The Annals of Statistics
1998, Vol. 26, No. 1, 315–334

ASYMPTOTIC COMPARISON OF (PARTIAL) CROSS-VALIDATION, GCV AND RANDOMIZED GCV IN NONPARAMETRIC REGRESSION

BY DIDIER A. GIRARD

CNRS and Université Joseph Fourier

When using nonparametric estimates of the mean curve, surface or image underlying noisy observations, the selection of "smoothing parameters" is generally crucial. This paper gives a theoretical comparison of the performances of generalized cross-validation (GCV) and of its fast randomized version (RGCV), as selection criteria. This is mainly done by studying the asymptotic distribution of the excess error for each selector, that is, the difference between the (data-driven) resulting average squared error (ASE) and the best possible ASE. We show here that, by using randomization, this distribution is dilated, as compared to that for CV or GCV, only by a factor always lower than $1 + 1/n_{\mathrm{R}}$, where n_{R} is the number of primary randomized trace estimates one uses in RGCV. We include in the compared selectors, the partial cross-validation (PCV) approach where only a fraction of all the possible "leave-one-out" validation tests are evaluated; so that PCV is a common practice to reduce the computational cost in many contexts. In this paper, PCV will in fact appear as quite inefficient as compared to RGCV from this computational point of view. Moreover, we show that a precise comparison (and interpretation of the gain of using $n_{\mathrm{R}} \geq 2$) is possible in terms of equivalent (in distribution) excess errors, if PCV uses a certain percentage of the test points greater than 50%. The obtained comparisons will be seen as quite reassuring on what is "sacrificed" in using randomized selectors. We give rigorous results mainly for the kernel regression setting as in the previous detailed study by Härdle, Hall and Marron of standard selectors, except that we do not restrict this one to an equidistant design.

1. Introduction. Let us first consider the problem of recovering a "smooth" function m from noisy data which satisfy the model

$$y_i = m(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where $x_i \in [0, 1]$, $i = 1, \ldots, n$, are known (design) points and ε_i are independent and identically distributed observation errors with mean zero and variance σ^2. To fix ideas, assume that $x_i = F^{-1}((i - 0.5)/n)$, $f = F'$, where F is some known distribution function (i.e., the design points are equispaced percentiles from the density f on [0, 1]). Then a very simple example of

Received November 1995; revised June 1997.

AMS 1991 *subject classifications*. Primary 62G07, 62G20; secondary 62J07, 62G09, 65U05.

Key words and phrases. Nonparametric regression, bandwidth selection, cross-validation, partial cross-validation, generalized cross-validation, C_L method, fast randomized versions of GCV or C_L, regularization.

explicit curve-estimate for $m(x)$ is the kernel estimate

$$\hat{m}_h(x) := \frac{1}{nhf(x)} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) y_j, \tag{1.2}$$

where h is the smoothing parameter (bandwidth) and K is a smooth "bell-shaped" symmetric function which satisfies $\int K(x)\,dx = 1$. Note that (1.2) is not satisfactory for x near the boundary unless m is smoothly periodic, in case of which one considers a circulant version of (1.2) [see Rice (1984), Eubank and Wang (1994)]. In the following, A_h will denote a generic smoother, that is, the matrix (often called the hat matrix) satisfying $\hat{\mathbf{m}}_h = A_h \mathbf{y}$, where $\hat{\mathbf{m}}_h = (\hat{m}_h(x_1), \ldots, \hat{m}_h(x_n))^T$. For the above kernel estimate, the (i, j)th entry of A_h is $[A_h]_{i,j} = (1/nhf(x_i))K((x_i - x_j)/h)$. Other popular examples of efficient function-estimators $\hat{m}_h$ are smoothing splines or "Lowess" estimates (local weighted least squares).

It is well known that the selection of h is crucial (much more than that of K). A commonly used measure of performance for an estimate $\hat{m}_h$ is the average squared error

$$\Delta(h) := n^{-1} \sum \left[\hat{m}_h(x_i) - m(x_i)\right]^2 u(x_i) = n^{-1} \|A_h \mathbf{y} - \mathbf{m}\|_u^2, \tag{1.3}$$

or its expectation $M(h) := E(\Delta(h))$. Here, u is a fixed weight function introduced, as is classical in asymptotic study, to eliminate the boundary effect: u will be assumed to be supported on a subinterval of $(0, 1)$; in the periodic case mentioned above, u may be taken to be identically 1. In the following, the diagonal matrix $\mathrm{diag}(u(x_i), i = 1, \ldots, n)$ will be denoted by U and the weighted inner product $\mathbf{x}^T U \mathbf{y}$ (resp. its associated weighted l_2 norm) will be interchangeably denoted by $\langle \mathbf{x}, \mathbf{y} \rangle_U$ or $\langle \mathbf{x}, \mathbf{y} \rangle_u$ (resp. $\|\cdot\|_U$ or $\|\cdot\|_u$). The optimal parameter may then be defined either as the minimizer of Δ (which will be denoted by $\hat{h}_0$) or as the minimizer of M (denoted by h_0). (For simplification of notations, the dependence of Δ, M, h_0, $\hat{h}_0$, $\hat{h}_{\mathrm{CV}}$, etc., on n is suppressed.)

Many data-driven procedures, like cross-validation, share the property of being *asymptotically optimal*, which "typically" (i.e., under "enough" regularity assumptions, as those of Section 2 for the simple one-dimensional kernel setting above; other settings are outlined in Remark 3.2) means that, for the bandwidth, say $\hat{h}_{\mathrm{X}}$, given by such a procedure X (in short, for the selector $\hat{h}_{\mathrm{X}}$), the ratio $\hat{h}_{\mathrm{X}}/\hat{h}_0$ comes close to 1 as the sample size n increases. Of course, the estimation of $\hat{h}_0$ or of h_0 is only a means to the end of estimating the unknown mean function m. As is mentioned in Hall and Johnstone (1992), to compare two selectors $\hat{h}_1$ and $\hat{h}_2$, for the same kernel K, one natural way is to attempt to compare the risks $E(\Delta(\hat{h}_1))$ and $E(\Delta(\hat{h}_2))$ [note that $E(\Delta(\hat{h}_1))$ is not $M(\hat{h}_1)$ since $\hat{h}_1$ is a function of $\mathbf{y}$]. It can be shown that *the dominant term in the risk difference* $E(\Delta(\hat{h}_1)) - E(\Delta(\hat{h}_2))$ *is typically proportional to* $E(\hat{h}_1 - \hat{h}_0)^2 - E(\hat{h}_2 - \hat{h}_0)^2$. So we will mainly focus in this paper on the asymptotic distribution of $\hat{h}_{\mathrm{X}} - \hat{h}_0$ for each considered selector $\hat{h}_{\mathrm{X}}$, similarly as in the work by Härdle, Hall and Marron (1988) (abbreviated

as HHM henceforth), which completed first results by Rice (1984) [see Hall and Marron (1987) for the related density estimation setting]; note that we will essentially use the notations of HHM and of Hall and Johnstone (1992).

Let us recall that, if σ^2 were known, a basic method for choosing h is to form the following unbiased estimate of $M(h)$:

$$\mathrm{CL}(h) := n^{-1}\|(I - A_h)\mathbf{y}\|_u^2 + 2\sigma^2 n^{-1} \operatorname{tr} UA_h, \tag{1.4}$$

and to select $\hat{h}_{\mathrm{CL}}$ so as to minimize the criterion $\mathrm{CL}(h)$. An early study of this selector is Mallows (1973).

Let us define

$$t(h) := \frac{\operatorname{tr} UA_h}{\operatorname{tr} U} = n^{-1}h^{-1}K(0)\frac{\sum[u(x_i)/f(x_i)]}{\sum u(x_i)}, \tag{1.5}$$

where the second expression holds for the above curve-estimate (1.2). Generalized cross-validation (GCV) is a member of a family of criteria that can be written as the product of the weighted sum of squared residuals by a correction factor $\Xi_{\mathrm{X}}(t(h))$:

$$G_{\mathrm{X}}(h) := n^{-1}\|(I - A_h)\mathbf{y}\|_u^2 \cdot \Xi_{\mathrm{X}}(t(h)), \tag{1.6}$$

where Ξ_{X} is a (penalization) function satisfying $\Xi_{\mathrm{X}}(t) = 1 + 2t + O(t^2)$, with Ξ''_{X} bounded on a neighborhood of 0. A list of usual penalizations Ξ_{X} is presented in HHM. GCV, which is defined by $\Xi_{\mathrm{GCV}}(t) := (1 - t)^{-2}$, is one of the most popular; see Craven and Wahba (1979), Rice (1984), Li (1985, 1986), Wahba (1985) for theoretical results, Kohn, Ansley and Tharm (1991), Thompson, Brown, Kay and Titterington (1991) for quite extensive experimentations. It is important to note that, when $u \not\equiv 1$, this family G is, in fact, different from the "classical" one, which would use $\Xi_{\mathrm{X}}(n^{-1} \operatorname{tr} A_h)$ as correction factor. As we shall see, the use of (1.5) as trace-term is more natural and, first of all, it restores the asymptotic optimality, which is not obtained in general with $\Xi_{\mathrm{X}}(n^{-1} \operatorname{tr} A_h)$ in the weighted setting [as was noticed by Härdle and Marron (1985a)].

In the case of equally spaced data [$f \equiv 1$, in case of which $\operatorname{tr} UA_h/\operatorname{tr} U = n^{-1} \operatorname{tr} A_h$ for the particular (fixed bandwidth) curve estimate (1.2)], it has been shown that all these selectors are asymptotically optimal under weak regularity condition on m and K, and, of much more importance, they all are *asymptotically equivalent up to second order* [let $\hat{h}$ denote a generic one of these GCV-like selectors (1.5)–(1.6)], in the sense that the difference $\hat{h} - \hat{h}_0$ has a limit distribution independent of the particular Ξ one uses (and identical to that of $\hat{h}_{\mathrm{CL}} - \hat{h}_0$) [Rice (1984), HHM]. We will first see that easy generalizations of the proofs of HHM yield a straightforward extension of the result of HHM to nonequidistant designs (Theorem 2.1) *provided the trace-term* (1.5) *is used*.

The popular "leave-one-out" approach (or ordinary cross-validation) yields the criterion

$$\begin{aligned} \mathrm{CV}(h) &= n^{-1}\|D_h^{-1}(I - A_h)\mathbf{y}\|_u^2, \\ D_h &\coloneqq \mathrm{diag}(1 - [A_h]_{i,i}, i = 1, \ldots, n) \end{aligned} \tag{1.7}$$

when one uses a natural definition for the estimate of m using $n - 1$ observations [e.g., Hastie and Tibshirani (1990), Section 3.4.3]. It is easy to see that, for the curve estimate (1.2), around the optimal h_0 [see the proof of (2.6) in the Appendix for a related statement on the derivative CV′],

$$\begin{aligned} &\mathrm{CV}(h)/n^{-1}\|(I - A_h)\mathbf{y}\|_u^2 \\ &\quad = 1 + 2n^{-1}h^{-1}K(0)\frac{\sum[u(x_i)/f(x_i)]\varepsilon_i^2}{\sum u(x_i)\varepsilon_i^2} + O_p(n^{-2}h^{-2}). \end{aligned} \tag{1.8}$$

However, the right-hand term *cannot* be replaced, in general, by $1 + 2t(h) + O_p(n^{-2}h^{-2})$. One exception is, of course, the particular case $f \equiv 1$, and this observation was used by HHM to treat CV similarly as a member of the G family. Anyway, we will see that, as in the equidistant case, the selector $\hat{h}_{\mathrm{CV}}$ remains asymptotically equivalent to the above G-selectors $\hat{h}$ up to second order in the sense of the following Theorem 2.1, Section 2.

The main purpose of this paper is to study how far from $\hat{h}_0$ are the bandwidths produced by the fast randomized versions of CL or of G, which are defined by (if n_{R}, defined below, is equal to 1) the general formulas

$$\begin{aligned} \mathrm{RCL}(h) &\coloneqq n^{-1}\|(I - A_h)\mathbf{y}\|_u^2 + 2\sigma^2 n^{-1}\langle \mathbf{w}, A_h\mathbf{w}\rangle_u, \\ {}^{\mathrm{R}}G_{\mathrm{X}}(h) &\coloneqq n^{-1}\|(I - A_h)\mathbf{y}\|_u^2 \cdot \Xi_{\mathrm{X}}\left(\frac{\langle \mathbf{w}, A_h\mathbf{w}\rangle_u}{\langle \mathbf{w}, \mathbf{w}\rangle_u}\right), \end{aligned}$$

where $\mathbf{w}$ is a simulated unitary "white noise" vector $\mathbf{w}$ of size n, that is, such that its components are iid with mean 0 and variance 1. These criteria have been introduced in Girard (1989) (in the unweighted case) as fast Monte Carlo-type approximations to the exact ones, for all the contexts where computing $\mathrm{tr}\, UA_h$ is not an easy task; typical examples are smoothing splines or penalized least squares procedures, additive modeling by backfitting, iterative image restorations, etc.; see Girard (1995) for references to various applications. Such a randomized version can always be computed at a cost similar to the cost of one "fit" and with no additional programming effort, since one only has to rerun computation of the function estimate with the original data replaced by a simulated noise $\mathbf{w}$ which "mimics" $\boldsymbol{\varepsilon}$ up the factor σ. We also denote by $\mathrm{RCL}(h)$ [resp. ${}^{\mathrm{R}}G_{\mathrm{X}}(h)$] the *averaged randomized criteria* obtained when $\langle \mathbf{w}, A_h\mathbf{w}\rangle_u$ [resp. $\langle \mathbf{w}, A_h\mathbf{w}\rangle_u/\langle \mathbf{w}, \mathbf{w}\rangle_u$] is replaced by a Monte Carlo average $(1/n_{\mathrm{R}})\sum_{k=1}^{n_{\mathrm{R}}}\langle \mathbf{w}^k, A_h\mathbf{w}^k\rangle_u$ [resp. $(1/n_{\mathrm{R}})\sum_{k=1}^{n_{\mathrm{R}}}\langle \mathbf{w}^k, A_h\mathbf{w}^k\rangle_u/\langle \mathbf{w}^k, \mathbf{w}^k\rangle_u$]. Here n_{R} is intended to be a *fixed* small number (e.g., 10). Note that we could as well use $\mathrm{tr}\, U$ instead of $\langle \mathbf{w}, \mathbf{w}\rangle_u$ in these ratios. We shall see that the asymptotic behavior is identical (but is a function of n_{R}). However,

the first definition above has some advantages for finite sample sizes [Girard (1989, 1995)]. We will denote by a generic $\hat{h}_{\mathrm{R}}$, the minimizer of such a criterion ${}^{\mathrm{R}}G_{\mathrm{X}}(h)$.

Of course, for the kernel estimate (1.2), the randomized versions do not offer any computational gain. We only consider this setting (and its multidimensional versions) because the theory is very well developed for this very simple estimate. Note that, however, we do not restrict this study to an equally spaced design (contrary to HHM) because, otherwise, A_h would always have a Toeplitz structure (constant down its diagonals) and so one could wrongly suspect that these Monte Carlo approximations are taking advantage of this too particular structure. We hope to make clear that, having in mind that many nonparametric curve, surface or image estimates are more or less asymptotically equivalent to certain kernel estimates, one may easily conjecture that the theoretical comparisons which follow also hold for such nonparametric procedures (see Remarks 3.1 and 3.2).

In this article, we shall show that the algebraic rates of convergence to zero of the errors $\hat{h}_{\mathrm{R}} - \hat{h}_0$ and $\hat{h} - \hat{h}_0$ are identical, even with $n_{\mathrm{R}} = 1$ (n_{R} is the number of simulations used in the randomized version). Of more importance for our comparisons, the asymptotic variance of $\hat{h}_{\mathrm{R}} - \hat{h}_0$ is always less than two times the one of $\hat{h} - \hat{h}_0$. In fact, the increase of variability is only by a factor of the form $1 + n_{\mathrm{R}}^{-1}C$, with $0 < C < 1$ (this is stated in Theorem 3.1). For example, for a Gaussian kernel and $u(\cdot) \equiv 1$, then $C = 0.6$. A first result of this type was announced in Girard (1992) for a more idealized setting, namely, a tapered Fourier series estimate from continuous time observation perturbed by a Gaussian white noise process (this context can be thought of as a continuous limit of the equidistant design case).

To gain insight into the value of this increase of variance, it is useful to also compare with common partial cross-validation criteria which enjoy a more "principled" motivation. For example, in the one-dimensional setting of (1.1), to define, say, PCV_k, only the points with index i multiple of k are successively left-out and compared to the corresponding leave-one-out function estimate. The computational cost is then divided by k compared to full CV when direct implementation is used. We shall show that the algebraic rate of convergence to zero of $\hat{h}_{\mathrm{PCV}_k} - \hat{h}_0$ is still the same, but the increase of variance is now by a factor comprised between $(k + 1)/2$ and k [a precise expression for this factor is $k(1 - C) + ((k + 1)/2)C$, where C is the same constant as above]. In all the cases, this means that randomized GCV using only $n_{\mathrm{R}} = 1$ (resp. 2) simulation(s) already has a better theoretical justification than PCV_k with $k = 3$ (resp. 2), which needs 33.33% (resp. 50%) of all the possible "leave-one-out validation" tests. We shall also see in Section 4 that the results on PCV_k also hold for "rational" $k \in (1, 2)$, that is, when one considers more than 50% of the points. So that, even more precise comparisons (and interpretations of the gain of using $n_{\mathrm{R}} > 2$) are given in Section 4.

2. Some "asymptotic background" for the standard selectors. For clarity, we shall state our main results throughout this paper for the very

simple one-dimensional, known regular deterministic design, setting (1.1)–(1.2), with second-order kernel, that is, satisfying $\int x^2 K(x)\,dx > 0$. We shall require the following classical assumptions:

1. The errors ε_i are iid with mean 0, variance σ^2 and all other moments finite.
2. K is symmetric, compactly supported and has a Hölder continuous second derivative.
3. m is $C^2[0,1]$.
4. f is $C^2[0,1]$ and $f(x) \geq c > 0$ on the support of u, which is assumed $C^1[0,1]$.

As is usual in asymptotic studies of kernel estimate, the minimization of the various selection criteria is assumed to be restricted to an interval $H_n = [n^{-1+\epsilon}, n^{-\epsilon}]$ for any (small) constant $\epsilon > 0$, so that h in H_n satisfies $h \to 0$ and $nh \to \infty$. Now, by standard Riemann sum approximations and Taylor expansions [e.g., Eubank (1988), Härdle and Marron (1985b)], it can be shown that, uniformly over H_n,

$$M(h) = n^{-1}h^{-1}C_1 + \frac{h^4}{4}C_2 + o(n^{-1}h^{-1} + h^4),$$

$$C_1 = \sigma^2 \int u \int K^2, \qquad C_2 = \left(\int x^2 K\right)^2 \int ((mf)'')^2 f^{-1} u, \tag{2.1}$$

$$\Delta(h) = M(h) + o_P(M(h)), \tag{2.2}$$

and so $h_0 \sim C_0 n^{-1/5}$ with $C_0 = (C_1/C_2)^{1/5}$ and $\hat{h}_0/h_0 \to 1$ in probability. For later reference, note also that

$$M''(h_0) \sim C_3 n^{-2/5}, \qquad C_3 = 5C_1/C_0^3.$$

It is also known [Härdle and Marron (1985b)] that, uniformly over H_n,

$$n^{-1}\|(I - A_h)\mathbf{y}\|_u^2 = n^{-1}\|\boldsymbol{\varepsilon}\|_u^2 + M(h) - 2\sigma^2 n^{-1}h^{-1}K(0)\int u + o_P(M(h)). \tag{2.3}$$

Thus, since $t(h) = O(M(h))$, we have, by Taylor's expansion of Ξ in (1.6),

$$G(h) = n^{-1}\|\boldsymbol{\varepsilon}\|_u^2 + M(h) - 2\sigma^2 n^{-1}h^{-1}K(0)\int u + 2t(h)\sigma^2 \int uf + o_P(M(h)). \tag{2.4}$$

This shows that $t(h)$ defined by (1.5) is a trace-term which produces, in this setting, a criterion uniformly close to $M(h)$ (up to $n^{-1}\|\boldsymbol{\varepsilon}\|_u^2$), and thus all of G give a bandwidth $\hat{h}$ satisfying $\hat{h}/h_0 \to 1$ in probability, also called an "asymptotically optimal" bandwidth.

Of more importance for our purpose, the asymptotic stochastic behavior of all of $\hat{h} - \hat{h}_0$ can be described as follows. Let us define the centered processes

$$D(h) := \Delta(h) - M(h), \qquad \delta(h) := \mathrm{CL}(h) - n^{-1}\|\boldsymbol{\varepsilon}\|_u^2 - \Delta(h).$$

As in HHM [see also Rice (1984)], a main intermediate result is that $\hat{h} - \hat{h}_0$ is correctly described by the two linearized equations (where $'$ denotes here differentiation with respect to h)

$$\begin{aligned} -M''(h_0)\big(\hat{h}_0 - h_0\big) &= D'(h_0) + o_P(n^{-7/10}), \\ -M''(h_0)\big(\hat{h} - h_0\big) &= D'(h_0) + \delta'(h_0) + o_P(n^{-7/10}), \end{aligned} \tag{2.5}$$

where both $n^{7/10}D'(h_0)$ and $n^{7/10}\delta'(h_0)$ weakly converge to normal variables. To obtain (2.5) for all the members of the G family, an important step in HHM was to prove that, uniformly over $[an^{-1/5}, bn^{-1/5}]$,

$$G'(h) = \Delta'(h) + \delta'(h) + o_P(n^{-7/10}). \tag{2.6}$$

It is easy to extend (2.6) to nonequidistant designs provided the trace-term (1.5) is used, and we show in the Appendix that this also holds for CV$'$ in place of G'. Then the main result of HHM can easily be extended (see the Appendix) to nonequidistant designs, and to the ordinary CV criterion (the "kernel" L is defined in the Appendix).

THEOREM 2.1. *Under* 1–4 (*i.e.*, *under the assumptions of HHM except that the deterministic design may be nonequidistant*, *its density* f *being* C^2 *and bounded from below on the support of* u), *we have*, *for any GCV-like selector* $\hat{h}$ *defined by* (1.5)–(1.6),

$$\begin{aligned} C_3 n^{3/10}\big(\hat{h} - \hat{h}_0\big) &\to N\big(0, B^2 + V_2\big), \\ n\big[\Delta(\hat{h}) - \Delta\big(\hat{h}_0\big)\big] &\to \frac{1}{2C_3}\big(B^2 + V_2\big)\chi_1^2, \\ C_3 n^{3/10}\big(h_0 - \hat{h}_0\big) &\to N\big(0, B^2 + V_1\big), \\ n\big[\Delta(h_0) - \Delta\big(\hat{h}_0\big)\big] &\to \frac{1}{2C_3}\big(B^2 + V_1\big)\chi_1^2, \end{aligned}$$

where

$$\begin{aligned} B^2 &= 4C_0^2\sigma^2\left(\int x^2K\right)^2\int((mf)'')^2 f^{-1}u^2, \\ V_1 &= \frac{8}{C_0^3}\sigma^4\int(K*K - K*L)^2\int u^2, \\ V_2 &= \frac{8}{C_0^3}\sigma^4\int(K - L)^2\int u^2. \end{aligned}$$

These asymptotic distributions also hold for $\hat{h}_{\mathrm{CL}}$ *in place of* $\hat{h}$. *Furthermore, they also hold for the ordinary CV selector since*

$$n^{3/10}\big(\hat{h}_{\mathrm{CV}} - \hat{h}\big) = o_P(1), \qquad n\big[\Delta\big(\hat{h}_{\mathrm{CV}}\big) - \Delta(\hat{h})\big] = o_P(1).$$

REMARK 2.1. As far as we know, this is the first time that, for a smoothing operator not constant down its main diagonal, GCV-like criteria are stated to be asymptotically equivalent up to second order to full CV (see Section 4 for partial CV criteria), without any further condition on u [notice that Härdle (1990) has studied other criteria obtained by replacing $(1 - [A_h]_{i,i})^{-2}$ in (1.7) by $\Xi([A_h]_{i,i})$: these too are asymptotically equivalent up to second order to CV but, although they use penalizing functions, they are not GCV-like inasmuch as they cannot be expressed as function of only the weighted residual sum of squares and some trace-term]. We believe that this second-order equivalence is not specific to the particular estimate (1.2) (see Remark 3.2 below). Since GCV has not, in fact, a so "principled" motivation as CV, this equivalence is quite reassuring and is of practical importance for all those numerous problems for which one presently can afford GCV but not CV; typical examples are thin-plate (or interaction) smoothing spline estimates [Gu, Bates, Chen and Wahba (1989), Hutchinson (1990), Gu (1993)] or regularized reconstructions in computerized tomography [Girard (1987)].

