Journal of Econometrics 184 (2015) 233–241

Contents lists available at ScienceDirect

Journal of Econometrics

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

Gradient-based smoothing parameter selection for nonparametric regression estimation☆

Daniel J. Henderson[a], Qi Li[b,c], Christopher F. Parmeter[d], Shuang Yao[e,*]

[a] Department of Economics, Finance and Legal Studies, University of Alabama, United States
[b] Department of Economics, Texas A&M University, United States
[c] ISEM, Capital University of Economics & Business, PR China
[d] Department of Economics, University of Miami, United States
[e] Economics and Management School, Wuhan University, PR China

ARTICLE INFO

Article history:
Received 20 June 2012
Received in revised form
2 May 2014
Accepted 11 September 2014
Available online 28 September 2014

Keywords:
Gradient estimation
Kernel smoothing
Least squares cross validation

ABSTRACT

Estimating gradients is of crucial importance across a broad range of applied economic domains. Here we consider data-driven bandwidth selection based on the gradient of an unknown regression function. This is a difficult problem given that direct observation of the value of the gradient is typically not observed. The procedure developed here delivers bandwidths which behave asymptotically as though they were selected knowing the true gradient. Simulated examples showcase the finite sample attraction of this new mechanism and confirm the theoretical predictions.

1. Overview

The success of nonparametric estimation hinges critically on the level of smoothing exerted on the unknown surface. Given this importance, a large literature has developed focusing on appropriate selection of the smoothing parameter(s) of the conditional mean. However, methods developed for recovering optimal smoothness levels for the conditional mean are not necessarily the proper surrogates when interest instead hinges on the *derivative* of the unknown function. Economic applications which require gradient estimation include estimates of heterogeneous individual attitudes toward risk (Chiapporis et al., 2009) and marginal willingness to pay within a two-stage hedonic regression (Bajari and Kahn, 2005; Heckman et al., 2010) to name a few.

The importance of appropriate smoothness selection for derivatives was illustrated by Wahba and Wang (1990) who showed in the smoothing spline setting that the ideal smoothing parameter depends on the derivative of the unknown function. A small strand of literature has developed focusing attention on smoothing parameter selection when interest hinges on the derivative. Within this literature there exist several different approaches for construction of the optimal bandwidth. To develop the intuition for existing approaches consider a univariate nonparametric regression model

$$y_j = g(x_j) + u_j \quad j = 1, \ldots, n. \tag{1}$$

Rice (1986) introduced a method for selecting a smoothing parameter optimal for construction of the derivative of $g(x)$. Rice's (1986) focus was univariate in nature. He suggested the use of a differencing operator (though this operator is not formally defined) and a criterion which was shown to be a nearly unbiased estimator of the mean integrated squared error (MISE) between the estimated derivative and the oracle. Building on the insight of Rice (1986), Müller et al. (1987) used Rice's noise-corrupted suggestion to select the bandwidth based on the natural extension of least-squares cross-validation (LSCV). Müller et al. (1987) also formally

☆ The authors are grateful for comments from three referees and an associate editor that greatly improved our paper. They are also thankful to the comments made in seminars at Clemson University, Columbia University, the London School of Economics, Michigan State University, North Carolina State University, the State University of New York at Binghamton, the University of Alabama, the University of California-Riverside, the University of Florida, the University of Miami, the University of Southern California and West Virginia University as well as by participants at the Latin American Meetings of the Econometrics Society in Lima, the Infometrics Workshop in Riverside, the International Symposium on Econometric Theory, Applications in Taipei and New York Camp Econometrics VII and Metro-Atlantic Study Group.

* Corresponding author.
E-mail addresses: djhender@cba.ua.edu (D.J. Henderson), qi@econmail.tamu.edu (Q. Li), cparmeter@bus.miami.edu (C.F. Parmeter), syaowhu@163.com (S. Yao).

http://dx.doi.org/10.1016/j.jeconom.2014.09.007
0304-4076/

proposed a differencing operator for calculating noise-corrupted observations of the gradients. Noting that the differencing operator deployed by Müller et al. (1987) possessed a high variance, Charnigo et al. (2011) suggested a differencing operator with more desirable variance properties as well as a generalized criterion to be used for selecting the optimal smoothing parameter.

As an alternative to noise-corrupted observations of the desired gradients, Müller et al. (1987) proposed a simpler approach by adjusting a bandwidth selected for $g(x)$ to account for the fact that the bandwidth for the gradient estimate needs to converge slower. The interesting aspect of the factor method is that, in the univariate setting, the ratio between the asymptotically optimal bandwidth for estimation of $g(x)$ and its derivative depends on the kernel. Using this fact, Müller et al. (1987) recovered an optimal bandwidth for the derivative eschewing difference quotients. Fan and Gijbels (1995) used this insight to first construct a plug-in estimator for the conditional mean and then adjust this bandwidth to have an optimal bandwidth for the derivative of the conditional mean.

Beyond the factor method, Fan and Gijbels (1995) also proposed a two-step bandwidth selector which consists of constructing empirical measures of the bias and conditional variance of the local-polynomial estimator. The unknown terms within the bias and variance are replaced with estimates found using the factor-method bandwidth. Once these measures are constructed, the final bandwidth, termed the refined bandwidth, is found by minimizing MISE. Fan et al. (1996) showed that this bandwidth selection mechanism has desirable properties both theoretically as well as in simulated settings.

