DEPARTMENT OF ECONOMICS

WORKING PAPER SERIES

2013-05

McMASTER UNIVERSITY

Department of Economics
Kenneth Taylor Hall 426
1280 Main Street West
Hamilton, Ontario, Canada
L8S 4M4

<http://www.mcmaster.ca/economics/>

INFINITE ORDER CROSS-VALIDATED LOCAL POLYNOMIAL REGRESSION

PETER G. HALL AND JEFFREY S. RACINE

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, finite-sample properties are examined, and an application highlights finite-sample improvements arising from the use of the proposed method.

1. INTRODUCTION

Nonparametric regression plays a key role in applied statistical analysis. Locally weighted polynomial regression (Fan (1992), Ruppert & 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

Date: January 17, 2013.

Key words and phrases. Model Selection, Efficiency, Rates of Convergence.

Racine would like to gratefully acknowledge support from the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).

derivative estimates are required.¹ 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 presents an illustrative application, section 4 considers a series of Monte Carlo simulations designed to assess the finite-sample behaviour of the proposed approach, while section 5 provides theoretical underpinnings. Detailed proofs can be found in the appendices.

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 $X_1, \dots, X_n$ are independent and identically distributed as X , with density f_X supported on a compact interval $\mathcal{I}$, and the experimental errors ϵ_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 σ and independent variables ϵ'_i with zero mean, and independent too of the X_i s, can be treated similarly.

¹Fan & 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 $p = \nu + 3$.” (ν is the order of the derivative required).

2.2. Methodology for function estimation. To estimate g , let $c = (c_0, \dots, c_p)^\top$ be a $(p+1)$ -vector, let $q(x|c) = c_0 + c_1 x + \dots + c_p x^p$ be a polynomial of degree p , and consider the problem of minimising the sum of squares

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

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

$$\begin{aligned} -\frac{1}{2} \frac{\partial}{\partial c_j} S(c) &= \sum_{i=1}^n \left\{ Y_i - q\left(\frac{x - X_i}{h} \mid c\right) \right\} \left(\frac{x - X_i}{h}\right)^j K\left(\frac{x - X_i}{h}\right) \\ &= \{V(x) - \widehat{M}(x)c\}_j, \end{aligned} \quad (2.2)$$

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

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

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

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

$$\hat{c}(x) = (\hat{c}_0(x), \dots, \hat{c}_p(x))^\top = \widehat{M}(x)^{-1} V(x). \quad (2.5)$$

Our estimator of g is

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

2.3. Cross-validation. The cross-validation “estimator” of integrated squared error weighted by the density f_X ,

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

is given by

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

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,

$$\text{CV}(h, p) = \frac{1}{n} \sum_{i=1}^n \{g(X_i) - \hat{g}_{-i}(X_i)\}^2 + \frac{2}{n} \sum_{i=1}^n \{g(X_i) - \hat{g}_{-i}(X_i)\} \epsilon_i + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2, \quad (2.8)$$

where the first term on the right-hand side of (2.8) is a good approximation to $\text{ISE}(h, p)$, 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 σ is as in section 2.1. Therefore it is reasonable to view $\text{CV}(h, p)$ as an approximation to $\text{ISE}(h, p) + \tau^2$.

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). Before providing the theoretical foundation for this approach, we first consider an illustrative application and assess finite-sample behaviour.

3. APPLICATION: MINCER EQUATIONS REVISITED

In a seminal paper Mincer (1974) proposed using a specific (parametric) function of years of potential experience to capture concavity of the age-earnings profile, a specification that endures in applied work to this day. The methods developed in this paper provide us with an opportunity to check, in an adaptive fashion without making the stringent assumptions that would be necessary for a formal hypothesis test, whether the concavity assumption, implicit in the theory of human capital, is satisfactory.

The so-called ‘Mincer Equation’ describes the relationship between (log) wages and potential experience, and has been widely studied. In fact, it presents a case where there is little disagreement among practitioners regarding which parametric functional form is most appropriate. The original formulation involves least squares parametric models that are quadratic with respect to potential experience. In a subsequent paper (the advice in which is often surprisingly ignored) Murphy & Welch (1990) demonstrated that the standard quadratic specification proposed by Mincer (1974) is not flexible enough to capture important features

of the data and suggested instead that practitioners adopt a quintic specification, although Mincer’s (1974) original specification remains popular in applied work. Nonparametric estimates of Mincer relationships have more recently found acceptance among practitioners, the local constant and local linear kernel variants being perhaps most popular. We revisit the Mincer relationship from a nonparametric perspective to see whether the proposed approach can inform this debate in a meaningful way, at least from the practitioner’s perspective.

We take up this issue using a popular data set considered in Ullah (1985) containing $n = 205$ observations consisting of a random sample of Canadian cross-section wage data taken from the 1971 Canadian Census Public Use Tapes for male individuals having common education (Grade 13). Our main focus here is how dominant ad hoc local polynomial variants (i.e. the local constant ($p = 0$) and local linear ($p = 1$) ones) perform relative to the proposed method. For comparison purposes we also estimate least squares polynomial fits of order $j = 2, \dots, 5$ which span Mincer’s (1974) original quadratic specification through Murphy & Welch’s (1990) quintic specification.

We first fit all models on the full sample and then, for what follows, use the functional form (or polynomial degree and/or bandwidths) from the full sample fits for the simulation exercise. The full-sample cross-validated bandwidths were 1.8961, 3.272, and 20.7612 for the local constant, local linear, and proposed method, while the cross-validated degree for the proposed method was $p = 4$. R^2 values (unadjusted for comparison purposes) from all models were 0.2308, 0.2496, 0.315, 0.3259, 0.3107, 0.3175, and 0.3195 for the parametric $j = 2$ through $j = 5$, local constant, local linear and proposed method, respectively. Values for the cross-validation objective function given in (2.7) were 0.3161, 0.3027, 0.2943 for the local constant, local linear and proposed method, respectively. Figure 1 presents the full sample fits from the range of models fitted for comparison purposes.

Having considered in-sample model summaries, we next split the sample into $n_1 = 200$ training observations and $n_2 = 5$ independent evaluation observations (i.e. first shuffle the data then split into two samples – note that results were qualitatively unchanged for splits

FIGURE 1. Model fits. The labels $\text{lm}.j$, $j=2, \dots, 5$ represent parametric models ranging from quadratic, $j=2$, through quintic, $j=5$. The labels np.lc , np.ll , and np.glp represent the local constant, $p = 0$, local linear, $p = 1$, and proposed estimator.

of $n_1 = 190/n_2 = 15$ and $n_1 = 195/n_2 = 10$). We then computed the predicted square error (PSE) for the independent evaluation observations as $n_2^{-1} \sum_{i=1}^{n_2} (Y_i - \hat{Y}_i)^2$ where $\hat{Y}_i$ denotes the predicted values computed from the X_i for the n_2 evaluation observations. We repeated this exercise $M = 10^5$ times and then compared the performance of the various models according to their average PSE over the M splits of the data. Figure 2 presents boxplots of the M PSEs for each model considered. This exercise was conducted to assess whether apparent improvements in in-sample performance (i.e the full sample, $n = 205$) reflect ‘overfitting’ or

FIGURE 2. Boxplots of the PSE for each model over the $M = 10^5$ sample splits. The labels $\text{lm}.j$, $j=2, \dots, 5$ represent parametric models ranging from quadratic, $j=2$, through quintic, $j=5$. The labels np.lc , np.ll , and np.glp represent the local constant ($p = 0$), local linear ($p = 1$), and proposed estimator.

not, and one would expect that the model with the lowest average PSE would be expected to be closest to the true DGP.

The mean PSEs over the M independent evaluation samples for each model were 0.5667, 0.5645, 0.5420, 0.5431, 0.5615, 0.5493, and 0.5414 for the parametric $j = 2$ through $j = 5$, local constant, local linear and proposed method, respectively, which delivers relative efficiencies (ratio of mean PSEs) of 1.0467, 1.0427, 1.0011, 1.0031, 1.0371, and 1.0147 for the parametric models $j = 2$ through $j = 5$, local constant and local linear methods, respectively

(numbers greater than one indicate poorer expected out-of-sample performance relative to the proposed approach).

Based on both in-sample and out-of-sample performance, the proposed method outperforms the ad hoc local constant and local linear variants. Moreover, a purely concave function is not a good fit to the data, and our very flexible, adaptive approach, as well as local constant and local polynomial methods, suggest that the true shape of the function is slightly bimodal. Interestingly, the Schwarz-Bayes Information Criterion ('SBIC', Schwarz (1978)) selects the quadratic parametric model ($j = 2$) which possesses the worst in-sample fit and also performs the poorest in terms of out-of-sample performance of all models considered. Here the proposed approach also outperforms the dominant quadratic parametric specification chosen by SBIC and is certainly competitive with specifications suggested by Murphy & Welch (1990). From a practical perspective, the proposed method may help resolve issues of functional form that continue to plague practitioners and suggests a quasi-global functional form in this case (local polynomial of degree $p = 4$, bandwidth $h = 20.7612$) that is consistent with subsequent refinements of Mincer's (1974) original specification.

4. FINITE-SAMPLE PERFORMANCE

In this section we summarize simulations designed to assess the finite-sample performance of the proposed approach. All computation was conducted in R Version 2.15.2 (R Core Team (2012)). A second order standard normal kernel was used throughout. Before proceeding, a few words on the numerical optimisation of $CV(p, h)$ defined in (2.7) are in order as this objective function is nonsmooth and non-differentiable with respect to p , hence minimisation of $CV(p, h)$ involves constrained mixed integer optimisation which is known to be computationally challenging ($p \geq 0$ is an integer, $h \geq 0$ is real-valued). For what follows we adopt the 'Nonsmooth Optimisation by Mesh Adaptive Direct Search' (NOMAD) approach proposed by Abramson, Audet, Couture, Dennis Jr. & Le Digabel (2011) which is implemented in the

R package ‘crs’ (Racine & Nie (2012)). The NOMAD library is an open source C++ implementation of the ‘Mesh Adaptive Direct Search’ (MADS) algorithm designed for constrained optimisation of blackbox functions, and is without peer for the optimisation problem we wish to solve.

We consider four DGPs which include a simple low order polynomial, a periodic function, the ‘Bump’ function, and the Doppler function defined as follows:

$$\text{Polynomial: } g(X_i) = X_i + X_i^2,$$

$$\text{sin: } g(X_i) = \sin(2\pi X_i),$$

$$\text{Bump: } g(X_i) = e^{-32(X_i-0.5)^2} + 2X_i - 1,$$

$$\text{Doppler: } g(X_i) = \sqrt{X_i(1 - X_i)} \times \sin\left(\frac{2\pi(1 + 2^{-7/5})}{X_i + 2^{-7/5}}\right).$$

For the polynomial DGP $X_i \sim U[-2, 2]$ while for the remaining DGPs $X_i \sim U[0, 1]$. We divide the signal component $g(X_i)$ for each DGP by its standard deviation $\sigma_{g(X_i)}$ to control the signal/noise ratio, generate $Y_i = g(X_i)/\sigma_{g(X_i)} + \epsilon_i$ and let $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma = 0.25, 0.50, 1.00, 2.00$.² This standardization delivers an expected R^2 for the Oracle estimator of 0.95, 0.80, 0.50 and 0.20, respectively. Figure 3 presents a random sample drawn from each DGP along with the true DGP and the fit determined by the proposed method for $n = 1,600$ and $\sigma = 0.50$.

For the simulations that follow we consider samples of size $n = 100, 200, 400, 800, 1600, 3200$ and consider $M = 1,000$ draws from each of the DGPs outlined above. For comparison purposes we consider local polynomial estimators of fixed (ad hoc) order $p = 0, 1, \dots, 10$. Results are summarized in tables 1 through 3. Rates of convergence are then obtained from the regression of $\log(\text{RMSE})$ on $\log(n)$ and a constant, the coefficient on $\log(n)$ being the ‘realized’ rate of convergence. This can be compared with that for the local constant and local linear variants that are widely found in applied settings and possess theoretical RMSEs

²Full simulation results are available upon request and are not included here for space considerations.

FIGURE 3. One random sample drawn from the simulated DGPs along with the proposed estimator, $n = 1,600$, $\sigma = 0.50$.

of $O(n^{-0.4})$ (we report the value of α , the coefficient on $\log(n)$ which can then be compared with the benchmark -0.4). We also assess the relative efficiency of the ad hoc deterministic polynomials versus the proposed approach by reporting their relative mean RMSE values for the M Monte Carlo replications.

Table 1 summarizes the mean values of the cross-validated polynomial degree p and bandwidth h for two representative simulations ($\sigma = 0.50, 1.00$). It can be seen that for the polynomial DGP the mean value of the cross-validated p is slightly greater than that for the true polynomial while the median is equal to that for the true polynomial (not reported for space considerations). The largest values of h are selected for the polynomial DGP, the smallest for the Doppler DGP. For the non-polynomial DGPs it is evident that the value of p chosen by cross-validation increases with n , while for the polynomial DGP it appears to be stable with respect to n , as expected.