3. Randomized criteria. We have seen that the process $\delta(h)$ can be considered as the exact "intrinsic" error or CL. Simple algebraic manipulations show that

$$\begin{aligned}\delta(h) &= -2n^{-1}\big(\langle \boldsymbol{\varepsilon}, A_h\boldsymbol{\varepsilon}\rangle_u - \sigma^2 \operatorname{tr} UA_h\big) + 2n^{-1}\langle \boldsymbol{\varepsilon}, (I - A_h)\mathbf{m}\rangle_u \\ &= e_1(\boldsymbol{\varepsilon}, h) + e_2(\mathbf{m}, \boldsymbol{\varepsilon}, h),\end{aligned} \tag{3.1}$$

say. On the other hand, the intrinsic error of the randomized version of CL, ${}^{\mathrm{R}}\delta(h) := \delta_{\mathrm{RCL}}(h) = \mathrm{RCL}(h) - n^{-1}\|\boldsymbol{\varepsilon}\|_u^2 - \Delta(h)$, can be easily seen to satisfy

$${}^{\mathrm{R}}\delta(h) = e_1(\boldsymbol{\varepsilon}, h) + e_2(\mathbf{m}, \boldsymbol{\varepsilon}, h) - \frac{1}{n_{\mathrm{R}}}\sum_{k=1}^{n_{\mathrm{R}}} e_1(\boldsymbol{\varepsilon}^{*k}, h), \tag{3.2}$$

where $\boldsymbol{\varepsilon}^{*k} = \sigma\,\mathbf{w}^k$ are mutually independent, independent from $\boldsymbol{\varepsilon}$ and well satisfy all the conditions assumed on $\boldsymbol{\varepsilon}$ [condition 1 with same σ^2]. Comparing (3.1) and (3.2), this means that the worst (i.e., for $n_{\mathrm{R}} = 1$) additional "randomization error" is already similar to one of the two components of the intrinsic error of CL. Since, in the process of proving (2.3), both the two components of δ were shown uniformly negligible as compared to $M(h)$ (a sketch of the proof is given in the Appendix), this immediately implies the asymptotic optimality of RCL, like in the ridge regression setting [Girard (1991), Section 2]. In fact, in our kernel setting here, the asymptotic optimality of the ${}^{\mathrm{R}}G_{\mathrm{X}}$'s is also immediate; indeed, a similar Taylor argument shows that, even with $n_{\mathrm{R}} = 1$, (2.4) also holds with ${}^{\mathrm{R}}G(h)$ in place of $G(h)$ and ${}^{\mathrm{R}}t(h) = \langle \mathbf{w}, A_h\mathbf{w}\rangle_u / \langle \mathbf{w}, \mathbf{w}\rangle_u$ or ${}^{\mathrm{R}}t(h) = \langle \mathbf{w}, A_h\mathbf{w}\rangle_u / \operatorname{tr} U$ in place of $t(h)$, which states that ${}^{\mathrm{R}}G - \mathrm{RCL}$ is uniformly $o_P(M)$ over H_n.

Of more importance for our purpose, we shall prove in the Appendix the analog of (2.5) that one could expect, that is,

$$-M''(h_0)\left(\hat{h}_{\mathrm{R}} - h_0\right) = D'(h_0) + {}^{\mathrm{R}}\delta'(h_0) + o_P(n^{-7/10}), \tag{3.3}$$

so that $\hat{h}_{\mathrm{R}} - \hat{h}_0$ is asymptotically proportional to the centered variable ${}^{\mathrm{R}}\delta'(h_0)$, with the same (deterministic) proportionality factor as for $\hat{h} - \hat{h}_0$. To obtain such an equivalence, for a given n_{R}, between all the members of the ${}^{\mathrm{R}}G$ family (the same sequence $\mathbf{w}^1, \ldots, \mathbf{w}^{n_{\mathrm{R}}}$ being used in every one), an important step is to prove the following analog of (2.6):

$${}^{\mathrm{R}}G'(h) = \Delta'(h) + {}^{\mathrm{R}}\delta'(h) + o_P(n^{-7/10}) \tag{3.4}$$

uniformly over $[an^{-1/5}, bn^{-1/5}]$. Then, using (3.2), we easily obtain that the limit distributions of Theorem 2.1 are modified in a very simple way when using randomized versions.

THEOREM 3.1. *Under the assumptions of Theorem* 2.1, *assumptions on the mutually independent* $\mathbf{w}$*'s identical to that on* $\sigma^{-1}\boldsymbol{\varepsilon}$ *and assuming these* $\mathbf{w}$*'s independent from* $\boldsymbol{\varepsilon}$, *we have, for the* ${}^{\mathrm{R}}G_{\mathrm{X}}$ *selectors* $\hat{h}_{\mathrm{R}}$,

$$C_3 n^{3/10}\left(\hat{h}_{\mathrm{R}} - \hat{h}_0\right) \to N\left(0, B^2 + \left(1 + \frac{1}{n_{\mathrm{R}}}\right)V_2\right),$$

$$n\left[\Delta\left(\hat{h}_{\mathrm{R}}\right) - \Delta\left(\hat{h}_0\right)\right] \to \frac{1}{2C_3}\left(B^2 + \left(1 + \frac{1}{n_{\mathrm{R}}}\right)V_2\right)\chi_1^2,$$

where B *and* V_2 *are defined in Theorem* 2.1. *This also holds for the RCL selector*.

One may now combine Theorems 2.1 and 3.1 to express what is "sacrificed" when using a fast randomized version, in terms of relative risk regret:

$$\frac{\mathrm{E}\left[\Delta\left(\hat{h}_{\mathrm{R}}\right) - \Delta\left(\hat{h}_0\right)\right]}{\mathrm{E}\left[\Delta(\hat{h}) - \Delta\left(\hat{h}_0\right)\right]} = 1 + \frac{1}{n_{\mathrm{R}}}C, \qquad C = \frac{V_2}{B^2 + V_2} < 1, \tag{3.5}$$

where E denotes here the asymptotic expectation. We point out that, in the case $u \equiv 1$ (which requires periodicity of m and f), one can easily verify that C is a constant only function of K; for example, $C = 0.6$ is obtained when K is the Gaussian density.

REMARK 3.1. Note that the decompositions (2.6) and (3.4) are exact [the terms $o_P(n^{-7/10})$ vanish] for, respectively, CL and RCL. This implies, *even for finite* n, for example for Gaussian $\boldsymbol{\varepsilon}$, that, at any h, the standard error of $\mathrm{CL}'(h) - \Delta'(h)$ is at worst "widened" by a factor $\sqrt{1 + 1/n_{\mathrm{R}}}$ by randomization. So $n_{\mathrm{R}} = 10$, say, may be claimed as a sufficient simulation size for practice, as was discussed in Girard (1995). (In that paper, heuristics suggest that this also holds for RGCV, the randomized version of GCV, for any reasonable problem for which exact GCV is a relevant method.) The first part

of Theorem 3.1 says that such a guaranty of an "at worst 5% additional error" when using ten simulations is well propagated, asymptotically, on the standard error of $\hat{h} - \hat{h}_0$, while the second part says that the maximal inflation then becomes 10% when considering the distribution of the excess error $\Delta(\hat{h}) - \Delta(\hat{h}_0)$. Of course, these approximations may be inaccurate for small n. Further work (including simulation studies) would be useful to understand when this may happen. However, this theoretical "at worst 10% inflation" result is rather well in agreement with the simulation studies of which we are aware: for example, in Girard (1989), with $n_R = 10$, the (admittedly rough) estimate of the distribution of the inefficiencies $\Delta(\hat{h}_R)/\Delta(\hat{h}_0) - 1$ was almost indistinguishable from that of $\Delta(\hat{h})/\Delta(\hat{h}_0) - 1$ for $n = 50$ or 500 and various designs. Note that such a quite attractive behavior is in contrast with common Monte Carlo practices: for example, when using the classical bootstrap method to compute standard error of a given statistic, the required number of replications of the data set is typically of the order of n to have a 5% accuracy, as is discussed in Efron [(1987), Section 9].

REMARK 3.2. (a) As was the case in HHM, our theorems may be easily extended to higher-order kernels; see HHM for the kind of modifications of assumptions, rates and constants involved.

(b) Extensions to multidimensional settings could be stated; for detailed constants for the related multidimensional density estimation problem, see Hall and Marron [(1987), Remark 2.1]. But it should be pointed out that such a multidimensional generalization to the regression problem requires some care in the case of a deterministic regular design: as was already noticed by several authors, a "workable" asymptotic mean square error formula [like (2.1)] may not exist in too high dimensions for "natural" multidimensional kernel estimates because, briefly said, the quadrature error in approximating the expectation of $\hat{m}_h(x)$ by a convolution integral may dominate the required usual Taylor approximation (typically a constant multiple of h^2) of the bias; see, for example, Azari and Muller (1992). Notice, however, that this does not happen, for example, in dimension 2 for certain asymptotically regular designs and the GM (Gasser–Muller) kernel estimate of order 2, as studied by Herrmann, Wand, Engel and Gasser [(1995), Section 2]. For the setting of that paper, it can be verified that the proofs of Theorems 2.1 and 3.1 work, the results of these theorems hold with a single modification in the rates [the power $3/10$ is replaced everywhere by $(d + 2)/(8 + 2d)$ with $d = 2$] and with modifications in the constants which are obtained by computations which are now standard in the field. [It may be interesting to note that, as expected from the convolution form of the GM estimate, in the expression of C_0, B^2 and V_2, the term $((mf)'')^2 f^{-1}$ is now replaced by f times a linear combination of squared second-order partial derivatives of m, not of mf.]

(c) Among other possible settings, let us finally mention that a result of the form (3.5) can also be shown when discontinuities (jumps or cups) are allowed in m in the setting of Section 2, as is discussed in Kneip (1994) [with, again,

only changes in the constant B^2 and V_2 which can be computed similarly as in Van Es (1992)]. Such a robustness property with respect to lack of smoothness of m is quite reassuring for the reliability of the randomized versions.

Of course, it is not within the scope of this paper to list all the settings (with possibly random designs) where a limit distribution, with some rate of convergence, has been derived or conjectured for the excess error of the CL (or GCV) selector. The essential point is that, in fact, by considering the nature of the arguments above and in the Appendix [especially (3.2) and (3.3)], one can now conjecture that a result of the form (3.5) also holds for all these settings (they include many other nonparametric function estimates, like smoothing spline or penalized least squares estimates).

REMARK 3.3. The assumption $\operatorname{var} \varepsilon_i \equiv \sigma^2$ has a specific role in this study. Indeed, it is known that, while CV remains asymptotically optimal under heteroskedasticity, GCV needs a modification which requires knowledge of all the $\operatorname{var} \varepsilon_i$'s or estimates for them, up to a multiply. If they are known, simple extension of the results here could be made [see Girard (1995)], but this is rarely the case. In fact, fast randomized versions for such contexts would be of great interest and so deserve further study.

REMARK 3.4. It might be worth studying possible variance reduction techniques for the primary randomized trace-estimate one uses here. In Girard (1993), a simple correlated sampling technique greatly improved the accuracy of the trace-estimate for a particular image smoothing problem. Of course, such developments are useful in practice only for "hard" data analysis problems (in the sense that exact cross-validation already has quite a lot of variability) where one cannot easily afford, say, $n_R = 10$; see Girard [(1995), Section 5.1] for such an example. We will not develop this topic further here, since very different instances of variance reduction techniques could be elaborated.

4. Comparison with partial cross-validation. The points which are successively left out and compared to the corresponding leave-one-out function-estimate may be chosen as only a subset of the n observations. To fix the idea, consider a strategy which uses only the odd points in the one-dimensional setting of (1.1). The resulting partial cross-validation is then concisely defined by

$$\mathrm{PCV}_{\mathrm{odd}}(h) = 2n^{-1}\|D_h^{-1}(I - A_h)\mathbf{y}\|^2_{U_{\mathrm{odd}}}, \tag{4.1}$$

where $U_{\mathrm{odd}} = \operatorname{diag}(1, 0, 1, 0, \ldots)U$. Note that only one-half of the diagonal elements of A_h are required. So the computation cost of $\mathrm{PCV}_{\mathrm{odd}}(h)$ is one-half

that of full CV when, either computing all the diagonal elements has the dominant cost, or one uses a direct implementation which successively computes the leave-one-out estimates, as is typically the case when the function-estimate procedure is an iterative one.

As first sight, it seems natural to compare this bandwidth selector against the minimizer of the corresponding partially averaged square error

$$\Delta_{\text{odd}}(h) := 2n^{-1}\|A_h\mathbf{y} - \mathbf{m}\|^2_{U_{\text{odd}}} \tag{4.2}$$

or its expectation $M_{\text{odd}}(h) := E(\Delta_{\text{odd}}(h))$. It is easy to see that such partial averaging incurs relatively negligible modification on M or its derivative, provided m and f are smooth enough. So the first step here is to show (proof in the Appendix) that this is also the case for the random Δ, so that

$$|\hat{h}_0 - \hat{h}_{0\,\text{odd}}| = o_P(n^{-3/10}), \tag{4.3}$$

where $\hat{h}_{0\,\text{odd}}$ denotes the minimizer of Δ_{odd}.

It thus remains to study the stochastic behavior of $\hat{h}_{\text{PCV}_{\text{odd}}} - \hat{h}_{0\,\text{odd}}$. By exactly similar arguments as in the above sections (so the proof will be omitted), one can show that it is well characterized by the following linearized equation:

$$-M''(h_0)\left(\hat{h}_{\text{PCV}_{\text{odd}}} - \hat{h}_{0\,\text{odd}}\right) = \delta'_{\text{odd}}(h_0) + o_P(n^{-7/10}), \tag{4.4}$$

where δ_{odd} is the corresponding partial version of (3.1), that is, obtained with U replaced by $2U_{\text{odd}}$.

Before we state our results on the efficiency of PCV_{odd}, we now introduce one example of more general "partial averaging" approaches for which analog results also hold. Let l and k be two given integers such that $1 \le l \le k$. PCV_k^l is then defined as the average over only the points of index $1, 2, \ldots, l$ among the first k points, the points of index $k+1,\ k+2, \ldots, k+l$ among the second k points, and so on (PCV_k^1 then coincides with PCV_k defined in the Introduction). For n large enough, PCV_k^l can be called a partial CV using $100(l/k)\%$ of the possible test points for the validation step. For PCV_k^l, the proof of the analog of (4.3) is identical (see Appendix) and, by exactly similar proofs (which differ only by a need of more complex notations), one obtains the analog of (4.4) with a "δ-term" defined by replacing $2U_{\text{odd}}$ by $(k/l)U_{k,l}$, with

$$U_{k,l} := \text{diag}(u(x_1), \ldots, u(x_l), 0, \ldots, 0, u(x_{k+1}), \ldots, u(x_{k+l}), 0, \ldots, 0, \ldots).$$

By an analysis of this new δ-term (see the Appendix), the following theorem is then obtained.

THEOREM 4.1. *Under the assumptions of Theorem* 2.1 *and assuming m and f in* C^3 [0, 1],

$$C_3 n^{3/10}\left(\hat{h}_{\text{PCV}_{\text{odd}}} - \hat{h}_0\right) \to N\left(0, 2B^2 + \tfrac{3}{2}V_2\right),$$

$$n\left[\Delta\left(\hat{h}_{\text{PCV}_{\text{odd}}}\right) - \Delta\left(\hat{h}_0\right)\right] \to \frac{1}{2C_3}\left(2B^2 + \frac{3}{2}V_2\right)\chi^2_1,$$

where B, V_2 *are defined in Theorem* 2.1. *For any integers* $k \geq 2$ *and* $1 \leq l \leq k-1$, *this extends to* PCV_k^l *with* $2B^2 + \frac{3}{2}V_2$ *replaced by* $(k/l)B^2 + \frac{1}{2}(k/l + 1)V_2$.

The analog of (3.5) is thus

$$\frac{\mathrm{E}\left[\Delta\left(\hat{h}_{\mathrm{PCV}_k^l}\right) - \Delta\left(\hat{h}_0\right)\right]}{\mathrm{E}\left[\Delta\left(\hat{h}\right) - \Delta\left(\hat{h}_0\right)\right]} = \frac{k}{l}(1-C) + \frac{1}{2}\left(\frac{k}{l} + 1\right)C, \tag{4.5}$$

$$C = \frac{V_2}{B^2 + V_2} < 1.$$

REMARK 4.1. The additional smoothness conditions on m and f are for convenience but are not minimal; our goal here is not to give a detailed study of the partial CV approach (see Remarks 4.3 and 4.4). The "periodicity" in the subsampling is for notational convenience: it is easy to see that the above asymptotic behavior of PCV_k^l holds whatever locations the l distinct points may have in each neighborhood of size k; that is, in other words, it holds for any one of the "$100(l/k)\%$ partial CV" criteria which use such quasi-uniform distribution of the test points.

REMARK 4.2. The claims of Remark 3.2 may be repeated here with only (3.5) replaced by (4.5). But it must be pointed out that, because we have restricted ourselves to deterministic designs, easy extensions of (4.5) to multidimensional designs are claimed only for rectangular designs [as those considered in the simulation study of Herrmann et al. (1995)]; otherwise, the construction of such a PCV_k^l with "quasi-uniform" subsampling may actually be quite a headache.

REMARK 4.3. Our first goal here was to compare the randomization approach against the partial CV approach as two possible computational strategies, for example, when an iterative procedure is used to compute the function-estimate [see the discussion and rejoinder of Girard (1995) for references]. From a comparison of result (3.5) and result (4.5) with $l = 1$, one may now argue that, in many nonparametric settings (cf. Remarks 3.2 and 4.2), where we appeal to such partial CV criteria in order to alleviate the computational burden of a direct implementation (or in order the reduce the number of required $[A_h]_{ii}$'s), partial CV is, in fact, a very inefficient way of reducing the cost, as compared to randomized GCV.

REMARK 4.4. We can now interpret the value of the possible inflation factors caused by randomization, by comparing them with those caused by partial CV criteria using more than 50% of the test points. Consider any $n_{\mathrm{R}} \geq 1$. Simply by finding the ratio k/l which equates (3.5) and (4.5), one obtains the following: in order to be as efficient (in terms of asymptotic risk

regret) as randomized G or CL, the PCV_k^l strategy must use l, k satisfying

$$\frac{l}{k} \geq \frac{n_{\mathrm{R}}(1 - C/2)}{n_{\mathrm{R}}(1 - C/2) + C}.$$

In terms of percentage of points used in partial CV for the validation step, to be as efficient as ${}^{\mathrm{R}}G$ or RCL, a partial CV strategy must then use $100/(1 + n_{\mathrm{R}}^{-1}C')\%$ of the points, where $C' = C/(1 - C/2)$. For example, for $n_{\mathrm{R}} = 10$, if K is the Gaussian density and $u \equiv 1$, this percentage is as much as 92.1%. Note that the maximal value of C' is 2, and so the required percentage will always be greater than $100/(1 + n_{\mathrm{R}}^{-1}2)\%$ (i.e., greater than 83.33%, for $n_{\mathrm{R}} = 10$).

APPENDIX

Because all the studied selectors are asymptotically optimal procedures, to derive the asymptotic distribution of the selected bandwidth, we can assume without loss of generality that these procedures are redefined by restricting the minimizations over an interval $[an^{-1/5}, bn^{-1/5}]$ containing $C_0 n^{-1/5}$. For the unequally spaced design and kernel estimate (1.1)–(1.2), the analogs of the notations of HHM become

$$r_n(h) = n^{-1}h^{-1} + h^4, \qquad L(x) = -xK'(x)$$

and

$$b_h(x) = \frac{1}{nhf(x)} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) m(x_j) - m(x),$$

$$c_h(x) = \frac{1}{nhf(x)} \sum_{j=1}^{n} L\left(\frac{x - x_j}{h}\right) m(x_j) - m(x).$$

Let also B_h denote the smoothing operator associated with the kernel L, so that

$$\frac{d}{dh}[A_h]_{i,j} = -h^{-1}[A_h - B_h]_{i,j}.$$

Also, as in HHM, let us define $\delta_1(h) = 2n^{-1}\langle \boldsymbol{\varepsilon}, (I - A_h)\mathbf{m}\rangle_u - 2n^{-1}\langle \boldsymbol{\varepsilon}, A_h\boldsymbol{\varepsilon}\rangle_u$.

PROOF OF (2.6). First, as in HHM, the fact that $\Delta'(h)$ and $\delta_1'(h)$ are still $O_P(n^{-3/5})$ uniformly over $[an^{-1/5}, bn^{-1/5}]$ results from the extension of (A.7) of HHM to nonequidistant designs, which is proved in the following Lemma 1. Next, the proof for G' is very similar to that in HHM for the equispaced case. So we only consider the proof for CV′. Differentiating the sum (1.7) term by

term gives

$$\begin{aligned}\mathrm{CV}'(h) &= n^{-1}\sum u(x_i)\left\{\frac{d}{dh}\left[(I-A_h)\mathbf{y}\right]_i^2\right\}\left(1+O(n^{-1}h^{-1})\right)\\ &\quad - n^{-1}\sum u(x_i)\left[(I-A_h)\mathbf{y}\right]_i^2\left(f(x_i)^{-1}n^{-1}h^{-2}K(0)\right)\\ &\quad\quad \times\left(2+O(n^{-1}h^{-1})\right)\\ &= \Delta'(h)+\delta_1'(h)-2n^{-1}h^{-2}K(0)n^{-1}\|\boldsymbol{\varepsilon}\|_{u/f}^2+o_P(n^{-7/10}),\end{aligned}$$

where the third term of this second approximation can be obtained similarly as (2.3) with weights u/f in place of u. Noting that $n^{-1}\|\boldsymbol{\varepsilon}\|_{u/f}^2$ may be also replaced by $\sigma^2\int u$ in this approximation up to $o_P(n^{-7/10})$, yields (2.6) for CV′. □

PROOF OF (3.4). Let us recall that ${}^{\mathrm{R}}G(h) = n^{-1}\|(I-A_h)\mathbf{y}\|_u^2\cdot\Xi({}^{\mathrm{R}}t(h))$, where ${}^{\mathrm{R}}t(h) = \langle\mathbf{w}, A_h\mathbf{w}\rangle_u/\langle\mathbf{w},\mathbf{w}\rangle_u$ or ${}^{\mathrm{R}}t(h) = \langle\mathbf{w}, A_h\mathbf{w}\rangle_u/\operatorname{tr}U$. Thus,

$$\begin{aligned}{}^{\mathrm{R}}G'(h) &= \left(\Delta'(h)+\delta_1'(h)\right)\left(1+O_P(n^{-1}h^{-1})\right)\\ &\quad +\left(\Delta(h)+n^{-1}\|\boldsymbol{\varepsilon}\|_u^2+\delta_1(h)\right)\left[{}^{\mathrm{R}}t'(h)\left(2+O_P(n^{-1}h^{-1})\right)\right]\\ &= \Delta'(h)+\delta_1'(h)+2\left[{}^{\mathrm{R}}t'(h)n^{-1}\|\boldsymbol{\varepsilon}\|_u^2\right]+o_P(n^{-7/10}).\end{aligned}$$

Approximation (3.4) is then obtained by observing that ${}^{\mathrm{R}}t'(h)n^{-1}\|\boldsymbol{\varepsilon}\|_u^2$ may also be replaced by $(d/dh)(\sigma^2\langle\mathbf{w}, A_h\mathbf{w}\rangle_u)$, up to $o_P(n^{-7/10})$, since ${}^{\mathrm{R}}t'(h) = O_P(n^{-1}h^{-2})$. □

PROOFS OF THEOREMS 2.1 AND 3.1. We assume here the conditions of Theorem 3.1 (i.e., those of Theorem 2.1 when the statements do not concerne randomized terms). We will need the following lemmas.

LEMMA 1. *Lemmas* 1, 2, 3 *and* 5 *of HHM still hold.*

PROOF. Let $D_1(h) = -(h/2)D'(h)$. It is easy to see that $D_1 = S_1(h) + S_2(h)$, where

$$\begin{aligned}S_1(h) &= n^{-1}\langle A_h\boldsymbol{\varepsilon}, (A_h-B_h)\boldsymbol{\varepsilon}\rangle_u - \sigma^2\operatorname{tr}\left(A_h^TU(A_h-B_h)\right),\\ S_2(h) &= n^{-1}\langle A_h\boldsymbol{\varepsilon}, \mathbf{b}_h-\mathbf{c}_h\rangle_u + n^{-1}\langle(A_h-B_h)\boldsymbol{\varepsilon}, \mathbf{b}_h\rangle_u.\end{aligned}$$

Consider, for example,

$$S_{11}(h) = n^{-1}\langle A_h\boldsymbol{\varepsilon}, A_h\boldsymbol{\varepsilon}\rangle_u - \sigma^2n^{-1}\operatorname{tr}\left(A_h^TUA_h\right).$$

To show, for example, the second part of Lemma 1 of HHM, then, Theorem 2 of Whittle (1960) can still be applied, to give, for any integer $l\geq 1$ (C is a generic constant here and in the following),

$$\begin{aligned}&\mathrm{E}\left\{\left(r_n(h)^{-1}h^{-1/2}|S_{11}(h)-S_{11}(h')|\right)^{2l}\right\}\\ &\quad\leq Cr_n(h)^{-2l}h^{-l}n^{-2l}\left[\sum_i\sum_j\left|\left[A_h^TUA_h\right]_{ij}-\left[A_{h'}^TUA_{h'}\right]_{ij}\right|^2\right]^l,\end{aligned}$$

where $[A_h^T U A_h]_{ij} = n^{-2}h^{-2}\sum_{l=1}^n f(x_l)^{-1}K((x_l - x_i)/h)u(x_l)f(x_l)^{-1}K((x_l - x_j)/h)$. Now, since $r_n(h)^{-2l} \le n^{2l}h^{2l}$, it remains to see that, as in HHM, the sum $\sum_i\sum_j$ contains at most a multiple of n^2h' (for $h' \ge h$) nonzero terms, and that, uniformly over i, j,

$$\left|\left[A_h^T U A_h\right]_{ij} - \left[A_{h'}^T U A_{h'}\right]_{ij}\right| \le Cn^{-1}h^{-2}|h - h'|.$$

This stems from the fact that, because f is smooth and bounded from below, A_h and $(d/dh)A_h$ are still banded matrices with horizontal (and vertical) width bounded by Cnh and that it still holds that $|[A_h]_{ij}| \le Cn^{-1}h^{-1}$ and $|[(d/dh)A_h]_{ij}| \le Cn^{-1}h^{-2}$.

For the linear terms of $S_2(h)$ such as $n^{-1}\langle B_h\boldsymbol{\varepsilon}, \mathbf{b}_h\rangle$ (note that here we will use $r_n(h)^{-2l} \le (n^l h^l)h^{-4l}$ over H_n), it suffices to observe that the components $[\mathbf{b}_n]_i$ and $[\mathbf{c}_h]_i$ are still $O(h^2) + O(n^{-1})$ uniformly. The extension of the δ part of Lemma 1 of HHM is proved by similar bounds. Extension of Lemma 2 of HHM is obtained from the above extension of Lemma 1 using exactly the same partitioning arguments as in HHM. Extension of Lemma 3 of HHM is also derived from the extension of Lemma 2 of HHM using the same arguments as in HHM and using that $\hat{h}/h_0, \hat{h}_0/h_0 \to 1$. Lemma 5 of HHM [which is required for the results on $\Delta(\hat{h}) - \Delta(\hat{h}_0)$] resulted from an analog on D'' of Lemma 1 of HHM, namely, that $\sup_{h\in[an^{-1/5}, bn^{-1/5}]}\mathrm{E}|n^{1/2}D''(h)|^{2l} \le C$. Extension of this bound to nonequidistant design may be derived as above. □

From the expression (3.2) of ${}^{\mathrm{R}}\delta$, it is now immediate to prove the following.

LEMMA 2. *Lemma 2 of HHM holds with* ${}^{\mathrm{R}}\delta$ *in place of* δ.

LEMMA 3. *For some* $\varepsilon > 0$, $|\hat{h}_{\mathrm{R}} - h_0| = O_P(n^{-1/5-\varepsilon})$.

PROOF. Similarly as for Lemma 3 of HHM, this results from Lemma 2 and $\hat{h}_{\mathrm{R}}/h_0 \to 1$. □

The following lemma generalizes and completes Lemma 4 of HHM.

LEMMA 4. *Under the assumptions of Theorem* 3.1,

$$n^{7/10}\begin{bmatrix} D'(h_0) \\ \delta'(h_0) \\ {}^{\mathrm{R}}\delta'(h_0) \end{bmatrix} \to N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_3^2 & & \\ \sigma_{34} & \sigma_4^2 & \\ \sigma_{34} & \sigma_4^2 & \sigma_{\mathrm{R}}^2 \end{bmatrix}\right),$$

with $\sigma_3^2 = B^2 + V_1$, $\sigma_4^2 = B^2 + V_2$, $\sigma_{\mathrm{R}}^2 = B^2 + (1 + 1/n_{\mathrm{R}})V_2$, *where* B, V_1 *and* V_2 *are defined in Theorem* 2.1.

PROOF. Let us consider, for example, the first component and, as in the proof of Lemma 1, the associated decomposition $D_1(h) = S_1(h) + S_2(h)$. The derivation of the asymptotic normal distribution for the linear term $S_2(h_0)$ is very similar to that in HHM. So let us again consider the term $nS_1(h) = \langle\boldsymbol{\varepsilon}, Q_h\boldsymbol{\varepsilon}\rangle - \mathrm{E}\langle\boldsymbol{\varepsilon}, Q_h\boldsymbol{\varepsilon}\rangle$, where $Q_h = A_h^T U(A_h - B_h)$ is still a banded matrix of

width $O(nh)$. Note that, for any matrix Q, standard calculations show that $\operatorname{var}\langle \boldsymbol{\varepsilon}, Q\boldsymbol{\varepsilon}\rangle = \sigma^4 \operatorname{tr} Q^2 + \sigma^4 \operatorname{tr} QQ^T + (\mathrm{E}(\varepsilon_i^4) - 3\sigma^4)\sum_i Q_{ii}^2$. Let $\tilde{Q} = \frac{1}{2}(Q + Q^T) - \operatorname{diag}(Q_{ii}, i = 1, \ldots, n)$, so that $\tilde{Q}$ is symmetric and of zero diagonal. It is known [e.g., de Jong (1987)] that $\langle \boldsymbol{\varepsilon}, \tilde{Q}\boldsymbol{\varepsilon}\rangle$ has an asymptotic normal distribution under the simple condition $\operatorname{tr} \tilde{Q}^4/(\operatorname{tr} \tilde{Q}^2)^2 \to 0$.

Now, by Riemann sum approximation and first-order Taylor expansion of u/f, we obtain that

$$[Q_h]_{i,j} = \frac{u(x_i)}{nhf(x_i)}(K * K - K * L)\left(\frac{x_i - x_j}{h}\right) + O(n^{-1}) + O(n^{-2}h^{-2}),$$

uniformly over i and j. Thus, $\operatorname{tr} Q_h Q_h^T \sim h^{-1}\int (K * K - K * L)^2 \int u^2$, and similarly for $\operatorname{tr} Q_h^2$, and the third diagonal term $\sum_i [Q_h]_{i,i}^2$ is relatively negligible. Now it suffices to derive $\operatorname{tr} \tilde{Q}_h^4 \le Ch^{-1}$ from the banded structure of Q_h, to obtain the asymptotic normal distribution of $S_1(h_0)$. The limit distribution of $D_1(h_0)$ is finally obtained by verifying that the covariance of $S_1(h_0)$ and $S_2(h_0)$ is relatively negligible. The joint distribution of $[D'(h_0), \delta'(h_0)]^T$ can then be proved by similar modifications of the proof in HHM.

Finally, a simple way to derive the whole joint distribution is to consider any linear combination

$$[\alpha, \beta, \gamma]^T \begin{bmatrix} D'(h_0) \\ \delta'(h_0) \\ {}^{\mathrm{R}}\delta'(h_0) \end{bmatrix} = \alpha D'(h_0) + (\beta + \gamma)\delta'(h_0) - \frac{\gamma}{n_{\mathrm{R}}} \sum_{k=1}^{n_{\mathrm{R}}} e_1'(\boldsymbol{\varepsilon}^{*k}, h_0).$$

Indeed, from the joint distribution of $[D'(h_0), \delta'(h_0)]^T$, and the fact that the third term in this expression is an average of n_{R} independent terms already analyzed and is independent from the first two terms, we see that such a (nonzero) combination, multiplied by $n^{7/10}$, is asymptotically distributed as the centered normal distribution of variance $\alpha^2\sigma_3^2 + (\beta + \gamma)^2\sigma_4^2 + 2\alpha(\beta + \gamma)\sigma_{34} + (\gamma^2/n_{\mathrm{R}})V_2$, which yields the stated tridimensional normal law. □

From the above lemmas, and from (2.6) and (3.4), the proofs of the linearized characterizations (2.5) and (3.3), and of Theorems 2.1 and 3.1, use very similar lines of proof as in HHM and so are omitted.

PROOF OF THEOREM 4.1. We have first to prove the equivalence [extension of (4.3)] between $\hat{h}_0$ and the minimizer of the partial loss $\Delta_k^l(h) := (k/l)n^{-1}\|A_h\mathbf{y} - \mathbf{m}\|_{U_{k,l}}^2$. Only for notational simplicity, we consider the case $k = 2$, $l = 1$, that is, Δ_{odd}. For this, it suffices to prove the following.

