Estimating the accuracy of (local) cross-validation via randomised GCV choices in kernel or smoothing spline regression

Didier A. Girard

ESTIMATING THE ACCURACY OF (LOCAL) CROSS-VALIDATION VIA RANDOMIZED GCV CHOICES IN KERNEL OR SMOOTHING SPLINE REGRESSION

Didier A. Girard
CNRS and Université Joseph Fourier

Abstract

In nonparametric regression, it is generally crucial to select "nearly" optimal smoothing parameters for which a given (weighted) average squared error (Δ) is "nearly" minimized. The cross-validation (CV) selector, or the GCV selector, are popular for this task but it has been observed by many statisticians that these selectors may happen to be "not sufficiently" accurate in some situations. So a practical matter of great importance is the development of reliable estimates of this accuracy.

The purpose of this paper is to show that the simulation of the randomized GCV selector or a simple general variant using an "augmented-randomized-trace", can provide useful inferences, like consistent estimates of the standard error in the CV selector or of the expected increase of Δ due to this error. Furthermore this also provides a tool for constructing more parsimonious curve estimates having almost the same asymptotic justification as the CV estimate, namely with similar increase of Δ up to a given factor.

Rigorous proofs are given in the context of one-dimensional kernel regression. Simulated examples, also in this context, illustrate the usefulness of the methodology even at moderate sample sizes. Some direct extensions (for multi-dimensional kernels, equispaced splines) of the theoretical results are outlined. We give heuristics which indicate that the general methodology proposed in this article should be useful in many curve-, surface- or image-estimation problems when using spline-like smoothers.

Key words : Bootstrap, C_L criterion, Coverage probability, Generalized cross-validation, Kernel regression, Nonparametric regression, Randomized trace, Regularization, Smoothing spline.

1. Introduction

When regularization methods or nonparametric techniques are used to solve an inverse problem or to estimate the mean curve or surface underlying noisy observations, the choice of the regularization parameter(s), smoothing parameter(s) or bandwidth(s) is generally crucial. Cross-validation (CV), unbiased risk estimate (C_L) and generalized cross-validation (GCV) methods are very popular for this choice. They were studied and applied in a broad variety of contexts: see Craven and Wahba (1979), Rice (1984), Speckman (1985), Li (1985, 1986), Wahba (1985), Härdle and Marron (1985), Härdle, Hall and Marron (1988), Kneip (1994), Girard (1998), for theoretical results, Hutchinson (1990), Kohn et al. (1991), Thompson et al. (1989, 1991) for extensive experimentations, Eubank (1999) and Gu (2002) for recent surveys.

However, it is now well known that the resulting best bandwidth estimates may exhibit too large a sample-to-sample variability in certain applications, and thus the computed cross-validated curve or surface (from the single observed sample) should be interpreted with great caution in such a case. Various other bandwidth selectors have been recently proposed (especially for nonparametric techniques of the kernel type) which can give more stable estimates, at the price of some bias. However, their asymptotic justification needs stronger smoothness assumptions and in practice their use is not so 'automatic' (additional

parameters). As mentioned by several authors (e.g. Hart(1992) in the discussion of Hall and Johnstone (1992), Loader (1999)), cross-validation is more "assumptions-robust", and is also generally recognized as an unbiased selector: this makes cross-validation an irreplaceable tool. What is presently missing to any practitioner who uses GCV is a reasonable estimate (or bound) of how far from the optimal smoothing parameter is the GCV choice or at least some feeling about its usefulness, as was noticed by Nychka (1991) and others. Confidence interval would be even more appreciated but is a more ambitious goal. Such accuracy estimates would turn out to be very useful to guide the practitioner in his/her choices among curve-estimators of various degrees of sophistication. One instance is the choice between global and local bandwidths: Schukany (1995) aptly points out that the development of an indicator stating whether one has sufficient data to warrant more than a global bandwidth, would be very useful.

In this paper we propose a simulation-based methodology for producing estimates of the accuracy of the CV or GCV selector either in the bandwidth-space or in the curve- or surface-space. These estimators will be obtained as a simple by-product of the repeated use of the fast randomized version of GCV or of variants, where "fast" stands for meaning that one uses a single randomized-trace in place of the exact trace. In Girard (1995), computing several replications of the minimizer of the fast randomized GCV criterion was already advocated, even in contexts where exact GCV is computationally feasible, as a simple methodology for providing a useful indicator of certain "pathological" contexts: if the observable variability in the final solution (curve or surface) due to randomization, is "too large" for the problem at hand then (even exact) GCV should not be considered as "reliable" for this problem (see Sections 3.2 and 5.4 of Girard (1995)). The aim of this paper is to show that the distribution of the randomized GCV selector and a variant using an "augmented-randomized-trace", obtained by such a simulation (such a distribution will be called a randomization-based distribution) can actually provide considerably more precise inferences in contexts where GCV is relevant and where enough regularity is present. Before describing this new methodology, let us recall some definitions and previous results.

1.1. Simplified background and general definitions

Let us first consider the simple observational model

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

where m is an unknown smooth function observed at the points $x_i = i/n, i = 1, \cdots, n$ and ε_i are independent and identically distributed observation errors with mean zero and variance σ^2. Consider the widely studied kernel regression estimator for m of the form:

$$\hat{m}_h(x) = \frac{1}{nh} \sum_{j=1}^{n} K(\frac{x - x_j}{h}) y_j. \tag{1.2}$$

where K is a smooth "bell shaped" symmetric function chosen so that it satisfies $\int K(x)\ \mathrm{d}x = 1$ and $\int x^2 K(x)\ \mathrm{d}x > 0$, and $h > 0$ is the smoothing parameter (or bandwidth) to be chosen.

This curve estimator evaluated at the points x_i will also be matricially denoted by:

$$\hat{\mathbf{m}}_h = A_h \mathbf{y} \tag{1.3}$$

where the so-called smoother matrix A_h is explicit here $[A_h]_{i,j} = \frac{1}{nh} K(\frac{x_i - x_j}{h})$. This will permit us to concisely formulate some future *general* definitions and discussions which will be applicable to other common smoothers A_h like smoothing splines or "lowess" estimates (local weighted least squares).

A convenient measure for assessing $\hat{m}_h$ as an estimate of m, is the weighted average squared error

$$\Delta(h) := n^{-1} \sum \left[\hat{m}_h(x_i) - m(x_i)\right]^2 u(x_i) = n^{-1} \|A_h \mathbf{y} - \mathbf{m}\|_U^2 \tag{1.4}$$

(that we will sometimes simply call the "loss") or its expectation $M(h)$:=$E(\Delta(h))$. Here u is a fixed non-negative function (we denote by U the diagonal matrix $\mathrm{diag}(u(x_i), i = 1, \cdots, n)$ and $\|\cdot\|_U$ is the weighted l_2 norm associated with the inner product $\langle \mathbf{x}, \mathbf{y} \rangle_U = \mathbf{x}^T U \mathbf{y}$). By assuming u compactly supported on a subinterval of (0,1), the boundary effects are eliminated. By using a "localized" weight function u, like the characteristic function of a "small" subinterval, this also classically permits one to target a "local bandwidth". Let us denote by $\hat{h}_0$ the minimizer of Δ and h_0 the minimizer of M. The "*optimal bandwidth*" will refer to $\hat{h}_0$ in this paper. (For simplification of notation, the dependence of A_h, Δ, M, h_0, $\hat{h}_0$, $\hat{h}_{\mathrm{CV}}$, etc., on n are suppressed.) The popular "leave-one-out" approach (or ordinary cross-validation) for estimating the optimal bandwidth, consists, in our linear framework (1.3)-(1.4), of numerically minimizing the criterion

$$\mathrm{CV}(h) = n^{-1} \|D_h^{-1}(I - A_h)\mathbf{y}\|_U^2 \qquad D_h := \mathrm{diag}(1 - [A_h]_{i,i}, i = 1, \cdots, n) \tag{1.5}$$

which, in the particular setting here, also coincides with the popular generalized cross-validation criterion (e.g. Craven and Wahba (1979), Hastie and Tibshirani (1990) Section 3.4.3). Let $\hat{h}_{\mathrm{CV}}$ the resulting bandwidth. It is now well known that, for the context (1.1)-(1.2), under standard regularity conditions, we have $h_0 \sim c(K, m, \sigma, u) n^{-1/5}$ as $n \to \infty$ and both $\hat{h}_0/h_0$, $\hat{h}_{\mathrm{CV}}/h_0$ tend to 1 in probability, and moreover the "errors" in $\hat{h}_{\mathrm{CV}}$ or h_0 have asymptotic normal distributions:

$$\mathcal{L}\left(n^{3/10}(h_0 - \hat{h}_0)\right) \to \mathcal{N}(0, \sigma_1^2), \quad \mathcal{L}\left(n^{3/10}(\hat{h}_{\mathrm{CV}} - \hat{h}_0)\right) \to \mathcal{N}(0, \sigma_2^2),$$

where σ_1, σ_2 are constants which are known functions of K, m, σ, u; see Rice (1984), Härdle, Hall and Marron (1988) abbreviated by HHM in the following. In the latter paper, Remark 3.6, it is suggested that, by estimating the unknown terms in σ_2, this asymptotic result could be used to provide approximate confidence intervals for $\hat{h}_0$.

Before going on, we have to recall some other definitions. Let us define

$$t(h) := \mathrm{tr}UA_h/\mathrm{tr}U, \tag{1.6}$$

a trace-term which is simply $n^{-1}h^{-1}K(0)$ for the particular estimator (1.2) in the case of an equidistant design, and is thus independent of u in this case. Generalized cross-validation (GCV) is a member of a family of criteria that can be written as

$$G_{\mathrm{X}}(h) := n^{-1} \|(I - A_h)\mathbf{y}\|_U^2 \; \Xi_{\mathrm{X}}(t(h)) \tag{1.7}$$

where $\Xi_{\rm X}$ is a correction factor (or penalization) satisfying $\Xi_{\rm X}(t) = 1 + 2t + O(t^2)$ with $\Xi''_{\rm X}$ bounded on a neighborhood of 0. A list of usual penalizations $\Xi_{\rm X}$ is presented in HHM. GCV which is defined by $\Xi_{\rm GCV}(t) := (1-t)^{-2}$, is one of the most popular. It has been shown in HHM that, for the setting (1.1)-(1.4), each of these penalized criteria gives a bandwidth equivalent up to second order to $\hat{h}_{\rm CV}$ (that is, with the same normal limiting distribution as the one above). Let $\hat{h}$ generically denote any one of these GCV-type selectors.

Now, the fast randomized version ${}^{\rm R}G$ of G is obtained by using in place of $t(h)$, in the definition (1.7), the general *randomized trace function* :

$$ {}^{\rm R}t(h) := \frac{\langle \mathbf{w}, A_h \mathbf{w} \rangle_U}{\langle \mathbf{w}, \mathbf{w} \rangle_U}, $$

where $\mathbf{w}$ is a simulated unitary "white noise" vector $\mathbf{w}$ of size n, i.e. such that $E\mathbf{w} = 0$ and $\mathrm{Var}(\mathbf{w}) = I$ (the same $\mathbf{w}$ being used for every h). 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: this is not the case in the setting (1.1)-(1.4) above, but typical important 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.

