Article type: Focus Article

Bin Smoother **Article ID**

Jussi Klemelä

University of Oulu, Department of Mathematical Sciences

Keywords

nonparametric function estimation, recursive partitioning, regression analysis, regressogram

Abstract

Bin smoothers, or regressograms, are piecewise constant regression function estimators whose values are averages of the response variable over the sets of a partition of the space of the explanatory variables. First we review results about bin smoothers whose partition is regular, giving conditions for consistency and for achieving the optimal rate of convergence. Second we review representative results about bin smoothers whose partition is irregular, again giving conditions for consistency and for achieving the optimal rate of convergence. Third we give an exposition of recursive partitioning, main examples being greedy partitions and the CART methodology.

Bin Smoothing and Regressograms

Bin smoothers might be the simplest nonparametric estimators of a regression function. A bin smoother is a piecewise constant regression function estimator. The X-observation space is covered by disjoint bins and the value of a bin smoother in a bin is the average of the Y-values for the X-values inside that bin. The bins are typically rectangles but they can also be hexagons, for example. Bin smoothers are also called "regressograms". The name "regressogram" was coined by Tukey [1961]. The name is related to "histogram", which denotes a piecewise constant estimator of a density function, analogous to a regressogram. Below we use the term regressogram.

We define a regression function $f : \mathbf{R}^d \to \mathbf{R}$ as the conditional expectation $f(x) = E(Y \,|\, X = x)$, where $Y \in \mathbf{R}$ is the response variable and $X \in \mathbf{R}^d$ is the vector of explanatory variables. The regression function is estimated using data $(X_1, Y_1), \ldots, (X_n, Y_n)$, which is a sequence of identically distributed random vectors, each vector having the same distribution as (X, Y). Regressograms can also be applied in the case of a fixed design regression, where the regression function is the function $f : \mathbf{R}^d \to \mathbf{R}$ in the model $Y_i = f(X_i) + \epsilon_i$,

where $\epsilon_i \in \mathbf{R}$ are identically distributed error terms with $E\epsilon_i = 0$ and $X_i \in \mathbf{R}^d$ are fixed design points for $i = 1, \ldots, n$.

A regressogram is completely determined by defining a partition of the X-space. We discuss only partitions made of rectangles. We distinguish between regular and irregular partitions. In the one dimensional case a regular partition is a collection of intervals of length h and an irregular partition is a collection of intervals of differing lengths. In the multivariate case we can distinguish between isotropic and anisotropic regular partitions. An isotropic regular partition is a partition where all rectangles have the same side lengths h and thus the partition is a collection of cubes of volume h^d (cubic partition). An anisotropic regular partition is a partition where the side lengths of the rectangles are the same in one direction but differ across dimensions, having side lengths $h_1, \ldots, h_d$ and volumes $h_1 \cdots h_d$. In the multivariate case an irregular partition consists of rectangles, where each rectangle can have a different volume and shape. Figure 1 shows two irregular partitions. Panel (a) shows a dyadic partition (a partition that is obtained by midpoint splits) and panel (b) shows a partition that is obtained by allowing splits on a finer grid (a partition that is obtained with CART methodology).

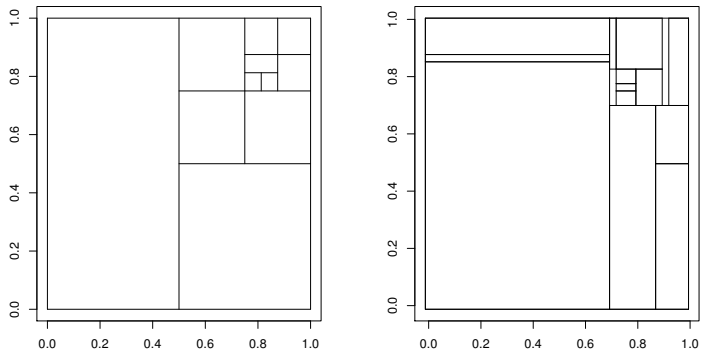

Figure 1: Irregular partitions; (a) a dyadic partition and (b) a CART partition.

Regular partitions depend on the data through the smoothing parameter h, or in the anisotropic case through smoothing parameters $h_1, \ldots, h_d$. The smoothing parameters can be chosen by cross validation or a plug-in method, for example. Irregular partitions depend more heavily on the data, because the shapes and volumes of the sets of the partition are chosen using data. We discuss cases where the partitions are chosen using penalized empirical risk minimization.

We start the article with the definition of a regressogram. After that, we give consistency and rate of convergence results for regular partitions, following

Stone [1977] and Györfi et al. [2002]. Next, results concerning irregular partitions are given. Consistency results are taken from Nobel [1996] and Györfi et al. [2002]. Rate of convergence results for irregular partitions are taken in the one dimensional case from Mammen and van de Geer [1997] and in the two dimensional case from Donoho [1997]. Finally, recursive partitioning schemes are discussed: a greedy partitioning is explained and CART approach to partition generation is introduced following Breiman et al. [1984].

Definition of a Regressogram

A regressogram, based on data $(X_1, Y_1), \ldots, (X_n, Y_n)$, is determined by a collection $A_1, \ldots, A_N \subset \mathbf{R}^d$ of sets such that they are disjoint and their union covers the observed explanatory variables:

1. $A_i \cap A_j = \emptyset$, when $i \neq j$,
2. $\{X_1, \ldots, X_n\} \subset U_{j=1}^{N} A_j$.