LEMMA 5. *We have*

$$\sup_{h \in [an^{-1/5}, bn^{-1/5}]} n^{7/10}\{|M'(h) - M'_{\mathrm{odd}}(h)| + |D'(h) - D'_{\mathrm{odd}}(h)|\} = o_P(1).$$

PROOF. Let

$$M(x,h) := \big(b_h(x)\big)^2 + \sigma^2 \sum_{j=1}^{n} \left[\frac{1}{nhf(x)} K\Big(\frac{x - x_j}{h}\Big)\right]^2,$$

the mean square error at x. By standard Riemann sum approximation and using that $n^{-1}h^{-1}\sum_j (d/dx)K((x - x_j/h)y_j$ is a kernel estimate of $(mf)'$ with a bias still $O(h^2)$, one verifies that $(d/dx)(n^{4/5}M(x,h))$ is bounded uniformly over x and $h \in [an^{-1/5}, bn^{-1/5}]$. Thus, the difference between the normalized partial average $n^{4/5}M_{\text{odd}}(h)$ and full average $n^{4/5}M(h)$ is well of the order n^{-1}, as is usual for the discretization of an integral. And this also holds for $n^{3/5}M'_{\text{odd}}(h) - n^{3/5}M'(h)$ by a similar proof.

For the random term $D' - D'_{\text{odd}}$, it suffices to prove that there exists some $\tau > 0$ and C_l such that

$$\sup_{h \in [an^{-1/5}, bn^{-1/5}]} \mathrm{E}|n^{7/10}\big(D'(h) - D'_{\text{odd}}(h)\big)|^{2l} \le C_l n^{\tau l}$$

and, next, to use a partitioning argument similarly as in HHM [see also Härdle and Marron (1985b) for another use of similar technique]. The above bound can be obtained again by using Theorem 2 of Whittle (1960) similarly as in the proof of Lemma 1 and, for the quadratic term in $D'(h) - D'_{\text{odd}}(h)$, by using bounds of the form

$$\big|\big[A_h^T U A_h\big]_{ij} - \big[2A_h^T U_{\text{odd}}\, A_h\big]_{ij}\big| = O(n^{-2}h^{-2}),$$

uniformly over i and j, which results from a standard Riemann sum approximation. For the linear terms such as $-h^{-1}n^{-1}\langle A_h\boldsymbol{\varepsilon}, \mathbf{b}_h\rangle_{U-2U_{\text{odd}}}$, we use the second Whittle inequality along with the bound

$$\big|\big[A_h^T U \mathbf{b}_h\big]_i - \big[2A_h^T U_{\text{odd}}\mathbf{b}_h\big]_i\big| = O(n^{-1}h^2),$$

uniformly over i, which holds because $(d/dx)(h^{-2}b_h(x))$ is uniformly bounded. □

We finally need the following:

LEMMA 6. *We have*

$$n^{7/10}\delta'_{\text{odd}}(h_0) \to N\big(0, 2B^2 + \tfrac{3}{2}V_2\big),$$

where B and V_2 are defined in Theorem 2.1.

PROOF. We can decompose $(h/2)\delta'_{\text{odd}}(h) = T_1(h) + T_2(h)$, with $T_1(h) := n^{-1}\langle (A_h - B_h)\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}\rangle_{2U_{\text{odd}}} - \sigma^2 n^{-1}\operatorname{tr} 2U_{\text{odd}}(A_h - B_h)$ and $T_2(h) := n^{-1}\langle (\mathbf{b}_h - \mathbf{c}_h), \boldsymbol{\varepsilon}\rangle_{2U_{\text{odd}}}$. As in the proof of Lemma 4, the derivation of the limit distribution of $T_2(h_0)$ is easier than that of the quadratic term $T_1(h_0)$. The essential modification, as compared to full CV, is in the asymptotic variance of $T_1(h_0)$ obtained simply by replacing $U(A_h - B_h)$ by $2U_{\text{odd}}(A_h - B_h)$. Denoting by Q_h this latter matrix and by F the difference $K - L$, and, for notational simplic-

ity, considering the equidistant case with $u \equiv 1$, we can write

$$\begin{aligned}
&\operatorname{tr}\left(\frac{Q_h + Q_h^T}{2}\right)^2 \\
&\quad = \sum_i \sum_j \left[\frac{2(1_{\{i \text{ odd}\}})n^{-1}h^{-1}F((i-j)/nh) + 2n^{-1}h^{-1}F((j-i)/nh)1_{\{j \text{ odd}\}}}{2}\right]^2 \\
&\quad = \frac{1}{4}n^{-2}h^{-2}\left[\sum_{i \text{ odd}} \sum_j 4F^2\left(\frac{i-j}{nh}\right) + \sum_i \sum_{j \text{ odd}} 4F^2\left(\frac{j-i}{nh}\right)\right. \\
&\qquad\qquad \left. + \sum_{i \text{ odd}} \sum_{j \text{ odd}} 8F\left(\frac{i-j}{nh}\right)F\left(\frac{j-i}{nh}\right)\right] \\
&\quad \sim \frac{3}{2}h^{-1}\int F^2. \qquad \square
\end{aligned}$$

The analog of this result for PCV_k^l [which will yield the factor $((k/l) + 1)/2$ in place of $3/2$] can be obtained similarly; the modifications are only a matter of notation and attention, after having replaced $2U_{\text{odd}}$ by $(k/l)U_{k,l}$.

REFERENCES

AZARI, A. S. and MÜLLER, H-G. 1992. Preaveraged localized orthogonal polynomial estimators for surface smoothing and partial differentiation. *J. Amer. Statist. Assoc.* **87** 1005–1017.

CRAVEN, P. and WAHBA G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.

DE JONG, P. (1987). A central limit theorem for generalized quadratic forms. *Probab. Theory* **75** 261–277.

EFRON, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **82** 171–200.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

EUBANK, R. L. and WANG, S. (1994). Confidence regions in non-parametric regression. *Scand. J. Statist.* **21** 147–157.

GIRARD, D. (1987). Optimal regularized reconstruction in computerized tomography. *SIAM J. Sci. Statist. Comput.* **8** 934–950.

GIRARD, D. (1989). A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data. *Numer. Math.* **56** 1–23.

GIRARD, D. (1991). Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. *Ann. Statist.* **19** 1950–1963.

GIRARD, D. (1992). Comment on "Empirical functionals and efficient smoothing parameter selection" by P. Hall and I. Johnstone. *J. Roy. Statist. Soc. Ser. B* **54** 521.

GIRARD, D. (1993). Comment on "Discretized Laplacian smoothing by Fourier method" by O'Sullivan. *J. Amer. Statist. Assoc.* **88** 1478–1479.

GIRARD, D. (1995). On the fast Monte-Carlo cross-validation and C_L procedures: comments, new results and application to image recovery problems (with discussion). *Comput. Statist.* **10** 205–258. (Errata: **11** 87–90.)

GU, C. (1993). Interaction splines with regular data: automatically smoothing digital images. *SIAM J. Sci. Statist. Comput.* **14** 218–230.

GU, C., BATES, D. M., CHEN, Z. and WAHBA, G. (1989). The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models. *SIAM J. Matrix Anal. Appl.* **10** 457–480.

HALL, P. and JOHNSTONE, I. (1992). Empirical functionals and efficient smoothing parameter selection (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 475–530.

HALL, P. and MARRON, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74** 567–581.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *J. Amer. Statist. Assoc.* **83** 86–101.

HÄRDLE, W. and MARRON, J. S. (1985a). Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression. *Biometrika* **72** 481–484.

HÄRDLE, W. and MARRON, J. S. (1985b). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

HERRMANN, E., WAND, M. P., ENGEL, J. and GASSER, T. (1995). A bandwidth selector for bivariate kernel regression. *J. Roy. Statist. Soc. Ser. B* **57** 171–180.

HUTCHINSON, M. F. (1990). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Comm. Statist. Simulation* **19** 433–450.

KNEIP, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866.

KOHN, R., ANSLEY, C. F. and THARM, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Amer. Statist. Assoc.* **86** 1042–1060.

LI, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross-validation. *Ann. Statist.* **13** 1352–1377.

LI, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

THOMPSON, A. M., BROWN, J. C., KAY, J. W. and TITTERINGTON, D. M. (1991). A study of methods of choosing the smoothing parameter in image restoration by regularization. *IEEE Trans. Pattern Anal. Machine Intell.* **13** 326–339.

VAN ES, B. (1992). Asymptotics for least squares cross-validation bandwidths in nonsmooth cases. *Ann. Statist.* **20** 1647–1657.

WAHBA, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13** 1378–1402.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

UMR 5523
LABORATOIRE DE MODÉLISATION ET CALCUL
TOUR IRMA, BP 53X
38041 GRENOBLE CEDEX
FRANCE
E-MAIL: didier.girard@imag.fr