TABLE 1. Mean degree and bandwidth for proposed estimator over $M = 1,000$ Monte Carlo replications.

n	Degree	Bandwidth	n	Degree	Bandwidth
Polynomial, $\sigma = 0.50$			Polynomial, $\sigma = 1.00$		
100	2.5	728.5	100	2.4	693.8
200	2.4	779.4	200	2.3	762.6
400	2.5	773.6	400	2.4	746.7
800	2.4	758.3	800	2.4	735.7
1600	2.4	791.1	1600	2.4	797.4
3200	2.4	762.4	3200	2.4	755.7
sin, $\sigma = 0.50$			sin, $\sigma = 1.00$		
100	3.7	138.0	100	3.1	143.2
200	3.9	149.2	200	3.2	159.7
400	4.5	164.1	400	3.6	153.2
800	4.9	172.3	800	4.1	159.7
1600	4.9	174.2	1600	4.5	160.3
3200	5.0	180.3	3200	4.6	161.0
Bump, $\sigma = 0.50$			Bump, $\sigma = 1.00$		
100	3.2	66.9	100	2.1	74.7
200	3.4	66.8	200	2.5	78.3
400	4.0	66.6	400	2.7	64.1
800	4.6	82.8	800	3.2	69.7
1600	5.1	79.9	1600	3.4	60.4
3200	5.3	82.3	3200	4.1	71.4
Doppler, $\sigma = 0.50$			Doppler, $\sigma = 1.00$		
100	10.4	35.7	100	6.7	41.2
200	10.8	43.4	200	9.1	63.9
400	10.8	61.1	400	10.0	62.9
800	11.5	74.9	800	11.2	75.1
1600	12.8	61.7	1600	11.7	80.1
3200	13.7	35.4	3200	12.6	66.7

Table 2 summarizes the realized rates of convergence computed from the M Monte Carlo replications for each DGP. First, recall that the theoretical rate of convergence of the popular local constant ($p = 0$) and local linear ($p = 1$) estimators is of $O(n^{-0.40})$ which is borne out by the simulations, though the rate of convergence of the local linear estimator is slightly higher than theory predicts for all but the Doppler DGP. For the proposed estimator it is evident that for the polynomial and periodic DGPs the realized rate is equal to the parametric rate for the Oracle estimator, $O(n^{-0.5})$. Note also that the rate of convergence of the proposed

TABLE 2. Realized rate of convergence (i.e. $-\alpha$ for $O(n^{-\alpha})$) of the proposed estimator ('glp') and deterministic estimators ($p = 0, 1, \dots$) over $M = 1,000$ Monte Carlo replications.

σ	glp	$p = 0$	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$
Polynomial												
0.25	-0.50	-0.39	-0.42	-0.49	-0.49	-0.49	-0.49	-0.50	-0.50	-0.50	-0.50	-0.50
0.50	-0.51	-0.39	-0.43	-0.49	-0.50	-0.50	-0.50	-0.51	-0.51	-0.51	-0.51	-0.50
1.00	-0.50	-0.40	-0.44	-0.50	-0.50	-0.50	-0.50	-0.50	-0.51	-0.51	-0.51	-0.50
2.00	-0.51	-0.41	-0.45	-0.50	-0.50	-0.50	-0.50	-0.51	-0.51	-0.51	-0.51	-0.51
sin												
0.25	-0.49	-0.40	-0.42	-0.45	-0.47	-0.48	-0.47	-0.49	-0.50	-0.50	-0.51	-0.50
0.50	-0.50	-0.40	-0.43	-0.46	-0.46	-0.48	-0.49	-0.50	-0.50	-0.50	-0.50	-0.50
1.00	-0.49	-0.40	-0.43	-0.47	-0.44	-0.46	-0.50	-0.50	-0.51	-0.51	-0.51	-0.51
2.00	-0.46	-0.41	-0.45	-0.48	-0.44	-0.47	-0.49	-0.50	-0.50	-0.51	-0.51	-0.51
Bump												
0.25	-0.48	-0.42	-0.42	-0.46	-0.46	-0.47	-0.47	-0.48	-0.47	-0.47	-0.48	-0.50
0.50	-0.47	-0.42	-0.43	-0.46	-0.46	-0.48	-0.47	-0.46	-0.47	-0.48	-0.49	-0.50
1.00	-0.46	-0.42	-0.44	-0.47	-0.46	-0.46	-0.46	-0.46	-0.48	-0.50	-0.50	-0.51
2.00	-0.44	-0.42	-0.43	-0.45	-0.44	-0.44	-0.45	-0.47	-0.49	-0.50	-0.51	-0.51
Doppler												
0.25	-0.48	-0.42	-0.40	-0.42	-0.45	-0.46	-0.47	-0.44	-0.38	-0.39	-0.39	-0.40
0.50	-0.47	-0.40	-0.40	-0.42	-0.44	-0.45	-0.46	-0.47	-0.44	-0.42	-0.42	-0.43
1.00	-0.47	-0.40	-0.40	-0.43	-0.44	-0.44	-0.46	-0.46	-0.47	-0.47	-0.47	-0.46
2.00	-0.45	-0.39	-0.40	-0.43	-0.43	-0.44	-0.45	-0.45	-0.46	-0.47	-0.47	-0.47

estimator improves upon that for local constant and local linear variants. Finally, it is evident that as p increases, the rate of the ad hoc (deterministic p) local polynomial estimator approaches the parametric rate for the Oracle estimator. Regarding this last observation, why not then simply use a large (ad hoc) value of p in applied settings? The answer will be apparent when we examine relative efficiency, to which we now turn.

Table 3 summarizes the efficiencies (ratio of RMSEs) of the ad hoc deterministic local polynomial estimator relative to the proposed estimator (numbers ≥ 1 indicate higher mean RMSE of the ad hoc local polynomial estimator with p equal to the column heading). Note that there will be some (unknown) deterministic value of p for each DGP and sample size that will be optimal and were one to use this value rather than the stochastic choice one would naturally dominate estimators that do not make use of this information. But in general the

TABLE 3. Mean relative efficiency (ratio of mean RMSE) of the ad hoc deterministic local polynomial versus the proposed estimator over $M = 1,000$ Monte Carlo replications (numbers ≥ 1 indicate higher mean RMSE of the ad hoc local polynomial estimator with p equal to the column heading).

n	$p = 0$	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$	$p = 7$	$p = 8$	$p = 9$	$p = 10$
Polynomial, $\sigma = 0.50$											
100	1.35	1.11	0.88	0.97	1.07	1.15	1.25	1.35	1.44	1.50	1.53
200	1.47	1.20	0.91	1.01	1.10	1.19	1.28	1.36	1.44	1.51	1.58
400	1.56	1.24	0.91	1.00	1.09	1.17	1.26	1.34	1.42	1.49	1.56
800	1.74	1.35	0.92	1.02	1.11	1.20	1.28	1.37	1.45	1.53	1.60
1600	1.85	1.40	0.93	1.01	1.09	1.18	1.27	1.35	1.44	1.51	1.57
3200	2.04	1.50	0.94	1.03	1.11	1.20	1.28	1.37	1.45	1.52	1.58
sin, $\sigma = 0.50$											
100	1.10	1.02	0.90	0.92	0.94	0.94	1.02	1.08	1.15	1.20	1.25
200	1.17	1.06	0.92	0.96	0.95	0.92	0.99	1.06	1.13	1.18	1.23
400	1.27	1.14	0.96	1.01	0.99	0.94	1.00	1.07	1.13	1.18	1.24
800	1.36	1.19	0.99	1.04	1.01	0.95	1.02	1.08	1.14	1.20	1.25
1600	1.47	1.26	1.02	1.05	1.01	0.96	1.02	1.08	1.15	1.20	1.26
3200	1.54	1.30	1.03	1.06	1.01	0.96	1.02	1.06	1.12	1.17	1.21
Bump, $\sigma = 0.50$											
100	0.93	0.94	0.92	0.97	0.98	1.00	0.99	1.04	1.07	1.12	1.16
200	0.96	0.96	0.93	0.98	0.99	1.01	1.00	1.04	1.05	1.09	1.13
400	0.98	0.98	0.92	0.96	0.96	1.01	1.02	1.03	1.02	1.06	1.09
800	1.02	1.01	0.93	0.97	0.96	1.00	1.03	1.03	1.01	1.05	1.07
1600	1.06	1.04	0.93	0.97	0.96	1.00	1.02	1.03	1.03	1.04	1.05
3200	1.10	1.07	0.94	0.97	0.95	0.99	1.00	1.02	1.03	1.03	1.01
Doppler, $\sigma = 0.50$											
100	1.08	1.05	1.02	1.04	1.04	1.05	1.03	1.00	0.98	0.97	0.97
200	1.15	1.11	1.05	1.05	1.04	1.05	1.04	0.99	0.98	0.97	0.97
400	1.21	1.18	1.09	1.08	1.07	1.06	1.05	1.00	0.99	0.98	0.98
800	1.27	1.23	1.11	1.10	1.07	1.05	1.04	1.00	1.00	1.00	0.98
1600	1.36	1.28	1.15	1.12	1.10	1.06	1.04	1.05	1.08	1.08	1.02
3200	1.37	1.34	1.19	1.16	1.12	1.07	1.05	1.08	1.15	1.14	1.13

optimal value of p is unknown and varies substantially (see e.g. Table 1 where, depending on n , the optimal value of p ranges from 2 through > 10). It can be seen that the relative efficiencies of the local constant ($p = 0$) and local linear ($p = 1$) variants are ≥ 1 (with the exception of the Bump DGP for small samples i.e. when $n \leq 400$). Furthermore, their relative efficiencies deteriorate uniformly as n increases. Finally, it is clear that setting p

in an ad hoc manner to some large value cannot be justified on efficiency grounds. These simulations suggest that the proposed method can lead to substantial improvements in the behaviour of local polynomial kernel regression via data-driven choice of both the polynomial degree p and bandwidth h , particularly with respect to the local constant ($p = 0$) and local linear ($p = 1$) variants that are dominant in applied settings.

5. THEORETICAL PROPERTIES

5.1. 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 $0 \leq j, k \leq p$. 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 summarize some of it below.

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

$$\text{ev}_p(K) \sim B p^{1/2} \exp \left\{ - \left(1 + 2^{1/2} \right) p \right\} \quad (5.9)$$

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

$$\text{ev}_p(K) \sim B p^{1/4} \exp \left(- 2^{3/2} p^{1/2} \right) \quad (5.10)$$

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

Cases more general than the Gaussian, including those where $K(u) = B_1 \exp(-B_4 |u|^\alpha)$, with $B_1, B_4 > 0$ and $\alpha > 1$, have been discussed by Chen & Lubinsky (2004). From their results it can be seen that if α is not an odd integer,

$$\text{ev}_p(K) \sim p^{(\alpha-1)/(2\alpha)} \exp \left\{ - B_3 n^{(\alpha-1)/\alpha} + \sum_{j=1}^{[(\alpha-1)/2]} D_j p^{-(2j+1)/\alpha} \right\}, \quad (5.11)$$

where $\lfloor a \rfloor$ denotes the largest integer not exceeding a , $B_3 > 0$, $D_1, D_2, \dots$ are nonzero constants (the values depending on α , B_1 and B_4), and $D_1 > 0$; and if α is an odd integer then the same formula holds, with the constants having the same properties, provided that the last term in the series is replaced by a constant multiple of $\log p$. It is possible to develop versions of our results for cases represented by (5.11), but for the sake of brevity we do not.

5.2. Moment bounds relating to particular kernel functions. If $s \geq 0$ is an integer then

$$\begin{aligned} \int_{-1}^1 u^{2r} (1-u^2)^s du &= 2 \int_0^1 u^{2r} (1-u^2)^s du = \int_0^1 v^r (1-v)^s v^{-1/2} dv \\ &= B\left(r + \frac{1}{2}, s+1\right) = \frac{\Gamma(r + \frac{1}{2}) \Gamma(s+1)}{\Gamma(r + s + \frac{3}{2})} \sim \frac{s!}{r^{s+1}}, \end{aligned}$$

where the asymptotic result holds as $r \rightarrow \infty$ for s fixed. Therefore, if Z has density $K(u) = c_s (1-u^2)^s$ with support $[-1, 1]$, where $s \geq 1$ is an integer and c_s is chosen so that $\int K(u) du = 1$, then $E(Z^{2r}) \sim c_s s! / r^{s+1}$ as r diverges. Note too that if Z has the standard normal distribution then

$$E(Z^{2r}) = (2r-1)!! = \frac{(2r)!}{r! 2^r} \sim 2^{1/2} (2r/e)^{2r} (r/e)^{-r} 2^{-r} = 2^{1/2} (2r/e)^r.$$

Combining the results in this paragraph we deduce that, for integers $r \geq 1$,

$$\kappa_{2r} \equiv \int u^{2r} K(u) du \leq \begin{cases} C_K r^{-(s+1)} & \text{if } K(u) = K_s(u) \\ C_K (2r/e)^r & \text{if } K(u) = (2\pi)^{-1/2} \exp(-\frac{1}{2} u^2), \end{cases} \quad (5.12)$$

where the constant C_K depends only on K , and K_s is the kernel

$$K_s(x) = c_s (1-x^2)^s, \quad x \in [-1, 1], \quad (5.13)$$

with c_s chosen so that $\int K_s = 1$. Analogous bounds for $\int |u|^{2r} K(u) du$ hold when $r \geq \frac{1}{2}$.

5.3. Properties of average summed squared error and mean integrated squared error. We take K to be a symmetric probability density having a bounded derivative on its support, and more particularly we suppose that, with $\lambda_r = \sup_x [|x|^r \{K(x) + |K'(x)|\}]$, the following is true:

$$\begin{aligned} &\text{either (a) } \text{ev}_p(K) \geq B_1 p^{-B_2} \exp(-B_3 p), \lambda_p \leq B_1 \text{ and } \kappa_{2r} \leq B_4 (2r)^{-(s+1)} \\ &\text{for } r = \frac{1}{2}, 1, \frac{3}{2}, 2, \dots, \text{ for constants } B_1, B_4, B_3 > 0 \text{ depending only on } \\ &K, \text{ and for an integer } s \geq 0; \text{ or (b) } \text{ev}_p(K) \geq B_1 p^{-B_2} \exp(-B_3 p^{1/2}), \\ &\lambda_p \leq B_4 p^{p/2} \text{ and } \int |u|^{2r} K(u) du \leq B_4 (2 B_5 r)^r, \text{ for } r = \frac{1}{2}, 1, \frac{3}{2}, 2, \dots \text{ and} \\ &\text{for constants } B_1, \dots, B_5 > 0 \text{ depending only on } K. \end{aligned} \quad (5.14)$$

The inequalities $\lambda_p \leq B_1$ and $\lambda_p \leq B_4 p^{2p}$ are respectively satisfied for the kernels $K_s(x) = c_s (1 - x^2)^s$ for $x \in [-1, 1]$, and $K(x) = (2\pi)^{-1/2} \exp(-\frac{1}{2} x^2)$ for $x \in \mathbb{R}$. Properties (5.9), (5.10) and (5.12) motivate the other aspects of (5.14). More generally, (5.14)(a) holds if $K = K_s$ (examples include the uniform kernel, with $s = 0$, the Epanechnikov kernel, with $s = 1$, and the biweight kernel, with $s = 2$), and (5.14)(b) obtains if K is the standard normal kernel.