Now the regressogram is defined as

$$\hat{f}_n(x) = \hat{Y}_{A_j}, \qquad \text{if } x \in A_j,$$

where $\hat{Y}_{A_j}$ is the average of those response variables whose corresponding explanatory variable is in A_j. We can write, using the notation $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$,

$$\hat{Y}_A = \frac{1}{n_A} \sum_{i=1}^{n} Y_i \, I_A(X_i), \tag{1}$$

where n_A is the number of explanatory variables inside A:

$$n_A = \sum_{i=1}^{n} I_A(X_i).$$

We can write the definition of a regressogram compactly by

$$\hat{f}_n(x, \mathcal{P}) = \sum_{j=1}^{N} \hat{Y}_{A_j} I_{A_j}(x), \qquad x \in \mathbf{R}^d, \tag{2}$$

where we have also made the dependence of the regressogram on the partition $\mathcal{P} = \{A_1, \ldots, A_N\}$ explicit. Changing the order of summation in (2) we get

$$\hat{f}_n(x) = \sum_{j=1}^{N} \left(\frac{1}{n_{A_j}} \sum_{i=1}^{n} Y_i \, I_{A_j}(X_i) \right) I_{A_j}(x) = \sum_{i=1}^{n} p_i(x) \, Y_i,$$

where

$$p_i(x) = \sum_{j=1}^{N} \frac{1}{n_{A_j}} I_{A_j}(X_i) I_{A_j}(x) = \frac{1}{n_{A_x}} I_{A_x}(X_i), \tag{3}$$

and $A_x \in \{A_1, \ldots, A_N\}$ is such that $x \in A_x$.[1] Thus the regressogram belongs to the class of local averaging estimators, that have the form

$$\hat{f}_n(x) = \sum_{i=1}^{n} p_i(x)\, Y_i,$$

where $p_i(x) = p_i(x, X_1, \ldots, X_n) \geq 0$ and $\sum_{i=1}^{n} p_i(x) = 1$. The weights $p_i(x)$ of a local averaging estimator should be such that the weight $p_i(x)$ is large when X_i is close to x and the weight $p_i(x)$ is small when X_i is far away from x.

Regular Partitions

In the one dimensional a regular partition is a collection of intervals of length h. In the multivariate case a regular partition is a collection of cubes of volume h^d (isotropic case) or a rectangle partition with side lengths $h_1, \ldots, h_d$ (anisotropic case). First we give conditions for consistency and then conditions that guarantee that the rate of convergence is optimal.

Consistency

We present a consistency theorem which implies that a sufficient condition for the weak universal consistency of a regressogram is

$$\lim_{n\to\infty} h_n = 0, \qquad \lim_{n\to\infty} n h_n^d = \infty,$$

when the regressogram has a cubic partition with the side lengths h_n for the cubes. A sequence of regression function estimates $\{\hat{f}_n\}$ is called weakly consistent for a certain distribution of (X, Y) with regression function $f : \mathbf{R}^d \to \mathbf{R}$, if

$$\lim_{n\to\infty} E \int \left(\hat{f}_n(x) - f(x)\right)^2 \mu(dx) = 0,$$

where μ is the distribution of X. A sequence of regression function estimates $\{f_n\}$ is called weakly universally consistent if it is weakly consistent for all distributions of (X, Y) with $EY^2 < \infty$.

The following theorem was proved in Györfi et al. [2002, Th. 4.2, p. 60] as a corollary of the consistency theorem of Stone [1977]. The theorem gives sufficient conditions for the weak universal consistency of regressograms. The first condition is a bias condition,

[1] By symmetry we can as well write $p_i(x) = I_{A_{X_i}}(x)/n_{A_{X_i}}$.

and it requires that the bins of the underlying partition shrink to zero inside a bounded set, so the estimate is local. The second condition is a variance condition, and it requires that the number of bins inside a bounded set is small with respect to the sample size n, which implies that with a large probability each cell contains many data points.

The theorem considers a sequence of partitions indexed with the sample size, and for sample size n we denote the sets of the partition by $A_{n,1}, A_{n,2}, \ldots$

Theorem 1 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed. If for each sphere S centered at the origin*

$$\lim_{n\to\infty} \max_{j: A_{n,j}\cap S\neq\emptyset} \text{diam}(A_{n,j}) = 0$$

and

$$\lim_{n\to\infty} \frac{\#\{j : A_{n,j} \cap S\} \neq \emptyset\}}{n} = 0,$$

then the regressogram is weakly universally consistent.[2]

The consistency theorem gives a mathematical interpretation to the phenomenon that choosing a too small bin width for the regressogram leads to an estimate with small bias but large variance (small bins do not contain enough observations) and a too large bin width leads to an estimate with small variance but large bias (large bins do not allow an accurate reproduction of the regression function).

Rates of Convergence

It can be proved that a regressogram is not only consistent estimator but that its mean integrated squared error converges to zero with a fast rate, uniformly over a certain collection of distributions of (X, Y). Let us denote with $\mathcal{D}$ the collection of distributions of (X, Y) such that

1. For a constant σ^2,
$$\text{Var}(Y \mid X = x) \leq \sigma^2, \qquad x \in \mathbf{R}^d,$$
2. the regression function $f(x) = E(Y \mid X = x)$ is Lipschitz continuous: for a constant C,
$$|f(x) - f(z)| \leq C\|x - z\|, \qquad x, z \in \mathbf{R}^d,$$
3. X has compact support $S \subset \mathbf{R}^d$.

The following theorem is proved in Györfi et al. [2002, Th. 4.3, p. 64].

