The Annals of Statistics
1991, Vol. 19, No. 4, 1950–1963

ASYMPTOTIC OPTIMALITY OF THE FAST RANDOMIZED VERSIONS OF GCV AND C_L IN RIDGE REGRESSION AND REGULARIZATION

BY DIDIER A. GIRARD

CNRS and Université Joseph Fourier

Ridge regression is a well-known technique to estimate the coefficients of a linear model. The method of regularization is a similar approach commonly used to solve underdetermined linear equations with discrete noisy data. When applying such a technique, the choice of the smoothing (or regularization) parameter h is crucial. Generalized cross-validation (GCV) and Mallows' C_L are two popular methods for estimating a good value for h, from the data. Their asymptotic properties, such as consistency and asymptotic optimality, have been largely studied [Craven and Wahba (1979); Golub, Heath and Wahba (1979); Speckman (1985)]. Very interesting convergence results for the actual (random) parameter given by GCV and C_L have been shown by Li (1985, 1986). Recently, Girard (1987, 1989) has proposed fast randomized versions of GCV and C_L. The purpose of this paper is to show that the above convergence results also hold for these new methods.

1. Introduction. Suppose that we observe an n-dimensional vector $\mathbf{y}_n$ of data satisfying the regression model

$$\mathbf{y}_n = \mathbf{f}_n + \mathbf{e}_n, \quad \text{where} \quad \mathbf{f}_n = X_n \mathbf{g}_n, \qquad \mathbf{e}_n \sim \mathcal{N}(0, \sigma^2 I),$$

where X_n is a known $n \times p_n$ (design) matrix, $\mathbf{f}_n$ and $\mathbf{g}_n$ are unknown deterministic vectors and $\mathbf{e}_n \sim \mathcal{N}(0, \sigma^2 I)$ means that $\mathbf{e}_n$ is a random vector of independent normal variables with mean 0 and common variance σ^2 [note that, in this paper, we take essentially the same notations as in Li (1986) since most of his results will be used in the following]. In addition we suppose that we observe (or generate) a noise vector $\mathbf{w}_n$, independent or not of $\mathbf{e}_n$, with the same probability law as $\mathbf{e}_n$, except for the factor σ:

$$\mathbf{w}_n \sim \mathcal{N}(0, I).$$

In the ridge regression approach (or standard regularized least squares approach) [e.g., Golub, Heath and Wahba (1979), Li (1985, 1986)], the estimates of $\mathbf{g}_n$ and $\mathbf{f}_n$ have the following form:

$$(1.1) \qquad \hat{\mathbf{g}}_n(h) = \left(X_n^t X_n + hI\right)^{-1} X_n^t \mathbf{y}_n, \qquad \hat{\mathbf{f}}_n(h) = M_n(h)\mathbf{y}_n,$$

Received October 1989; revised August 1990.

AMS 1980 *subject classifications*. Primary 62G05, 65U05; secondary 65D10, 65R20, 92A07.

Key words and phrases. GCV, C_L, ridge regression, regularization, smoothing splines, Monte Carlo techniques, randomized versions, asymptotic optimality.

where h is the smoothing (or ridge) parameter to be chosen and

$$M_n(h) = X_n\left(X_n^t X_n + hI\right)^{-1} X_n^t \tag{1.2}$$

is the smoothing (or influence) matrix corresponding to h. It is easy to show that any spline smoothing problem can take the form of a ridge regression problem after a suitable orthogonal transformation [e.g., Demmler and Reinsch (1975), Speckman (1985), Li (1985)].

Most of the procedures for choosing h consist of minimizing with respect to h a certain function of the residual $n^{-1}\|\mathbf{y}_n - \hat{\mathbf{f}}_n(h)\|^2$ and of the trace $\operatorname{tr} M_n(h)$. An exception is the residual method which is known to give typical oversmoothing [Craven and Wahba (1979), Hall and Titterington (1987)]. As in Li (1986), we consider two of the most popular selection methods. Mallows' C_L [Mallows (1973)] consists of choosing h by minimizing

$$\mathrm{CL}_n(h) = n^{-1}\left\|\left(I - M_n(h)\right)\mathbf{y}_n\right\|^2 + 2\sigma^2 n^{-1} \operatorname{tr} M_n(h). \tag{1.3}$$

If σ^2 is not known, a popular method is generalized cross-validation or GCV [Craven and Wahba (1979)] which selects h by minimizing

$$\mathrm{GCV}_n(h) = \frac{n^{-1}\left\|\left(I - M_n(h)\right)\mathbf{y}_n\right\|^2}{\left[n^{-1}\operatorname{tr}\left(I - M_n(h)\right)\right]^2}. \tag{1.4}$$

In most of the typical applications, any smoothing matrix of the family $\{M_n(h)\colon h \geq 0\}$ is neither explicitly known nor can it be easily computed and stored: Instead, any required estimate $M_n(h)\mathbf{y}_n$ is generally computed by solving a linear system of the form (1.1) by a direct or an iterative method. In practice, there exist many important applications where computation for given h of such an estimate, and of the associated residual, costs much less than order n^3 (for instance, one-dimensional spline smoothing problem, multidimensional iterative smoothing techniques, or, more generally, least-squares problems with sparse, banded or well-structured matrices). In some special cases, efficient algorithms have been established for the computation of the GCV function; see Elden (1984) for the banded matrix case, Hutchinson and de Hoog (1985) for the one-dimensional spline smoothing problem [earlier, Utreras (1980) has provided an approximate algorithm for the equally spaced data case], Girard (1988) for the one-dimensional partial spline (e.g., discontinuities-preserving smoothing) problem, Girard (1987a) for well-structured tomographical reconstruction problems. Unfortunately, for many other applications the evaluation of the denominator of the GCV function (the trace-term) generally costs much more than the computation of the residual term, so the use of exact GCV requires an enormous computational task for large data set.