Of the function g we assume that, for constants $B_6, B_7 > 0$, and whenever $j \geq 0$,

$$\sup_{x \in \mathcal{I}} |g^{(j)}(x)| \leq B_6 B_7^j, \quad (5.15)$$

where $\mathcal{I}$ is the support of the design density f_X . Condition (5.15) clearly holds with $B_7 = |C|$ if $g(x) = \exp(Cx)$ for a constant C , and it also obtains if g is a trigonometric function; and it holds for $B_7 > 1$ in many other settings, for example if either $g(x) = \exp\{\pi(x)\}$ and π is a polynomial, or $g(x) = \exp\{\pi(|x|^\alpha)\}$ for some $\alpha > 0$, provided in the latter case that the compact interval $\mathcal{I}$ does not contain the origin. See Appendix A.

Our assumptions of the design density f_X , and the distribution of the experimental errors in the model at (2.1), are conventional:

f_X is bounded away from zero on the compact interval $\mathcal{I}$, vanishes outside that set, and has a bounded derivative on $\mathcal{I}$ if f is considered to be restricted to that set; and $\epsilon_1, \epsilon_2, \dots$ are independent and identically distributed with zero mean and uniformly bounded variances. (5.16)

The following theorem describes properties of the local polynomial estimator $\hat{g}$ for increasingly large values of the degree of the polynomial. It implies that, under conditions (5.14)–(5.16), $\int_{\mathcal{I}} (\hat{g} - g)^2$ and $\hat{g}(x) - g(x)$ are respectively of orders $n^{-(1-\delta)}$ and $n^{-(1-\delta)/2}$ in probability, where in the case of $\hat{g}(x) - g(x)$ the result holds for each $x \in \mathcal{I}$, and where $\delta = \delta(n)$ converges to zero as the degree p of the locally fitted polynomial increases. The theorem gives details of the appropriate sizes of p and the bandwidth, h .

Theorem 5.1. *Assume that (5.14)–(5.16) hold. (a) In the case of (5.14)(a), if $p \sim \frac{1}{2} \{(\log n)/B_3\}^{1/2}$ and*

$$h^{2p+3} \asymp n^{-1} p^{s+2} \exp(2p B_3) [p/\{\exp(B_3 + 1) B_7\}]^{2p} \quad (5.17)$$

then, for each $x \in \mathcal{I}$ and each uniformly bounded, nonnegative weight function w ,

$$\frac{1}{n} \sum_{i: X_i \in \mathcal{I}} \{\hat{g}(X_i) - g(X_i)\}^2 w(X_i) = O_p(d_n), \quad (5.18)$$

$$\int_{\mathcal{I}} (\hat{g} - g)^2 w = O_p(d_n), \quad \hat{g}(x) - g(x) = O_p(d_n^{1/2}), \quad (5.19)$$

where $d_n = n^{-1} \exp[2(B_3 \log n)^{1/2} \{1 + o(1)\}]$. (b) In the case of (5.14)(b), if $p \sim \{(\log n)/\log \log n\}^{1/2}$ and $h^{2p+3} \asymp n^{-1} (B_7 e)^{-2p} p^{p+1}$ then, for each $x \in \mathcal{I}$, (5.18) and (5.19) hold with $d_n = n^{-1} \exp[(\log n \cdot \log \log n)^{1/2} \{1 + o(1)\}]$.

5.4. Properties of cross-validation. For simplicity in this section we confine attention to the case where the error distribution is essentially bounded and the kernel is the s -weight kernel:

- (a) the distribution of $\epsilon_1, \epsilon_2, \dots$ satisfies $P(|\epsilon_i| \leq C) = 1$ for a constant $C > 0$; and
 - (b) the kernel is $K = K_s$, defined at (5.13).
- (5.20)

Let $\mathcal{S}_0$, depending on $\eta_3, \eta_4 \in (0, 1)$, $\eta_5 \in (0, \frac{2}{3})$ and a positive sequence ϵ_n converging to 0, denote the set of pairs (h, p) such that $p \leq \eta_3^{-1} (\log n)^{1-\eta_4}$ and

$$\eta_3 \max \left[n^{-1} p^{1/2} \exp \left\{ (1 + 2^{1/2}) p \right\}, n^{\eta_5 - 2/3} \right] \leq h \leq \epsilon_n p^{1/2} \exp \left\{ - (1 + 2^{1/2}) p \right\}. \quad (5.21)$$

If p is held fixed then $\mathcal{S}_0$ includes all pairs (h, p) for which h lies between constant multiples of $n^{\eta_5 - 2/3}$ and ϵ_n , where ϵ_n can decrease to zero arbitrarily slowly. In particular, if η_5 is sufficiently small then it includes all bandwidths of size $n^{-1/(2p+3)}$, which, for at least odd p , would be the appropriate size of h if g had just $p + 1$ bounded derivatives.

As is well known, the asymptotic variance and squared bias of $\hat{g}$ are of size $(nh)^{-1}$ and $h^{2(p+1)}$, respectively, provided that g has $p + 1$ bounded derivatives. Reflecting this fact, average summed squared error is of order $(nh)^{-1} + h^{2(p+1)}$, in probability, as $n \rightarrow \infty$. Arguments similar to those leading to Theorem 5.1 show that this remains true, modulo multiplication by a function of p that increases with p , if p diverges sufficiently slowly as n increases. Theorem 5.2, below, demonstrates that, for pairs (h, p) in $\mathcal{S}_0$, the cross-validation criterion $CV(h, p)$ captures average summed squared error, up to a remainder that is of strictly smaller order than $(nh)^{-1} + h^{2(p+1)}$, and a term $n^{-1} \sum_i \epsilon_i^2$ that depends on neither h nor p .

Theorem 5.2. *If (5.15), (5.16) and (5.20) hold then, for some $\eta > 0$,*

$$CV(h, p) = \left\{ 1 + O_p(h^{1/2} n^{-\eta}) \right\} \frac{1}{n} \sum_{i=1}^n \{ \hat{g}(X_i) - g(X_i) \}^2 + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 + o_p \left\{ (nh)^{-1} + h^{2(p+1)} \right\}, \quad (5.22)$$

uniformly in $(h, p) \in \mathcal{S}_0$.

The set $\mathcal{S}_0$ does not include arbitrarily large bandwidths, and in particular, $h = \infty$ is not allowed. This makes it awkward to deduce, from our theory, the way in which the cross-validation criterion behaves for very large h . However, if $\hat{g}$ is a local polynomial estimator then, for any fixed but arbitrarily large p , we have, as $h \rightarrow \infty$ for a fixed sample,

$$\text{CV}(h, p) \rightarrow \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_{ip, -i})^2,$$

where $\hat{Y}_{ip, -i}$ is a leave-one-out, global polynomial fit of degree p to the mean of Y_i given X_i .

Suppose that the true g is a polynomial of degree p_0 , say; call this model M_1 . Then, if the degree p of the local linear fit satisfies $p \geq p_0$, $\hat{g}$ is unbiased for all choices of $h > 0$, and the variance of $\hat{g}$ is minimised by taking $p = p_0$ and $h = \infty$. Still treating the case of model M_1 , if p is held fixed at a value greater than or equal to p_0 , then the value of h that minimises $\text{CV}(h, p)$ is finite with positive probability, and the same is true if h and p are allowed to vary together.

To appreciate why, let us suppose that the density of the distribution of the experimental errors ϵ_i is continuous and strictly positive on the entire real line, and let M_2 be a model where g is a very smooth function not representable as a polynomial, for example the standard normal density function. Then there is a nonzero probability that data pairs $(X_1, Y_1), \dots, (X_n, Y_n)$, drawn under M_1 , more closely resemble pairs from the incorrect model M_2 , where “more closely resemble” can be interpreted in any of many ways, for example in terms of the likelihood ratio statistic for testing M_1 against M_2 . Therefore, there is positive probability that the pair (h, p) that minimises $\text{CV}(h, p)$ has properties that would arise for M_2 rather than for M_1 . Those properties include a recommendation by cross-validation that tuning parameter choices reflect bias, even in cases where $\hat{g}$ is unbiased. In the presence of bias, choosing a finite h reflects a bias-variance trade-off not unlike that found when using shrinkage methods.

5.5. Relationship between Theorems 5.1 and 5.2. We claim that the pair (h, p) reported in part (a) of Theorem 5.1 lies in the set $\mathcal{S}_0$ that is addressed by Theorem 5.2. Therefore the latter theorem shows that cross-validation correctly captures the main effects of average summed squared error, up to terms that either do not depend on h or p , or are of strictly smaller order than $(nh)^{-1} + h^{2(p+1)}$.

To appreciate that this is correct, note that if $K = K_s$, and p satisfies $p \sim \frac{1}{2} \{(\log n)/(1 + 2^{1/2})\}^{1/2}$ as asserted in part (a) of Theorem 5.1 (recall from (5.9) that $B_3 = 1 + 2^{1/2}$), then by (5.17),

$$h \asymp n^{-1/(2p+3)} p \asymp (\log n)^{1/2} \exp \left[- \{1 + o(1)\} (1 + 2^{1/2})^{1/2} (\log n)^{1/2} \right].$$

These formulae for the sizes of h and p imply that $(h, p) \in \mathcal{S}_0$ for all sufficiently large n , provided that ϵ_n decreases sufficiently slowly.

REFERENCES

- Abramson, M., Audet, C., Couture, G., Dennis Jr., J. & Le Digabel, S. (2011), The NOMAD project, Technical report.
URL: *Software available at <http://www.gerad.ca/nomad>*
- Chen, Y. & Lubinsky, D. (2004), ‘Smallest eigenvalues of Hankel matrices for exponential weights’, *Journal of Mathematical Analysis and Applications* **293**, 476–495.
- Fan, J. (1992), ‘Design-adaptive nonparametric regression’, *Journal of the American Statistical Association* **87**, 998–1004.
- Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.
- Hitchenko, P. (1990), ‘Best constants in martingale version of Rosenthal’s inequality’, *Annals of Probability* **18**, 1656–68.
- Mincer, J. A. (1974), *Schooling, Experience, and Earnings*, number minc74-1 in ‘NBER Books’, National Bureau of Economic Research, Inc.
- Murphy, K. M. & Welch, F. (1990), ‘Empirical age-earnings profiles’, *Journal of Labor Economics* **8**(2), 202–229.
- Nadaraya, E. A. (1965), ‘On nonparametric estimates of density functions and regression curves’, *Theory of Applied Probability* **10**, 186–190.
- R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: <http://www.R-project.org/>
- Racine, J. S. & Nie, Z. (2012), *crs: Categorical Regression Splines*. R package version 0.15-18.
URL: <http://CRAN.R-project.org/package=crs>
- Ruppert, D. & Wand, M. P. (1994), ‘Multivariate locally weighted least squares regression’, *Annals of Statistics* **22**, 1346–1370.
- Schwarz, G. (1978), ‘Estimating the dimension of a model’, *The Annals of Statistics* **6**, 461–464.

- Szegő, G. (1936), 'On some hermitian forms associated with two given curves of the complex plane', *Transactions of the American Mathematical Society* **40**, 450–461.
- Ullah, A. (1985), 'Specification analysis of econometric models', *Journal of Quantitative Economics* **1**(2), 187–209.
- Watson, G. S. (1964), 'Smooth regression analysis', *Sankhya* **26:15**, 359–372.
- Widom, H. & Wilf, H. (1966), 'Small eigenvalues of large Hankel matrices', *Proceedings of the American Mathematical Society* **17**, 338–344.

APPENDIX A. PROOF THAT (5.15) HOLDS IF g IS AN EXPONENTIATED POLYNOMIAL

We shall give a proof when $g(x) = \exp(x^k)$, $k \geq 1$ is an integer and $\mathcal{I}$ is a compact interval. Only slight modifications are necessary when $g(x) = \exp\{\pi(x)\}$ and π is a polynomial, or when $g(x) = \exp\{\pi(|x|^\alpha)\}$ for some $\alpha > 0$, provided in this case that the compact interval $\mathcal{I}$ does not contain the origin.

If $g(x) = \exp(x^k)$ and $k \geq 1$ is an integer then

$$g^{(\ell)}(x) = \sum_{j=1}^{\infty} \frac{1}{j!} (kj) (kj-1) \dots (kj-\ell+1) x^{kj-\ell}. \quad (\text{A.1})$$

Note that in the sum over j here, terms for which $j \leq (\ell-1)/k$ vanish. In interpreting series below it should be understood that such terms are excluded for sums over j ; that is, it is implicitly assumed that $j > (\ell-1)/k$.