[2]We denote with $\text{diam}(A) = \sup\{\|x - y\| : x, y \in A\}$ the diameter of set $A \subset \mathbf{R}^d$, where $\|\cdot\|$ is the Euclidean distance, and with $\#I$ we denote the cardinality of a finite set I.

Theorem 2 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed. Let $\hat{f}_n$ be a regressogram with a cubic partition with side length*

$$h_n = C' \left(\frac{\sigma^2 + \sup_{x \in S} |f(x)|^2}{C^2} \right)^{1/(d+2)} C^{2d/(d+2)} n^{-2/(d+2)},$$

where C' is a positive constant depending on d and on the diameter of S. Then,

$$\limsup_{n \to \infty} n^{2/(d+2)} \sup_{(X,Y) \in \mathcal{D}} E \int \left(\hat{f}_n(x) - f(x) \right)^2 \mu(dx) < \infty,$$

where μ is the distribution of X.

The previous theorem shows that a regressogram can achieve the rate of convergence $O(n^{-2/(d+2)})$, for Lipschitz continuous functions and for the L_2 error. It can be proved that this rate is fastest possible for this class of distributions, see Györfi et al. [2002, Th. 3.2, p. 38] for a lower bound that shows that rate $O(n^{-2/(d+2)})$ cannot be improved for Lipschitz continuous functions. However, it can be shown that smoother regression functions, for example regression functions with s continuous derivatives can be estimated with rate $O(n^{-2s/(2s+d)})$ for the L_2 error. The estimators achieving this faster rate can be chosen as piecewise polynomials of order $s-1$ or as kernel estimators with a kernel of order s. Thus regressogram is optimal only for $s = 1$.

Irregular Partitions

We define an irregular partition to be a partition that consists of sets (rectangles) that have different shapes and volumes at different parts of the X-space. First we give a general consistency result for the regressograms with an irregular partition, then we give results about rates of convergence separately for one- and two-dimensional cases. Finally, we give an introduction to recursive partitioning and, in particular, to the CART methodology.

Consistency

Let Π be a family of partitions of $\mathbf{R}^d$. Define the partition number

$$\Delta_n(\Pi) = \max \left\{ \Delta(x_1^n, \Pi) : x_1, \ldots, x_n \in \mathbf{R}^d \right\},$$

where $\Delta(x_1^n, \Pi)$ is the number of distinct partitions of $x_1^n = \{x_1, \ldots, x_n\} \subset \mathbf{R}^d$ induced by elements of Π, i.e., the number of different partitions $\{x_1^n \cap A : A \in \mathcal{P}\}$ of x_1^n for $\mathcal{P} \in \Pi$. Let

$$M(\Pi) = \max\{\#\mathcal{P} : \mathcal{P} \in \Pi\}$$

be the maximal number of sets contained in a partition $\mathcal{P} \in \Pi$.

A method to choose a data-dependent partition is a mapping P_n that assigns to observation $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbf{R}^d \times \mathbf{R}$ a partition of $\mathbf{R}^d$. This mapping induces the family

$$\Pi_n = \left\{ P_n((x_1, y_1), \ldots, (x_n, y_n)) : (x_1, y_1), \ldots, (x_n, y_n) \in \mathbf{R}^d \times \mathbf{R} \right\}$$

of data-dependent partitions. Thus for a given observation we obtain a partition $\mathcal{P}_n \in \Pi_n$, and from this partition we obtain the regressogram $\hat{f}_n$.

The following theorem was proved in Györfi et al. [2002, Th. 13.1, p. 237] extending the results of Nobel [1996]. In contrast to Theorem 1, which considered weak consistency, Theorem 3 considers strong consistency.

Theorem 3 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed. Let Π_n be a family of data-dependent partitions and let $\hat{f}_n$ be the corresponding regressograms. Let $\bar{f}_n$ be a truncated regressogram, defined by*

$$\bar{f}_n(x) = \begin{cases} \hat{f}_n(x), & \text{if } |\hat{f}_n(x)| \leq \beta_n, \\ \beta_n \cdot \text{sign}(\hat{f}_n(x)), & \text{otherwise}, \end{cases}$$

where $\beta_n > 0$. Assume that $\lim_{n\to\infty} \beta_n \to \infty$, $\lim_{n\to\infty} \beta_n^4 / n^{1-\delta} = 0$ for some $\delta > 0$,

$$\lim_{n\to\infty} \frac{M(\Pi_n)\, \beta_n^4 \log \beta_n}{n} = 0, \tag{4}$$

$$\lim_{n\to\infty} \frac{\log(\Delta_n(\Pi_n))\, \beta_n^4}{n} = 0, \tag{5}$$

and

$$\lim_{n\to\infty} \inf_{S : S \subset \mathbf{R}^d, \mu(S) \geq 1-\delta} \mu(\{x : \text{diam}(A_n(x) \cap S) > \gamma\}) = 0 \tag{6}$$

almost surely for all $\gamma > 0$, $\delta \in (0, 1)$, where μ is the distribution of X and $A_n(x)$ is the bin $A \in \mathcal{P}_n$ of the partition which contains x. Then,

$$\lim_{n\to\infty} \int \left(\bar{f}_n(x) - f(x) \right)^2 \mu(dx) = 0$$

almost surely.

Conditions (4) and (5) require that the set of partitions from which the data-dependent partition is chosen is not too complex, i.e., the maximal number of bins in a partition and the logarithm of the partition number are small compared to the sample size. Condition (6) requires that the diameters of the bins of the data-dependent partition converge in a certain sense to zero.

