Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions

Salvatore Ingrassia
Università di Catania, Italy

Simona C. Minotti
Università di Milano-Bicocca, Italy

Giorgio Vittadini
Università di Milano-Bicocca, Italy

Abstract: Cluster-weighted modeling (CWM) is a mixture approach to modeling the joint probability of data coming from a heterogeneous population. Under Gaussian assumptions, we investigate statistical properties of CWM from both theoretical and numerical point of view; in particular, we show that Gaussian CWM includes mixtures of distributions and mixtures of regressions as special cases. Further, we introduce CWM based on Student-\(t\) distributions, which provides a more robust fit for groups of observations with longer than normal tails or noise data. Theoretical results are illustrated using some empirical studies, considering both simulated and real data. Some generalizations of such models are also outlined.

Keywords: Cluster-weighted modeling; Mixture models; Model-based clustering.

The authors sincerely thank the referees for their interesting comments and valuable suggestions. We also thank Antonio Punzo for helpful discussions.

Authors’ Addresses: S. Ingrassia, Dipartimento di Economia e Impresa, Università di Catania, Corso Italia 55, Catania, Italy, e-mail: s.ingrassia@unict.it; S.C. Minotti, Dipartimento di Statistica, Università di Milano-Bicocca, Via Bicocca degli Arcimboldi 8 - 20126 Milano, Italy, e-mail: simona.minotti@unimib.it; G. Vittadini, Dipartimento di Metodi Quantitativi per l’Economia e le Scienze Aziendali, Università di Milano-Bicocca, Via Bicocca degli Arcimboldi 8 - 20126 Milano, Italy, e-mail: giorgio.vittadini@unimib.it.

Published online 23 August 2012
1. Introduction

Finite mixture models provide a flexible approach to the statistical modeling of a wide variety of random phenomena characterized by unobserved heterogeneity. In these models, dating back to the work of Newcomb (1886) and Pearson (1894), the observations in a sample are assumed to arise from unobserved groups in the population; the purpose is to identify the groups and estimate the parameters of the conditional-group density functions. If no exogenous variables explain the means and the variances of each component, we refer to unconditional mixture models, i.e., the so-called finite mixtures of distributions (FMD), developed for both normal and non-normal components (e.g. Everitt and Hand 1981; Titterington, Smith and Makov 1985; McLachlan and Basford 1988; McLachlan and Peel 2000; Frühwirth-Schnatter 2005). Otherwise, we refer to conditional mixture models, i.e., finite mixtures of regression models (FMR) and finite mixtures of generalized linear models (FMGLM), (e.g. DeSarbo and Cron 1988; Jansen 1993; Wedel and DeSarbo 1995; McLachlan and Peel 2000; Frühwirth-Schnatter 2005). These models are also known as mixture-of-experts models in machine learning (Jordan and Jacobs 1994; Peng, Jacobs, and Tanner 1996; Ng and McLachlan 2007, 2008), switching regression models in econometrics (Quandt 1972), latent class regression models in marketing (DeSarbo and Cron 1988; Wedel and Kamakura 2000), and mixed models in biology (Wang, Puterman, Cockburn and Le 1996). An extension of FMR are the so-called finite mixtures of regression models with concomitant variables (FMRC) (Dayton and Macready 1988; Wedel 2002), where the weights of the mixture functionally depend on a set of concomitant variables, which may be different from the explanatory variables and are usually modeled by a multinomial logistic distribution.

The present paper focuses on a different mixture approach to modeling the joint probability of a response variable and a set of explanatory variables. The original formulation, proposed by Gershenfeld (1997) under Gaussian and linear assumptions and called cluster-weighted modeling (CWM), was developed in the context of media technology to build a digital violin with traditional inputs and realistic sound (Gershenfeld, Schöner and Metois 1999; Gershenfeld 1999, Schöner 2000; Schöner and Gershenfeld 2001). Wedel (2002)) refers to such a model as the saturated mixture regression model. Moreover, Wedel and DeSarbo (2002) propose tests of dependencies that lead to a set of nested models for market segment derivation and profiling. However, this overview leaves out general distributional assumptions and is based only on independence properties, without analyses in terms of probability density functions and posterior probabilities.