Journal of Classification 17:273-296 (2000)
DOI: 10.1007/s003570000022

Identifiability of Models for Clusterwise Linear Regression

Christian Hennig

Universität Hamburg

Abstract: Identifiability of the parameters is a necessary condition for the existence of consistent estimators. In this paper the identifiability of the parameters of models for data generated by different linear regression distributions with Gaussian errors is investigated. It turns out that such models cause other identifiability problems than do simple Gaussian mixtures. This problem was heretofore ignored; thus there are no satisfying consistency proofs in this area. Three different models are treated: Finite mixture models with random and fixed covariates and a fixed partition model. Counterexamples and sufficient conditions for identifiability are given, including an example for nonidentifiable parameters with an invertible information matrix.

The model choice and the interpretation of the parameters are discussed as well as the use of the identifiability concept for fixed partition models. The concept is generalized to "partial identifiability".

Keywords: Partial identifiability; Switching regression; Mixture model; Fixed partition model; Change point problem; Gaussian mixtures with covariates

Author's Address: Christian Hennig, Universität Hamburg, Fachbereich Mathematik - SPST, Bundesstr. 55, D-20146 Hamburg, Germany; e-mail: hennig@math.uni-hamburg.de

1. Introduction

This paper treats the problem of estimating the parameters of linear regression models, where different subsets ("clusters") of the entities corresponding to the data set (henceforth, "data points") follow different linear relations between a covariate $\mathbf{x}$ and a dependent variable Y, and the cluster membership of the data points is unknown. That is, I assume for each cluster a linear regression distribution of the form

$$\begin{aligned} &Y = \mathbf{x}'\boldsymbol{\beta} + U, \qquad \mathcal{L}(U) = \mathcal{N}_{0,\sigma^2} \text{ i.i.d.}, \\ &Y\ I\!R - \text{valued r.v.}, \qquad \mathbf{x} = (x_1, \ldots, x_p, 1) \in I\!R^p \times \{1\}, \\ &\boldsymbol{\beta} \in I\!R^{p+1},\ \sigma^2 \in I\!R_0^+,\ i \in I. \end{aligned} \tag{1}$$

The $p + 1$st component of $\boldsymbol{\beta}$ denotes the intercept parameter. $\mathcal{L}(U) = \mathcal{N}_{0,\sigma^2}$ means that U is distributed according to the Gaussian distribution with mean 0 and variance σ^2. I is some index set. Distinct regression and scale parameters $(\boldsymbol{\beta}, \sigma^2)$ define distinct clusters. $\mathbf{x}$ and Y are observable, but the parameters and cluster memberships are not. Note that two linear regression distributions with distinct parameters do not necessarily lead to well separated observations. I somewhat informally use the term "cluster" because the mixture and fixed partition model have in common that they are used in partitioning cluster analysis. This practice should avoid confusion between "mixture components" (here: "clusters") and "components" of p-dimensional vectors. The Gaussian distribution is considered as the distribution of U because it is the most familiar for this purpose. All results carry over to arbitrary univariate location-scale families that generate identifiable mixtures.

Examples can be found, e.g., in biology and economy:

Example 1.1.

1. Animals or plants can sometimes be grouped according to relationships between their properties. For example, male and female halibuts can be divided by considering the relationship between age and length (Hosmer 1974).

2. Seber and Wild (1989, p. 435 ff.) give biological examples for situations where a relation, e.g., between quantity of fertilizer and yield of corn, changes at some time or quantity.

3. In marketing consumers or suppliers rate the quality of products or events. Markets can be segmented by finding groups with respect to the relation between the rating and the features of the product (DeSarbo and Cron 1988; Kamarkura 1988; Wedel and Steenkamp 1991). Other economical applications can be found in Fair and Jaffee (1972) and Quandt and Ramsey (1978).

The paper is not on particular estimators and their consistency, but instead it starts one step earlier: What can be estimated consistently? I doubt that there is presently any valid consistency proof for any estimator at any of the models treated here, and I try to make the reasons clearer. The model choice and the meaning of the identifiability concept are discussed, with some hopefully surprising findings.

For the whole data set there are various models, of which the mixture model with fixed covariates is the most popular:

Model 1.

$$\mathcal{L}\left((Y_i)_{i\in I}\right) = \bigotimes_{i\in I} F_{\mathbf{x}_i,J}, \text{ where}$$

$$F_{\mathbf{x},J}(y) = \textstyle\int_{T_1} \Phi_{0,\sigma^2}(y - \mathbf{x}'\boldsymbol{\beta})dJ(\boldsymbol{\beta},\sigma^2), \quad T_1 := \mathbb{R}^{p+1} \times \mathbb{R}_0^+,$$
$$J \in \Omega_1 := \mathcal{J}(T_1).$$

Φ_{a,σ^2} stands for the cumulative distribution function (cdf) of the Gaussian distribution with mean a and variance σ^2. φ_{a,σ^2} is the corresponding density. All other distributions are denoted with the same letter as their cdf. $\mathcal{J}(T)$ denotes the set of mixing distributions with finite support on the parameter set T. $S(J)$ is the support set of $J \in \mathcal{J}(T)$. Thus, $s := |S(J)|$ is the number of mixture components, informally called "clusters" here. That is, the members of $\mathcal{J}(T)$ are distributions generating parameter values $(\boldsymbol{\beta}_1,\sigma_1^2),\ldots,(\boldsymbol{\beta}_s,\sigma_s^2)$ for s clusters with probabilities $J(\boldsymbol{\beta}_1,\sigma_1^2),\ldots,J(\boldsymbol{\beta}_s,\sigma_s^2)$. I is some index set, e.g., $I = \{1,\ldots,n\}$ if there are n observations with distinct covariates $\mathbf{x}_i$. "$\bigotimes$" denotes the independent product of distributions, i.e., the observations are modeled as independent in Model 1. For each covariate point $\mathbf{x}_i$, modeled as nonrandom, $F_{\mathbf{x}_i,J}$ is a finite mixture of one-dimensional Gaussian distributions with means $\mathbf{x}'\boldsymbol{\beta}_1,\ldots,\mathbf{x}'\boldsymbol{\beta}_s$ and variances $\sigma_1^2,\ldots,\sigma_s^2$.

Before I discuss Model 1, I introduce some further notation. The letter $\mathbf{x}$ is always used for the covariates, which are $p+1$-dimensional points with $(p+1)$st component equal to 1, i.e., the $(p+1)$st component of a regression parameter denotes the intercept and $\mathbf{x}^-$ denotes the first p components of $\mathbf{x}$. $\tilde{\mathbf{X}}$ is used for the sequence of all covariate values $(\mathbf{x}_i)_{i\in I}$. This notation implies that "$\dim\langle\{\mathbf{x}_i : i \in I\}\rangle < p+1$" is equivalent to "$\mathbf{x}_i^-, i \in I$ lie on a common $(p-1)$-dimensional hyperplane $H := \{\mathbf{x}^- \in \mathbb{R}^p : \boldsymbol{\alpha}'\mathbf{x}^- = a\}$, $\mathbb{R}^p \ni \boldsymbol{\alpha} \neq 0$", where $\langle A\rangle$ denotes the linear hull of A. $\mathcal{H}_{p-1}$ denotes the space of $(p-1)$-dimensional hyperplanes of $\mathbb{R}^p$.

"$\mathbf{t} \leq \mathbf{x}$", where $\mathbf{t},\mathbf{x}$ are d-dimensional vectors, means component-wise $\leq$. $\mathcal{P}_d$ denotes the space of distributions on $\mathbb{R}^d$.

Random variables are usually denoted by capitals. If $\mathcal{L}(\mathbf{Z}) = G$ for some random variable $\mathbf{Z} = (\mathbf{Z}_1,\ldots,\mathbf{Z}_k)$, $G^{\mathbf{Z}_1}$ stands for the marginal distribution of

$\mathbf{Z}_1$ under G. Sometimes I write $G^{\mathbf{Z}}$ for G to indicate the corresponding random variable.

Two further models will be introduced in later sections. Ω_i always denotes the parameter set corresponding to the Model i for linear regression clusters. It defines a family of the parameterized distributions $C_i = (F_\omega)_{\omega \in \Omega_i}$. If needed, T_i is the corresponding parameter set for a single linear regression. Some equivalence relations on Ω_i will be considered. They are denoted by "$\sim_{ij}$", where i is the model number and $j = 0, 1, 2$ distinguishes between different relations on Ω_i.

Returning to Model 1: The interest lies in the estimation of the mixing distribution J, i.e., s and all parameter vectors and cluster probabilities. For fixed s there are some proposals, e.g., Maximum Likelihood (ML) estimation (Hosmer 1974; DeSarbo and Cron 1988) or estimation via the Moment Generating Function (MGF, Quandt and Ramsey 1978). These estimators are believed to be consistent, at least if the σ_i^2 are bounded away from 0, but this paper shows that this claim is not in general true.

Only identifiable parameters can be estimated consistently. "Identifiability" means that, knowing the data distribution $\mathcal{L}\left((Y_i)_{i\in I}\right)$, one can identify uniquely the mixing distribution J. That is, no two distinct sets of parameters $(\boldsymbol{\beta}_{1i}, \sigma_{1i}^2, J(\boldsymbol{\beta}_{1i}, \sigma_{1i}^2)), \ldots, (\boldsymbol{\beta}_{si}, \sigma_{si}^2, J(\boldsymbol{\beta}_{si}, \sigma_{si}^2))$, $i = 1, 2$, lead to the same data distribution.