Using Stirling's formula we deduce that there exist absolute constants $C_1, C_2, C_3 > 0$ such that, whenever $1 \leq j \leq \ell-1$,

$$\frac{j^\ell}{j!} \leq C_1 \frac{j^\ell}{j^{1/2} (j/e)^j} \leq C_2 \frac{j^\ell (e/j)^\ell}{(j-\ell)^{1/2} \{(j-\ell)/e\}^j} \leq C_3 \frac{e^\ell}{(j-\ell)!}.$$

Therefore,

$$\begin{aligned} S_{1,\ell}(x) &\equiv \sum_{j=1}^{\ell-1} \frac{1}{j!} (kj) (kj-1) \dots (kj-\ell+1) x^{kj-\ell} \\ &\leq C_3 (ek x^{k-1})^\ell \sum_{j=1}^{\ell-1} \frac{x^{k(j-\ell)}}{(j-\ell)!} \leq C_3 (ek x^{k-1})^\ell \exp(x^k). \end{aligned} \quad (\text{A.2})$$

Note too that

$$\begin{aligned}
S_{2,\ell}(x) &\equiv \sum_{j=\ell}^{\infty} \frac{1}{j!} (kj) (kj-1) \dots (kj-\ell+1) x^{kj-\ell} \\
&= k^\ell \sum_{j=\ell}^{\infty} \frac{x^{kj-\ell}}{j!} j(j-1) \dots (j-\ell+1) \prod_{s=1}^{\ell-1} \left(1 + \frac{s(k-1)}{k(j-s)}\right) \\
&\leq k^\ell \sum_{j=\ell}^{\infty} \frac{x^{kj-\ell}}{(j-\ell)!} \exp \left\{ \sum_{s=1}^{\ell-1} \frac{s(k-1)}{k(j-s)} \right\}.
\end{aligned} \tag{A.3}$$

Approximating the sum of $s/(j-s)$ over $1 \leq s \leq \ell-1$ by the integral of $s/(j-s)$ over $0 \leq s \leq \ell-1$, and noting that the indefinite integral of $u/(1-u)$ is $-\log(1-u) - u$, we deduce that, for $j \geq \ell$,

$$\begin{aligned}
\sum_{s=1}^{\ell-1} \frac{s}{j-s} &\leq -\log[1 - \{(\ell-1)/j\}] - \{(\ell-1)/j\} + C_4 \ell \\
&\leq \frac{(\ell/j)^2}{2[1 - \{(\ell-1)/j\}]} + C_4 \ell = \frac{\ell^2}{2j(j-\ell+1)} + C_4 \ell \leq \left(\frac{1}{2} + C_4\right) \ell,
\end{aligned}$$

where $C_4 > 0$ is an absolute constant. Substituting this formula into the far right-hand side of (A.3), and defining $C_5 = k \exp(\frac{1}{2} + C_4)$, we obtain:

$$S_{2,\ell}(x) \leq C_5^\ell \sum_{j=\ell}^{\infty} \frac{x^{kj-\ell}}{(j-\ell)!} = (C_5 x^{k-1})^\ell \exp(x^k). \tag{A.4}$$

Combining (A.1), (A.2) and (A.4) we deduce that

$$g^{(\ell)}(x) = S_{1,\ell}(x) + S_{2,\ell}(x) \leq C_6 C_7^\ell, \tag{A.5}$$

where $C_6 = (C_3 + 1) \sup_{x \in \mathcal{I}} \exp(x^k)$ and $C_7 = (ek + C_5) \sup_{\mathcal{I}} x^{k-1}$. The desired result (5.15) follows from (A.5).

APPENDIX B. PROOF OF THEOREM 5.1 (NOT FOR PUBLICATION)

Step 1: First upper bound to $|\hat{g} - g|$. The bound is given at (B.6). To derive it, note that by Taylor expansion,

$$g(X_i) = \sum_{k=0}^p \frac{\{-(x - X_i)\}^k}{k!} g^{(k)}(x) + \frac{\{-(x - X_i)\}^{p+1}}{(p+1)!} g^{(p+1)}(x_i),$$

where x_i lies between x and X_i . Therefore, defining

$$G = (g^{(0)}, -h g^{(1)}, \dots, (-h)^p g^{(p)} / p!)^T,$$

we have:

$$\begin{aligned} \bar{g}_j(x) &\equiv \frac{1}{nh} \sum_{i=1}^n g(X_i) \left(\frac{x - X_i}{h} \right)^j K\left(\frac{x - X_i}{h} \right) = \sum_{k=0}^p \hat{m}_{jk}(x) g^{(k)}(x) \frac{(-h)^k}{k!} + R_j(x) \\ &= \{\widehat{M}(x) G(x)\}_j + R_j(x), \end{aligned}$$

where, here and below, notation such as $\{\widehat{M}(x) G(x)\}_j$ denotes the j th component of the $(p+1)$ -vector $\widehat{M}(x) G(x)$, and

$$R_j(x) = \frac{1}{nh} \frac{(-h)^{p+1}}{(p+1)!} \sum_{i=1}^n g^{(p+1)}(x_i) \left(\frac{x - X_i}{h} \right)^{j+p+1} K\left(\frac{x - X_i}{h} \right). \quad (\text{B.1})$$

Define

$$\bar{\epsilon}_j(x) = \frac{1}{nh} \sum_{i=1}^n \epsilon_i \left(\frac{x - X_i}{h} \right)^j K\left(\frac{x - X_i}{h} \right), \quad (\text{B.2})$$

$\bar{\epsilon} = (\bar{\epsilon}_0, \dots, \bar{\epsilon}_p)^T$, $U_j = \bar{\epsilon}_j + R_j$, $U = (U_0, \dots, U_p)^T$ and $\bar{g} = (\bar{g}_0, \dots, \bar{g}_p)^T$. Then, by (2.1), and (B),

$$V = \bar{g} + \bar{\epsilon} = \widehat{M} G + U.$$

Hence, by (2.5) and (2.6), $\hat{c} = G + \widehat{M}^{-1} U$ and

$$\hat{g}(x) = g(x) + \{\widehat{M}(x)^{-1} U(x)\}_0, \quad (\text{B.3})$$

where, here and below, the subscript 0 indicates that we take the zeroth component.

Note too that, if $\widehat{\text{ev}}_p$ denotes the smallest eigenvalue of $\widehat{M}$, then

$$|(\widehat{M}^{-1} U)_0| \leq \frac{\|U\|}{\widehat{\text{ev}}_p} \leq \frac{2^{1/2}}{\widehat{\text{ev}}_p} \left\{ \sum_{j=0}^p (\bar{\epsilon}_j^2 + R_j^2) \right\}^{1/2}. \quad (\text{B.4})$$

Recall that $m_{jk} = \int u^{j+k} K(u) du$, that $M = (m_{jk})$ and that $\text{ev}_p(K)$ is the smallest eigenvalue of M , where neither M nor $\text{ev}_p(K)$ depends on x . Given a $(p+1) \times (p+1)$ matrix $A = (a_{jk})$, define $\|A\| = (\sum_j \sum_k a_{jk}^2)^{1/2}$. Let v denote a $(p+1)$ -vector of unit length, write CV for the class of all such vectors, and observe that if $v \in \text{CV}$ then

$$v^T \widehat{M} v = f v^T M v - v^T (\widehat{M} - f M) v \geq f \text{ev}_p(K) - \|\widehat{M} - f M\|.$$

Taking the infimum of the far left-hand side over all $v \in \text{CV}$ we deduce that, for each $x \in \mathcal{I}$,

$$\widehat{\text{ev}}_p(x) \geq f(x) \text{ev}_p(K) - \|\widehat{M}(x) - f(x) M\|. \quad (\text{B.5})$$

Let $\mathcal{E}(x)$ denote the event that $\|\widehat{M}(x) - f(x) M\| < f(x) \text{ev}_p(K)$. If $\mathcal{E}(x)$ holds then it follows from (B.3), (B.4) and (B.5) that

$$\begin{aligned} |\hat{g}(x) - g(x)| &= |\{\widehat{M}(x)^{-1} U(x)\}_0| \\ &\leq 2^{1/2} \left\{ f(x) \text{ev}_p(K) - \|\widehat{M}(x) - f(x) M\| \right\}^{-1} \\ &\quad \times \left[\sum_{j=0}^p \{\bar{\epsilon}_j(x)^2 + R_j(x)^2\} \right]^{1/2}. \end{aligned} \quad (\text{B.6})$$

Step 2: Bounds for expected value of term in square brackets in (B.6). The bounds are given at (B.12) and (B.13). It follows from (B.1) that

$$|R_j(x)| \leq \frac{h^{p+1}}{(p+1)!} \left\{ \sup_{u \in \mathcal{I}} |g^{(p+1)}(u)| \right\} \frac{1}{nh} \sum_{i=1}^n \left| \frac{x - X_i}{h} \right|^{j+p+1} K\left(\frac{x - X_i}{h}\right). \quad (\text{B.7})$$

Let $s_p = \{(p+1)!\}^{-2} \sup_{u \in \mathcal{I}} |g^{(p+1)}(u)|^2$, $s_{\sup} = \max\{\sup f_X, (\sup f_X)^2\}$,

$$\begin{aligned} s_{\text{sum}} &= \frac{B_4 2^{-(s+1)}}{nh} \sum_{j=0}^p (j+p+1)^{-(s+1)} + C_1 \left\{ B_4 \sum_{j=0}^p (j+p+1)^{-(s+1)} \right\}^2 \\ &\leq \max(B_4, B_4^2) C_2 p^{-s} \end{aligned} \quad (\text{B.8})$$

if (5.14)(a) holds, and

$$\begin{aligned} s_{\text{sum}} &= \frac{B_4}{nh} \sum_{j=0}^p \{2 B_5 (j+p+1)\}^{j+p+1} + C_1 B_4 \left[\sum_{j=0}^p \{B_5 (j+p+1)\}^{(j+p+1)/2} \right]^2 \\ &\leq C_2 \max(B_4, B_4^2) \left[(nh)^{-1} \{2 B_5 (2p+1)\}^{2p+1} + \{B_5 (2p+1)\}^{2p+1} \right] \end{aligned} \quad (\text{B.9})$$

if (5.14)(b) prevails, where the constants C_j in (B.8), (B.9) and below are absolute, and in particular do not depend on h , n or p , and we employed (5.12) to derive (B.8) and (B.9).

Using (5.15) to bound $\sup_{u \in \mathcal{I}} |g^{(p+1)}(u)|$, in the definition of s_p , we deduce that

$$\begin{aligned} \sum_{j=0}^p E\{R_j(x)^2\} &\leq h^{2(p+1)} s_p \left[(nh)^{-1} (\sup f_X) \sum_{j=0}^p \int u^{2(j+p+1)} K(u) du \right. \\ &\quad \left. + \left\{ (\sup f_X) \sum_{j=0}^p \int |u|^{j+p+1} K(u) du \right\}^2 \right] \\ &\leq \{(p+1)!\}^{-2} B_6^2 (B_7 h)^{2(p+1)} s_{\sup} s_{\text{sum}} . \end{aligned} \quad (\text{B.10})$$

Writing $\sigma^2 = \sup_{j \geq 1} \text{var}(\epsilon_j)$ we obtain:

$$\begin{aligned} nh \sum_{j=0}^p E\{\bar{\epsilon}_j(x)^2\} &\leq (\sup f_X) \sigma^2 \sum_{j=0}^p \int u^{2j} K(u)^2 du \\ &\leq C_3 (\sup f_X) \sigma^2 \times \begin{cases} \sup K & \text{if (5.14)(a) holds} \\ B_4 (B_5 p)^p & \text{if (5.14)(b) holds} . \end{cases} \end{aligned} \quad (\text{B.11})$$

Combining (B.8)–(B.11) we deduce that, uniformly in $x \in \mathcal{I}$,

$$\sum_{j=0}^p \left[E\{\bar{\epsilon}_j(x)^2\} + E\{R_j(x)^2\} \right] = O\left[\{(p+1)!\}^{-2} (B_7 h)^{2(p+1)} p^{-s} + (nh)^{-1}\right] \quad (\text{B.12})$$

if (5.14)(a) holds, and equals

$$O\left[\{(p+1)!\}^{-2} (B_7 h)^{2(p+1)} (2 B_5 p)^{2p+1} + (nh)^{-1} (2 B_5 p)^p\right] \quad (\text{B.13})$$

if (5.14)(b) holds, where the orders of magnitude in (B.12) and (B.13) are valid uniformly in x .

Step 3: Uniform bound for the term involving $\text{ev}_p(K)$ in (B.6). The bound is given at (B.31) below. Recall that $\widehat{M} = (\widehat{m}_{jk})$, where the function $\widehat{m}_{jk}$ is given by (2.4), and that $M_p = (m_{jk})$ is as defined in the first paragraph of section 5.1. Now,

$$|\widehat{m}_{jk} - f_X m_{jk}| \leq |\widehat{m}_{jk} - E(\widehat{m}_{jk})| + |E(\widehat{m}_{jk}) - f_X m_{jk}|, \quad (\text{B.14})$$

where $\widehat{m}_{jk}$ and f_X are interpreted as functions, and m_{jk} is not a function. Furthermore,

$$\begin{aligned} |E\{\widehat{m}_{jk}(x)\} - f_X(x) m_{jk}| &\leq h (\sup |f'_X|) \int |u|^{j+k+1} K(u) du \\ &\leq h (\sup |f'_X|) B_4 \times \begin{cases} (j+k+1)^{-(s+1)} & \text{if (5.14)(a) holds} \\ \{B_5 (j+k+1)\}^{(j+k+1)/2} & \text{if (5.14)(b) holds} \end{cases} \end{aligned} \quad (\text{B.15})$$

and if $r \geq 1$ is an integer,

$$\begin{aligned}
& E \left[\left| \widehat{m}_{jk}(x) - E\{\widehat{m}_{jk}(x)\} \right|^{2r} \right] \\
& \leq A_1(r) \left(\left[\text{var}\{\widehat{m}_{jk}(x)\} \right]^r + n E \left| \frac{1}{nh} (1 - E) \left\{ \left(\frac{x - X_1}{h} \right)^{j+k} K \left(\frac{x - X_1}{h} \right) \right\} \right|^{2r} \right) \\
& \leq A_2(r) \left[\left\{ \frac{(\sup f_X)(\sup K)}{nh} \int u^{2(j+k)} K(u) du \right\}^r \right. \\
& \quad \left. + \frac{(\sup f_X)(\sup K)^{2r-1}}{(nh)^{2r-1}} \int u^{2r(j+k)} K(u) du \right] \tag{B.16}
\end{aligned}$$

where, here and below, the positive constants $A_j(r)$ depend only on r and, for any random variable R , $(1 - E) R$ denotes $R - E(R)$.

