Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5247–5252
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Fitting finite mixtures of generalized linear regressions in R

Bettina Grün[a,*], Friedrich Leisch[b]

[a] *Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10/1071, A-1040 Wien, Austria*
[b] *Department of Statistics, University of Munich, Ludwigstraße 33, D-80539 München, Germany*

Available online 31 August 2006

Abstract

R package **flexmix** provides flexible modelling of finite mixtures of regression models using the EM algorithm. Several new features of the software such as fixed and nested varying effects for mixtures of generalized linear models and multinomial regression for a priori probabilities given concomitant variables are introduced. The use of the software in addition to model selection is demonstrated on a logistic regression example.

Keywords: Concomitant variable; Finite mixture; Fixed effect; Generalized linear model; R

1. Introduction

Finite mixtures of regression models are a popular method to model unobserved heterogeneity or to account for overdispersion in data. They are flexible models and in theory it is easy to modify and extend them by using more complex models for the component distribution functions and estimate the corresponding parameters, e.g., using the EM algorithm.

R (R Development Core Team, 2006) features several extension packages for estimation of mixture regression models, e.g., **fpc** for mixtures of linear regression models (Hennig, 2000) and **mmlcr** for mixed-mode latent class regression (Buyske, 2006). However, like virtually all other (non-R) implementations, they consider only a few particular types of mixture models and do not reflect the generality of the theoretical model class in the software design. R package **flexmix** (Leisch, 2004) tries to fill this gap by encapsulating the abstract statistical objects of interest into S4 classes and methods such that the resulting software can be easily extended.

This paper is organized as follows: Section 2 gives notation and the model class, the main new functions of **flexmix** are presented in Section 3, and we end with a short demonstration in Section 4. The latest development version of the package sources and all R code necessary to reproduce the results in this article are available from http://www.ci.tuwien.ac.at/research/mixtures.

* Corresponding author. Tel.: +43 1 58801 10716; fax: +43 1 58801 10798.
E-mail addresses: Bettina.Gruen@ci.tuwien.ac.at (B. Grün), Friedrich.Leisch@stat.uni-muenchen.de (F. Leisch).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.08.014

2. Model specification

We consider finite mixtures of regression models of form

$$H(y|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{w}, \boldsymbol{\alpha}) F\left(y|\boldsymbol{x}, \boldsymbol{\beta}_k, \phi_k\right),$$

where $\boldsymbol{\Theta}$ denotes the vector of all parameters, y the dependent, $\boldsymbol{x}$ the independent, $\boldsymbol{w}$ the concomitant variable, and F is the component specific distribution function. For component-wise generalized linear models (GLMs), F must be a member of the exponential family (McCullagh and Nelder, 1989). The component specific parameters are the regression coefficients $\boldsymbol{\beta}_k$ and dispersion parameters ϕ_k. The component weights π_k need to satisfy

$$\sum_{k=1}^{K} \pi_k(\boldsymbol{w}, \boldsymbol{\alpha}) = 1 \quad \text{and} \quad \pi_k(\boldsymbol{w}, \boldsymbol{\alpha}) > 0 \quad \forall k, \boldsymbol{w}, \boldsymbol{\alpha}, \tag{1}$$

where $\boldsymbol{\alpha}$ are the parameters of the concomitant variable model.

Different concomitant variable models are possible to determine the component weights (Dayton and Macready, 1988), as the mapping function only has to fulfill condition (1). In the following a multinomial logit model for the π_k is assumed with the first component as baseline.

This class of finite mixtures of GLMs with concomitant variable models is given in McLachlan and Peel (2000, p. 145). Special cases are for example random intercept models (see Follmann and Lambert, 1989; Aitkin, 1999) where the coefficients of all independent variables are assumed to be equal over the mixture components.

Our software implementation allows to specify such equality constraints for parameters over mixture components: $\left(\boldsymbol{\beta}_k', \phi_k\right)$ may be restricted to be equal over all components, to vary between groups of components, or to be different for all components. Variation between groups is referred to as varying effects with one level of nesting. In addition, each (group of) components may use different sets of covariates. Due to space restrictions we cannot give full details of parameter estimation, but extension from standard linear models (Grün and Leisch, 2006) to GLMs is rather straightforward.

3. Design principles

Functions and model formulae are first class objects in the S language, which allows in combination with the lexical scoping rules of R (Gentleman and Ihaka, 2000) for very modular software design. Rather than using text mode arguments used as switches within function bodies, **flexmix** uses driver functions to specify all aspects of the mixture model. Users can either use the growing collection of drivers distributed as part of **flexmix** or write and use their own drivers.

In a first step the (unfitted) component specific model $F\left(y|\boldsymbol{x}, \boldsymbol{\beta}_k, \phi_k\right)$ and the concomitant variable model $\pi(\boldsymbol{w}, \boldsymbol{\alpha})$ have to be specified. For this no data are needed, only the names of the independent and dependent variables and their respective interaction structure are defined.

`FLXglm()` only allows varying effects for the coefficients and the dispersion parameters. In this case the likelihood can be maximized separately for each component in the M-step of the EM algorithm. If there are also fixed and nested varying effects for the regression coefficients and dispersion parameters, our new driver `FLXglmFix()` has to be used and the likelihood is maximized simultaneously for all components. The design matrix is constructed by replicating the observations K times with suitable columns of zeros added. Model formulae for the varying, nested varying and fixed effects have to be provided. These are evaluated by successively updating the formula of the random effects with the formula for the fixed and then the nested varying effects.

The concomitant variable model is specified in a similar fashion. The default dummy driver `FLXconstant()` uses no concomitant variables and acts only as a placeholder. For multinomial logistic regression our new function `FLXmultinom()` can be used (see example section). The main estimation engine of **flexmix** has changed to be able to use the new functionality, however these are changes behind the scenes in unexported functions, all existing user code should run unaffected.