I do not know of any consistency proof for Model 1 that takes the question of identifiability adequately into account, presumably because it is believed that identifiability for linear regression mixtures with Gaussian errors follows directly from the identifiability of Gaussian mixtures (proven by Yakowitz and Spragins 1968). DeSarbo and Cron (1988, p. 255) make that claim explicitly. I discuss Model 1 and the various consistency proofs in Section 2 and give identifiability conditions and a counterexample. The example has the interesting property that the inverse of the Fisher information matrix of the nonidentifiable parameters exists and disproves the belief that an invertible information matrix implies consistent ML-estimation. The counterexample may seem somewhat pathological, and indeed there are sufficient conditions for identifiability that presumably hold in most applications – but not always (see the discussion in the Conclusion). Furthermore, I give arguments for the belief that the identifiability problems can lead to serious complications for consistency proofs.

Model 1 seems inadequate for most of the applications because it assumes "assignment independence": The probability for a point to be generated by one of the s cluster distributions has to be the same for all covariate values $\mathbf{x}$; the assignment of the data points to the clusters has to be independent of the covariates as illustrated by Figure 1. For change point problems like Example 1.1.2, the opposite is true: The covariate value *determines* the cluster mem-

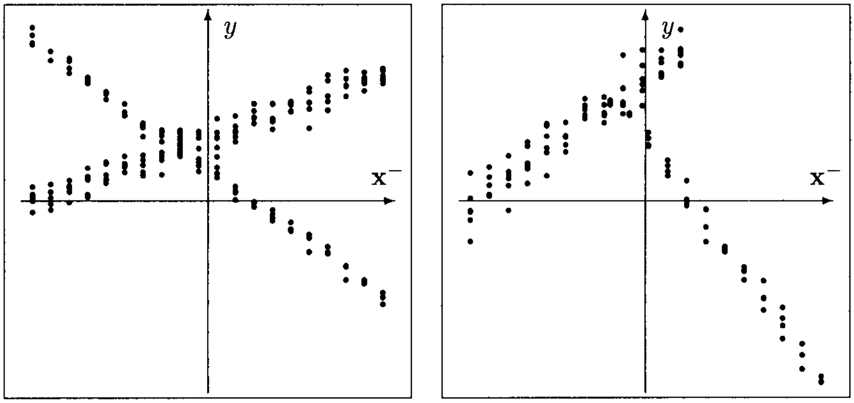

Figure 1: In Figure 1a two regression clusters are differentiated by the slope. Their x-values are distributed similarly and give no information about the cluster membership. In Figure 1b the distribution of x is quite different for the two clusters. These cases are referred to as "assignment independence" and "assignment dependence".

bership. But also intermediate situations are possible – and in my experience likely – e.g., if the lifespan distribution for male and female halibuts would not be equal in Example 1.1.1. In Example 1.1.3 assignment independence would be fulfilled if all products would be rated by the same proportions of members of the market segments, i.e., only if the product choice would be independent of the segment.

There are two reasonable models for linear regression clusters that do not assume assignment independence. One strategy replaces the fixed covariates by covariate distributions that are allowed to differ between the clusters, as in Section 3. Alternatively, one can use a fixed partition model where the cluster membership of the points is not random but explicitly parameterized. This approach was first proposed by Fair and Jaffee (1972) and is discussed in Section 4. Consistency gets even more cumbersome here; at least the ML-estimator is known to be inconsistent (Oberhofer 1980).

Up to now there are no consistent estimators for these models, and the investigation of identifiability leads to some new difficulties. In both cases there are additional parameters that are not identifiable in general, namely the covariate distributions, respectively the membership parameters. Concerning the question of estimability, it would be interesting if one could identify *some* of the parameters, e.g., only regression and scale.

For this purpose I introduce a more general concept of identifiability. Given a parameterized family of distributions C, one can define an equivalence relation "$\sim$" on the parameter set Ω so that two parameters are equivalent if the components to be estimated are equal. "Identifiability" then means that the distribution determines the *equivalence class* of the parameters. For the usual identifiability problem for mixture distributions J_1 is treated as equivalent to J_2 if $J_1 = J_2$. That is, the equivalence classes correspond to the full parameters, and the distribution is also determined by the equivalence class of parameters (corresponding to the "$\Leftarrow$"-direction of the equivalence in the definitions below).

If only a part of the parameter is to be estimated, the appropriate equivalence classes do not determine the distribution. Examples appear in the later sections. The identifiability of the interesting parameter part is then called "partial identifiability".

Definition 1.2. *Let Ω be an arbitrary parameter space, $\mathcal{P}$ be some space of distributions,*

$$C := (F_\omega)_{\omega \in \Omega} \in \mathcal{P}^\Omega, \tag{2}$$

and "$\sim$" be an equivalence relation on Ω. Then C is identifiable with respect to "$\sim$" if

$$\forall \omega, \hat{\omega} \in \Omega : \qquad F_\omega = F_{\hat{\omega}} \Leftrightarrow \omega \sim \hat{\omega}.$$

Definition 1.3. *C is called partially identifiable w.r.t. "$\sim$" if*

$$\forall \omega, \hat{\omega} \in \Omega : \qquad F_\omega = F_{\hat{\omega}} \Rightarrow \omega \sim \hat{\omega}.$$

To my knowledge, there is no previous treatment of identifiability of the parameters of any fixed partition model, even in the simple Gaussian case. The interpretation of the concept is less obvious in this setup than for mixture models and is discussed in Section 4.

In the context of mixture models the concept of identifiability goes back to Teicher (1961). Some familiar classes of distributions have been shown to generate identifiable mixtures, e.g., multivariate Gaussian distributions (Yakowitz and Spragins 1968). Binomial and uniform distributions are examples for the opposite (Titterington, Smith, and Makov, 1985). More results on identifiability of mixture distributions can be found in Chandra (1977), Prakasa Rao (1992) and Lindsay (1995). Prakasa Rao (1992, p. 149) defines "partial identifiability" for identifiability problems apart from mixture parameters, but his interpretation is analogous. Lindsay (1995, p. 44) uses the term informally. Li and Sedransk (1988) generalize the concept of identifiability for finite mixtures in other directions. Yakowitz (1969) and Chen (1995) prove the existence of consistent estimators for identifiable mixture models under certain further assumptions. A reviewer has also recommended Gordon (1990) and Feng and McCulloch (1996).

Redner (1981) appears to be the first to introduce equivalence classes in relation with identifiability. His context is a setup where consistent estimation fails because of nonidentifiability and he then shows that the *equivalence class of parameters giving rise to the same distribution* is estimable consistently by ML in the quotient space $\Omega/\sim$ under certain conditions. C is trivially identifiable with respect to that equivalence relation according to Definition 1.2. Redner's consistent estimate of such an equivalence class is only interpretable properly if all its members are known. Therefore the identifiability considerations presented here do not become superfluous by estimating such equivalence classes of parameters.

2. Mixture Model, Fixed Covariates

Model 1 is treated in this section. To my knowledge there are four papers containing consistency arguments: Kiefer (1978) and DeSarbo and Cron (1988) treat the consistency of the ML-estimator; Quandt and Ramsey (1978) as well as Kiefer (1978) investigate the MGF-estimator. Huang and Pao (1991) propose a minimum distance-estimator. All proofs base on a generalization of results for i.i.d. random variables. The Y_i, $i \in I$, are not identically distributed and one has to inspect the sequence $\tilde{\mathbf{X}} = (\mathbf{x}_i)_{i \in I}$. The choice of I is discussed at the end of this section. Clearly the covariates $\mathbf{x}_i^-, i \in I$ must not lie on a common $(p-1)$-dimensional hyperplane because otherwise even a single regression would not be identified. Proofs of consistency for a single linear regression need to prevent the covariate sequence not only from being collinear, but also from being *too near* to collinearity: Consider the setup of (1) and recall that under Gaussian errors the LS-estimator is distributed Gaussian with variance $\sigma^2(\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n)^{-1}$, and $\tilde{\mathbf{X}}_n$ being the covariate matrix of n observations. Thus, $(\tilde{\mathbf{X}}_n'\tilde{\mathbf{X}}_n)^{-1} \to 0$ is necessary even for convergence in probability, and in the Gaussian case it suffices for strong consistency (Anderson and Taylor 1976). None of the papers cited above states assumptions of that kind, which suggests that the authors generalized the i.i.d.-results too superficially. Kiefer (1978, p. 430, footnote 5) mentions that for this reason his proof, strictly speaking, is not complete, but he does not seem to expect ensuing problems. But there are some. A single regression line is determined by two covariate points, but two clusters are not:

To apply the notation of Definition 1.2, define for given $\tilde{\mathbf{X}}$:

$$C_1 := \left(F_{\tilde{\mathbf{X}},J} : F_{\tilde{\mathbf{X}},J} = \bigotimes_{i \in I} F_{\mathbf{x}_i,J} \right)_{J \in \Omega_1},$$

where $F_{\mathbf{x}_i,J}$ is defined as in Model 1,

$$J \sim_{10} \hat{J} :\Leftrightarrow J = \hat{J} \quad \forall J, \hat{J} \in \Omega_1.$$