If (5.14)(a) holds then (B.16) above implies that

$$\begin{aligned}
E \left[\left| \widehat{m}_{jk}(x) - E\{\widehat{m}_{jk}(x)\} \right|^{2r} \right] & \leq A_3(r) \left[\left\{ \frac{(\sup f_X)(\sup K)}{nh} B_4(j+k)^{-(s+1)} \right\}^r \right. \\
& \quad \left. + \frac{(\sup f_X)(\sup K)^{2r-1}}{(nh)^{2r-1}} B_4(j+k)^{-(s+1)} \right]. \tag{B.17}
\end{aligned}$$

On the other hand, if (5.14)(b) obtains then

$$\begin{aligned}
E \left[\left| \widehat{m}_{jk}(x) - E\{\widehat{m}_{jk}(x)\} \right|^{2r} \right] & \leq A_3(r) \left[\left\{ \frac{(\sup f_X)(\sup K)}{nh} B_4 \{2 B_5(j+k)\}^{j+k} \right\}^r \right. \\
& \quad \left. + \frac{(\sup f_X)(\sup K)^{2r-1}}{(nh)^{2r-1}} B_4 \{2 B_5 r(j+k)\}^{r(j+k)} \right]. \tag{B.18}
\end{aligned}$$

By Markov's inequality,

$$\begin{aligned}
\pi_{pr}(x) &\equiv \left\{ \frac{1}{3} f_X(x) \text{ev}_p(K) \right\}^{2r} P \left[\left\| \widehat{M}(x) - E\{\widehat{M}(x)\} \right\| > \frac{1}{3} f_X(x) \text{ev}_p(K) \right] \\
&\leq E \left[\left\| \widehat{M}(x) - E\{\widehat{M}(x)\} \right\|^{2r} \right] \\
&\leq \left[\sum_{j=1}^{p+1} \sum_{k=1}^{p+1} \left\{ E |\widehat{m}_{jk}(x) - E\widehat{m}_{jk}(x)|^{2r} \right\}^{1/r} \right]^r.
\end{aligned} \tag{B.19}$$

If (5.14)(a) holds then, by (B.17) and (B.19), and for each fixed integer $r \geq 1$,

$$\begin{aligned}
\sup_{x \in \mathcal{I}} \pi_{pr}(x) &= O \left\{ \left(\sum_{j=1}^{p+1} \sum_{k=1}^{p+1} \left\{ \frac{1}{nh} (j+k)^{-(s+1)} + \frac{1}{(nh)^{2-(1/r)}} (j+k)^{-(s+1)/r} \right\} \right)^r \right\} \\
&= O \left\{ \left(\frac{p}{nh} + \frac{p^{2-(1/r)}}{(nh)^{2-(1/r)}} \right)^r \right\} = O \left\{ \left(\frac{p}{nh} \right)^r \right\},
\end{aligned} \tag{B.20}$$

where the last identity holds provided that $p = O(nh)$. In the case of (5.14)(b), (B.18) and (B.19) imply that

$$\begin{aligned}
\sup_{x \in \mathcal{I}} \pi_{pr}(x) &= O \left\{ \left(\sum_{j=1}^{p+1} \sum_{k=1}^{p+1} \left[\frac{1}{nh} \{2 B_5 (j+k)\}^{j+k} \right. \right. \right. \\
&\quad \left. \left. \left. + \frac{1}{(nh)^{2-(1/r)}} \{2 B_5 r (j+k)\}^{j+k} \right] \right)^r \right\} \\
&= O \left[\left\{ \frac{(4 B_5 p)^{2p}}{nh} + \frac{(4 B_5 r p)^{2p}}{(nh)^{2-(1/r)}} \right\}^r \right] = O \left[\left\{ \frac{(4 B_5 p)^{2p}}{nh} \right\}^r \right],
\end{aligned} \tag{B.21}$$

where the last identity holds provided that $(4 B_5 p)^{2p} = O(nh)$ (this property entails $C^p/nh = o(1)$ for any $C > 0$).

Properties (B.20) and (B.21), and the lower bounds on $\text{ev}_p(K)$ in (5.14)(a) and (5.14)(b), imply that

$$\begin{aligned} \sup_{x \in \mathcal{I}} P \left[\left\| \widehat{M}(x) - E\{\widehat{M}(x)\} \right\| > \frac{1}{3} f_X(x) \text{ev}_p(K) \right] \\ = O \left[\left\{ p^{B_2+1} \exp(2 B_3 p) / nh \right\}^r \right], \end{aligned} \quad (\text{B.22})$$

$$\begin{aligned} \sup_{x \in \mathcal{I}} P \left[\left\| \widehat{M}(x) - E\{\widehat{M}(x)\} \right\| > \frac{1}{3} f_X(x) \text{ev}_p(K) \right] \\ = O \left(\left[\left\{ 4 B_5 p^{B_2+1} \exp(B_3 p^{-1/2}) \right\}^{2p} / nh \right]^r \right), \end{aligned} \quad (\text{B.23})$$

in cases where (5.14)(a) and (5.14)(b) holds, respectively.

Assume that:

$$\begin{aligned} &\text{in the respective contexts of (5.14)(a) and (5.14)(b), and for some } \delta_1 > 0, \\ &\exp(2 B_3 p) / nh = O(n^{-\delta_1}) \text{ and } p^{2p} / nh = O(n^{-\delta_1}). \end{aligned} \quad (\text{B.24})$$

Then the assumptions $p = O(nh)$ and $(4 B_5 p)^{2p} = O(nh)$, which were imposed as preludes to (B.20) and (B.21), respectively, hold. Moreover, for some $\delta_2 > 0$ not depending on r , the right-hand sides of (B.22) and (B.23) equal $O(n^{-r\delta_2})$ for all $r > 0$. Therefore, if $\mathcal{I}_n$ denotes a set of points on a regular grid in $\mathcal{I}$, of only polynomial fineness as a function of n , then

$$P \left[\sup_{x \in \mathcal{I}_n} f_X(x)^{-1} \left\| \widehat{M}(x) - E\{\widehat{M}(x)\} \right\| > \frac{1}{3} \text{ev}_p(K) \right] = O(n^{-C}) \quad (\text{B.25})$$

for all $C > 0$.

If $r \geq 1$ is an integer then, when (5.14)(a) holds, $|x|^r K(x)$ and $|x|^r |K'(x)|$ are both uniformly bounded, and when (5.14)(b) obtains are both bounded above by r^{2r} . Using these properties it can be proved that, in the case of (5.14)(a), $|x_1^r K(x_1) - x_2^r K(x_2)| \leq C_1 |x_1 - x_2|$, and when (5.14)(b) obtains, $|x_1^r K(x_1) - x_2^r K(x_2)| \leq C_1 (2p)^p |x_1 - x_2|$, where $C_1 > 0$ is a constant not depending on r , and both bounds hold uniformly in $x_1, x_2 \in \mathbb{R}$ and positive

integers $r \leq 2p$. Hence, if (B.24) holds then, in the case of (5.14)(a), with probability 1,

$$\sup_{x_1, x_2 \in \mathbb{R}} |x_1 - x_2|^{-1} \max_{0 \leq j, k \leq p} |\widehat{m}_{jk}(x_1) - \widehat{m}_{jk}(x_2)| \leq \frac{C_2 n}{h} \times \begin{cases} 1 & \text{if (5.14)(a) holds} \\ p^{2p} & \text{if (5.14)(b) holds} \end{cases}$$

Hence, if $C_3 > 0$ is given, if (B.24) holds, and if $\mathcal{I}_n$ is a grid with edge width n^{-C_4} , then for sufficiently large C_4 ,

$$P \left\{ \sup_{x_1 \in \mathcal{I}} \sup_{x_2 \in \mathcal{I}_n} \|\widehat{M}(x_1) - \widehat{M}(x_2)\| > n^{-C_3} \right\} = 0. \quad (\text{B.26})$$

It follows from (5.14) and (B.24) that, for some $C_4 > 0$, $\text{ev}_p(K)$ is bounded below by a constant multiple of n^{-C_4} , and so if C_3 is chosen larger than C_4 then (B.26) implies the version of (B.25) with $\mathcal{I}$ replacing $\mathcal{I}_n$:

$$P \left[\sup_{x \in \mathcal{I}} f_X(x)^{-1} \|\widehat{M}(x) - E\{\widehat{M}(x)\}\| > \frac{1}{3} \text{ev}_p(K) \right] = O(n^{-C}) \quad (\text{B.27})$$

for all $C > 0$.

More simply, the bound at (B.15) implies that

$$\begin{aligned} & \sup_{x \in \mathcal{I}} \sum_{j=1}^p \sum_{k=1}^p |E\{\widehat{m}_{jk}(x)\} - f_X(x) m_{jk}|^2 \\ & \leq \{h (\sup |f'_X|) B_4\}^2 \\ & \quad \times \begin{cases} \sum_{j \leq p} \sum_{k \leq p} (j+k+1)^{-2(s+1)} & \text{if (5.14)(a) holds} \\ \sum_{j \leq p} \sum_{k \leq p} \{B_5 (j+k+1)\}^{j+k+1} & \text{if (5.14)(b) holds} \end{cases} \quad (\text{B.28}) \\ & \leq C_5 h^2 \times \begin{cases} 1 & \text{if (5.14)(a) holds} \\ (2 B_5)^{2p+3} & \text{if (5.14)(b) holds.} \end{cases} \end{aligned}$$

Using the lower bounds for $\text{ev}_p(K)$ given in (5.14) we deduce from (B.28) that, provided that

$$\begin{aligned} p^{2B_2} h^2 \exp(2 B_3 p) &\rightarrow 0 \text{ when (5.14)(a) holds, or, in the case of (5.14)(b),} \\ p^{2B_2} h^2 (2 B_5)^{2p+3} \exp(2 B_3 p^{1/2}) &\rightarrow 0, \end{aligned} \quad (\text{B.29})$$

we have:

$$P \left[\left\| E \{ \widehat{M}(x) \} - f_X(x) M_p \right\| > \frac{1}{3} f_X(x) \text{ev}_p(K) \quad \text{for some } x \in \mathcal{I} \right] = 0 \quad (\text{B.30})$$

for all sufficiently large n . Combining (B.14), (B.27) and (B.30) we deduce that

$$P \left\{ \left\| \widehat{M}(x) - f_X(x) M_p \right\| > \frac{2}{3} f_X(x) \text{ev}_p(K) \quad \text{for some } x \in \mathcal{I} \right\} = O(n^{-C}). \quad (\text{B.31})$$

for all $C > 0$. Result (B.31) also implies that the probability that $\mathcal{E}(x)$ holds for all $x \in \mathcal{I}$, equals $1 - O(n^{-C})$ for all $C > 0$.

Step 4: Rate of convergence of $\hat{g}$ to g . The rates are given by (B.34)–(B.36). Define

$$d(h, n, p) = p^{2B_2} \exp(2 p B_3) \left[\{(p+1)!\}^{-2} (B_7 h)^{2(p+1)} p^{-s} + (nh)^{-1} \right] \quad (\text{B.32})$$

if (5.14)(a) holds, and

$$\begin{aligned} d(h, n, p) = p^{2B_2} \exp(2 B_3 p^{1/2}) &\left[\{(p+1)!\}^{-2} (B_7 h)^{2(p+1)} (2 B_5 p)^{2p+1} \right. \\ &\left. + (nh)^{-1} (2 B_5 p)^p \right] \end{aligned} \quad (\text{B.33})$$

when (5.14)(b) obtains. Combining (B.3), (B.6), (B.12), (B.13) and (B.31) we deduce that if (B.24) and (B.29) hold then so too do the following properties, whenever w is a nonnegative,

uniformly bounded function:

$$\frac{1}{n} \sum_{i: X_i \in \mathcal{I}} \{\hat{g}(X_i) - g(X_i)\}^2 w(X_i) = O_p\{d(h, n, p)\}, \quad (\text{B.34})$$

$$\int_{\mathcal{I}} \{\hat{g}(x) - g(x)\}^2 w(x) dx = O_p\{d(h, n, p)\}, \quad (\text{B.35})$$

$$\text{for each } x \in \mathcal{I}, \quad \hat{g}(x) - g(x) = O_p\{d(h, n, p)^{1/2}\}. \quad (\text{B.36})$$

Step 5: Conclusion when (5.14)(a) holds. Note that $\{(p+1)!\}^{-2} \asymp p^{-3} (e/p)^{2p}$ as $p \rightarrow \infty$, and so if (5.14)(a) obtains then $d(h, n, p)$, given at (B.32), is bounded above and below by constant multiples of

$$p^{2B_2-(s+3)} \{\exp(B_3+1) B_7/p\}^{2p} h^{2(p+1)} + p^{2B_2} \exp(2p B_3) (nh)^{-1}, \quad (\text{B.37})$$

which is minimised by a bandwidth for which

$$h^{2p+3} \asymp n^{-1} p^{s+2} \exp(2p B_3) [p/\{\exp(B_3+1) B_7\}]^{2p}. \quad (\text{B.38})$$

If h satisfies (B.38) then the the first and second terms in (B.37) are respectively bounded above and below by constant multiples of $p^{-1} \psi(p)$ and $\psi(p)$, where $\psi(p) = p^{2B_2-(s+2)} (2 B_5 p)^p \exp(2p B_3) [p/\{\exp(B_3+1) B_7\}]^{2p}$. In particular, the contribution of bias is asymptotically negligible, the main impact comes from variance, and the quantity in (B.37) is asymptotic to a constant multiple of

$$p^{2B_2} \exp(2p B_3) n^{-1} \cdot n^{1/(2p+3)} p^{-1} = p^{2B_2-1} \exp(2p B_3) n^{-2(p+1)/(2p+3)}. \quad (\text{B.39})$$

Taking the derivative with respect to p of the negative logarithm of the right-hand side of (B.39), and equating it to zero, we obtain the equation

$$\frac{2 \log n}{(2p+3)^2} - 2 B_3 + \frac{1-2 B_2}{p} = 0,$$

the solution of which satisfies

$$p = \frac{1}{2} \left(\frac{\log n}{B_3} \right)^{1/2} \{1 + o(1)\}. \quad (\text{B.40})$$

If p is given by (B.40) then the negative logarithm of the right-hand side of (B.39) equals

$$(1 - 2B_2) \log p - 2pB_3 + \frac{2(p+1)}{2p+3} \log n = \log n - 2(B_3 \log n)^{1/2} \{1 + o(1)\},$$

where the identity follows on substituting into the left-hand side the formula for p at (B.40).

