Calibrating the Degrees of Freedom for Automatic Data Smoothing and Effective Curve Checking

Chunming ZHANG

Curve fitting and curve checking based on the local polynomial regression technique are commonly used data-analytic methods in statistics. This article examines, in nonparametric settings, both the asymptotic expressions and empirical formulas for degrees of freedom (DF), a notion introduced by Hastie and Tibshirani, of linear smoothers. The asymptotic results give useful insights into the nonparametric modeling complexity. Meanwhile, by substituting the exact DFs by the empirical formula, an empirical version of the generalized cross-validation (EGCV) is obtained. An automatic bandwidth selection method based on minimizing EGCV is proposed for conducting local smoothing. This procedure preserves full benefits of the ordinary and generalized cross-validation, but offers a substantial reduction in computational burden. Furthermore, the EGCV-minimizing bandwidth can be extended in a very simple manner to fit multivariate models, such as the varying-coefficient models. Applications of calibrating DFs to important inferential issues, such as assessing the validity of useful model assumptions and measuring the significance of predictor variables based on the generalized likelihood ratio statistics are also discussed. Simulation studies are presented to illustrate the performance of the proposed procedures in a range of statistical problems.

KEY WORDS: Bandwidth selection; Cross-validation; Goodness of fit; Local polynomial regression; Varying coefficient model.

1. INTRODUCTION

Curve fitting and curve checking, based on the scatterplot smoothing, are commonly used data-analytic methods in statistics. Their utilities and potential areas of applications for a wide variety of smoothing techniques have been summarized by Eubank (1988), Wahba (1990), Hastie and Tibshirani (1990), Green and Silverman (1994), Wand and Jones (1995), Fan and Gijbels (1996), and others. An important practical problem is choosing the appropriate amount of smoothing when carrying out data smoothing and significance assessment. In principle, one seeks the smoothing parameter that trades-off the bias and variance of the resulting estimator, leading to the optimal (global) smoothing parameter that minimizes criteria such as the mean integrated squared error (MISE). In this article, the discussion is confined mainly to the local polynomial regression technique, the smoothing parameter of which is referred to as *bandwidth*.

A number of data-based methods have been developed for the automatic choice of bandwidth. The most frequently used procedure for bandwidth selection is cross-validation (CV) (Allen 1974; Stone 1974). The asymptotic equivalence of bandwidth selectors based on CV and some other criteria was briefly discussed by Rice (1984). Theoretically, the CV-minimizing bandwidth converges to the true MISE-optimal bandwidth. (See Härdle, Hall, and Marron 1992 for CV bandwidth selector applied to kernel regression.) To further improve the convergence rate of CV-based bandwidth selectors, various alternative methods based on the plug-in idea have been proposed. Gasser, Kneip, and Köhler (1991) and Ruppert, Sheather, and Wand (1995) developed plug-in bandwidth selection methods whose motivations are based on asymptotic theory (the former was developed for a different nonparametric estimator, but applies also to local linear regression), whereas the plug-in method of Fan and Gijbels (1995) relies on nonasymptotic expressions. Although refinements based on the plug-in approaches are elegant, the choice of some other extraneous parameters is required in addition to the bandwidth parameter itself; moreover, implementation of plug-in methods need special programming efforts. Ruppert, Wand, Holst, and Hössjer (1997) also pointed out that plug-in bandwidths, such as those of Ruppert et al. (1995) and Fan and Gijbels (1995), restrict to odd-degree local polynomials, because the bias expression of even-degree local fitting is more complex. In contrast, the CV method can be implemented with relatively minimal effort and is easily applicable to both odd and even degrees.

In this article, the proposed bandwidth selector applies a variant of CV, based on the notion of generalized cross-validation (GCV) criterion and the calibrating formulas of *degrees of freedom* (DFs). The idea of GCV appeared originally in the context of smoothing splines (see Craven and Wahba 1979 and references cited therein). An in-depth discussion of the theoretical properties of GCV has been given by for example, Li (1985, 1986). Operationally, CV requires evaluation of all diagonal entries of an associated smoother matrix, whereas GCV relaxes this need, instead computing only the trace of that matrix. Compared with CV, GCV improves computational speed. However, the development herein is distinct from ordinary CV and GCV methods in the following aspects:

1. Asymptotic expressions are derived for the matrix traces. Then some empirical formulas for the traces, or DFs, are proposed. Substituting the actual traces by the closed-form formulas leads to the empirical GCV (EGCV). The bandwidth selector chooses the bandwidth as the minimizer of the EGCV function. In the simulation study, these empirical formulas in the random design perform well for sample sizes around 400 and even better for larger sizes. (In the fixed design, a sample size around 200 or smaller could be fine.) Indeed, as demonstrated by Table 4 in random design, the variability of those traces decreases rapidly as sample size grows. Although direct computation of either CV or GCV is not a major concern for a dataset of small size (say 50, as used in Lee and Solo 1999) or moderate size, it can potentially become a problem

Chunming Zhang is Assistant Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: *cmzhang@stat.wisc.edu*). The author thanks the associate editor and two anonymous referees for constructive suggestions and comments on earlier versions of this article, and Kam-Wah Tsui for helpful discussions.

Journal of the American Statistical Association
September 2003, Vol. 98, No. 463, Theory and Methods
DOI 10.1198/016214503000000521

for the large and huge sample sizes common in data-mining tasks. Typical examples include processing massive financial data (Stanton 1997), functional data (Ramsay and Silverman 1997), and longitudinal data (Müller 1988). Nonetheless, with the derived empirical formulas for DFs, the full benefits of CV and GCV can be retained while greatly reducing the computational burden.

2. Although the principle of (G)CV is generally applicable, most of the work related to bandwidth selection is concentrated on the canonical nonparametric regression model with a univariate predictor variable, for which a single smooth curve must be estimated. For fitting a useful class of multivariate models, such as varying-coefficient models (Hastie and Tibshirani 1993) with multiple low- or one-dimensional smooth curves, bandwidth selection based on EGCV continues to be extensible in a very simple manner, whereas conventional (G)CV demands greater amount of computation as the number of covariates increases; see Section 3. Similarly, extensions of plug-in bandwidth selectors to multivariate smoothing, though theoretically feasible (Müller and Prewitt 1993), will encounter more inconvenience when put into practical application.

3. Many important inferential issues need to be addressed after applying nonparametric smoothing methods. Various hypothesis testing problems are of interest, including checking the suitability of some parametric/nonparametric models versus the nonparametric alternatives. In particular, practical aspects of problems arising from model validation would call for substantial developments. This article reports on methodologies based on the generalized likelihood ratio statistic (GLR) proposed by Fan, Zhang, and Zhang (2001). With the calibration formulas for DFs of local polynomial smoother, one can directly conduct the proposed GLR test by comparing with the null distribution percentiles of some familiar reference distributions. This procedure works well for large datasets. Meanwhile, a bootstrap procedure is proposed to estimate the null distribution of the test statistic with small sample sizes. The efficacies of both procedures are examined thoroughly in the simulation studies of Section 5.4. Thus calibrating DFs helps statistical modelers achieve two goals simultaneously: automatic data smoothing and effective curve checking.

In this article, calibrating DF is concentrated on commonly used linear smoothers. A generalized notion of "degrees of freedom" has been given by Ye (1998), which addresses both linear and nonlinear smoothers. Following his arguments, it is anticipated that calibrating DFs will have domains of applications broader than those discussed here.

The article is organized as follows. Section 2 begins with the DFs and EGCV in the nonparametric regression model. Section 3 addresses DFs of local regression for fitting varying-coefficient models. Section 4 discusses applications of DFs to model assessment. Section 5 presents simulation evaluation of the empirical formulas for DFs with applications to curve fitting and model assessment. Section 6 provides some further discussions on the proposed method and points to some possible extensions and improvements. The Appendix provides technical conditions and proofs.

2. NONPARAMETRIC REGRESSION MODEL

2.1 Degrees of Freedom of Linear Smoothers and Empirical Generalized Cross-Validation

The DF of local polynomial regression estimates considered in this article can be defined for general linear smoothers, a notion introduced by Hastie and Tibshirani (1990, sec. 3.5). Consider the situation where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a sample of random pairs described by the nonparametric regression model

$$Y = m(X) + \varepsilon, \tag{1}$$

where ε represents the background noise with $E(\varepsilon|X) = 0$ and $\text{var}(\varepsilon|X) = \sigma^2$, X is a scalar regressor variable with a sampling density f with a known compact support Ω, and $m(x)$ is some unknown response function of interest. Call $\hat{m}(\cdot)$ a linear estimator if there is a square matrix $\mathbf{S}$, independent of all Y_i, $i = 1, \ldots, n$, that transforms the vector of responses, $\mathbf{y} = (Y_1, \ldots, Y_n)^T$, to the vector of fitted values, $\hat{\mathbf{m}} = (\hat{m}(X_1), \ldots, \hat{m}(X_n))^T$, according to

$$\hat{\mathbf{m}} = \mathbf{S}\mathbf{y}. \tag{2}$$

Call $\mathbf{S}$ the smoother matrix. Examples of linear smoothers include smoothing splines, regression splines (see, e.g., Hastie and Tibshirani 1990), and wavelet estimators. Local polynomial regression estimation is also linear, with extreme dependence on the bandwidth h [see (9) in Sec. 2.2]. For later use, denote the resulting mean function estimate by $\hat{m}_h(x)$ and the smoother matrix by $\mathbf{S}_h$. The theoretically optimal constant bandwidth minimizes the conditional MISE criterion, where

$$\text{MISE}(h) = \int E[\{\hat{m}_h(x) - m(x)\}^2 | X_1, \ldots, X_n] f(x)\, dx.$$

The asymptotically optimal bandwidth that minimizes the asymptotic expression for MISE is given by

$$h_{\text{AMISE}} = \text{constant} \times \left[\frac{\sigma^2 |\Omega|}{\int_\Omega \{m^{(p+1)}(x)\}^2 f(x)\, dx} \right]^{1/(2p+3)} n^{-1/(2p+3)}, \tag{3}$$

where p is the degree of the local polynomial and $|\Omega|$ measures the length of Ω (see Fan and Gijbels 1996, pp. 67–68, for details). Clearly, this formula contains several unknown quantities and cannot readily serve for the purpose of automatic data smoothing.

Perhaps the most well-known data-driven bandwidth selection method is CV (see Wong 1983; Rice 1984; Fan and Gijbels 1996, sec. 4.10.2). This method selects the bandwidth that minimizes the CV score, defined as

$$\text{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \hat{m}_{h,-i}(X_i)\}^2, \tag{4}$$

where $\hat{m}_{h,-i}(X_i)$ stands for the usual "leave-one-out" estimate of $m(X_i)$ obtained by removing the ith observation pair (X_i, Y_i). In spline smoothing, an alternative expression of CV is widely used. To show that this sort of simplification works for the local

polynomial smoothers, the Appendix verifies that CV(h) is also equivalent to

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i - \hat{m}_h(X_i)\}^2}{\{1 - \mathbf{S}_h(i,\, i)\}^2}. \tag{5}$$

Evidently, the advantage of (5) relative to (4) allows significant savings in computational efforts in that one can obtain all of the fitted responses based on the original one sample instead of building n distinct subsamples, each of size $n-1$. The GCV approach, proposed by Craven and Wahba (1979), replaces all of the diagonal entries $\mathbf{S}(i,i)$ by their average, $n^{-1}\sum_{i=1}^{n}\mathbf{S}(i,i) = \mathrm{tr}(\mathbf{S})/n$, where "tr" denotes the matrix trace. This idea, when applied to (5), yields the GCV function given by

$$\mathrm{GCV}(h) = \frac{n^{-1}\sum_{i=1}^{n}\{Y_i - \hat{m}_h(X_i)\}^2}{\{1 - \mathrm{tr}(\mathbf{S}_h)/n\}^2},$$

the minimum of which can be found by optimization methods or by a grid search.

This article proposes a bandwidth selector based on minimizing an empirical asymptotic version of GCV. That is, substitute $\mathrm{tr}(\mathbf{S}_h)$ by its empirical asymptotic formula. The resulting bandwidth selector is termed the "EGCV-minimizing bandwidth."

Traces of smoother matrices also naturally serve to estimate the noise variance. As in the parametric regression model, a nonparametric variance estimator of the form

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\{Y_i - \hat{m}(X_i)\}^2}{n - \mathrm{tr}(2\mathbf{S} - \mathbf{S}^T\mathbf{S})} \tag{6}$$

was considered by Buckley, Eagleson, and Silverman (1988) and Cleveland and Devlin (1988). This motivates the need to compute $\mathrm{tr}(\mathbf{S})$, $\mathrm{tr}(\mathbf{S}^T\mathbf{S})$, and $\mathrm{tr}(2\mathbf{S} - \mathbf{S}^T\mathbf{S})$. Evaluation of the last quantity is also a crucial part of the model checking process, as is described further in Section 4. Hastie and Tibshirani (1990, sec. 3.5) considered the foregoing three quantities as three definitions of DFs used in estimating m. Of these, the naive calculation of $\mathrm{tr}(\mathbf{S})$ is the easiest to carry out, at a cost of $O(n)$ operations for many of the smoothers, whereas $\mathrm{tr}(\mathbf{S}^T\mathbf{S})$ costs $O(n^2)$ operations.

2.2 Degrees of Freedom of Local Polynomial Regression Smoother

This section derives asymptotic formulas of $\mathrm{tr}(\mathbf{S}_h)$ and $\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h)$ for local polynomial smoothers. The main result is presented in Theorem 1 under both random and fixed designs.

For expositional convenience, the derivation begins by describing the local polynomial regression, of degree p. Let x_0 be an interior point of Ω, the support of f. Denote $\beta_j = m^{(j)}(x_0)/j!$, $j = 0, \ldots, p$, where the dependence of β_j's on x_0 is suppressed for notational simplicity. Then the local polynomial regression estimates $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$, of degree p, are defined to be the minimizer of the weighted sum of squared residuals

$$\sum_{i=1}^{n}\{Y_i - \beta_0 - (X_i - x_0)\beta_1 - \cdots - (X_i - x_0)^p\beta_p\}^2 \times K_h(X_i - x_0). \tag{7}$$

Here the weight function $K_h(\cdot) = K(\cdot/h)/h$ is rescaled from a nonnegative kernel $K(\cdot)$, usually taken to be a probability density function. The bandwidth $h > 0$ specifies the size of the local neighborhood and governs the amount of smoothing or local averaging. Clearly, the resulting $\hat{\beta}_0$ gives the pth degree local polynomial regression estimate; call it $\hat{m}_h(x_0)$. The kernel regression and local linear methods correspond to $p = 0$ and $p = 1$.

A more systematic study of the local polynomial smoother matrix $\mathbf{S}_h$ draws on some matrix notations (Fan and Gijbels 1996, chap. 3). Put $S_n(x_0) = \mathbf{X}(x_0)^T\mathbf{W}(x_0)\mathbf{X}(x_0)$ and $T_n(x_0) = \mathbf{X}(x_0)^T\mathbf{W}(x_0)$, where

$$\mathbf{X}(x_0) = \begin{bmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{bmatrix}$$

and $\mathbf{W}(x_0)$ is a diagonal matrix with diagonal entries $K_h(X_i - x_0)$. Then, according to (7),

$$\begin{aligned}\hat{\boldsymbol{\beta}}(x_0) &= \arg\min_{\boldsymbol{\beta}}\{\mathbf{y} - \mathbf{X}(x_0)\boldsymbol{\beta}\}^T\mathbf{W}(x_0)\{\mathbf{y} - \mathbf{X}(x_0)\boldsymbol{\beta}\} \\ &= \{S_n(x_0)\}^{-1}T_n(x_0)\mathbf{y}. \end{aligned}\tag{8}$$

This expression immediately results in

$$\begin{aligned}\hat{m}_h(x_0) &= \hat{\beta}_0 = \boldsymbol{e}_{1,\,p+1}^T\hat{\boldsymbol{\beta}}(x_0) \\ &= \sum_{j=1}^{n} W_0^n\left(x_0, \frac{X_j - x_0}{h}\right)Y_j, \end{aligned}\tag{9}$$

with

$$W_0^n(x,t) = \boldsymbol{e}_{1,\,p+1}^T\{S_n(x)\}^{-1}H(1, t, \ldots, t^p)^T K(t)/h, \tag{10}$$

defined for any real-valued x and t, where $H = \mathrm{diag}\{1, h, \ldots, h^p\}$ is a diagonal matrix. Here the vector notation $\boldsymbol{e}_{k,\,\ell}$ represents the kth column of an $\ell \times \ell$ identity matrix; later, the second subscript may be dropped wherever it is clear from the context. Replicating the foregoing estimation procedure for each of n observations X_i, all of the fitted responses $\hat{m}_h(X_i)$ can be obtained. Thus from (9), the (i,j)th entry of the smoother matrix $\mathbf{S}_h$ is represented by

$$\mathbf{S}_h(i,j) = W_0^n\left(X_i, \frac{X_j - X_i}{h}\right), \qquad i,j = 1, \ldots, n, \tag{11}$$

and the entries on the diagonal are $\mathbf{S}_h(i,i) = W_0^n(X_i, 0)$, $i = 1, \ldots, n$. Obviously, $\mathbf{S}_h$ is neither symmetric nor idempotent. Using (11), the explicit expressions for DFs are obtained:

$$\mathrm{tr}(\mathbf{S}_h) = \sum_{i=1}^{n}\boldsymbol{e}_1^T\{S_n(X_i)\}^{-1}\boldsymbol{e}_1 K(0)/h, \tag{12}$$

and

$$\begin{aligned}\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h) = \sum_{i=1}^{n}\sum_{j=1}^{n}\big[\boldsymbol{e}_1^T\{S_n(X_i)\}^{-1}\{1, (X_j - X_i), \\ \ldots, (X_j - X_i)^p\}^T\big]^2 K^2\left(\frac{X_j - X_i}{h}\right)\Big/h^2. \end{aligned}\tag{13}$$

Therefore, when the sample size n increases, naive calculations of the traces for any particular h are computationally intensive.

Conceptually, the DFs, such as $\mathrm{tr}(2\mathbf{S}_h - \mathbf{S}_h^T\mathbf{S}_h)$ in (6), should be positive to be meaningful. To verify that this desired property holds requires only checking relations between DFs. To this end, first the preliminary nonasymptotic results on $\mathbf{S}_h$ are obtained; for any $h > 0$, the inequalities

$$p + 1 \le \mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h) \le \mathrm{tr}(\mathbf{S}_h) \le \mathrm{tr}(2\mathbf{S}_h - \mathbf{S}_h^T\mathbf{S}_h) < n \tag{14}$$

hold for any nonnegative kernel K under a mode condition, $K(0) = \sup_x K(x)$, which is satisfied by virtually all symmetric and unimodal kernels used in practice; the lower bound is for DFs with $h \to \infty$, whereas the upper bound is with $h \to 0$. The proof of (14) can be found in the Appendix.

To facilitate presentations of $\mathrm{tr}(\mathbf{S}_h)$ and $\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h)$ in their asymptotic forms, now $\mathcal{K}(t)$ defines the "equivalent kernel" for the local polynomial smoother (9), namely

$$\mathcal{K}(t) = \boldsymbol{e}_1^T S^{-1}(1, t, \ldots, t^p)^T K(t), \tag{15}$$

with the matrix $S = (\mu_{i+j-2})_{1\le i,j\le p+1}$, where $\mu_\ell = \int t^\ell K(t)\,dt$ (see Fan and Gijbels 1996, p. 64; Müller 1987, p. 233). Straightforward calculations lead to the following mappings useful for presenting Theorem 1:

$$\mathcal{K}(0) = K(0)\boldsymbol{e}_1^T S^{-1}\boldsymbol{e}_1 \tag{16}$$

and

$$\mathcal{K} * \mathcal{K}(0) = \boldsymbol{e}_1^T S^{-1} S^* S^{-1}\boldsymbol{e}_1, \tag{17}$$

where $S^* = (\nu_{i+j-2})_{1\le i,j\le p+1}$, with $\nu_\ell = \int t^\ell K^2(t)\,dt$, and $*$ on the left side of (17) denotes the convolution operator. In practical applications, multiweight kernel functions of the following form are commonly used:

$$\frac{1}{\mathrm{beta}(1/2, \ell+1)}(1-t^2)^\ell I(|t| \le 1), \qquad \ell = 0, 1, \ldots.$$

Table 1 summarizes the values of $\mathcal{K}(0)$, $\mathcal{K} * \mathcal{K}(0)$, and $(2\mathcal{K} - \mathcal{K} * \mathcal{K})(0)$ for the Epanechnikov kernel ($\ell = 1$), biweight kernel ($\ell = 2$), and triweight kernel ($\ell = 3$).

Theorem 1 presents the asymptotic expressions for DFs. Here and in the sequel, $|\Omega|$ denotes the length of Ω; in the random design, the probability measure in the term $o_P(1)$ is taken with respect to the distribution of X.

Theorem 1. For random designs, assume condition (A); see the Appendix. When $n \to \infty$, $h \to 0$, and $nh \to \infty$,

$$\mathrm{tr}(\mathbf{S}_h) = \mathcal{K}(0)|\Omega|/h\{1 + o_P(1)\}, \tag{18}$$

$$\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h) = \mathcal{K} * \mathcal{K}(0)|\Omega|/h\{1 + o_P(1)\}, \tag{19}$$

and

$$\mathrm{tr}(2\mathbf{S}_h - \mathbf{S}_h^T\mathbf{S}_h) = (2\mathcal{K} - \mathcal{K} * \mathcal{K})(0)|\Omega|/h\{1 + o_P(1)\}. \tag{20}$$

For fixed designs, assume Condition (A*); see the Appendix. When $n \to \infty$, $h \to 0$, and $nh \to \infty$,

$$\mathrm{tr}(\mathbf{S}_h) = \mathcal{K}(0)|\Omega|/h\{1 + o(1)\}, \tag{21}$$

$$\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h) = \mathcal{K} * \mathcal{K}(0)|\Omega|/h\{1 + o(1)\}, \tag{22}$$

and

$$\mathrm{tr}(2\mathbf{S}_h - \mathbf{S}_h^T\mathbf{S}_h) = (2\mathcal{K} - \mathcal{K} * \mathcal{K})(0)|\Omega|/h\,\{1 + o(1)\}. \tag{23}$$

Theorem 1 demonstrates that the asymptotic DFs are inversely proportional to the bandwidth h. Fan and Gijbels (1996, pp. 7–8) gave a more graphically oriented illustration of nonparametric modeling complexity by displaying local polynomial fits with varying amounts of h, but did not assess quantitatively the extent to which h carries information of DFs. Here this linkage is made more transparent. Results in Theorem 1 also deliver the asymptotic nondependence of DFs on the design density f. In comparison, the asymptotic DFs for the smoothing spline smoother rely on the knowledge of f; see Theorem 2.

In the Appendix, the higher-order approximation formulas are given in (A.9) for $\mathrm{tr}(\mathbf{S}_h)$ and in (A.17) for $\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h)$, where the kernel-dependent constants, $\overline{\mathcal{K}}(0)$, $\underline{\mathcal{K}}(0)$, $\ell_1(K)$, and $\ell_2(K)$, are as collected in Table 1.

Table 1. Kernel-Dependent Constants From the pth Degree Local Polynomial Fit

Kernel	p	$\mathcal{K}*\mathcal{K}(0)$	$\mathcal{K}(0)$	$(2\mathcal{K}-\mathcal{K}*\mathcal{K})(0)$	$\overline{\mathcal{K}}(0)$	$\underline{\mathcal{K}}(0)$	$\ell_1(K)$	$\ell_2(K)$	$r_\mathcal{K}$
Epanechnikov	0	.6000	.7500	.9000	0	.1500	0	.1543	2.1153
	1	.6000	.7500	.9000	.1500	.1500	.1543	.1543	2.1153
	2	1.2500	1.4062	1.5625	0	.1562	0	.1569	1.9755
	3	1.2500	1.4062	1.5625	.1562	.1562	.1569	.1569	1.9755
	4	1.8930	2.0508	2.2085	0	.1578	0	.1580	1.9336
	5	1.8930	2.0508	2.2085	.1578	.1578	.1580	.1580	1.9336
Biweight	0	.7143	.9375	1.1607	0	.1339	0	.1391	2.3061
	1	.7143	.9375	1.1607	.1339	.1339	.1391	.1391	2.3061
	2	1.4073	1.6406	1.8739	0	.1491	0	.1502	2.1283
	3	1.4073	1.6406	1.8739	.1491	.1491	.1502	.1502	2.1283
	4	2.0712	2.3071	2.5431	0	.1538	0	.1542	2.0620
	5	2.0712	2.3071	2.5431	.1538	.1538	.1542	.1542	2.0620
Triweight	0	.8159	1.0938	1.3716	0	.1215	0	.1269	2.3797
	1	.8159	1.0938	1.3716	.1215	.1215	.1269	.1269	2.3797
	2	1.5549	1.8457	2.1365	0	.1420	0	.1432	2.1946
	3	1.5549	1.8457	2.1365	.1420	.1420	.1432	.1432	2.1946
	4	2.2435	2.5378	2.8322	0	.1493	0	.1498	2.1219
	5	2.2435	2.5378	2.8322	.1493	.1493	.1498	.1498	2.1219

2.3 Degrees of Freedom of the Smoothing Spline Estimator

The asymptotic expressions for DFs based on the smoothing spline are developed herein. The main result is addressed in Theorem 2. There only the fixed-design is considered for ease of technical manipulation; extensions to random designs will be an interesting future work.

The smoothing spline estimator, denoted by $\hat{m}_\lambda$, minimizes the penalized sum of squared errors,

$$n^{-1}\sum_{i=1}^{n}\{Y_i - m(x_i)\}^2 + \lambda\int_0^1 \{m^{(q)}(x)\}^2\,dx, \qquad \lambda > 0, \quad (24)$$

over all functions $m \in W_2^q[0,1]$, where $W_2^q[0,1]$, the qth order Sobolev space, is defined as

$$W_2^q[0,1] = \Big\{m : m^{(j)} \text{ is absolutely continuous for } j = 0, 1, \ldots, q-1;\ \int_0^1 \{m^{(q)}(x)\}^2\,dx < \infty\Big\},$$

for some fixed integer $q \geq 1$. The commonly used cubic smoothing spline corresponds to $q = 2$. The support of design points x_i, taken to be $\Omega = [0,1]$, is merely for simplicity. The smoothing parameter or the penalty factor λ, on which the smoothing spline estimator depends, regulates the "rate of exchange" between fidelity to the data and smoothness of the fitted curve. The smoother matrix, as a result of (24), is denoted by $\mathbf{S}_\lambda$ to stress its dependence on λ. (See Eubank 1988 and Wahba 1990 for detailed descriptions of smoothing splines.)

To derive expressions for $\mathrm{tr}(\mathbf{S}_\lambda)$ and $\mathrm{tr}(\mathbf{S}_\lambda^T\mathbf{S}_\lambda)$, an explicit representation of $\mathbf{S}_\lambda$ is needed; this can be found in Eubank (1984) among others. For convenience, we assume that the x_j's have been ordered, so that $x_1 < \cdots < x_n$. It is well known (see, e.g., Reinsch 1967) that $\hat{m}_\lambda$ belongs to $\mathcal{S}_n^q$, the n-dimensional space of natural splines,

$$\mathcal{S}_n^q = \big\{m : m \in C^{2q-2}[0,1],\ m \text{ is a polynomial of degree } 2q-1 \text{ on } [x_i, x_{i+1}],\ i = 1, \ldots, n-1, \text{ and of degree } q-1 \text{ on } [0, x_1] \text{ and } [x_n, 1]\big\}.$$

An explicit expression for $\hat{m}_\lambda(x)$ can be obtained via the basis functions $\{\phi_{jn}, j = 1, \ldots, n\}$ of $\mathcal{S}_n^q$ introduced by Demmler and Reinsch (1975). These functions satisfy the conditions

$$\frac{1}{n}\sum_{i=1}^{n}\phi_{jn}(x_i)\phi_{kn}(x_i) = \delta_{jk}$$

and

$$\int_0^1 \phi_{jn}^{(q)}(x)\phi_{kn}^{(q)}(x)\,dx = \gamma_{kn}\delta_{jk},$$

for $j, k = 1, \ldots, n$, with $0 = \gamma_{1n} = \cdots = \gamma_{qn} < \gamma_{(q+1)n} \leq \cdots \leq \gamma_{nn}$, and δ_{jk} as Kronecker's delta. Denote by $\boldsymbol{\Phi}_{jn} = (\phi_{jn}(x_1), \ldots, \phi_{jn}(x_n))^T, j = 1, \ldots, n$, the basis vectors evaluated at the design observations. Then the solution of (24) can be expressed as

$$\hat{m}_\lambda(x) = \sum_{j=1}^{n}\frac{n^{-1}\boldsymbol{\Phi}_{jn}^T\mathbf{y}}{1+\lambda\gamma_{jn}}\phi_{jn}(x).$$

The smoother matrix $\mathbf{S}_\lambda$, associated with $\hat{m}_\lambda$, allows for a spectral decomposition

$$\mathbf{S}_\lambda = \mathbf{X}\,\mathrm{diag}\{(1+\lambda\gamma_{jn})^{-1}\}_{j=1}^{n}\mathbf{X}^T, \quad (25)$$

where the square matrix $\mathbf{X} = n^{-1/2}[\boldsymbol{\Phi}_{1n}, \ldots, \boldsymbol{\Phi}_{nn}]$ is orthonormal. Clearly, $\mathbf{S}_\lambda$ is symmetric (i.e., $\mathbf{S}_\lambda^T = \mathbf{S}_\lambda$), and the DFs take the forms

$$\mathrm{tr}(\mathbf{S}_\lambda) = q + \sum_{j=q+1}^{n}(1+\lambda\gamma_{jn})^{-1} \quad (26)$$

and

$$\mathrm{tr}(\mathbf{S}_\lambda^T\mathbf{S}_\lambda) = q + \sum_{j=q+1}^{n}(1+\lambda\gamma_{jn})^{-2}, \quad (27)$$

based on the fact $\gamma_{jn} = 0$, for $j = 1, \ldots, q$. It is then apparent that

$$q \leq \mathrm{tr}(\mathbf{S}_\lambda^T\mathbf{S}_\lambda) \leq \mathrm{tr}(\mathbf{S}_\lambda) \leq \mathrm{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda^T\mathbf{S}_\lambda) < n, \quad (28)$$

as given by Hastie and Tibshirani (1990, p. 54). Compared with (14), these types of inequalities for DFs hold similarly for local polynomial smoothers.

For cubic smoothing spline fit with equally spaced design, Green and Silverman (1994, p. 36) and Hastie and Tibshirani (1990, pp. 305–306) have established some approximation formulas for DFs. Under more general fixed designs [see Condition (B)], asymptotic results on DFs are stated in Theorem 2.

Theorem 2. Let $\boldsymbol{K}(x) = (2\pi)^{-1}\int_{-\infty}^{+\infty}(1+t^{2q})^{-1}\exp(-itx)\,dt$. Set $c(f) = \int_0^1 f(t)^{1/(2q)}\,dt$, where f denotes the design density. For fixed designs, assume condition (B); see the Appendix. Then, for $q \geq 2$, as $n \to \infty$, $\lambda \to 0$, and $n\lambda \to \infty$, it holds that

$$\mathrm{tr}(\mathbf{S}_\lambda) = q + \boldsymbol{K}(0)c(f)/\lambda^{1/(2q)}\{1+o(1)\}, \quad (29)$$

$$\mathrm{tr}(\mathbf{S}_\lambda^T\mathbf{S}_\lambda) = q + \boldsymbol{K}*\boldsymbol{K}(0)c(f)/\lambda^{1/(2q)}\{1+o(1)\}, \quad (30)$$

and

$$\mathrm{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda^T\mathbf{S}_\lambda) = q + (2\boldsymbol{K} - \boldsymbol{K}*\boldsymbol{K})(0)c(f)/\lambda^{1/(2q)} \times \{1+o(1)\}. \quad (31)$$

Theorem 2 reveals the asymptotic relationships between the DFs and the smoothing parameter λ. Notice that in Theorem 2, if $\lambda \to \infty$, where the smoothing spline is actually a polynomial regression function of degree $q-1$, then the limiting DFs coincide with q, the DFs defined under the classical linear model framework. Therefore, conclusions of Theorem 2 cover situations broader than those indicated in Theorem 2.

The function $\boldsymbol{K}$ specified in Theorem 2 is known as the "equivalent kernel function" for smoothing splines (Silverman 1984). Although $\boldsymbol{K}$ itself, implicitly expressed as a Fourier transform of $1/(1+t^{2q})$, appears to be complicated, analytic formulas for $\boldsymbol{K}(0)$ and $\boldsymbol{K}*\boldsymbol{K}(0)$ are rather simple. Using Lemma A.2 (see the Appendix) and the identities

$$\int_0^\infty \frac{dy}{1+y^{2q}} = \frac{1}{2q\sin\{\pi/(2q)\}}\pi$$

and

$$\int_0^\infty \frac{dy}{(1+y^{2q})^2} = \frac{(2q-1)}{4q^2\sin\{\pi/(2q)\}}\pi,$$

Table 2. Constants of $\boldsymbol{K}(0)$, $\boldsymbol{K}*\boldsymbol{K}(0)$, and $2\boldsymbol{K}(0)-\boldsymbol{K}*\boldsymbol{K}(0)$ for the Smoothing Spline Smoother

q	$\boldsymbol{K}*\boldsymbol{K}(0)$	$\boldsymbol{K}(0)$	$(2\boldsymbol{K}-\boldsymbol{K}*\boldsymbol{K})(0)$
1	.2500	.5000	.7500
2	.2652	.3536	.4419
3	.2778	.3333	.3889
4	.2858	.3266	.3675
5	.2912	.3236	.3560
6	.2951	.3220	.3488

the following are obtained:

$$\boldsymbol{K}(0)=\frac{1}{2q\sin\{\pi/(2q)\}}$$

and (32)

$$\boldsymbol{K}*\boldsymbol{K}(0)=\frac{2q-1}{4q^2\sin\{\pi/(2q)\}}.$$

To facilitate computations, the quantities $\boldsymbol{K}(0)$, $\boldsymbol{K}*\boldsymbol{K}(0)$, and $(2\boldsymbol{K}-\boldsymbol{K}*\boldsymbol{K})(0)$, for $1\le q\le 6$, are tabulated in Table 2.

2.4 Comparison of Theorems 1 and 2

To make a more reasonable comparison of Theorem 1 and Theorem 2, consider the fixed-design points $X_i=x_i$. Recall that the smoothing spline estimator at an interior point x behaves roughly as a kernel-type method with kernel $\boldsymbol{K}$ and variable bandwidth $h(x)=\{\lambda/f(x)\}^{1/(2q)}$ (Silverman 1984). In this perspective, Theorem 2 parallels conclusions conveyed from Theorem 1.

In a common respect, both theorems indicate that DFs are asymptotically monotone decreasing in smoothing parameters. Thus, for the purpose of curve fitting, use of the DFs and use of the smoothing parameters can produce nearly the same effect. From the standpoint of smoothing, working with DFs are relatively easy to handle and interpret, because they do not rely on the configuration of the response variable.

The major distinction is that in Theorem 1, an additive term $p+1$ is not incorporated, whereas an additive term q enters into Theorem 2. As can be seen clearly, when smoothing parameters tend to infinity, both the local polynomial fit and smoothing spline become a polynomial regression function of degree p and degree $q-1$. Thus the DFs tend to $p+1$ in the former and q in the latter, agreeing with the number of model parameters in each limiting case. The apparent difference is due to the fact that Theorem 2 works directly with (26) and (27), which, whenever $\lambda>0$, encompass the quantity q. In contrast, Theorem 1, derived under the imposition $h\to 0$, is based primarily on each entry of the matrices $S_n(x_i)$, $i=1,\ldots,n$, appearing in (12) and (13). Accordingly, the asymptotics for DFs assuming $h\to 0$ may not carry over to those requiring $h\to\infty$. A similar phenomenon was noted by Stone (1984, p. 1292) in another context in the least-squares CV selection of bandwidth for multivariate kernel density estimates. Stone pointed out that "small, moderate and large values of the coordinates of h must be handled separately." Indeed, in the simplest case of kernel regression method ($p=0$), apparently $\{S_n(x_i)\}^{-1}=1/\sum_{j=1}^n K_h(x_j-x_i)$. When $h\to 0$, this quantity is asymptotic to $\{nf(x_i)\}^{-1}$, leading to $\mathrm{tr}(\mathbf{S}_h)\approx\mathcal{K}(0)|\Omega|/h$ and $\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h)\approx\mathcal{K}*\mathcal{K}(0)|\Omega|/h$, as given in Theorem 1. However, when $h\to\infty$, the same quantity tends to $\{nK(0)/h\}^{-1}$ and thus, by (12) and (13) again, it follows that $\mathrm{tr}(\mathbf{S}_h)\approx 1$ and $\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h)\approx 1$, both of which agree with $p+1$.

The foregoing discussion suggests that for the local smoothing method, the inclusion of $p+1$ in (21)–(23) has the advantage of affecting the asymptotic expressions of DFs less when $h\to 0$ [because $p+1$ has a smaller magnitude than with $O(h^{-1})$], while allowing DFs to be more interpretable and well defined even in the case $h\to\infty$; similar adjustment can be applied to (18)–(20) for random design. This consideration, inspired from DFs of spline fitting, is in turn absorbed into the empirical formulas (33)–(35) for local fitting presented in the next section.

2.5 Empirical Formulas for Degrees of Freedom

For almost all applications encountered in practice, cases of smoothing parameters approaching 0 are of primary interest when applying local modeling techniques. Of course, it is hoped that formulas for DFs will accommodate a broader range of smoothing parameters and at the same time be reasonably accurate for applications to finite-sample situations. Guided by this motivation and aided by the finite-sample lower bounds for DFs given in (14) and (28), the following empirical formulas for DFs of local polynomial fit and smoothing spline are proposed:

$$\mathrm{tr}(\mathbf{S}_h)\simeq(p+1-a)+\mathcal{C}n/(n-1)\mathcal{K}(0)|\Omega|/h, \tag{33}$$

$$\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h)\simeq(p+1-a)+\mathcal{C}n/(n-1)\mathcal{K}*\mathcal{K}(0)|\Omega|/h, \tag{34}$$

$$\mathrm{tr}(2\mathbf{S}_h-\mathbf{S}_h^T\mathbf{S}_h)\simeq(p+1-a)+\mathcal{C}n/(n-1)(2\mathcal{K}-\mathcal{K}*\mathcal{K})(0)|\Omega|/h, \tag{35}$$

$$\mathrm{tr}(\mathbf{S}_\lambda)\simeq(q-b)+\boldsymbol{K}(0)c(f)/\lambda^{1/(2q)}, \tag{36}$$

$$\mathrm{tr}(\mathbf{S}_\lambda^T\mathbf{S}_\lambda)\simeq(q-b)+\boldsymbol{K}*\boldsymbol{K}(0)c(f)/\lambda^{1/(2q)}, \tag{37}$$

and

$$\mathrm{tr}(2\mathbf{S}_\lambda-\mathbf{S}_\lambda^T\mathbf{S}_\lambda)\simeq(q-b)+(2\boldsymbol{K}-\boldsymbol{K}*\boldsymbol{K})(0)c(f)/\lambda^{1/(2q)}, \tag{38}$$

where a and b are some small scalars correcting potential sources of bias. In (33)–(35), the factor $n/(n-1)$ arise from (A.10) and (A.18) in the proof of Theorem 1. The use of a slope correction factor $\mathcal{C}\ge 1$ may alleviate the undersmoothing tendency of EGCV-minimizing bandwidth; a similar idea applied to choosing GCV-minimizing λ in spline smoothing was given by Cummins, Filloon, and Nychka (2001, sec. 2.1), where $\mathcal{C}$ was put directly before the actual $\mathrm{tr}(\mathbf{S}_\lambda)$. More generally, $\mathcal{C}$ in (33)–(34) may differ. In (36)–(38), we reduce the additive term q by certain amount, to adjust for the numerical quadrature error in approximating the sum of finite-term series by an integral; see the proof of Theorem 2 for the full details. Further details on how to tune a, b, and $\mathcal{C}$ by the simple least squares method are given in Section 5.1; to obtain rough estimates of DFs, using $a=0$, $b=1$, and $\mathcal{C}=1$ are the simplest choices. In summary, these empirical formulas lend themselves to simple hand calculations. Although other styles of elaborate scheme also may be useful for improving the qualities of the empirical formulas, the simple ways of bias/slope correction suggested

earlier suffice for the simulations conducted herein. The performance of these handy formulas is illustrated with simulation studies in Section 5.

Calibrating degrees of freedom can also be used to make different smoothers comparable with the amount of smoothing that they produce. This can be achieved by prespecifying the target DFs for the smooth, then selecting the values of the corresponding smoothing parameters. Hastie and Tibshirani (1990, sec. 3.5) illustrated a graphical procedure consisting of plotting the exact DF as a function of the smoothing parameter. In such instance, using the empirical formulas above, the DFs can be directly converted into the smoothing parameters. Section 5.1 discusses simulation studies comparing the local linear fit and cubic smoothing spline.

3. VARYING-COEFFICIENT MODEL

This section explores the possibilities of calibrating DFs from the preceding univariare nonparametric regression model to models allowing multivariate predictors. In particular, varying-coefficient models are considered. These models provide a flexible framework for semiparametric and nonparametric regression and generalized regression analysis and do not suffer from the "curse of dimensionality." They arise naturally when one wishes to examine how regression coefficients change over different groups characterized by certain covariates, such as age or time (see the seminal works of Cleveland, Grosse, and Shyu 1992 and Hastie and Tibshirani 1993).

The varying-coefficient model for the scalar response variable Y assumes the following conditional linear structure:

$$Y = a_1(U)X_1 + \cdots + a_d(U)X_d + \varepsilon, \tag{39}$$

for given covariates U and $\mathtt{X} = (X_1, \ldots, X_d)^T$, where ε is the random error with $E(\varepsilon|U, \mathtt{X}) = 0$ and $\operatorname{var}(\varepsilon|U, \mathtt{X}) = \sigma^2$. The $r \times 1$ covariate vector U is assumed to have a sampling density f_U with a known bounded support Ω, and the case $r = 1$ often is practically more useful; $\mathtt{X}$ is assumed to be random. To ensure identifiability of model (39), the $d \times d$ matrix $E(\mathtt{X}\mathtt{X}^T|U = u)$ is assumed to be positive definite for each $u \in \Omega$. Of interest is to estimate the unknown smooth curves $a_j(u)$, $j = 1, \ldots, d$, and the population mean regression function, $m(u, x_1, \ldots, x_d) = \sum_{j=1}^{d} a_j(u)x_j$. If $d = 1$ and $X_1 \equiv c$ (say $c = 1$), then (39) reduces to the nonparametric regression model (1).

Given n independent pairs of measurements $\{(U_i, X_{1i}, \ldots, X_{di}, Y_i)_{i=1}^n\}$ from the model, only a couple of techniques have been proposed for fitting a varying-coefficient model. One plausible way of estimating coefficient functions, $a_j(u)$, applies the smoothing spline approach proposed by Hastie and Tibshirani (1993); minimize

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{d} a_j(U_i)X_{ji}\right]^2 + \sum_{j=1}^{d}\lambda_j \int \{a_j''(u)\}^2\,du,$$

with respect to functions $a_j(u)$, for some positive-valued smoothing parameters $\lambda_1, \ldots, \lambda_d$. As indicated by Fan and Zhang (1999), this method has a number of problems in that it involves selecting multiple smoothing parameters simultaneously, contains burden of computation, and sampling properties of estimates are not easy to obtain. For time-varying coefficient models, Hoover, Rice, Wu, and Yang (1998) discussed the asymptotic properties of kernel regression estimators.

This section focuses on local polynomials because of their intuitiveness and simplicity. Interestingly, the results of Theorem 1 and (5) can be flexibly extended to varying-coefficient models; see Theorem 3 and (54). Therefore, the EGCV-minimizing bandwidth selector continues to be applicable for producing smooth estimates of the varying-coefficient functions. Again, the function estimation procedure for $a_j(u)$ are described first. To characterize the solution, some additional notations are needed. Setting $\boldsymbol{A}(u) = (a_1(u), \ldots, a_d(u))^T$, model (39) can be expressed as $Y = \boldsymbol{A}(U)^T\mathtt{X} + \varepsilon$. Put $\boldsymbol{A}^{(\ell)}(u) = (a_1^{(\ell)}(u), \ldots, a_d^{(\ell)}(u))^T$, $\ell = 0, \ldots, p$. Then for the ith datum point U_i close to a fitting point u_0, via the Taylor series approximation,

$$\boldsymbol{A}(U_i)^T\mathtt{X}_i \approx \sum_{\ell=0}^{p}(U_i - u_0)^\ell \frac{\boldsymbol{A}^{(\ell)}(u_0)^T}{\ell!}\mathtt{X}_i, \tag{40}$$

where $\mathtt{X}_i = (X_{1i}, \ldots, X_{di})^T$. Define by $\boldsymbol{\beta}(u_0) = (\boldsymbol{A}(u_0)^T, \boldsymbol{A}^{(1)}(u_0)^T, \ldots, \frac{\boldsymbol{A}^{(p)}(u_0)^T}{p!})^T$ the $d(p+1)$ by 1 vector of coefficients with their derivatives, and set

$$\boldsymbol{Z}_i(u_0) = (1, (U_i - u_0), \ldots, (U_i - u_0)^p)^T \otimes \mathtt{X}_i,$$

where $\otimes$ denotes the Kronecker product. Then (40) can be written as $\boldsymbol{A}(U_i)^T\mathtt{X}_i \approx \boldsymbol{Z}_i(u_0)^T\boldsymbol{\beta}(u_0)$. Put

$$\begin{aligned}\widetilde{S}_n(u_0) &= \widetilde{\mathbf{X}}(u_0)^T\widetilde{\mathbf{W}}(u_0)\widetilde{\mathbf{X}}(u_0) \qquad \text{and}\\ \widetilde{T}_n(u_0) &= \widetilde{\mathbf{X}}(u_0)^T\widetilde{\mathbf{W}}(u_0),\end{aligned} \tag{41}$$

where

$$\begin{aligned}\widetilde{\mathbf{X}}(u_0) &= (\boldsymbol{Z}_1(u_0), \ldots, \boldsymbol{Z}_n(u_0))^T \qquad \text{and}\\ \widetilde{\mathbf{W}}(u_0) &= \operatorname{diag}\{K_h(U_1 - u_0), \ldots, K_h(U_n - u_0)\}.\end{aligned}$$

Then the pth degree local polynomial estimate $\hat{\boldsymbol{\beta}}(u_0)$, which minimizes the criterion $\sum_{i=1}^{n}\{Y_i - \boldsymbol{Z}_i(u_0)^T\boldsymbol{\beta}\}^2K_h(U_i - u_0)$, can be written explicitly as

$$\begin{aligned}\hat{\boldsymbol{\beta}}(u_0) &= \arg\min_{\boldsymbol{\beta}}\{\mathbf{y} - \widetilde{\mathbf{X}}(u_0)\boldsymbol{\beta}\}^T\widetilde{\mathbf{W}}(u_0)\{\mathbf{y} - \widetilde{\mathbf{X}}(u_0)\boldsymbol{\beta}\}\\ &= \{\widetilde{S}_n(u_0)\}^{-1}\widetilde{T}_n(u_0)\mathbf{y}.\end{aligned} \tag{42}$$

Apparently, the first d entries of $\hat{\boldsymbol{\beta}}(u_0)$ supply estimates $\hat{a}_j(u_0)$ of the coefficient functions $a_j(u_0)$. Write $\widehat{\boldsymbol{A}}(u_0) = (\hat{a}_1(u_0), \ldots, \hat{a}_d(u_0))^T$; that is, $\widehat{\boldsymbol{A}}(u_0) = (\boldsymbol{e}_1^T \otimes \mathbf{I}_d)\hat{\boldsymbol{\beta}}(u_0)$ where $\mathbf{I}_d$ represents a $d \times d$ identity matrix. The corresponding estimate of the mean regression, $m(u_0, \mathtt{x})$, is then given by

$$\begin{aligned}\hat{m}_h(u_0, \mathtt{x}) &= \sum_{j=1}^{d}\hat{a}_j(u_0)x_j = \widehat{\boldsymbol{A}}(u_0)^T\mathtt{x}\\ &= (\boldsymbol{e}_{1,p+1}^T \otimes \mathtt{x}^T)\{\widetilde{S}_n(u_0)\}^{-1}\widetilde{\mathbf{X}}(u_0)^T\widetilde{\mathbf{W}}(u_0)\mathbf{y},\end{aligned} \tag{43}$$

where $\mathtt{x} = (x_1, \ldots, x_d)^T$.

Now expressions are sought for the smoother matrix $\widetilde{\mathbf{S}}_h$, by which it is meant, as in (2), that $(\hat{m}_h(U_1, \mathtt{X}_1), \ldots, \hat{m}_h(U_n, \mathtt{X}_n))^T = \widetilde{\mathbf{S}}_h \mathbf{y}$. Combining the identity $\widetilde{\mathbf{X}}(u_0)^T \widetilde{\mathbf{W}}(u_0) \boldsymbol{e}_{j,n} = \mathbf{Z}_j(u_0) \times K_h(U_j - u_0)$ with (43) leads to

$$\hat{m}_h(u_0, \mathtt{x}) = \sum_{j=1}^{n} \widetilde{W}_0^n\left(u_0, \frac{U_j - u_0}{h}; \mathtt{x}, \mathtt{X}_j\right) Y_j,$$

where

$$\widetilde{W}_0^n(u, t; \mathtt{x}, \mathtt{X}) = (\boldsymbol{e}_{1,p+1}^T \otimes \mathtt{x}^T)\{\widetilde{S}_n(u)\}^{-1}(H \otimes \mathbf{I}_d) \times \{(1, t, \ldots, t^p)^T \otimes \mathtt{X}\}K(t)/h. \quad (44)$$

Consequently, the (i, j)th entry of the smoother matrix $\widetilde{\mathbf{S}}_h$ becomes

$$\widetilde{\mathbf{S}}_h(i, j) = \widetilde{W}_0^n\left(U_i, \frac{U_j - U_i}{h}; \mathtt{X}_i, \mathtt{X}_j\right), \quad i, j = 1, \ldots, n, \quad (45)$$

and, in particular, the entries along the diagonal are $\widetilde{\mathbf{S}}_h(i, i) = \widetilde{W}_0^n(U_i, 0; \mathtt{X}_i, \mathtt{X}_i)$. Thus the explicit expressions for DFs are given by

$$\mathrm{tr}(\widetilde{\mathbf{S}}_h) = \sum_{i=1}^{n} (\boldsymbol{e}_1^T \otimes \mathtt{X}_i^T)\{\widetilde{S}_n(U_i)\}^{-1}(\boldsymbol{e}_1 \otimes \mathtt{X}_i)K(0)/h$$

and

$$\mathrm{tr}(\widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) = \sum_{i=1}^{n}\sum_{j=1}^{n} \big[(\boldsymbol{e}_1^T \otimes \mathtt{X}_i^T)\{\widetilde{S}_n(U_i)\}^{-1}\{(1, (U_j - U_i), \ldots, (U_j - U_i)^p)^T \otimes \mathtt{X}_j\}\big]^2 K^2\left(\frac{U_j - U_i}{h}\right) \Big/ h^2.$$

Similar to Lemma A.1, some nonasymptotic results can be obtained for $\widetilde{\mathbf{S}}_h$. That is, for a nonnegative kernel K satisfying $K(0) = \sup_x K(x)$, $\sum_{j=1}^{n}\{\widetilde{\mathbf{S}}_h(i, j)\}^2 \le \widetilde{\mathbf{S}}_h(i, i)$ for $i = 1, \ldots, n$; for any matrix P whose column space is generated by the vectors $(U_1^{k_j} X_{j1}, \ldots, U_n^{k_j} X_{jn})^T$, for integers $1 \le j \le d$ and $0 \le k_j \le p$, $\widetilde{\mathbf{S}}_h P = P$ and $(\widetilde{\mathbf{S}}_h^T + \widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h)^\ell P = P$ for $\ell = 0, 1, \ldots$; and 1 is an eigenvalue of $\widetilde{\mathbf{S}}_h$ corresponding to $d(p + 1)$ distinct eigenvectors, $(U_1^{k_j} X_{j1}, \ldots, U_n^{k_j} X_{jn})^T$. This means, that for any $h > 0$,

$$d(p + 1) \le \mathrm{tr}(\widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) \le \mathrm{tr}(\widetilde{\mathbf{S}}_h) \le \mathrm{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) < n. \quad (46)$$

Clearly, with the increasing number d of covariates, the demand for trace computations becomes more intensive. However, after the foregoing preliminaries, Theorem 3 presents simple closed-form asymptotic representations of DFs based on $\widetilde{\mathbf{S}}_h$. Again, both random and fixed designs of U_i are considered, but we opt not to state the conclusions separately because $\mathtt{X}_i$, $i = 1, \ldots, n$, are usually assumed to be random regressors in either case.

Theorem 3. For random design, assume condition (C) (see the Appendix); for fixed design, assume condition (C*) (see the Appendix). In either case, when $n \to \infty$, $h \to 0$, and $nh \to \infty$,

$$\mathrm{tr}(\widetilde{\mathbf{S}}_h) = d\mathcal{K}(0)|\Omega|/h\{1 + o_P(1)\}, \quad (47)$$

$$\mathrm{tr}(\widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) = d\mathcal{K} * \mathcal{K}(0)|\Omega|/h\{1 + o_P(1)\}, \quad (48)$$

and

$$\mathrm{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) = d(2\mathcal{K} - \mathcal{K} * \mathcal{K})(0)|\Omega|/h\{1 + o_P(1)\}, \quad (49)$$

where Ω denotes the support of U.

The DFs are asymptotically proportional to the number d of regressor covariates in the varying-coefficient model (39). In particular, if $d = 1$, then the results of Theorem 3 reduce to those of Theorem 1. In this sense, calibration formulas of DF are as simple as those in a single regressor case. Using Theorem 3 along with (46), in a similar spirit to Section 2.5, the empirical formulas for DFs can be proposed:

$$\mathrm{tr}(\widetilde{\mathbf{S}}_h) \simeq d\{(p + 1 - a) + \mathcal{C}n/(n - d)\mathcal{K}(0)|\Omega|/h\}, \quad (50)$$

$$\mathrm{tr}(\widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) \simeq d\{(p + 1 - a) + \mathcal{C}n/(n - d)\mathcal{K} * \mathcal{K}(0)|\Omega|/h\}, \quad (51)$$

and

$$\mathrm{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) \simeq d\{(p + 1 - a) + \mathcal{C}n/(n - d) \times (2\mathcal{K} - \mathcal{K} * \mathcal{K})(0)|\Omega|/h\}. \quad (52)$$

The bandwidth parameter for fitting varying-coefficient models can also be selected based on minimizing (G)CV criterion. The usual leave-one-out CV score in the current setup of modeling has the form

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{h,-i}(U_i, \mathtt{X}_i)\}^2. \quad (53)$$

In the Appendix we show that the foregoing CV function has an alternative expression,

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i - \hat{m}_h(U_i, \mathtt{X}_i)\}^2}{\{1 - \widetilde{\mathbf{S}}_h(i, i)\}^2}, \quad (54)$$

and thus the GCV function is given by

$$\mathrm{GCV}(h) = \frac{n^{-1}\sum_{i=1}^{n}\{Y_i - \hat{m}_h(U_i, \mathtt{X}_i)\}^2}{\{1 - \mathrm{tr}(\widetilde{\mathbf{S}}_h)/n\}^2}. \quad (55)$$

Replacing $\mathrm{tr}(\widetilde{\mathbf{S}}_h)$ in (55) by the empirical formula (50), the EGCV(h) is obtained. Call $\hat{h}_{\mathrm{EGCV}}$ the EGCV-minimizing bandwidth. See Section 5 for the performance of this data-determined bandwidth.

4. MODEL ASSESSMENT

Calibrating DFs not only is useful as a graphical tool for exploratory data analysis, but also provides a helpful diagnostic tool for checking the agreement between a proposed parametric/nonparametric model with the observed dataset. For the purpose of exposition, this section discusses the varying-coefficient model, whose generality includes the nonparametric regression model as a special case.

For varying-coefficient model (39), two types of useful null hypotheses that arise naturally from statistical applications are considered. The first of these tests whether the coefficients that describe the effect of regressors, $X_1, \ldots, X_d$, are really varying as a function of another factor U. This is equivalent to assessing the adequacy of a linear model, with the null hypothesis established as

$$H_0: \quad a_1(u) \equiv c_1, \ldots, a_d(u) \equiv c_d, \quad (56)$$

for some unknown constants $c_1, \ldots, c_d$. To tackle this problem, one can construct the GLR statistic,

$$\lambda_n = \frac{n}{2} \log \frac{\mathrm{RSS}_0}{\mathrm{RSS}_1}, \tag{57}$$

where $\mathrm{RSS}_0 = \sum_{i=1}^n \{Y_i - \sum_{j=1}^d \hat{c}_j X_{ji}\}^2$, and $\mathrm{RSS}_1(h) = \sum_{i=1}^n \{Y_i - \sum_{j=1}^d \hat{a}_j(U_i) X_{ji}\}^2$, with the estimates $\hat{c}_j$ given by the least squares method and $\hat{a}_j(U_i)$ by the local polynomial approach described in the preceding section. [Certainly, in the event that $\{c_j\}$ in (56) are given, their true values will be used for obtaining RSS_0.] For notational clarity, the symbol $\widetilde{\mathbf{S}}_h$ represents the local polynomial smoother matrix, based on d covariates in the full model. The theoretical justification of this testing procedure was given by Fan et al. (2001), who showed that when the null hypothesis (56) holds, $\frac{r_{\mathcal{K}}\lambda_n(h) - 2^{-1} r_{\mathcal{K}} d(2\mathcal{K} - \mathcal{K}*\mathcal{K})(0)|\Omega|/h}{\{r_{\mathcal{K}} d(2\mathcal{K} - \mathcal{K}*\mathcal{K})(0)|\Omega|/h\}^{1/2}}$ converges in law to a standard normal distribution, as $h \to 0$ at a certain rate. The normalizing quantity before $\lambda_n(h)$ is given by $r_{\mathcal{K}} = \frac{(\mathcal{K} - 2^{-1}\mathcal{K}*\mathcal{K})(0)}{\int (\mathcal{K} - 2^{-1}\mathcal{K}*\mathcal{K})^2(t)\,dt}$; refer to Table 1 for the values of $r_{\mathcal{K}}$, which are close to 2. Applying the asymptotic DF formulas given in Theorem 3, this sampling distribution can be stated equivalently as

$$\frac{r_{\mathcal{K}}\lambda_n(h) - 2^{-1} r_{\mathcal{K}} \operatorname{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h)}{\{r_{\mathcal{K}} \operatorname{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h)\}^{1/2}} \xrightarrow{\mathcal{L}} N(0,1), \tag{58}$$

where $\xrightarrow{\mathcal{L}}$ denotes converges in distribution. The presence of $r_{\mathcal{K}}$ guarantees that $r_{\mathcal{K}}\lambda_n(h)$ has its asymptotic mean and variance above in a 1 : 2 ratio. In this sense, $r_{\mathcal{K}}\lambda_n(h)$ can be viewed as asymptotically chi-squared distributed, with DFs equal to $2^{-1} r_{\mathcal{K}} \operatorname{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h)$.

In another formulation of model assumptions, the null contains many nuisance functions. For instance, to assess whether the variables $X_1, \ldots, X_{d_1}$, $1 \le d_1 < d$, are significant or not involves testing whether certain coefficient functions are identically 0s

$$H_0: \quad a_1(u) = 0, \ldots, a_{d_1}(u) = 0 \tag{59}$$

(without placing restrictions on the effects of the remaining variables). In this case, the GLR statistics can be constructed in similar ways; namely, under the null (59), obtain the local polynomial estimates of $a_{d_1+1}(\cdot), \ldots, a_d(\cdot)$. Call these estimates $\hat{a}_j^0(\cdot)$, $j = d_1 + 1, \ldots, d$. Denote by $\widetilde{\mathbf{S}}_h^0$ the corresponding smoother matrix, based on the $d - d_1$ covariates in the reduced/null model. After that, put $\mathrm{RSS}_0(h) = \sum_{i=1}^n \{Y_i - \sum_{j=d_1+1}^d \hat{a}_j^0(U_i) X_{ji}\}^2$ into $\lambda_n(h)$. According to Fan et al. (2001), it follows that under the null hypothesis (59), $\frac{r_{\mathcal{K}}\lambda_n(h) - 2^{-1} r_{\mathcal{K}} d_1(2\mathcal{K} - \mathcal{K}*\mathcal{K})(0)|\Omega|/h}{\{r_{\mathcal{K}} d_1(2\mathcal{K} - \mathcal{K}*\mathcal{K})(0)|\Omega|/h\}^{1/2}}$ converges in distribution to a standard normal as $h \to 0$ at certain rate. Once again, this sampling distribution indicates that

$$\frac{r_{\mathcal{K}}\lambda_n(h) - 2^{-1} r_{\mathcal{K}} \{\operatorname{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) - \operatorname{tr}(2\widetilde{\mathbf{S}}_h^0 - \widetilde{\mathbf{S}}_h^{0T} \widetilde{\mathbf{S}}_h^0)\}}{[r_{\mathcal{K}} \{\operatorname{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) - \operatorname{tr}(2\widetilde{\mathbf{S}}_h^0 - \widetilde{\mathbf{S}}_h^{0T} \widetilde{\mathbf{S}}_h^0)\}]^{1/2}}$$

$$\xrightarrow{\mathcal{L}} N(0,1). \tag{60}$$

Indeed, the result (60) unifies the result (58). This can be understood from the observation; the smoother matrix $\widetilde{\mathbf{S}}_h^0$, corresponding to (56), is actually a usual projection matrix with $\operatorname{tr}(2\widetilde{\mathbf{S}}_h^0 - \widetilde{\mathbf{S}}_h^{0T} \widetilde{\mathbf{S}}_h^0) = d$, whose magnitude is asymptotically smaller than the counterpart of $\widetilde{\mathbf{S}}_h$, and is thus ignored in (58). Hence for the problem posed in either (56) or (59),

$$2^{-1} r_{\mathcal{K}} \{\operatorname{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T \widetilde{\mathbf{S}}_h) - \operatorname{tr}(2\widetilde{\mathbf{S}}_h^0 - \widetilde{\mathbf{S}}_h^{0T} \widetilde{\mathbf{S}}_h^0)\} \tag{61}$$

is the observed degrees of freedom (ODF) of the test statistic $r_{\mathcal{K}}\lambda_n(h)$; similarly, the version of (61) evaluated from the empirical formulas is the EDF.

It should be stressed that the difference between the EDF and its asymptotic form may be practically large. Working with the EDF has the advantage of making the distributional results (58) and (60) more closely reflected in finite-sample situations. Therefore, for practical applications of GLR tests, use of the EDF is recommended in place of its asymptotic form given by Fan et al. (2001).

5. SIMULATIONS

5.1 Assessing the Empirical Formulas for Degrees of Freedom

This section presents some finite-sample simulation studies. The purposes are two-fold: to assess the extent to which the empirical formulas for DFs approximate their exact values and to illustrate numerical comparisons of different smoothers in which the smoothing parameters are chosen based on the empirical formulas. To simplify the programming, assume that the design is on the interval $\Omega = [0, 1]$, so that $|\Omega| = 1$ (see Theorem 1) and $c(f) = 1$ (see Theorem 2). For reasons of computational efficiency, the Epanechnikov kernel is used throughout the simulations.

As an illustration, first consider the fixed uniform design points, $x_i = (i - .5)/n$, $i = 1, \ldots, n$, in which case the DFs are nonrandom. Determine a and $\mathcal{C}$ in the empirical formulas (33)–(35) based on the local linear smoother. With a medium sample size $n = 200$, the simple least squares estimates of three sets, $(h_j^{-1}, \operatorname{tr}(\mathbf{S}_{h_j}))$, $(h_j^{-1}, \operatorname{tr}(\mathbf{S}_{h_j}^T \mathbf{S}_{h_j}))$, and $(h_j^{-1}, \operatorname{tr}(2\mathbf{S}_{h_j} - \mathbf{S}_{h_j}^T \mathbf{S}_{h_j}))$, with respect to 20 bandwidths h_j, logarithmically evenly spaced between .025 and .20, give rise to the estimates of intercept, 1.4531, 1.4603, and 1.4458 and the estimates of slope, .7513, .6033, and .8993. These estimates, combined with (33)–(35) and Table 1, suggest that for $p = 1$, $p + 1 - a \simeq 1.45$ or $a \simeq .55$, and $\mathcal{C} \simeq 1$. The larger the n, the better the approximation provided by the empirical formulas. Using this strategy for local polynomial smoother of other degrees p, the recommended choices for a and $\mathcal{C}$ are given in Table 3.

Analogously, when the cubic smoothing spline method is applied to the foregoing $\{x_i\}$, the simple least squares estimates of three sets, $(\lambda_j^{-4}, \operatorname{tr}(\mathbf{S}_{\lambda_j}))$, $(\lambda_j^{-4}, \operatorname{tr}(\mathbf{S}_{\lambda_j}^T \mathbf{S}_{\lambda_j}))$, and $(\lambda_j^{-4}, \operatorname{tr}(2\mathbf{S}_{\lambda_j} - \mathbf{S}_{\lambda_j}^T \mathbf{S}_{\lambda_j}))$, yield the intercept estimates, 1.0038, 1.0015, and 1.0061, and slope estimates .3533, .2651, and .4416. Thus for $q = 2$, the choice $b = 1$ is adopted in (36)–(38). There the range of λ_j values is chosen so as to obtain an agreement in range between the empirical $\operatorname{tr}(\mathbf{S}_{\lambda_j})$ and the empirical $\operatorname{tr}(\mathbf{S}_{h_j})$. In each panel of Figure 1, the dots denote the actual values of the traces and the centers of the circles denote those evaluated from the empirical formulas. All plots provide convincing evidence that the empirical formulas track the evolution of the DFs as a function of the smoothing parameters nearly perfectly.

Table 3. Choices of a and $\mathcal{C}$, in the Empirical Formulas (33)–(35) and (50)–(52), for the pth-Degree Local Polynomial Smoother

Design type	p	a	$\mathcal{C}$
Fixed	0	.55	1
	1	.55	1
	2	1.55	1
	3	1.55	1
Random	0	.30	.99
	1	.70	1.03
	2	1.30	.99
	3	1.70	1.03

To see how use of the DF formulas make it easy to compare the amount of smoothing by different types of smoothers, consider Figure 2. This figure displays the local linear fit and the cubic smoothing spline fit to a sequence of observations Y_i at fixed-design points $x_i = (i - .5)/n$, simulated from model

$$Y_i = m(x_i) + \sigma\varepsilon_i, \qquad i = 1, \ldots, n, \tag{62}$$

where $m(x) = .6 + .3\cos(2\pi x)$ and ε_i are independent standard normal random variables. The noise variance σ^2 is chosen so that the signal-to-noise ratio (SNR), defined by $\text{var}\{m(x_1), \ldots, m(x_n)\}/\sigma^2$, is roughly equal to 4, a median amount of SNR. Figures 2(a) and (b) correspond to the smoothing parameters h and λ, which are chosen so that the empirical formulas, (33) for $\text{tr}(\mathbf{S}_h)$ and (36) for $\text{tr}(\mathbf{S}_\lambda)$, are set at 5 and 11. It can be observed clearly that the empirical DF formulas can produce two types of nonparametric fits comparable in a very simple fashion. Similar plots based on specifying $\text{tr}(\mathbf{S}_h^T\mathbf{S}_h)$ and $\text{tr}(\mathbf{S}_\lambda^T\mathbf{S}_\lambda)$ have been obtained in Figures 2(c) and (d).

Now consider random designs, in which case the observed traces are random quantities and the choices a and $\mathcal{C}$ will necessarily differ slightly from their counterparts in the fixed design. For the local linear method, the least squares estimates, based on 400 independent $U(0, 1)$ random variables X_i, give $a = .70$ and $\mathcal{C} = 1.03$. (These choices are adopted throughout the remaining simulations in random design.) Table 3 collects the choices of a and $\mathcal{C}$ for other degrees of the local polynomial regression method. Figure 3 presents typical plots of DFs based on one sample path. These plots show that the empirical formulas capture the observed patterns of the DFs reasonably well. Table 4 summarizes the sample mean and variance of the observed DFs, based on 100 sets of independent samples $\{X_i, i = 1, \ldots, n\}$, for $n = 400$ and 1,000. These summary statistics demonstrate that the average values of the observed DFs are slowly varying with sample size, whereas the variabilities of these random quantities decrease quickly with sample size, and that for fixed n, the larger the value of h, the smaller amount of variability in the DFs.

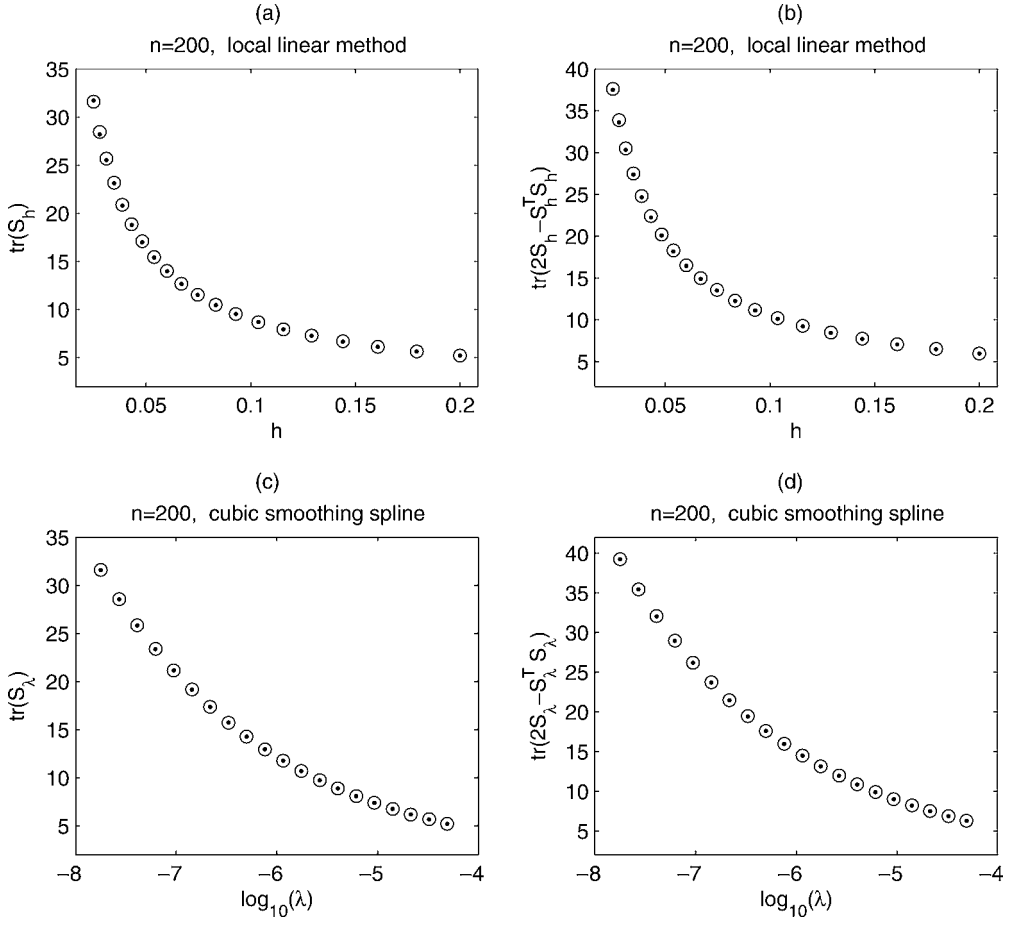

Figure 1. Plots of DFs Versus Smoothing Parameters, Under Fixed Uniform Design. Dots denote the actual values, and centers of circles represent the values using the empirical formulas (33) and (35) for the local linear smoother with a = .55 and $\mathcal{C} = 1$ [(a) and (b)] and (36) and (38) for the cubic smoothing spline with b = 1 [(c) and (d)].

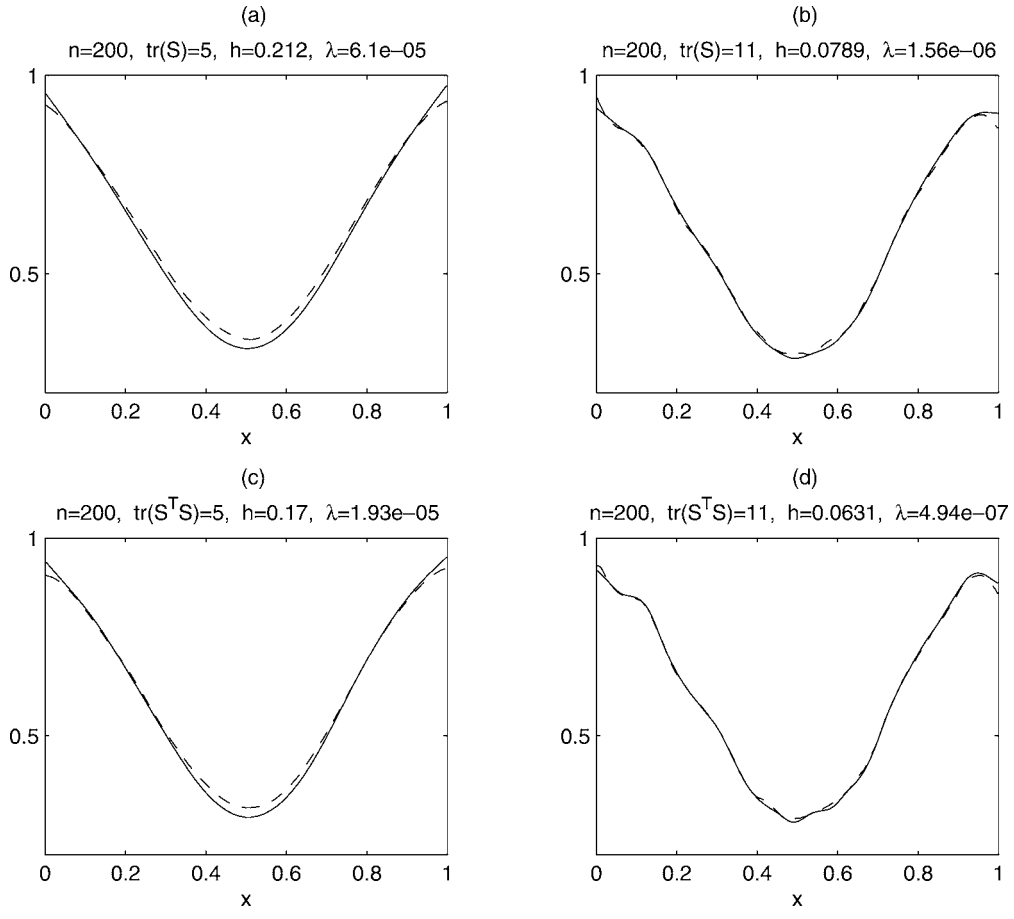

Figure 2. Comparison Between Local Linear Fit (dashed curve) and Cubic Smoothing Spline Fit (solid curve) of the Regression Curve in (62), Under Fixed Uniform Design. In (a) and (b), h and λ are chosen so that the empirical formulas of tr($\mathbf{S}_h$) and tr($\mathbf{S}_\lambda$) are set to be 5 and 11; in (c) and (d), h and λ are set based on the empirical formulas of tr($\mathbf{S}_h^T\mathbf{S}_h$) and tr($\mathbf{S}_\lambda^T\mathbf{S}_\lambda$).

5.2 Nonparametric Regression Model: Smoothing Parameter Selection

This section reports a simulation study done to evaluate the practical performance of the proposed EGCV-based bandwidth selector, as well as some existing bandwidth selectors, for local linear regression. For ease of comparison, consider two sets of regression functions,

$$\text{Example 1:} \quad m(x) = \sin(10\pi x)$$

and

$$\text{Example 2:} \quad m(x) = (4x-2) + 2\exp\{-16(4x-2)^2\},$$

in the model $Y = m(X) + \sigma\varepsilon$, with $X \sim U(0,1)$, $\varepsilon \sim N(0,1)$, and ε independent of X. The noise variance σ^2 in each case is chosen so that SNR equals 4.

A total of 400 random samples are drawn per setting with sample size $n = 400$. For each of these simulated datasets, four automatically selected bandwidths are computed:

$\hat{h}_{\text{EGCV}}$, a bandwidth that minimizes the EGCV;

$\hat{h}_{\text{GCV}}$, a bandwidth that minimizes the GCV;

$\hat{h}_{\text{FG}}$, the (global) refined bandwidth selector of Fan and Gijbels (1995);

and

$\hat{h}_{\text{RSW}}$, the direct plug-in bandwidth selector of Ruppert, Sheather, and Wand (1995).

Then h_{AMISE} in (3), the bandwidth asymptotically optimal but in practice unknown, is calculated. The first three bandwidth selectors are searched over an interval $[h_{\min}, h_{\max}]$ at a geometric grid of points, $h_j = C^{j-1}h_{\min}$, $j = 1, 2, \ldots$, with $C > 1$. The present implementation uses $C = 1.2$, $h_{\max} = .50$, and $h_{\min} = \max[5/n, \max_{2\le j\le n}\{X_{(j)} - X_{(j-1)}\}]$, where $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$. Figure 4 compares the relative errors of these bandwidth selectors to h_{AMISE}. As expected, there is little difference in the behaviors of $\hat{h}_{\text{EGCV}}$ and $\hat{h}_{\text{GCV}}$. In most cases, $\hat{h}_{\text{FG}}$ and $\hat{h}_{\text{RSW}}$ tend to oversmooth; this tendency is most pronounced for $\hat{h}_{\text{RSW}}$ in Example 1. This observation is similar to that obtained from small-sample simulation studies by Lee and Solo (1999), who compared the CV-minimizing bandwidth selector with $\hat{h}_{\text{FG}}$ and $\hat{h}_{\text{RSW}}$. Among the four selectors, $\hat{h}_{\text{FG}}$ has less variation and $\hat{h}_{\text{EGCV}}$ is closer to the asymptotically optimal bandwidth. Furthermore, numerical experience suggests that $\hat{h}_{\text{FG}}$ is occasionally unstable when using kernels with bounded support, such as the Epanechnikov kernel; that is, zero values of a matrix trace may occur in the denominator of equation (2.3) of Fan and Gijbels (1995). Figure 4

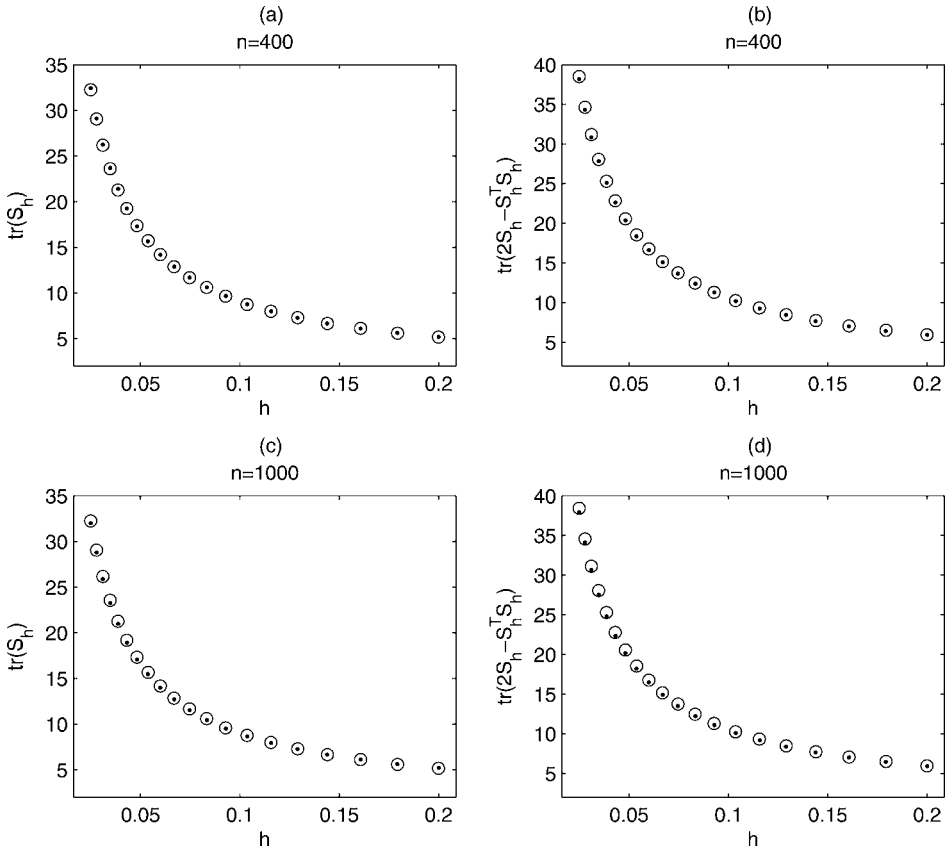

Figure 3. Plots of tr($\mathbf{S}_h$) [(a) and (c)] and tr($2\mathbf{S}_h - \mathbf{S}_h^T\mathbf{S}_h$) [(b) and (d)] Versus Bandwidth h for the Local Linear Method, Under Random Uniform Design. Dots denote the observed values, and centers of circles represent the values using the empirical formulas, (33) and (35), with $a = .7$ and $C = 1.03$.

also shows the boxplot of the averaged squared error (ASE), where $\text{ASE} = n^{-1}\sum_{i=1}^{n}\{\hat{m}_h(X_i) - m(X_i)\}^2$. In this aspect, the ASEs for all methods exhibit quite similar behavior; however, the ASEs alone may be less informative in distinguishing between bandwidth selectors that produce oversmoothed and undersmoothed fits.

5.3 Varying-Coefficient Model: Fitting Coefficient Functions

For the varying-coefficient model, the following example illustrates the performance of the EGCV-based bandwidth selector in curve fitting by the local linear method:

$$\text{Example 1:}\quad Y = \sin(3\pi U)X_1 + \sin(2\pi U)X_2 + \sigma\varepsilon, \qquad (63)$$

where U follows a uniform distribution on $[0, 1]$ and X_1 and X_2 are normally distributed with mean 0, unit variance, and correlation coefficient $2^{-1/2}$. Furthermore, U, (X_1, X_2), and ε are independent. The noise ε follows a standard normal distribution; σ is chosen so that the SN ratio is about 4 : 1.

First, examine the approximation of the empirical formulas (50)–(52). Generate from model (63) a three-covariate random sample $\{(U_i, X_{1i}, X_{2i})_{i=1}^{n}\}$, with each sample consisting of 400

Table 4. Sample Mean and Variance (%, in brackets) of tr($\mathbf{S}_h$), tr($\mathbf{S}_h^T\mathbf{S}_h$), and tr($2\mathbf{S}_h - \mathbf{S}_h^T\mathbf{S}_h$), Based on 100 Independent Samplings, Each of Which Contains n Independent Uniform Random Variables

		h									
n	*Statistic*	*.0250*	*.0311*	*.0387*	*.0482*	*.0600*	*.0747*	*.0930*	*.1157*	*.1440*	*.1793*
400	tr($\mathbf{S}_h$)	32.41	26.16	21.19	17.24	14.11	11.60	9.59	7.98	6.70	5.67
		(6.44)	(2.69)	(1.35)	(.85)	(.61)	(.42)	(.34)	(.24)	(.16)	(.13)
	tr($\mathbf{S}_h^T\mathbf{S}_h$)	26.61	21.48	17.42	14.20	11.67	9.64	8.02	6.72	5.69	4.87
		(7.94)	(3.01)	(1.35)	(.87)	(.60)	(.38)	(.31)	(.19)	(.12)	(.11)
	tr($2\mathbf{S}_h - \mathbf{S}_h^T\mathbf{S}_h$)	38.21	30.83	24.97	20.28	16.54	13.56	11.16	9.24	7.70	6.47
		(5.56)	(2.75)	(1.52)	(.95)	(.69)	(.51)	(.41)	(.31)	(.21)	(.16)
1,000	tr($\mathbf{S}_h$)	31.83	25.80	20.98	17.13	14.03	11.55	9.56	7.97	6.69	5.66
		(1.01)	(.52)	(.30)	(.20)	(.16)	(.12)	(.10)	(.08)	(.06)	(.04)
	tr($\mathbf{S}_h^T\mathbf{S}_h$)	25.92	21.05	17.17	14.06	11.57	9.58	7.98	6.70	5.68	4.86
		(1.05)	(.51)	(.29)	(.20)	(.14)	(.11)	(.09)	(.07)	(.04)	(.03)
	tr($2\mathbf{S}_h - \mathbf{S}_h^T\mathbf{S}_h$)	37.74	30.55	24.80	20.19	16.49	13.52	11.14	9.23	7.70	6.47
		(1.04)	(.59)	(.34)	(.23)	(.19)	(.15)	(.12)	(.10)	(.08)	(.06)

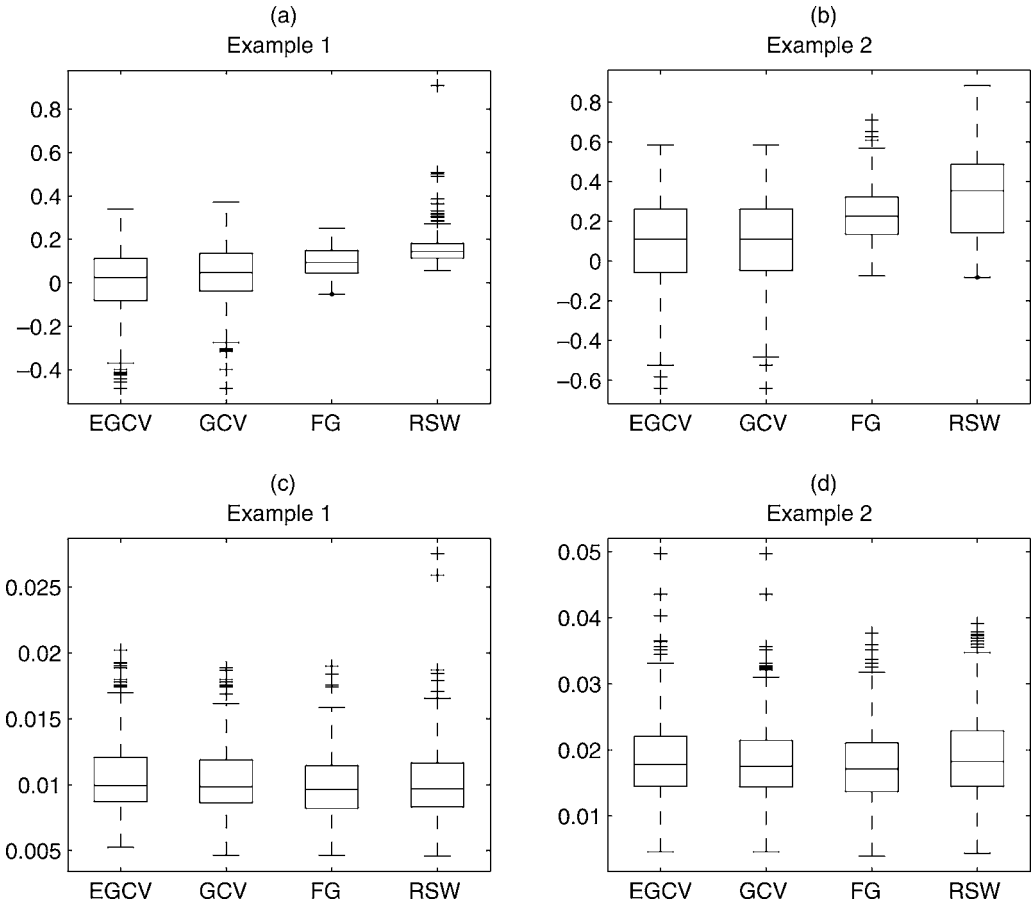

Figure 4. Comparison of Various Bandwidth Selectors $\hat{h}$. (a) and (b) Boxplots of the relative errors $(\hat{h} - h_{AMISE})/h_{AMISE}$. (c) and (d) Boxplots of the average squared errors $n^{-1}\sum_{i=1}^{n}\{\hat{m}(X_i) - m(X_i)\}^2$. In each panel, the boxplots correspond to (from left to right) $\hat{h}_{EGCV}$, $\hat{h}_{GCV}$, $\hat{h}_{FG}$, and $\hat{h}_{RSW}$.

observations. Figure 5 compares the exact values of $\mathrm{tr}(\widetilde{\mathbf{S}}_h)$ and $\mathrm{tr}(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T\widetilde{\mathbf{S}}_h)$ with their empirical formulas based on the local linear smoother. As sample size n grows, the overall patterns of the EDFs do resemble those actually observed.

Now, 100 independent samples are generated with sample size 400 from (63), and the local linear technique is used to fit the varying-coefficient model. The bandwidth is chosen to minimize the EGCV function. Figure 6 depicts the local linear estimates of the varying-coefficient functions $a_1(u)$ and $a_2(u)$ for Example 1, in which the smoothness of $a_1(u)$ and the smoothness of $a_2(u)$ are comparable. In each panel, the solid curves denote the true coefficient functions. Three typically estimated curves are presented, corresponding to the 10th (the dotted curve), 50th (the dashed curve), and 90th (the dash-dotted curve) percentiles among the ASE-based curves, where $\mathrm{ASE} = n^{-1}\sum_{i=1}^{n}\{\hat{m}_h(U_i, \mathrm{X}_i) - m(U_i, \mathrm{X}_i)\}^2$. The performance of $\hat{h}_{\mathrm{EGCV}}$, when applied to recovering multiple smooth curves in varying-coefficient models, is reasonable.

The local polynomial regression method discussed in Section 3 assumes implicitly the similarity between the degrees of smoothness of functions $a_j(u)$, $j = 1, \ldots, d$. To achieve the optimal rates of convergence for $a_j(u)$ with differing smoothness, the two-step iterative estimation procedure proposed by Fan and Zhang (1999) offers a flexible alternative and improvement over the one-step procedure. However, practical implementation of this approach relies on seeking, in the first-step, a simple and good pilot bandwidth to estimate a_j jointly, and identifying certain functions a_j, which are actually significantly smoother than the rest of functions and thus need to be reestimated individually in the second step. In such instances, $\hat{h}_{\mathrm{EGCV}}$ can be easily used in the initial stage for pilot bandwidth; a visual inspection of the preliminary fitted curves provides a quick check on the inhomogeneous smoothness across $a_j(u)$. Again, a simple choice of bandwidth in the second-step smoothing is to minimize the EGCV function. Illustrative examples are omitted here.

5.4 Varying-Coefficient Model: Hypothesis Testing

The objective in this section is to use simulations to study the discriminatory power of the testing procedure described in Section 4. For illustration, consider a four-covariate varying-coefficient model,

$$Y = a_1(U)X_1 + a_2(U)X_2 + a_3(U)X_3 + \sigma\varepsilon. \tag{64}$$

Set $X_3 \equiv 1$, and let $(U, X_1, X_2, \varepsilon)$ have the same types of joint distributions as specified in the previous section. Suppose that one is interested in the following typical forms of assertions about model (64):

$$H_0^{(1)}: \quad a_1(u) \equiv c_1,\ a_2(u) \equiv c_2,\ a_3(u) \equiv c_3 \tag{65}$$

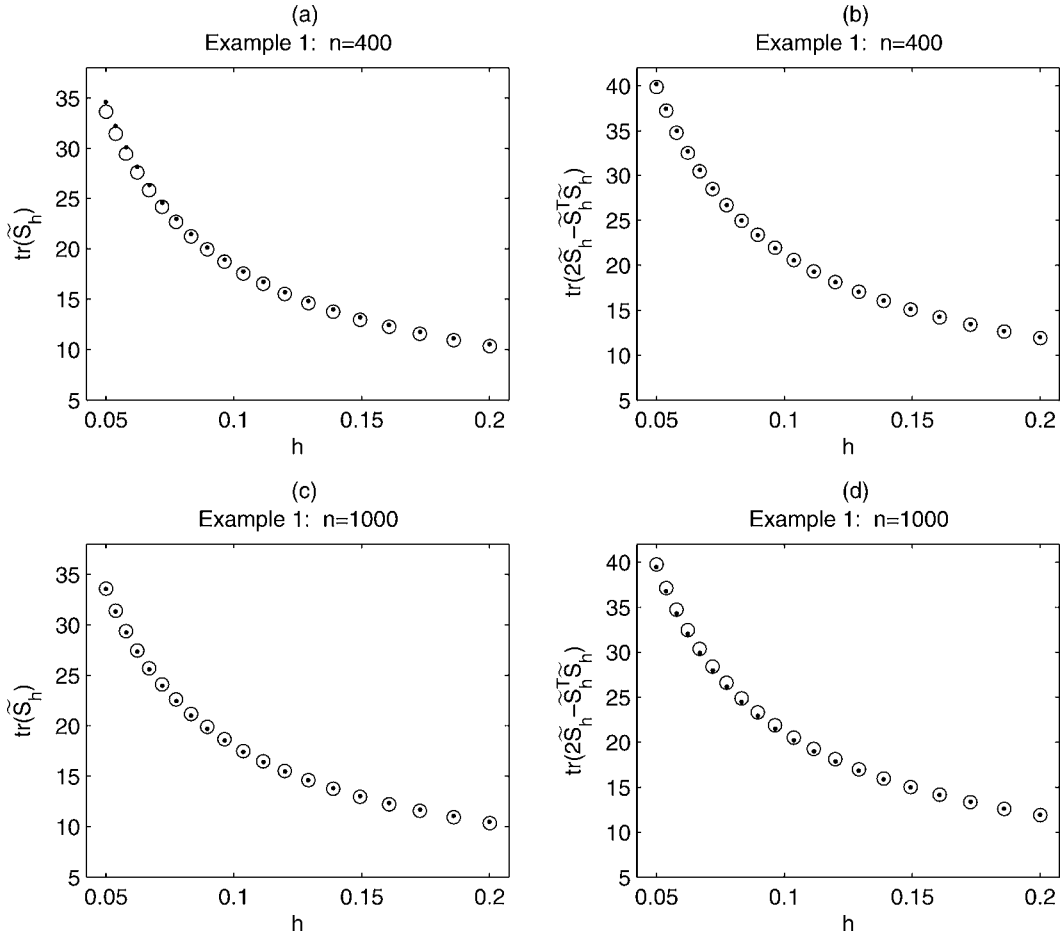

Figure 5. Plots of [(a) and (c)] $tr(\widetilde{\mathbf{S}}_h)$ and [(b) and (d)] $tr(2\widetilde{\mathbf{S}}_h - \widetilde{\mathbf{S}}_h^T\widetilde{\mathbf{S}}_h)$ Versus Bandwidth h, When the Local Linear Method is Applied. Dots denote the observed values, and centers of circles represent the values using the empirical formulas (50) and (52), with $a = .7$ and $C = 1.03$.

