Computational Statistics and Data Analysis 94 (2016) 75–85

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A semiparametric scale-mixture regression model and predictive recursion maximum likelihood

Ryan Martin [a,*], Zhen Han [b]

[a] Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, 851 S. Morgan St, 322 Science and Engineering Offices, Chicago, IL 60607, United States

[b] Department of Statistics, North Carolina State University, 5109 SAS Hall, 2311 Stinson Drive, Raleigh, NC 27695, United States

ARTICLE INFO

Article history:
Received 14 June 2013
Received in revised form 5 August 2015
Accepted 6 August 2015
Available online 13 August 2015

Keywords:
EM algorithm
Dirichlet process
Marginal likelihood
Nonparametric maximum likelihood
Normal scale mixture
Profile likelihood

ABSTRACT

To avoid specification of a particular distribution for the error in a regression model, we propose a flexible scale mixture model with a nonparametric mixing distribution. This model contains, among other things, the familiar normal and Student-t models as special cases. For fitting such mixtures, the predictive recursion method is a simple and computationally efficient alternative to existing methods. We define a predictive recursion-based marginal likelihood function, and estimation of the regression parameters proceeds by maximizing this function. A hybrid predictive recursion–EM algorithm is proposed for this purpose. The method's performance is compared with that of existing methods in simulations and real data analyses.

1. Introduction

Consider the standard linear regression model,

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is a $n \times 1$ vector of response variables, $\mathbf{X}$ is a $n \times p$ matrix of predictor variables, with ith row $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is an $n \times 1$ vector of independent and identically distributed (i.i.d.) errors with common density f. In classical linear model applications, one assumes that f is a normal distribution with mean zero and unknown variance σ^2. In this case, the ordinary least squares method provides the optimal estimates of $(\boldsymbol{\beta}, \sigma^2)$. However, if f happens to be non-normal, in particular, if f has heavier-than-normal tails, then the accuracy of the ordinary least squares solutions is lost.

When the error density f may be non-normal, one might consider an alternative to the normal model and ordinary least squares. Model-free alternatives based on M-estimation (e.g., Huber, 1973, 1981) include methods based on minimizing an objective function different from the sum of squared residuals, such as least absolute deviation, or L_1 regression. Surveys of these standard techniques are given in Rousseeuw and Leroy (1987) and Ryan (2009). If a likelihood-based method is preferred, then one common approach is to model the errors by a heavy-tailed Student-t distribution; see, for example, Lange et al. (1989), Liu (1996), and Pinheiro et al. (2001). The standard implementation of this approach uses the expectation–maximization (EM) algorithm (Dempster et al., 1977), which is based on a representation of the Student-t distribution as a scale mixture of normals (Andrews and Mallows, 1974; West, 1987). The goal of this paper is to explore a more general version of this latter heavy-tailed model.

* Corresponding author.
E-mail addresses: rgmartin@uic.edu (R. Martin), zhan3@ncsu.edu (Z. Han).

http://dx.doi.org/10.1016/j.csda.2015.08.005

Motivated by the Student-t's normal scale mixture representation, we consider a more general regression model specified by an arbitrary normal scale mixture for the error distribution. Specifically, we write the error density f as a mixture

$$f(\varepsilon) = \int_0^\infty \mathsf{N}(\varepsilon \mid 0, u^2)\, \Psi(du), \tag{1}$$

where Ψ is an unspecified mixing distribution supported on $(0, \infty)$. By symmetry of the normal kernel, the density f is symmetric. Moreover, (1) contains both the normal model, $\mathsf{N}(0, \sigma^2)$ and the Student-t model, $\mathsf{t}_\nu(0, \sigma)$, as special cases, corresponding to Ψ a point-mass at σ and a scaled inverse chi-square distribution, respectively. Since Ψ is completely unspecified, an additional scale parameter would not be identifiable so, without loss of generality, Ψ fully characterizes the error distribution in our regression model.