In this case the right-hand side of (B.39) is given by

$$n^{-1} \exp [2(B_3 \log n)^{1/2} \{1 + o(1)\}]. \quad (\text{B.41})$$

Therefore, in view of (B.34)–(B.36) and the argument in the first paragraph of this section, the quantity in (B.41) is an upper bound for the order of magnitude, in probability, of both $\int_{\mathcal{I}} (\hat{g} - g)^2$ and $\{\hat{g}(x) - g(x)\}^2$. Theorem 5.1, in the case of (5.14)(a), follows directly from this result.

Note too that if h and p satisfy (B.38) and (B.40), respectively, then

$$h \asymp \exp [- (B_3 \log n)^{1/2} \{1 + o(1)\}],$$

$$p^c \exp(2B_3 p) = \exp [(B_3 \log n)^{1/2} \{1 + o(1)\}],$$

where the latter result holds for any real number c , positive or negative. It follows that the parts of (B.24) and (B.29) that pertain to (5.14)(a) hold.

Step 6: Conclusion when (5.14)(b) holds. In this instance $d(h, n, p)$, at (B.33), is bounded above and below by constant multiples of

$$p^{2B_2} (2B_5)^p \exp(2B_3 p^{1/2}) \left\{ p^{-2} (B_7 e)^{2p} h^{2(p+1)} + (nh)^{-1} p^p \right\}. \quad (\text{B.42})$$

Differentiating with respect to h , and equating to zero, we deduce that the bandwidth that minimises the expression at (B.42) satisfies

$$h^{2p+3} \asymp n^{-1} (B_7 e)^{-2p} p^{p+1}. \quad (\text{B.43})$$

As in Step 5 the contribution of the term in $h^{2(p+1)}$ in (B.42), representing bias, for any bandwidth satisfying (B.43), is asymptotically negligible, and so when h satisfies (B.43) (and hence has the property $h \asymp n^{-1/(2p+3)} p^{1/2}$), the quantity at (B.42) is bounded above and below by constants multiples of

$$p^{2B_2-(1/2)} (2 B_5 p)^p \exp(2 B_3 p^{1/2}) n^{-2(p+1)/(2p+3)}. \quad (\text{B.44})$$

Differentiating the negative logarithm of this expression, and equating to zero, we obtain the equation

$$\frac{2 \log n}{(2p+3)^2} - \log(2 B_5) - B_3 p^{-1/2} - \log p + \frac{(1/2) - 2 B_2}{p} - 1 = 0, \quad (\text{B.45})$$

the solution of which satisfies

$$p = \left(\frac{\log n}{\log \log n} \right)^{1/2} \{1 + o(1)\}. \quad (\text{B.46})$$

For such a p the negative logarithm of the quantity at (B.44) equals

$$\begin{aligned} & \left(\frac{1}{2} - 2 B_2 \right) \log p - p \log p - p \log(2 B_5) - 2 B_3 p^{1/2} + \frac{2(p+1)}{2p+3} \log n \\ & = \log n - \{(\log n) \log \log n\}^{1/2} \{1 + o(1)\}. \end{aligned}$$

It follows, as before, that an upper bound for the order of magnitude, in probability, of both $\int_{\mathcal{I}} (\hat{g} - g)^2$ and $\{\hat{g}(x) - g(x)\}^2$ is given by

$$n^{-1} \exp \left[\{(\log n) \log \log n\}^{1/2} \{1 + o(1)\} \right].$$

This result, and (B.34)–(B.36), imply Theorem 5.1 in the case of (5.14)(b).

If h and p satisfy (B.43) and (B.46) then

$$h \asymp \exp \left[-\frac{1}{2} \{(\log n) \log \log n\}^{1/2} \{1 + o(1)\} \right] \quad (\text{B.47})$$

and, for all real c ,

$$\begin{aligned} p^c h^2 (2 B_5)^{2p+3} \exp(2 B_3 p^{1/2}) / p^{1/2} \\ = \exp \left[-\{(\log n) \log \log n\}^{1/2} \{1 + o(1)\} \right]. \end{aligned} \quad (\text{B.48})$$

(The quantity on the left-hand side here appears in (B.29) in the case of (5.14)(b).) Properties (B.47) and (B.48) imply that (B.24) and (B.29) hold.

APPENDIX C. PROOF OF THEOREM 5.2 (NOT FOR PUBLICATION)

Step 1: Bound for moments of the first part of the cross-product term in (2.8).

Recall the definition of $\hat{g}_{-i}$ given in section 2.3. The cross-product term is proportional to

$$T \equiv \frac{1}{n} \sum_{i=1}^n \{g(X_i) - \hat{g}_{-i}(X_i)\} \epsilon_i = T_1 - T_2, \quad (\text{C.1})$$

where

$$T_1 = \frac{1}{n} \sum_{i=1}^n \{g(X_i) - \tilde{g}_{-i}(X_i)\} \epsilon_i, \quad T_2 = \frac{1}{n} \sum_{i=1}^n \Delta_i \epsilon_i, \quad (\text{C.2})$$

$$\tilde{g}_{-i}(x) = E\{\hat{g}_{-i}(x) \mid \mathcal{X}\} \quad (\text{C.3})$$

and $\Delta_i = \hat{g}_{-i}(X_i) - \tilde{g}_{-i}(X_i)$. The term T_1 in (C.2) represents the first part of the cross-product term and the bound is given at (C.4).

Conditional on $\mathcal{X} = \{X_1, \dots, X_n\}$, T_1 equals a sum of n independent random variables, and thus it can be proved using Rosenthal's inequality that, for each integer $r \geq 1$,

$$\begin{aligned} E(T_1^{2r} \mid \mathcal{X}) &\leq \left\{ \frac{A r}{n \log(2r)} \right\}^{2r} \left(\left[E(\epsilon_1^2) \sum_{i=1}^n \{g(X_i) - \tilde{g}_{-i}(X_i)\}^2 \right]^r \right. \\ &\quad \left. + E(\epsilon_1^{2r}) \max_{1 \leq i \leq n} \{g(X_i) - \tilde{g}_{-i}(X_i)\}^{2r} \right) \\ &\leq \left\{ \frac{C_1 r}{n \log(2r)} \right\}^{2r} \left[\sum_{i=1}^n \{g(X_i) - \tilde{g}_{-i}(X_i)\}^2 \right]^r, \end{aligned} \quad (\text{C.4})$$

where $A > 0$ is an absolute constant and, in (C.4) and below, $C_1, C_2, \dots$ denote positive constants not depending on h or n . See Hitchenko (1990) for discussion of the constant in Rosenthal's inequality. In deriving (C.4) we used the fact that $E(\epsilon_1^{2r}) \leq C^{2r}$ for a constant $C > 0$.

Step 2: Bound for moments of the second part of the cross-product term in (2.8).

The second part of the cross-product term is T_2 , at (C.2), and the bound is at (C.11).

We can write

$$\Delta_i = \{ \widehat{M}_{-i}(X_i)^{-1} W_{-i}(X_i) \}_0,$$

i.e. Δ_i equals the zeroth component of the $(p+1)$ -vector $\widehat{M}_{-i}(x)^{-1} W_{-i}(x)$, where $\widehat{M}_{-i}$, a $(p+1) \times (p+1)$ matrix, is the version of $\widehat{M}$ when the i th data pair is omitted from the sample, and $W_{-i} = (W_{0,-i}, \dots, W_{p,-i})^T$, with the scalar $W_{j,-i}(x)$ given by

$$W_{j,-i}(x) = \frac{1}{(n-1)h} \sum_{i_1: i_1 \neq i} \epsilon_{i_1} \left(\frac{x - X_{i_1}}{h} \right)^j K \left(\frac{x - X_{i_1}}{h} \right).$$

Thus,

$$\Delta_i = \sum_{i_1: i_1 \neq i} w_{i_1 i} \epsilon_{i_1}, \quad (\text{C.5})$$

where

$$w_{i_1 i} = \frac{1}{(n-1)h} \sum_{j=0}^p \widehat{m}_{0j,-i}^{(-1)}(X_i) \left(\frac{X_i - X_{i_1}}{h} \right)^j K \left(\frac{X_i - X_{i_1}}{h} \right) \quad (\text{C.6})$$

and $\widehat{m}_{jk,-i}^{(-1)}$ denotes the (j, k) th component of $\widehat{M}_{-i}^{(-1)}$.

Define too

$$Z_i = \sum_{i_1=1}^{i-1} (w_{ii_1} + w_{i_1i}) \epsilon_{i_1} \epsilon_i,$$

where $2 \leq i \leq n$. Then $E(Z_i | \mathcal{X}; \epsilon_1, \dots, \epsilon_{i-1}) = 0$, implying that the Z_i s are martingale differences with respect to the increasing sequence of sigma-fields $\mathcal{F}_i$ generated by $\mathcal{X}$ and $\epsilon_1, \dots, \epsilon_i$. Moreover, in view of (C.5),

$$n T_2 = \sum_{i=1}^n \Delta_i \epsilon_i = \sum_{i=1}^n Z_i.$$

Therefore, using Rosenthal's inequality again, we deduce that

$$\begin{aligned} E\{(n T_2)^{2r} | \mathcal{X}\} &\leq \left\{ \frac{A r}{\log(2r)} \right\}^{2r} \left(E \left[\left\{ \sum_{i=1}^n E(Z_i^2 | \mathcal{F}_{i-1}) \right\}^r \middle| \mathcal{X} \right] \right. \\ &\quad \left. + \max_{1 \leq i \leq n} E(Z_i^{2r} | \mathcal{X}) \right). \end{aligned} \tag{C.7}$$

Defining $\sigma^2 = \text{var}(\epsilon_1)$, it can be shown that

$$E(Z_i^2 | \mathcal{F}_{i-1}) = \sigma^2 \left\{ \sum_{i_1=1}^{i-1} (w_{ii_1} + w_{i_1i}) \epsilon_{i_1} \right\}^2.$$

Therefore, employing Rosenthal's inequality once more, this time for a sum of independent random variables,

$$\begin{aligned} E \left[\left\{ E(Z_i^2 | \mathcal{F}_{i-1}) \right\}^r \middle| \mathcal{X} \right] &= \sigma^{2r} E \left[\left\{ \sum_{i_1=1}^{i-1} (w_{ii_1} + w_{i_1i}) \epsilon_{i_1} \right\}^{2r} \middle| \mathcal{X} \right] \\ &\leq \left\{ \frac{A r \sigma}{\log(2r)} \right\}^{2r} \left[\sigma^{2r} \left\{ \sum_{i_1=1}^{i-1} (w_{ii_1} + w_{i_1i})^2 \right\}^r \right. \\ &\quad \left. + E(\epsilon_1^{2r}) \max_{1 \leq i_1 \leq i-1} (w_{ii_1} + w_{i_1i})^{2r} \right]. \end{aligned}$$

Hence, since $E(\epsilon_1^{2r}) \leq C^{2r}$ for a constant $C > 0$,

$$\begin{aligned}
E \left[\left\{ \sum_{i=1}^n E(Z_i^2 \mid \mathcal{F}_{i-1}) \right\}^r \mid \mathcal{X} \right] &\leq \left\{ \sum_{i=1}^n \left(E \left[\{ E(Z_i^2 \mid \mathcal{F}_{i-1}) \}^r \mid \mathcal{X} \right] \right)^{1/r} \right\}^r \\
&\leq \left\{ \frac{A r \sigma}{\log(2r)} \right\}^{2r} \left\{ \sum_{i=1}^n \left[\sigma^2 \sum_{i_1=1}^{i-1} (w_{ii_1} + w_{i_1 i})^2 \right. \right. \\
&\quad \left. \left. + (E\epsilon_1^{2r})^{1/r} \max_{1 \leq i_1 \leq i-1} (w_{ii_1} + w_{i_1 i})^2 \right] \right\}^r \\
&\leq \left\{ \frac{C_2 r \sigma}{\log(2r)} \right\}^{2r} \left(\sum_{1 \leq i_1 \neq i_2 \leq n} w_{i_1 i_2}^2 \right)^r.
\end{aligned} \tag{C.8}$$

Similarly but more simply, for each $i = 2, \dots, n$ the following properties hold:

$$\begin{aligned}
E(Z_i^{2r} \mid \mathcal{X}) &= E(\epsilon_1^{2r}) E \left[\left\{ \sum_{i_1=1}^{i-1} (w_{ii_1} + w_{i_1 i}) \epsilon_{i_1} \right\}^{2r} \mid \mathcal{X} \right] \\
&\leq E(\epsilon_1^{2r}) \left\{ \frac{A r}{\log(2r)} \right\}^{2r} \left[\left\{ \sigma^2 \sum_{i_1=1}^{i-1} (w_{ii_1} + w_{i_1 i})^2 \right\}^r \right. \\
&\quad \left. + E(\epsilon_1^{2r}) \max_{1 \leq i \leq n} (w_{ii_1} + w_{i_1 i})^{2r} \right].
\end{aligned} \tag{C.9}$$

Combining (C.7)–(C.9), and observing that $E(\epsilon_1^{2r}) \leq C_2^{2r}$, we deduce that

$$E(T_2^{2r} \mid \mathcal{X}) \leq \left\{ \frac{C_4 r}{\log(2r)} \right\}^{4r} \left(\frac{1}{n^2} \sum_{1 \leq i_1 \neq i_2 \leq n} w_{i_1 i_2}^2 \right)^r. \tag{C.10}$$

Noting the definition of $w_{i_1 i}$ at (C.6), and defining $v_{i_1 i}$ to be the $(p+1)$ -vector of which the j th component is $\{(X_i - X_{i_1})/h\}^j K\{(X_i - X_{i_1})/h\}$ for $j = 0, \dots, p$, we deduce that

$$|(n-1)h w_{i_1 i}| = \left| \sum_{j=0}^p \widehat{M}_{0j,-i}^{(-1)}(X_i) \left(\frac{X_i - X_{i_1}}{h} \right)^j K \left(\frac{X_i - X_{i_1}}{h} \right) \right| \leq \widehat{\text{ev}}_{p,-i}^{-1} \|v_{i_1 i}\|,$$

where $\widehat{\text{ev}}_{p,-i}$ is the smallest eigenvalue of $\widehat{M}_{-i}$. Since the support of K equals the interval $[-1, 1]$ then the j th component of $v_{i_1 i}$ is dominated by $K\{(X_i - X_{i_1})/h\}$, for $0 \leq j \leq p$,

and so

$$\sum_{1 \leq i_1 \neq i_2 \leq n} \|v_{i_1 i}\|^2 \leq (p+1) \sum_{1 \leq i_1 \neq i_2 \leq n} K\left(\frac{X_i - X_{i_1}}{h}\right)^2.$$

Therefore, defining

$$\tilde{\text{ev}}_p = \min_{1 \leq i \leq n} \hat{\text{ev}}_{p,-i},$$

we deduce from (C.10) that

$$\begin{aligned} E(T_2^{2r} \mid \mathcal{X}) &\leq \left\{ \frac{C_5 r^4 p}{n^2 h \tilde{\text{ev}}_p^2 (\log 2r)^4} \right\}^r \left\{ \frac{1}{n^2 h} \sum_{1 \leq i_1 \neq i_2 \leq n} K\left(\frac{X_i - X_{i_1}}{h}\right) \right\}^r \\ &\leq \left\{ \frac{C_5 r^4 p}{n^2 h \tilde{\text{ev}}_p^2 (\log 2r)^4} \right\}^r \left\{ \sup_{x \in \mathcal{I}} \hat{f}(x) \right\}^r, \end{aligned} \quad (\text{C.11})$$

where

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \quad (\text{C.12})$$

is a conventional estimator of the density f_X of the design points X_i .