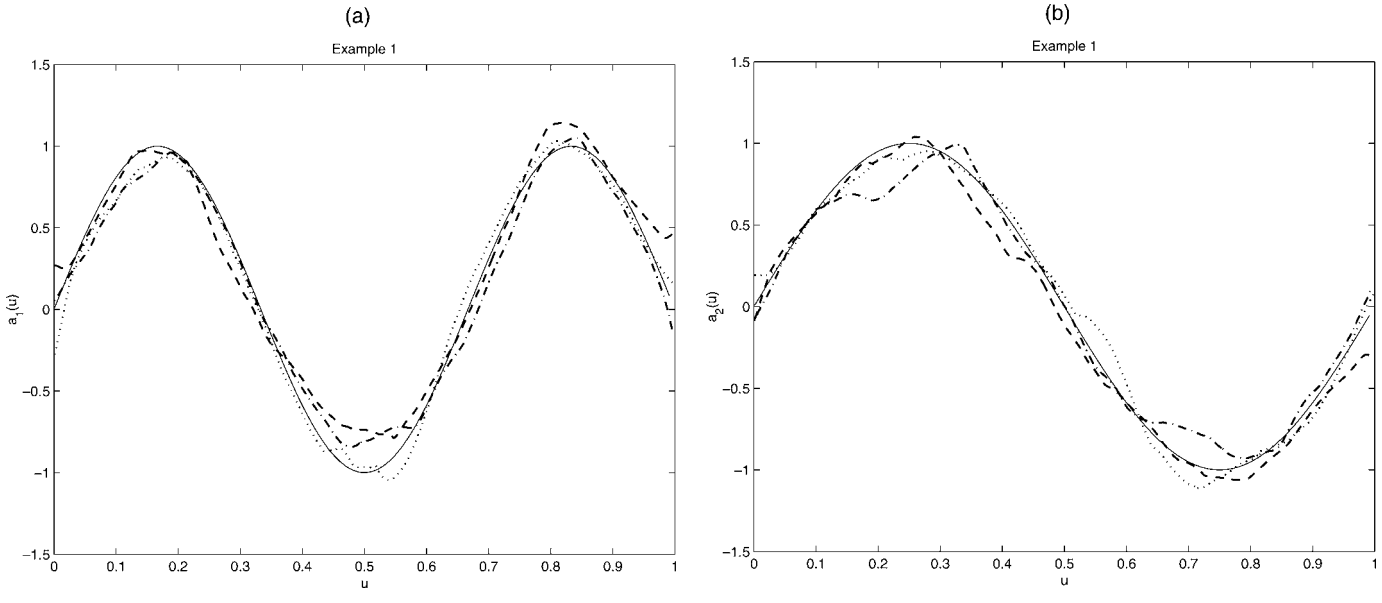

Figure 6. Use of the EGCV-Minimizing Bandwidth Selector for Fitting the Coefficient Functions (a) $a_1(u)$ and (b) $a_2(u)$ in (63). Three typical estimated curves are presented, corresponding to the 10th (dotted curve), the 50th (dashed curve), and the 90th (dashed-dotted curve) percentiles among the ASE-ranked curves. The solid curves denote the true coefficient functions.

and

$$H_0^{(2)}: \quad a_2(u) \equiv 0, \tag{66}$$

where in $H_0^{(1)}$ the three constants c_j are unknown. No other testing procedures deal with the foregoing model-checking problems in varying-coefficient models, so comparisons with published results are impossible.

To obtain the critical values and powers of the GLR test statistics $r_K\lambda_n(h)$, the data are simulated as follows. To test (65), observations are generated from model (64) using the varying-coefficient functions,

$$a_1(u) = 1 + \theta \sin(2\pi u), \qquad a_2(u) = 1/2 - \cos(3\pi\theta u),$$
$$a_3(u) = 4u(1-u)\theta - 1, \qquad \theta \geq 0. \tag{67}$$

Analogously, to test (66), the data are generated from model (64) with

$$a_1(u) = \sin(2\pi u), \qquad a_2(u) = \theta u^2,$$
$$a_3(u) = 4u(1-u) - 1, \qquad \theta \geq 0. \tag{68}$$

In both cases, θ indexes structural deviations of the alternative models from the null models. Particularly, let $\theta = 0$ characterize the model from which data are simulated under the null hypotheses. Again, the magnitude of σ in (64) is determined so that SNR under each null hypothesis equals 4.

Note the following remarks about bandwidth selection in model checking. Take the null hypothesis (65), for instance. If a sample of observed data indeed comes from model (64) satisfying this null assumption (linear model), then the optimal bandwidth for local fitting of the coefficient functions should be close to infinity, and data-driven bandwidth selectors will lead to a large bandwidth. However, the distributional properties in (58) and (60) rely implicitly on the assumption $h \to 0$. Hence the bandwidth well suited for producing visually smooth estimates of the underlying curves may not in general be appropriate for checking model assumptions. Based on this consideration, an empirical bandwidth formula,

$$h^* = \text{std}(U) \times n^{-2/(4p+5)}, \tag{69}$$

is proposed for model checking based on the pth degree local polynomial regression. The rate $n^{-2/(4p+5)}$ was given by Fan et al. (2001); in practice, the standard deviation of the covariate U can be simply replaced by its sample standard deviation.

5.4.1 Large Datasets. First, methodologies for processing large datasets are developed. Experience indicates that under the null hypotheses, the sampling distribution of the test statistic, $r_K\lambda_n(h^*)$, is close to a chi-squared distribution, $\chi^2_{\text{EDF}+2}$, where EDF represents the empirical degrees of freedom for hypothesis testing, defined at the end of Section 4. To see this, simulate, from the null hypotheses (65) and (66), 400 random samples each of size 400. Figure 7(a) and Figure 8(a) plot the kernel density estimates (Fan and Gijbels 1996, p. 47) of $\{r_K\lambda_n(h^*)\}$, for (65) and (66). For comparison, two reference density functions are also included: the normal density with mean EDF and standard deviation $\sqrt{2\,\text{EDF}}$ and the chi-squared density $\chi^2_{\text{EDF}+2}$. Table 5 compares the simulated rejection rates of $r_K\lambda_n(h^*)$ exceeding the $100(1-\alpha)$th percentiles of the reference distributions. Note that the chi-squared approximation gives better agreement with the nominal type-I errors than does the normal approximation. Moreover, using EDF instead of ODF in the parameters of the reference distributions makes little difference in approximating the tail distribution of $r_K\lambda_n(h^*)$, but has the added merit of being quickly implemented.

To measure the GLR test's ability to detect departures from the null, the following power studies were performed.

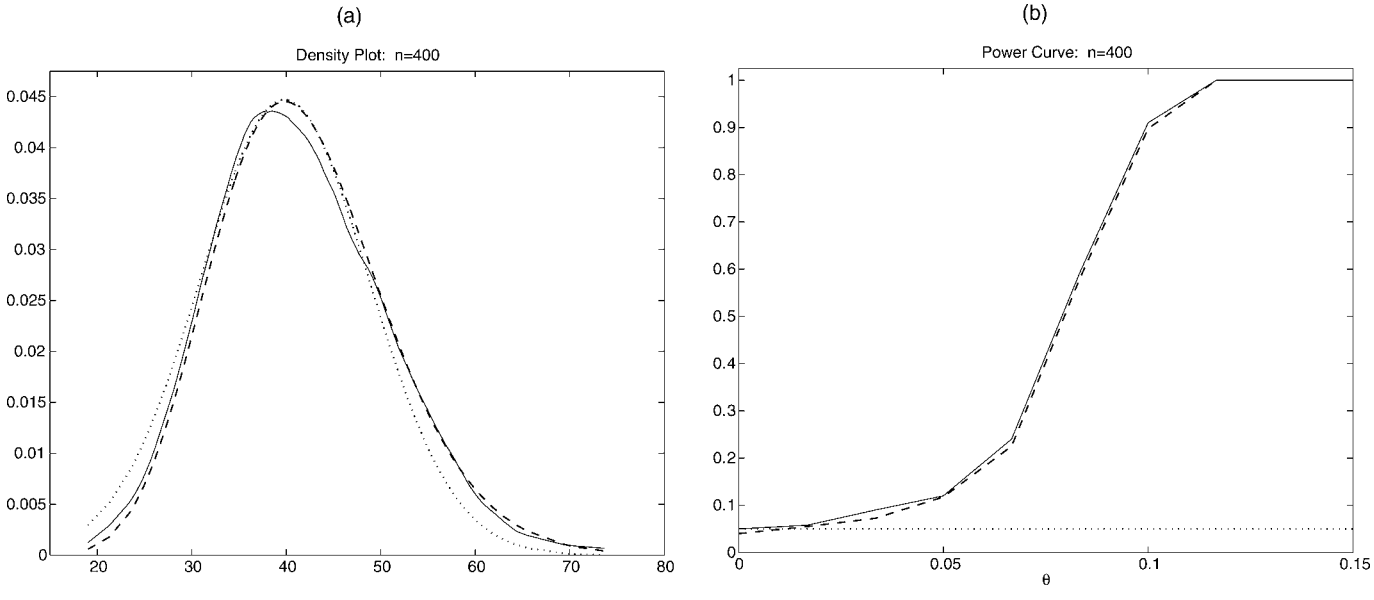

Figure 7. (a) Kernel density estimate of the test statistics, $r_K\lambda_n(h^*)$, under the null hypothesis (65), based on 400 random samples each of size 400. The solid curve represents kernel density estimate; the dotted curve, density of N(EDF, $\sqrt{2\text{EDF}}$); and the dashed curve, density of χ^2_{EDF+2}. (b) Estimated power curves when the data are simulated from (67). The solid curve is based on the simulated null critical values, and the dashed curve is based on the percentile of χ^2_{EDF+2}. The bottom dotted line denotes the nominal level of significance.

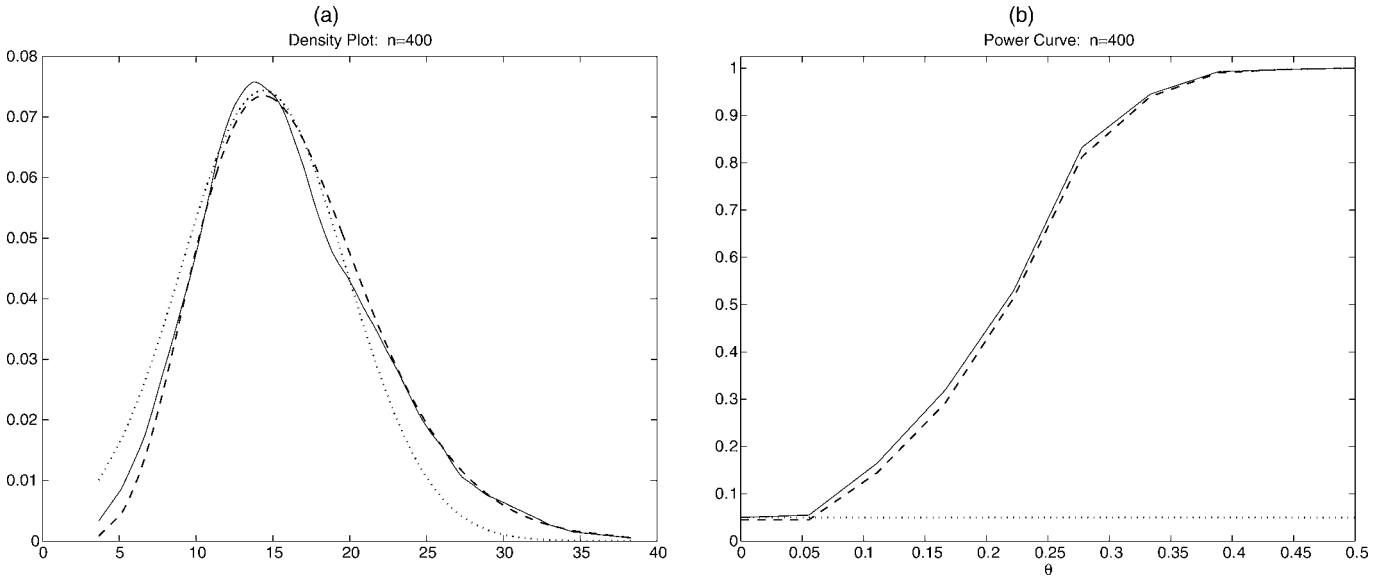

Figure 8. Same as Figure 7, Except That (a) Shows the Null Hypothesis (66) and in (b) the Data Are Simulated From (68).

A total of 400 independent samples were simulated from the alternative models indexed by θ, with each sample containing 400 observations; for simplicity, 10 values of θ evenly spaced on an interval $[0, .15]$ with (67) and $[0, .5]$ with (68) were considered. At a nominal level 5%, the empirical powers are estimated through simulation by the proportion of observed test statistics $r_{\mathcal{K}}\lambda_n(h^*)$, across 400 samples, exceeding the 95th percentile of the $\chi^2_{\mathrm{EDF}+2}$ distribution. For comparison, also included is the rate of the observed test statistics exceeding their 95th sample percentile, under the null, across 400 samples. Figures 7(b) and 8(b) display the estimated power curves. It is clear that the GLR tests are powerful, besides holding their correct sizes.

Simulation studies of model checking with small datasets, via the bootstrap procedure, are omitted here.

6. DISCUSSION

This article has provided a simple, yet flexible extension of the criterion based on GCV, termed EGCV, for smoothing parameter selection in nonparametric regression. By using an empirical formulas of DFs, the computational burden associated with smoothing large datasets is improved considerably, and the procedure for model checking is also easier to carry out. It is hoped that the methodology presented here will be useful in mining large collections of data.

In closing, several points bear mentioning. First, an application to bandwidth selectors based on principles other than CV or GCV would also be valuable. Hurvich, Simonoff, and Tsai (1998) used an improved Akaike information criterion, to which the empirical formulas of DFs proposed here can also apply. Second, if there is a clear indication that the design is nonuniform, then the data-based form, $n^{-1}\sum_{i=1}^{n} f^{-1}(X_i)$, could be used directly in place of its asymptotic form, $|\Omega|$, in the empirical formulas (33)–(35); refer to (A.10) and (A.18) for further derivational details. In such instances, if f is unknown, then those $f(X_i)$ can be replaced by their kernel density estimates, which are computationally far more economic than evaluating the actual traces, but will surely improve the finite-sample performance of the empirical approximations. Indeed, experiments

Table 5. Simulated Rejection Rates of the Test Statistics, Based on 400 Random Samples of Size $n = 400$, From Null Hypotheses $H_0^{(1)}$ and $H_0^{(2)}$. The significance levels are $\alpha = .01, .025, .05, .10$

			Rejection rate at level α			
Null	*Test statistic*	*Reference distribution*	$\alpha = .01$	$\alpha = .025$	$\alpha = .05$	$\alpha = .10$
$H_0^{(1)}$	$r_{\mathcal{K}}\lambda_n(\mathrm{h}^*)$	$N(EDF, \sqrt{2\mathrm{EDF}})$	.0250	.0500	.0750	.1375
		$N(ODF, \sqrt{2\mathrm{ODF}})$	.0250	.0550	.0875	.1425
		$N(EDF+2, \sqrt{2(\mathrm{EDF}+2)})$	.0125	.0300	.0550	.0950
		$N(ODF+2, \sqrt{2(\mathrm{ODF}+2)})$	.0150	.0300	.0550	.0975
		$\chi^2_{\mathrm{EDF}+2}$	.0075	.0200	.0400	.0900
		$\chi^2_{\mathrm{ODF}+2}$	.0075	.0200	.0475	.0950
$H_0^{(2)}$	$r_{\mathcal{K}}\lambda_n(\mathrm{h}^*)$	$N(EDF, \sqrt{2\mathrm{EDF}})$	.0450	.0725	.1150	.1700
		$N(ODF, \sqrt{2\mathrm{ODF}})$	.0450	.0750	.1175	.1700
		$N(EDF+2, \sqrt{2(\mathrm{EDF}+2)})$	.0200	.0375	.0575	.0950
		$N(ODF+2, \sqrt{2(\mathrm{ODF}+2)})$	.0200	.0375	.0575	.0925
		$\chi^2_{\mathrm{EDF}+2}$	.0075	.0200	.0450	.0850
		$\chi^2_{\mathrm{ODF}+2}$	.0050	.0200	.0450	.0875

with this modifications have yielded results comparable with those in the uniform design. Third, if f is other than uniformity but known, then the higher-order approximations given in (A.9) and (A.17) may perform better than those in (A.10) and (A.18). On the other hand, even a decent approximation, as simple as the first-order empirical formulas, is likely to be satisfactory. Fourth, some robust procedures may be applied before the empirical formulas are used; for example, remove a few design observations at the edges of the sample space. This preprocessing of the raw data before the analyses are performed is also common in other contexts of statistical applications.