Example 2.1. Let $I = \{1,2\}$, $p = 1$, $|S(J)| = 2$, $\mathbf{x}_1 = (0,1)$, $\mathbf{x}_2 = (1,1)$. Define the joint distribution F of (Y_1, Y_2) as the independent product of the distribution functions

$$F_{\mathbf{x}_1} = \tfrac{1}{2}\Phi_{1,\sigma^2} + \tfrac{1}{2}\Phi_{2,\sigma^2}, \tag{3}$$

$$F_{\mathbf{x}_2} = \tfrac{1}{2}\Phi_{1,\sigma^2} + \tfrac{1}{2}\Phi_{2,\sigma^2}, \tag{4}$$

$$\sigma^2 \geq 0, \tag{5}$$

and consider the following different parameter choices:

$$j = 1,2: \ \sigma_j^2 = \sigma^2, \boldsymbol{\beta}_1 = (0,1), \boldsymbol{\beta}_2 = (0,2),$$

and $J\{(\boldsymbol{\beta}_j, \sigma_j^2)\} = \frac{1}{2}$, or

$$j = 1,2: \ \hat{\sigma}_j^2 = \sigma^2, \hat{\boldsymbol{\beta}}_1 = (1,1), \hat{\boldsymbol{\beta}}_2 = (-1,2),$$

and $\hat{J}\{(\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)\} = \frac{1}{2}$. Observe that $F = F_{\tilde{\mathbf{X}},J} = F_{\tilde{\mathbf{X}},\hat{J}}$. C_1 is not identifiable w.r.t. "$\sim_{10}$"; see the left side of Figure 2.

Consequently, no consistency is possible at Model 1 if the covariates concentrate on (presumably even near to) two $(p-1)$-dimensional hyperplanes. The right side of Figure 2 shows also that three clusters are not in general identifiable with covariates from three lower-dimensional hyperplanes. The authors of all four papers cited above did not exclude these cases from their consistency considerations.

Example 2.1 deals with parameters that are nonidentifiable and therefore not consistently estimable. One may be surprised that nevertheless their Fisher information matrix is invertible:

Example 2.1. Continued. Consider F of Example 2.1. Assume $\sigma_1^2 = \sigma_2^2 = 1$ as well as $\beta_{12} = 1$, $\beta_{22} = 2$ as known and consider the problem of estimating β_{11}, β_{21}, i.e., the parameters where J and $\hat{J}$ of Example 2.1 differ. Compute the information matrix for the parameters $(\beta_{11}, \ \beta_{21})$ that may have the values $(0, \ 0)$ in case of J or $(1, \ -1)$ in case of $\hat{J}$ as shown above.

Here is the density of $F = F_{\beta_{11},\beta_{21}}$:

$$f_{\beta_{11},\beta_{21}}(y_1, y_2) = \left[\frac{1}{2}\varphi_{1,1}(y_1) + \frac{1}{2}\varphi_{2,1}(y_1)\right]\left[\frac{1}{2}\varphi_{\beta_{11}+1,1}(y_2) + \frac{1}{2}\varphi_{\beta_{21}+2,1}(y_2)\right].$$

Use $\frac{\partial}{\partial a}\varphi_{a,1}(y) = (y-a)\varphi_{a,1}(y)$ to get

$$\frac{\partial}{\partial \beta_{11}} \log f_{\beta_{11},\beta_{21}}(y_1, y_2) = \frac{(y_2-\beta_{11}-1)\varphi_{\beta_{11}+1,1}(y_2)}{\varphi_{\beta_{11}+1,1}(y_2)+\varphi_{\beta_{21}+2,1}(y_2)}, \text{ and}$$

$$\frac{\partial}{\partial \beta_{21}} \log f_{\beta_{11},\beta_{21}}(y_1, y_2) = \frac{(y_2-\beta_{21}-2)\varphi_{\beta_{21}+2,1}(y_2)}{\varphi_{\beta_{11}+1,1}(y_2)+\varphi_{\beta_{21}+2,1}(y_2)}.$$

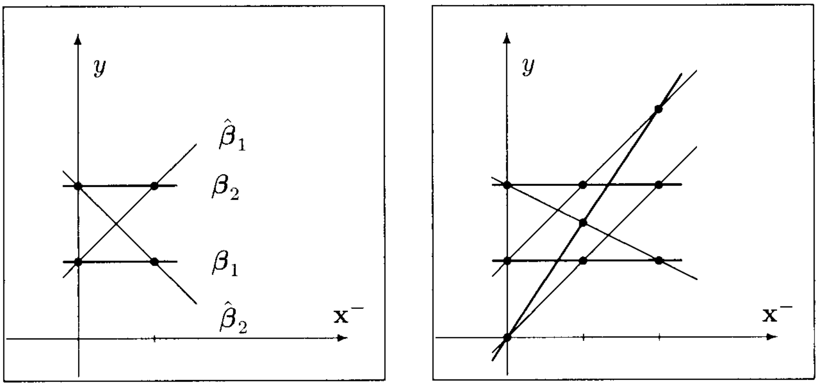

Figure 2: The black circles represent the means of the Gaussian distributions of y at different x-values. In Figure 2a, (β_1, β_2) defines one two-cluster parameterization and $(\hat{\beta}_1, \hat{\beta}_2)$ defines an alternative (Example 2.1). In Figure 2b the nine means can be represented by three clusters in two different ways, corresponding to either the light lines or the heavy lines.

Now the information matrix $\mathcal{I}(\beta_{11}, \beta_{21})$ can be evaluated at $\beta_{11} = \beta_{21} = 0$:

$$\int \left(\frac{\partial}{\partial \beta_{11}} \log f_{\beta_{11},\beta_{21}} \bigg|_{\beta_{11}=\beta_{21}=0} \right)^2 dF_{0,0} = 0.2861$$

$$= \int \left(\frac{\partial}{\partial \beta_{21}} \log f_{\beta_{11},\beta_{21}} \bigg|_{\beta_{11}=\beta_{21}=0} \right)^2 dF_{0,0},$$

$$\int \left(\frac{\partial}{\partial \beta_{11}} \log f_{\beta_{11},\beta_{21}} \bigg|_{\beta_{11}=\beta_{21}=0} \right)$$

$$\left(\frac{\partial}{\partial \beta_{21}} \log f_{\beta_{11},\beta_{21}} \bigg|_{\beta_{11}=\beta_{21}=0} \right) dF_{0,0} = 0.1144.$$

Thus,

$$\mathcal{I}(0,0) = \begin{pmatrix} 0.2861 & 0.1144 \\ 0.1144 & 0.2861 \end{pmatrix},$$

which is invertible. The reason is that the information matrix is based on derivatives. Thus it is a local concept. The uniqueness of $(\beta_{11}, \beta_{21}) = (0,0)$ is not distorted in the neighborhood of $(0,0)$, but is at $(1,-1)$, so that the information matrix does not need to detect any abnormalities at $(0,0)$.

Let h denote the minimum number of $(p-1)$-dimensional hyperplanes to cover the covariates. The key problem of the examples of Figure 2 is that h is too small compared to the number of clusters. The following theorem shows that it has to be larger than the number of clusters $|S(J)|$ to guarantee identifiability.

Theorem 2.2. *In Model 1, C_1 is identifiable w.r.t. "$\sim_{10}$" under the additional restriction $|S(J)| < h \, \forall J \in \Omega_1$, where*

$$h := \min\left\{ q : \{\mathbf{x}_i^- : i \in I\} \subseteq \bigcup_{i=1}^{q} H_i : H_i \in \mathcal{H}_{p-1} \right\}.$$

All proofs are given in the Appendix. If $|I| = \infty$, C_1 can be identifiable without restrictions to the number of clusters:

Corollary 2.3. *In Model 1, C_1 is identifiable w.r.t. "$\sim_{10}$", if $\{\mathbf{x}_i^- : i \in I\}$ cannot be covered by any A of the form $A = \bigcup_{i=1}^{m} H_i$, $m \in I\!N$, where $H_i \in \mathcal{H}_{p-1}$, $i = 1, \ldots, m$.*

I offer two possibilities to interprete the index set I:

1. "Observation model": $I = I\!N$, each $(\mathbf{x}_i, Y_i)$ models one of the observations of a (potentially) infinite sequence of observations. Index sets for a single linear regression are usually interpreted as observation models and lead to consistency conditions about the limit behavior of $(\tilde{\mathbf{X}}_n' \tilde{\mathbf{X}}_n)^{-1}$. Analogous conditions for Model 1 could be cumbersome, because one presumably has to exclude the case that the covariate sequence comes too near to nonidentifiability situations, as shown above. The problem might be only of technical nature: Consistency may hold under mild assumptions, which, however, may be hard to derive.

2. "Repeatable design": A theoretically easier but not necessarily realistic approach would be to choose $I = \{1, \ldots, m\}$, m being the number of distinct available covariate points. If observations could be repeated at all m points, the whole experimental design $(\mathbf{x}_i, Y_i)_{i \in I}$ could be repeated i.i.d. and would possibly enable easier consistency proofs. I can also be interpreted as a repeatable design, if the $\mathbf{x}_i$, $i \in I$, are not pairwise distinct. Such an interpretation means that some covariate values occur more often than others in the design to be repeated.