Rates of Convergence

Univariate Partitions through Total Variation Penalties

Mammen and van de Geer [1997] define a class of penalized least squares estimators that contain as a special case piecewise constant estimators that are very close to regressograms. Let

$$Y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{7}$$

where $0 \leq x_1 < \cdots < x_n \leq 1$ are nonrandom and ϵ_i are independent, identically distributed, and have mean zero. Define the estimator $\hat{f}_n : [0,1] \to \mathbf{R}$ as the minimizer of

$$\sum_{i=1}^{n} (Y_i - f(x_i))^2 + \alpha \cdot \mathrm{TV}(f),$$

where $\alpha > 0$ and $TV(f)$ is the total variation metric defined by

$$\mathrm{TV}(f) = \sup \sum_{j=1}^{p-1} |f(t_{j+1}) - f(t_j)|,$$

where the supremum is over all $p \geq 2$ and all points $0 < t_1 < \ldots < t_p < 1$. In particular, for a piecewise constant right continuous function f with jump points $0 < t_1 < \ldots < t_p < 1$, $\mathrm{TV}(f) = \sum_{j=1}^{p-1} |f(t_{j+1}) - f(t_j)|$. Also, for differentiable f, $TV(f) = \int_0^1 |f'(x)|\, dx$.

Mammen and van de Geer [1997, Prop. 8] show that the estimator $\hat{f}_n$ is almost a regressogram with the partition $A_j = [t_j, t_{j+1})$, $j = 1, \ldots, p-1$, where the jump points $0 < t_1 < \cdots < t_p < 1$ of the estimate are among the design points $x_2, \ldots, x_n$. More precisely, it holds that

$$\hat{f}(x) = \hat{Y}_{A_j}, \qquad \text{for } x \in A_j,$$

where $\hat{Y}_{A_j}$ is defined in (1), unless $\hat{f}(t_j)$ is a local maximum, local minimum, minimum at the boundary, or maximum at the boundary. At local maxima the local average is moved downwards and at local minima the local average is moved upwards. For α large enough, at monotone pieces of f the estimate $\hat{f}$ behaves like an isotonic least squares estimate.[3] The partition can be calculated with an iterative algorithm based on stepwise addition and deletion of the endpoints of the intervals. For a given α the algorithm takes $O(n \log(n))$ steps and the estimate can be calculated for all α with $O(n^2)$ steps. Mammen and van de Geer [1997] consider also more generally the total variation metric of the kth derivative of the regression function as the penalty, and this leads to an estimate which is a spline of order $k-1$.

The estimator achieves optimal rates of convergence in bounded variation function classes $\mathcal{F}_C = \{f : \mathrm{TV}(f) \leq C\}$, $0 < C < \infty$. The following theorem follows from Mammen and van de Geer [1997, Th. 10].

[3] The isotonic least squares estimate is the nonparametric least squares estimator under the monotonicity restriction for the regression function.

Theorem 4 *Assuming the model (7), if $f \in \mathcal{F}_C$, then*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_n(x_i) - f(x_i)\right)^2 = O_p\left(n^{-2/3}\right).$$

The rate $O(n^{-2/3})$ is the minimax rate for the averaged squared error, because classes $\mathcal{F}_C$ are larger than the Sobolev classes $\{f : \int_0^1 (f')^2 \le C\}$, and the minimax rate of convergence has been proved to be $O(n^{-2/3})$ for the Sobolev classes, see Ibragimov and Hasminskii [1980], Stone [1982], and Nemirovskii et al. [1985]. In the total variation classes linear estimates (regressograms with a regular partition) do not achieve the optimal rate. To achieve the optimal rate the smoothing must be locally adaptive (the interval lengths of the regressogram have to change).

Dyadic CART

Dyadic CART was introduced in Donoho [1997], where the two-dimensional case $d = 2$, for the fixed equidistant design, and for the Gaussian errors is considered. Let $f : [0,1]^2 \to \mathbf{R}$ and

$$Y_i = \bar{f}(i) + \sigma\,\epsilon_i, \tag{8}$$

where $i = (i_1, i_2)$ are fixed equispaced design points, $i_1, i_2 = 0, \ldots, m-1$, m is dyadic (an integral power of 2), $\bar{f}(i)$ is the cell average over the cell C_i: $\bar{f}(i) = \int_{C_i} f/\text{volume}(C_i)$, with $C_i = [i_1/m, (i_1+1)/m) \times [i_2/m, (i_2+1)/m)$. Furthermore, ϵ_i are independent and identically distributed Gaussian random variables with mean zero and unit variance, and $\sigma > 0$. The number of observations is $n = m^2$.

Dyadic CART can be defined in two steps.

1. Let $\mathcal{P}^*$ be the largest possible dyadic partition. A dyadic partition is a partition that is obtained by midpoint splits of $[0,1]^2$. When the side length of a rectangle is m^{-1}, then this side is not allowed to be split. Thus the largest dyadic partition consists of the rectangles with volume m^{-2}.

2. Let $\mathcal{P}_\alpha$ be the partition of $[0,1]^2$ that minimizes

$$\sum_{i_1,i_2=1}^{m}\left(Y_i - \hat{f}_n(x_i, \mathcal{P})\right)^2 + \alpha \cdot \#\mathcal{P}$$

among all dyadic subpartitions of $\mathcal{P}^*$, where x_i is the midpoint of cell C_i, $\hat{f}(\,\cdot\,, \mathcal{P})$ is the regressogram with partition $\mathcal{P}$, $\alpha > 0$ and $\#\mathcal{P}$ is the cardinality of partition $\mathcal{P}$. Define the dyadic CART estimator by

$$\hat{f}_n(x) = \hat{f}_n(x, \mathcal{P}_\alpha).$$

Donoho [1997] proposes an algorithm with $O(n)$ steps for the calculation of the optimal partition, where $n = m^2$ is the number of observations.

