigma_0^2)} \{w_{\beta,\sigma^2} = 1\} \nearrow \{(x, y) : \; (y - x'\beta_0)^2 < c\sigma_0^2\},$$

therefore (11.18).

11.6 Proof of Theorem 6.1

Observe under $\sigma_1^2 = 0$

$$\sigma_P^2(\beta_1, \sigma^2) = 0 < \sigma_P^2(\beta, \sigma^2) \; \forall \sigma^2, \beta \neq \beta_1,$$

i.e., $(\beta_1, 0)$ is the only fixed point of f_P. In the following $\sigma_1^2 > 0$. Because of the equivariance properties of LS-FPCI and LS-FPCV from Remark 3.4, assume w.l.o.g. $\beta_1 = 0, \sigma_1^2 = 1$, i.e., x and y are stochastically independent under P. The proof proceeds as follows[11]:

Step 1: h has a unique zero.

[11] This is the proof of Theorem 12.1 of Hennig (1997), shortened here.

Step 2: w_{0,σ^2} is FPCI w.r.t. P iff $\sigma^2 = k$.

Step 3: If $\beta \neq 0$, w_{β,σ^2} is not FPCI w.r.t. P.

Proof of step 1: Use

$$h(s^2) = 0 \Leftrightarrow h_0(s) := (1-s^2)[\Phi(\sqrt{c}s) - \Phi(-\sqrt{c}s)] - 2\sqrt{c}s\varphi(\sqrt{c}s) = 0.$$

Observe $s \geq 1 \Rightarrow h_0(s) < 0,\ h_0(0) = 0$, and show $h_0(s) > 0$ in some neighborhood of 0:

$$h_0'(s) = 2s[\sqrt{c}s(c-1)\varphi(\sqrt{c}s) - (\Phi(\sqrt{c}s) - \Phi(-\sqrt{c}s))] >$$
$$> 2\sqrt{c}s^2[(c-1)\varphi(\sqrt{c}s) - 2\varphi(0)] > 0$$

in some neighborhood of 0 since $c > 3,\ (c-1)\varphi(\sqrt{c}s) - 2\varphi(0) > 0$. Continuity of h_0 ensures the existence of some zero argument > 0.

To show uniqueness of this zero, use

$$h_0'(s) = 0 \Leftrightarrow h_1(s) := \sqrt{c}s(c-1)\varphi(\sqrt{c}s) - (\Phi(\sqrt{c}s) - \Phi(-\sqrt{c}s)) = 0,$$
$$\lim_{s\to\infty} h_1(s) = -1.$$

Notice $h_1(0) = 0$. h_1 has the same sign as h_0' for all positive arguments. Calculate

$$h_1'(s) = \sqrt{c}\varphi(\sqrt{c}s)\left[(c-1)(1-cs^2) - 2\right],$$
$$h_1'(0) = \sqrt{c}\varphi(0)(c-3) > 0.$$

$h_1'(s) < 0$ iff $0 > (c-1)(1-cs^2) - 2 < 0$, which is strictly monotone decreasing in s^2. Thus, h_1' has a unique zero s_2, which is local maximum of h_1, $h_1(s_2) > 0$. h_1 decreases strictly monotone for $s > s_2$ and must have some unique zero, which is the unique local extremum of h_0. Thus, h_0 can only have a unique zero.

Proof of step 2: If $\sigma^2 = 0$, then $Pw_{\beta,\sigma^2} = 0$ and w_{β,σ^2} cannot be LS-FPCI because of (3.1). Consider $\sigma^2 > 0$. $\beta_P(0,\sigma^2) = 0$ because of Proposition 11.1.

$$\sigma_P^2(0,\sigma^2) = \frac{\mathcal{N}y^2 1[y^2 \leq c\sigma^2]}{\mathcal{N}1[y^2 \leq c\sigma^2]} = 1 - \frac{2\sqrt{c\sigma^2}\varphi(\sqrt{c\sigma^2})}{\Phi(\sqrt{c\sigma^2}) - \Phi(-\sqrt{c\sigma^2})},$$

i.e., $(0,\sigma^2)$ is fixed point of f_P iff $\sigma^2 = k$.

