Computational Statistics and Data Analysis 57 (2013) 336–348

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Statistical inference and visualization in scale-space using local likelihood

Cheolwoo Park[a], Jib Huh[b,*]

[a] Department of Statistics, University of Georgia, Athens, GA 30602, USA

[b] Department of Statistics, Duksung Women's University, Seoul 132-714, Republic of Korea

ARTICLE INFO

Article history:
Received 20 June 2011
Received in revised form 25 June 2012
Accepted 27 June 2012
Available online 2 July 2012

Keywords:
Generalized linear models
Local likelihood
Local polynomial smoothing
Scale-space
Statistical significance

ABSTRACT

SiZer (SIgnificant ZERo crossing of the derivatives) is a graphical scale-space visualization tool that allows for exploratory data analysis with statistical inference. Various SiZer tools have been developed in the last decade, but most of them are not appropriate when the response variable takes discrete values. In this paper, we develop a SiZer for finding significant features using a local likelihood approach with local polynomial estimators. This tool improves the existing one (Li and Marron, 2005) by proposing a theoretically justified quantile in a confidence interval using advanced distribution theory. In addition, we investigate the asymptotic properties of the proposed tool. We conduct a numerical study to demonstrate the sample performance of SiZer using Bernoulli and Poisson models using simulated and real examples.

1. Introduction

Nelder and Wedderburn (1972) introduced generalized linear models as a means of applying techniques used in ordinary linear regression to more general settings. Green and Silverman (1994) studied an extension of the smoothing spline methodology to generalized linear models. Additionally, Fan et al. (1995) investigated the extension of the nonparametric regression technique of local polynomial fitting with a kernel weight to generalized linear models and quasi-likelihood contexts. In the case of the multiple-covariates, Carroll et al. (1997) and Huh and Park (2002) considered semiparametric and nonparametric versions respectively with kernel regression and a single-index model. Typically these methods produce a nonparametric estimate of the mean function in the form of a smooth curve. Visual inspection of such a nonparametric curve can suggest the existence of trends in the true mean function, but it does not provide statistical inference to determine the statistical significance of such trends.

SiZer (SIgnificant ZERo crossing of the derivatives), originally proposed by Chaudhuri and Marron (1999), is a powerful exploratory data analysis tool equipped with statistical inference based on nonparametric kernel estimates. It provides a new way to look at data in scale-space so that analysts are able to discover any meaningful structure by testing the data against underlying assumptions or potential models while doing exploratory analysis. By doing so, SiZer addresses the critical question in scientific research of which features observed are really there, or represent an important underlying structure.

Several other versions of SiZer have been developed since the seminal work of Chaudhuri and Marron (1999), Hannig and Marron (2006) improved statistical inference of the original SiZer using extreme value theory. Hannig and Lee (2006) proposed a robust version of SiZer which can identify outliers, and Kim and Marron (2006) developed a tool for detecting

[*] Corresponding author.
E-mail address: jhuh@duksung.ac.kr (J. Huh).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.06.023

discontinuities in the data. Park and Kang (2008) proposed a SiZer that compares multiple curves with independent data based on their differences of smooths. Park et al. (2010) studied a SiZer which targets the quantile composition of the data instead of the mean structure. Park et al. (2004) extended the conventional SiZer to dependent errors, which conducts a goodness-of-fit test by comparing the observed data with an assumed time series model. Rondonotti et al. (2007) developed SiZer for time series that estimates an autocovariance function in order to detect significant features in a time series. Later, Park et al. (2009a) improved this SiZer tool with new quantile and autocovariance function estimator. Park et al. (2009b) introduced a SiZer that puts forth a method for comparing two or more time series. In addition, various Bayesian versions of SiZer have also been proposed as an approach to Bayesian multiscale smoothing (Erästö and Holmström, 2005; Godtliebsen and Oigard, 2005; Oigard et al., 2006; Erästö and Holmström, 2007; Sørbye et al., 2009). Note that all of these tools are restricted to data with a continuous response variable, and thus they are not readily applicable to discrete data. In scale-space, Li and Marron (2005) proposed the local likelihood SiZer map that is more efficient in distinguishing features than the original SiZer for discrete data. Ganguli and Wand (2007) considered the problem of determining the significance of features such as peaks or valleys in observed covariate effects under an additive model. They worked with low rank radial spline smoothers to allow for handling of sparse designs and large sample sizes.