The dyadic CART estimator has the optimal convergence rate in anisotropic Nikolskii smoothness classes. For functions $f : [0,1]^2 \to \mathbf{R}$, define the finite difference operators

$$D_h^1 f(x,y) = f(x+h,y) - f(x,y), \qquad D_h^2 f(x,y) = f(x,y+h) - f(x,y).$$

Let $0 \le \delta_1, \delta_2 \le 1$, $0 < C < \infty$, and let p be such that $1/p < \rho + 1/2$, where

$$\rho = \frac{\delta_1 \delta_2}{\delta_1 + \delta_2}\,.$$

Let the Nikolskii class of functions $f : [0,1]^2 \to \mathbf{R}$ with mixed smoothness (anisotropic class) be

$$\mathcal{F}_p^{\delta_1,\delta_2}(C) = \left\{ f : \|f\|_p \le C,\ \sup_{h \in (0,1)} h^{-\delta_k} \|D_h^k f\|_{L^p(Q_h^k)} \le C,\ k = 1,2 \right\},$$

where $Q_h^1 = [0, 1-h) \times [0,1]$ and $Q_h^2 = [0,1] \times [0, 1-h)$.[4] These classes are discussed in Nikol'skii [1969] and Temlyakov [1993].

Theorem 5 *Assume model (8) and let $\alpha \asymp \sigma^2 \log_e n$.[5] Then,*

$$\sup_{f \in \mathcal{F}_p^{\delta_1,\delta_2}(C)} E\,\frac{1}{n}\sum_{i=1}^{n} \left(\hat{f}_n(x_i) - \bar{f}(x_i) \right)^2 \le C' \left(\frac{\sigma^2 \log n}{n} \right)^{2\rho/(2\rho+1)},$$

where C' is a positive constant.

Cross validation is not needed for the choice of the smoothing parameter α, since a choice for α given in Theorem 5 gives the optimal rate of convergence, up to a logarithmic factor. The rate in Theorem 5 is the minimax rate, as proved in Donoho [1997]. We may call parameter p a "spatial inhomogeneity parameter", because for $p \ge 2$ it would suffice to use a regressogram with a regular partition which has a different number of bins in each direction (to handle the anisotropicity) but to obtain the near minimax rate for $p < 2$ it is required that the bin widths are locally adaptive, that is, the partition is irregular. A regression estimate for random design regression based on similar ideas than the Dyadic CART estimate but using piecewise polynomials was analyzed for univariate data in Kohler [1999]. Figure 1(a) shows an example of a dyadic partition.

[4] We define the L_p-norm by $\|f\|_{L_p(Q)} = (\int_Q |f(x)|^p\,dx)^{1/p}$.

[5] Notation $a_n \asymp b_n$ for positive sequences a_n, b_n means that there exists positive constants C_1 and C_2 such that $C_1 \le \liminf_{n\to\infty} a_n/b_n \le \limsup_{n\to\infty} a_n/b_n \le C_2$.

Recursive Partitioning

We review popular methods for choosing a partition of a regressogram. These methods obtain the partition by recursive splitting of the space of explanatory variables. First we define a greedy partition and second a CART partition. Figure 1(b) shows an example of a partition that can be either a greedy partition or a CART partition.

Greedy Partition

A greedy partition is a partition of the space of the explanatory variables which is found by a stepwise algorithm, which recursively splits the space to finer sets. This algorithm is called greedy, or stepwise, because we do not try to find a global minimum for the optimization problem but find the optimizer one step at a time. Morgan and Sonquist [1963] presented this type of algorithm, although they did not restrict themselves to binary splits but allowed a large number of splits to be made simultaneously.

First we define the split points over which we search the best splits. The splits are made parallel to the coordinate axes and thus we have to define a grid of possible split points for each direction. Let us denote the sets of possible split points by

$$\mathcal{G}_1, \ldots, \mathcal{G}_d, \tag{9}$$

where $\mathcal{G}_k \subset \mathbf{R}$ is a finite grid of split points in direction k. A natural possibility for choosing $\mathcal{G}_k$ is to take it to be the collection of the midpoints of the coordinates of the observations: $\mathcal{G}_k = \{Z_1^k, \ldots, Z_{n-1}^k\}$, $k = 1, \ldots, d$, where Z_i^k is the midpoint of $X_{(i)}^k$ and $X_{(i+1)}^k$:

$$Z_i^k = \frac{1}{2}\left(X_{(i)}^k + X_{(i+1)}^k\right),$$

where $X_{(1)}^k, \ldots, X_{(n)}^k$ is the order statistic of the kth coordinate of the observations $X_1, \ldots, X_n$. This choice of possible split points guarantees that all the cells, even at the finest resolution level, contain observations.

