Computational Statistics and Data Analysis 93 (2016) 177–191

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Laplace mixture of linear experts

Hien D. Nguyen *, Geoffrey J. McLachlan

School of Mathematics and Physics, University of Queensland, Australia

ARTICLE INFO

Article history:
Received 14 April 2014
Received in revised form 14 October 2014
Accepted 16 October 2014
Available online 23 October 2014

Keywords:
Laplace distribution
Minorization–maximization algorithm
Mixture of experts
Robust regression

ABSTRACT

Mixture of Linear Experts (MoLE) models provide a popular framework for modeling nonlinear regression data. The majority of applications of MoLE models utilizes a Gaussian distribution for regression error. Such assumptions are known to be sensitive to outliers. The use of a Laplace distributed error is investigated. This model is named the Laplace MoLE (LMoLE). Links are drawn between the Laplace error model and the least absolute deviations regression criterion, which is known to be robust among a wide class of criteria. Through application of the minorization–maximization algorithm framework, an algorithm is derived that monotonically increases the likelihood in the estimation of the LMoLE model parameters. It is proven that the maximum likelihood estimator (MLE) for the parameter vector of the LMoLE is consistent. Through simulation studies, the robustness of the LMoLE model over the Gaussian MOLE model is demonstrated, and support for the consistency of the MLE is provided. An application of the LMoLE model to the analysis of a climate science data set is described.

1. Introduction

Mixture of experts (MoE) models were first introduced in Jacobs et al. (1991) as a model for nonlinear regression relationships; see Jordan and Jacobs (1994) and Section 5.12 of McLachlan and Peel (2000) for details. Since their inception, the development in MoE research has been rich, and the framework has been successfully applied to problems of clustering, classification, and regression in a variety of fields. A review of the current state of the art can be found in Yuksel et al. (2012). The MoE framework can be defined as follows.

Let $Z \in \{1, \ldots, g\}$ be a categorical random variable such that

$$\mathbb{P}(Z = i|\boldsymbol{v}) = \begin{cases} \dfrac{\exp\left(\alpha_i^T \boldsymbol{v}\right)}{1 + \sum_{i'=1}^{g-1} \exp\left(\alpha_{i'}^T \boldsymbol{v}\right)} & \text{if } i = 1, \ldots, g-1, \\ \dfrac{1}{1 + \sum_{i'=1}^{g-1} \exp\left(\alpha_{i'}^T \boldsymbol{v}\right)} & \text{otherwise,} \end{cases}$$
$$= \pi_i(\boldsymbol{v}; \alpha), \quad (1)$$

for some covariate $\boldsymbol{v} \in \mathbb{R}^p$, and let $Y \in \mathbb{R}$ be a random variable such that $Y|Z = i$ (for $i = 1, \ldots, g$) has density function $f_i(y|\boldsymbol{x})$, which we refer to as the component density, for some covariate $\boldsymbol{x} \in \mathbb{R}^q$. Here, T denotes matrix transposition and $\alpha = \left(\alpha_1^T, \ldots, \alpha_{g-1}^T\right)^T \in \mathbb{R}^{(g-1)p}$.

* Correspondence to: Department of Mathematics, University of Queensland, St. Lucia, 4072, Australia. Tel.: +61 7 3346 7958.
E-mail addresses: h.nguyen7@uq.edu.au (H.D. Nguyen), g.mclachlan@uq.edu.au (G.J. McLachlan).

http://dx.doi.org/10.1016/j.csda.2014.10.016
0167-9473/

If one observes Y without observing Z (i.e. Z is a latent variable), then the density function of $Y|\boldsymbol{v}, \boldsymbol{x}$ can be written as

$$f_Y(y|\boldsymbol{v}, \boldsymbol{x}) = \sum_{i=1}^{g} \pi_i(\boldsymbol{v}; \alpha) f_i(y|\boldsymbol{x}). \tag{2}$$

Density functions of form (2) are known as MoE models. In this article, we concentrate on the case when $Y \in \mathbb{R}$ and $\mathbb{E}(Y|Z = i, \boldsymbol{x}) = \boldsymbol{\beta}_i^T \boldsymbol{x}$, where $\boldsymbol{\beta}_i \in \mathbb{R}^q$. We shall refer to such densities as mixture of linear experts (MoLE) models.

MoLE models have recently received strong interest from the computational statistics and neural computation communities. For example, Wedel (2002) and Grun and Leisch (2008) considered MoLE densities for modeling concomitant variables; Ingrassia et al. (2012) showed it to be related to cluster-weighted modeling; and Chamroukhi et al. (2009, 2010), and Same et al. (2011) applied it to fit, classify, and cluster time-series data, respectively.