In a separate approach, Ruppert (1997) developed empirical-bias bandwidth selection. A key difference from Ruppert's (1997) approach is that instead of fitting a local-polynomial to obtain estimates for the unknown components in the bias expansion for the gradient, he instead estimates the gradient for several different bandwidths and then uses least-squares to fit a Taylor expansion to the estimated unknown components of the bias. A benefit of this approach over the aforementioned methods is that it requires estimation of fewer components in practice.

Each of the existing methods leaves something to be desired in a multivariate setting. The factor method requires bandwidth selection on the conditional mean followed by calculation of a scaling factor dependent upon the kernel function (in the univariate setting) which can be tedious. The calculation of noise-corrupted derivatives also requires computing the number of neighboring observations to construct the estimates prior to minimizing the criterion function. In high dimensional settings this may not be feasible. Lastly, plug-in approaches, while having desirable theoretical properties, require the calculation of numerous unknown quantities, neutering the ability of having a completely automatic procedure. All plug-in approaches require estimation of unknown functions and their derivatives prior to the formal selection of the bandwidth. Moreover, the plug-in formula for the optimal bandwidths can become quite complicated in high dimensional settings. The framework laid out here does not require adjustment, calculation of noise-corrupted derivatives or unknown quantities related to the underlying data generating process. The method also does not hinge on a pilot bandwidth nor a set of estimates being supplied to the criterion function, streamlining the process.

Our approach begins with the oracle LSCV setup for the gradient as in Müller et al. (1987), with a local-linear estimator. We then show that replacing the oracle gradient with a local-cubic estimator produces bandwidths which behave asymptotically as though the oracle was used. The intuition for this result is that the bias of the local-cubic estimator is of sufficiently smaller order relative to the local-linear estimator that the only aspect of the local-cubic estimator which appears in our asymptotic expansion of the LSCV criterion is the variance of the difference between these estimators (local-linear and local-cubic). In the limit, the variance of this difference behaves (up to a constant depending on the kernel) exactly as the case with the oracle gradient. Thus, bandwidths selected replacing the oracle gradient with the local-cubic estimator are asymptotically equivalent to those selected with the unknown oracle gradient.

The gradient-based cross-validation (GBCV) approach studied here has several appealing features. First, the computational burden is dramatically decreased given that pilot bandwidths and first differences are not necessary to make the procedure operational. Further, the approach readily scales to the multivariate setting and is firmly entrenched within the data-driven bandwidth selection arena. Lastly, the method is intuitively appealing as it represents an easily explained procedure which mimics the traditional LSCV approach to bandwidth selection, albeit for gradients.

The remainder of the paper is as follows. Section 2 provides the formal details of our new cross-validation procedure and the asymptotic justification for our proposed method. Section 3 contains a set of simulations to show the performance of our bandwidth selection method for estimation of derivative functions compared with the oracle selection method. Concluding remarks appear in Section 4.

2. The gradient-based cross-validation method and its asymptotic behavior

We consider the problem of using a data-driven method to select the smoothing parameters for estimation of the derivative of a function. Here we describe our gradient-based cross-validation method first in the univariate setting and then for the general multivariate case.

2.1. The univariate case

To motivate the idea and keep the notational burden to a minimum, in this section we focus on the univariate nonparametric regression model in (1):

$$y_j = g(x_j) + u_j, \quad j = 1, \ldots, n, \tag{2}$$

where the functional form of $g(\cdot)$ is not specified and the error term u_j satisfies $E(u_j|x_j) = 0$. Let $\beta(x) = dg(x)/dx$ denote the first order derivative function of $g(\cdot)$ with respect to x. Let $\hat{\beta}_{LL}(x)$ be the local-linear estimator of $\beta(x)$. Ideally, we would like to choose the smoothing parameter h to minimize the estimation mean squared error $E\{[\hat{\beta}_{LL}(x) - \beta(x)]^2\}$, or the sample analog of it:

$$CV(h) \stackrel{def}{=} \frac{1}{n}\sum_{j=1}^{n}[\hat{\beta}_{LL}(x_j) - \beta(x_j)]^2 M(x_j), \tag{3}$$

where $M(\cdot)$ is a weight function with bounded support that trims out data near the boundary of the support of x.

Following the same arguments as in Racine and Li (2004) and Hall et al. (2007), one can show that

$$CV(h) = \int E[\hat{\beta}_{LL}(x) - \beta(x)]^2 M(x)f(x)dx + (s.o.),$$

where $f(x)$ denotes the density function of x and $(s.o.)$ captures terms having probability orders smaller than the leading term $\int E\left[\hat{\beta}_{LL}(x) - \beta(x)\right]^2 M(x)f(x)dx$. Let $Bias^0\left(\hat{\beta}_{LL}(x)\right)$ and $Var^0\left(\hat{\beta}_{LL}(x)\right)$ denote the leading bias and leading variance terms of $\hat{\beta}_{LL}(x)$. Then the leading term of $CV(h)$ is given by