In the setting above, it has been shown in Girard (1998) that

$$ \mathcal{L}\left(n^{3/10}(\hat{h}_{\rm R} - \hat{h}_0) \right) \to \mathcal{N}(0, \sigma^2_{\rm R}), $$

with $\sigma^2_{\rm R} < 2\sigma^2_2$, where $\hat{h}_{\rm R}$ denotes a generic minimizer of any one of these randomized GCV-type criteria.

1.2. Outline of results and structure of the article

In this paper, we propose a rather general method for building up inferences concerning the optimally smoothed curve, optimal in terms of the loss $\Delta(\cdot)$. It simply consists of repeatedly rerunning, for example, the randomized GCV procedure, and, in order to approximate the desired distribution of $\hat{h}_{\rm GCV} - \hat{h}_0$, of using the empirical distribution of the so-obtained randomized choices $\hat{h}_{\rm RGCV}$ which is thus *conditional to the data* $\mathbf{y}$, centered about its mean or about $\hat{h}_{\rm GCV}$ (these two centerings being asymptotically equivalent).

We claim (see Section 2, Corollary 2.3, and Remark 8.2 for extensions of the theory) that this permits to construct, under classical regularity conditions, a consistent estimate (in probability) of the distribution of $\hat{h}_{\rm GCV} - \hat{h}_0$ as $n \to \infty$. To be more precise, the obtained conditional distribution *slightly underestimates* the targeted distribution in the sense

$$ \mathcal{L}\left(\hat{h}_{\rm RGCV} - \hat{h}_{\rm GCV} \middle| \mathbf{y} \right) \text{ and } \mathcal{L}\left(\underline{\kappa}(\hat{h}_{\rm GCV} - \hat{h}_0) \right) \text{ have the same limit} $$

in probability with respect to $\mathbf{y}$, where the constant $\underline{\kappa} < 1$ is function of K and in general of m and u (and of the density of the x_i in the nonequispaced case). We used the word "slightly" because $\underline{\kappa}$ will be shown to be often "reasonably close" to 1 (Section 3) : for example, for second order kernel, if u is a characteristic function, then $1/\sqrt{3} \le \underline{\kappa} \le 1$. However, in some cases (e.g. kernel of high order) $\underline{\kappa}$ may be too small.

We then show in Sections 4 and 5 that this possible trouble can be completely eliminated via a second simulation method similar as the above one, except that the randomized trace is appropriately modified so that it has an "augmented randomness". Note, that the first simulation study, with second order kernel, that we will describe in Section 8, demonstrates that even without such a correction, these randomization-based estimates of var $\left(\hat{h}_{\mathrm{GCV}} - \hat{h}_0\right)$ are satisfactory. After some comments on how to use confidence bounds on $\hat{h}_0$, we also describe in Section 7 two other useful byproducts of a simulated population of randomized choices. An additional attractive property is that the variation of the GCV criterion over such a population yields a simple consistent estimate of the supplement of risk $\mathrm{E}\left(\Delta(\hat{h}_{\mathrm{GCV}})\right) - \mathrm{E}\left(\Delta(\hat{h}_0)\right)$ from estimation of $\hat{h}_0$ (or "risk regret"). An appealing feature of the methodology is that its theoretical justification does not require any additional smoothness assumption on m.

1.3. General heuristics and other definitions

The underlying idea can be more easily described by considering instead the Mallows unbiased risk estimate which is a basic criterion for choosing h when σ^2 is known:

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

(Mallows (1973)). The fast randomized version of CL is

$$\mathrm{RCL}(h) := n^{-1}\|(I - A_h)\mathbf{y}\|_U^2 + 2\sigma^2 n^{-1}\langle \mathbf{w}, A_h\mathbf{w}\rangle_U$$

The following discussion will remain relevant for cross-validation or GCV-type criteria, because it has been shown that typically CL (respectively RCL) produces a smoothing parameter equivalent up to second order (and essentially equivalent in practice; see e.g. Hastie and Tibshirani (1990) Section 3, Girard 1995) to any one of the G-selectors (1.6)-(1.7) (respectively ${}^{\mathrm{R}}G$-selectors). Now algebraic manipulations show that:

$$\begin{aligned} \mathrm{CL}(h) - n^{-1}\|\boldsymbol{\varepsilon}\|_U^2 - \Delta(h) &= -2n^{-1}\left(\langle \boldsymbol{\varepsilon}, A_h\boldsymbol{\varepsilon}\rangle_U - \sigma^2\mathrm{tr}UA_h\right) + 2n^{-1}\langle \boldsymbol{\varepsilon}, (I - A_h)\mathbf{m}\rangle_U \\ &= e_1(\boldsymbol{\varepsilon}, h) + e_2(\mathbf{m}, \boldsymbol{\varepsilon}, h), \end{aligned} \tag{1.8}$$

say, while the analog intrinsic error of the randomized version is

$$\mathrm{RCL}(h) - n^{-1}\|\boldsymbol{\varepsilon}\|_U^2 - \Delta(h) = e_1(\boldsymbol{\varepsilon}, h) + e_2(\mathbf{m}, \boldsymbol{\varepsilon}, h) - e_1(\boldsymbol{\varepsilon}^*, h) \tag{1.9}$$

where $\boldsymbol{\varepsilon}^* = \sigma\mathbf{w}$ is independent and distributed identically to $\boldsymbol{\varepsilon}$ at least for the first and second moments. This means that the additional "randomization error" $-e_1(\boldsymbol{\varepsilon}^*, h)$ is similar to one of the two components of the intrinsic error of CL (see Girard(1995) for an extension to the randomized GCV criterion).

Now, for the case A_h a kernel smoother, it is well known that, under classical regularity conditions, the asymptotic distribution of $\hat{h}$ is typically obtained by the following linearization of the equation $G'(\hat{h}) = 0$, where " $'$ " denotes differentiation w.r.t. h:

$$0 = G'(\hat{h}) \approx G'(h_0) + (\hat{h} - h_0)M''(h_0)$$

where the error in "≈" is small enough relatively to the centered normal limit distribution of $G'(h_0)$; and it is known that this also typically holds (with linearization errors of the same order) for the couple $G', \hat{h}$ replaced by $\mathrm{CL}', \hat{h}$ or $\Delta', \hat{h}_0$ (Rice (1984), HHM), and also by $\mathrm{RCL}', \hat{h}_{\mathrm{R}}$ as may be expected from the above comments (Girard 1998). As discussed in HHM, Nychka (1991), Girard (1998, Section3) and others, one can anticipate that the so-obtained three linearizations can yield satisfying approximations also for numerous other classes of linear smoothers A_h, provided A_h is twice continuously differentiable w.r.t. h. Now by linearly combining these three approximations using the relations (1.8) and (1.9), we obtain

$$-M''(h_0)(\hat{h} - \hat{h}_0) \approx e_1'(\boldsymbol{\varepsilon}, h_0) + e_2'(\mathbf{m}, \boldsymbol{\varepsilon}, h_0), \tag{1.10}$$

$$-M''(h_0)(\hat{h}_{\mathrm{R}} - \hat{h}) \approx -e_1'(\boldsymbol{\varepsilon}^*, h_0). \tag{1.11}$$

It is important to note that the right hand term of this last approximation is independent of $\boldsymbol{\varepsilon}$. This will permit us, under usual regularity conditions, to show the following: the observable fluctuations of $\hat{h}_{\mathrm{R}}$ for a given $\mathbf{y}$, about the associated $\hat{h}$, tend (in probability) to have the same distribution as the unconditional difference $\hat{h}_{\mathrm{R}} - \hat{h}$. Now it is often already known that the two components $e_1'(\boldsymbol{\varepsilon}, h_0)$ and $e_2'(\mathbf{m}, \boldsymbol{\varepsilon}, h_0)$ are asymptotically distributed as two independent, centered, normal variables of the same order (identical to that of $G'(h_0)$). We shall show that moreover, in the second order kernel setting, their asymptotic variances are numerically close for typical weight functions u, independently of m. So that lower bounds on the underestimation factor $\underline{\kappa}$, as the one mentioned above, will be obtained (Section 3).

A second set of heuristics is the following. In order to automatically drop this underestimation factor, the expressions (1.8)-(1.9) suggest that $\mathrm{RCL}(h)$ should ideally have a second additional error $\pm e_2(\mathbf{m}, \boldsymbol{\varepsilon}^*, h)$. If $\mathbf{m}$ were known, it would suffice to add such a term to $\mathrm{RCL}(h)$. We propose here to add $-e_2(\hat{\mathbf{m}}_g, \boldsymbol{\varepsilon}^*, h)$ where g, a pilot smoothing parameter, has to be chosen. We choose to use the sign minus and the same $\boldsymbol{\varepsilon}^*$ as the one already used for generating RCL because the whole additional error now "exactly mimics" up to the factor -1 the intrinsic error of RCL, and because the corresponding new randomized criterion, that we call the "augmented randomized" version of CL (or ARCL), then has the natural expression

$$\begin{aligned} \mathrm{ARCL}(h) &:= \mathrm{RCL}(h) - e_2(\hat{\mathbf{m}}_g, \boldsymbol{\varepsilon}^*, h) \\ &= n^{-1}\|(I - A_h)\mathbf{y}\|_U^2 + 2n^{-1}\langle \boldsymbol{\varepsilon}^*, A_h\left(A_g\mathbf{y} + \boldsymbol{\varepsilon}^*\right) - A_g\mathbf{y}\rangle_U. \end{aligned} \tag{1.12}$$

where the adjustment of the residual can be recognized similar to "a bootstrap realization of the optimism" in the work of Efron (1986). There, the average of such bootstrap realizations was used to estimate an "expected optimism" whose definition was not limited to linear $A_h(\cdot)$ (nor to quadratic Δ) in case of which it is a function of $\mathbf{m}$. As far as we know, there is no published paper touching on the methodology proposed here for kernel or spline-like smoothers where each bootstrap realization is used, via numerical minimization of ARCL, to produce a randomized choice.

The corresponding penalized criterion ${}^{\mathrm{AR}}G(h)$ naturally uses the following augmented-randomized-trace function in place of $t(h)$ in the definition (1.6) − (1.7) :

$${}^{\mathrm{AR}}t(h) := \frac{\langle \boldsymbol{\varepsilon}^*, A_h\left(A_g\mathbf{y} + \boldsymbol{\varepsilon}^*\right) - A_g\mathbf{y}\rangle_U}{\langle \boldsymbol{\varepsilon}^*, \boldsymbol{\varepsilon}^*\rangle_U}. \tag{1.13}$$

Note that contrary to ${}^{\mathrm{R}}t(h)$, this is no longer invariant by a renormalisation of $\boldsymbol{\varepsilon}^*$ and an estimate of σ is thus required for generating ${}^{\mathrm{AR}}G(h)$. In practice many good estimates exist for estimating σ. We fell that the arguments which follow should still hold with σ^2 replaced by any one of these variance estimates.

We consider in this paper, the natural automatic choice $g = \hat{h}_{\mathrm{GCV}}$ for the pilot bandwidth. Other choices might be "better" but this one is very appealing since it does *not* add any complications in practice.

Now, with enough regularity assumptions and n large enough, the following stochastic representation typically approximates well the true behavior of the minimizer of ARCL, say $\hat{h}_{\mathrm{AR}}$,

$$\begin{aligned} -M''(h_0)\left(\hat{h}_{\mathrm{AR}} - \hat{h}\right) &\approx \left(\mathrm{ARCL}' - \mathrm{CL}'\right)(h_0) \\ &\approx -e_1'(\boldsymbol{\varepsilon}^*, h_0) - e_2'(A_{h_0}\mathbf{y}, \boldsymbol{\varepsilon}^*, h_0) \\ &\approx -e_1'(\boldsymbol{\varepsilon}^*, h_0) - e_2'(\mathbf{m} + A_{h_0}\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^*, h_0) - e_2'(A_{h_0}\mathbf{m} - \mathbf{m}, \boldsymbol{\varepsilon}^*, h_0). \end{aligned} \tag{1.14}$$

The third term is typically negligible relatively to the second. Let us define

$$\begin{aligned} v_1 &:= 8n^{-2}\sigma^4 \mathrm{tr}((A'_{h_0}A_{h_0})^T U^2 A'_{h_0}A_{h_0}), \\ v_2 &:= 8n^{-2}\sigma^4 \mathrm{tr}({A'_{h_0}}^T U^2 A'_{h_0}) \\ b^2 &:= 4n^{-2}\sigma^2 \|A'_{h_0}\mathbf{m}\|^2_{U^2} \end{aligned} \tag{1.15}$$

(these are analogs of V_1, V_2 and B^2 of Theorem 2.1; note that Nychka (1991, Section 3), in addition of the bootstrap strategy, also studies an analytic strategy which uses similar formulae for general spline smoothers and replaces the unknowns terms by consistent estimates). Straightforward differentiation of e_1 and e_2 and variance calculus show, assuming for example that $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}^*$ are Gaussian, that, conditionally to a given $\mathbf{y}$, the first term and the second one are independently distributed with an approximate law for the first typically given by $\mathcal{N}(0, v_2)$ and an exact law for the second equal to $\mathcal{N}(0, 4n^{-2}\sigma^2\|A'_{h_0}\mathbf{m} + A'_{h_0}A_{h_0}\boldsymbol{\varepsilon}\|^2_{U^2})$ which, for large n, will be closed to $\mathcal{N}(0, b^2 + (1/2)v_1)$ with a high probability; and thus the conditional distribution of the sum of the three terms will typically be an inflated version of the "target" $\mathcal{N}(0, b^2 + v_2)$, i.e. the approximate unconditionned law of $-M''(h_0)(\hat{h} - \hat{h}_0)$, with inflation (or overestimation) factor given by

$$\overline{\kappa} = \sqrt{\left(b^2 + v_2 + (1/2)v_1\right) / \left(b^2 + v_2\right)} \leq \sqrt{1 + (1/2)(v_1/v_2)}.$$

We will see (Sections 5-6) how to easily eliminate this biais. However it is worth to point out that $\overline{\kappa}$ will be not much larger than 1 in many context. For example, for spline-like smoother where A_h and A'_h are symmetric matrices with a common base of eigenvectors, and where the eigenvalues of A_h are bounded by 1, we have, for example for $U = I$, that $v_1/v_2 = \mathrm{tr}(A'_{h_0}A_{h_0})^2/\mathrm{tr}(A'_{h_0})^2 \leq 1$ and thus $1 \leq \overline{\kappa} \leq \sqrt{3/2}$.

All this will be rigorously proved in the kernel context, Section 4, and extended to other contexts Remark 8.2, for large sample. Moreover we will prove in Section 5 (and 6) that we can in fact linearly combine the variance estimate obtained from the simulated $\hat{h}_{\mathrm{R}}$'s and the one from the simulated $\hat{h}_{\mathrm{AR}}$'s to obtain an asymptotically unbiased estimate of $\mathrm{var}\left(\hat{h} - \hat{h}_0\right)$.

1.4. Comparison with a more classical bootstrap methodology

Nychka (1991) has proposed a parametric bootstrap approach to approximate the distribution of $\hat{h} - \hat{h}_0$. His approach seems quite natural from a practical point of view: after having estimated m and σ using cross-validation, one can generate new independent data sets by taking these estimates in place of the true m and σ. Since typically these estimates are consistent as $n \to \infty$, one may think that these data sets will have a distribution similar as that of the observed data set for large n. One can then compute the empirical distribution of the differences between cross-validation choice and optimal choice associated to these simulated data sets and use this distribution as an approximation of the desired distribution of $\hat{h}_{\mathrm{CV}} - \hat{h}_0$. Simulations with some typical test functions m in Nychka (1991) demonstrate that this method may be rather effective. However, note that the theoretical results in Nychka (1991) state that the mean of the bootstrap distribution actually converges to a constant (< 1) multiply of h_0. Asymptotic theory suggests that this may be corrected by using an oversmoothed $\hat{m}_g$ for the generation of the pseudo-data; but in practice choosing such a pilot bandwidth g is a difficult matter. This (uncorrected) bootstrap methodology is compared to our proposal in the simulation study, Section 8. The randomized-choices methodology will prove to be more accurate in these experiments.

2. Statement of results for kernel regression estimates

For the sake of clarity and simplicity, we shall state our results for the case of one-dimensional data, in the widely studied setting of second order kernel as in Section 1.1. However, as in Girard (1998) we relax the condition of equispaced design: we assume that there exists a smooth distribution function F over $[0,1]$ such that $x_i := F^{-1}((i-0.5)/n), i = 1, \cdots, n$, and, for simplicity, that the density $f = F'$ is known. Then we can still use a simple explicit modified version of (1.2): $\hat{m}_h(x) := \frac{1}{nhf(x)} \sum K(\frac{x-x_j}{h}) y_j$, for which an asymptotic study remains relatively easy. We shall require the following classical assumptions.

a) The errors ε_i are iid with mean 0 and all other moments finite. b) K is symmetric, compactly supported and has a Hölder continuous second derivative. c) m is $C^2[0,1]$. d) 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 estimates, the minimization of the various selection criteria is assumed to be restricted to an interval $H_n = [n^{-1+\epsilon}, n^{-\epsilon}]$ for an arbitrary (small) constant $\epsilon > 0$. Let us define

$$J_u(m) = \left[\int \left((mf)'' \right)^2 f^{-1} u \right] \Big/ \int u.$$

Then it is known (e.g. Härdle and Marron (1985)), that $h_0 \sim C_0 n^{-1/5}$ with $C_0 = (C_1/C_2)^{1/5}$ and

$$C_1 = \sigma^2 \int u \int K^2, \quad C_2 = \left(\int x^2 K \right)^2 J_u(m) \int u, \tag{2.1}$$

and that $\frac{\hat{h}}{h_0} \to 1$ and $\frac{\hat{h}_0}{h_0} \to 1$ in probability. And setting $C_3 := 5C_1/C_0^3$ we have

$$M(h_0) \sim \frac{5}{4} \frac{C_1}{C_0} n^{-4/5}, \qquad M''(h_0) \sim C_3 n^{-2/5}.$$

Well known important works by Rice (1984) and HHM have established the following asymptotic stochastic behaviors of the non-randomized selectors (the extension to nonequidistant designs is analyzed in Girard 1998) that we recall for the sake of completeness. To simplify the statements of properties that are shared by various selectors, in the following, $\hat{h}$ (respectively $\hat{h}_{\mathrm{R}}$) also denotes the minimizer of $\mathrm{CV}(h)$ or the one of $\mathrm{CL}(h)$ (resp. the minimizer of $\mathrm{RCL}(h)$) in addition to the G-selectors (resp. the ${}^{\mathrm{R}}G$-selectors).

Theorem 2.1. *(HHM, Girard (1998)) Under a), b), c) and d) (i.e. the assumptions of HHM except that the design may be non-equidistant, its density f being C^2 and bounded from below on the support of u), we have*

$$\mathcal{L}\left(C_3 n^{3/10}(h_0 - \hat{h}_0)\right) \to \mathcal{N}(0, B^2 + V_1), \qquad \mathcal{L}\left(C_3 n^{3/10}(\hat{h} - \hat{h}_0)\right) \to \mathcal{N}(0, B^2 + V_2)$$

with, denoting by L the kernel $L(x) = -xK'(x)$,

$$B^2 = 4C_0^2\sigma^2 \left(\int x^2 K\right)^2 J_{u^2}(m) \int u^2,$$
$$V_1 = \frac{8}{C_0^3}\sigma^4 \int (K * K - K * L)^2 \int u^2, \qquad V_2 = \frac{8}{C_0^3}\sigma^4 \int L^2 \int u^2.$$

In the following, the convergence of the conditional distribution of a randomized quantity X_n given $\mathbf{y}$ (i.e. the randomization-based distribution of X_n) toward X will be stated in probability, in the sense that for any fixed t where the distribution of X is continuous, $|P(X_n \le t|\mathbf{y}) - P(X \le t)| \to 0$ in probability.

Theorem 2.2. *Under the assumptions of Theorem 2.1, assumptions on the generated* $\mathbf{w}$ *identical to those on $\sigma^{-1}\boldsymbol{\varepsilon}$ and assuming* $\mathbf{w}$ *independent from $\boldsymbol{\varepsilon}$,*

$$\mathcal{L}\left(C_3 n^{3/10}(\hat{h}_{\mathrm{R}} - \hat{h})\middle|\mathbf{y}\right) \to \mathcal{N}(0, V_2)$$

in probability, where the constants C_3 and V_2 are the same ones as in Theorem 2.1.

Proof: Let X_n denotes $C_3 n^{3/10}(\hat{h}_{\mathrm{R}} - \hat{h})$. The unconditional version of this result, that is, the convergence in distribution of X_n toward $\mathcal{N}(0, V_2)$, can be classically shown from the approximation

$$X_n = n^{7/10} e_1'(\boldsymbol{\varepsilon}^*, h_0) + o_P(1), \tag{2.2}$$

where e_1 is defined in (1.18), which is immediately obtained by combining the two linearizations stated in Girard (1998, equations (2.5) and (3.3)), and from the convergence in distribution (Lemma 4 of the Appendix of Girard (1998)):

$$n^{7/10} e_1'(\boldsymbol{\varepsilon}^*, h_0) \to \mathcal{N}(0, V_2). \tag{2.3}$$

Recall that one classical way to prove the well known fact that (2.2) and (2.3) are sufficient for $X_n \to \mathcal{N}(0, V_2)$, is to obtain the following inequality which holds for any $\eta > 0$ (e.g. section 20.6 of Cramer 1970):

$$|P(X_n \le t) - F(t)| \le \max\{|F_n(t+\eta) - F(t)|, |F_n(t-\eta) - F(t)|\} + 2P\left(|r_n(h_0, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^*)| > \eta\right),$$

where F_n is the distribution of $n^{7/10}e_1'(\boldsymbol{\varepsilon}^*, h_0)$, F is the distribution $\mathcal{N}(0, V_2)$, and $r_n(h_0, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^*) = X_n - n^{7/10}e_1'(\boldsymbol{\varepsilon}^*, h_0)$; next, to observe that this bound can be made arbitrarily small with an appropriate choice of η and with n large enough.

Concerning the conditional X_n, by noticing that F_n is also the distribution of $n^{7/10}e_1'(\boldsymbol{\varepsilon}^*, h_0)|\mathbf{y}$, it is easy to see that this inequality actually also holds with the probability $P(\cdot)$ replaced by the conditional probability $P(\cdot|\mathbf{y})$. Now, it suffices to observe that the so-obtained bound of $|P(X_n \leq t|\mathbf{y}) - F(t)|$ can also be made arbitrarily small in probability since, of course, $\alpha_n(\mathbf{y}) = P(|r_n(h_0, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^*)| > \eta|\mathbf{y}) \to_P 0$ (which results from $E|\alpha_n(\mathbf{y})| = P(|r_n(h_0, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^*)| > \eta) \to 0$ and Markov inequality). ■

Recalling that all the selectors are restricted to a compact set, one can define $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$ by

$$\widehat{\mathrm{SD}}_{\mathrm{RAND}} := \mathrm{var}^{\frac{1}{2}}\left(\hat{h}_{\mathrm{R}}\middle|\mathbf{y}\right). \tag{2.4}$$

We could have used $\sqrt{E\left((\hat{h}_{\mathrm{R}} - \hat{h})\middle|\mathbf{y}\right)^2}$ instead of (2.4). For large samples, this would be equivalent. However, for finite samples this is, of course, always greater than $\mathrm{var}^{\frac{1}{2}}\left(\hat{h}_{\mathrm{R}}\middle|\mathbf{y}\right)$, and might then improve it in view of the underestimator factor we discuss in the following. We come back to this point in the simulation study.

By simply combining Theorem 2.1 and Theorem 2.2 and using the expression of C_0^5 given by (2.1), we now obtain the following corollary which describes the behavior of the variances of the normal asymptotic laws (that we simply call asymptotic variances):

Corollary 2.3. *Under the assumptions of Theorem 2.2,* $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$*, defined in* (2.4)*, is a consistent estimate of* $\underline{\kappa}\,\mathrm{var}^{\frac{1}{2}}\left(\hat{h} - \hat{h}_0\right)$*, in the sense that* $\hat{h}_{\mathrm{R}}|\mathbf{y}$ *has the same asymptotic variance than* $\underline{\kappa}(\hat{h} - \hat{h}_0)$*, where*

$$\underline{\kappa} = \left(\frac{V_2}{B^2 + V_2}\right)^{1/2} = \left(\frac{1}{2}\frac{J_{u^2}(m)}{J_u(m)}\frac{\int K^2}{\int L^2} + 1\right)^{-1/2} \leq 1.$$

We will study in more depth this downward bias of $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$ in the next section.

3. The underestimation factor $\underline{\kappa}$ is never "too small" for second order kernel

At first look, one may think that the underestimation factor $\underline{\kappa}$ should be estimated in order to appropriately inflate $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$. However, we show in this section that for second order kernel and typical u, V_2 can never be a small part of the whole variance $B^2 + V_2$ and thus $\underline{\kappa}$ is reasonably close to 1. From the expression of $\underline{\kappa}$ in Corollary 2.3, we see that an important case can be first considered: the case u^2 proportional to u since $\underline{\kappa}$ is then independent of m and of the density f. We thus give in Table 3.1 the values of this factor $\underline{\kappa}$ in this case, for four typical kernels K, with a now standard terminology (e.g. Priestley and Chao (1972)). Notice that the first and last columns are extremal cases (see Lemma 3.1 below). It is seen that these values of $\underline{\kappa}$ are all reasonably close to 1.

K	"Rectangle"	Quadratic	Biweight	Gaussian	$\exp(-\lvert x\rvert)$	"$K_{\min}$"
$\frac{\int K^2}{\int L^2}$	0	$\frac{2}{3}$	1	$\frac{4}{3}$	2	4
$\underline{\kappa}$	1	.8660	.8165	.7746	.7071	.5774

Table 3.1. Case u^2 proportional to u

In order to derive general lower bounds for $\underline{\kappa}$ (like the last column of Table 3.1), let us state the two following lemmas:

Lemma 3.1. *Assume that K is a continuously differentiable function with compact support. Then*

$$\int K^2 \leq 4 \int |xK'|^2 = 4 \int L^2.$$

Proof: By integration by parts, we have $\int K^2 = -\int x(2K'K)$ whose square is bounded by $\int (2xK')^2 \int K^2$ by Cauchy-Scharwz.

Furthermore, it can be easily shown that this inequality is "sharp", i.e. there exists a minimizing sequence of kernels (denoted "$K_{\min}$" in Table 3.1) for which the ratio $\int |xK'|^2 / \int K^2$ converges to the lower bound 1/4.

Lemma 3.2. *Under the assumptions of Theorem 2.1,*

$$\frac{J_{u^2}(m)}{J_u(m)} \leq \frac{\int u}{\int u^2} \max u.$$

Proof: Note that this ratio can be written $\frac{\int u}{\int u^2} \frac{\int gu}{\int g}$ for a certain positive function g.

It immediately follows that:

Theorem 3.3. *If $u^2 \propto u$ then the underestimation factor $\underline{\kappa}$ which appears in Corollary 2.3 always satisfies:*

$$1/\sqrt{3} \leq \underline{\kappa} \leq 1.$$

Otherwise, this still holds, for any m and f, with $\sqrt{3}$ replaced by $\sqrt{1 + 2\frac{\int u}{\int u^2} \max u}$. When the chosen u is a Gaussian density of any width, $\frac{\int u}{\int u^2} \max u = \sqrt{2}$ if the integrals are taken over $(-\infty, +\infty)$.

The case u Gaussian considered here is of practical interest since it is a natural weight function often used to define local bandwidths (e.g. Vieu (1991)). Note that if the width of a Gaussian u is chosen to be small enough relatively to the local variation of $(mf)'' f^{-1}$, then $J_{u^2}(m)/J_u(m)$ will actually be closer to 1 than to the approximate upper bound $\sqrt{2}$.

4. Augmenting the randomization error in $\hat{h}_{\mathrm{R}}$ for another accuracy estimate

We have seen in the previous section, that the intrinsic error (1.9) of the randomized criterion RCL has an additional error (compared to the error (1.8) of CL) which actually may be not large enough for

our precise objective here. And this becomes even more acute for kernel of higher order, as detailed in Remark 8.2. So we consider the augmented-randomized-version of CL (i.e. ARCL defined in (1.12)) as well as ${}^{\rm AR}G(h)$ using the augmented-randomized-trace function defined in (1.13).

To simplify the presentation, we assume σ is known in (1.12)-(1.13). However many estimates with good asymptotic property exist for estimating σ and the following result should remain true when replacing σ by one of them, for example the estimate of Rice (1984).

As discussed in Section 1.3, we use here the natural automatic choice $g = \hat{h}_{\rm GCV}$ for the pilot bandwidth. Note that the condition on g in Theorem 4.1 below is then satisfied.

Assuming also the standard conditions of Theorem 2.2, it can be shown that the minimizer of ARCL(h) and those of any one criterion of the family ${}^{\rm AR}G(h)$ are all equivalent up to second order. So, let $\hat{h}_{\rm AR}$ denote a generic one of these selectors.

We claim that the simulation of $\hat{h}_{\rm AR}$ given $\mathbf{y}$, provides useful inferences on the underlying distribution of $\hat{h} - \hat{h}_0$ (by comparing Theorem 4.1 to Theorem 2.1). It is important to note that we do not require any additional smoothness condition on the underlying $m(\cdot)$. Note that our approach is quite different from the natural bootstrap approach where generated pseudo-data $\mathbf{y}^* = \hat{\mathbf{m}}_g + \boldsymbol{\varepsilon}^*$ would produce "bootstrap replications" of both $\hat{h}$ and $\hat{h}_0$, as considered by Nychka (1991).

Theorem 4.1. *Under the assumptions of Theorem 2.2, and assuming that the pilot bandwidth g (random or not) used in* ARCL *or in* ${}^{\rm AR}G$ *satisfies, for some $\epsilon > 0$, $|g - h_0|/h_0 = O_p(n^{-\epsilon})$, then*

$$\mathcal{L}\left(C_3 n^{3/10}(\hat{h}_{\rm AR} - \hat{h})\big|\mathbf{y}\right) \to \mathcal{N}(0, V_2 + B^2 + \frac{1}{2}V_1) \quad \textit{in probability,}$$

where the constant C_3, B, V_1 and V_2 are the same ones as in Theorem 2.1.

Proof: The first step is to show that there still exists a approximate linearized form for expressing $\hat{h}_{\rm AR}$. We only sketch its derivation because it uses similar lines of proof as for $\hat{h}$ in HHM or $\hat{h}_{\rm R}$ in Girard (1998). For example, letting B_h denote the smoothing operator associated with the kernel L, what is required in supplement of (A.8) of HHM is now

$$\sup_{\substack{h,g \\ |h-h_0|+|g-h_0| \le n^{-1/5-\epsilon}}} \left|\frac{1}{nh}\langle (A_h - B_h)A_g\mathbf{y}, \boldsymbol{\varepsilon}^*\rangle_U - \frac{1}{nh_0}\langle (A_{h_0} - B_{h_0})A_{h_0}\mathbf{y}, \boldsymbol{\varepsilon}^*\rangle_U\right| = o_P(n^{-7/10})$$

which can be shown by similar bounds and partitioning argument to those used in HHM. This permits us to replace both $\hat{h}_{\rm AR}$ and g by h_0 in the right hand term of the approximation $\hat{h}_{\rm AR} - h_0 \approx -\big(\text{ARCL}' - M'\big)\,(\hat{h}_{\rm AR})/M''(h_0)$ obtained by standard first order Taylor expansion and then, subtracting the known analog Taylor approximation of $\hat{h} - h_0$, to derive (with e_1 and e_2 defined as in (1.8))

$$-M''(h_0)\left(\hat{h}_{\rm AR} - \hat{h}\right) = -e_1'(\boldsymbol{\varepsilon}^*, h_0) - e_2'(A_{h_0}\mathbf{y}, \boldsymbol{\varepsilon}^*, h_0) + o_P(n^{-7/10}).$$

The second step is to establish the limiting normal distribution of $e_2'(A_{h_0}\mathbf{y}, \boldsymbol{\varepsilon}^*, h_0)$ conditioned to $\mathbf{y}$. First, we can imitate the proof for $\mathcal{L}\left(e_2'(\mathbf{m}, \boldsymbol{\varepsilon}, h)/\text{var}^{1/2}e_2'(\mathbf{m}, \boldsymbol{\varepsilon}, h)\right) \to \mathcal{N}(0,1)$ which uses the Lindeberg condition

(e.g. Eubank and Wang 1994). For this, it suffices to observe that $e_2'(A_{h_0}\mathbf{y},\boldsymbol{\varepsilon}^*,h_0) = 2n^{-1}{h_0}^{-1}\langle(B_{h_0} - A_{h_0})A_{h_0}\mathbf{y},\boldsymbol{\varepsilon}^*\rangle_U$ where the vector $(B_{h_0} - A_{h_0}))A_{h_0}\mathbf{y}$ has ith element

$$ {h_0}^2\left(\frac{1}{2}\int x^2(K-L)\right)(\hat{m}_{h_0}f)''(x_i)f^{-1}(x_i) + o_P({h_0}^2), $$

uniformly in i, and quadratic mean

$$ n^{-1}||(B_{h_0} - A_{h_0}))A_{h_0}\mathbf{y}||_{U^2}^2 = h_0^4\left(\frac{1}{2}\int x^2(K-L)\right)^2 J_{u^2}(\hat{m}_{h_0})\int u^2(1+o_P(1)). $$

Second, the additional term $\frac{1}{2}V_1$ results from $n^{-1}||(B_{h_0} - A_{h_0})A_{h_0}\mathbf{y}||_{U^2}^2 \sim n^{-1}||(B_{h_0} - A_{h_0})A_{h_0}\mathbf{m}||_{U^2}^2 + \sigma^2 n^{-1}{h_0}^{-1}\int(K*K - K*L)^2\int u^2$, in probability, where the two terms are of the same order and can be checked to be proportional to B^2 and $\frac{1}{2}V_1$ (for this, note $\int x^2(L-K) = 2\int x^2K$). ■

Let us now define $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$ by

$$ \widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}} := \mathrm{var}^{\frac{1}{2}}\left(\hat{h}_{\mathrm{AR}}\middle|\mathbf{y}\right), \tag{4.1} $$

then

Corollary 4.2. *Under the assumptions of Theorem 4.1, $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$, defined in (4.1), is a consistent estimate of $\overline{\kappa}\,\mathrm{var}^{1/2}(\hat{h} - \hat{h}_0)$, in the sense that $\hat{h}_{\mathrm{AR}}|\mathbf{y}$ has the same asymptotic variance than $\overline{\kappa}(\hat{h} - \hat{h}_0)$, where*

$$ \overline{\kappa}^2 = \frac{B^2 + V_2 + \frac{1}{2}V_1}{B^2 + V_2} = 1 + \frac{1}{2}\frac{V_1}{V_2}\underline{\kappa}^2 \geq 1. $$

It is important to point out that the overestimation factor $\overline{\kappa}$ is also typically reasonably close to 1. For example, for any positive K, it can be shown that $V_1 \leq V_2$ (e.g. HHM) and thus $\overline{\kappa} \leq \sqrt{3/2} = 1.2247$. If K is the Gaussian kernel, we obtain the numerical value $V_1/V_2 = \sqrt{2}/8$ and thus $\overline{\kappa} \leq \sqrt{1 + \sqrt{2}/16} = 1.0432$. Such an appealing closeness to 1 will be also obtained for the cubic smoothing spline setting, Section 6.

5. Estimating $\mathrm{var}(\hat{h} - \hat{h}_0)$ without asymptotic bias

In the general case where $u^2 \not\propto u$, neither $\overline{\kappa}$ nor $\underline{\kappa}$ are known, but they are known functions of the functionals $J_u(m)$ and $J_{u^2}(m)$. So, one might be tempted to work with standard estimates of these functionals. In fact, we do not need such estimates to derive a consistent estimate of $\mathrm{var}(\hat{h} - \hat{h}_0)$. Indeed,

Theorem 5.1. *Under the assumptions of Theorem 2.1, and for $\underline{\kappa}$ and $\overline{\kappa}$ given in Corollaries 2.3 and 4.2, let us assume that $\left(n^{3/10}\widetilde{\mathrm{SD}}_{\mathrm{RAND}}/\underline{\kappa}\right)^2$ and $\left(n^{3/10}\widetilde{\mathrm{SD}}_{\mathrm{AUG-RAND}}/\overline{\kappa}\right)^2$ both converge in probability toward the asymptotic variance of $n^{3/10}(\hat{h} - \hat{h}_0)$; then*

$$ \left(n^{3/10}\widetilde{\mathrm{SD}}_{\mathrm{AUG-RAND}}\right)^2 - \frac{1}{2}\frac{\int(K*K - K*L)^2}{\int L^2}\left(n^{3/10}\widetilde{\mathrm{SD}}_{\mathrm{RAND}}\right)^2 \tag{5.1} $$

converges in probability toward the asymptotic variance of $n^{3/10}(\hat{h} - \hat{h}_0)$.

Proof: It can be derived from Corollary 4.2 that $\overline{\kappa}^2/\underline{\kappa}^2 = 1/\underline{\kappa}^2 + (1/2)(V_1/V_2)$, where $\frac{V_1}{V_2}$ is a constant depending only on K. On the other hand, $\widetilde{\mathrm{SD}}_{\mathrm{AUG-RAND}}/\widetilde{\mathrm{SD}}_{\mathrm{RAND}}$ is a consistent estimate of $\frac{\overline{\kappa}}{\underline{\kappa}}$. Thus a consistent estimate of $\underline{\kappa}$ (or of $J_{u^2}(m)/J_u(m)$) is deduced of this ratio. Plugging it in $\left(n^{3/10}\widetilde{\mathrm{SD}}_{\mathrm{RAND}}/\underline{\kappa}\right)^2$, it is checked this is equivalent to use the stated linear combination of $\widetilde{\mathrm{SD}}^2_{\mathrm{AUG-RAND}}$ and $\widetilde{\mathrm{SD}}^2_{\mathrm{RAND}}$. ■

Note that the convergence in probability of $n^{3/10}\widehat{\mathrm{SD}}_{\mathrm{RAND}}$ or $n^{3/10}\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$ is not claimed in Corollaries 2.3 or 4.2; indeed it is well known that convergence in distribution of a random sequence does not imply convergence in moment. A classical technical condition sufficient for this, is a uniform integrability condition that we were not able to obtain here. Nevertheless Theorem 5.1 is already useful for percentile-based variance estimates (such as those in (7.3) below) in place of empirical moments: such estimates are classical in case of asymptotic normality and are even often proposed as robust standard deviation estimates (e.g. Efron (1982)); their consistency holds true in our setting (see Section 7.2).

6. Extension to other nonparametric regression settings

From the heuristics of Section 1.3 (mathematical rigor as for the theorems above, is clearly possible in some other settings as outlined in Remark 8.2) these accuracy estimators have a large potential but the estimator given Theorem 5.1 is, at first glance, not as easily applicable as $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$ or $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$ since the appropriate general factor $-(1/2)(v_1/v_2)$ in this linear combination (where v_1 and v_2 are defined in (1.5)) is specific to the setting. However, note that for all the settings where v_1/v_2 is independent of h_0, a general version of this estimate could be constructed simply be replacing this constant by an approximation that can be obtained, for a given family A_h of smoothing operators, by a natural simulation based approach as follows: choose one (or several) realistic "true" function m's and a corresponding σ, simulate several data-sets for each m, and adjust the above linear combination so as to optimize its observed average performance (in a least-squares sense for example) on these simulated problems. This should work in the general smoothing spline setting where $A_h = (I + h\Omega)^{-1}$ with Ω a given symmetric ≥ 0 matrix, at least for $u \equiv 1$. Indeed it can be checked, by using that $A_h' = -h(I - A_h)A_h$ in the formulae given Section 1.3, that the appropriate constant in this linear combination becomes

$$-(1/2)(v_1/v_2) = -(1/2)\mathrm{tr}A_h^4(I - A_h)^2/\mathrm{tr}A_h^2(I - A_h)^2 \tag{6.1}$$

where, strictly speaking, h should be set to the unknown h_0; but it is known in the spline literature that such a ratio of traces is typically asymptotic to a constant independent of h (e.g. Nychka (1990) for a proof of this for the one-dimensional polynomial smoothing spline). In fact, v_1/v_2 can be simply evaluated form the expression (6.1) where the replacement of h_0 by $\hat{h}_{\mathrm{GCV}}$ should not degrade the approximation.

Moreover numerical values close to a few percents can often be expected for this ratio, in case of which the correction of $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$ is not really mandatory. Indeed for example in the cubic spline setting one obtains (e.g. Kou 2003) $(1/2)\mathrm{tr}A_h^4(I - A_h)^2/\mathrm{tr}A_h^2(I - A_h)^2 \sim (1/2)\left[\Gamma(15/4)/\Gamma(6)\right] / \left[\Gamma(7/4)/\Gamma(4)\right] = 0.118$.

7. Various uses of these accuracy estimates

7.1. The empirical distribution of the randomized GCV choices, and the one for the augmented-randomized version, both obtained by simulation, can be used in several different ways. The first one is classical and was already developed in Nichka (1991) : even if, strictly speaking, such a simulation furnishes a consistent, say 95%, confidence interval for the difference $\hat{h} - \hat{h}_0$, it is natural to very simply transform it (by translating it by $\hat{h}$) in a prediction (called "confidence" in Nichka (1991)) interval for $\hat{h}_0$ which can be interpreted as follows. For the data in hand, the Δ-optimal $\hat{h}_0$ can be claimed to belong to the produced interval with a confidence level of 95%, that is, the claim being incorrect for only 5% (assuming that the large sample approximation is correct) of the possible replicated data sets from the stochastic structure of the problem at hand. If a data analyst draws the curve estimates for every bandwidth in this interval and observes "little" changes at each design point of the essential support of the weight function u, then he/she may be assured, with a confidence level of 95%, that he/she has found the Δ-optimal fit (up to the mentioned "little changes" in it). The epithet "little" for a change in the curve at a design point may be made more precise by adding the phrase "relatively to the width of a confidence band for the curve around this point" (we come back to the topic of confidence bands in Section 7.3) since perturbing the center of a such band by "little" changes will then have "little" consequence on its coverage probability.

As advocated by Nychka (1991), a good practice is to report the 2 curve- or surface-estimates which correspond to the lower and upper points of the produced interval. When, contrarily to the first situation, the two fits associated with the endpoints of such a $(1-\alpha)100\%$ confidence interval have rather different aspects as α passes from, say, 30% to 5%, the choice of α has a large influence on the report for the data analyst. Note that, concerning two different bandwidths which are in a such 95% confidence interval for $\hat{h}_0$, one can say that there is no objective reasons to think that one of the two corresponding fits should be better (for the Δ criterion) than the other. In the following we state a further interpretation of the two fits reported, which may give a guide for the choice of α in terms of expected loss.

7.2. Consider, again to simplify the presentation, the setting of Sections 2-5. Let τ a given constant and let $\hat{C}$ any statistic (i.e. a function of $\mathbf{y}$) assumed to be a consistent estimate of $\tau(B^2+V_2)^{1/2}/C_3$; and consider $\hat{h}_{\mathrm{GCV}}(\tau) = \hat{h}_{\mathrm{GCV}} + n^{-3/10}\hat{C}$ a perturbed version of $\hat{h}_{\mathrm{GCV}}$. Then it can be shown that $C_3 n^{3/10}(\hat{h}_{\mathrm{GCV}}(\tau) - \hat{h}_0)$ is asymptotically distributed as $(B^2+V_2)^{1/2}\mathcal{N}(\tau,1)$ and that the known asymptotic laws for the "excess loss" (i.e. the difference between the (data-driven) resulting Δ and the best possible Δ)

$$\mathcal{L}\left(n\left[\Delta(\hat{h}) - \Delta(\hat{h}_0)\right]\right) \to \frac{1}{2C_3}(B^2+V_2)\chi_1^2 \tag{7.1}$$

which are stated in HHM and Girard (1998), can be easily generalized for $\hat{h}_{\mathrm{GCV}}(\tau)$ in place of $\hat{h}$: the standard χ_1^2 must simply be replaced by decentred chi-square $[\mathcal{N}(\tau,1)]^2$ (the proofs are very similar to the ones in HHM, Girard (1998)) and are not repeated here).

On the other hand, for $\alpha, \beta \in (0, 0.5)$ let us denote by $\underline{\hat{h}_{\mathrm{R}}}(\beta)$ and $\overline{\hat{h}_{\mathrm{R}}}(\alpha)$ the (β)100th and $(1-\alpha)$100th percentiles of the randomization-based distribution, i.e. defined by

$$P(\hat{h}_{\mathrm{R}} \leq \underline{\hat{h}_{\mathrm{R}}}(\beta)\big|\mathbf{y}) = \beta, \quad P(\hat{h}_{\mathrm{R}} \leq \overline{\hat{h}_{\mathrm{R}}}(\alpha)\big|\mathbf{y}) = 1 - \alpha \tag{7.2}$$

Besides providing a standard simulation-based $(1-(\alpha+\beta))100\%$ confidence interval, these percentiles also furnish variance estimates whose consistency can be proven by classical techniques, namely

$$n^{3/10}\frac{\overline{\hat{h}_{\mathrm{R}}}(\alpha) - \underline{\hat{h}_{\mathrm{R}}}(\beta)}{z_{1-\alpha} - z_{\beta}}, \quad n^{3/10}\frac{\overline{\hat{h}_{\mathrm{R}}}(\alpha) - \hat{h}}{z_{1-\alpha}} \quad \text{or} \quad n^{3/10}\frac{\hat{h} - \underline{\hat{h}_{\mathrm{R}}}(\beta)}{-z_{\beta}} \tag{7.3}$$

where z_{β} and $z_{1-\alpha}$ are the (β)100th and $(1-\alpha)$100th percentiles of the $\mathcal{N}(0,1)$, both converge in probability toward $\underline{\kappa}(B^2+V_2)^{1/2}/C_3$. By combining the above statement, using $z_{1-\alpha}$ times the second of this standard deviation estimates in place of $\hat{C}$ in the above perturbation, which implies the coincidence $\hat{h} + n^{-3/10}\hat{C} = \overline{\hat{h}_{\mathrm{R}}}(\alpha)$, we obtain

Theorem 7.1. *Let $\alpha \in (0, 0.5)$. Under the assumptions of Theorem 2.2,*

$$\mathcal{L}\left(n\left[\Delta\left(\overline{\hat{h}_{\mathrm{R}}}(\alpha)\right) - \Delta(\hat{h}_0)\right]\right) \to \frac{1}{2C_3}(B^2+V_2)\Big(\mathcal{N}(z_{1-\alpha}\underline{\kappa}, 1)\Big)^2$$

where the constants C_3, B, V_2 and $\underline{\kappa}$ are the same ones as in Theorem 2.1 and Corollary 2.3.

By combining this with the expression (7.1) for the asymptotic law of $\Delta(\hat{h}) - \Delta(\hat{h}_0)$, one obtain a very simple bound about what is sacrificed, comparatively to the CV choice, in the fit corresponding to the upper point $\overline{\hat{h}_{\mathrm{R}}}(\alpha)$, in terms of relative increase of the excess loss, in average, also called relative "risk regret" :

$$\frac{\mathrm{E}\left[\Delta\left(\overline{\hat{h}_{\mathrm{R}}}(\alpha)\right) - \Delta(\hat{h}_0)\right]}{\mathrm{E}\left[\Delta(\hat{h}) - \Delta(\hat{h}_0)\right]} = 1 + (z_{1-\alpha}\underline{\kappa})^2 \leq 1 + z_{1-\alpha}^2 \tag{7.4}$$

where E denotes here the asymptotic variance.

For example, a data analyst concerned by data reduction will look after a kind of "parsimonious yet near optimal" fit of his data : Theorem 7.1 implies that the fit corresponding to the upper point $\overline{\hat{h}_{\mathrm{R}}}(0.16)$ has a theoretical justification "almost" as good as the CV fit, namely the relative risk regret is bounded by $1 + z_{0.84}^2 = 2$. As another example, by choosing $\overline{\hat{h}_{\mathrm{R}}}(0.32)$, the bound becomes $1 + z_{0.68}^2 = 1.21$.

Recall that modifying the CV choice toward more smoothing is commonly made in practice (e.g. Gu (2002) Section 6.3.2): the standard modification consists of multipying the trace term in $\mathrm{CL}(h)$ (or $t(h)$ in the GCV-like criteria) by a factor larger than 1 (e.g. 1.4). This is had hoc but often works well. It would certainly be interesting to compare these two ways of producing "more parsimonious yet near optimal" fits.

It is obvious that an analog of Theorem 7.1 exists for a lower point : the Δ performance of the fit obtained with $\underline{\hat{h}_{\mathrm{R}}}(\beta)$ in place of $\overline{\hat{h}_{\mathrm{R}}}(\alpha)$ has the same expression with $\mathcal{N}(z_{1-\alpha}\underline{\kappa}, 1)$ replaced by $\mathcal{N}(-z_{\beta}\underline{\kappa}, 1)$.

Using such a fit might be useful for a data analyst concerned by discovering features (like peaks) in a data set.

Analogs of these results can also easily be stated for an upper point and a lower point (denoted respectively by $\overline{\hat{h}_{\mathrm{AR}}}(\alpha)$ and $\underline{\hat{h}_{\mathrm{AR}}}(\beta)$) of the empirical conditional distribution of repeated bandwidth choices by the augmented-randomized-trace version. The only change is that $\underline{\kappa}$ is then replaced by $\overline{\kappa}$.

7.3. One *can* in fact estimate the absolute value of the risk excess itself, simply from the randomization-based distribution. For this task one might estimate the curvature of $\Delta(\cdot)$ near $\hat{h}_0$, estimate $\mathrm{E}\left(\hat{h}-\hat{h}_0\right)^2$ and invoke the usual second order Taylor approximation $\Delta(\hat{h})-\Delta(\hat{h}_0) \approx (1/2)\left(\hat{h}-\hat{h}_0\right)^2 \Delta''(h_0)$. But it can be shown, as a consequence of the previous results, that the variations of, say, $\mathrm{GCV}(\cdot)$ over the population of the randomized choices, thus conditional to the observed $\mathbf{y}$, furnishes (at least) as much information :

Theorem 7.2. *Let G be any one of the G-selectors defined in (1.6)-(1.7) and $\hat{h}$ (resp. $\hat{h}_{\mathrm{R}}$ and $\hat{h}_{\mathrm{AR}}$) the minimizer of G (resp. of its randomized-trace version and its augmented-randomized-trace version). Under the assumptions of Theorem 4.1,*

$$\mathcal{L}\left(n\left[G(\hat{h}_{\mathrm{R}})-G(\hat{h})\right]\Big|\mathbf{y}\right) \to \underline{\kappa}^2 \frac{1}{2C_3}(B^2+V_2)\chi_1^2$$

$$\mathcal{L}\left(n\left[G(\hat{h}_{\mathrm{AR}})-G(\hat{h})\right]\Big|\mathbf{y}\right) \to \overline{\kappa}^2 \frac{1}{2C_3}(B^2+V_2)\chi_1^2$$

in probability, where the constants C_3, B, V_2 , $\underline{\kappa}$ and $\overline{\kappa}$ are the same ones as in Theorem 2.1 and Corollaries 2.3 and 4.2.

Proof: Similar steps as in the proof of Theorems 2.2 and 4.1, along with the stochastic approximations for $\Delta(h)-\Delta(\hat{h}_0)$ developed in HHM and Girard (1998) for $h=\hat{h},\hat{h}_{\mathrm{R}}$, give the results. ■

Either by sample moment estimates or by appropriate percentiles-based estimates, one can then produce a "sandwich" of estimates of lower and upper bounds reasonably (from the "nearness" of $\underline{\kappa}$ and $\overline{\kappa}$ to 1) closed to the risk regret due to the inaccuracy of CV. One example of percentiles-based estimate, denoted by $\widetilde{\mathrm{RR}}_{\mathrm{RAND}}$ for future references, is simply the median of the empirical conditional distribution of $G(\hat{h}_{\mathrm{R}})-G(\hat{h})$ divided by the median (≈ 0.45) of χ_1^2. $\widetilde{\mathrm{RR}}_{\mathrm{AUG-RAND}}$ can be defined similarly with $\hat{h}_{\mathrm{R}}$ replaced by $\hat{h}_{\mathrm{AR}}$. From Theorem 7.2, these 2 estimates can be easily proven to be consistent up to the bias factor $\underline{\kappa}^2$ and $\overline{\kappa}^2$ respectively. By an argument similar as in the proof of Theorem 5.1, it is seen that

$$\widetilde{\mathrm{RR}}_{\mathrm{AUG-RAND}} - \frac{1}{2}\frac{V_1}{V_2}\widetilde{\mathrm{RR}}_{\mathrm{RAND}} \tag{7.5}$$

is an asymptotically unbiased estimate of the asymptotic risk regret from using $\hat{h}$, $\mathrm{E}\left[\Delta(\hat{h})-\Delta(\hat{h}_0)\right]$. Note that when $u^2 \propto u$, one can check that the ratio $(B^2+V_2)/C_3$ is in fact independent of m (e.g. Section 4.4.5 of Hall and Johnstone 1992 for the case $f\equiv 1$); but this does not hold any longer when $u^2 \not\propto u$.

Let us mention how these simple risk-regret estimators might be used in the production of inferences about m. Of course for any selector $h(\mathbf{y})$ the risk $E\Delta(h(\mathbf{y}))$ measures the (local) performance of $\hat{m}_{h(\mathbf{y})}$; one of the possible uses of an estimate of this risk is the following:

A methodology which is now very popular, in the smoothing spline setting, for accompanying the point estimate $\hat{m}_{\hat{h}}(x)$, with an interval estimate, is the so-called 90% Bayesian confidence interval

$$\hat{m}_{\hat{h}}(x) \pm z_{0.05}\sigma\sqrt{\mathrm{tr}UA_{\hat{h}}/\mathrm{tr}U} \tag{7.6}$$

which is well known, at least for the one-dimensional polynomial spline and $u \equiv 1$, to have good frequentist "across the curve" properties for m deterministic provided $\hat{h}$ is a good estimate of $\hat{h}_0$ or of h_0. Such an interval indeed takes account the fact that $\hat{m}_{\hat{h}}$ is biaised. However a possibly very serious weakness is that the error in $\hat{h}$ is not taken in account. Indeed the true $(1-\gamma)100\%$ interval for any $h(\mathbf{y})$ should be

$$\hat{m}_{h(y)}(x) \pm z_{\gamma/2}\sqrt{E\Delta(h(\mathbf{y})/\mathrm{tr}U} \tag{7.7}$$

while the justification of (7.6) stems on a consistency result $\sigma^2\mathrm{tr}UA_{\hat{h}} \approx E\Delta(\hat{h})$ which is in fact a consistency toward $E\Delta(h_0)$ and thus neglect the risk-regret term(s). Before to see how this could be attacked, we point out that the theoretical justification of (7.7) developed by Nychka (1990) can be easily seen to also holds for a weight function $u \not\equiv 1$ and more general smoothing spline or kernel estimates: the only required assumptions are that the associated smoothing operator is symmetric and that the chosen weight function u is smooth enough so that $A_h\mathbf{u} \approx \mathbf{u}$ (where $\mathbf{u} = (u(x_1, \cdots, u(x_n))^T$) over the domain of considered h's. The meaning of "across the curve" is now that the coverage probability of the interval (7.7) is not only relatively to the distribution of $\boldsymbol{\varepsilon}$ but also for x drawn with the discrete probability which assigns the mass $u(x_i)/\mathrm{tr}U$ at each x_i.

Now by decomposing $\Delta(h(\mathbf{y}) = \Delta(\hat{h}_0) + \Big(\Delta(h(\mathbf{y}) - \Delta(\hat{h}_0)\Big)$ we see that

1. The risk regret from using $h(\mathbf{y})$, i.e. the expectation of the second term, is a minorant of $E\Delta(h(\mathbf{y})$. Of course for any one of the selectors $h(\mathbf{y})$ studied theoretically here, the relative size of this minorant, compared to $E\Delta(h(\mathbf{y})$, tends to 0 as $n \to \infty$, but we must keep in mind that it may take very large sample sizes before this risk regret is a small part of $E\Delta(h(\mathbf{y})$; indeed for the second order kernel setting of Section 2, this relative size decreases to zero at a rate only of $n^{-1}/n^{-4/5} = n^{-1/5}$ (and $n^{-1/(2k+1)}$ for kernel of order k, see Remark 8.2); in fact this becomes $n^{-1/10}$ if we come back (by taking square roots) to the correct scale for the width of the confidence band. Moreover it can be checked that, for any fixed n, one can obtain a setting for which the constant factor c, in the approximation $cn^{-1/10}$ of the relative size of this minorant, is arbitrarily large simply by taking a more concentrated weight function u. Thus a first use of risk-regret estimates like (7.5) may be as follows, using the fact that, for the trivial no-smoothing curve estimator (or "$h(\mathbf{y}) = 0$"), the value of the risk, $E\Delta(h(\mathbf{y})$, is known (and equal to $\sigma^2\mathrm{tr}U$) : indeed as soon as the data analyst observes that the estimated lower bound (7.5) for $E\Delta(\hat{h}_{\mathrm{CV}})$ is not a small part of $\sigma^2\mathrm{tr}U$, the CV-choice should be considered as likely not better than

the raw data in terms of Δ-performance. (this may be called a "first diagnostic for a failure of (local) CV"). A similar assessment can clearly also be constructed for the toward-more-parcimony selectors $\overline{\hat{h}_{\mathrm{R}}}(\alpha)$ (or $\overline{\hat{h}_{\mathrm{AR}}}(\alpha)$) proposed section 7.2. For example let us state this for $\overline{\hat{h}_{\mathrm{AR}}}(\alpha)$: a consistent estimate of the associated risk regret $\mathrm{E}\left(\Delta(\overline{\hat{h}_{\mathrm{AR}}}(\alpha) - \Delta(\hat{h}_0)\right) \sim (2nC_3)^{-1}[B^2 + V_2 + z_{1-\alpha}^2(B^2 + V_2 + (1/2)V_1]$ is

$$\widetilde{\mathrm{RR}}_{\mathrm{AUG-RAND}} - \frac{1}{2}\frac{V_1}{V_2}\widetilde{\mathrm{RR}}_{\mathrm{RAND}} + z_{1-\alpha}^2\widetilde{\mathrm{RR}}_{\mathrm{AUG-RAND}}.$$

2. By decomposing further $\Delta(\hat{h}_0) = \Delta(h_0) - \left(\Delta(h_0) - \Delta(\hat{h}_0)\right)$ and using the previous asymptotic laws, we see that, for example, again for the choice $h(\mathbf{y}) := \overline{\hat{h}_{\mathrm{AR}}}(\alpha)$,

$$E\left(\Delta(\overline{\hat{h}_{\mathrm{AR}}}(\alpha))\right) \approx M(h_0) - \frac{1}{2nC_3}(B^2 + V_1) + \frac{1}{2nC_3}(B^2 + V_2)(1 + (z_{1-\alpha}\overline{\kappa})^2),$$

the error in this approximation being $o(n^{-1})$ if we take for granted that the invoked moments at finite size n are asymptotic to the moments of the asymptotic laws. The sum of the second and third terms is easily seen, from the above results, to be consistently (thus up to $o_P(n^{-1})$) estimated by

$$(1 - \frac{V_1}{V_2})\widetilde{\mathrm{RR}}_{\mathrm{RAND}} + z_{1-\alpha}^2\widetilde{\mathrm{RR}}_{\mathrm{AUG-RAND}}. \tag{7.8}$$

An as much accurate estimate, say $\hat{M}_0$, of the minimum $M(h_0)$ is available, in theory, provided the well known asymptotic version $(nh)^{-1}C_1 + (1/4)C_2h^4$ of $M(h)$ is accurate enough and, as is well known too, some additional smoothness can be assumed so that $\sqrt{n}$-consistent estimates of C_2 and h_0 are available. Unfortunately such accurate $\hat{M}_0$ requires a good choice of pilot bandwidths which is not easy. This is beyond the scope of this article, and we do not include any experimental study.

3. In some contexts, the $M(h)$ function can be bounded, say by $B(h)$, from a priori bound on the "roughness" of m (like on $J_u(m)$ as in Speckman (1985)). There, another possible use of the above correction (7.8) of $M(h_0)$ may be to apply it to the minimum of $B(h)$ over the considered h (since $M(h_0) \leq M(\mathrm{argmin}_h B(h)) \leq \min_h B(h)$) to yield an accurate *conservative* confidence band centered at $\hat{m}_{\overline{\hat{h}_{\mathrm{AR}}(\alpha)}}$.

Further work is certainly useful to develop a methodology for constructing confidence bands or regions ready for practical use.

It is noteworthy that the produced confidence-bands for the toward-more-parcimony selectors $\overline{\hat{h}_{\mathrm{R}}}(\alpha)$ and $\overline{\hat{h}_{\mathrm{AR}}}(\alpha)$, have a width which increases with $z_{1-\alpha}^2$: this is very satisfactory and is in contrast with the standard procedure when one uses the mentioned standard had hoc modification of the criterion to translate $\hat{h}$ toward a little bit more smoothing and still retains the above formula (7.6) for the width of a 90% confidence-band (indeed $\mathrm{tr}UA_h$ is generally a decreasing function of h).

8. Simulations, discussion and further remarks

In practice, one may thus propose the following general methodology: First compute $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$ by repeatedly rerunning the fast randomized version of, for example, GCV. Then, if this estimated lower

bound (for the accuracy of $\hat{h}_{\mathrm{GCV}}$) is reasonably "small" (otherwise the use of GCV is very questionable), one next computes $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$ by repeating the whole previous procedure with, this time, the augmented randomized GCV using the cross-validation choice as pilot bandwidth g and one of the usual estimates of σ: one thus obtains a less optimistic estimate of the variance of $\hat{h}_{\mathrm{GCV}} - \hat{h}_0$ which is thus more reliable for providing a conservative confidence interval for $\hat{h}_0$. Note that this second stage has the same cost as the first one and requires no further analytic or programming effort (if computing exact GCV is costly, g may be taken as the mean of the randomized choices of the first stage, instead of the exact $\hat{h}_{\mathrm{GCV}}$).

8.1. Simulation study

For our experiments, we used a setting which has already been used in Rice (1984), HHM, and many others. So, the chosen mean function was

$$m(x) = x^3(1-x)^3,$$

and we used equispaced designs for simulating data sets. We consider a gaussian noise, i.e. $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$. A periodic version of the kernel estimate (1.2) was considered, as this is described e.g. in HHM. This is appropriate because m is "smoothly" periodic with period 1. We used $u \equiv 1$ as weight function, which is possible in this periodic setting. The computational cost of the simulation is then greatly reduced by using fast Fourier transforms (the randomized criteria are also easily generated in the Fourier space) as in the studies mentioned above and in Nychka (1991).

The kernel function was taken as the biweight

$$K(x) = (15/8)(1-4x^2)^2 1_{[-.5,.5]}(x).$$

In this simulation study, among all the selectors which are asymptotically equivalent to GCV, we only consider the Mallows' C_L selector for which we assume σ is known (since this criterion is then the only one which is exactly unbiased).

We report here the results that we obtained for the following two sample sizes: $n = 128$ and $n = 512$; and two noise levels: $\sigma = 0.0015$ and $\sigma = 0.011$ (the first one corresponds to a peak signal-to-noise equal to 10), these four cases being rather representative of "small" and "large" samples with "small" and "large" noise levels, respectively.

For each of the so-obtained four cases, 100 samples (i.e. 100 data sets) of size n are simulated. A reasonably good approximation of the targeted $\mathrm{var}^{\frac{1}{2}}(\hat{h}_{\mathrm{CL}} - \hat{h}_0)$ is first computed by its empirical version (denoted SD in the following) over these 100 samples. For each of these 100 samples, first, we generated 200 realizations of the randomized selector RCL, each one using a $\mathbf{w}$ drawn from $\mathcal{N}(0, I)$ to compute $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$; next, using the exact CL selector as pilot bandwidth, we generated 200 realizations of the augmented randomized version (ARCL), also using $\boldsymbol{\varepsilon}^*$'s from $\mathcal{N}(0, \sigma^2 I)$, to compute $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$.

	$\hat{h}_0$	$\hat{h}_{\rm CL}$	$\hat{h}_{\rm CL} - \hat{h}_0$	$\widehat{\rm SD}_{\rm RAND}$	$\widehat{\rm SD}_{\rm AUG-RAND}$	$\widehat{\rm SD}_{\rm BOOT}$	$.5796\,\hat{h}_{\rm CL}^{3/2}$
$n = 128$	.209	.200	$-.009$	.052	.060	.107	.053
	(.033)	(.038)	(SD = .054)	(.014)	(.014)	(.034)	(.014)
$n = 512$	.156	.151	$-.005$	.035	.040	.050	.034
	(.025)	(.026)	(SD = .040)	(.010)	(.010)	(.013)	(.008)

Table 8.1. $\sigma = .0015$. Summary (mean and, in brackets, standard deviation) over 100 samples

	$\hat{h}_0$	$\hat{h}_{\rm CL}$	$\hat{h}_{\rm CL} - \hat{h}_0$	$\widehat{\rm SD}_{\rm RAND}$	$\widehat{\rm SD}_{\rm AUG-RAND}$	$\widehat{\rm SD}_{\rm BOOT}$	$.5796\,\hat{h}_{\rm CL}^{3/2}$
$n = 128$	.480	.457	$-.023$	.168	.195	.255	.185
	(.093)	(.129)	(SD = .188)	(.044)	(.042)	(.044)	(.072)
$n = 512$	.346	.326	$-.020$	.109	.130	.155	.111
	(.072)	(.090)	(SD = .133)	(.027)	(.028)	(.032)	(.042)

Table 8.2. $\sigma = .011$. Summary (mean and, in brackets, standard deviation) over 100 samples

First, it should be observed that, as suggested from the theory, it is reasonable here to neglect the "bias" in quantifying the accuracy of the CL selector: more precisely, one see in the third column of these tables, that the mean of $\hat{h}_{\rm CL} - \hat{h}_0$ is rather "negligible", in absolute value, as compared to SD.

Second, one observes in Tables 8.1 and 8.2 that the "sandwich" $(\widehat{\rm SD}_{\rm RAND}, \widehat{\rm SD}_{\rm AUG-RAND})$, is quite satisfactory: the expected value of this interval approximately includes SD and it has reasonably small width in view of the variability of each of these two "bounds". The "noise" in these randomization estimates is of course not negligible : in each of the four cases, their standard deviation is roughly one fourth of SD. But in view of the rather large range of the different values for the scale SD, even with such a noise, these randomization-based estimates prove to be useful in these experiments.

In this simulation study, we also considered the bootstrap approach mentioned in the Introduction, using, essentially as in Nychka (1991), the exact CL choice for each data set as a simple pilot bandwidth for generating 200 pseudo-data sets (each one drawn from $\mathcal{N}(\hat{\mathbf{m}}_{\hat{h}_{\rm CL}}, \sigma^2 I)$). A striking fact, summarized in Tables 8.1 and 8.2, is that $\widehat{\rm SD}_{\rm BOOT}$ tends to over-estimate SD, even more than $\widehat{\rm SD}_{\rm AUG-RAND}$, especially for small samples. The noise in all these accuracy estimates is rather similar, except in the "small sample, small noise" case where $\widehat{\rm SD}_{\rm BOOT}$ is much more variable than $\widehat{\rm SD}_{\rm RAND}$ or $\widehat{\rm SD}_{\rm AUG-RAND}$.

As another benchmark, we also consider the "plug-in" variance estimate proposal of HHM, mentioned in the Introduction, which is based on the asymptotic analysis of this specific setting: from Theorem 2.1 and the expressions of C_0 and C_3 (given in Girard (1998) for the case $f \not\equiv 1$), the following expression for the targeted asymptotic variance is easily derived:

$$\mathrm{var}\left(\hat{h}_{\rm CL} - \hat{h}_0\right) \sim h_0^3 \frac{\int u^2}{\int K^2 \left(\int u\right)^2} \frac{4}{25} \left(\frac{J_{u^2}(m)}{J_u(m)} + 2\frac{\int L^2}{\int K^2}\right).$$

So, in the case $u^2 \propto u$, a natural and very simple approach is to plug-in $\hat{h}_{\rm CL}$ in place of h_0 in this expression, and to compute the required constant : this yields the estimate $(.5796)^2 \hat{h}_{\rm CL}^3$ for our choice of K which is displayed in the last column of Tables 8.1 and 8.2. It is remarkable that this plug-in estimate, although it uses a detailed analysis of this specific setting, has a noise which is either only comparable to or worse than that of the randomization-based estimates. Note that, in the general case $u^2 \not\propto u$, this plug-in proposal requires estimates of the ratio $J_{u^2}(m)/J_u(m)$ and an additional noise can then be apprehended.

Notice that in these experiments, we observed relatively small differences between $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$ and $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$, and so computing $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$ actually was not very useful. For the same reason, we have also not implemented here the more sophisticated estimate of Section 5. However, since this second stage has the same cost as the first one, it may well be worth implementing it in other applications.

8.2. Further remarks

Remark 8.1. *Extensions (and limitations) of the proposed randomization-based methodology.* Various variants and extensions are possible. For example

1. In this article, we focus on the estimation of $\mathrm{var}^{1/2}(\hat{h} - \hat{h}_0)$. Other measures for the "accuracy" of $\hat{h}$ could have been chosen, or a logarithmic scale (or other monotone transformations, like the one giving the popular equivalent degree of freedom, e.g. Hastie and Tibshirani (1990)) could have been taken as well. Of course, a standard deviation is a good summary only for an error distribution which is roughly normal. Note that this is rather typically the case for kernel bandwidths, as shown by the simulations in HHM, Section 4, and e.g. in Hall and Johnstone (1992, Section 5 and Remark 7.2).
2. When considering the generality of the GCV-type criteria and their randomized versions, and the underlying heuristics (Section 1.3), one may guess that this methodology can be applied to many other curve-, surface- or image-estimation techniques. However, one must keep in mind that the approximate stochastic representations of $\hat{h}-\hat{h}_0$ and $\hat{h}_{\mathrm{R}}-\hat{h}$ of Section 1.3 are based on local quadratic approximation of the considered criteria at their minima, and thus, at least second order differentiability of A_h as a function of h seems to be required as in e.g. Kneip (1995, Section 5.2). Note also that the theory here does not deal with the behavior near zero of the distribution of the GCV selector (since this selector has been appropriately restricted to the interval H_n); for a discussion of this disturbing behavior for small samples, see Wahba and Wang (1995). One may also guess that the augmented randomization technique should generally provide a certain correction of the downward bias of $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$; but in situations where such a correction would be mandatory (theory says this might be the case for kernels of very high order, see Remark 8.2 below), the upper-bound property of $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$ stems on the consistency properties of the GCV-selector. Note that, if the practitioner desires that $\widehat{\mathrm{SD}}_{\mathrm{AUG-RAND}}$ be reliably conservative, then a pilot parameter which would undersmooth, compared to the optimal $\hat{h}_0$, seems more appropriate than one which oversmoothes.
3. Nonparametric curve or surface estimators using a multidimensional smoothing parameter could be considered as well. One of the popular multidimensional examples is additive modeling with back-fitting. There, the randomization-based methodology could naturally produce accuracy estimates, simultaneously for each component of the *vector* $\hat{h}$ of GCV bandwidth estimates; an estimated covariance matrix (or even a confidence region for the vector $\hat{h}_0$ of the optimal bandwidths) might be constructed. This proposal would, of course, need to be investigated carefully.

Remark 8.2. *Extensions of the theory.* The theoretical results above can be extended to several contexts.

We point out that typically the underestimation factor $\underline{\kappa}$ still remains reasonably close to 1.

1. Consider for example the setting of Section 2, with only K replaced by a higher order kernel, that is
$$\int K = 1, \quad \int xK = \cdots = \int x^{k-1}K = 0, \quad \int x^k K > 0,$$
and assume that m and f have uniformly continuous kth derivatives. Then all the theoretical results of Sections 2, 4, 5 and 6, still hold with appropriate changes in the exponents of convergence ($n^{3/10}$ becomes $n^{3/2(2k+1)}$) and with modifications in the constants classically calculated. A very simple modification can be easily proved for the expression of $\underline{\kappa}$ in Corollary 2.3, which becomes:
$$\underline{\kappa} = \left(\frac{k}{4} \frac{J_{u^2}(m)}{J_u(m)} \frac{\int K^2}{\int L^2} + 1 \right)^{-1/2},$$
where again $L(x) = -xK'(x)$ but $J_u(m) := [\int \left((mf)^{(k)}\right)^2 f^{-1}u] / \int u$, for which Lemmas 3.1 and 3.2 are also valid, and thus Theorem 3.3 still holds with $\sqrt{3}$ replaced by $\sqrt{1+k}$, and the factor 2 simply replaced by k in the expression $\sqrt{1 + 2\frac{\int u}{\int u^2} \max u}$. This implies thus reasonable lower bounds for $\underline{\kappa}$ when the order k remains small.
2. A justification for this methodology is also possible in the context of smoothing spline estimate. For one-dimensional cubic smoothing spline, a convincing development in Nychka (1991, Section 3) indicates that the asymptotic formulae agree with those of a fourth-order kernel regression estimate. In the case of equidistant design, by exploiting that the smoothing spline framework is very well approximated by a known Gaussian white noise framework using Butterworth filter (whose range is parametered by h) in the frequency domain, to which correspond a particular kernel for which compactness condition (a) in Section 3 can be dropped, a rigorous development can be made as in Hall and Johnstone (1992, section 6.2).
3. Extensions to multidimensional settings could also be stated. See HHM and Girard (1998, Remark 3.2) for a two dimensional example using Gasser-Muller kernel estimate for which the result of this paper can be extended with classical modifications in the rate (the power 3/10 replaced everywhere by $(d+2)/(8+2d)$ with $d = 2$) and in the expressions for the constants; see Girard (1998, Remark 3.2) for the definition of the analog of $J_u(m)$.

Remark 8.3. It would be interesting to study the rate of convergence for these accuracy estimates: this deserves further study.

Remark 8.4. *Computational cost.* Since the theoretical results in this paper only describe the behavior of e.g. $\widehat{\mathrm{SD}}_{\mathrm{RAND}}$ which uses an infinite number of randomized choices, it is natural to wonder how many repeated minimizations one has to simulate. A complete theoretical analysis of this question would first require an answer to Remark 7.3. In our above experiments where $n = 128$ or $n = 512$, sufficiently stabilized estimates were already obtained by simulating 50 randomized choices : Simulating further 150 randomized choices, as was done here, did not actually produce great improvement.

References

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

CRAMER, H. (1970). Random variable and probability distribution. 3rd ed. Cambridge University Press, Cambridge, UK.

EFRON, B. (1982) The Bootstrap, the Jacknife and Other Resampling Plans. SIAM, Philadelphia.

EFRON, B. (1986) How biased is the apparent error rate of a prediction rule? J.A.S.A. **81** 461-470.

EUBANK, R. (1999). Nonparametric regression and spline smoothing. New York : Marcel Dekker.

EUBANK, R.L. and SUOJIN WANG (1994). Confidence regions in Non-parametric Regression. **21** 147-157.

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. (1995). On the fast Monte-Carlo cross-validation and C_L procedures: comments, new results and application to image recovery problems (with discussion). Computational Statistics **10** 205-258.

GIRARD, D. (1998). Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression. Ann. Statist. **26**, 315-334

GU, C. (2002). Smoothing spline ANOVA models. New-York : Springer.

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

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

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

HART, J.D. (1992). Comment on "Empirical functionals and efficient smoothing parameter selection" by HALL, P. and JOHNSTONE, I. . J. R. Statist. Soc. B. **54** No. 2. 518.

HASTIE, T. and TIBSHIRANI, R. (1990). Generalized additive models. Chapman and Hall, London

HUTCHINSON, M.F. (1990). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. Commun. Statist. -Simula. **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.A.S.A. **86** 1042-1060.

KOU S.C. (2003). On the efficiency of selection criteria in spline regression. Probab. Theory Relat. Fields **127**, 153-176.

LOADER, C. (1999). Bandwidth selection : Classical or plug-in ? Ann. Statist. **27** 415-438.

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.

NYCHKA, D.(1990). The average posterior variance of a smoothing spline and a consistent estimate of the average squared error. Ann. Statist. **18** 415-428.

NYCHKA, D.(1991). Choosing a range for the amount of smoothing in nonparametric regression. J.A.S.A. **86** 653-664.

PRIESTLEY, M.B. and CHAO, M.T. (1972). Non-parametric function fitting. J. R. Statist. Soc. B. **34** 385-392.

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

SPECKMAN, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. Ann. Statist. **13** 970-983.

SCHUCANY, W.R. (1995). Adaptive bandwidth choice for kernel regression. J.A.S.A. **90** 535-540.

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.

THOMPSON, A.M., KAY, J.W. and TITTERINGTON, D.M. (1989). A cautionary note about crossvalidatory choice. J. Statis. Comput. Simul. **33** 199-216.

VIEU, P. (1991). Nonparametric regression: local optimal bandwidth choice. J. R. Statist. Soc. B. **53** 453-464.

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.

WAHBA, G., AND WANG, Y. (1995). Behavior near zero of the distribution of GCV smoothing parameter estimates. Statist. and Probability Letters **25** 105-111.

Résumé

En régression non-paramétrique, il est généralement crucial d'utiliser des paramètres de lissage "à peu près optimaux". Il a été observé par de nombreux statisticiens que le choix validation croisée (CV), comme le choix GCV, se révèle dans certaines applications ne pas être "très précis" en tant qu'estimateur du choix optimal. Aussi le développement d'estimateurs fiables pour cette précision, serait très utile pour la pratique. Dans cette article nous montrons que la simulation du choix 'GCV randomisée', et d'une variante naturelle, peut fournir d'utiles inférences sur la précision du choix CV (ou GCV), comme un intervalle de confiance consistant pour le paramètre optimal ou un estimateur consistant de l'excès d'erreur (dans l'espace des courbes) due à l'imprécision du choix CV. Des preuves rigoureuses, ainsi qu'une vérification expérimentale, sont données pour les estimateurs de courbes du type noyau. Un ensemble d'heuristiques montre que la méthodologie générale donnée ici, pourrait être utile pour de nombreux autres techniques d'estimation d'une courbe -ou surface ou image- moyenne sous-jacente à des observations bruitées.

CNRS, Laboratoire de Modélisation et Calcul Tour IRMA, BP 53, 38041 Grenoble Cedex 9, France

email : didier.girard@imag.fr