Step 3: Bounds for $\tilde{g} - \tilde{g}_{-i}$ and $\hat{g} - \hat{g}_{-i}$. Put

$$U_j(x) = \frac{1}{nh} \sum_{i=1}^n g(X_i) \left(\frac{x - X_i}{h}\right)^j K\left(\frac{x - X_i}{h}\right), \quad (\text{C.13})$$

$$U_{j,-i}(x) = \frac{1}{(n-1)h} \sum_{i_1: i_1 \neq i} g(X_{i_1}) \left(\frac{x - X_{i_1}}{h}\right)^j K\left(\frac{x - X_{i_1}}{h}\right). \quad (\text{C.14})$$

Analogously to the definition of $\tilde{g}_{-i}$ at (C.3), define $\tilde{g}$ by

$$\tilde{g} = E(\hat{g} \mid \mathcal{X}) = (\widehat{M}^{-1} U)_0, \quad (\text{C.15})$$

where (compare with (2.5) and (2.6)) $U = (U_0, \dots, U_p)^T$. In this step we shall derive a bound, given at (C.23), for $\tilde{g} - \tilde{g}_{-i}$, and a bound, at (C.24), for $\hat{g} - \hat{g}_{-i}$.

Observe first that

$$|U_j - U_{j,-i}| \leq \frac{|U_j|}{n-1} + \frac{(\sup |g|)(\sup K)}{(n-1)h} \leq \left(\frac{\sup \hat{f}_X}{n-1} + \frac{\sup K}{(n-1)h} \right) \sup |g|, \quad (\text{C.16})$$

where $\hat{f}_X$ is as at (C.12). Recall that $\hat{m}_{jk}$, at (2.4), is the (j, k) th component of $\widehat{M}$, let $\hat{m}_{jk,-i}$ be the (j, k) th component of $\widehat{M}_{-i}$, and note that, for all $j, k \geq 0$,

$$|\hat{m}_{jk} - \hat{m}_{jk,-i}| \leq \frac{\hat{m}_{jk}}{n-1} + \frac{\sup K}{(n-1)h} \leq \frac{\sup \hat{f}_X}{n-1} + \frac{\sup K}{(n-1)h}. \quad (\text{C.17})$$

Here we have used the fact that, since K is supported on $[-1, 1]$, $0 \leq \hat{m}_{jk} \leq \hat{f}_X$.

Let $\widehat{\text{ev}}(x)$ denote the smallest eigenvalue of $\widehat{M}(x)$. If

$$\delta = \delta(h, n) \equiv \frac{\sup \hat{f}_X}{n-1} + \frac{\sup K}{(n-1)h} \leq \frac{\widehat{\text{ev}}_p}{2p} \quad (\text{C.18})$$

then, by (C.17),

$$\begin{aligned} \|\widehat{M}_{-i} - \widehat{M}\| \widehat{M}^{-1} &\leq \|\widehat{M}_{-i} - \widehat{M}\| \|\widehat{M}^{-1}\| \\ &\leq \left\{ \sum_{j=0}^p \sum_{k=0}^p (\hat{m}_{jk} - \hat{m}_{jk,-i})^2 \right\}^{1/2} \widehat{\text{ev}}_p^{-1} \leq \frac{(p+1)\delta}{\widehat{\text{ev}}_p} \leq \frac{1}{2}. \end{aligned} \quad (\text{C.19})$$

Writing $D_i = (\widehat{M}_{-i} - \widehat{M}) \widehat{M}^{-1}$, a $(p+1) \times (p+1)$ matrix, we have if (C.18) holds,

$$\widehat{M}^{-1} U - \widehat{M}_{-i}^{-1} U_{-i} = \widehat{M}^{-1} \{I - (I + D_i)^{-1}\} U - \widehat{M}^{-1} (I + D_i)^{-1} (U_{-i} - U),$$

and therefore,

$$\begin{aligned} \|\widehat{M}^{-1} U - \widehat{M}_{-i}^{-1} U_{-i}\| &\leq \|\widehat{M}^{-1}\| \left\{ \|D_i (I + D_i)^{-1}\| \|U\| + \|(I + D_i)^{-1}\| \|U_{-i} - U\| \right\} \\ &\leq \frac{2}{\widehat{\text{ev}}_p} (\|D_i\| \|U\| + \|U_{-i} - U\|) \\ &\leq \frac{2(p+1)^{1/2} \delta}{\widehat{\text{ev}}_p} \left\{ \frac{(p+1) \sup \hat{f}_X}{\widehat{\text{ev}}_p} + 1 \right\} \sup |g|. \end{aligned} \quad (\text{C.20})$$

Here we have used the fact that $|U_j| \leq \hat{f}_X \sup |g|$, whence it follows that $\|U\| \leq (p+1)^{1/2} \hat{f}_X \sup |g|$; that by (C.19), $\|D_i\| \leq (p+1) \delta / \widehat{\text{ev}}_p$; and that by (C.16),

$$\|U_{-i} - U\| = \left\{ \sum_{j=0}^p (U_j - U_{j,-i})^2 \right\}^{1/2} \leq (p+1)^{1/2} \delta \sup |g|.$$

Noting the definitions of $\tilde{g}$ and $\tilde{g}_{-i}$ at (C.15) and (C.3), we deduce that

$$\tilde{g} - \tilde{g}_{-i} = (\widehat{M}^{-1} U - \widehat{M}_{-i}^{-1} U_{-i})_0, \quad (\text{C.21})$$

where $U_{-i} = (U_{0,-i}, \dots, U_{p,-i})^T$. Provided that (C.18) holds, it follows from (C.20) that

$$\begin{aligned} |(\widehat{M}^{-1} U - \widehat{M}_{-i}^{-1} U_{-i})_0| &\leq \|\widehat{M}^{-1} U - \widehat{M}_{-i}^{-1} U_{-i}\| \\ &\leq \frac{2(p+1)^{1/2} \delta}{\widehat{\text{ev}}_p} \left\{ \frac{(p+1) \sup \hat{f}_X}{\widehat{\text{ev}}_p} + 1 \right\} \sup |g|. \end{aligned} \quad (\text{C.22})$$

Together, (C.21) and (C.22) imply the following bound for $\tilde{g} - \tilde{g}_{-i}$: If (C.18) holds then

$$\sup_{x \in \mathcal{I}} |\tilde{g}(x) - \tilde{g}_{-i}(x)| \leq \frac{2(p+1)^{1/2} \delta}{\widehat{\text{ev}}_p} \left\{ \frac{(p+1) \sup \hat{f}_X}{\widehat{\text{ev}}_p} + 1 \right\} \sup |g|. \quad (\text{C.23})$$

Next we give a bound for $\hat{g} - \hat{g}_{-i}$. The analogues in this setting of U_j and $U_{j,-i}$, defined at (C.13) and (C.14), are respectively V_j , defined at (2.3), and $V_{j,-i}$, given by

$$V_{j,-i}(x) = \frac{1}{(n-1)h} \sum_{i_1: i_1 \neq i} Y_{i_1} \left(\frac{x - X_{i_1}}{h} \right)^j K \left(\frac{x - X_{i_1}}{h} \right).$$

In this notation, and with $V = (V_0, \dots, V_p)^T$ and $V^{-i} = (V_{-i,0}, \dots, V_{-i,p})^T$, $\hat{g}$ and $\hat{g}_{-i}$ are given by

$$\hat{g} = (\widehat{M}^{-1} V)_0, \quad \hat{g}_{-i} = (\widehat{M}_{-i}^{-1} V_{-i})_0.$$

The argument leading to (C.23) can now be repeated to show that, provided that (C.18) holds,

$$\begin{aligned} & \max_{1 \leq i \leq n} \sup_{x \in \mathcal{I}} |\hat{g}(x) - \hat{g}_{-i}(x)| \\ & \leq \frac{2(p+1)^{1/2} \delta}{\widehat{\text{ev}}_p} \left\{ \frac{(p+1) \sup \hat{f}_X}{\widehat{\text{ev}}_p} + 1 \right\} (C + \sup |g|), \end{aligned} \quad (\text{C.24})$$

where C is a constant such that $P(|\epsilon_i| \leq C) = 1$.

Step 4: Bound for remainder term in (5.22). The bound is given at (C.28). Observe that

$$\begin{aligned} & \left| \text{CV}(h, p) - \left[\frac{1}{n} \sum_{i=1}^n \{\hat{g}(X_i) - g(X_i)\}^2 + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \right] \right| \\ & = \left| \frac{1}{n} \sum_{i=1}^n \left[\{\hat{g}(X_i) - \hat{g}_{-i}(X_i)\}^2 + 2\{g(X_i) - \hat{g}(X_i)\} \{\hat{g}(X_i) - \hat{g}_{-i}(X_i)\} \right] \right. \\ & \quad \left. + 2(T_1 - T_2) \right| \\ & \leq \left\{ \max_{1 \leq i \leq n} \sup_{x \in \mathcal{I}} |\hat{g}(x) - \hat{g}_{-i}(x)| \right\}^2 \\ & \quad + 2 \left[\frac{1}{n} \sum_{i=1}^n \{\hat{g}(X_i) - g(X_i)\}^2 \right]^{1/2} \max_{1 \leq i \leq n} \sup_{x \in \mathcal{I}} |\hat{g}(x) - \hat{g}_{-i}(x)| \\ & \quad + 2(|T_1| + |T_2|). \end{aligned} \quad (\text{C.25})$$

Define

$$S_1^2 = n^{-1} \sum_{i=1}^n \{\hat{g}(X_i) - g(X_i)\}^2, \quad S_2^2 = \frac{p^3}{(nh)^2 \widehat{\text{ev}}_p^4}, \quad (\text{C.26})$$

and assume that

$$\sup \hat{f}_X \leq C_6 \quad (\text{C.27})$$

for a constant $C_6 > 0$. If (C.27) holds then, noting the definition of δ at (C.18), we deduce that $\delta \leq C_7 (nh)^{-1}$ for all $n \geq 2$, where C_7 is a positive constant. Therefore, in view of (C.24) and (C.25), if $\mathcal{S}_1$ is a set of values of (h, p) such that the inequalities (C.18) and

(C.27) hold for all $(h, p) \in \mathcal{S}_1$, then

$$\sup_{(h,p) \in \mathcal{S}_1} \left| \text{CV}(h, p) - \left(S_1^2 + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \right) \right| = O_p(S_2^2 + S_1 S_2 + |T_1| + |T_2|). \quad (\text{C.28})$$

Step 5: Bound for $n^{-1} \sum_i \{g(X_i) - \tilde{g}_{-i}(X_i)\}^2$. Recall that $\tilde{g}$ and $\tilde{g}_{-i}$ are given at (C.15) and (C.3), respectively. We define

$$\begin{aligned} S_4^2 &= \frac{1}{n} \sum_{i=1}^n \{g(X_i) - \tilde{g}_{-i}(X_i)\}^2, \quad S_5^2 = \frac{1}{n} \sum_{i=1}^n \{g(X_i) - \tilde{g}(X_i)\}^2, \\ S_6^2 &= \frac{1}{n} \sum_{i=1}^n \{\tilde{g}(X_i) - \tilde{g}_{-i}(X_i)\}^2, \quad S_7^2 = \{B_7 \exp(B_3 + 1)\}^{2p} p^{2B_2 - (2p+3+s)} h^{2(p+1)}, \end{aligned}$$

and, at (C.32) below, we give a bound for S_4 . This helps to elucidate the moment bound at (C.4).

Note that

$$S_4^2 \leq 2(S_5^2 + S_6^2). \quad (\text{C.29})$$

If the inequalities at (C.18) and (C.27) hold then it follows from (C.23) that

$$S_6^2 = O_p(S_2^2). \quad (\text{C.30})$$

Let $\mathcal{E}(x)$ denote the event that $\|\widehat{M}(x) - f(x) M\| < f(x) \text{ev}_p(K)$. The argument leading to (B.6) in Appendix B implies that if the event $\mathcal{E}(x)$ obtains then

$$|g(x) - \tilde{g}(x)| \leq 2^{1/2} \left\{ f(x) \text{ev}_p(K) - \|\widehat{M}(x) - f(x) M\| \right\}^{-1} \left\{ \sum_{j=0}^p R_j(x)^2 \right\}^{1/2}.$$

Then, pursuing the argument leading to (B.32) and thence (B.34), we deduce that if the part of (B.29) pertaining to (5.14)(a) holds then

$$S_5^2 = O_p \left[p^{2B_2} \exp(2p B_3) \{(p+1)!\}^{-2} (B_7 h)^{2(p+1)} p^{-s} \right] = O_p(S_7^2). \quad (\text{C.31})$$

(In this instance the requirement at (B.24), which relates to the contribution to (B.32) and (B.34) from stochastic error terms, is not relevant.) Combining (C.29)–(C.31) we obtain:

$$S_4^2 = O_p(S_2^2 + S_7^2). \quad (\text{C.32})$$

Step 6: Bound for $|T_1| + |T_2|$. The bound is given at (C.41) below.

