Commentationes Mathematicae Universitatis Carolinae

Gérard Collomb; Pascal Sarda; Philippe Vieu

Weak pointwise consistency of the cross validatory window estimate in non parametric regression estimation

Commentationes Mathematicae Universitatis Carolinae, Vol. 26 (1985), No. 4, 789--798

Persistent URL: http://dml.cz/dmlcz/106415

COMMENTATIONES MATHEMATICAE UNIVERSITATIS CAROLINAE

26,4 (1985)

WEAK POINTWISE CONSISTENCY OF THE CROSS VALIDATORY WINDOW ESTIMATE IN NON PARAMETRIC REGRESSION ESTIMATION

G. COLLOMB, P. SARDA and P. VIEU

Abstract: Let (X,Y), (X_i,Y_i), $i=1,\dots,n$ be independent identically distributed $\mathbb{R}^2$ valued random vectors and let $r(\cdot) = E(Y|X=\cdot)$ be the regression of Y on X that has to be estimated from the (X_i,Y_i), $i=1,\dots,n$. We prove the weak pointwise consistency of the cross-validatory window estimate r_n defined for all real x by $\hat{r}_n(x) = r_n(x;\hat{h},(X_i,Y_i),\ i=1,\dots,n) =$ $=$ "average of Y_i, $i,..,n$: $X_i \in [x-\hat{h}/2, x+\hat{h}/2]$", with $\hat{h}$ such that $Q(\hat{h}) = \min_{h>0} Q(h)$ where

$$Q(h) = \sum_{j=1}^{n} (Y_j - r_{n-1}(X_j; h,(X_i,Y_i),\ i=1,\dots,n,\ i\neq j))^2 \mathbb{1}_{[X_j\in A]},$$

A being a compact interval and the distribution of (X,Y) being submitted to very unrestrictive conditions defining a nonparametric model.

Key words and phrases: Cross-validation, nonparametric regression, kernel estimate, bandwidth choice, weak pointwise consistency, convergence in probability.

Classification: 62G05

---

1. Introduction. Let (X,Y) be a random vector which is valued in $\mathbb{R}^2$ and let r denote the regression function of Y on X

$$r(x) = E(Y|X=x), \quad \forall x\in\mathbb{R}.$$

Let (X_i,Y_i), $i=1,\dots,n$ be a random sample from the distribution of (X,Y). The most popular nonparametric estimate of r is the kernel estimate, proposed by Nadaraya (1964) and Watson (1964) and defined by (with the convention $c/0 = 0$)

$$(1.1)\quad r_n(x) = r_n(x;h) = r_n(x;h,(X_j,Y_j),\ j=1,\dots,n) = \\ = \sum_{j=1}^{n} Y_j K((x-X_j)/h) \Big/ \sum_{k=1}^{n} K((x-X_k)/h), \quad \forall x\in\mathbb{R},$$

where K is a kernel (see e.g. Rosenblatt, 1956) and the bandwidth $h \in \mathbb{R}_*^+$, with $h=h_n$, $\forall n \in \mathbb{N}$, and $h_n \longrightarrow 0$ as $n \longrightarrow \infty$. Several pointwise or norm (including uniform convergence) properties of this estimate have been obtained by many authors: see the reviews of Collomb (1981, §3) or Collomb (1985, §3). Such results connect these properties to the asymptotic behaviour of $(h_n)_{\mathbb{N}}$. For instance Collomb (1977) [resp. Collomb, 1978] shows that the pointwise [resp. uniform, on an appropriate compact] convergence in probability of r_n towards r is satisfied if and only if $n\,h_n \longrightarrow \infty$ [resp. $n\,h_n/\mathrm{Log}\,n \longrightarrow \infty$] as $n \to \infty$ when the model is a sufficiently large class of distributions for (X,Y). However all these results do not lead to a procedure providing a value for h in (1.1). The most popular procedure for such a choice of the bandwidth h is the <u>cross validation</u> method, defining an $\hat{h}$ such that

$$(1.2) \qquad \alpha_n(\hat{h}) = \min_{h>0} \alpha_n(h)$$

with

$$(1.3) \qquad \alpha_n(h) = n^{-1} \sum_{i=1}^{n} (Y_i - r_{n,i}(X_i))^2 \, 1_A(X_i)$$

where A is a fixed interval in $\mathbb{R}$ corresponding to the domain of interest for the estimation of the regression and where

$$r_{n,i}(X_i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} Y_j \, K((X_i-X_j)/h) \Big/ \sum_{\substack{j=1 \\ j\neq i}}^{n} K((X_i-X_j)/h), \quad i=1,\ldots,n$$

[i.e. $r_{n,i}(X_i) = r_{n-1}(X_i) = r_{n-1}(X_i; h, (X_j,Y_j), j=1,\ldots,n, j\neq i)$ defined by (1.1)].