When rectangle $R \subset \mathbf{R}^d$ is splitted through the point $s \in \mathbf{R}$ in direction $k = 1, \ldots, d$, then we obtain sets

$$R_{k,s}^{(0)} = \{(x_1, \ldots, x_d) \in R : x_k \leq s\} \tag{10}$$

and

$$R_{k,s}^{(1)} = \{(x_1, \ldots, x_d) \in R : x_k > s\}. \tag{11}$$

The split point s satisfies

$$s \in S_{R,k} \stackrel{def}{=} \mathcal{G}_k \cap \text{proj}_k(R), \tag{12}$$

where $\text{proj}_k(R) = R_k$, when $R = R_1 \times \cdots \times R_d$. We say that partition $\mathcal{P}$ is grown if it is replaced by partition

$$\mathcal{P}_{R,k,s} = \mathcal{P} \setminus \{R\} \cup \left\{R_{k,s}^{(0)}, R_{k,s}^{(1)}\right\}, \tag{13}$$

where rectangle $R \in \mathcal{P}$ is splitted in direction $k = 1, \ldots, d$ through the point $s \in S_{R,k}$.

A greedy partition $\mathcal{P}$ is one of the partitions in the sequence of partitions found by the following procedure.

- Start with the partition $\{\mathbf{R}^d\}$ and split the rectangles of the partition as long as some rectangle contains a sufficient number of observations.
- Make splits so that the empirical risk of the corresponding regressogram is minimized. The minimization is done over all rectangles in the current partition, over all directions, and over all split points in the given rectangle and in the given direction.

We describe the procedure more precisely in the following definition. The partition is grown by minimizing an empirical risk of the estimator, which is typically defined as the sum of squared errors of the estimator $\hat{f}$.

Definition 1 Greedy Partitions *A sequence of greedy partitions $\mathcal{P}_1, \ldots, \mathcal{P}_M$, with minimal observation number $m \geq 1$, is defined recursively by the following rules.*

1. *Start with the partition $\mathcal{P}_1 = \{R\}$, where $R = \mathbf{R}^d$.*
2. *Assume that we have constructed partitions $\mathcal{P}_1, \ldots, \mathcal{P}_L$, where $L \geq 1$.*
   (a) *If all $R \in \mathcal{P}_L$ satisfy $\#\{X_i \in R\} \leq m$, then partition $\mathcal{P}_L$ is the final partition.*
   (b) *Otherwise, we construct next partition $\mathcal{P}_{\hat{R},\hat{k},\hat{s}}$, where*

$$\left(\hat{R}, \hat{k}, \hat{s}\right) = \text{argmin}_{(R,k,s)\in I} \sum_{i=1}^{n} \left(Y_i - \hat{f}(X_i, \mathcal{P}_{R,k,s})\right)^2, \tag{14}$$

where

$$I = \{(R,k,s) \quad : \quad R \in \mathcal{P}_L,\ \#\{X_i \in R\} \geq m, \\ k = 1, \ldots, d,\ s \in S_{R,k}\},$$

$S_{R,k}$ is the set of split points defined in (12), $\mathcal{P}_{R,k,s}$ is the partition defined in (13), and $\hat{f}(\cdot, \mathcal{P})$ is the regressogram defined in (2).

Let

$$\hat{\mathcal{P}} \in \{\mathcal{P}_1, \ldots, \mathcal{P}_M\},$$

be a greedy partition, where $\mathcal{P}_1, \ldots, \mathcal{P}_M$ are defined in Definition 1. The greedy regressogram is defined by

$$\hat{f} = \hat{f}\left(\cdot, \hat{\mathcal{P}}\right),$$

where $\hat{f}$ is defined in (2). We can use sample splitting to find a good partition $\hat{\mathcal{P}}$ and thus a good regressogram. Let $n^* = [n/2]$ and use the data (X_i, Y_i), $i = 1, \ldots, n^*$, to construct the sequence $\mathcal{P}_1, \ldots, \mathcal{P}_M$ and the corresponding sequence of estimators $\hat{f}_1, \ldots, \hat{f}_M$. Then we calculate for each estimate the sum of squared residuals using the second part of the data:

$$\text{SSR}_m = \sum_{i=n^*+1}^{n} \left(Y_i - \hat{f}_m(X_i)\right)^2, \qquad m = 1, \ldots, M.$$

The final estimate is $\hat{f}_{\hat{m}}$, where $\hat{m} = \text{argmin}_{m=1,\ldots,M} \text{SSR}_m$.

CART

CART (classification and regression tress) procedure was introduced in Breiman et al. [1984]. In the previous section a sequence of partitions was constructed in a stepwise manner and then one partition was selected from this sequence, using sample splitting, to define the regressogram, CART constructs the sequence of partitions in a different way. First a fine partition is grown with stepwise optimization and then the sequence of partitions is found by a complexity penalized pruning. The new way of constructing the sequence opens up the possibility for using cross validation to choose the final partition, instead of sample splitting. Also, the complexity penalized pruning may increase the quality of partitions in the sequence. In contrast to dyadic CART the large partition $\mathcal{P}^*$ is now data dependent. Otherwise the final estimate is obtained analogously as in dyadic CART, by minimizing a complexity penalized sum of squared residuals. The CART sequence is found by the following steps.

1. Choose a large partition $\mathcal{P}^*$. This partition is the largest partition $\mathcal{P}_M$ from the sequence of greedy partitions defined in Definition 1.