APPENDIX: PROOFS OF MAIN RESULTS

Condition (A)

(A1) The design variable X has a bounded support Ω; the marginal density f of X is Lipschitz continuous and bounded away from 0.

(A2) $m(x)$ has the continuous $p+1$th derivative in Ω.

(A3) The kernel function $K(t)$ is a symmetric probability density function with bounded support and is Lipschitz continuous.

(A4) $0 < E(\varepsilon^4) < \infty$.

Condition (A*) is similar to condition (A), except that (A1) is replaced by (A1*).

(A1*) The design observations $X_i = x_i$, $i = 1, \ldots, n$, are generated by $x_i = F^{-1}((i-.5)/n)$, where F has a probability density function f with a bounded support Ω; f is Lipschitz continuous and bounded away from 0.

Condition (B)

The design points x_i, $i = 1, \ldots, n$, are generated from a continuous and strictly positive density f, on a finite interval [0, 1] without loss of generality, through the relation $\int_0^{x_i} f(x)\,dx = (i-.5)/n$.

Condition (C)

(C1) The covariate U has a bounded support Ω; the marginal density f_U of U is Lipschitz continuous and bounded away from 0.

(C2) $a_j(u), j = 1, \ldots, d$, has the continuous $p+1$th derivative in Ω.

(C3) The kernel function $K(t)$ is a symmetric probability density function with bounded support and is Lipschitz continuous.

(C4) $0 < E(\varepsilon^4) < \infty$.

(C5) The matrix $E(\mathrm{X}\mathrm{X}^T|U=u)$ is positive definite for each $u \in \Omega$, and each entry is Lipschitz continuous.

Condition (C*) is similar to condition (C), except that (C1) is replaced by (C1*).

(C1*) The covariate observations $U_i = u_i$ are generated by $u_i = F_U^{-1}((i-.5)/n)$, where F_U has a probability density function f_U with a bounded support Ω; f_U is Lipschitz continuous and bounded away from 0.

Proof of (14)

This proof begins with Lemma A.1.

Lemma A.1. 1. For a nonnegative kernel K satisfying $K(0) = \sup_x K(x)$, $\sum_{j=1}^n \{\mathbf{S}_h(i,j)\}^2 \le \mathbf{S}_h(i,i)$ for $i = 1, \ldots, n$, and thus $0 \le \mathbf{S}_h(i,i) \le 1$, and $|\mathbf{S}_h(i,j)| \le 1$ for $i \ne j$.

2. For any integer $k = 0, 1, \ldots, p$, $\mathbf{S}_h(X_1^k, \ldots, X_n^k)^T = (X_1^k, \ldots, X_n^k)^T$. Thus for any matrix P whose column space is generated by the vectors $(X_1^k, \ldots, X_n^k)^T$, $k = 0, 1, \ldots, p$, it follows that $\mathbf{S}_h P = P$ and $(\mathbf{S}_h^T + \mathbf{S}_h - \mathbf{S}_h^T\mathbf{S}_h)^\ell P = P$ for integers $\ell \ge 0$.

Proof. To show part 1, it is seen that

$$\begin{aligned}\{\mathbf{S}_h(i,j)\}^2 &= \boldsymbol{e}_1^T\{S_n(X_i)\}^{-1}\{1, (X_j - X_i), \ldots, (X_j - X_i)^p\}^T \\ &\quad \times \{1, (X_j - X_i), \ldots, (X_j - X_i)^p\} \\ &\quad \times \{S_n(X_i)\}^{-1}\boldsymbol{e}_1 K_h^2(X_j - X_i) \\ &\le \boldsymbol{e}_1^T\{S_n(X_i)\}^{-1}\{1, (X_j - X_i), \ldots, (X_j - X_i)^p\}^T \\ &\quad \times \{1, (X_j - X_i), \ldots, (X_j - X_i)^p\} \\ &\quad \times K_h(X_j - X_i)\{S_n(X_i)\}^{-1}\boldsymbol{e}_1 K_h(0),\end{aligned}$$

and thus $\sum_{j=1}^n \{\mathbf{S}_h(i,j)\}^2 \le \boldsymbol{e}_1^T\{S_n(X_i)\}^{-1}\boldsymbol{e}_1 K_h(0) = \mathbf{S}_h(i,i)$.

To show part 2, recall that $\sum_{j=1}^n (X_j - x)^k W_0^n(x, \frac{X_j - x}{h}) = I(k=0)$ holds for any integer $k = 0, 1, \ldots, p$ (Fan and Gijbels 1996, p. 103). Applying this result and the binomial expansion leads to

$$\begin{aligned}\sum_{j=1}^n X_j^k W_0^n\Big(x, \frac{X_j - x}{h}\Big) &= \sum_{j=1}^n \sum_{\ell=0}^k \binom{k}{\ell}(X_j - x)^\ell x^{k-\ell} W_0^n\Big(x, \frac{X_j - x}{h}\Big) \\ &= \sum_{\ell=0}^k \binom{k}{\ell} x^{k-\ell} \sum_{j=1}^n (X_j - x)^\ell W_0^n\Big(x, \frac{X_j - x}{h}\Big) \\ &= x^k.\end{aligned}$$

Thus for each $i = 1, \ldots, n$, $\sum_{j=1}^n \mathbf{S}_h(i,j)X_j^k = \sum_{j=1}^n X_j^k W_0^n(X_i, \frac{X_j - X_i}{h}) = X_i^k$. Lemma A.1 finishes.

To show (14), observe from the first part of Lemma A.1 that

$$\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h) = \sum_{i=1}^n \sum_{j=1}^n \{\mathbf{S}_h(i,j)\}^2 \le \mathrm{tr}(\mathbf{S}_h) \tag{A.1}$$

holds under the kernel-mode condition. The second part of Lemma A.1 asserts that 1 is an eigenvalue of $\mathbf{S}_h$ corresponding to a number $p+1$ of distinct eigenvectors. Let $\lambda_i(\mathbf{S}_h)$, $i = 1, \ldots, n$, denote all of the eigenvalues of $\mathbf{S}_h$. Then Schur's inequality (Marcus and Minc 1964, p. 142) says that $\mathrm{tr}(\mathbf{S}_h^T\mathbf{S}_h) \ge \sum_{i=1}^n |\lambda_i(\mathbf{S}_h)|^2 \ge p+1$. Henceforth, this inequality, together with (A.1) and $\mathrm{tr}\{(\mathbf{I}_n - \mathbf{S}_h)^T(\mathbf{I}_n - \mathbf{S}_h)\} > 0$, implies the expected results.

Proof of Theorem 1

For each $i = 1, \ldots, n$, because the ith row of $\mathbf{X}(X_i)$ is $\boldsymbol{e}_1^T$, and the ith diagonal entry of $\mathbf{W}(X_i)$ is $K_h(0)$, $S_n(X_i) = K_h(0)\boldsymbol{e}_1\boldsymbol{e}_1^T + A_i$ can be rewritten, where the (ℓ, r)th entry of the matrix A_i is given by

$$A_i(\ell, r) = \sum_{k:\,k\ne i} (X_k - X_i)^{\ell+r-2} K_h(X_k - X_i), \qquad \ell, r = 1, \ldots, p+1. \tag{A.2}$$

Set $d_{ij} = \frac{X_j - X_i}{h}$. Thus the use of $\{S_n(X_i)\}^{-1} = A_i^{-1} - \frac{A_i^{-1}\boldsymbol{e}_1\boldsymbol{e}_1^T A_i^{-1}}{h/K(0) + \boldsymbol{e}_1^T A_i^{-1}\boldsymbol{e}_1}$, together with (10) and (11), implies

$$\mathbf{S}_h(i,j) = \frac{B_{ij}}{h/K(0) + B_{ii}}, \tag{A.3}$$

where

$$B_{ij} = \boldsymbol{e}_1^T A_i^{-1} H(1, d_{ij}, \ldots, d_{ij}^p)^T K(d_{ij})/K(0), \qquad i,j = 1, \ldots, n. \tag{A.4}$$

Writing $f_i = f(X_i)$, $f_i' = f'(X_i)$, $f_i'' = f''(X_i)$, $g_{i1} = f_i'/f_i$, and $g_{i2} = 2^{-1}(f_i''/f_i)$, then the higher-order Taylor expansion of $A_i(\ell, r)$, $\ell, r = 1, \ldots, p+1$, leads to

$$A_i = (n-1)f_i H(S + hg_{i1}\widetilde{S} + h^2 g_{i2}\underline{S})H\{1 + o_P(1)\}, \tag{A.5}$$

where $o_P(1)$ is uniform in $i = 1, \ldots, n$ (abbreviated as $\cup i$), $\widetilde{S} = (\mu_{\ell+r-1})_{1\le\ell,\, r\le p+1}$, and $\underline{S} = (\mu_{\ell+r})_{1\le\ell,\, r\le p+1}$. Denote $M_1 = S^{-1}\widetilde{S}S^{-1}$, $M_2 = S^{-1}\widetilde{S}S^{-1}\widetilde{S}S^{-1}$, and $M_3 = S^{-1}\underline{S}S^{-1}$; then

$$A_i^{-1} = \frac{1}{(n-1)f_i} H^{-1}\{S^{-1} - hg_{i1}M_1 + h^2(g_{i1}^2 M_2 - g_{i2}M_3)\}H^{-1} \times \{1 + o_P(1)\}, \quad (A.6)$$

$\cup i$. Set $\widetilde{\mathcal{K}}(t) = \boldsymbol{e}_1^T M_1(1, t, \ldots, t^p)^T K(t)$, $\overline{\mathcal{K}}(t) = \boldsymbol{e}_1^T M_2(1, t, \ldots, t^p)^T K(t)$, and $\underline{\mathcal{K}}(t) = \boldsymbol{e}_1^T M_3(1, t, \ldots, t^p)^T K(t)$. Equations (A.4) and (A.6), together with the fact that $\widetilde{\mathcal{K}}(0) = 0$, give

$$B_{ii} = \frac{1}{(n-1)K(0)f_i}\{\mathcal{K}(0) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(0)\} \times \{1 + o_P(1)\}, \quad (A.7)$$

$\cup i$, an immediate consequence of which is

$$\mathbf{S}_h(i,i) = \frac{\mathcal{K}(0) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(0)}{(n-1)hf_i + \{\mathcal{K}(0) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(0)\}} \times \{1 + o_P(1)\}, \quad (A.8)$$

$\cup i$, and therefore,

$$\operatorname{tr}(\mathbf{S}_h) = \sum_{i=1}^n \frac{\mathcal{K}(0) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(0)}{(n-1)hf_i + \{\mathcal{K}(0) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(0)\}} \times \{1 + o_P(1)\} \quad (A.9)$$

$$= \frac{\mathcal{K}(0)}{(n-1)h}\left(\sum_{i=1}^n f_i^{-1}\right)\{1 + o_P(1)\}. \quad (A.10)$$

This finishes the proof of the first part.

For $\operatorname{tr}(\mathbf{S}_h^T\mathbf{S}_h)$, the fact is that

$$\operatorname{tr}(\mathbf{S}_h^T\mathbf{S}_h) = \sum_{i=1}^n\{\mathbf{S}_h(i,i)\}^2 + \sum_{1\le i\ne j\le n}\sum\{\mathbf{S}_h(i,j)\}^2. \quad (A.11)$$

For the first additive term of (A.11), from (A.8),

$$\sum_{i=1}^n\{\mathbf{S}_h(i,i)\}^2 = \sum_{i=1}^n \frac{\mathcal{K}^2(0)}{(n-1)^2h^2f_i^2}\{1 + o_P(1)\} = O_P\{(nh^2)^{-1}\}. \quad (A.12)$$

Consider the second additive term of (A.11). According to (A.4) and (A.6), it holds that

$$B_{ij} = \frac{1}{(n-1)K(0)f_i}\{\mathcal{K}(d_{ij}) - hg_{i1}\widetilde{\mathcal{K}}(d_{ij}) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(d_{ij})\}\{1 + o_P(1)\}, \quad (A.13)$$

where the term $o_P(1)$ is independent of j and uniform in $i = 1, \ldots, n$ (abbreviated as $\mathtt{I}j\cup i$). It can be deduced that from (A.3), (A.7), and (A.13),

$$\mathbf{S}_h(i,j) = \frac{\mathcal{K}(d_{ij}) - hg_{i1}\widetilde{\mathcal{K}}(d_{ij}) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(d_{ij})}{(n-1)f_i + \{\mathcal{K}(0) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(0)\}} \times \{1 + o_P(1)\}, \quad (A.14)$$

$\mathtt{I}j\cup i$. For the numerator of (A.14),

$$\{\mathcal{K}(d_{ij}) - hg_{i1}\widetilde{\mathcal{K}}(d_{ij}) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(d_{ij})\}^2 = \mathcal{K}^2(d_{ij}) - 2hg_{i1}\mathcal{K}\widetilde{\mathcal{K}}(d_{ij}) + h^2\{g_{i1}^2(2\mathcal{K}\overline{\mathcal{K}} + \widetilde{\mathcal{K}}^2) - 2g_{i2}\mathcal{K}\underline{\mathcal{K}}\}(d_{ij})\{1 + o_P(1)\},$$

in which this $o_P(1)$ can also be $\mathtt{I}j\cup i$, under the smoothness assumption on f. Similar to the derivation of (A.5),

$$\sum_{j:j\ne i}(1, d_{ij}, \ldots, d_{ij}^p)^T(1, d_{ij}, \ldots, d_{ij}^p)K^2(d_{ij}) = (n-1)hf_i(S^* + hg_{i1}\widetilde{S^*} + h^2g_{i2}\underline{S^*})\{1 + o_P(1)\}, \quad (A.15)$$

$\cup i$, where $S^* = (\nu_{i+j-2})_{1\le i,j\le p+1}$, $\widetilde{S^*} = (\nu_{i+j-1})_{1\le i,j\le p+1}$, and $\underline{S^*} = (\nu_{i+j})_{1\le i,j\le p+1}$. Analogously, using (17), $\boldsymbol{e}_1^TS^{-1}\widetilde{S^*}S^{-1}\boldsymbol{e}_1 = 0$, and $\boldsymbol{e}_1^TS^{-1}S^*M_1\boldsymbol{e}_1 = \mathcal{K} * \widetilde{\mathcal{K}}(0) = 0$, it can be deduced that

$$\sum_{j:j\ne i}\mathcal{K}^2(d_{ij}) = (n-1)hf_i\{\mathcal{K} * \mathcal{K}(0) + h^2g_{i2}\boldsymbol{e}_1^TS^{-1}\underline{S^*}S^{-1}\boldsymbol{e}_1\} \times \{1 + o_P(1)\},$$

$$\sum_{j:j\ne i}\mathcal{K}\widetilde{\mathcal{K}}(d_{ij}) = (n-1)hf_i\{hg_{i1}\boldsymbol{e}_1^TS^{-1}\widetilde{S^*}M_1\boldsymbol{e}_1\}\{1 + o_P(1)\},$$

$$\sum_{j:j\ne i}\widetilde{\mathcal{K}}^2(d_{ij}) = (n-1)hf_i\widetilde{\mathcal{K}} * \widetilde{\mathcal{K}}(0)\{1 + o_P(1)\},$$

$$\sum_{j:j\ne i}\mathcal{K}\overline{\mathcal{K}}(d_{ij}) = (n-1)hf_i\mathcal{K} * \overline{\mathcal{K}}(0)\{1 + o_P(1)\},$$

and

$$\sum_{j:j\ne i}\mathcal{K}\underline{\mathcal{K}}(d_{ij}) = (n-1)hf_i\mathcal{K} * \underline{\mathcal{K}}(0)\{1 + o_P(1)\},$$

$\cup i$. Hence, uniformly in i,

$$\sum_{j:j\ne i}\{\mathcal{K}(d_{ij}) - hg_{i1}\widetilde{\mathcal{K}}(d_{ij}) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(d_{ij})\}^2 = (n-1)hf_i[\mathcal{K} * \mathcal{K}(0) + h^2\{g_{i1}^2\ell_1(K) - g_{i2}\ell_2(K)\}] \times \{1 + o_P(1)\}, \quad (A.16)$$

again due to the fact $\mathcal{K} * \widetilde{\mathcal{K}}(0) = 0$, where $\ell_1(K) = 2\mathcal{K} * \overline{\mathcal{K}}(0) + \widetilde{\mathcal{K}} * \widetilde{\mathcal{K}}(0) - 2\boldsymbol{e}_1^TS^{-1}\widetilde{S^*}M_1\boldsymbol{e}_1$ and $\ell_2(K) = 2\mathcal{K} * \underline{\mathcal{K}}(0) - \boldsymbol{e}_1^TS^{-1}\underline{S^*}S^{-1}\boldsymbol{e}_1$. Applying (A.14) and (A.16), it can be observed that

$$\sum_{i=1}^n\sum_{j:j\ne i}\{\mathbf{S}_h(i,j)\}^2 = \sum_{i=1}^n \frac{[\mathcal{K} * \mathcal{K}(0) + h^2\{g_{i1}^2\ell_1(K) - g_{i2}\ell_2(K)\}](n-1)hf_i}{[(n-1)hf_i + \{\mathcal{K}(0) + h^2(g_{i1}^2\overline{\mathcal{K}} - g_{i2}\underline{\mathcal{K}})(0)\}]^2} \times \{1 + o_P(1)\} \quad (A.17)$$

$$= \frac{\mathcal{K} * \mathcal{K}(0)}{(n-1)h}\left(\sum_{i=1}^n f_i^{-1}\right)\{1 + o_P(1)\}. \quad (A.18)$$

This, together with (A.11) and (A.12), finishes the proof of the second part. Proof of the third part is trivial from the first two parts.

For fixed designs, the proofs follow arguments analogous to the foregoing. For example, (A.5) and (A.15) follow from the theory of Riemann sums, with $o_P(1)$ replaced by $o(1)$.

Proof of Theorem 2

The proof of Theorem 2 depends on two lemmas. The proof of Lemma A.2 was given by Ramil Novo and González Manteiga (2000).

Lemma A.2. Let $\boldsymbol{K}(x) = (2\pi)^{-1}\int_{-\infty}^{+\infty}(1 + t^{2q})^{-1}\exp(-itx)\,dt$, with $q = 1, 2, \ldots$. Denote by $\overbrace{\boldsymbol{K} * \cdots * \boldsymbol{K}}^{r}(x)$ the r-times convolution product of $\boldsymbol{K}(x)$. Then

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty}(1 + t^{2q})^{-r}\,dt = \overbrace{\boldsymbol{K} * \cdots * \boldsymbol{K}}^{r}(0), \qquad r = 1, 2, \ldots.$$

In particular, $(2\pi)^{-1}\int_{-\infty}^{+\infty}(1+t^{2q})^{-2\ell}\,dt = \int(\overbrace{\boldsymbol{K}*\cdots*\boldsymbol{K}}^{\ell})^2(x)\,dx$ for $\ell = 1, 2, \ldots$.