The cross <u>validatory estimate</u> $\hat{r}_n$ is defined from $\hat{h}$ by

$$(1.4) \qquad \hat{r}_n(x) = r_n(x,\hat{h}) = \sum_{j=1}^{n} Y_j \, K((x-X_j)/\hat{h}) \Big/ \sum_{j=1}^{n} K((x-X_j)/\hat{h}), \quad \forall x \in \mathbb{R}.$$

This procedure is only a formalization of an intuitive approach: most practicians make approximatively this choice of $\hat{h}$

when choosing a number h such that on a graphical display the curve $r_n(\cdot;h,(X_i,Y_i),\ i=1,\ldots,n)$ seems to be "well-centered" inside the set (X_i,Y_i), $i=1,\ldots,n$.

Notwithstanding its practical importance but seemingly for mathematical difficulties arising in the investigation of (1.4), this cross validatory estimate $\hat{r}_n$ has originated few mathematical results. These results are now discussed.

Note: in this short communication we focus our attention on cross validation for the kernel estimate. Cross validation for other estimates (splines, k-NN,...) is not at all discussed.

2. Discussion and result. The cross validation method for regression estimation (and not for curve fitting, i.e. the case of unrandom X_i, $i=1,\ldots,n$) is investigated by Hall (1984): all his results do not imply directly that

$$\hat{r}_n(x) \xrightarrow{M} r(x) \quad \text{as} \quad n \longrightarrow \infty \tag{2.1}$$

for some classical stochastic mode M, i.e. M = "P.", "L_2", "w.p.1" or completely, for x fixed or in a norm (e.g. L_2 or L_∞) sense[x]. Here we prove (2.1) with M = "P." by combining a result of Hall (1984) and a general lemma of Collomb (1979). We shall suppose the existence of the density f of the distribution of X.

Hall (1984) considers the kernel

$$K = \mathbb{1}_{[-0.5,0.5]} \tag{2.2}$$

[therefore (1.4) is the definition given in the summary] and develops his study from the results of Collomb (1976, 1977) on the

---

x) Such results follow from the more precise results of Härdle and Marron (1984): however we note that this work deals with a Lipschitz kernel K excluding (2.2) and therefore the window estimate investigated in the present paper.

bias and variance of $r_n(x;h)$ when $h=h_n$ is unrandom. These results lead to

(2.3) $I_n(h) = \int_A E(r_n(x;h) - r(x))^2\, dx =$

$$= c_1(nh)^{-1} + c_2 h^4 + o((nh)^{-1}) + h^4), h \in \mathbb{R}_*^+ ,$$

with $0 < c_1 < \infty$ and $0 < c_2 < \infty$, under appropriate conditions on the distribution of (X,Y). The minimization of this function I_n shows that

(2.4) $h^0 = h_n^0$ with $h_n^0 = (c_1/4c_2)^{1/5}\, n^{-1/5}$

satisfies

$$I_n(h^0) = \min_{h>0} I_n(h) + o(n^{-4/5}).$$

Hall (1984, p. 178) defines $\hat{h}$ from (1.3) by

(2.5) $\alpha_n(\hat{h}) = \min\{\alpha_n(h),\ \zeta n^{-1/5} \leq h \leq \lambda n^{-1/5}\},$

$$\zeta n^{-1/5} \leq \hat{h} \leq \lambda n^{-1/5},$$

where ζ and λ are sufficiently small and large constants with $\zeta < (c_1/4c_2)^{1/5} < \lambda$, and proves that