2. For $\alpha \geq 0$, let

$$\mathcal{P}_\alpha = \text{argmin}_{\mathcal{P} \subset \mathcal{P}^*} \left[\sum_{i=1}^{n} \left(Y_i - \hat{f}(X_i, \mathcal{P})\right)^2 + \alpha \cdot \#\mathcal{P} \right], \tag{15}$$

   where $\hat{f}$ denotes a regressogram as defined in (2). For $\alpha = 0$, $\mathcal{P}_\alpha = \mathcal{P}^*$, and for large enough α, $\mathcal{P}_\alpha = \{\mathbf{R}^d\}$. Since there are a finite number of subsets of $\mathcal{P}^*$, there are a finite number of values $0 = \alpha_1 < \cdots < \alpha_M$ such that

$$\mathcal{P}_\alpha = \mathcal{P}_{\alpha_i}, \text{ when } \alpha_i \leq \alpha < \alpha_{i+1}, \tag{16}$$

   for $i = 1, \ldots, M$, and we denote $\alpha_{M+1} = \infty$. Now $\mathcal{P}_{\alpha_1} = \mathcal{P}^*$ and $\mathcal{P}_{\alpha_M} = \{\mathbf{R}^d\}$.

Definition 2 CART Partitions *A sequence of CART partitions $\mathcal{P}_1, \ldots, \mathcal{P}_M$ is defined, with an abuse of notation, by $\mathcal{P}_i = \mathcal{P}_{\alpha_i}$, $i = 1, \ldots, M$, where $\alpha_1, \ldots, \alpha_M$ is defined by (16).*

We can use cross validation to find a good partition and thus a good regressogram. In the case of greedy partitions we had to use sample splitting, that is, twofold cross validation, but in the case of CART partitions the penalization parameter α can be used to connect different partitions and we can use K fold cross validation for $2 \leq K \leq n$. Let us denote by $I_1, \ldots, I_K$ a partition of the index set $\{1, \ldots, n\}$, where $2 \leq K \leq n$. Typically we partition observations into $K = 10$ subsets (ten fold cross validation) but at most we can partition the observations to n subsets and at least to two subsets. Observations (X_i, Y_i), $i \notin I_k$, are used to construct sequence $\hat{f}_{\alpha_{1,k}}, \ldots, \hat{f}_{\alpha_{M_k,k}}$, where $\alpha_{1,k} < \cdots < \alpha_{M_k,k}$, $k = 1, \ldots, K$. For each estimate in the sequence we calculate the average of squared residuals (ASR) using (X_i, Y_i), $i \in I_k$:

$$\text{ASR}_{j,k} = \frac{1}{\#I_k} \sum_{i \in I_k} \left(Y_i - \hat{f}_{\alpha_{j,k}}(X_i) \right)^2, \qquad j = 1, \ldots, M_k, \ k = 1, \ldots, K.$$

Finally we use the complete data to find a sequence $\hat{f}_{\alpha_1}, \ldots, \hat{f}_{\alpha_M}$ and a grid $\alpha_1, \ldots, \alpha_M$. We make a partition of $(0, \infty) = U_{m=1}^{M} A_m$, where $\alpha_m \in A_m$ and estimate

$$\text{ASR}_{\alpha_m} = \frac{\sum \{SSR_{j,k} : \alpha_{j,k} \in A_m\}}{\#\{(j,k) : \alpha_{j,k} \in A_m\}}, \qquad m = 1, \ldots, M.$$

The final estimate is $\hat{f}_{\alpha_{\hat{m}}}$, where $\hat{m} = \text{argmin}_{m=1,\ldots,M} \text{ASR}_{\alpha_m}$.

We need two algorithms to find the sequence $\mathcal{P}_1, \ldots, \mathcal{P}_M$: a growing algorithm for growing the large partition $\mathcal{P}^*$ and a pruning algorithm for producing the sequence from this large partition. Both algorithms use the fact that the partitions which we consider can be represented as binary trees, where the rectangles of the partition are the nodes of the tree. The representation as a binary tree follows from the stepwise splitting procedure. We take the whole space to be the root of the tree. After that, when a node (a rectangle) is splitted, the two obtained rectangles are taken to be the child nodes of the splitted node.

We choose $\mathcal{P}^*$ as the largest partition $\mathcal{P}_M$ from Definition 1. We can now use a faster algorithm to obtain $\mathcal{P}^*$ than the algorithm described in Definition 1, since this algorithm uses unnecessary time to optimize the order in which the partition is grown, and we are interested only in the final partition and not in the intermediate partitions. Thus we can use an algorithm based on the following recursion. Let the minimal observation number be $m \geq 1$.

1. Start with the partition $\mathcal{P} = \{\mathbf{R}^d\}$. The rectangle $\mathbf{R}^d$ is taken to be the root node of the initial binary tree.

2. Assume that we have constructed partition $\mathcal{P}$. This partition is interpreted as a binary tree.

   (a) If all child nodes $R \in \mathcal{P}$ satisfy $\#\{X_i \in R\} \leq m$, then we finish the splitting.

(b) Otherwise, choose a child node $R \in \mathcal{P}$ with $\#\{X_i \in R\} > m$. Construct new partition $\mathcal{P}_{R,\hat{k},\hat{s}}$, where

$$\left(\hat{k}, \hat{s}\right) = \text{argmin}_{(k,s)\in I_R} \sum_{i=1}^{n} \left(Y_i - \hat{f}(X_i, \mathcal{P}_{R,k,s})\right)^2,$$

where $I_R = \{(k,s) : k = 1,\ldots,d,\ s \in S_{R,k}\}$, $S_{R,k}$ is the set of split points defined in (12), $\mathcal{P}_{R,k,s}$ is the partition defined in (13), and $\hat{f}(\cdot, \mathcal{P})$ is the regressogram defined in (2). Partition $\mathcal{P}_{R,\hat{k},\hat{s}}$ is interpreted as a binary tree, where rectangle $R^{(0)}_{\hat{k},\hat{s}}$ is the left child node of node R and rectangle $R^{(1)}_{\hat{k},\hat{s}}$ is the right child node of node R, where we use the notation of (13).