3. Mixture Model, Random Covariates

I do not know of any attempt to solve the problem of consistent estimation of clusterwise linear regressions using random covariates, even if the execution of Kiefers (1978, p. 430, footnote 5) proposal to complete his proof would result in considering such a model. The mixture model with random covariates can be stated as follows:

Model 2.

$$\begin{aligned} \mathcal{L}(\mathbf{X},Y) &= F_J, \\ F_J(\mathbf{x},y) &= \textstyle\int_{T_2} F(\mathbf{x},y,\boldsymbol{\theta})dJ(\boldsymbol{\theta}), \text{ where} \\ F(\mathbf{x},y,\boldsymbol{\theta}) &= \textstyle\int 1(t \leq \mathbf{x})\Phi_{0,\sigma^2}(y - t'\boldsymbol{\beta})dG(t), \\ \boldsymbol{\theta} := (\boldsymbol{\beta},\sigma^2,G) &\in T_2 := I\!R^{p+1} \times I\!R_0^+ \times \mathcal{G},\ J \in \Omega_2, \end{aligned}$$

$\mathcal{G}$ being some set of distributions of $\mathbf{X} \in I\!R^{p+1}$ where for $G \in \mathcal{G}: \ G(\mathbf{X}_{p+1} = 1) = 1$, $\Omega_2 \subseteq \mathcal{J}(T_2)$, i.e., $C_2 := (F_J)_{J \in \Omega_2}$.

For the sake of simplicity in the following $G \in \mathcal{G}$ is treated as a p-dimensional distribution and $\mathbf{X}_{p+1} = 1$ as fixed.

That is, a single linear regression distribution F is defined by $(\boldsymbol{\beta},\sigma^2)$, and the covariate distribution G, and it occurs with probability $J(\boldsymbol{\beta},\sigma^2,G)$ in the mixture distribution F_J. Mixing over G does not impose measurability problems as can be seen from Chapter 12 of Hinderer (1970). Instead of $\mathcal{J}(T_2)$, it is more convenient to work with

$$\Omega_2 := \{J \in \mathcal{J}(T_2) : \ (\boldsymbol{\beta},\sigma^2,G) \in S(J), G \neq \hat{G} \Rightarrow (\boldsymbol{\beta},\sigma^2,\hat{G}) \notin S(J)\}.$$

This approach allows only mixing distributions where distinct clusters are characterized by distinct regression or scale parameters. Otherwise, the term "linear regression clusters" would not be justified.

The model has two advantages compared to Model 1:

1. The random variables $(\mathbf{X}_i, Y_i),\ i = 1,\ldots,n$ for n observations are i.i.d., and the corresponding theory may be applied.
2. Assignment independence is not assumed as long as J also mixes over G, i.e., the covariate distributions of the clusters are allowed to be distinct.

Random covariates occur in many applications, e.g., in Example 1.1, the age of the halibuts caught cannot be controlled, and most economical surveys, while they are of course not adequate if time is the covariate or there is some controlled experimental design. One could wonder as well in the latter situations, if some estimator of Model 2 without the assignment independence assumption could be better as one of Model 1 if this assumption would not be valid. But Section 4 provides a more adequate model for this case.

What can be estimated? I consider the case in which the covariate distributions are not of primary interest. There are models where such "nuisance" distributions can be estimated as well (see Kiefer and Wolfowitz 1956), and indeed they are sometimes identifiable (Theorem 3.2). If this is not the case, the regression parameters and the cluster proportions $J(\beta, \sigma^2, G)$ are of interest. Different estimation problems lead to different equivalence relations on Ω_2, which are chosen such that two mixing distributions are equivalent if the parameter parts of interest are equal:

(a) Estimation of the covariate distributions, cluster proportions, regression and scale parameters, i.e., the whole mixing distribution:

$$J \sim_{20} \hat{J} :\Leftrightarrow J = \hat{J}.$$

(b) Estimation of cluster proportions, regression and scale parameters, but not of the covariate distributions ($J^{(\beta,\sigma^2)}$ denotes the marginal distribution of (β, σ^2) under J):

$$J \sim_{21} \hat{J} :\Leftrightarrow J^{(\beta,\sigma^2)} = \hat{J}^{(\beta,\sigma^2)}.$$

(c) Estimation of only the regression and scale parameters:

$$\begin{aligned} J \sim_{22} \hat{J} :\Leftrightarrow & \{(\beta, \sigma^2);\ (\beta, \sigma^2, G) \in S(J)\} \\ = & \{(\beta, \sigma^2);\ (\beta, \sigma^2, G) \in S(\hat{J})\}. \end{aligned}$$

The equivalence classes of "$\sim_{21}$" and "$\sim_{22}$" do not determine the members of C_2, i.e., $J \sim_{2i} \hat{J}$ does not imply $F_J = F_{\hat{J}}$, $i = 1, 2$. Thus, one can only be interested in partial identifiability w.r.t. these equivalence relations.

If the covariate distributions concentrate on too few $(p-1)$-dimensional hyperplanes, counterexamples against identifiability can be obviously constructed as in Sections 2 and 4, Figures 2 and 4. But another problem can appear:

Example 3.1. C_2 is not identifiable w.r.t. "$\sim_{20}$" nor even partially identifiable w.r.t. "$\sim_{21}$" in general, if $\mathcal{G}$ contains distributions which have positive probability on a $(p-1)$-dimensional hyperplane

$$H_{\boldsymbol{\alpha}} := \{\mathbf{x}^- \in I\!R^p : \mathbf{x}'\boldsymbol{\alpha} = 0\}, \quad \boldsymbol{\alpha} \in I\!R^{p+1} \setminus \{0\},$$

as in Figure 3 ($H_{\boldsymbol{\alpha}}$ is a single point in $I\!R^1$). In this case the cluster proportions are not identified because the points on $H_{\boldsymbol{\alpha}}$ (distributed according to G_H below) may be assigned to the cluster with regression parameter β, $\beta + \boldsymbol{\alpha}$ respectively:

$\mathbf{x}'\beta = \mathbf{x}'(\beta + \boldsymbol{\alpha})$ holds for $\mathbf{x} \in H_{\boldsymbol{\alpha}}$. Let $G_1 \in \mathcal{G}$, where

$$1 > \epsilon := G_1(H_{\boldsymbol{\alpha}}) > 0, \quad G_H(B) := G_1(B|H_{\boldsymbol{\alpha}}) \forall \text{ Borel sets } B \in I\!B^p,$$
$$G_2 := \tfrac{1}{1-\epsilon}(G_1 - \epsilon G_H) \text{ (supposed to be } \in \mathcal{G}).$$

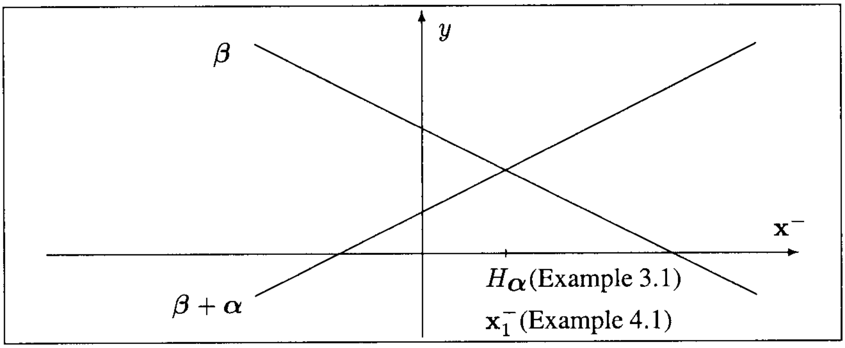
y
$\boldsymbol{\beta}$
$\boldsymbol{\beta}+\boldsymbol{\alpha}$
$\mathbf{x}^-$
$H_{\boldsymbol{\alpha}}$(Example 3.1)
$\mathbf{x}_1^-$(Example 4.1)

Figure 3: Example 3.1/Example 4.1: The regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\beta}+\boldsymbol{\alpha}$ lead to the same distribution of y for covariates from $H_{\boldsymbol{\alpha}}$.

G_2 corresponds to G_1 conditional on $H^c_{\boldsymbol{\alpha}}$. Let $\delta_{\mathbf{a}}$ denote the Dirac measure in $\mathbf{a}$. Consider

$$J := \tfrac{1}{2-\epsilon}\delta_{(\boldsymbol{\beta},\sigma^2,G_1)} + \tfrac{1-\epsilon}{2-\epsilon}\delta_{(\boldsymbol{\beta}+\boldsymbol{\alpha},\sigma^2,G_2)},$$
$$\hat{J} := \tfrac{1-\epsilon}{2-\epsilon}\delta_{(\boldsymbol{\beta},\sigma^2,G_2)} + \tfrac{1}{2-\epsilon}\delta_{(\boldsymbol{\beta}+\boldsymbol{\alpha},\sigma^2,G_1)}.$$