The Monte Carlo cross-validation procedure proposed by Girard (1987b, 1989) is a general approximate version of GCV which eliminates this drawback. It consists of generating a few independent pseudorandom vectors $\mathbf{w}_n^1, \ldots, \mathbf{w}_n^m$ each following $\mathcal{N}(0, I)$, and replacing $n^{-1}\operatorname{tr} M_n(h)$ in (1.4) by the average of the m estimates $\langle \mathbf{w}_n^k, M_n(h)\mathbf{w}_n^k\rangle / \langle \mathbf{w}_n^k, \mathbf{w}_n^k\rangle$, $k = 1, \ldots, m$. In

Girard (1987b, 1989), it is shown, both theoretically and practically, that using a few of such estimates (say $m = 10$ or even $m = 1$ if n is large enough) is sufficient to obtain an approximation of the GCV function with very good *relative* accuracy, in typical spline smoothing problems; see Section 3.1.1 and also Girard and Laurent (1989) for some other applications. Here we consider this method without averaging (i.e., $m = 1$) not as an approximation but in its own right, and we call it randomized GCV (or RGCV) since the name Monte Carlo generally refers to iterated processes intended to approximate deterministic quantities. So we define RC_L as the procedure which selects h by minimizing (see Remark 2.1 for a variant)

$$\mathrm{RCL}_n(h) = n^{-1}\|(I - M_n(h))\mathbf{y}_n\|^2 + 2\sigma^2 n^{-1}\langle \mathbf{w}_n, M_n(h)\mathbf{w}_n\rangle.$$

And we define RGCV as the procedure which selects h by minimizing

$$\mathrm{RGCV}_n(h) = \frac{n^{-1}\|(I - M_n(h))\mathbf{y}_n\|^2}{\left[n^{-1}\langle \mathbf{w}_n, (I - M_n(h))\mathbf{w}_n\rangle\right]^2}.$$

The main goal of this paper is to show that RGCV and RC_L, the randomized version of GCV and C_L, will always possess the same asymptotic validity as the classical exact GCV and C_L, respectively [at least for the convergence properties established in Li (1986)].

As in Li (1986), let $\hat{h}_M$ and $\hat{h}_G$ denote the h selected by C_L and GCV, respectively. Let also $\hat{h}_{RM}$ and $\hat{h}_{RG}$ denote the h selected by their randomized versions, RC_L and RGCV, respectively. Note that all these parameters are random variables, as functions of $\mathbf{y}_n$; but the latters are also function of $\mathbf{w}_n$. Define $L_n(h)$ as the true loss while estimating $\mathbf{f}_n$ by $\hat{\mathbf{f}}_n(h)$:

$$L_n(h) = n^{-1}\|\mathbf{f}_n - \hat{\mathbf{f}}_n(h)\|^2.$$

Li (1986) has shown that under weak assumptions, C_L and GCV are asymptotically optimal (a.o.) in the sense that as $n \to \infty$, the inefficiency tends to 1, that is,

$$\frac{L_n(\hat{h})}{\inf_{h \geq 0} L_n(h)} \to 1, \quad \text{in probability,}$$

for $\hat{h} = \hat{h}_M, \hat{h}_G$. These results are much stronger than the previous ones of Craven and Wahba where only the deterministic minimizer of $E\,\mathrm{GCV}_n$ (expected function which actually is not observable) was shown to possess an expectation inefficiency (i.e., with EL_n in place L_n) that tends to 1. Note that this expectation a.o. can be immediately extended to any criteria whose expectation can be written as the sum of EL_n (or $E\,\mathrm{GCV}_n$) and a term independent of h (e.g., this is the case for RCL_n).

In this paper we shall show the a.o. of RC_L (Section 2) and RGCV (Section 3) in the same sense and under the same assumptions as in Li (1986).

For the sake of completeness, let us recall these required assumptions. The only condition needed for C_L or RC_L to be a.o. is

$$\text{(A.1)} \qquad \inf_{h \geq 0} nEL_n(h) \to \infty,$$

that is, the rate of convergence of the minimal expected error $\inf_{h \geq 0} EL_n(h)$ is slower than $1/n$. Li (1986) points out that (A.1) implies more or less that $p_n \to \infty$ and that, without (A.1), it seems that no selection procedure can be a.o. For many problems, this condition is equivalent to the condition that $\mathbf{f}_n$ is not infinitely "smooth". For example, in typical polynomial spline smoothing problems, $\inf_{h \geq 0} EL_n(h)$ tends to 0 at the rate $n^{-1+\delta}$ for some small constant $\delta > 0$ except if $\mathbf{f}_n$ happens to be the discretization of a low order polynomial [see, e.g., Wahba (1985)].

In addition to (A.1), a second condition is required for the a.o. of GCV or RGCV. This condition says, roughly speaking, that $X_n X_n^t$ must be ill-conditioned for large n. Specifically, the eigenvalues $\lambda_{1,n} \geq \lambda_{2,n} \geq \cdots \geq \lambda_{n_p,n} \geq 0$ of $X_n^t X_n$ must satisfy [as in Li (1986)] the following condition:

For any m such that $m/n \to 0$, we have

$$\text{(A.2)} \qquad \left(n^{-1} \sum_{i=m+1}^{n} \lambda_{i,n} \right)^2 \Bigg/ \left(n^{-1} \sum_{i=m+1}^{n} \lambda_{i,n}^2 \right) \to 0,$$

where we write $\lambda_{i,n} = 0$ for $i = p_n + 1, \ldots, n$. Note that (A.2) is satisfied in typical spline smoothing problems [see Li (1986)].