Proof of step 3: Suppose that w_{β,σ^2} with $\beta \neq 0$ is FPCI w.r.t. P. $\sigma^2 = 0$ is impossible since $P\{w_{\beta,0} = 1\} = 0$. Define

$$F_\beta(t) := P(y - x't)^2 1\left((y - x'\beta)^2 \leq cs^2\right),$$

i.e., $\beta_P(\beta,\sigma^2) = \arg\min_t F_\beta(t)$. With $v := \frac{\beta}{\|\beta\|}$ get

$$\tfrac{\partial}{\partial v}F_\beta(t) = -2\sum_{i=1}^p v_i P x_i (y - x't) 1\left((y - x'\beta)^2 \leq c\sigma^2\right),$$
$$\tfrac{\partial}{\partial v}F_\beta(\beta) = -\tfrac{2}{\|\beta\|}G[x'\beta J(x'\beta)], \text{ where}$$
$$J(u) := \mathcal{N}(y-u)1\left((y-u)^2 \leq c_0^2\right), \quad c_0 := \sqrt{cs^2}.$$

Show $uJ(u) < 0$ for $u \neq 0$:

$$\begin{gathered} uJ(u) = \int u(y-u)1[|y-u| \leq c_0]\varphi(y)dy = \\ = \int u|y-u|1[|y-u| \leq c_0](1[y>u] - 1[y<u])\varphi(y)dy = \\ = \int u|t|1[0 < t \leq c_0](\varphi(t+u) - \varphi(-t+u))dt = \\ = \int |u||t|1[0 < t \leq c_0](\varphi(t+|u|) - \varphi(-t+|u|))dt, \\ \text{since } \varphi(t+u) = \varphi(-t+|u|), \quad \varphi(-t+u) = \varphi(t+|u|) \end{gathered}$$

for $u < 0$. Get $uJ(u) < 0$ by

$$t > 0, w > 0 \Rightarrow wt[\varphi(t+w) - \varphi(-t+w)] < 0.$$

Since Gxx' was supposed to be invertible (see 4.2), $G\{x'\beta = 0\} < 1$. That is, $\frac{\partial}{\partial v}F_\beta(\beta) > 0$, and $\beta \neq \beta_P(\beta, \sigma^2)$. The proof is completed.

11.7 Preparations for the proof of Lemma 6.3

Theorem 11.16 *Let K be some compact convex subset of $I\!R^p$, $C^1(K) \ni f = (f_1, \dots, f_q) : K \mapsto I\!R^q$. Then,*

$$\begin{gathered} \forall x, y \in K : \ \|f(x) - f(y)\| \leq \|df\|_K \|x-y\|, \ \textit{where} \\ \|df\|_K := \sup_{x \in K} \left(\max_{i=1,\dots,q} \sum_{j=1}^{p} \left| \left[\frac{\partial}{\partial x_j} f_i \right](x) \right| \right). \end{gathered}$$

Proof e.g. in Königsberger (1993).

Proposition 11.17 *Let $h_1 : I\!R^{p+1} \mapsto I\!R$ be continuous, where h_1, $h_{1i}(x) := x_i h_1(x)$ $G-$integrable for $i = 1, \dots, p+1$, $h_2 : I\!R \mapsto I\!R$ be continuous, $\mathcal{N}$-integrable and $\exists y_0 < \infty : \ |h_2\varphi| \leq y_0$. Then,*

$$l : I\!R^{p+2} \mapsto I\!R, \ l(a,b) := \int h_1(x)h_2(y)1[(y-x'b)^2 \leq a^2]\varphi(y)d[\lambda(y) \otimes G(x)]$$

is continuously differentiable.