J assigns $H_{\boldsymbol{\alpha}}$ to the first cluster, $\hat{J}$ assigns it to the second one. This does not change the common distribution of $\mathbf{X}, Y$:

$$\begin{aligned}
F_J^{(\mathbf{X},Y)} &:= \tfrac{1}{2-\epsilon}(G_1^{\mathbf{X}} \otimes \mathcal{N}^Y_{\mathbf{X}'\boldsymbol{\beta},\sigma^2}) + \tfrac{1-\epsilon}{2-\epsilon}(G_2^{\mathbf{X}} \otimes \mathcal{N}^Y_{\mathbf{X}'(\boldsymbol{\beta}+\boldsymbol{\alpha}),\sigma^2}) \\
&= \tfrac{1-\epsilon}{2-\epsilon}(G_2^{\mathbf{X}} \otimes \mathcal{N}^Y_{\mathbf{X}'\boldsymbol{\beta},\sigma^2}) + \tfrac{\epsilon}{2-\epsilon}(G_H^{\mathbf{X}} \otimes \mathcal{N}^Y_{\mathbf{X}'\boldsymbol{\beta},\sigma^2}) + \tfrac{1-\epsilon}{2-\epsilon}(G_2^{\mathbf{X}} \otimes \mathcal{N}^Y_{\mathbf{X}'(\boldsymbol{\beta}+\boldsymbol{\alpha}),\sigma^2}) \\
&= \tfrac{1-\epsilon}{2-\epsilon}(G_2^{\mathbf{X}} \otimes \mathcal{N}^Y_{\mathbf{X}'\boldsymbol{\beta},\sigma^2}) + \tfrac{1}{2-\epsilon}(G_1^{\mathbf{X}} \otimes \mathcal{N}^Y_{\mathbf{X}'(\boldsymbol{\beta}+\boldsymbol{\alpha}),\sigma^2}) =: F_{\hat{J}}^{(\mathbf{X},Y)},
\end{aligned}$$
$$J \not\sim_{20} \hat{J}, \quad J \not\sim_{21} \hat{J}.$$

"$\sim_{21}$" is introduced only to illustrate that if there are problems with the covariate distributions, also the proportions are not identified. The identifiability results concern either "$\sim_{20}$" or "$\sim_{22}$": To get identifiability w.r.t. "$\sim_{20}$", the covariate distributions must not give positive probability to any $(p-1)$-dimensional hyperplane:

Theorem 3.2. *In Model 2, C_2 is identifiable w.r.t. "$\sim_{20}$" if $\mathcal{G} \subseteq \{P \in \mathcal{P}_p : \; P(H) = 0 \quad \forall H \in \mathcal{H}_{p-1}\}$.*

The regression and scale parameters are (partially) identified, if no $G \in \mathcal{G}$ concentrates on less than $h+1$ hyperplanes of dimension $(p-1)$, h being an upper bound for the number of clusters $|S(J)|$.

Theorem 3.3. *In Model 2, C_2 is partially identifiable w.r.t. "$\sim_{22}$" under the additional restriction that $|S(J)| < h \, \forall J \in \Omega_2$, if*

$$\mathcal{G} \subset \mathcal{P}_p, \text{ where } \forall G \in \mathcal{G},\ m \leq h \in I\!N,\ A = \bigcup_{i=1}^{m} H_i : \quad G(A) < 1 \quad (6)$$

for arbitrary $H_i \in \mathcal{H}_{p-1}$, $i = 1, \ldots, m$.

4. Fixed Partition Model

The fixed partition model with fixed covariates was considered first by Fair and Jaffee (1972) in the regression context. "Fixed partition" means that the cluster membership of each point is not treated as random, but parameterized by some γ that determines the particular regression and scale parameter values for the cluster of each single point $(\mathbf{x}_i, y_i)$:

Model 3.

$$\mathcal{L}\left((Y_i)_{i \in I}\right) = \bigotimes_{i \in I} F_{\mathbf{x}_i, \gamma(i)}, \text{ where}$$
$$\gamma : I \mapsto I\!R^{p+1} \times I\!R_0^+, \quad |\gamma(I)| < \infty,$$
$$F_{\mathbf{x}, \boldsymbol{\beta}, \sigma^2}(y) = \Phi_{0,\sigma^2}(y - \mathbf{x}'\boldsymbol{\beta}) \quad \forall (\boldsymbol{\beta}, \sigma^2) \in \gamma(I).$$

This model is most flexible in the sense that it can be interpreted as one of the other two models *conditional* on a given assignment of a single point $(\mathbf{x}_i, Y_i)$ to $\gamma(i)$, which occurs with probability $J(\gamma(i))$ in Model 1, respectively $J(\gamma(i), G)$ in Model 2, and on the given covariate values in Model 2. There are further applications, where some deterministic covariate value or the position i in the experiment determines the cluster membership as in change point problems as Example 1.1.2. Furthermore, the cluster proportions may change over the time.

This flexibility is paid for by the fact that the number of parameters tends with the number of observations to infinity, if I is interpreted as an observation model.

I do not know of any publication where identifiability of some fixed partition model is discussed. The reason is that in linear regression there is an identifiability problem, while in other situations there is none. Consider, e.g., one-dimensional Gaussian distributions, i.e., $p = 0$, $\mathbf{x} \equiv 1$ above. For each i there is only a single Gaussian distribution, and its location parameter $\boldsymbol{\beta}$ and the scale σ^2 are clearly identified as expectation and variance. However, this result is not very useful if I is interpreted as an observation model, because then there is only one observation available for each parameter $\gamma(i)$. This result does not answer the question if $\gamma(I)$, i.e., the *set* of location and scale parameters, can be estimated consistently from the non-i.i.d. sequence $(Y_i)_{i \in I}$. The assignment of each single point Y_i is identifiable, but it cannot be estimated consistently, unless I is a repeatable design. Nevertheless, identifiability is a necessary condition for the existence of consistent estimators of $\gamma(I)$, and in the linear regression case it is not such trivial, as will be shown.

Note that one can investigate the limit behavior of ML-estimators of the parameters of the fixed partition model at least in the location/scale-case, but this problem is usually approached assuming i.i.d. observations from some mixture model (e.g., Bryant and Williamson 1978). To my knowledge there is no asymptotic theory assuming *observations* from a fixed partition model. Furthermore, the ML-estimator is known to be inconsistent as shown by Marriott (1975), Bryant and Williamson (1978) and for the linear regression case by Oberhofer (1980).

I define two equivalence relations, namely "$\sim_{30}$" for the identification of the whole parameterization, and "$\sim_{31}$" for identifying only the regression and scale parameters: For given $\tilde{\mathbf{X}}$ let

$$\Omega_3 := \left\{\gamma : I \mapsto \mathbb{R}^{p+1} \times \mathbb{R}_0^+, \quad |\gamma(I)| < \infty\right\},$$

$$C_3 := \left(F_{\tilde{\mathbf{X}},\gamma}\right)_{\gamma \in \Omega_3}, \; F_{\tilde{\mathbf{X}},\gamma} := \bigotimes_{i \in I} F_{\mathbf{x}_i,\gamma(i)},$$

$$\gamma \sim_{30} \hat{\gamma} :\Leftrightarrow \gamma = \hat{\gamma} \quad \forall \gamma, \hat{\gamma} \in \Omega_3,$$

$$\gamma \sim_{31} \hat{\gamma} :\Leftrightarrow \gamma(I) = \hat{\gamma}(I).$$

Observe first that the cluster membership of the single points is not identified, as opposed to the Gaussian location case: Example 4.1 (see Figure 3). C_3 is not identifiable w.r.t. "$\sim_{30}$": Choose $\boldsymbol{\alpha} \in \mathbb{R}^{p+1} \setminus \{0\}$ such that $\mathbf{x}_1'\boldsymbol{\alpha} = 0$. Let

$$\gamma(I) = \{(\boldsymbol{\beta}, \sigma^2), (\boldsymbol{\beta} + \boldsymbol{\alpha}, \sigma^2)\},$$

so that there are two distinct clusters corresponding to the regression parameters $\boldsymbol{\beta}$, $\boldsymbol{\beta} + \boldsymbol{\alpha}$, respectively. Then, $(\mathbf{x}_1, y_1)$ may be assigned to each of them: $F_{(\mathbf{x}_i)_{i \in I},\gamma}$ is equal for $\gamma(1) = (\boldsymbol{\beta}, \sigma^2)$, and $\gamma(1) = (\boldsymbol{\beta} + \boldsymbol{\alpha}, \sigma^2)$.

As a consequence of the assignment independence, all clusters are supported by all covariate points in Model 1. This situation is not the case for the Models 2 and 3. Therefore, there are counterexamples against identifiability of the regression and scale parameters with more distinct covariate points ($(p-1)$-dimensional hyperplanes, respectively) than in Example 2.1:

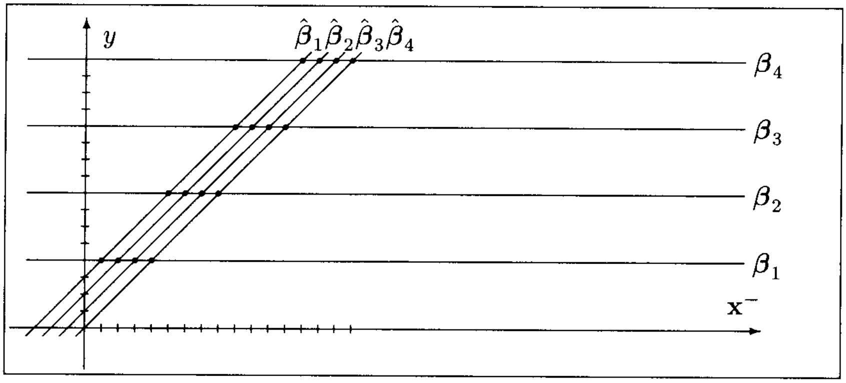

Figure 4: Example 4.2: The black circles represent again the means of the Gaussian distributions of y for the different s^2 covariate points. $(\beta_1, \ldots, \beta_4)$ and $(\hat\beta_1, \ldots, \hat\beta_4)$ define different four-cluster parameterizations of the same model.