In this paper, we develop a SiZer tool using local likelihood, which utilizes a local polynomial estimator with multiple bandwidths to determine the significance of features for discrete data. Because it considers a wide range of bandwidths, it circumvents the classical problem of bandwidth selection, which allows one to do statistical inference and detect all the information that is available at each individual level of resolution. Also, it focuses on smoothed curves depending on bandwidths rather than a true underlying curve because a scale-space approach views that truth exists at each scale. This allows one to avoid a bias problem that occurs in estimating a true underlying function.

Our work is differentiated from that of Li and Marron (2005) in three aspects. First, we improve global inference by proposing a theoretically justified quantile in a confidence interval using advanced distribution theory. This approach was proposed by Hannig and Marron (2006), but we extend their work to discrete data. Second, we provide the asymptotic properties of the proposed SiZer tool, which was not studied in Li and Marron (2005). Therefore, the proposed SiZer can provide more accurate and informative analysis with statistical inference and visualization for a vast range of statistical problems. Third, we discuss how the proposed SiZer can be extended to multiple-covariate cases.

The rest of the paper is organized as follows. Section 2 reviews a local likelihood approach in generalized linear models and proposes a SiZer tool using local likelihood. Section 3 investigates the performance of the local likelihood SiZer using both simulated and real examples. We study the asymptotic properties of the proposed SiZer in Section 4. We also briefly discuss how a local likelihood SiZer for multiple covariates can be constructed in Section 5. Finally, details on the new quantile estimator in the SiZer based on advanced distribution theory are provided in the Appendix.

2. Local likelihood SiZer

SiZer is based on scale-space ideas from computer vision, see Lindeberg (1994), where it refers to a family of smooths of a digital image. A scale-space approach regards no particular level of smoothing as correct and considers that each smooth provides information about the underlying image structure at a particular scale. In SiZer, scale-space is a family of kernel smooths indexed by the bandwidth. The idea is that this approach uses all the information that is available in the data at each given bandwidth. SiZer extends the usefulness of a family of smooths plot by adding a SiZer map, which displays results of statistical inference. A SiZer map visually displays the significance of features over both location and scale (i.e., bandwidth). Multiple comparison tests based on confidence intervals for the derivatives of the underlying curve are involved in flagging significant features. Therefore, SiZer is a more advanced version of a basic statistical graphic, such as a plot or chart, that simultaneously looks at data with different scopes with statistical inference.

In what follows we propose the local likelihood SiZer for the one covariate cases. We illustrate how to extend the proposed tool to the multiple-covariate cases in Section 5.

Suppose we observe a random sample (X_i, Y_i) of (X, Y) where Y_i's are real valued responses associated with covariates X_i's having density f with support $[0, 1]$ without loss of generality for $i = 1, 2, \ldots, n$. Assume that the conditional distribution of $Y|X = x$ belongs to the following one-parameter exponential family:

$$f_{Y|X}(y|x) = \exp\{y\tau(x) - b(\tau(x)) + c(y)\} \tag{1}$$

where b and c are some known functions. It is of interest to estimate the regression function $m(x) \equiv E(Y|X = x) = b'(\tau(x))$. In parametric generalized linear models, the function $m(x)$ is modeled linearly via a link function g by

$$\eta(x) \equiv g(m(x)) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p,$$

where p is the degree of a polynomial function. If $g = (b')^{-1}$, then g is called the canonical link (McCullagh and Nelder, 1989). Then, the conditional density $f_{Y|X}(y|x)$ in (1) can be written in terms of $\eta(x)$ as

$$f_{Y|X}(y|x) = \exp\{y(g \circ b')^{-1}(\eta(x)) - b((g \circ b')^{-1}(\eta(x))) + c(y)\} \tag{2}$$

where $\circ$ denotes the composition of functions.