After growing the large partition $\mathcal{P}^*$ we need an algorithm to find the CART sequence of Definition 2. To solve for a given α the complexity penalized minimization problem (15), we can use a dynamic programming algorithm which starts at the leaves of the binary tree T^* corresponding to $\mathcal{P}^*$. If t is a node of T^*, denote the sum of squared residuals associated with this node by

$$\text{ssr}(t) = \sum_{i:X_i \in R_t} \left(Y_i - \bar{Y}_{R_t}\right)^2,$$

where R_t is the rectangle associated with node t. Denote with $Q(t)$ the sum of $\text{ssr}(t')$ over the leafs t' of the subtree T_t whose root is t. Starting at the leaf nodes, we compare at each node t whether

$$Q(t) + \alpha \cdot \#T_t < \text{ssr}(t) + \alpha, \tag{17}$$

where $\#T_t$ is the number of leaves in the subtree T_t. If this holds, then the subtree whose root is t should be kept, because the complexity penalized error is smaller than obtained by making t a leaf node. Otherwise, the tree is pruned at node t and t is made a leaf node. The value $Q(t)$ can be calculated during the pruning process.

To extend this idea to find the complete CART sequence and the corresponding values $\alpha_1, \ldots, \alpha_M$, note that we have for every nonterminal node t of T^* that $Q(t) < \text{ssr}(t)$. As long as (17) holds, branch T_t has a smaller error-complexity than the single node $\{t\}$, but at some critical value of α, the two error-complexities become equal. At this point the subbranch $\{t\}$ is smaller than T_t, has the same error-complexity, and is therefore preferable. To find this α, solve (17) to get

$$\alpha < \frac{\text{ssr}(t) - Q(t)}{|T_t| - 1}.$$

The algorithm is based on finding the "weakest links", which are the nodes minimizing

$$g_k(t) = \begin{cases} \frac{\text{ssr}(t) - Q(t)}{|T_t| - 1}, & t \text{ is not a leaf in } T_k \\ \infty, & t \text{ is a leaf in } T_k, \end{cases} \tag{18}$$

$k = 1, \ldots, K$. Let $t_1 = \text{argmin}_{t\in T_0} g_0(t)$, $T_0 = T^*$. Then t_1 is the root node and $\alpha_1 = g_0(t_1) = 0$. Let T_1 be the subtree of T^* obtained by making t_1 a leaf node. We continue

in this way: $t_k = \text{argmin}_{t \in T_{k-1}} g_{k-1}(t)$ and $\alpha_k = g_{k-1}(t_k)$ for $k = 1, \ldots, M$.[6] We get a sequence where $\alpha_1 = 0$, and for $\alpha_k \leq \alpha < \alpha_{k+1}$, the corresponding partition $\mathcal{P}_{\alpha_k}$ is the collection of rectangles associated with the leaf nodes of tree T_k.

Cross-References

Nonparametric curve estimation, Nonparametric regression, Wavelet methods

Conclusion

Regressograms with a regular partition can have the optimal rate of convergence for estimating functions in Sobolev classes with smoothness index $s = 1$. Regressograms with irregular partition can have the optimal rate of convergence for estimating functions in total variation classes or functions in Nikolskii classes when the spatial inhomogeneity parameter p is smaller than two.

In the one dimensional case, when $X \in \mathbf{R}$, a regressogram is useful in estimating functions with jumps. Estimation of function with jumps is related to change-point estimation. In the multivariate case, when $X \in \mathbf{R}^d$ with $d > 1$, regressogram is useful in estimating piecewise constant functions, like images ($d = 2$). For high dimensional analysis recursive partition is an viable alternative in multivariate regression function estimation, because CART type procedures perform implicit variable selection.

References

L. Breiman, J. Friedman, R. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall, New York, 1984.

D. L. Donoho. Cart and best-ortho-basis: A connection. *Ann. Statist.*, 25:1870–1911, 1997.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, Berlin, 2002.

I. A. Ibragimov and R. Z. Hasminskii. On the nonparametric estimation of regression. *Soviet Math. Dokl.*, 21:810–814, 1980.

M. Kohler. Nonparametric estimation of piecewise smooth regression functions. *Statist. Probab. Lett.*, 43:49–55, 1999.

E. Mammen and S. van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1):387–413, 1997.

J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.*, 58:415–434, 1963.

[6] If at any stage there is a multiplicity of weakest links, for example $g_{k-1}(t_k) = g_{k-1}(t'_k)$, then define $T_k = T_{k-1} - T_{t_k} - T_{t'_k}$.

A. S. Nemirovskii, B. T. Polyak, and A. B. Tsybakov. Rate of convergence of nonparametric estimators of maximum likelihood type. *Probl. Inf. Transmiss.*, 21:258–272, 1985.

S. M. Nikol'skii. *Approximation of Functions of Several Variables and Imbedding Theorems.* Springer, Berlin, 1969.

A. Nobel. Histogram regression estimation using data dependent partitions. *Ann. Statist.*, 24:1084–1105, 1996.

C. J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5:595–645, 1977.

C. J. Stone. Optimal rates of convergence for nonparametric regression. *Ann. Statist.*, 10:1040–1053, 1982.

V. N. Temlyakov. *Approximation of Periodic Functions*. Nova Press, 1993.

J. Tukey. Curves as parameters and toch estimation. In *Proc. 4th Berkeley Symposium*, pages 681–694, 1961.