Example 4.2. Let $I = \{1, \ldots, n\}$, $p = 1$, $n = s^2$, $s := |\gamma(I)|$ being the number of clusters in Model 3. Choose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ equidistant (which would not be necessary), say $\mathbf{x}_i = (i, 1)$. Then the joint distribution of the $(Y_i)_{i \in I}$ is defined by

$$F_{\mathbf{x}_i}(y_i) = \Phi_{0,\sigma^2}(y_i - js), \quad i \in \{(j-1)s+1, \ldots, js\},\ j = 1, \ldots, s,$$

simultaneously for the following parameter choices (see Figure 4):

$$\gamma(i) = (0, js, \sigma^2) :\Leftrightarrow i \in \{(j-1)s+1, \ldots, js\},\ j = 1, \ldots, s,$$

or

$$\hat\gamma(i) = (1, s-j, \sigma^2) :\Leftrightarrow i \in \{ks + j : k \in \{0, \ldots, s-1\}\},\ j = 1, \ldots, s.$$

That is, C_3 is not partially identifiable w.r.t. "$\sim_{31}$".

The example is perhaps interesting also in change point theory, because it illustrates that this kind of a change point setup (with $s-1$ change points parameterized by γ) cannot be separated from a situation with s linear regression clusters parameterized by $\hat\gamma$ without clusterwise connected domains of covariates.

The solution of the identifiability problem in Model 3 as well as in Model 2 (Theorem 3.3) is that the number of clusters (here $|\gamma(I)|$) has to be exceeded by the number of distinct $(p-1)$-dimensional hyperplanes to cover the covariate values for *each* cluster (denoted by $h(\beta, \sigma^2)$):

Theorem 4.3. *C_3 is partially identifiable w.r.t. "$\sim_{31}$" if the restriction $|\gamma(I)| < \min_{(\boldsymbol{\beta},\sigma^2)\in\gamma(I)} h(\boldsymbol{\beta},\sigma^2)$ is added to Ω_3, where*

$$h(\boldsymbol{\beta},\sigma^2) := \min\left\{q:\ \{\mathbf{x}_i^- : i \in I, \gamma(i) = (\boldsymbol{\beta},\sigma^2)\} \subseteq \bigcup_{i=1}^{q} H_i : H_i \in \mathcal{H}_{p-1}\right\}. \tag{7}$$

Unfortunately this result implies that one cannot determine the maximum number of clusters guaranteeing identifiability by considering the covariate design $(\mathbf{x}_i)_{i\in I}$ alone, as opposed to Model 1.

5. Conclusion

The problem of consistent estimation of the parameters of linear regression clusters turned out to be difficult. It was shown that the existing consistency proofs for estimators of the mixture model with fixed covariates do not take identifiability problems adequately into account. Furthermore, this model suffers from the restrictive assumption of assignment independence. I discussed two more flexible models in which no consistent estimators are known up to now. A lot of work remains to be done to prove consistency for any estimator in any of the models. Identifiability conditions were given for all three models, accompanied by some discussion of (partial) identifiability for the fixed partition model, where the identifiability question only starts to get interesting for the regression case. In general the regression and scale parameters are identifiable if the number of clusters is exceeded by the number of distinct $(p-1)$-dimensional hyperplanes which one needs to cover the covariates of each cluster. These conditions seem to be rather mild, but they may be violated in applications where the covariate variables can only take a small number of values. This happens

(a) if they are dummy variables, e.g., in ANOVA,

(b) at optimal designs for *single* linear regression, or

(c) if they reflect a small number of possible answers to questionnaires as often in marketing research.

For example, the three-dimensional covariates of the first data example of Kamarkura (1988) can be covered by two two-dimensional hyperplanes. (Kamarkura performs a clusterwise regression on these data. He does not discuss identifiability, but the counterexamples here do not directly apply because he imposes an additional side condition for his clustering.)

It would be useful to have an algorithm to compute the minimal number of $(p-1)$-dimensional hyperplanes needed to cover the n points of dimension p. In Model 1 the answer is "one plus the maximal number of clusters that can be estimated identifiable with a given covariate design of n points". In the other models one could verify if the covariates of every single cluster would suffice to identify the given number of clusters after having assigned the points.

Edelsbrunner (1987, p. 278 ff.) gives an algorithm to report all subsets of more than p points that can be covered by $(p-1)$-dimensional hyperplanes. This approach will suffice often to compute the requested minimum number. Unfortunately even the existence of such subsets is an NP-complete problem (Khachiyan 1995).

ML-estimators for the Models 1 and 3 and "Fixed Point Clustering" for Model 2 are discussed in Hennig (1998).

Appendix

Proof of Theorem 2.2. Only $F_{\tilde{\mathbf{X}},J} = F_{\tilde{\mathbf{X}},\hat{J}} \Rightarrow J = \hat{J}$ has to be shown, because J contains all information to define the common distribution $F_{\tilde{\mathbf{X}},J}$ of $(Y_i)_{i\in I}$. Let $F_{\tilde{\mathbf{X}},J} = F_{\tilde{\mathbf{X}},\hat{J}}$, $J \neq \hat{J}$, without loss of generality

$$|S(J)| \geq |S(\hat{J})|, \quad J\{(\beta_1,\sigma_1^2)\} \neq \hat{J}\{(\beta_1,\sigma_1^2)\}. \tag{8}$$

$F_{\tilde{\mathbf{X}},J} = F_{\tilde{\mathbf{X}},\hat{J}}$ implies the equality of the marginal Gaussian mixtures at all $\mathbf{x}_i$, $i \in I$:

$$\begin{aligned} F_{\mathbf{x}_i,J} &= \int_{T_1} \mathcal{N}_{(\mathbf{x}_i'\beta,\sigma^2)} dJ(\beta,\sigma^2) = \\ &= F_{\mathbf{x}_i,\hat{J}} = \int_{T_1} \mathcal{N}_{(\mathbf{x}_i'\beta,\sigma^2)} d\hat{J}(\beta,\sigma^2). \end{aligned} \tag{9}$$

By identifiability of finite Gaussian mixtures, for $i \in I$:

$$\hat{J}\{(\hat{\beta},\hat{\sigma}^2) : (\mathbf{x}_i'\hat{\beta},\hat{\sigma}^2) = (\mathbf{x}_i'\beta_1,\sigma_1^2)\} = J\{(\beta,\sigma^2) : (\mathbf{x}_i'\beta,\sigma^2) = (\mathbf{x}_i'\beta_1,\sigma_1^2)\}. \tag{10}$$

A crucial idea in the proof is that the restriction to $|S(\hat{J})|$ ensures the existence of a covariate point $\mathbf{x}_i$ where the marginal mixture proportion of $\mathcal{N}_{(\mathbf{x}_i'\beta_1,\sigma_1^2)}$, parameterized by J, cannot be explained by $(\hat{\beta},\hat{\sigma}^2) \in S(\hat{J})$ with $\hat{\beta} \neq \beta_1$, as will be shown. Therefore, $S(\hat{J})$ must contain (β_1,σ_1^2). Covariate points with similar properties are considered in the other proofs. The argument can be taken further to contradict (8):

The assumption $|S(J)| < h$ would be in contradiction to the existence of some $(\beta,\sigma^2) \in S(J)$ such that

$$\bigcup_{(\hat{\beta},\hat{\sigma}^2)\in S(\hat{J}):\hat{\beta}\neq\beta} \{\mathbf{x}^- : \mathbf{x}'\beta = \mathbf{x}'\hat{\beta}\} \supset \{\mathbf{x}_i^- : i \in I\},$$

because then $h \leq |S(\hat{J})| \leq |S(J)|$.

Thus, for all $(\boldsymbol{\beta},\sigma^2) \in S(J)$, and in particular for $(\boldsymbol{\beta}_1,\sigma_1^2)$, there exists $i(\boldsymbol{\beta}) \in I$ such that

$$\forall(\hat{\boldsymbol{\beta}},\hat{\sigma}^2) \in S(\hat{J}) : \ \mathbf{x}'_{i(\boldsymbol{\beta})}\boldsymbol{\beta} = \mathbf{x}'_{i(\boldsymbol{\beta})}\hat{\boldsymbol{\beta}} \Rightarrow \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}. \tag{11}$$

Put $i = i(\boldsymbol{\beta}_1)$ in (9). The definition of $\mathbf{x}_i = \mathbf{x}_{i(\boldsymbol{\beta}_1)}$ implies that

$$\forall S(\hat{J}) \ni (\hat{\boldsymbol{\beta}},\hat{\sigma}^2) \neq (\boldsymbol{\beta}_1,\sigma_1^2) : \ (\mathbf{x}'_i\hat{\boldsymbol{\beta}},\hat{\sigma}^2) \neq (\mathbf{x}'_i\boldsymbol{\beta}_1,\sigma_1^2). \tag{12}$$

Thus, using (10),

$$\hat{J}\{(\boldsymbol{\beta}_1,\sigma_1^2)\} = J\{(\boldsymbol{\beta},\sigma^2) : (\mathbf{x}'_i\boldsymbol{\beta},\sigma^2) = (\mathbf{x}'_i\boldsymbol{\beta}_1,\sigma_1^2)\} \tag{13}$$

implying $(\boldsymbol{\beta}_1,\sigma_1^2) \in S(\hat{J})$.

