Computational Statistics and Data Analysis 71 (2014) 138–150

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A multivariate linear regression analysis using finite mixtures of t distributions

Giuliano Galimberti *, Gabriele Soffritti

Department of Statistical Sciences, University of Bologna, Italy

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form 22 November 2012
Accepted 20 January 2013
Available online 25 January 2013

Keywords:
EM algorithm
Maximum likelihood
Model identifiability
Non-normal error distribution
Unobserved heterogeneity

ABSTRACT

Recently, finite mixture models have been used to model the distribution of the error terms in multivariate linear regression analysis. In particular, Gaussian mixture models have been employed. A novel approach that assumes that the error terms follow a finite mixture of t distributions is introduced. This assumption allows for an extension of multivariate linear regression models, making these models more versatile and robust against the presence of outliers in the error term distribution. The issues of model identifiability and maximum likelihood estimation are addressed. In particular, identifiability conditions are provided and an Expectation–Maximisation algorithm for estimating the model parameters is developed. Properties of the estimators of the regression coefficients are evaluated through Monte Carlo experiments and compared to the estimators from the Gaussian mixture models. Results from the analysis of two real datasets are presented.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Linear regression analysis (see, e.g., Srivastava, 2002) is a technique that allows for the study of the dependence of D responses $\boldsymbol{Y} = (Y_1, \ldots, Y_d, \ldots, Y_D)'$ on P regressors $(X_1, \ldots, X_p, \ldots, X_P)'$, where $D \geq 1$ and $P \geq 1$. Linear regression is based on the following statistical model:

$$\boldsymbol{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{B}'\boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where the symbol i is used to denote a sample unit; $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{id}, \ldots, Y_{iD})'$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip}, \ldots, x_{iP})'$ are the D-dimensional vector of the response variables and the P-dimensional vector of the fixed regressor values for the ith unit, respectively; $\boldsymbol{\beta}_0$ is a D-dimensional vector containing the intercepts for the D responses; $\boldsymbol{B}$ is a matrix of dimension $P \times D$ whose (p, d)th element, β_{pd}, is the regression coefficient of the pth regressor on the dth response; finally, $\boldsymbol{\epsilon}_i$ denotes the D-dimensional random vector of the error terms corresponding to the ith unit. In the classical linear regression model, it is additionally assumed that $\boldsymbol{\epsilon}_i,\ i = 1, \ldots, I$, are independent and identically distributed random vectors with a Gaussian distribution with a D-dimensional zero mean vector and a positive definite covariance matrix $\boldsymbol{\Sigma}$ of dimension $D \times D$:

$$\boldsymbol{\epsilon}_i \sim N_D(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{2}$$

Many extensions of this classic model have been proposed to broaden the applicability of linear regression analysis to situations where the Gaussian error term assumption may be inadequate, for example, because of outlying values in the responses or datasets involving errors with longer than normal tails. Some such extensions rely on the use of the t distribution (see, e.g., Lange et al., 1989; Sutradhar and Ali, 1986; Zellner, 1976). In particular, a linear regression analysis

* Correspondence to: Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy. Tel.: +39 051 2098227; fax: +39 051 232153.

E-mail address: giuliano.galimberti@unibo.it (G. Galimberti).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2013.01.017

has been developed by replacing (2) with the assumption

$$\boldsymbol{\epsilon}_i \sim t_D(\mathbf{0}, \boldsymbol{\Sigma}, \nu), \tag{3}$$

where $t_D(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ denotes the D-dimensional t distribution with location parameter $\boldsymbol{\mu} \in \mathbb{R}^D$, dispersion matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{\boldsymbol{\Sigma}_D}$ and degrees of freedom $\nu \in \mathbb{R}^+$, where $\mathbb{R}^{\boldsymbol{\Sigma}_D}$ is the set of all positive definite matrices in $\mathbb{R}^{D\times D}$.

However, in practice, when nothing is known about the true distribution of the error terms, a linear regression analysis based on any of the above models may be performed using an incorrectly specified model. Furthermore, there may be situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data. To overcome these problems, solutions that use finite mixture models have been recently proposed. Namely, Bartolucci and Scaccia (2005) and Soffritti and Galimberti (2011) have developed methods for linear regression analysis by assuming a finite mixture of Gaussian components for the error terms. More specifically, in the linear regression model obtained using this approach, the assumption (2) is replaced with

$$\boldsymbol{\epsilon}_i \sim \sum_{k=1}^{K} \pi_k N_D(\boldsymbol{\delta}_k, \boldsymbol{\Sigma}_k), \tag{4}$$

where π_k's are positive weights that sum to 1, the $\boldsymbol{\delta}_k$'s are D-dimensional mean vectors that satisfy the constraint $\sum_{k=1}^{K} \pi_k \boldsymbol{\delta}_k = \mathbf{0}$ and the $\boldsymbol{\Sigma}_k$'s are positive definite covariance matrices.