To fit this new semiparametric regression model, estimation of both $\boldsymbol{\beta}$ and Ψ is required and, even though the mixing distribution Ψ is a nuisance parameter, care is needed. Maximum likelihood and Bayes approaches can be developed, and we discuss the computational challenges faced by these in Section 2.1. The main contribution of this paper is a computationally efficient alternative, an extension of the predictive recursion (PR) method discussed in Newton et al. (1998), Newton (2002), Ghosh and Tokdar (2006), Martin and Ghosh (2008), Tokdar et al. (2009), and Martin and Tokdar (2009). The PR algorithm was originally designed for fast nonparametric estimation of a mixing distribution of a mixture model, but Martin and Tokdar (2011) developed a PR-driven marginal likelihood approach for estimating structural parameters in semiparametric mixture models; see Section 2.2 for a brief review of the PR algorithm and related methods. Previous applications of PR focused on location mixtures, and the special scale mixture formulation in this paper requires new ideas. After writing down the PR marginal likelihood for the semiparametric regression problem, in Section 3.2 we propose a hybrid PR–EM strategy that takes advantage of the latent scale parameter structure in the mixture model (1). This hybrid algorithm is fast and easy to compute, and in Section 3.3 we provide some theoretical support for its ascent property. Some remarks on the robustness of the PR method are given in Section 3.4. Section 4 demonstrates numerically that our proposed approach provides accurate estimation of $\boldsymbol{\beta}$ compared to existing methods across a range of different error distributions. Section 5 provides some concluding remarks.

2. Background

2.1. Challenges faced by standard approaches

There are two natural likelihood-based approaches that one could consider for fitting the semiparametric regression model with error distribution (1). The first is via nonparametric maximum likelihood. Start by writing a joint likelihood function for $(\boldsymbol{\beta}, \Psi)$:

$$L(\boldsymbol{\beta}, \Psi) = \prod_{i=1}^n \int_0^\infty \mathsf{N}(y_i - \boldsymbol{x}_i\boldsymbol{\beta} \mid 0, u^2)\Psi(du).$$

Next, define a profile likelihood $L_p(\boldsymbol{\beta}) = L(\boldsymbol{\beta}, \hat{\Psi}_{\boldsymbol{\beta}})$, where $\hat{\Psi}_{\boldsymbol{\beta}}$ is the conditional maximum likelihood estimator of Ψ for the given $\boldsymbol{\beta}$. Then $L_p(\boldsymbol{\beta})$ can be treated like a usual likelihood function, to produce estimators, tests, or confidence regions for $\boldsymbol{\beta}$. Existing algorithms for nonparametric maximum likelihood estimation of mixing distributions (e.g., Wang, 2007) can be used to compute $\hat{\Psi}_{\boldsymbol{\beta}}$ and, in turn, the profile likelihood $L_p(\boldsymbol{\beta})$. We claim that this profile likelihood function is generally rough, so optimization over $\boldsymbol{\beta}$ is unstable and computationally expensive. To justify this claim, we consider a simple special case of the regression problem with no predictor variables, i.e., i.i.d. data with location β. In this case, using Wang's algorithm, we can easily evaluate and plot the profile likelihood function. An independent sample of size $n = 100$ was drawn from a Student-t distribution with df $= 2$, centered at $\beta = 0$, and the corresponding likelihood functions for β are plotted in Fig. 1. The profile likelihood has a number of local modes, so numerical optimization is unstable. On the other hand, the likelihood function for our proposed method, described in Section 3, is smooth with one global mode, so optimization is fast and easy.

A second approach is based on nonparametric Bayes, where a prior distribution for Ψ is introduced. A reasonable choice would be to take a Dirichlet process prior for Ψ (e.g., Ferguson, 1973; Lo, 1984; Müller and Quintana, 2004). The idea is to integrate out Ψ from the joint likelihood $L(\boldsymbol{\beta}, \Psi)$ with respect to the prior, leaving a marginal likelihood function $L_m(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$. Markov chain Monte Carlo algorithms (e.g., Escobar and West, 1995; MacEachern and Müller, 1998; Neal, 2000; Carvalho et al., 2010), as well as software (e.g., Jara et al., 2011), are available for evaluating this marginal likelihood but this is too expensive because each marginal likelihood evaluation requires its own Monte Carlo run, and optimization requires several such runs. One can avoid repeatedly running Monte Carlo if $\boldsymbol{\beta}$ is assigned a proper prior. That is, one can employ the technique in Chib (1995) to get a marginal likelihood for $\boldsymbol{\beta}$ from a single, joint Monte Carlo run for $(\boldsymbol{\beta}, \Psi)$. This single joint Monte Carlo is generally more expensive than our proposed PR-driven strategy, so we do not explore this further here.

2.2. Review of predictive recursion

PR is a fast algorithm designed for recursive estimation of mixing distributions in nonparametric mixture models. It was first proposed as an alternative to Markov chain Monte Carlo methods in fitting Bayesian Dirichlet process mixture models