Journal of Econometrics 185 (2015) 510–525

Contents lists available at ScienceDirect

Journal of Econometrics

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

# Infinite order cross-validated local polynomial regression

Peter G. Hall [a], Jeffrey S. Racine [b,*]

[a] Department of Mathematics and Statistics, University of Melbourne, Melbourne, Australia

[b] Department of Economics, Graduate Program in Statistics, McMaster University, Hamilton, Canada

ARTICLE INFO

*Article history:*
Received 11 June 2013
Received in revised form
16 December 2013
Accepted 14 June 2014
Available online 23 June 2014

*Keywords:*
Model selection
Efficiency
Rates of convergence

ABSTRACT

Many practical problems require nonparametric estimates of regression functions, and local polynomial regression has emerged as a leading approach. In applied settings practitioners often adopt either the local constant or local linear variants, or choose the order of the local polynomial to be slightly greater than the order of the maximum derivative estimate required. But such ad hoc determination of the polynomial order may not be optimal in general, while the joint determination of the polynomial order and bandwidth presents some interesting theoretical and practical challenges. In this paper we propose a data-driven approach towards the joint determination of the polynomial order and bandwidth, provide theoretical underpinnings, and demonstrate that improvements in both finite-sample efficiency and rates of convergence can thereby be obtained. In the case where the true data generating process (DGP) is in fact a polynomial whose order does not depend on the sample size, our method is capable of attaining the $\sqrt{n}$ rate often associated with correctly specified parametric models, while the estimator is shown to be uniformly consistent for a much larger class of DGPs. Theoretical underpinnings are provided and finite-sample properties are examined.



## 1. Introduction

Nonparametric regression plays a key role in applied statistical analysis. Locally weighted polynomial regression (Fan (1992); Ruppert and Wand (1994)) has proven extremely popular and is the most studied and widely used nonparametric regression method. The seminal work of Nadaraya (1965) and Watson (1964) examined the 'local constant' variant (which is a limiting case of the local polynomial estimator with polynomial order $p = 0$), while the local linear variant ($p = 1$) is dominant in applied settings as it possesses one of the best boundary correction methods available while it is also minimax efficient. Practitioners sometimes consider polynomials of order $p > 1$, but typically this is only done when higher order derivative estimates are required.[1] There exists work on nonparametric regression estimators of very high order (see e.g. Golubev et al. (1996); Lepski and Levit (1998)), but it involves using methods that are designed specifically for the very high order case, and, unlike local polynomial techniques, are unattractive in lower order settings. There are also ingenious, minimax optimal approaches to choosing smoothing parameters (see e.g. Lepski et al. (1997)), potentially useful in high order settings. However, they too are not attractive in practice, and for this reason are not used to analyse real data. In reality a practitioner does not know whether a low or high order method is going to be required, and so finds it attractive to use a relatively conventional, tried-and-tested construction that is sufficiently flexible to address both low and high order cases. The techniques suggested in this paper are of that type; they employ local polynomial methods to construct the estimator, and cross-validation to choose the bandwidth.

From this perspective, the order of the local polynomial used in many applications appears to be somewhat ad hoc. However, the order of the polynomial can have a noticeable impact on the quality of the resulting approximation, while the appropriate order will in general depend on the underlying DGP, as will be seen. But how to best tailor the order of the polynomial to the data at hand remains an open question. In this paper we propose using delete-one cross-validation for jointly determining the bandwidth $h$ and polynomial order $p$. The rest of this paper proceeds as follows: Section 2 presents the proposed approach, Section 3 provides theoretical underpinnings, Section 4 considers a series of Monte

* Corresponding author.
*E-mail address:* racinej@mcmaster.ca (J.S. Racine).

[1] Fan and Gijbels (1996, page 59) write "Another issue in local polynomial fitting is the choice of the order of the local polynomial. Since the modelling bias is primarily controlled by the bandwidth, this issue is less crucial however. […] Since the bandwidth is used to control the modelling complexity, we recommend the use of the lowest odd order, i.e. $p = \nu + 1$, or occasionally $p = \nu + 3$". ($\nu$ is the order of the derivative required). See also Fan and Gijbels (1995).

http://dx.doi.org/10.1016/j.jeconom.2014.06.003
0304-4076/

Carlo simulations designed to assess the finite-sample behaviour of the proposed approach, while Section 5 presents some concluding remarks.

## 2. Methodology

### 2.1. Model

Data pairs $(X_i, Y_i)$ are assumed to be generated by the model

$$Y_i = g(X_i) + \epsilon_i, \tag{2.1}$$

where $X_1, \ldots, X_n$ are independent and identically distributed as $X$, with density $f_X$ supported on a compact interval $\mathcal{I}$, and the experimental errors $\epsilon_i$ are independent and identically distributed with zero mean, independent too of the $X_i$s. The case where $\epsilon_i = \sigma(X_i)\,\epsilon_i'$, for a bounded function $\sigma$ and independent variables $\epsilon_i'$ with zero mean, independent of the $X_i$s, can be treated similarly.

