Computational Statistics and Data Analysis 78 (2014) 218–234

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A sampling algorithm for bandwidth estimation in a nonparametric regression model with a flexible error density

Xibin Zhang [a,*], Maxwell L. King [a], Han Lin Shang [b]

[a] Department of Econometrics and Business Statistics, Monash University, Australia

[b] Research School of Finance, Actuarial Studies and Applied Statistics, Australian National University, Australia

ARTICLE INFO

Article history:
Received 5 March 2013
Received in revised form 24 April 2014
Accepted 25 April 2014
Available online 4 May 2014

Keywords:
Bayes factors
Kernel-form error density
Metropolis–Hastings algorithm
Posterior predictive density
State-price density
Value-at-risk

ABSTRACT

The unknown error density of a nonparametric regression model is approximated by a mixture of Gaussian densities with means being the individual error realizations and variance a constant parameter. Such a mixture density has the form of a kernel density estimator of error realizations. An approximate likelihood and posterior for bandwidth parameters in the kernel-form error density and the Nadaraya–Watson regression estimator are derived, and a sampling algorithm is developed. A simulation study shows that when the true error density is non-Gaussian, the kernel-form error density is often favored against its parametric counterparts including the correct error density assumption. The proposed approach is demonstrated through a nonparametric regression model of the Australian All Ordinaries daily return on the overnight FTSE and S&P 500 returns. With the estimated bandwidths, the one-day-ahead posterior predictive density of the All Ordinaries return is derived, and a distribution-free value-at-risk is obtained. The proposed algorithm is also applied to a nonparametric regression model involved in state-price density estimation based on S&P 500 options data.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

A simple and commonly used estimator of the regression function in a nonparametric regression model is the Nadaraya–Watson (NW) estimator, whose performance is mainly determined by the choice of bandwidths. A large literature exists on bandwidth selection for the NW estimator, and the most popular approaches are the rule-of-thumb, cross-validation (CV), plug-in and bootstrapping methods (see for example, Härdle, 1990; Herrmann et al., 1995; Hall et al., 1995). Even though the NW estimator does not require an assumption on the analytical form of the error density, it is often of great interest to investigate the distribution of the response around the estimated mean. Such a distribution is characterized by the error density, estimation of which is a fundamental issue in statistical inference for any regression model. This issue was extensively discussed by Efromovich (2005), who developed a nonparametric approach to error-density estimation in a nonparametric regression model using residuals as proxies of errors.

A simple approach to error density estimation is the kernel density estimator of residuals, whose performance is mainly determined by the choice of bandwidth. This density estimator depends on residuals fitted through the NW estimator of the regression function. Moreover, the resulting density estimator of residuals provides no information for the purpose of choosing bandwidths in the NW regression estimator, although bandwidth selection in this situation depends on the

* Correspondence to: 900 Dandenong Road, Caulfield East, Victoria 3145, Australia. Tel.: +61 3 99032130; fax: +61 3 99032007.
E-mail addresses: xibin.zhang@monash.edu (X. Zhang), max.king@monash.edu (M.L. King), hanlin.shang@anu.edu.au (H.L. Shang).

http://dx.doi.org/10.1016/j.csda.2014.04.016
0167-9473/© 2014 Elsevier B.V. All rights reserved.

error distribution (see for example, Zhang et al., 2009). Therefore, there is a lack of a data-driven procedure for choosing bandwidths for the two estimators simultaneously. This motivates the study reported in this paper.

Our investigation of error density estimation is also motivated by its practical applications. In financial econometrics, an important use of the estimated error density in modeling an asset return is to estimate the value-at-risk (VaR). At a given confidence level $100(1-\alpha)\%$ with α being a small probability value, the VaR is a threshold loss of an investment on an asset over a given period of time in the sense that there is a $100\alpha\%$ chance of a loss as great as or greater than the VaR loss for the given time period. When the density of this asset's return is obtained, the VaR is defined as the magnitude of this density's lower $100\alpha\%$ quantile. The VaR is an important measure of the risk involved in holding the investment and is used to help manage risk (see Jorion, 1997, among others). In the nonparametric regression model that we consider, any mis-specification of the error density may produce an inaccurate estimate of the VaR and make it harder to manage the risk. Therefore, being able to estimate the error density can be just as important as being able to estimate the mean of the regression model.