Proof:

$$l(a,b) = Gj(a,b,x), \ \text{where } j(a,b,x) := \int_{-a+x'b}^{a+x'b} h_1(x)h_2(y)\varphi(y)dy.$$

By continuity of h and φ get

$$\begin{gathered} \tfrac{\partial}{\partial a}j(a,b,x) = h_1(x)[h_2(-a+x'b)\varphi(-a+x'b) + h_2(a+x'b)\varphi(a+x'b)], \\ i = 1, \dots, p+1 : \ \tfrac{\partial}{\partial b_i}j(a,b,x) = \\ = x_i[h_1(x)h_2(a+x'b)\varphi(a+x'b) - h_1(x)h_2(-a+x'b)\varphi(-a+x'b)]. \end{gathered}$$

The derivative w.r.t. a can be bounded $\forall a, b$ by $|h_1|y_0$, the derivatives w.r.t. b_i by $2|h_{1i}|y_0$. The bounds are $G-$integrable as well as $|j(a,b,\bullet)|$ $\forall a, b$, which can be bounded by $|h_1||\int h_2 d\mathcal{N}|$. Thus, integration and differentiation are exchangeable at l and all partial derivatives are continuous by continuity of h_1, h_2 and φ. The proposition follows.

11.8 Proof of Lemma 6.3

Because of the equivariance properties of FPCI and FPCV from the Remarks 3.4 and 5.4, assume w.l.o.g. $\beta_1 = 0, \sigma_1^2 = 1$, i.e., x and y are stochastically independent under $P_{0,1,G}$. Consider (6.3) and $\mathcal{L}(y) = \mathcal{N}_{(0,1)}$ under $P_{0,1,G}$, and get (5.1) by continuity of $\mathcal{L}(y)$, (5.2) by (6.2), and (5.3) by (6.2). $w_{0,k}$ is LS-FPCI w.r.t. P and k is uniquely defined by Corollary 6.2, which requires $P^*\{y^2 \leq ck\} = 0$ by (6.3), thus (5.5) is fulfilled. Show (5.7) by application of Theorem 11.16 and Proposition 11.17:

To prove (5.7), it suffices to have $\|df_P\|_{B_{\epsilon_0}(0,k)} =: \alpha < 1$ for some $\epsilon_0 > 0$. Complete the proof by showing

$$f_P \text{ is continuously differentiable in } (0,k), \tag{11.21}$$

$$\left(\max_{i=1,\dots,p+2} \sum_{j=1}^{p+2} \left| \left[\frac{\partial}{\partial x_j} f_{Pi} \right] (0,k) \right| \right) < 1. \tag{11.22}$$

At first, observe that $1[(y - x'\beta)^2 \leq c\sigma^2] dP^* \equiv 0$ for $(\beta, \sigma^2) \in B_{\epsilon_1}(0,k)$, therefore $f_P|_{B_{\epsilon_1}(0,k)} = f_{P_{0,1,G}}|_{B_{\epsilon_1}(0,k)}$. Apply Proposition 11.17 and get explicitly

$$\tfrac{\partial}{\partial a} l(a,b) = G h_1(x)[h_2(-a+x'b)\varphi(-a+x'b) + h_2(a+x'b)\varphi(a+x'b)], \tag{11.23}$$

$$i = 1, \dots, p+1 : \; \tfrac{\partial}{\partial b_i} l(a,b) =$$

$$= G x_i [h_1(x) h_2(a+x'b)\varphi(a+x'b) - h_1(x) h_2(-a+x'b)\varphi(-a+x'b)]. \tag{11.24}$$

Now compute the partial derivatives of $f_{P_{0,1,G}}$ in $(0,k)$. $\beta_{P_{0,1,G}}(\beta, \sigma^2)$ is defined by

$$F(\beta, \sigma^2, \beta_{P_{0,1,G}}(\beta, \sigma^2)) = 0, \text{ where}$$

$$F(\beta, \sigma^2, t) := (P_{0,1,G} x x' 1[(y - x'\beta)^2 \leq c\sigma^2]) t - P_{0,1,G} x y 1[(y - x'\beta)^2 \leq c\sigma^2].$$