By (8), $\hat{J}\{(\boldsymbol{\beta}_1,\sigma_1^2)\} \neq J\{(\boldsymbol{\beta}_1,\sigma_1^2)\}$, therefore

$$\exists S(J) \ni (\boldsymbol{\beta}_2,\sigma_2^2) \neq (\boldsymbol{\beta}_1,\sigma_1^2) : (\mathbf{x}'_i\boldsymbol{\beta}_2,\sigma_2^2) = (\mathbf{x}'_i\boldsymbol{\beta}_1,\sigma_1^2). \tag{14}$$

Consider $\mathbf{x}_i = \mathbf{x}_{i(\boldsymbol{\beta}_2)}$ and apply the arguments above again to get $(\boldsymbol{\beta}_2,\sigma_2^2) \in S(\hat{J})$. This result leads to a contradiction between (12) and (14).

Proof of Theorem 3.2.
a) Preparation. Only

$$F_J = F_{\hat{J}} \Rightarrow J = \hat{J} \tag{15}$$

has to be shown, because J contains all information to define F_J. Let $s := |S(J)|, \hat{s} := |S(\hat{J})|$, where

$$S(J) = \{(\boldsymbol{\beta}_i,\sigma_i^2,G_i), i = 1,\ldots,s\}, \quad \pi_i := J\{(\boldsymbol{\beta}_i,\sigma_i^2,G_i)\}, \tag{16}$$

$$S(\hat{J}) = \{(\boldsymbol{\beta}_i,\sigma_i^2,G_i), i = 1,\ldots,\hat{s}\}, \quad \hat{\pi}_i := \hat{J}\{(\hat{\boldsymbol{\beta}}_i,\hat{\sigma}_i^2,\hat{G}_i)\}. \tag{17}$$

With this notation,

$$F_J = \sum_{i=1}^{s} \pi_i F(\bullet,\boldsymbol{\beta}_i,\sigma_i^2,G_i),$$

F defined as in Model 2. Define

$$\mu := \sum_{i=1}^{s} G_i + \sum_{i=1}^{\hat{s}} \hat{G}_i, \tag{18}$$

$$F^{(\mathbf{X},Y)} = F_J = F_{\hat{J}}, \quad g_i(\mathbf{x}^-) := \tfrac{dG_i}{d\mu}(\mathbf{x}^-), \quad \hat{g}_i(\mathbf{x}^-) := \tfrac{d\hat{G}_i}{d\mu}(\mathbf{x}^-). \tag{19}$$

With that, on a set of $\mathbf{x}$-values of probability 1 under $F^{\mathbf{X}}$,

$$F^{Y|\mathbf{X}=\mathbf{x}} = \sum_{i=1}^{s} \pi_i \frac{g_i(\mathbf{x}^-)}{\sum_{j=1}^{s} \pi_j g_j(\mathbf{x}^-)} \mathcal{N}_{(\mathbf{x}'\boldsymbol{\beta}_i, \sigma_i^2)}, \tag{20}$$

and by analogy, considering $F_{\hat{J}}$,

$$F^{Y|\mathbf{X}=\mathbf{x}} = \sum_{i=1}^{\hat{s}} \hat{\pi}_i \frac{\hat{g}_i(\mathbf{x}^-)}{\sum_{j=1}^{\hat{s}} \hat{\pi}_j \hat{g}_j(\mathbf{x}^-)} \mathcal{N}_{(\mathbf{x}'\hat{\boldsymbol{\beta}}_i, \hat{\sigma}_i^2)}. \tag{21}$$

Define the set of all covariate points $\mathbf{x}$ which can be used to distinct different $\boldsymbol{\beta}$-parameters by different values of $\mathbf{x}'\boldsymbol{\beta}$:

$$\begin{aligned} M := \{\mathbf{x} : \forall j, k \in \{1, \ldots, s\}, l, m \in \{1, \ldots, \hat{s}\} : \\ \mathbf{x}'\boldsymbol{\beta}_j = \mathbf{x}'\boldsymbol{\beta}_k \Rightarrow \boldsymbol{\beta}_j = \boldsymbol{\beta}_k, \quad \mathbf{x}'\boldsymbol{\beta}_j = \mathbf{x}'\hat{\boldsymbol{\beta}}_l \Rightarrow \boldsymbol{\beta}_j = \hat{\boldsymbol{\beta}}_l, \\ \mathbf{x}'\hat{\boldsymbol{\beta}}_l = \mathbf{x}'\hat{\boldsymbol{\beta}}_m \Rightarrow \hat{\boldsymbol{\beta}}_l = \hat{\boldsymbol{\beta}}_m\}. \end{aligned} \tag{22}$$

M is complement of a finite union of elements of $\mathcal{H}_{p-1}$. Therefore, from the assumption to $\mathcal{G}$,

$$F^{\mathbf{X}} = \sum_{i=1}^{s} \pi_i G_i \Rightarrow F^{\mathbf{X}}(M) = 1.$$

For $\mathbf{x} \in M$, all $(\mathbf{x}'\boldsymbol{\beta}_i, \sigma_i^2)$, $i = 1, \ldots, s$, are pairwise distinct, because all $(\boldsymbol{\beta}_i, \sigma_i^2)$, $i = 1, \ldots, s$, are pairwise distinct for $J \in \Omega_2$.

b) Identification of $\boldsymbol{\beta}_i, \sigma_i^2$. For $\mathbf{x} \in M$, let $J_{\mathbf{x}}$ be an empirical distribution on $I\!R \times I\!R_0^+$ defined by

$$J_{\mathbf{x}}\{(\mathbf{x}'\boldsymbol{\beta}_i, \sigma_i^2)\} := \pi_i \frac{g_i(\mathbf{x}^-)}{\sum_{j=1}^{s} \pi_j g_j(\mathbf{x}^-)}, \quad i = 1, \ldots, s, \tag{23}$$

$$S(J_{\mathbf{x}}) = \{(\mathbf{x}'\boldsymbol{\beta}_i, \sigma_i^2) : i = 1, \ldots, s, \quad g_i(\mathbf{x}^-) > 0\}, \tag{24}$$

such that (20) can be written as

$$F^{Y|\mathbf{X}=\mathbf{x}} = \int_{I\!R \times I\!R_0^+} \mathcal{N}_{\boldsymbol{\theta}} dJ_{\mathbf{x}}(\boldsymbol{\theta}),$$

and, with $\hat{J}_{\mathbf{x}}$ defined by analogy to (23),

$$F^{Y|\mathbf{X}=\mathbf{x}} = \int_{I\!R \times I\!R_0^+} \mathcal{N}_{\boldsymbol{\theta}} d\hat{J}_{\mathbf{x}}(\boldsymbol{\theta})$$

by (21). Thus, with probability 1 under $F^{\mathbf{X}}$,

$$J_{\mathbf{x}} = \hat{J}_{\mathbf{x}}, \tag{25}$$

because Gaussian distributions are identifiable. Because of $G_i(M) = 1$, there exists $\mathbf{x}(i) \in M : g_i(\mathbf{x}(i)^-) > 0$ and, by analogy, $\hat{\mathbf{x}}(j) \in M : \hat{g}_j(\hat{\mathbf{x}}(j)^-) > 0$.

For given $i \in \{1, \dots, s\}$, (25) implies

$$\exists j \in \{1, \dots, \hat{s}\} : \mathbf{x}(i)'\boldsymbol{\beta}_i = \mathbf{x}(i)'\hat{\boldsymbol{\beta}}_j, \quad \sigma_i^2 = \hat{\sigma}_j^2.$$

The definition of M yields

$$(\boldsymbol{\beta}_i, \sigma_i^2) \in \{(\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2) : j \in \{1, \dots, \hat{s}\}\},$$

and by the same argument applied to $\hat{\mathbf{x}}(j) \quad \forall j \in \{1, \dots, \hat{s}\}$:

$$(\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2) \in \{(\boldsymbol{\beta}_i, \sigma_i^2) : i \in \{1, \dots, s\}\}.$$

Therefore

$$\{(\boldsymbol{\beta}_i, \sigma_i^2) : i \in \{1, \dots, s\}\} = \{(\hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2) : j \in \{1, \dots, \hat{s}\}\}.$$

The definition of Ω_2 enforces pairwise distinctness of the elements, therefore $s = \hat{s}$ and w.l.o.g $(\boldsymbol{\beta}_i, \sigma_i^2) = (\hat{\boldsymbol{\beta}}_i, \hat{\sigma}_i^2), \quad i = 1, \dots, s$.

c) Identification of G_i, ϵ_i. For $i = 1, \dots, s$, there are unique G_i and $\pi_i = J\{(\boldsymbol{\beta}_i, \sigma_i^2, G_i)\}$ for given $(\boldsymbol{\beta}_i, \sigma_i^2)$ such that $(\boldsymbol{\beta}_i, \sigma_i^2, G_i) \in S(J)$ and analogously $\hat{G}_i$, $\hat{\pi}_i$ because of the definition of Ω_2. Define for $\mathbf{x} \in M, i = 1, \dots, s$:

$$\epsilon_i(\mathbf{x}) := J_{\mathbf{x}}\{(\mathbf{x}'\boldsymbol{\beta}_i, \sigma_i^2)\}, \quad \hat{\epsilon}_i(\mathbf{x}) := \hat{J}_{\mathbf{x}}\{(\mathbf{x}'\boldsymbol{\beta}_i, \sigma_i^2)\}$$

as defined in (23). Then, by (25),

$$\forall \mathbf{x} \in M, \ i = 1, \dots, s : \epsilon_i(\mathbf{x}) = \hat{\epsilon}_i(\mathbf{x}).$$

With this result and $F^{\mathbf{X}}(M) = 1$,

$$\forall i = 1, \dots, s : \int_M g_i(\mathbf{x}^-) d\mu(\mathbf{x}^-) = 1$$

$$\Rightarrow \pi_i = \int_M \pi_i g_i(\mathbf{x}^-) d\mu(\mathbf{x}^-) = \int_M \epsilon_i(\mathbf{x}) \sum_{j=1}^{s} \pi_j g_j(\mathbf{x}^-) d\mu(\mathbf{x}^-)$$

$$= \int_M \epsilon_i(\mathbf{x}) dF^{\mathbf{X}}(\mathbf{x}) = \int_M \hat{\epsilon}_i(\mathbf{x}) dF^{\mathbf{X}}(\mathbf{x}) = \hat{\pi}_i.$$

At last, with probability 1 under $F^{\mathbf{X}} : g_i(\mathbf{x}^-) = \frac{\epsilon_i(\mathbf{x})}{\pi_i} \sum_{j=1}^{s} \pi_j g_j(\mathbf{x}^-)$

$$= \frac{\hat{\epsilon}_i(\mathbf{x})}{\hat{\pi}_i} \frac{dF^{\mathbf{X}}}{d\mu}(\mathbf{x}^-) = \hat{g}_i(\mathbf{x}^-) \Rightarrow G_i = \hat{G}_i$$

$$\Rightarrow J = \hat{J}.$$

Proof of Theorem 3.3. $F_{\mathbf{x},J} = F_{\mathbf{x},\hat{J}} \Rightarrow J \sim_{22} \hat{J}$ has to be shown. Let $F_{\mathbf{x},J} = F_{\mathbf{x},\hat{J}}$. According to the definition of "$\sim_{22}$" it suffices to show that for arbitrary $(\beta_0, \sigma_0^2, G_0) \in S(J)$

$$\exists G \in \mathcal{G} : \; (\beta_0, \sigma_0^2, G) \in S(\hat{J}). \tag{26}$$

Define $\mathcal{X}_{(\hat{\beta},\hat{\sigma}^2)} := \{\mathbf{x}_i : i \in I, (\mathbf{x}_i'\beta_0, \sigma_0^2) = (\mathbf{x}_i'\hat{\beta}, \hat{\sigma}^2)\}$. If $\dim\langle\mathcal{X}_{(\hat{\beta},\hat{\sigma}^2)}\rangle = p+1$ for some $(\hat{\beta}, \hat{\sigma}^2, \hat{G}) \in S(\hat{J})$, then $(\beta_0, \sigma_0^2) = (\hat{\beta}, \hat{\sigma}^2)$, i.e., (26).

Otherwise $\forall(\hat{\beta}, \hat{\sigma}^2, \hat{G}) \in S(\hat{J}) : \dim\langle\mathcal{X}_{(\hat{\beta},\hat{\sigma}^2)}\rangle < p+1$. Let $\mathcal{X}^-_{(\hat{\beta},\hat{\sigma}^2)}$ the set of the corresponding $\mathbf{x}_i^-$, let $H_{(\hat{\beta},\hat{\sigma}^2)}$ be a $(p-1)$-dimensional hyperplane covering $\mathcal{X}^-_{(\hat{\beta},\hat{\sigma}^2)}$, consider

$$F_J = F_{\hat{J}} \Rightarrow G_0\{\mathbf{x}_i : \; (\mathbf{x}_i'\beta_0, \sigma_0^2) \neq (\mathbf{x}_i'\hat{\beta}, \hat{\sigma}^2) \quad \forall(\hat{\beta}, \hat{\sigma}^2, \hat{G}) \in S(\hat{J})\} = 0,$$

and conclude

$$G_0\left(\bigcup_{(\hat{\beta},\hat{\sigma}^2,\hat{G})\in S(\hat{J})} H_{(\hat{\beta},\hat{\sigma}^2)}\right) = 1$$

in contradiction to (6).

Proof of Theorem 4.3: The proof is analogous to that of Theorem 3.3. Ignore the members of $\mathcal{G}$, and replace J by γ and $S(J)$ by $\gamma(I)$. Observe that $(\beta_0, \sigma_0^2) = (\hat{\beta}, \hat{\sigma}^2)$ for some $(\hat{\beta}, \hat{\sigma}^2) \in \hat{\gamma}(I)$:

Else $\bigcup_{(\hat{\beta},\hat{\sigma}^2)\in\hat{\gamma}(I)} H_{(\hat{\beta},\hat{\sigma}^2)}$ covers $\{\mathbf{x}_i^- : \; i \in I, \gamma(i) = (\beta_0, \sigma_0^2)\}$ in contradiction to $|\gamma(I)| < \min\{h(\beta, \sigma^2)\}$.

References

ANDERSON, T. W., and TAYLOR, J. B. (1976), "Strong consistency of least squares estimates in normal linear regression", *Annals of Statistics, 4,* 788-790.

BRYANT, P. G., and WILLIAMSON, J. A. (1978), "Asymptotic Behavior of Classification Maximum Likelihood Estimators", *Biometrika, 65,* 273-281.

CHANDRA, S. (1977), "On the mixtures of probability distributions", *Scandinavian Journal of Statistics, 4,* 105-112.

CHEN, J. (1995), "Optimal rate of convergence for finite mixture models", *Annals of Statistics, 23,* 221-233.

DESARBO, W. S., and CRON, W. L. (1988), "A maximum likelihood methodology for clusterwise linear regression", *Journal of Classification, 5,* 249-282.

EDELSBRUNNER, H. (1987), *Algorithms in Combinatorial Geometry,* Heidelberg: Springer-Verlag.

FAIR, R. C., and JAFFEE, D. M. (1972), "Methods of Estimation for Markets in Disequilibrium", *Econometrica, 40,* 497-514.

FENG, Z. D., and McCULLOCH, C. E. (1996), "Using bootstrap likelihood ratio in finite mixture models", *Journal of the Royal Statistical Society B, 58,*609-617.

GORDON, N. H. (1990), "Application of the theory of finite mixtures for the estimation of 'cure' rates of treated cancer patients, *Statistics in Medicine, 9,* 397-407.

HENNIG, C. (1999), "Models and Methods for Clusterwise Linear Regression", in *Classification in the Information Age*, ed. W. Gaul, and H. Locarek-Junge, Heidelberg: Springer-Verlag, 179-187.

HINDERER, K. (1970), *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter,* Heidelberg: Springer-Verlag.

HOSMER, D. W. jr. (1974), "Maximum Likelihood estimates of the parameters of a mixture of two regression lines", *Communications in Statistics, 3,* 995-1006.

HUANG, W.-T., and PAO, K.-M. (1991), "Minimum Distance Estimations in a Switching Regression Model", *Information and Management Sciences, 2,* 119-128.

KAMARKURA, W. (1988), "A least squares procedure for benefit segmentation with conjoint experiments", *Journal of Marketing Research, 25,* 157-167.

KHACHIYAN, L. (1995), "On the complexity of approximating extremal determinants in matrices", *Journal of Complexity, 11,* 138-153.

KIEFER, J., and WOLFOWITZ, J. (1956), "Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters", *Annals of Mathematical Statistics, 27,* 887-906.

KIEFER, N. M. (1978), "Discrete Parameter Variation: Efficient Estimation of a Switching Regression Model", *Econometrica, 46,* 427-434.

LI, L. A., and SEDRANSK, N. (1988), "Mixtures of distributions: A topological approach", *Annals of Statistics, 16,* 1623-1634.

LINDSAY, B. G. (1995), *Mixture Models: Theory, Geometry and Applications,* Hayward: NSF-CBMS Series, Institute of Mathematical Statistics.

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

OBERHOFER, W. (1980), "Die Nichtkonsistenz der ML-Schätzer im Switching Regression-Problem", *Metrika, 27,* 1-13.

PRAKASA RAO, B. L. S. (1992), *Identifiability in Stochastic Models: Characterizations of Probability Distributions,* San Diego: Academic Press.

QUANDT, R. E., and RAMSEY, J. B. (1978), "Estimating Mixtures of Normal Distributions and Switching Regressions", *Journal of the American Statistical Association, 73,* 730-752.

REDNER, R. (1981), "A note on the consistency of the maximum likelihood estimate for non-identifiable distributions", *Annals of Statistics, 9,* 225-228.

SEBER, G. A. F., and WILD, C. J. (1989), *Nonlinear Regression,* New York: Wiley.

TEICHER, H. (1961), "Identifiability of mixtures", *Annals of Mathematical Statistics, 34,* 244-248.

TITTERINGTON, D. M., SMITH, A. F. M., and MAKOV, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions,* Chichester: Wiley.

WEDEL, M., and STEENKAMP, J.-B. E. M. (1991), "A clusterwise regression method for simultaneous fuzzy market structuring and benefit segmentation", *Journal of Marketing Research, 28,* 385-396.

YAKOWITZ, S. J. (1969), "A consistent estimator for the identification of finite mixtures", *Annals of Mathematical Statistics, 40,* 1728-1735.

YAKOWITZ, S. J., and SPRAGINS, J. D. (1968), "On the identifiability of finite mixtures", *Annals of Mathematical Statistics, 39,* 209-214.