The arguments leading to (B.5) and (B.31) can be used to prove that

$$\widehat{\text{ev}}_p(x) \geq f(x) \text{ev}_p(K) - \left\| \widehat{M}_{-i}(x) - f(x) M \right\|,$$

$$P \left\{ \max_{1 \leq i \leq n} \left\| \widehat{M}_{-i}(x) - f_X(x) M_p \right\| > \frac{2}{3} f_X(x) \text{ev}_p(K) \text{ for some } x \in \mathcal{I} \right\} = O(n^{-C}),$$

for all $C > 0$. Hence,

$$\widehat{\text{ev}}_p^{-1} + \widehat{\text{ev}}^{-1} = O_p(\text{ev}_p^{-1}) = O_p\{p^{B_2} \exp(B_3 p)\}, \quad (\text{C.33})$$

uniformly in h and p satisfying the parts of (B.24) and (B.29) pertaining to (5.14)(a). (The bound in (C.33) for $\widehat{\text{ev}}^{-1}$ follows from (B.5) and (B.31).)

Observe that

$$E(T_1^{2r} \mid \mathcal{X}) \leq \left\{ \frac{C_8 r}{n^{1/2} \log(2r)} \right\}^{2r} \left\{ \frac{p^3}{(nh)^2 \widehat{\text{ev}}_p^4} + \phi_1(h, p) \right\}^r, \quad (\text{C.34})$$

$$E(T_2^{2r} \mid \mathcal{X}) \leq \left\{ \frac{C_8 r^4 p}{n^2 h \widehat{\text{ev}}_p^2 (\log 2r)^4} \right\}^r, \quad (\text{C.35})$$

where

$$\phi_1(h, p) = \{B_7 \exp(B_3 + 1)\}^{2p} p^{2B_2 - (2p+3+s)} h^{2(p+1)},$$

we used (C.4) and (C.32) to derive (C.34), and (C.11) to obtain (C.35), and (C.34) and (C.35) hold uniformly in values of h such that (C.27) holds. Hence, by (C.33),

$$E(T_1^{2r} \mid \mathcal{X}) = O_p \left(\left\{ \frac{C_9 r}{n^{1/2} \log(2r)} \right\}^{2r} \times \left[p^3 (nh)^{-2} \{p^{B_2} \exp(B_3 p)\}^4 + \phi_1(h, p) \right]^r \right), \quad (\text{C.36})$$

$$E(T_2^{2r} \mid \mathcal{X}) = O_p \left(\left[\frac{C_9 r^4 p \{p^{B_2} \exp(B_3 p)\}^2}{n^2 h (\log 2r)^4} \right]^r \right), \quad (\text{C.37})$$

hold uniformly in $r \geq 1$ and in h and p such that (B.24)(a), (B.29)(a) and (C.27) hold. That is, if we write (C.36) and (C.37) as $\text{LHS} = O(\text{RHS})$ then the supremum of LHS/RHS , over $r \geq 1$ and h and p such that (B.24)(a), (B.29)(a) and (C.27) hold, equals $O_p(1)$. Note, however, that in view of (5.20)(b), the constants B_2 and B_3 in (B.24)(a) equal $-\frac{1}{2}$ and $1+2^{1/2}$, respectively. Therefore, (B.24)(a) and (B.29)(a) follow from the left- and right-hand inequalities in (5.21), respectively. Moreover, it can be proved from (5.16) and (5.20)(b) that

$$P \left\{ \sup_{h: (h,p) \in \mathcal{S}_0} \sup_{x \in \mathcal{I}} \hat{f}_X(x) \leq C_9 \right\} = 1 - O(n^{-C}) \quad (\text{C.38})$$

for all $C > 0$. Property (C.38) implies that the probability that (C.27) holds converges to 1 as $n \rightarrow \infty$. Therefore,

$$\sup_{(h,p) \in \mathcal{S}_0} \frac{\text{LHS}}{\text{RHS}} = O_p(1). \quad (\text{C.39})$$

Let $a > 0$ and take $r = r(n)$, in (C.36) and (C.37), to be the smallest integer not less than $t = t(n) = a(\log n)/(\log \log n)$. For this r , define $a_j = a_j(h, n, p)$ by

$$a_j = \begin{cases} \left\{ C_9 r / n^{1/2} \log(2r) \right\} \left[p^3 (nh)^{-2} \{p^{B_2} \exp(B_3 p)\}^4 + \phi_1(h, p) \right]^{1/2} & \text{if } j = 1 \\ \left[C_9 r^4 p \{p^{B_2} \exp(B_3 p)\}^2 / n^2 h (\log 2r)^4 \right]^{1/2} & \text{if } j = 2. \end{cases}$$

Then (C.39) implies that, for each $\eta_1 > 0$,

$$P\{|T_j| > (\log n)^{\eta_1} a_j \mid \mathcal{X}\} \leq \{(\log n)^{\eta_1} a_j\}^{-2r} E(T_1^{2r} \mid \mathcal{X}) = O_p\{(\log n)^{-2\eta_1 r}\},$$

uniformly in $(h, p) \in \mathcal{S}_0$. Now, $(\log n)^{-2\eta_1 r} \leq (\log n)^{-2\eta_1 t} = n^{-2\eta_1}$. Therefore, if $\mathcal{S}_2 \subseteq \mathcal{S}_0$ satisfies $\#\mathcal{S}_2 = o(n^{2\eta_1})$, then

$$P\left\{\sup_{(h,p) \in \mathcal{S}_2} |T_j| > (\log n)^{\eta_1} a_j \mid \mathcal{X}\right\} = o_p(1). \quad (\text{C.40})$$

Since a can be any fixed positive number, and the kernel K is Hölder continuous, then, if a is chosen sufficiently large, standard lattice arguments allow (C.40) to be extended from $\mathcal{S}_2$ to the set

$$\mathcal{S}_3 = \{(h, p) \in \mathcal{S}_0 : n^{\eta-1} \leq h \leq 1 \text{ and } 1 \leq p \leq \log n\},$$

for any fixed $\eta > 0$. However, it follows from the definition of $\mathcal{S}_0$, immediately above Theorem 5.2, that $\mathcal{S}_3 = \mathcal{S}_0$. Hence, for any fixed $\eta_1 > 0$, and with

$$\begin{aligned} \phi_2(p) &= \{B_7 \exp(B_3 + 1)\}^p p^{\{2B_2 - (2p+3+s)\}/2}, \\ b_j &= \begin{cases} C_{10} (\log n) (\log \log n)^{-2} \\ \quad \times \left[(n^{3/2}h)^{-1} p^{2B_2+(3/2)} \exp(2B_3 p) + \phi_2(p) n^{-1/2} h^{p+1} \right] & \text{if } j = 1 \\ (nh^{1/2})^{-1} C_{10} (\log n)^2 p^{B_2+(1/2)} \exp(B_3 p) / (\log \log n)^4 & \text{if } j = 2, \end{cases} \end{aligned}$$

where $C_{10} > 0$ depends on a , we have:

$$P\left\{\sup_{(h,p) \in \mathcal{S}_0} |T_j| > (\log n)^{\eta_1} b_j \mid \mathcal{X}\right\} = o_p(1).$$

Therefore,

$$\sup_{(h,p) \in \mathcal{S}_0} |T_j| = O_p\{(\log n)^{\eta_1} b_j\}. \quad (\text{C.41})$$

Step 7: Conclusion. Recall from just above (C.28) that $\mathcal{S}_1$ is any set of pairs (h, p) such that (C.18) and (C.27) hold for all $(h, p) \in \mathcal{S}_1$. Put $\mathcal{S}_4 = \mathcal{S}_1 \cap \mathcal{S}_0$. Noting the definition of

S_2 in (C.26), and combining (C.28), (C.33) and (C.41), we deduce that for each $\eta_1 > 0$,

$$\begin{aligned} \sup_{(h,p) \in \mathcal{S}_4} \left| \text{CV}(h,p) - \left(S_1^2 + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \right) \right| \\ = O_p \left\{ \phi_3(h,p) + S_1 \phi_3(h,p)^{1/2} + (\log n)^{\eta_1} (b_1 + b_2) \right\}, \end{aligned} \quad (\text{C.42})$$

where $\phi_3(h,p) = (nh)^{-2} p^{3+4B_2} \exp(4 B_3 p)$.

Note from (5.9) (valid since (5.20)(b) holds), and (B.5) and (B.31) in Appendix B, that

$$\widehat{\text{ev}}_p(x) \geq f(x) \text{ev}_p(K) - \left\| \widehat{M}(x) - f(x) M \right\| \geq C_{11} p^{1/2} \exp \left\{ - (1 + 2^{1/2}) p \right\},$$

where the first inequality holds with probability 1 if $x \in \mathcal{I}$, and the second inequality holds with probability $1 - O(n^{-C})$, uniformly in $x \in \mathcal{I}$, for all $C > 0$. From this result and (C.38) we see that the probability that both (C.18) and (C.27) hold, the latter for some $C_6 > 0$, converges to 1 at rate $O(n^{-C})$, for all $C > 0$, as $n \rightarrow \infty$. Therefore the probability that $\mathcal{S}_4 = \mathcal{S}_0$ converges to 1, and so we can replace $\mathcal{S}_4$ by $\mathcal{S}_0$ in (C.42).

It follows from the definition of $\mathcal{S}_0$ that, for all $(h,p) \in \mathcal{S}_0$,

$$\begin{aligned} \text{(a) } n^{\eta_5-2/3} \leq h \leq 1, \text{ and (b) } p \text{ is bounded above by a constant multiple} \\ \text{of } (\log n)^{1-\eta_4}. \end{aligned} \quad (\text{C.43})$$

Property (C.43)(a) implies that $(nh)^{-2} \leq (nh^{1/2})^{-1} n^{-\eta_5}$, where $\eta_6 = 3\eta_5/2$, and from (C.43)(b) that $\exp(Cp) = o(n^\eta)$ for all $C, \eta > 0$. Therefore, (C.43) entails $\phi_3(h,p) = O\{(nh)^{-2} n^\eta\}$ for all $\eta > 0$, whence it follows that $\phi_3(h,p) = O\{(nh^{1/2})^{-1} n^{-\eta_6}\}$ for each $\eta_6 < \eta_5$. Hence, using again (C.43)(b), noting the definitions of b_1 and b_2 , and defining $\eta_7 = \eta_6/2$, we deduce that

$$\begin{aligned} \phi_3(h,p) + S_1 \phi_3(h,p)^{1/2} + (\log n)^{\eta_1} (b_1 + b_2) \\ \leq C_{12} \left[(nh^{1/2})^{-1} (\log n)^{2+\eta_1} p^{B_2+(1/2)} \exp(B_3 p) + S_1 (nh^{1/2})^{-1/2} n^{-\eta_7} \right. \\ \left. + n^{-1/2} h^{p+1} (\log n)^{1+\eta_1} \{B_7 \exp(B_3 + 1)\}^p p^{\{2B_2-(2p+3+s)\}/2} \right]. \end{aligned} \quad (\text{C.44})$$

However, by assumption,

$$h^{1/2} (\log n)^{2+\eta_1} p^{B_2+(1/2)} \exp(B_3 p) = o(1), \quad (\text{C.45})$$

$$h^{2/3} (\log n)^{1+\eta_1} \{B_7 \exp(B_3 + 1)\}^p p^{\{2B_2-(2p+3+s)\}/2} = o(1). \quad (\text{C.46})$$

Note too that for each $a > 0$, $2n^{-1/2} h^{p+1} \leq (nh^a)^{-1} + h^{2(p+a+1)}$. Taking $a = \frac{1}{3}$ in this formula we see that the property

$$n^{-1/2} h^{p+1} (\log n)^{1+\eta_1} \{B_7 \exp(B_3 + 1)\}^p p^{\{2B_2-(2p+3+s)\}/2} = o\{(nh)^{-1} + h^{2(p+1)}\} \quad (\text{C.47})$$

follows from (C.46). More simply, the property

$$(nh^{1/2})^{-1} (\log n)^{2+\eta_1} p^{B_2+(1/2)} \exp(B_3 p) \rightarrow 0 \quad (\text{C.48})$$

follows from (C.45).

Combining (C.44), (C.47) and (C.48) we deduce that

$$\begin{aligned} \phi_3(h, p) + S_1 \phi_3(h, p)^{1/2} + (\log n)^{\eta_1} (b_1 + b_2) \\ = o_p\{(nh)^{-1} + h^{2(p+1)}\} + O_p\left\{S_1 (nh^{1/2})^{-1/2} n^{-\eta_7}\right\}. \end{aligned} \quad (\text{C.49})$$

Combining (C.42), with $\mathcal{S}_0$ replacing $\mathcal{S}_4$, and (C.49) we deduce that, uniformly in $(h, p) \in \mathcal{S}_0$,

$$\begin{aligned} \text{CV}(h, p) - \left(S_1^2 + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2\right) \\ = o_p\{(nh)^{-1} + h^{2(p+1)}\} + O_p\left\{S_1 (nh^{1/2})^{-1/2} n^{-\eta_7}\right\}. \end{aligned} \quad (\text{C.50})$$

Now, $S_1 (nh^{1/2})^{-1/2} n^{-\eta_7} = o_p\{S_1^2 h^{1/2} n^{-\eta_8} + (nh)^{-1}\}$, for some $\eta_8 > 0$, uniformly in $(h, p) \in \mathcal{S}_0$. Result (5.22) Theorem 5.2 follows from this result and (C.50).

JEFFREY S. RACINE: DEPARTMENT OF ECONOMICS, GRADUATE PROGRAM IN STATISTICS, MCMASTER
UNIVERSITY, HAMILTON, CANADA