By Proposition 11.17, F is continuously differentiable w.r.t. β, σ^2, t; the $G-$integrability-conditions follow from (6.2). If $\sigma^2 > 0$ then $P_{0,1,G}\{w_{\beta,\sigma^2} = 1\} > 0$ for arbitrary β. $\frac{\partial}{\partial t} F(\beta, \sigma^2, t) = P_{0,1,G} x x' 1[(y - x'\beta)^2 \leq c\sigma^2]$ is invertible by (6.2), $\beta_{P_{0,1,G}}$ is continuous at β, σ^2 (Proposition 11.2). Thus, by differentiation of implicit functions,

$$\frac{\partial}{\partial(\beta, \sigma^2)} \beta_{P_{0,1,G}}(\beta, \sigma^2) = -\left(\frac{\partial F(\beta, \sigma^2, t)}{\partial t} \right)^{-1} \left(\frac{\partial F(\beta, \sigma^2, t)}{\partial(\beta, \sigma^2)} \right) \Bigg|_{t = \beta_{P_{0,1,G}}(\beta, \sigma^2)} \tag{11.25}$$

continuously in some neighborhood of $(0,k)$. Notice $\beta_{P_{0,1,G}}(0,k) = 0$ by step 2 of the proof of Theorem 6.1. Evaluate with help of (11.23), (11.24)

$$\tfrac{\partial F(\beta,\sigma^2,t)}{\partial t}\Big|_{(\beta,\sigma^2,t)=(0,k,0)} = [\Phi(\sqrt{ck}) - \Phi(\sqrt{-ck})] G x x',$$

$$\tfrac{\partial F(\beta,\sigma^2,t)}{\partial \sigma^2}\Big|_{(\beta,\sigma^2,t)=(0,k,0)} = -\tfrac{\partial}{\partial \sigma^2} P_{0,1,G} x y 1[(y - x'\beta)^2 \leq c\sigma^2]\Big|_{(\beta,\sigma^2)=(0,k)} =$$

$$= \tfrac{\partial}{\partial a} l(a,b)\Big|_{(a,b)=(\sqrt{ck},0)} \tfrac{\sqrt{c}}{2\sqrt{k}} =$$

$$= \tfrac{\sqrt{c}}{2\sqrt{k}} [G x(-\sqrt{ck})\varphi(-\sqrt{ck}) + G x(\sqrt{ck})\varphi(\sqrt{ck})] = 0,$$

$$\tfrac{\partial F(\beta,\sigma^2,t)}{\partial \beta}\Big|_{(\beta,\sigma^2,t)=(0,k,0)} = -\tfrac{\partial}{\partial \beta} P_{0,1,G} x y 1[(y - x'\beta)^2 \leq c\sigma^2]\Big|_{(\beta,\sigma^2)=(0,k)} =$$

$$= -\Big(G x_i [x\sqrt{ck}\varphi(\sqrt{ck}) - x(-\sqrt{ck})\varphi(-\sqrt{ck})] \Big)_{i=1,\dots,p+1} =$$

$$= -2\sqrt{ck}\varphi(\sqrt{ck}) G x x' \Rightarrow$$

$$\Rightarrow \tfrac{\partial}{\partial \beta} \beta_{P_{0,1,G}}(\beta, \sigma^2)\Big|_{(\beta,\sigma^2)=(0,k)} = \tfrac{2\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck}) - \Phi(\sqrt{-ck})} \mathbf{I}_{p+1}, \quad \tfrac{\partial}{\partial \sigma^2} \beta_{P_{0,1,G}}(\beta, \sigma^2)\Big|_{(\beta,\sigma^2)=(0,k)} = 0.$$

Notice $1 > 1 - E_{\mathcal{N}}(y^2|y^2 \leq ck) = \frac{2\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck})-\Phi(\sqrt{-ck})} > 0$ (Hennig (1997), Proposition 11.7). Now consider

$$\begin{gathered}\tfrac{\partial}{\partial(\beta,\sigma^2)}\sigma^2_{P_{0,1,G}}(\beta,\sigma^2) = \tfrac{\partial}{\partial(a,b_0,b_1)}l_0(a,b_0,b_1)\Big|_{(a,b_0,b_1)=r(\beta,\sigma^2)}\tfrac{\partial}{\partial(\beta,\sigma^2)}r(\beta,\sigma^2),\ \text{where}\\ r(\beta,\sigma^2) := (\sqrt{c\sigma^2},\beta,\beta_{P_{0,1,G}}(\beta,\sigma^2)), l_0(a,b_0,b_1) := \tfrac{l_1(a,b_0,b_1)}{l_2(a,b_0)},\\ l_2(a,b_0) := P_{0,1,G}\{(y-x'b_0)^2 \leq a^2\},\\ l_1(a,b_0,b_1) := P_{0,1,G}(y-x'b_1)^2 1[(y-x'b_0)^2 \leq a^2],\ (a,b_0,b_1) \in I\!R \times I\!R^{p+1} \times I\!R^{p+1}.\end{gathered}$$

l_1 and therefore l_0 are also continuously differentiable if $b_1 \in B_\eta(0)$, $\eta < \infty$: Continuous differentiability w.r.t. (a, b_0) follows from Proposition 11.17, Exchangeability of integration and differentiation w.r.t. b_1 holds with $j_0(a,b_0,b_1,x) := \int_{-a+x'b_0}^{a+x'b_0}(y-x'b_1)^2\varphi(y)dy$ since

$$\left\|\frac{\partial}{\partial b_1}j_0(a,b_0,b_1,x)\right\| = \left\|-2x\int_{-a+x'b_0}^{a+x'b_0}(y-x'b_1)\varphi(y)dy\right\| \leq 2(\|x\|E_{\mathcal{N}}|y| + \eta\|x\|^2),$$

which is $G-$integrable as well as $j_0(a,b_0,b_1,\bullet)$ $\forall(a,b_0,b_1)$. $l_2(a,b) > 0$ is continuously differentiable by Proposition 11.17 as well as r. Together with (11.25) get (11.21).

Observe by symmetry considerations $\sigma^2_{P_{0,1,G}}(\beta,\sigma^2) = \sigma^2_{P_{0,1,G}}(-\beta,\sigma^2)$, thus $\frac{\partial}{\partial\beta}\sigma^2_{P_{0,1,G}}(\beta,\sigma^2)\Big|_{(\beta,\sigma^2)=(0,k)} = 0$. Use for $a > 0$

$$\begin{gathered}l_1(a,0,0) = \Phi(a) - \Phi(-a) - 2a\varphi(a),\quad \tfrac{\partial}{\partial a}l_1(a,0,0) = 2a^2\varphi(a),\\ l_2(a,0) = \Phi(a) - \Phi(-a),\quad \tfrac{\partial}{\partial a}l_2(a,0) = 2\varphi(a).\end{gathered}$$

to compute

$$\begin{gathered}\tfrac{\partial}{\partial\sigma^2}\sigma^2_{P_{0,1,G}}(\beta,\sigma^2)\Big|_{(\beta,\sigma^2)=(0,k)} =\\ \tfrac{\partial}{\partial(a,b_0,b_1)}l_0(a,b_0,b_1)\Big|_{(a,b_0,b_1)=r(0,k)}\begin{pmatrix}\frac{\sqrt{c}}{2\sqrt{k}}\\ 0\\ \frac{\partial}{\partial\sigma^2}\beta_{P_{0,1,G}}(\beta,\sigma^2)\Big|_{(\beta,\sigma^2)=(0,k)}\end{pmatrix} =\\ = \tfrac{\sqrt{c}}{2\sqrt{k}}\ \tfrac{\partial}{\partial a}l_0(a,b_0,b_1)\Big|_{(a,b_0,b_1)=(\sqrt{ck},0,0)} =\\ = \frac{\sqrt{c^3k}\varphi(\sqrt{ck})[\Phi(\sqrt{ck})-\Phi(-\sqrt{ck})]-[\Phi(\sqrt{ck})-\Phi(-\sqrt{ck})-2\sqrt{ck}\varphi(\sqrt{ck})]\frac{\sqrt{c}}{\sqrt{k}}\varphi(\sqrt{ck})}{[\Phi(\sqrt{ck})-\Phi(-\sqrt{ck})]^2} =\\ = \left(c - \tfrac{1}{k} + \tfrac{2\sqrt{ck}\varphi(\sqrt{ck})}{k[\Phi(\sqrt{ck})-\Phi(-\sqrt{ck})]}\right)\tfrac{\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck})-\Phi(-\sqrt{ck})} = (c-1)\tfrac{1-k}{2}\end{gathered}$$

by definition of k. Get $|\frac{\partial}{\partial\sigma^2}\sigma^2_{P_{0,1,G}}(\beta,\sigma^2)\Big|_{(\beta,\sigma^2)=(0,k)}| < 1$ by (6.4).