In this paper, we extend this approach by assuming that the distribution of each component belongs to the class of t distributions. The rationale of such an approach is that quite complex distributions can be modelled through a finite mixture model, and thus, a more flexible modelling of the unknown error distribution of a linear regression model can be obtained. In addition, using a finite mixture model makes it possible to capture the effect of omitting relevant nominal regressors from the model. In this case, the source of unobserved heterogeneity introduced in the model will affect the error terms, whose distribution will be a mixture of K components, where K equals the number of categories obtained from the cross classification of the omitted nominal regressors. Thus, an approach based on the finite mixture model should detect the presence of such unobserved heterogeneity in the linear regression model. The model obtained under this new assumption may be particularly suitable whenever the tails of the distribution of the error terms in each component of the mixture model are heavier than those of the Gaussian distribution (Peel and McLachlan, 2000); furthermore, this model protects against the presence of outlying residuals.

The remainder of the paper is organised as follows. Section 2 provides the details of this novel class of models. In Section 2.1, we describe the multivariate linear regression model in which the error term distribution is a finite mixture of t distributions; model identifiability and maximum likelihood (ML) estimation using an Expectation–Maximisation (EM) algorithm are addressed in Section 2.2 (proofs of some results are provided in Appendices A and B). In Section 3, we present the results of Monte Carlo experiments, which provide numerical evaluations of the main properties of the estimators of the model regression coefficients. In Section 4, we report results obtained by applying the proposed methodology and other existing methods to two real datasets. Properties concerning the t distribution that are used in this paper are summarised in Appendix C.

2. Linear regression through finite mixtures of *t* distributions

2.1. The general model

We assume that the distribution of the error terms in the model (1) is the following mixture of K multivariate t distributions:

$$\boldsymbol{\epsilon}_i \sim \sum_{k=1}^{K} \pi_k t_D(\boldsymbol{\delta}_k, \boldsymbol{\Sigma}_k, \nu_k), \tag{5}$$

where π_k's are positive weights that sum to 1, the $\boldsymbol{\delta}_k$'s are D-dimensional mean vectors that satisfy the constraint $\sum_{k=1}^{K} \pi_k \boldsymbol{\delta}_k = \mathbf{0}$, the $\boldsymbol{\Sigma}_k$'s are positive definite dispersion matrices and the ν_k's are the degrees of freedom of the K mixture components, with $\nu_k \in \mathbb{R}^+\ \forall k$. In the special case where $K = 1$, this model results in the linear regression model based on the assumption (3) proposed by Lange et al. (1989). The limiting form of the probability distribution in Eq. (5) as $\nu_k \to \infty\ \forall k$ coincides with the forms considered in Eqs. (2) and (4) when $K = 1$ and $K \geq 1$, respectively.

Given Eqs. (1) and (5), the probability density function (p.d.f.) of the D-dimensional random vector $\boldsymbol{Y}_i$ is

$$\sum_{k=1}^{K} \pi_k f(\boldsymbol{y}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_k, \nu_k), \quad \boldsymbol{\mu}_{ik} = \boldsymbol{\lambda}_k + \boldsymbol{B}'\boldsymbol{x}_i,\ \boldsymbol{y}_i \in \mathbb{R}^D, \tag{6}$$

where $\boldsymbol{\lambda}_k = \boldsymbol{\delta}_k + \boldsymbol{\beta}_0$ and $f(\boldsymbol{y}_i; \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_k, \nu_k)$ is the p.d.f. of the distribution $t_D(\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_k, \nu_k)$ evaluated at $\boldsymbol{y}_i$ (see Eq. (C.1) in Appendix C). The vector of the model parameters is $\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{b}', \boldsymbol{\lambda}', \boldsymbol{\sigma}', \boldsymbol{\nu}')'$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K-1})'$, $\boldsymbol{b} = \text{vec}(\boldsymbol{B})$ denotes the vector formed by stacking the columns of the matrix $\boldsymbol{B}$, one underneath the other, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_K')'$, $\boldsymbol{\sigma} =$

دریافت فوری ← متن کامل مقاله

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات