Statistics and Probability Letters 130 (2017) 32–39

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Robust mixture multivariate linear regression by multivariate Laplace distribution

Xiongya Li[a], Xiuqin Bai[b], Weixing Song[a,*]

[a] Department of Statistics, Kansas State University, Manhattan, KS 66506, United States
[b] Department of Mathematics, Eastern Washington University, Cheney, WA, 99004, United States

ARTICLE INFO

Article history:
Received 30 January 2017
Received in revised form 29 June 2017
Accepted 29 June 2017
Available online 8 July 2017

MSC 2010:
primary 62F35
secondary 62F10

Keywords:
Finite mixtures
Multivariate linear regression
Robust estimation
Multivariate Laplace distribution
EM algorithm

ABSTRACT

Assuming that the error terms follow a multivariate Laplace distribution, we propose a robust estimation procedure for mixture of multivariate linear regression models in this paper. Using the fact that the multivariate Laplace distribution is a scale mixture of the multivariate standard normal distribution, an efficient EM algorithm is designed to implement the proposed robust estimation procedure. The performance of the proposed algorithm is thoroughly evaluated by some simulation and comparison studies.

1. Introduction

Finite mixture regression modeling is an efficient tool to investigate the relationship between a response variable and a set of predictors when the underlying population consists of several unknown latent homogeneous groups, and it has been already applied for more than a hundred years since Newcomb (1886). More real examples on finite mixture modeling can be found in Jiang and Tanner (1999), Böhning (2000), McLachlan and Peel (2004), Wedel and Kamakura (2012) and the references therein. Statistical inferences have been discussed extensively for finite mixture modeling when the normality is assumed for the regression error in each cluster. Due to the untractable likelihood function for normal mixture regression models, the unknown regression parameters are often estimated via the expectation and maximization (EM) algorithm. However, the unweighted least squares nature makes the maximum likelihood estimate (MLE) of the regression parameters susceptible of non-robustness to the outliers and the data with heavy tails. Because of its wide application in practice, how to design robust estimation procedures in the finite mixture regression models has attracted much attention from statisticians.

Extensive research has been done for linear or mixture of linear regression models when the response variable is univariate. For examples, Neykov et al. (2007) proposed a trimmed likelihood estimator (TLE) to robustly estimate the mixtures and the breakdown points of the TLE for the mixture component parameters is also characterized; replacing the least square criterion in the M step of EM algorithm designed for normal mixtures, Bai et al. (2012) achieved robustness using Tukey's bisquare and Huber's ψ-functions; a class of S-estimators were introduced in Bashir and Carter (2012) and Farcomeni and Greco (2015) which exhibit certain robustness and the parameter estimation is achieved via an

* Corresponding author.
E-mail address: weixing@ksu.edu (W. Song).

http://dx.doi.org/10.1016/j.spl.2017.06.028
0167-7152/

expectation–conditional maximization algorithm. Inspired by Peel and McLachlan (2000), Yao et al. (2014) proposed a new robust estimation method for mixture of linear regression by assuming that the mixtures have t-distributions, the EM algorithm is made possible by the fact that t-distribution is a scale mixture of a normal distribution. Due to the selection of degrees of freedom, the procedure in Yao et al. (2014) requires relatively heavy computation although the choice of degrees of freedom provides certain adaptivity to the data. Realizing that the Laplace distribution is also a scale mixture of normal distribution, Song et al. (2014) proposed an alternative robust estimation procedure by assuming the random error has a Laplace distribution, which has a natural connection with the least absolute deviation (LAD) procedure, see Dielman (1984), Li and Arce (2004), and Dielman (2005) for more detail on LAD methodology.

Comparing to the relatively extensive discussion for the univariate response cases, there are fewer work having been done for the multivariate linear regressions. Lin (2010) designed a robust estimation procedure using the multivariate skewed t-distribution, which offers a great deal of flexibility that accommodates asymmetry and heavy tails simultaneously. Xian Wang et al. (2004) proposed a mixture of multivariate t-distribution to fit the multivariate continuous data with a large number of missing values. We have not seen any work on developing robust estimation procedures for the multivariate linear regression with the multivariate Laplace distribution. We wish there is a multivariate version of Song et al.'s (2014) procedure which should perform equally well in the multivariate linear regression. This is the motivation of the research conducted in the current paper.

The paper is organized as follows. Section 2 introduces the mixture of multivariate linear regression models, and also the definition of the multivariate Laplace distribution, some essential properties of the multivariate Laplace distribution is also discussed. The EM algorithm will be developed in Section 3 for the mixture of multivariate linear regression models. Section 4 includes some simulation and comparison studies to evaluate the performance of the proposed methods.

2. Statistical model and multivariate Laplace distribution

We begin with a brief introduction on the mixture of multivariate linear regression models, and a definition of multivariate Laplace distribution.

2.1. Mixture of multivariate linear regression

Let G be a latent class variable such that given $G = j, j = 1, 2, \ldots, g, g \geq 1$, a p-dimensional response Y and a q-dimensional predictor X are in one of the following multivariate linear regression models

$$Y = \beta_j' X + \varepsilon_j, \tag{1}$$

where, for each j, β_j is a $q \times p$ unknown regression coefficient matrix, and ε_j is a p-dimensional random error. Assume ε_j's are independent of X and it is commonly assumed that the density functions f_j of ε_j's are members in a location-scale family with mean 0 and covariance Σ_j. If we further assume that $P(G = j) = \pi_j, j = 1, \ldots, g$. Then conditioning on X, the density function of Y is given by

$$f(y|x, \theta) = \sum_{j=1}^{g} \pi_j f_\varepsilon(y - \beta_j' x, 0, \Sigma_j), \tag{2}$$

where $\theta = \{\pi_1, \beta_1, \Sigma_1, \ldots, \pi_g, \beta_g, \Sigma_g\}$. The model (2) is the so called mixture multivariate regression models. The unknown parameters could be estimated by the maximum likelihood estimator (MLE), which maximizes the log-likelihood function (3) based on an independent sample $(X_i, Y_i), i = 1, \ldots, n$ from (2),

$$L_n(\theta) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{g} \pi_j f_\varepsilon(Y_i, \beta_j' X_i, \Sigma_j) \right]. \tag{3}$$

If $g = 1$, then the mixture linear regression model is simply a multivariate linear regression model.

The traditional maximum likelihood estimation procedure is based on the normality assumption. However, no explicit solution is available due to the untractable expression of (3), and EM algorithm thus developed to obtain its the maximizer. As we mentioned in Section 1, the MLE based on the normality assumption is sensitive to outliers or heavy-tailed error distribution, and we shall develop a robust estimation procedure by assuming that the error distributions are Laplacian.

2.2. Multivariate Laplace distribution

There are multiple forms of definitions of the multivariate Laplace distribution. For example, the bivariate case was introduced by Ulrich and Chen (1987), and the first form in larger dimensions was discussed in Fang et al. (1990). Later, the multivariate Laplace was introduced as a special case of a multivariate Linnik distribution in Anderson (1992), and the multivariate power exponential distribution in Fernandez et al. (1995) and Ernst (1998). Portilla et al. (2003) presented multivariate Laplace distribution as a Gaussian scale mixture. Kotz et al. (2012) presented the multivariate Laplace distribution formally and thoroughly discussed its probability properties. The multivariate Laplace distribution is