Altogether, $\|df_P\|_{B_{\epsilon_0}(0,k)} \leq \max\left(\frac{2\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck})-\Phi(-\sqrt{ck})}, (c-1)\frac{1-k}{2}\right) < 1$ proving (11.22). This completes the proof.

References

Azzalini, A. and Bowman, A. W. (1990): A look at some data on the Old Faithful geyser, *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 39, p. 357-365.

Banfield, J. D. and Raftery, A. E. (1993): Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics* 49, p. 803-821.

Barnett, V. and Lewis, T. (1994): *Outliers in Statistical Data (3rd Ed.),* Wiley, Chichester.

Boscher, H. (1992): *Behandlung von Ausreißern in linearen Regressionsmodellen,* dissertation, Universität Dortmund.

Bozdogan, H. (1987): Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions, *Psychometrika,* 52, p. 345-370.

Byers, S. and Raftery, A. E. (1998): Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes, *Journal of the American Statistical Association,* 93, p. 577-584.

Collatz, L. (1966): *Functional Analysis and Numerical Mathematics,* Academic Press, New York.

Cook, R. D. and Weisberg, S. (1982): *Residuals and Influence in Regression,* Chapman and Hall, London.

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

DasGupta, A. and Raftery, A. E. (1998): Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering, *Journal of the American Statistical Association,* 93, p. 294-302.

Davies, P. L. (1988): Consistent Estimates for Finite Mixtures of Well Separated Elliptical Distributions in Bock, H.-H. (Ed.): *Classification and Related Methods of Data Analysis,* Elsevier Science Publishers, Amsterdam, p. 195-202.

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

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

Garcia-Escudero, L. A., and Gordaliza, A. (1999): Robustness Properties of k Means and Trimmed k Means, *Journal of the American Statistical Association,* 94, 956-969.

Hampel, F. R., Ronchetto, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986): *Robust Statistics. The Approach Based on Influence Functions,* Wiley, New York.

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

Hennig, C. (1998): Clustering and Outlier Identification: Fixed Point Cluster Analysis in Rizzi, A., Vichi, M., and Bock, H.-H.. (Eds.): *Advances in Data Science and Classification,* Springer, Berlin, p. 37-42.

Hennig, C. (1999): Models and Methods for Clusterwise Linear Regression in Gaul, W. and Locarek-Junge, H. (Eds.): *Classification in the Information Age,* Springer, Berlin, p. 179-187.

Hennig, C. (2000): What Clusters are Generated by Normal Mixtures? To appear in the Proceedings of IFCS-2000, Namur.

Hosmer, D. W. jr. (1974): Maximum Likelihood Estimates of the Parameters of a Mixture of Two Regression Lines, *Communications in Statistics,* 3, p. 995-1006.

Huber, P. J. (1981): *Robust Statistics,* Wiley, New York.

Königsberger, K. (1993): *Analysis 2,* Springer, Berlin.

Morgenthaler, S. (1990): Fitting redescending M-estimators in regression in Lawrence, H. D. and Arthur, S. (Eds.): *Robust Regression* , Dekker, New York, p. 105-128.

Pollard, D. (1984): *Convergence of Stochastic Processes,* Springer, New York.

Rousseeuw, P. J. (1994): Unconventional features of positive-breakdown estimators, *Statistics and Probability Letters,* 19, p. 417-431.

Schwarz, G. (1978): Estimating the dimension of a model, *Annals of Statistics* 6, p. 461-464.

van der Vaart, A. W. and Wellner, J. A. (1996): *Weak Convergence and Empirical Processes,* Springer, New York.

Wedel, M. and DeSarbo, W. S. (1995): A mixture likelihood approach for generalized linear models, *Journal of Classification,* 12, p. 21-56.