(2.6) $\beta_n(\hat{h})/I_n(h^0) \xrightarrow{P.} 1$ and $\beta_n(h^0)/I_n(h^0) \xrightarrow{P.} 1$ as $n \longrightarrow \infty$

where β_n is the random function defined on $\mathbb{R}_*^+$ by

(2.7) $\beta_n(h) = n^{-1} \sum_{i=1}^{n} (r_n(X_i,h) - r(X_i))^2.$

This very precise result shows that the optimal estimate $r_n(\cdot;h^0)$ and the cross validatory estimate $\hat{r}_n(\cdot) = r_n(\cdot,\hat{h})$ [$\hat{h}$ defined by 2.5] are in probability asymptotically equivalent according to the <u>empirical criterion</u> of comparison defined by (2.7). However (2.6) does not imply that (2.1) is satisfied for any classical stochastic mode of convergence M. More clearly the property

$$\int_A (\check{r}_n(x) - r(x))^2 f(x)dx \xrightarrow{P.} 0 \quad \text{as} \quad n \longrightarrow \infty$$

which is satisfied by $\check{r}_n(\cdot) = r_n(\cdot,h^0)$ - since $I_n(h^0) \longrightarrow 0$ as $n \longrightarrow \infty$ is not proved for $\check{r}_n = \hat{r}_n$ - but suggested by (2.6) -

(2.7).

However Hall (1984, p. 178) pointed out that his results imply

$$\hat{h}/h^0 = \hat{h}_n/h_n^0 \xrightarrow{P.} 1 \quad \text{as} \quad n \to \infty . \tag{2.8}$$

This property of $\hat{h}$ and a general result of Collomb (1979, lemme, p. 162) lead to the following theorem dealing with the weak pointwise consistency of $\hat{r}_n$ on A. We suppose $A = [a,b]$ and denote $A^{\sigma'} = [a-\sigma', b+\sigma']$, $\sigma' > 0$.

Theorem. If the distribution of (X,Y) is such that for some $\varepsilon > 0$

(i) the density f is bounded away from zero on A, is twice differentiable on A^{ε} and satisfies

$$\sup\{|f''(x+y) - f''(x)|, x \in A^{\varepsilon/2},\ 0 < y < \sigma'\} = O(\sigma'^{1/2}) \text{ as } \sigma' \to 0,$$

and

(ii) the r.r.v. Y is bounded and the regression r has two continuous derivatives on A^{ε},

then the cross validatory window estimate $\hat{r}_n$ [defined by (1.4), (2.2) and (2.5)] satisfies

$$\forall x \in A,\ \hat{r}_n(x) \xrightarrow{P.} r(x) \quad \text{as} \quad n \to \infty . \tag{2.9}$$

This result is now commented in connection with two other problems involving the kernel method and cross validation techniques.

Remarks on cross validation in the curve fitting problem. In the curve-fitting context with the following additional conditions on the unrandom X_i, $i=1,\ldots,n$,

$$X_i = i/n, \quad i=1,\ldots,n, \tag{2.10}$$

Wong (1983) proves that for $A = [0,1]$

$$\beta_n(\hat{h}_n) \xrightarrow{\text{w.p.1}} 0 \quad \text{as} \quad n \to \infty .$$

This result which involves the empirical criterion (2.7) is sub-

ject to the same criticism: it does not imply the convergence (2.1) for some classical stochastic criterion M (see also Collomb, Math. Reviews # 720259).
The same remark also concerns more precise results of Rice (1984) who proved (2.8) under assumption (2.10).

Remark on cross validation in density estimation. Our mathematical approach of cross validation for the regression kernel estimation is similar to the approach of Devroye and Penrod (1984) who investigate the properties of the kernel density estimate

$$(2.11) \qquad f_n(x;h) = (n\,h)^{-1} \sum_{i=1}^{n} K((x-X_i)/h),\ h>0$$

when (-, Theorem 2, p. 1232) the bandwidth h is a r.r.v. $h(X_1,\ldots,X_n)$. Devroye and Penrod (1984, p. 1236-1237) apply their results to cross validation procedures maximizing an empirical likelihood for the choice of h in (2.11) from $X_1,\ldots,X_n$: for such an automatic density estimate they derive some classical convergence properties of the type (M) considered just after (2.1). Lastly we note that the result of Collomb (1979, p. 162) stated below is in the spirit of these results of Devroye and Penrod: for instance Collomb (1979, p. 170) gives an application to the convergence of an heuristic regression estimate closely related to the "direct nonparametric density estimate" considered by Devroye and Penrod (1984, p. 1235, §3).

For the future of cross validation investigation we remark that this mathematical tool involved in the following proof remains valid for M = "w.p.1" or "completely" and for X_i which are $\mathbb{R}^p$ valued, $p>1$, and also does not suppose any i.i.d. type condition on (X_i,Y_i), $i=1,\ldots,n$.

3. Proof. Result (2.9) is deduced from (2.8): we use the following lemma given by Collomb (1979, p. 162) to prove the strong pointwise convergence of the k-NN kernel regression estimate (-,p. 165). This general result is repeated in the following self-contained paragraph.

3.1. Preliminary result (Collomb, 1979, lemme, p. 162). Let (A_i,B_i), i=1,...,n be random variables which are valued in $(\mathbb{E} \times \mathbb{R}^+, \mathcal{A} \otimes \mathcal{B}_{\mathbb{R}^+})$ where $(\mathbb{E},\mathcal{A})$ is a measurable space. Let M denote one of the following convergence modes: in probability ("P."), almost surely or complete. Let k be a real positive measurable function on $\mathbb{R} \times \mathbb{E}$ such that for every t and t' in $\mathbb{R}$

(3.0) $\qquad t \leq t' \Longrightarrow \forall z \in \mathbb{E},\ k(t,z) \leq k(t',z).$

Let c denote a positive real number, for any integer n and for any r.r.v. T let

$$c_n(T) = \sum_{i=1}^{n} B_i\, k(T,A_i) \Big/ \sum_{i=1}^{n} k(T,A_i).$$

Lemma. Let $(D_n)_{\mathbb{N}}$ be a sequence of r.r.v. If for any $\beta \in]0,1[$, there exist no sequences $D_n^-(\beta)$ and $D_n^+(\beta)$ of r.r.v. $(n \in \mathbb{N})$ such that

(3.1) $\qquad D_n^-(\beta) \leq D_n^+(\beta),\ \forall n \in \mathbb{N}$, and $\mathbb{1}_{\{D_n^-(\beta) \leq D_n \leq D_n^+(\beta)\}} \xrightarrow{M} 1,$

(3.2) $\qquad \sum_{i=1}^{n} K(D_n^-(\beta),A_i) \Big/ \sum_{i=1}^{n} k(D_n^+(\beta),A_i) \xrightarrow{M} \beta,$

(3.3) $\qquad c_n(D_n^-(\beta)) \xrightarrow{M} c$ and $c_n(D_n^+(\beta)) \xrightarrow{M} c,$

as $n \longrightarrow \infty$, then we have

(3.4) $\qquad c_n(D_n) \xrightarrow{M} c$ as $n \longrightarrow \infty$.

3.2. Proof of the theorem. Let x be any given element of A. We apply the lemma above to M = "P.", $\mathbb{E} = \mathbb{R}$; $\mathcal{A} = \mathcal{B}_{\mathbb{R}}$; $k(t,z) = K((x-z)/t)$, $\forall z \in \mathbb{R}$, $\forall t > 0$; $\forall n \in \mathbb{N}$, $D_n = \hat{h}_n$ as defined by

(2.5) and

$$(A_i,B_i) = (X_i,Y_i),\ i=1,\dots,n,$$

so that

$$\hat{r}_n(x) = c_n(D_n). \tag{3.5}$$

The condition "B_i valued in $\mathbb{R}_*^+$" comes from the condition "Y bounded" by a simple translation argument, and (2.2) implies (3.0). For any β fixed in $]0,1[$, the sequences

$$D_n^-(\beta) = \beta^{1/2}\, h_n^o \text{ and } D_n^+(\beta) = \beta^{-1/2}\, h_n^o, \qquad \forall n \in \mathbb{N}, \tag{3.6}$$

satisfy (3.1),(3.2) and (3.3) with M = "P." :

Proof of (3.1): the result (2.8), proved by Hall (1984) under the conditions of the theorem, implies that

$$E_n = \{D_n^-(\beta) \leq D_n \leq D_n^+(\beta)\} = \{\beta^{1/2} \leq \hat{h}_n/h_n^o \leq \beta^{-1/2}\}$$

satisfies

$$P({}^cE_n) \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty .$$

This result and the trivial equality

$$\forall \varepsilon > 0 \qquad P(|\mathbb{1}_{\{E_n\}} - 1| \geq \varepsilon) = P({}^cE_n)$$

lead to

$$\mathbb{1}_{E_n} \xrightarrow{P.} 1 \quad \text{as} \quad n \longrightarrow \infty .$$

Proof of (3.2): we remark that

$$\sum_{i=1}^{n} k(D_n^-(\beta),A_i) \Big/ \sum_{i=1}^{n} k(D_n^+(\beta),A_i) = \beta f_n(x,D_n^-(\beta))/f_n(x,D_n^+(\beta))$$

where $f_n(x,\cdot)$ is defined by (2.11). The condition (3.2) follows from the classical result of Rosenblatt (1956) ensuring that under our conditions on f

$$f_n(x,a_n) \xrightarrow{P.} f(x) \neq 0 \quad \text{as} \quad n \longrightarrow \infty$$

for $a_n = D_n^-(\beta)$ and $a_n = D_n^+(\beta)$, since formulae (2.4) and (3.6) imply $a_n \longrightarrow 0$ and $n\, a_n \longrightarrow \infty$ as $n \longrightarrow \infty$.

Proof of (3.3): this last argument implies also (see e.g.

Collomb, 1976, 1977) that

$r_n(x,D_n^-(\beta)) \xrightarrow{P.} r(x)$ and $r_n(x,D_n^+(\beta)) \xrightarrow{P.} r(x)$ as $n \to \infty$.

Hence the condition (3.3) is satisfied with $c = r(x)$ and therefore gives (2.9) from (3.4) and (3.5).

References

COLLOMB G. (1976): Estimation non paramétrique de la régression par la méthode du noyau, Thèse, Université Paul Sabatier, Toulouse.

COLLOMB G. (1977): Quelques propriétés de la méthode du noyau pour l'estimation non paramétrique de la régression en un point fixé, Comptes Rendus à l'Académie des Sciences de Paris 285, Série A, 289-292.

COLLOMB G. (1978): Conditions nécessaires et suffisantes de convergence uniforme d'un estimateur de la régression, estimation des dérivées de la régression, Comptes Rendus à l'Académie des Sciences de Paris 288,Série A,161-164.

COLLOMB G. (1979): Estimation de la régression par la méthode des k points les plus proches avec noyau: quelques propriétés de convergence ponctuelle, Lectures Notes in Mathematics 821, 159-175.

COLLOMB G. (1981): Estimation non paramétrique de la régression: revue bibliographique, International Statistical Review 49, 75-93.

COLLOMB G. (1985): Non parametric regression: an up-to-date bibliography, Mathematische Operationsforschung und Statistik, Ser. Statistics, to appear.

DEVROYE L. and PENROD C.S. (1984): The consistency of automatic kernel density estimates, Annals of Statistics 12, 4, 1231-1249.

HALL P. (1984): Asymptotic properties of integrated square error and cross-validation for kernel estimation of a regression function, Zeitschrift für Wahrscheinlichkeitstheorie u. verw. Gebiete 67, 175-196.

HÄRDLE W. and MARRON J.S. (1984): Optimal bandwidth selection in non parametric regression function estimation, preprint, Universität Heidelberg.

NADARAYA E.A. (1964): On estimating regression, Theory of Probability and its applications 9, 141-142.

RICE J. (1984): Bandwidth choice for nonparametric regression, Annals of Statistics 12, 4, 1215-1230.

ROSENBLATT M. (1956): Remarks on some nonparametric estimates of a density function, Annals of Mathematical Statistics 27, 642-669.

WATSON G.S. (1964): Smooth regression analysis, Sankhya, Ser. A, vol. 26, 359-372.

WONG W.W. (1983): On the consistency of cross validation in kernel nonparametric regression, Annals of Statistics 11, 4, 1136-1141.

Laboratoire de Statistique et Probabilités, U.A.-C.N.R.S. 745 - Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex, France

(Oblatum 30.4. 1985)