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 $x$ 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

$$Y = x' \beta + U, \quad \mathcal{L}(U) = \mathcal{N}_{0,\sigma^2} \text{ i.i.d.,}$$

$$Y \text{ IR-valued r.v.,} \quad x = (x_1, \ldots, x_p, 1) \in \mathbb{IR}^p \times \{1\},$$

$$\beta \in \mathbb{IR}^{p+1}, \quad \sigma^2 \in \mathbb{IR}^+_0, \quad i \in I. \tag{1}$$

The $p+1$st component of $\beta$ 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 $\sigma^2$. $I$ is some index set. Distinct regression and scale parameters $(\beta, \sigma^2)$ define distinct clusters. $x$ 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).