### 2.2. Methodology for function estimation

To estimate $g$, let $c = (c_0, \ldots, c_p)^{\mathrm{T}}$ be a $(p+1)$-vector, let $q(x \mid c) = c_0 + c_1 x + \cdots + c_p x^p$ be a polynomial of degree $p$, and consider the problem of minimising the sum of squares

$$S(c) = \frac{1}{nh} \sum_{i=1}^{n} \Big\{ Y_i - q\Big( \frac{x - X_i}{h} \;\Big|\; c \Big) \Big\}^2 K\Big( \frac{x - X_i}{h} \Big),$$

where $K$ is a kernel function and $h$ is a bandwidth. Now,

$$\begin{aligned} -\frac{1}{2} \frac{\partial}{\partial c_j} S(c) &= \frac{1}{nh} \sum_{i=1}^{n} \Big\{ Y_i - q\Big( \frac{x - X_i}{h} \;\Big|\; c \Big) \Big\} \\ &\qquad \times \Big( \frac{x - X_i}{h} \Big)^j K\Big( \frac{x - X_i}{h} \Big) \\ &= \big\{ V(x) - \widehat{M}(x)\, c \big\}_j, \end{aligned} \tag{2.2}$$

where $V = (V_0, \ldots, V_p)^{\mathrm{T}}$ is a $(p+1)$-vector, $\widehat{M} = (\widehat{m}_{jk})$ is a $(p+1) \times (p+1)$ matrix,

$$V_j(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i \Big( \frac{x - X_i}{h} \Big)^j K\Big( \frac{x - X_i}{h} \Big), \tag{2.3}$$

$$\widehat{m}_{jk}(x) = \frac{1}{nh} \sum_{i=1}^{n} \Big( \frac{x - X_i}{h} \Big)^{j+k} K\Big( \frac{x - X_i}{h} \Big). \tag{2.4}$$

Equating to zero the derivative at (2.2), and solving for $c$, we obtain:

$$\hat{c}(x) = (\hat{c}_0(x), \ldots, \hat{c}_p(x))^{\mathrm{T}} = \widehat{M}(x)^{-1} V(x). \tag{2.5}$$

Our estimator of $g$ is

$$\hat{g}(x) = \hat{c}_0(x). \tag{2.6}$$

### 2.3. Cross-validation

The cross-validation "estimator" of integrated squared error weighted by the density $f_X$,

$$\mathrm{ISE}(h, p) = \int_{\mathcal{I}} (\hat{g} - g)^2 f_X,$$

is given by

$$\mathrm{CV}(h, p) = \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - \hat{g}_{-i}(X_i) \}^2, \tag{2.7}$$

where $\hat{g}_{-i}$ denotes the version of $\hat{g}$, defined as at (2.6), when the data pair $(X_i, Y_i)$ is removed from the sample. In fact,

$$\begin{aligned} \mathrm{CV}(h, p) &= \frac{1}{n} \sum_{i=1}^{n} \{ g(X_i) - \hat{g}_{-i}(X_i) \}^2 + \frac{2}{n} \sum_{i=1}^{n} \{ g(X_i) - \hat{g}_{-i}(X_i) \}\, \epsilon_i \\ &\qquad + \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2, \end{aligned} \tag{2.8}$$

where the first term on the right-hand side of (2.8) is a good approximation to $\mathrm{ISE}(h, p)$, the second term is generally negligibly small, and the third term does not depend on $h$ or $p$ and converges to $\tau^2 \equiv E\{\sigma(X)^2\}$ where the function $\sigma$ is as in Section 2.1. Therefore it is reasonable to view $\mathrm{CV}(h, p)$ as an approximation to $\mathrm{ISE}(h, p) + \tau^2$.

We proceed by minimising (2.7) jointly with respect to $h$ and $p$, and then use the resulting values for constructing the estimator of $g$ given in (2.6) (particulars of this mixed integer optimisation are described in Section 4).

## 3. Theoretical properties

### 3.1. Overview

Since we are treating high order local polynomial methods, where the degree of the polynomial diverges with sample size, then in technical arguments we must compute inverses of high order matrices of covariance type. These are Hankel matrices, and so in Section 3.2 we introduce properties of such quantities, governed by the particular kernels we shall use. The properties of smallest eigenvalues discussed in Section 3.2 will prove invaluable when assessing expressions involving inverses of Hankel matrices, and, as discussed in Section 3.4, they motivate our regularity conditions. The components of our Hankel matrices depend to a large extent on moments of distributions whose respective densities are kernel functions, and so in Section 3.3 we develop basic properties of those moments. (We discuss the properties there, rather than later in the paper, since again they are needed to motivate our regularity conditions, given in Section 3.4.) In Sections 3.4 and 3.5, respectively, we describe theoretical properties of function estimators and cross-validation in high order settings, and in Section 3.6 we discuss these properties together, describing their ramifications.

### 3.2. Hankel matrices

Let $M_p = M_p(K)$ denote the $(p+1) \times (p+1)$ matrix with $(j, k)$th element $\int u^{j+k} K(u)\, du$, for $0 \le j, k \le p$. Such matrices are distinctly patterned (in particular, the components down any anti-diagonal are identical), and are in the class of Hankel matrices. There is a literature on properties of the smallest eigenvalues of Hankel matrices, and we summarise some of it below.

In what follows, if $a_n$ and $b_n$ are sequences of positive numbers, we write $a_n \sim b_n$ to mean that the ratio $c_n = a_n / b_n$ converges to one as $n \to \infty$, and we write $a_n \asymp b_n$ to mean that $c_n$ is bounded away from zero and infinity as $n \to \infty$.

Let $\mathrm{ev}_p = \mathrm{ev}_p(K)$ denote the smallest eigenvalue of $M_p$. If $K$ is bounded, symmetric and has support equal to $[-1, 1]$, and if $K(u) \ge C\,(1 - u^2)^s$ for constants $C > 0$ and $s \ge 0$, then a result of Widom and Wilf (1966) implies that

$$\mathrm{ev}_p(K) \sim B\, p^{1/2}\, \exp\big\{ -\big(1 + 2^{1/2}\big)\, p \big\} \tag{3.9}$$

where $B > 0$. If $K$ is the standard normal density,

$$\mathrm{ev}_p(K) \sim B\, p^{1/4}\, \exp\big( -2^{3/2}\, p^{1/2} \big) \tag{3.10}$$

(Szegö (1936)), where $B = 2^{13/4}\, \pi^{3/2}\, e$.