Although a rich class, the current research in MoLE models has been restrictive in the sense that $f_i(y|\boldsymbol{x})$ is always considered to be Gaussian (i.e. $f_i(y|\boldsymbol{x}) = \phi\left(y; \boldsymbol{\beta}_i^T \boldsymbol{x}, \sigma_i^2\right)$, where $\phi\left(y; \mu, \sigma^2\right)$ is the Gaussian density function with mean $\mu \in \mathbb{R}$, and variance $\sigma^2 \in (0, \infty)$). We will call this model the Gaussian MoLE (GMoLE).

When misapplied, the Gaussian assumption is known to incur problems in mixture models (e.g. misspecification error, and outlier sensitivity; see p. 221 of McLachlan and Peel (2000) for a brief discussion), which can often lead to incorrect inference making. As a remedy to these problems, the mixture model literature has extended in scope to using robust generalizations of the Gaussian densities as component functions (e.g. Lee and McLachlan (2013) recently reviewed a variety of skewed-generalizations of Gaussian mixture models). Outside of Gaussian generalizations, Jones and McLachlan (1990) have considered Laplace distribution components for modeling data that departs from the Gaussian assumption and Franczak et al. (2014) have considered mixtures of asymmetric Laplace distributions for density estimation.

In the mixtures of linear regression context, Galimberti and Soffritti (2014), Ingrassia et al. (2014), and Yao et al. (2014) have suggested the use of the t-distribution as an error model in various settings; and Song et al. (2014) have considered the use of Laplace distributed errors. When $g \geq 2$, the density function for the Laplace mixture from Song et al. (2014) can be expressed in the form of (2) by taking $\boldsymbol{v} = 1$, and setting $f_i(y|\boldsymbol{x}) = \lambda\left(y; \boldsymbol{\beta}_i^T \boldsymbol{x}, \xi_i\right)$, where

$$\lambda(y; \mu, s) = \frac{\exp\left(-|y - \mu| / \xi\right)}{2\xi}, \tag{3}$$

is the Laplace density function with mean $\mu \in \mathbb{R}$ and scale parameter $\xi \in (0, \infty)$. Unlike MoLE models, the aforementioned mixtures of linear regression models do not allow for covariate dependencies in the component probabilities.

Considering the state of current research in MoLE models (with respect to distributional assumptions), we believe that an extension of the Laplace mixture to the more general MoLE setting (i.e. model (2) with $f_i(y|\boldsymbol{x}) = \lambda\left(y; \boldsymbol{\beta}_i^T \boldsymbol{x}, \xi\right)$) is timely and pertinent. We name our new model the Laplace MoLE (LMoLE). The following considerations regarding the model shall be discussed in this article.

Firstly, we will discuss maximum likelihood estimation (MLE) for the LMoLE model parameters. Like other MoE models, MLE for LMoLE models cannot be conducted in closed form; as such, an iterative numerical scheme is required for MLE. Unlike previous works (e.g. Jordan and Jacobs (1994) and Grun and Leisch (2008)), we do not use an expectation–maximization (EM) algorithm for the task; see McLachlan and Krishnan (2008) for a treatment on EM algorithms. This is because EM algorithms require specialist knowledge of probabilistic characterizations in order to express the iterative updates (e.g. Song et al. (2014) required a Gaussian scale mixture representation to express their updates). Furthermore, to the best of our knowledge, all current EM algorithms for MLE of MoE model parameters require a Newton or quasi-Newton update step (e.g. Jordan and Jacobs (1994) and Grun and Leisch (2008), respectively), which can violate the usual monotonicity property of EM algorithms.

Instead of an EM algorithm, we suggest a monotonic iterative scheme using the minorization–maximization (MM) algorithm framework; see Hunter and Lange (2004) for a concise introduction to MM algorithms. The MM algorithms are attractive due to their use of analytic inequalities, rather than probabilistic characterizations, in order to construct iterative schemes.

We then show the relationship between LMoLE and least absolute deviations (LAD) regression; treatments on LAD and related regression methods can be found in Maronna et al. (2006) and Section 2.3 of Amemiya (1985). Such a relationship indicates that LMoLE models should be more robust than GMoLE models, in the sense of Huber and Ronchetti (2009, Ch. 7).

Next, we show that the maximum likelihood estimator (MLE) for the parameter vector of the LMoLE model is consistent. Estimates for various quantities of interest are also given. We then use simulations to provide empirical validation that the estimates appear to converge to their true values (as predicted by the consistency of the MLE), and demonstrate situations whereby the LMoLE is robust in comparison to the GMoLE.

Lastly, we demonstrate the LMoLE model via an application to climate science data. Here, we describe an analysis of temperature anomalies data from Hansen et al. (2001).

The article will proceed as follows. The LMoLE is defined, and the MM algorithm for its MLE is presented in Section 2. The relationship between LMoLE and LAD regressions is also given here. Theoretical results and derivations of quantities of interest are then given in Section 3. Empirical evidence of practical and theoretical claims is provided via simulations in Section 4. A short application of LMoLE to climate data is described in Section 5. Finally, conclusions are drawn in Section 6.