Lemma A.3. Let $\boldsymbol{K}(x) = (2\pi)^{-1}\int_{-\infty}^{+\infty}(1+t^{2q})^{-1}\exp(-itx)\,dt$. Set $c(f) = \int_0^1 f(t)^{1/(2q)}\,dt$. Then for $q \geq 2$, as $n \to \infty$, $\lambda \to 0$, and $n\lambda \to \infty$, it holds that

$$\operatorname{tr}(\mathbf{S}_\lambda^r) = q + \lambda^{-1/(2q)}c(f)(2\pi)^{-1}\int_{-\infty}^{+\infty}(1+t^{2q})^{-r}\,dt\{1+o(1)\}, \quad r \geq 1. \tag{A.19}$$

Proof. From (25), first observe that $\operatorname{tr}(\mathbf{S}_\lambda^r) = q + \sum_{j=q+1}^n(1+\lambda\gamma_{jn})^{-r}$, for integers $r \geq 1$. Speckman (1981) showed that if $q \geq 2$, then $\gamma_{jn} = j^{2q}c_0\{1+o(1)\}$ for $1 \leq j \leq n$, where $c_0 = \pi^{2q}\{\int_0^1 f(t)^{1/(2q)}\,dt\}^{-2q}$, and the $o(1)$ term is uniform for $j = o(n^{2/5})$. Combining this result, it can be deduced that

$$\sum_{q+1\leq j\leq n^{3/(4q)}}(1+\lambda\gamma_{jn})^{-r} = \sum_{q+1\leq j\leq n^{3/(4q)}}(1+\lambda c_0 j^{2q})^{-r}\{1+o(1)\} = (\lambda c_0)^{-1/(2q)}\int_0^\infty \frac{dt}{(1+t^{2q})^r}\{1+o(1)\}.$$

On the other hand, the sequence $\{\gamma_{jn}\}_{j=1}^n$ is nondecreasing, and therefore $\gamma_{jn} \geq O(n^{3/2})$ for $j \geq n^{3/(4q)}$, so that

$$\sum_{n^{3/(4q)}<j\leq n}(1+\lambda\gamma_{jn})^{-r} \leq O\{n(n^{3/2}\lambda)^{-r}\}. \tag{A.20}$$

The upper bound in (A.20) is thus $o\{\lambda^{-1/(2q)}\}$ if $r \geq 2$. Hence $\operatorname{tr}(\mathbf{S}_\lambda^r) = q + \lambda^{-1/(2q)}c(f)(2\pi)^{-1}\int_{-\infty}^{+\infty}(1+t^{2q})^{-r}\,dt\{1+o(1)\}$ for $r \geq 2$. For $r = 1$, a similar expression for $\operatorname{tr}(\mathbf{S}_\lambda)$ was given by Eubank (1988, p. 327). This finishes the proof of Lemma A.3.

To obtain the asymptotic representations of $\operatorname{tr}(\mathbf{S}_\lambda)$ and $\operatorname{tr}(\mathbf{S}_\lambda^2)$, in terms of Silverman's kernel function $\boldsymbol{K}$, a direct application of Lemma A.2 to Lemma A.3 leads to the desired conclusions.

Proof of Theorem 3

First, according to the definition (41), $\widetilde{S}_n(U_i) = K_h(0)(\boldsymbol{e}_1\otimes \mathtt{X}_i)(\boldsymbol{e}_1^T\otimes \mathtt{X}_i^T) + A_i$, where $A_i = \sum_{k:\,k\neq i}\mathbf{Z}_k(U_i)\{\mathbf{Z}_k(U_i)\}^T K_h(U_k - U_i)$. Set $d_{ij} = \frac{U_j - U_i}{h}$. Thus the use of

$$\{\widetilde{S}_n(U_i)\}^{-1} = A_i^{-1} - \frac{A_i^{-1}(\boldsymbol{e}_1\otimes \mathtt{X}_i)(\boldsymbol{e}_1^T\otimes \mathtt{X}_i^T)A_i^{-1}}{h/K(0) + (\boldsymbol{e}_1^T\otimes \mathtt{X}_i^T)A_i^{-1}(\boldsymbol{e}_1\otimes \mathtt{X}_i)},$$

together with (44) and (45), implies that

$$\widetilde{\mathbf{S}}_h(i,j) = \frac{B_{ij}}{h/K(0) + B_{ii}}, \tag{A.21}$$

where

$$B_{ij} = (\boldsymbol{e}_1^T\otimes \mathtt{X}_i^T)A_i^{-1}(H\otimes \mathbf{I}_d)\{(1, d_{ij}, \ldots, d_{ij}^p)^T\otimes \mathtt{X}_j\}K(d_{ij})/K(0), \quad i,j = 1,\ldots,n. \tag{A.22}$$

It can be shown via some standard arguments that

$$A_i = (n-1)\{(HSH)\otimes\Gamma(U_i)\}\{1+o_P(1)\} \tag{A.23}$$

and

$$A_i^{-1} = (n-1)^{-1}[(H^{-1}S^{-1}H^{-1})\otimes\{\Gamma(U_i)\}^{-1}]\{1+o_P(1)\}, \tag{A.24}$$

$\mathtt{U}i$, where $\Gamma(u) = f_U(u)E(\mathtt{X}\mathtt{X}^T|U=u)$. This results in

$$B_{ii} = \frac{\mathcal{K}(0)}{(n-1)K(0)}\mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i\{1+o_P(1)\}, \tag{A.25}$$

$$\widetilde{\mathbf{S}}_h(i,i) = \frac{\mathcal{K}(0)\mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i}{(n-1)h + \mathcal{K}(0)\mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i}\{1+o_P(1)\}, \tag{A.26}$$

$\mathtt{U}i$, and

$$\operatorname{tr}(\widetilde{\mathbf{S}}_h) = \sum_{i=1}^n\widetilde{\mathbf{S}}_h(i,i) = \frac{\mathcal{K}(0)}{(n-1)h}\Big(\sum_{i=1}^n \mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i\Big)\{1+o_P(1)\}, \tag{A.27}$$

which, combined with $E\{\mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i\} = d|\Omega|$, finishes the proof of the first part.

Now consider $\operatorname{tr}(\widetilde{\mathbf{S}}_h^T\widetilde{\mathbf{S}}_h)$. According to (A.22) and (A.24), it can be deduced that

$$B_{ij} = \frac{\mathcal{K}(d_{ij})}{(n-1)K(0)}\mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i\{1+o_P(1)\} \tag{A.28}$$

and

$$\widetilde{\mathbf{S}}_h(i,j) = \frac{\mathcal{K}(d_{ij})\mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_j}{(n-1)h + \mathcal{K}(0)\mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i}\{1+o_P(1)\}, \tag{A.29}$$

$\mathtt{I}j\mathtt{U}i$. Verification of

$$\sum_{j:\,j\neq i}\mathcal{K}^2(d_{ij})\mathtt{X}_j\mathtt{X}_j^T = (n-1)h\Gamma(U_i)\mathcal{K}*\mathcal{K}(0)\{1+o_P(1)\},$$

$\mathtt{U}i$, implies that

$$\sum_{j:\,j\neq i}\{\widetilde{\mathbf{S}}_h(i,j)\}^2 = \frac{\mathcal{K}*\mathcal{K}(0)(n-1)h\mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i}{[(n-1)h + \mathcal{K}(0)\mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i]^2}\{1+o_P(1)\},$$

$\mathtt{U}i$, and thus

$$\operatorname{tr}(\widetilde{\mathbf{S}}_h^T\widetilde{\mathbf{S}}_h) = \frac{\mathcal{K}*\mathcal{K}(0)}{(n-1)h}\Big(\sum_{i=1}^n \mathtt{X}_i^T\{\Gamma(U_i)\}^{-1}\mathtt{X}_i\Big)\{1+o_P(1)\}.$$

This completes the proof of the second part.

Proof of Expressions (5) and (54)

It is necessary to show only (54), which includes (5) as a special case. To ease the notation, consider first the local linear fit, the conclusion of which [(54)] can be extended straightforwardly to that of higher-degree local polynomial fit. Recall from (43) that $\hat{m}_h(\cdot,\cdot)$ represents the local polynomial estimates of the regression function based on the raw data $\{(U_\ell, \mathtt{X}_\ell, Y_\ell)_{\ell=1}^n\}$.

For $1 \leq j \leq d$, let $\hat{a}_{j,-i}(\cdot)$ and $\hat{a}_{j,-i}^{[1]}(\cdot)$ denote the local linear estimates of the varying-coefficient function $a_j(\cdot)$ and its derivative, based on the data $\{(U_\ell, \mathtt{X}_\ell, Y_\ell)_{\ell=1}^n\}$ with the ith pair $(U_i, \mathtt{X}_i, Y_i)$ removed. Let $\hat{m}_{h,-i}(U_i, \mathtt{X}_i) = \sum_{j=1}^d \hat{a}_{j,-i}(U_i)X_{ji}$ represent the resulting estimated response at $(U_i, \mathtt{X}_i)$. Analogously, let $\hat{a}_{j,*i}(\cdot)$ and $\hat{a}_{j,*i}^{[1]}(\cdot)$ denote similar quantities as before, except based on the data $\{(U_\ell, \mathtt{X}_\ell, Y_\ell)_{\ell=1}^n\}$, with the ith response Y_i replaced by $\hat{m}_{h,-i}(U_i, \mathtt{X}_i)$. In this case, let $\hat{m}_{h,*i}(U_i, \mathtt{X}_i) = \sum_{j=1}^d \hat{a}_{j,*i}(U_i)X_{ji}$ represent the resulting estimated response at $(U_i, \mathtt{X}_i)$.

Put $K_{ji} = K_h(U_j - U_i)$ and $g_{\ell,i}(a_0, a_1) = \sum_{j=1}^d\{a_0 + (U_\ell - U_i)a_1\}X_{j\ell}$. For each $i = 1, \ldots, n$,

$$\begin{aligned}\sum_{1\leq\ell\leq n:\,\ell\neq i}&\big[Y_\ell - g_{\ell,i}(\hat{a}_{j,-i}(U_\ell), \hat{a}_{j,-i}^{[1]}(U_\ell))\big]^2K_{\ell i}\\ &= \min_{\{a_j\},\,\{a_j^{[1]}\}}\sum_{1\leq\ell\leq n:\,\ell\neq i}\big[Y_\ell - g_{\ell,i}(a_j, a_j^{[1]})\big]^2K_{\ell i}\\ &\leq \sum_{1\leq\ell\leq n:\,\ell\neq i}\big[Y_\ell - g_{\ell,i}(\hat{a}_{j,*i}(U_i), \hat{a}_{j,*i}^{[1]}(U_i))\big]^2K_{\ell i}.\end{aligned} \tag{A.30}$$

In contrast,

$$\begin{aligned}
&\sum_{1\le\ell\le n:\,\ell\neq i}\left[Y_\ell - g_{\ell,i}(\hat{a}_{j,-i}(U_\ell), \hat{a}^{[1]}_{j,-i}(U_\ell))\right]^2 K_{\ell i}\\
&\quad\ge \min_{\{a_j\},\,\{a^{[1]}_j\}}\left\{\sum_{1\le\ell\le n:\,\ell\neq i}\left[Y_\ell - g_{\ell,i}(a_j, a^{[1]}_j)\right]^2 K_{\ell i}\right.\\
&\qquad\qquad\left.+\left[\hat{m}_{h,-i}(U_i, \mathrm{X}_i) - g_{i,i}(a_j, a^{[1]}_j)\right]^2 K_{ii}\right\}\\
&\quad=\sum_{1\le\ell\le n:\,\ell\neq i}\left[Y_\ell - g_{\ell,i}(\hat{a}_{j,*i}(U_\ell), \hat{a}^{[1]}_{j,*i}(U_\ell))\right]^2 K_{\ell i}\\
&\qquad+\{\hat{m}_{h,-i}(U_i, \mathrm{X}_i) - \hat{m}_{h,*i}(U_i, \mathrm{X}_i)\}^2 K_{ii}. \qquad \text{(A.31)}
\end{aligned}$$

Inequalities (A.30) and (A.31) imply that $\{\hat{m}_{h,-i}(U_i, \mathrm{X}_i) - \hat{m}_{h,*i}(U_i, \mathrm{X}_i)\}^2 K_{ii} = 0$ for each $i = 1, \ldots, n$, and thus, according to $K(0) > 0$, indicate $\hat{m}_{h,-i}(U_i, \mathrm{X}_i) = \hat{m}_{h,*i}(U_i, \mathrm{X}_i)$. This equality in turn yields

$$\begin{aligned}
\hat{m}_{h,-i}(U_i, \mathrm{X}_i) &= \hat{m}_{h,*i}(U_i, \mathrm{X}_i)\\
&= \sum_{1\le j\le n:\,j\neq i}\widetilde{\mathbf{S}}_h(i,j)Y_j + \widetilde{\mathbf{S}}_h(i,i)\hat{m}_{h,-i}(U_i, \mathrm{X}_i)\\
&= \sum_{j=1}^{n}\widetilde{\mathbf{S}}_h(i,j)Y_j + \widetilde{\mathbf{S}}_h(i,i)\{\hat{m}_{h,-i}(U_i, \mathrm{X}_i) - Y_i\}\\
&= \hat{m}_h(U_i, \mathrm{X}_i) + \widetilde{\mathbf{S}}_h(i,i)\{\hat{m}_{h,-i}(U_i, \mathrm{X}_i) - Y_i\}.
\end{aligned}$$

Hence $Y_i - \hat{m}_{h,-i}(U_i, \mathrm{X}_i) = Y_i - \hat{m}_h(U_i, \mathrm{X}_i) + \widetilde{\mathbf{S}}_h(i,i)\{Y_i - \hat{m}_{h,-i}\times(U_i, \mathrm{X}_i)\}$, which leads to the stated result, $Y_i - \hat{m}_{h,-i}(U_i, \mathrm{X}_i) = \{Y_i - \hat{m}_h(U_i, \mathrm{X}_i)\}/\{1 - \widetilde{\mathbf{S}}_h(i,i)\}$, for $i = 1, \ldots, n$.

[Received May 2002. Revised January 2003.]

REFERENCES

Allen, D. M. (1974), "The Relationship Between Variable and Data Augmentation and a Method of Prediction," *Technometrics*, 16, 125–127.

Buckley, M. J., Eagleson, G. K., and Silverman, B. W. (1988), "The Estimation of Residual Variance in Nonparametric Regression," *Biometrika*, 75, 189–200.

Cleveland, W., and Devlin, S. (1988), "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, 83, 596–610.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992), "Local Regression Models," in *Statistical Models in S*, eds. J. M. Chambers and T. J. Hastie, Pacific Grove, CA: Wadsworth/Brooks Cole, pp. 309–376.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, 31, 377–403.

Cummins, D. J., Filloon, T. G., and Nychka, D. (2001), "Confidence Intervals for Nonparametric Curve Estimates: Toward More Uniform Pointwise Coverage," *Journal of the American Statistical Association*, 96, 233–246.

Demmler, A., and Reinsch, C. (1975), "Oscillation Matrices With Spline Functions," *Numerische Mathematik*, 24, 375–382.

Eubank, R. L. (1984), "The Hat Matrix for Smoothing Spline," *Statistics and Probability Letters*, 2, 9–14.

——— (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Fan, J., and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society*, Ser. B, 57, 371–394.

——— (1996), *Local Polynomial Modeling and Its Applications*, London: Chapman & Hall.

Fan, J., Zhang, C. M., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193.

Fan, J., and Zhang, W. Y. (1999), "Statistical Estimation in Varying Coefficient Models," *The Annals of Statistics*, 27, 1491–1518.

Gasser, T., Kneip, A., and Köhler, W. (1991), "A Flexible and Fast Method for Automatic Smoothing," *Journal of the American Statistical Association*, 86, 643–652.

Green, P. J., and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*, London: Chapman & Hall.

Härdle, W., Hall, P., and Marron, J. S. (1992), "Regression Smoothing Parameters That Are Not Far From Their Optimum," *Journal of the American Statistical Association*, 87, 227–233.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman & Hall.

——— (1993), "Varying-Coefficient Models" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 757–796.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. P. (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data," *Biometrika*, 85, 809–822.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society*, Ser. B, 60, 271–293.

Lee, T. C. M., and Solo, V. (1999), "Bandwidth Selection for Local Linear Regression: A Simulation Study," *Computational Statistics*, 14, 515–532.

Li, K.-C. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross-Validation," *The Annals of Statistics*, 13, 1352–1377.

——— (1986), "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression With Application to Spline Smoothing," *The Annals of Statistics*, 14, 1101–1112.

Marcus, M., and Minc, H. (1964), *A Survey of Matrix Theory and Matrix Inequalities*, Boston: Allyn and Bacon.

Müller, H. G. (1987), "Weighted Local Regression and Kernel Methods for Nonparametric Curve Fitting," *Journal of the American Statistical Association*, 82, 231–238.

——— (1988), *Nonparametric Regression Analysis of Longitudinal Data* (Lecture Notes in Statistics, Vol. 46), Berlin: Springer-Verlag.

Müller, H. G., and Prewitt, K. A. (1993), "Multiparameter Bandwidth Processes and Adaptive Surface Smoothing," *Journal of Multivariate Analysis*, 47, 1–21.

Ramil Novo, L. A., and González Manteiga, W. (2000), "F-Tests and Regression ANOVA Based on Smoothing Spline Estimators," *Statistica Sinica*, 10, 819–837.

Ramsay, J. O., and Silverman, B. W. (1997), *Functional Data Analysis*, New York: Springer-Verlag.

Reinsch, C. (1967), "Smoothing by Spline Functions," *Numerische Mathematik*, 10, 177–183.

Rice, J. (1984), "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics*, 20, 712–736.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Ruppert, D., Wand, M. P., Holst, U., and Hössjer, O. (1997), "Local Polynomial Variance-Function Estimation," *Technometrics*, 39, 262–273.

Silverman, B. W. (1984), "Spline Smoothing: the Equivalent Variable Kernel Method," *The Annals of Statistics*, 19, 898–916.

Speckman, P. (1981), "The Asymptotic Integrated Mean Squared Error for Smoothing Noisy Data by Spline Functions," unpublished manuscript, University of Missouri, Dept. of Statistics.

Stanton, R. (1997), "A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk," *Journal of Finance*, 52, 1973–2002.

Stone, C. (1984), "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *The Annals of Statistics*, 12, 1285–1297.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman & Hall.

Wong, W. H. (1983), "On the Consistency of Cross-Validation in Kernel Nonparametric Regression," *The Annals of Statistics*, 11, 1136–1141.

Ye, J. M. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120–131.