Let $\boldsymbol{y}$ denote the response and $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)'$ a set of explanatory variables or regressors. Given observations $(y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \ldots, n$, a nonparametric regression model is expressed as

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i, \tag{1}$$

where ε_i, for $i = 1, 2, \ldots, n$, are assumed to be independent and identically distributed (iid) with an unknown density denoted as $f(\varepsilon)$. Let the NW estimator of the regression function be denoted as $\widehat{m}(\boldsymbol{x}; \boldsymbol{h})$ with $\boldsymbol{h}$ a vector of bandwidths. In this paper, we assume that the unknown $f(\varepsilon)$ is approximated by a kernel-form density given by

$$f(\varepsilon; b) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{b}\phi\left(\frac{\varepsilon - \varepsilon_i}{b}\right), \tag{2}$$

where $\phi(\cdot)$ is the probability density function of the standard Gaussian distribution.

The density function given by (2) is a mixture of n Gaussian densities, and the component densities have a common standard deviation b and means ε_i, for $i = 1, 2, \ldots, n$. From the viewpoint of kernel smoothing, this error density is of the form of a kernel density estimator of the errors (rather than residuals) with $\phi(\cdot)$ the kernel function and b the bandwidth. Consequently, it is reasonable to expect that $f(\varepsilon; b)$ can approximate $f(\varepsilon)$ well when $f(\varepsilon)$ is unknown. We call (2) the kernel-form error density, and b is referred to as the bandwidth.

We aim to develop a sampling algorithm, through which the bandwidths, $\boldsymbol{h}$ and b, can be simultaneously estimated. We treat bandwidths as parameters and conduct our investigation in a parametric way although the underlying model is nonparametric. Our main contribution is to construct an approximate likelihood and therefore, the posterior of bandwidth parameters for the nonparametric regression model with its unknown error density approximated by the kernel-form error density given by (2).

When the iid errors follow a Gaussian distribution, Zhang et al. (2009) derived an approximate posterior of $\boldsymbol{h}$ for given $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, where the likelihood of $\boldsymbol{y}$ for given $\boldsymbol{h}$ is the product of the Gaussian densities of y_i with its mean approximated by the leave-one-out NW estimator denoted as $\widehat{m}_i(\boldsymbol{x}_i; \boldsymbol{h})$, for $i = 1, 2, \ldots, n$. The error density can be assumed to be of other parametric forms such as a mixture of Gaussian densities. However, any parametric assumption of the error density is likely to be wrong, and subsequent inference might be misleading. The contribution of this paper is not only a relaxation of the Gaussian error assumption of Zhang et al. (2009), but also a novel sampling algorithm under a flexible error density in regression models.

There is a growing literature on the estimation of the error density in a nonparametric regression model. Efromovich (2005) presented the so-called Efromovich–Pinsker estimator of the error density and showed that this estimator is asymptotically as accurate as an oracle that knows the underlying errors. Cheng (2004) showed that the kernel density estimator of residuals is uniformly, weakly and strongly consistent. When the regression function is estimated by the NW estimator and the error density is estimated by the kernel estimator of residuals, Samb (2011) proved the asymptotic normality of the bandwidths in both estimators and derived the optimal convergence rates of the two types of bandwidths. Linton and Xiao (2007) proposed a kernel estimator based on residuals obtained through local polynomial fitting of the unknown regression function. They showed that their estimator is adaptive and concluded that the adaptive estimation is possible in local polynomial fitting, which includes the NW estimator as a special case. In a class of nonlinear regression models, Yuan and de Gooijer (2007) constructed an approximate likelihood through the kernel density estimator of pre-fitted residuals with its bandwidth pre-chosen by the rule-of-thumb. They proved that under some regularity conditions, the resulting maximum likelihood estimates of parameters are consistent, asymptotically normal and efficient. Jaki and West (2008) proposed using the kernel density estimator of the pre-fitted residuals to construct an approximate likelihood, which they called the kernel likelihood.

In all these investigations, residuals were commonly used as proxies of errors, and the bandwidth for the kernel density estimator of residuals was pre-chosen. To our knowledge, there is no method that can simultaneously estimate the bandwidths for the NW estimator of the regression function and the kernel-form error density.

Our proposed kernel-form error density is robust in terms of different specifications of the error density in a nonparametric regression model. In order to understand the relative gains and losses that result from this robust assumption against other parametric assumptions, we conduct simulation studies by simulating samples through a nonlinear regression function, where the error densities are respectively, the Gaussian and several mixture densities of two Gaussians. We find that the

Download English Version:

https://daneshyari.com/en/article/415408

Download Persian Version:

https://daneshyari.com/article/415408

Daneshyari.com