Finally a third condition [which is a natural condition to be able to prove consistency of any $\hat{\mathbf{f}}_n(\hat{h}_n)$] is also required: We must assume that there exists a deterministic sequence h_n such that $\hat{\mathbf{f}}_n(h_n)$ is consistent, in the sense

$$\text{(A.3)} \qquad EL_n(h_n) \to 0.$$

2. Optimality of randomized Mallows' C_L. For the sake of completeness, we will also recall the main results of Li (1986).

THEOREM 2.1 [Li (1986)]. *Under* (A.1), C_L *is a.o.*

The main steps of the proof of Li can be outlined as follows. As it is classical, it is enough to prove that $\mathrm{CL}_n(h)$, possibly corrected by a term c_n independent of h, approximates the true loss $L_n(h)$ with a relative accuracy which tends to zero uniformly in h, that is,

$$\text{(2.1)} \quad \mathrm{CL}_n(h) - c_n = L_n(h)(1 + \varepsilon_n(h)), \quad \text{where} \quad \sup_{h \geq 0} |\varepsilon_n(h)| = o_P(1),$$

where $o_P(1)$ denotes a random quantity which tends to 0 in probability as $n \to \infty$. (Throughout this paper, all the convergences of random quantities will

be in probability.) Now, since we can write

$$\begin{aligned} &\mathrm{CL}_n(h) - n^{-1}\|\mathbf{e}_n\|^2 - L_n(h) \\ &\quad = 2n^{-1}\langle \mathbf{e}_n, (I - M_n(h))\mathbf{f}_n\rangle \\ &\qquad + 2n^{-1}\big(\sigma^2 \operatorname{tr} M_n(h) - \langle \mathbf{e}_n, M_n(h)\mathbf{e}_n\rangle\big), \end{aligned} \tag{2.2}$$

the above uniform relative closeness for large n of $\mathrm{CL}_n(h) - n^{-1}\|\mathbf{e}_n\|^2$ and $L_n(h)$ is easily deduced from the following results that we recall from Li (1986).

PROPOSITION 2.1 [Li (1986)]. *Under* (A.1), *we have the following convergences in probability as* $n \to \infty$:

$$\sup_{h \geq 0} \frac{|n^{-1}\langle \mathbf{e}_n, (I - M_n(h))\mathbf{f}_n\rangle|}{EL_n(h)} \to 0, \tag{2.3}$$

$$\sup_{h \geq 0} \frac{|\sigma^2 n^{-1} \operatorname{tr} M_n(h) - n^{-1}\langle \mathbf{e}_n, M_n(h)\mathbf{e}_n\rangle|}{EL_n(h)} \to 0, \tag{2.4}$$

$$\sup_{h \geq 0} \left| \frac{L_n(h)}{EL_n(h)} - 1 \right| \to 0. \tag{2.5}$$

Let us turn now to RC_L. We can write

$$\mathrm{RCL}_n(h) = \mathrm{CL}_n(h) + 2\sigma^2\big(n^{-1}\langle \mathbf{w}_n, M_n(h)\mathbf{w}_n\rangle - n^{-1} \operatorname{tr} M_n(h)\big).$$

So, to show the uniform relative closeness of $\mathrm{RCL}_n(h) - n^{-1}\|\mathbf{e}_n\|^2$ and $L_n(h)$ under (A.1), it is enough to show that $\sigma^2 n^{-1}\langle \mathbf{w}_n, M_n(h)\mathbf{w}_n\rangle - \sigma^2 n^{-1} \operatorname{tr} M_n(h)$ is also negligible, compared to $EL_n(h)$, uniformly over $h \geq 0$. But this is simply the second convergence result of Proposition 2.1 applied to the problem of estimating $\mathbf{f}_n$ from $\mathbf{y}_n^* = \mathbf{f}_n + \sigma\mathbf{w}_n$ in place of $\mathbf{y}_n = \mathbf{f}_n + \mathbf{e}_n$ (in other words, Proposition 2.1 is also true with $\sigma\mathbf{w}_n$ in place of $\mathbf{e}_n$). Thus we also have:

THEOREM 2.2. *Under* (A.1), RC_L *is a.o.*

REMARK 2.1. In Girard (1989), it is shown that, if we look for a better finite sample estimate of $\mathrm{CL}_n(h)$, then it is better to use the normalized estimate $\langle \mathbf{w}_n, M_n(h)\mathbf{w}_n\rangle / \langle \mathbf{w}_n, \mathbf{w}_n\rangle$ in place of $n^{-1}\langle \mathbf{w}_n, M_n(h)\mathbf{w}_n\rangle$ in RCL_n. Theorem 2.2 also holds for this normalized RC_L. This can be easily proved by splitting $\langle \mathbf{w}_n, M_n(h)\mathbf{w}_n\rangle / \langle \mathbf{w}_n, \mathbf{w}_n\rangle - n^{-1} \operatorname{tr} M_n(h)$ into two components as in the proof of Proposition 3.2.

3. Optimality of randomized GCV. The extension of the results of Li for GCV will not be so easy as for C_L [although it will be more easy if we make a further assumption, see (A.4) defined below]. We first recall the main results of Li (1986).

THEOREM 3.1 [Li (1986)]. *Under* (A.1), (A.2) *and* (A.3), *GCV is a.o.*

For the proof of this theorem, Li considers the following simplified version of the Stein estimate of $\mathbf{f}_n$, defined by

$$\tilde{\mathbf{f}}_n(h) = \mathbf{y}_n - \sigma^2 \frac{\operatorname{tr}(I - M_n(h))}{\|(I - M_n(h))\mathbf{y}_n\|^2}(I - M_n(h))\mathbf{y}_n \tag{3.1}$$

[see Li and Hwang (1984), Li (1985), for discussions on some advantages of $\tilde{\mathbf{f}}_n(h)$ over $\hat{\mathbf{f}}_n(h)$ related to certain robustness concepts]. And Li makes use of an interesting connection between GCV and the associated Stein's unbiased risk estimate (SURE)

$$\mathrm{SURE}_n(h) = \sigma^2 - \sigma^4 \frac{\left[n^{-1}\operatorname{tr}(I - M_n(h))\right]^2}{n^{-1}\|(I - M_n(h))\mathbf{y}_n\|^2}, \tag{3.2}$$

initially proposed as an estimate of the risk $E\tilde{L}_n$, where

$$\tilde{L}_n(h) = n^{-1}\|\mathbf{f}_n - \tilde{\mathbf{f}}_n(h)\|^2$$

is the true loss while estimating $\mathbf{f}_n$ by $\tilde{\mathbf{f}}_n(h)$. Indeed, it is clear, by comparing (3.2) and (1.4), that $\hat{h}_G$ also minimizes $\mathrm{SURE}_n(h)$ over $h \geq 0$.

A first key result used in the proof of Li (1986) is the following:

THEOREM 3.2 [consistency of GCV, Li (1985)]. *Under* (A.3) *and the condition on the eigenvalues* (A.2), $\hat{\mathbf{f}}_n(\hat{h}_G)$, *where* $\hat{h}_G$ *is the GCV choice, is a consistent estimate of* $\mathbf{f}_n$, *that is,*

$$L_n(\hat{h}_G) \to 0.$$

REMARK 3.1. To show this, Li first established the interesting fact that, even without the assumptions (A.1), (A.2) or (A.3), $\mathrm{SURE}_n(h)$ is always a consistent estimate of $\tilde{L}_n(h)$ uniformly over both $\mathbf{f}_n \in \mathbb{R}^n$ and $h \geq 0$ [Li (1985), Lemma 4.2]. Next, since (A.3) implies $\tilde{L}_n(h_n) \to 0$ [Li and Hwang (1984)], one can conclude [as in the proof of Theorem 4.1 of Li (1985)] that, with only (A.3), $\tilde{\mathbf{f}}_n(\hat{h}_G)$ selected by GCV is always consistent, that is, $\tilde{L}_n(\hat{h}_G) \to 0$. Note that, from the consistency of $\mathrm{SURE}_n(h)$, this also implies that with only (A.3) we have $\mathrm{SURE}_n(\hat{h}_G) \to 0$ or equivalently

$$\mathrm{GCV}_n(\hat{h}_G) \to \sigma^2. \tag{3.3}$$

REMARK 3.2. These results have actually been proved under a weaker assumption [condition (5.6) of Li (1985)] on the eigenvalues of $X_n^t X_n$. The fact that (5.6) of Li (1985) is weaker than (A.2), has been used by Li (1986) without giving its proof. However, a rigorous proof can be obtained from this author [Li (1990)]. The interested reader could check that this weaker assumption is also sufficient for the consistency of RGCV.

A second key result is Proposition 3.1 of Li (1986), which states that for some appropriate sequences $\hat{h}$ [in particular, under (A.1) and (A.2), the consistency of $\hat{h}$ can be seen to be sufficient, cf. the proof of Theorem 3.4] we have a certain relative closeness of $\mathrm{SURE}_n(\hat{h})$ and $L_n(\hat{h})$.

PROPOSITION 3.1 [Li (1986)]. *Under* (A.1), *for any* $\hat{h}$ *random or not, such that*

$$n^{-1}\left\|\left(I - M_n(\hat{h})\right)\mathbf{y}_n\right\|^2 \to \sigma^2$$

and

$$\frac{\left(n^{-1}\operatorname{tr} M_n(\hat{h})\right)^2}{n^{-1}\operatorname{tr} M_n^2(\hat{h})} \to 0,$$

we have

$$\frac{\left|SURE_n(\hat{h}) - \tilde{L}_n(\hat{h}) - n^{-1}\|\mathbf{e}_n\|^2 + \sigma^2\right|}{L_n(\hat{h})} \to 0$$

and

$$\tilde{L}_n(\hat{h})/L_n(\hat{h}) \to 1.$$

We will see [cf. the proof of the a.o. of RGCV in Theorem 3.4] that these are the two results of Li that we have to extend to our randomized GCV and to the corresponding randomized version of Stein's unbiased risk estimate, RSURE_n, that we define by

$$\mathrm{RSURE}_n(h) = \sigma^2 - \sigma^4\frac{\left(n^{-1}\left\langle \mathbf{w}_n, \left(I - M_n(h)\right)\mathbf{w}_n\right\rangle\right)^2}{n^{-1}\left\|\left(I - M_n(h)\right)\mathbf{y}_n\right\|^2}$$

or its normalized version NRSURE_n defined by

$$\mathrm{NRSURE}_n(h) = \sigma^2 - \sigma^4\left(\frac{\left\langle \mathbf{w}_n, \left(I - M_n(h)\right)\mathbf{w}_n\right\rangle}{\|\mathbf{w}_n\|^2}\right)^2 \Bigg/ n^{-1}\left\|\left(I - M_n(h)\right)\mathbf{y}_n\right\|^2.$$

Note that these two criteria have the same minimizer $\hat{h}_{RG}$ as RGCV_n.

Before proceeding with our proof, we state some other useful results that are direct extensions of Lemma 5.1 and Lemma 5.2 of Li (1985).

LEMMA 3.1. *For any sequence* $\hat{h}$, *random or not, such that* $GCV_n(\hat{h}) \to \sigma^2$, $\hat{\mathbf{f}}_n(\hat{h})$ *is consistent if and only if* $n^{-1}\operatorname{tr} M_n(\hat{h}) \to 0$.

PROOF. First, the consistency of $\hat{\mathbf{f}}_n(\hat{h})$ implies $n^{-1}\|(I - M_n(\hat{h}))\mathbf{y}_n\|^2 \to \sigma^2$ [because $L_n(\hat{h}) = n^{-1}\|(I - M_n(\hat{h}))\mathbf{y}_n - \mathbf{e}_n\|^2$ and $n^{-1}\|\mathbf{e}_n\|^2 \to \sigma^2$] and thus $[n^{-1}\operatorname{tr}(I - M_n(\hat{h}))]^2 = n^{-1}\|(I - M_n(\hat{h}))\mathbf{y}_n\|^2/\mathrm{GCV}_n(\hat{h}) \to 1$. Conversely, since $\mathrm{GCV}_n(\hat{h}) \to \sigma^2$ implies the consistency of $\tilde{\mathbf{f}}_n(\hat{h})$ [by the uniform closeness of

$\mathrm{SURE}_n(h)$ and $\tilde{L}_n(h)$, cf. Remark 3.1], then, by writing from (3.1),

$$n^{-1}\left\|\tilde{\mathbf{f}}_n(h) - \hat{\mathbf{f}}_n(h)\right\|^2 = \left|1 - \frac{\sigma^2\,\mathrm{GCV}_n^{-1}(h)}{n^{-1}\operatorname{tr}(I - M_n(h))}\right|^2 \mathrm{GCV}_n(h)\left[n^{-1}\operatorname{tr}(I - M_n(h))\right]^2,$$

we see that $n^{-1}\operatorname{tr} M_n(\hat{h}) \to 0$ implies the consistency of $\hat{\mathbf{f}}_n(\hat{h})$. □

LEMMA 3.2. *For any sequence $\hat{h}$, random or not, such that $GCV_n(\hat{h}) \to \sigma^2$, we have*

$$\frac{n^{-1}\operatorname{tr}\left(I - M_n(\hat{h})\right)^2}{\left[n^{-1}\operatorname{tr}\left(I - M_n(\hat{h})\right)\right]^2} \to 1.$$

Under (A.2), *from this it follows that $n^{-1}\operatorname{tr} M_n(\hat{h}) \to 0$.*

PROOF. Li has stated the first part of this lemma [Li (1985), Lemma 5.2] in the particular case where $\hat{h} = \hat{h}_G$. But it suffices to note that the only condition on $\hat{h}_G$ used in his proof [Li (1985), pages 1374–1376] is that $\mathrm{SURE}_n(\hat{h}_G) \to 0$, or equivalently, $\mathrm{GCV}_n(\hat{h}_G) \to \sigma^2$. The second statement of this lemma is proved in Li [(1985), page 1365], under a weaker assumption on the eigenvalues than (A.2) (see Remark 3.2). □

3.1. *Consistency of* RGCV. In this section we will show, under the only conditions (A.2) and (A.3), the consistency of the ridge-regression estimate $\hat{\mathbf{f}}_n(\hat{h}_{RG})$ selected by RGCV. From Lemmas 3.1 and 3.2, we see that it suffices to prove that $\mathrm{GCV}_n(\hat{h}_{RG}) \to \sigma^2$. This will be proved here by recalling that $\mathrm{GCV}_n(\hat{h}_G) \to \sigma^2$ [cf. (3.3) in Remark 3.1], and by showing a certain uniform closeness between $\mathrm{GCV}_n(h)$ and $\mathrm{RGCV}_n(h)$ over $h \geq 0$.

We will first consider (Section 3.1.1) an additional assumption with which the proof of the consistency of RGCV is easy. Next we will see (Section 3.1.2) that this assumption is unnecessary.

3.1.1. *With an additional assumption*:

Let us define $c_{0,n}$ as in Girard [(1989), Theorem 2.4, in the particular case $D = I$, $\Omega = I$], with $r_n = \operatorname{rank}(X_n)$,

$$c_{0,n} = \begin{cases} \left(n/(n - r_n)\right)^{1/2}, & \text{if } r_n < n, \\ \left(n^{-1}\sum_{i=1}^{n} \lambda_{i,n}^{-2}\right)^{1/2} \Big/ \left(n^{-1}\sum_{i=1}^{n} \lambda_{i,n}^{-1}\right), & \text{if } r_n = n. \end{cases}$$

It was shown in Girard [(1989), Corollary 2.5] that the *relative* precision (i.e., relative standard deviation) of $\langle \mathbf{w}_n, (I - M_n(h))\mathbf{w}_n \rangle$, as an approximation of $\operatorname{tr}(I - M_n(h))$, is a nonincreasing function of h, uniformly bounded by $\sqrt{2}\, n^{-1/2} c_{0,n}$.

It is thus natural to consider the following condition on the asymptotic behavior of the eigenvalues of $X_n^t X_n$:

(A.4) $$n^{-1/2} c_{0,n} \to 0.$$

Then, assumption (A.4) guarantees that the uniform relative closeness of $\langle \mathbf{w}_n, (I - M_n(h))\mathbf{w}_n \rangle$ and $\operatorname{tr}(I - M_n(h))$ [and equivalently, of $\mathrm{RGCV}_n^{-1/2}(h)$ and $\mathrm{GCV}_n^{-1/2}(h)$] will hold and thus $\mathrm{GCV}_n^{-1/2}(\hat{h}_{RG})/\sup_{h \geq 0} \mathrm{GCV}_n^{-1/2}(h) \to 1$. Note that, in typical spline smoothing problems and for some integral equations, $c_{0,n}$ can be easily shown to tend to a constant [cf. Girard (1987b, 1989)], so (A.4) is satisfied.

3.1.2. *Without* (A.4):

PROPOSITION 3.2. *We always have the following convergence*:

$$\sup_{h \geq 0} \left| GCV_n^{-1/2}(h) - RGCV_n^{-1/2}(h) \right| \to 0.$$

PROOF. We have to show that for any $\delta_1, \delta_2 > 0$, there exists an integer N such that for $n > N$,

$$P\left\{ \sup_{h \geq 0} \frac{\left| n^{-1} \operatorname{tr}(I - M_n(h)) - n^{-1} \langle \mathbf{w}_n, (I - M_n(h))\mathbf{w}_n \rangle \right|}{n^{-1/2} \| (I - M_n(h))\mathbf{y}_n \|} \geq \delta_1 \right\} \leq \delta_2. \tag{3.4}$$

A sufficient condition for (3.4) is that there exists $\alpha_n > 0$ such that

$$P\left\{ \inf_{h \geq 0} \frac{\| (I - M_n(h))\mathbf{y}_n \|}{Q_n^{1/2}(h)} \leq \alpha_n \right\} \leq \frac{\delta_2}{2}, \tag{3.5}$$

$$P\left\{ \sup_{h \geq 0} \frac{\left| n^{-1} \operatorname{tr}(I - M_n(h)) - n^{-1} \langle \mathbf{w}_n, (I - M_n(h))\mathbf{w}_n \rangle \right|}{n^{-1/2} Q_n^{1/2}(h)} \geq \delta_1 \alpha_n \right\} \leq \frac{\delta_2}{2}, \tag{3.6}$$

where $Q_n(h) = E(\| (I - M_n(h))\, \mathbf{y}_n \|^2)$. Now, Li has shown [cf. proof of (7.2.2), page 1370, in the proof of Lemma 4.2 of Li (1985)] that, with L denoting any number greater than 0, we have

$$P\left\{ \inf_{h \geq 0} \frac{\| (I - M_n(h))\mathbf{y}_n \|}{Q_n^{1/2}(h)} \leq \alpha_n \right\} \leq K \alpha_n L^{1/2} + C\left(1 - 2\alpha_n^2\right)^{-2} L^{-1},$$

where K and C are constants independent of n and of L. Thus, setting $L = 8C\delta_2^{-1}$, (3.5) holds for large n as soon as $\alpha_n \to 0$. Turning to (3.6), one

can check that in the proof of (7.2.4) of Li [(1985), page 1371], it has been also shown that (3.6) holds for large n provided $\alpha_n^4 n \to \infty$. Thus, there exists α_n such that (3.5) and (3.6) hold for large n. □

Now, we can write, with $\delta_n(h) = \mathrm{GCV}_n^{-1/2}(h) - \mathrm{RGCV}_n^{-1/2}(h)$:

$$\begin{aligned} 0 &\le \mathrm{GCV}_n^{-1/2}(\hat{h}_G) - \mathrm{GCV}_n^{-1/2}(\hat{h}_{RG}) \\ &= \mathrm{GCV}_n^{-1/2}(\hat{h}_G) - \mathrm{RGCV}_n^{-1/2}(\hat{h}_{RG}) - \delta_n(\hat{h}_{RG}) \\ &\le \mathrm{GCV}_n^{-1/2}(\hat{h}_G) - \mathrm{RGCV}_n^{-1/2}(\hat{h}_G) - \delta_n(\hat{h}_{RG}) \\ &= \delta_n(\hat{h}_G) - \delta_n(\hat{h}_{RG}) \to 0 \end{aligned}$$

by Proposition 3.2, and thus we conclude that, as soon as $\mathrm{GCV}_n(\hat{h}_G)$ tends to a constant greater than 0, $\mathrm{GCV}_n(\hat{h}_{RG})$ tends to the same constant.

Thus we have proved:

THEOREM 3.3 (Consistency of RGCV). *Under* (A.3) *and the condition on the eigenvalues* (A.2), $\hat{\mathbf{f}}_n(\hat{h}_{RG})$, *where* $\hat{h}_{RG}$ *is the RGCV choice, is a consistent estimate of* $\mathbf{f}_n$, *that is,*

$$L_n(\hat{h}_{RG}) \to 0.$$

REMARK 3.3. Note that, since (A.3) is sufficient for $\mathrm{GCV}_n(\hat{h}_G) \to \sigma^2$ [cf. Remark 3.1], we have also proved that with only (A.3) we have $\mathrm{GCV}_n(\hat{h}_{RG}) \to \sigma^2$ or equivalently, the consistency of $\tilde{\mathbf{f}}_n(\hat{h}_{RG})$.

3.2. *Optimality of* RGCV. Since the a.o. of RGCV will be shown using similar lines of proof as in Li [(1986), Theorem 2], we have to extend Proposition 3.1 to NRSURE_n. For this it suffices to establish:

PROPOSITION 3.3. *Under* (A.1), *for any* $\hat{h}$ *random or not, such that*

$$n^{-1}\left\|\left(I - M_n(\hat{h})\right)\mathbf{y}_n\right\|^2 \to \sigma^2,$$

we have

$$\frac{NRSURE_n(\hat{h}) - SURE_n(\hat{h})}{L_n(\hat{h})} \to 0.$$

PROOF. By (2.5) of Proposition 2.1, it is enough to show that

$$\frac{\left|\left(\|\mathbf{w}_n\|^{-2}\langle \mathbf{w}_n, \left(I - M_n(\hat{h})\right)\mathbf{w}_n\rangle\right)^2 - \left(n^{-1}\operatorname{tr}\left(I - M_n(\hat{h})\right)\right)^2\right|}{n^{-1}\left\|\left(I - M_n(\hat{h})\right)\mathbf{y}_n\right\|^2 EL_n(\hat{h})} \to 0.$$

Since the eigenvalues of $M_n(\hat{h})$ are all between 0 and 1, we can write

$$\left|\left(\frac{\langle \mathbf{w}_n, (I - M_n(\hat{h}))\mathbf{w}_n\rangle}{\|\mathbf{w}_n\|^2}\right)^2 - \left(n^{-1}\operatorname{tr}(I - M_n(\hat{h}))\right)^2\right|$$

$$\leq 2\left|\frac{\langle \mathbf{w}_n, M_n(\hat{h})\mathbf{w}_n\rangle}{\|\mathbf{w}_n\|^2} - n^{-1}\operatorname{tr} M_n(\hat{h})\right|$$

$$\leq 2\frac{n}{\|\mathbf{w}_n\|^2}\left(\left|n^{-1}\langle \mathbf{w}_n, M_n(\hat{h})\mathbf{w}_n\rangle - n^{-1}\operatorname{tr} M_n(\hat{h})\right| + \left|\left(1 - \frac{\|\mathbf{w}_n\|^2}{n}\right)n^{-1}\operatorname{tr} M_n(\hat{h})\right|\right).$$

Thus, since $n\|\mathbf{w}_n\|^{-2} \to 1$ and $n^{-1}\|(I - M_n(\hat{h}))\mathbf{y}_n\|^2 \to \sigma^2$, it is enough to show that

$$\frac{\left|\left(1 - \left(\|\mathbf{w}_n\|^2/n\right)\right)n^{-1}\operatorname{tr} M_n(\hat{h})\right|}{EL_n(\hat{h})} \to 0$$

and

$$\frac{\left|n^{-1}\langle \mathbf{w}_n, M_n(\hat{h})\mathbf{w}_n\rangle - n^{-1}\operatorname{tr} M_n(\hat{h})\right|}{EL_n(\hat{h})} \to 0.$$

Now, observing that $n^{1/2}(1 - n^{-1}\|\mathbf{w}_n\|^2)$ has a bounded variance, we have by (A.1),

$$\frac{\left|1 - \left(\|\mathbf{w}_n\|^2/n\right)\right|}{\left(EL_n(\hat{h})\right)^{1/2}} \leq \frac{n^{1/2}\left|1 - \left(\|\mathbf{w}_n\|^2/n\right)\right|}{\inf_{h \geq 0}\left(nEL_n(h)\right)^{1/2}} \to 0,$$

and it suffices to use the inequality

$$n^{-1}\frac{\operatorname{tr} M_n(\hat{h})}{\left(EL_n(\hat{h})\right)^{1/2}} \leq n^{-1}\frac{\operatorname{tr} M_n(\hat{h})}{\left(\sigma^2 n^{-1}\operatorname{tr} M_n^2(\hat{h})\right)^{1/2}} \leq \frac{1}{\sigma}$$

to obtain the first required convergence. The second one results from (2.4) of Proposition 2.1. □

Now let us recall Lemma 3.1 of Li (1986) which states that the assumption (A.2) of a large "variability" in the eigenvalues of $X_n X_n^t$ is required for the

desirable behavior of $(n^{-1}\operatorname{tr} M_n(\hat{h}))^2/n^{-1}\operatorname{tr} M_n^2(\hat{h})$ when $\hat{\mathbf{f}}_n(\hat{h})$ is a consistent sequence:

LEMMA 3.3 [Li (1986)]. *Under* (A.1) *and* (A.2), *for any* $\hat{h}$ *random or not, such that*

$$L_n(\hat{h}) \to 0,$$

we have

$$\frac{\left(n^{-1}\operatorname{tr} M_n(\hat{h})\right)^2}{n^{-1}\operatorname{tr} M_n^2(\hat{h})} \to 0.$$

We are now ready to prove the following analog of Theorem 3.1.

THEOREM 3.4. *Under* (A.1), (A.2) *and* (A.3), *RGCV is a.o.*

PROOF. It is immediate to see that for any sequence $\hat{h}$ such that $L_n(\hat{h}) \to 0$, we have $n^{-1}\|(I - M_n(\hat{h}))\mathbf{y}_n\|^2 \to \sigma^2$ (see proof of Lemma 3.1), and thus by Lemma 3.3, we can apply the results of both Proposition 3.1 and Proposition 3.3. By combining them, we see that the results of Proposition 3.1 will still hold with $\operatorname{SURE}_n(\hat{h})$ replaced by $\operatorname{NRSURE}_n(\hat{h})$. In particular, this holds for $\hat{h}$ equal to the minimizer of $L_n(h)$, say h_n^* [since $\inf_{h \geq 0} L_n(h) \leq L_n(h_n) \to 0$, by (A.3)], that is,

$$\operatorname{NRSURE}_n(h_n^*) - n^{-1}\|\mathbf{e}_n\|^2 + \sigma^2 = L_n(h_n^*)\big(1 + o_p(1)\big).$$

On the other hand, by Theorem 3.3, this also holds for $\hat{h} = \hat{h}_{RG}$

$$\operatorname{NRSURE}_n\big(\hat{h}_{RG}\big) - n^{-1}\|\mathbf{e}_n\|^2 + \sigma^2 = L_n\big(\hat{h}_{RG}\big)\big(1 + o_p(1)\big).$$

Now from $\operatorname{NRSURE}_n(\hat{h}_{RG}) \leq \operatorname{NRSURE}_n(h_n^*)$ and $L_n(h_n^*) \leq L_n(\hat{h}_{RG})$, we obtain $L_n(\hat{h}_{RG})/L_n(h_n^*) \to 1$. □

4. Remarks.

REMARK 4.1. The normality assumption for the errors may be unnecessary, as noted by Speckman (1985) for the classical GCV. In the same way, weaker (or different) assumptions on the distribution of the simulated vector $\mathbf{w}_n$ might then be sufficient. Recently, Hutchinson (1989) has studied this randomized GCV and proposed a variant using a vector $\mathbf{u}_n$ of independent binomial variables equal to plus or minus one with probability 1/2, in place of $\mathbf{w}_n$. This proposal is based on the fact that $(1/n)\langle \mathbf{u}_n, M_n(h)\,\mathbf{u}_n\rangle$ has a smaller variance than any unbiased estimator of $n^{-1}\operatorname{tr} M_n(h)$ of the form $(1/n)\langle \mathbf{v}_n, M_n(h)\mathbf{v}_n\rangle$ with $\mathbf{v}_n$ a vector of iid variables [Hutchinson (1989)]. Note that $\langle \mathbf{w}_n, M_n(h)\mathbf{w}_n\rangle/\langle \mathbf{w}_n, \mathbf{w}_n\rangle$ (cf. Remark 2.1) is not of this form. However, it can be shown that the variance of $(1/n)\langle \mathbf{u}_n, M_n(h)\mathbf{u}_n\rangle$ is bounded by $(n+2)/n$ times the variance of $\langle \mathbf{w}_n, M_n(h)\mathbf{w}_n\rangle/\langle \mathbf{w}_n, \mathbf{w}_n\rangle$, and that this

bound becomes sharp for well regular problems (e.g., smoothing splines with equidistant data). Various numerical experiments in Hutchinson (1989) confirm that these two estimators, without averaging, give results (in the minimization of GCV) essentially identical to the exact computation for n as small as a few hundred. However, concerning asymptotic theory, the possible a.o. of this variant seems not at all obvious, since a binomial error does not even satisfy the condition [see (A.2) of Theorem 3.1 in Li (1985)] necessary for a good behavior of $\tilde{\mathbf{f}}_n(h)$ and $\mathrm{SURE}_n(h)$.

REMARK 4.2. The asymptotic optimality results that we have obtained for the fast randomized versions of GCV (or C_L), are very encouraging since they are exactly of the same type as the known results for the classical versions. However, we are aware that in practice, these convergence results are not sufficient to imply that the randomized version of GCV (or C_L) and the classical one will give similar performances with data sets of realistic size. The rate of convergence toward the optimal parameter is also crucial to quantify these performances. It would thus be useful to compare these rates for the two versions of GCV (or C_L). Note that such rates have recently been investigated for standard parameter estimates in the context of kernel regression, see Härdle, Hall and Marron (1988).

REFERENCES

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

DEMMLER, A. and REINSCH, C. (1975). Oscillation matrices with spline smoothing. *Numer. Math.* **24** 375–382.

ELDEN, L. (1984). A note on the computation of the generalized cross-validation function for ill-conditioned least squares problems. *BIT* **24** 467–472.

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

GIRARD, D. (1987b). Un algorithme simple et rapide pour la validation croisée généralisée sur des problèmes de grande taille. RR 669-M, Institut IMAG, Grenoble.

GIRARD, D. (1988). Détection de discontinuités dans un signal (ou une image) par inf-convolution spline et validation croisée: Un algorithme rapide nonparamétré. RR 702-I-M, Institut IMAG, Grenoble.

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. and LAURENT, P. J. (1989). Splines and estimation of nonlinear parameters. In *Mathematical Methods in CAGD* (T: Lyche and L. L. Schumaker, eds.) 273–298. Academic, New York.

GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–224.

HALL, P. and TITTERINGTON, D. M. (1987). Common structure of techniques for choosing smoothing parameters in regression problems. *J. Roy. Statist. Soc. Ser. B* **49** 184–198.

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

HUTCHINSON, M. F. (1989). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Comm. Statist. Simulation Comput.* **18** 1059–1076.

HUTCHINSON, M. F. and DE HOOG, F. R. (1985). Smoothing noisy data with spline functions. *Numer. Math.* **47** 99–106.

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.

LI, K.-C. (1990). Personal communication.

LI, K.-C. and HWANG, J. (1984). The data smoothing aspect of Stein estimates. *Ann. Statist.* **12** 887–897.

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

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

UTRERAS, F. (1980). Sur le choix du paramètre d'ajustement dans le lissage par fonctions spline. *Numer. Math.* **34** 15–28.

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.

LABORATOIRE DE MODÉLISATION ET CALCUL
51 RUE DES MATHÉMATIQUES
DOMAINE UNIVERSITAIRE
BP 53 X
38041 GRENOBLE CEDEX
FRANCE