I N S T I T U T D E

S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N

P A P E R

0334

MONOTONE REGRESSION

I. GIJBELS

http://www.stat.ucl.ac.be

MONOTONE REGRESSION

Abstract. In nonparametric regression the objective is to explore the relationship between an explanatory variable and the variable of interest. Sometimes it is plausible to assume that the regression function is monotone. We review nonparametric methods for estimating a monotone regression function, as well as testing procedures for testing for monotonicity. A brief discussion on related estimation and testing problems is given.

Keywords and Phrases. Concordance, constrained estimation, critical bandwidth, isotonic regression, local methods, monotone estimate, nonparametric regression, ranks, running gradients, spline smoothing, testing for monotonicity, tilting method.

AMS Subject Classification. Primary: 62G08, 62G10, 62G05; Secondary: 62G09.

INTRODUCTION

In univariate regression studies the main interest is in investigating the influence of an explanatory variable X on the variable of interest Y. Often it is rather natural to assume that a regression relationship is monotone. For example, it is to be expected that a response increases or decreases when increasing dose of a drug. If monotonicity is a valid assumption, then the estimated regression function should be monotone too. If the monotonicity assumption is not a valid one, then imposing the monotonicity constraint on a regression estimate is a wrong action. In situations of doubts about the validity of the monotonicity assumption there is a need for testing for monotonicity before proceeding with the, unconstrained or constrained, estimation of the regression function.

More formally, with Y the response variable and X the explanatory variable, one is interested in the mean regression function $E(Y|X = x)$, based on an i.i.d. sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ from (X, Y). In nonparametric regression one is faced with the situation that no prior information on a particular form of the regression function is available. For example, there is no justification for assuming that Y changes linearly, say, with X, such as is the case in simple linear regression. There are various methods available for estimating an unknown regression function $E(Y|X = x)$ nonparametrically. Among the most simple methods is one that simply averages Y_i-observations for which the X_i-observations are close to the value x. Such a method is known as a local averaging method. In the next section we briefly review nonparametric methods for estimating a regression function without imposing the constraint of monotonicity of the function. This leads to unconstrained estimators.

ESTIMATING A MONOTONE REGRESSION FUNCTION

The main question addressed in this section is that of how to estimate a regression function when it is assumed that this is a monotone function. In other words, how to obtain a constrained estimator — an estimator that satisfies the (natural) monotonicity constraint? Before giving an overview of the most important approaches for estimating a regression function under the constraint of monotonicity, we briefly review the most commonly-used methods for unconstrained regression estimation. The next question is then whether these methods can readily be adapted to accommodate the monotonicity constraint or whether more work is needed to obtain constrained estimators.

Unconstrained estimators

We briefly discuss a few methods for estimating a regression function nonparametrically. These methods are meant as illustrative examples and also serve the discussion on constrained estimation.

Kernel estimators and local polynomial fitting

The intuitive idea behind kernel estimation is as follows. Since no parametric form of the regression function is assumed, estimation of this regression function at a fixed point x will mainly be influenced by observations 'close' to that point. Therefore the mean regression function $m(x) = E(Y|X = x)$ is estimated by an average of Y_i-observations for which the X_i-observation falls in a neighbourhood of x. It is also natural here to not give equal weights to each observation in that neighbourhood: the observations closest to x should get more weight than more distant observations.

Kernel estimators typically have a form

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} w_i(x) Y_i}{\sum_{j=1}^{n} w_j(x)} = \sum_{i=1}^{n} \left(\frac{w_i(x)}{\sum_{j=1}^{n} w_j} \right) Y_i \ , \tag{1}$$

where $w_i(x)$ is the weight associated to the observation X_i that is 'close' to x, and where the denominator is just the sum of all these weights. These are estimates of a local averaging type.

If one would give weight $1/n$ to all observations then this would result in a simple averaging of the observations (what would be done if the aim is to estimate $E(Y)$).

In classical kernel estimation, the weights are determined by a kernel function (often chosen to be a symmetric probability density function) and a bandwidth or smoothing parameter that determines, together with the support of the kernel function, the size of the neighbourhood. More precisely commonly-used weights are

$$w_i(x) = \frac{1}{h} K\left(\frac{X_i - x}{h}\right) \equiv K_h(X_i - x) \ , \tag{2}$$

where K denotes the kernel function and h the bandwidth parameter. Note that the function $K_h(\cdot)$ denotes a rescaled version of K.

The estimator $\widehat{m}(x)$ in (1) with weights as in (2) is the well-known Nadaraya-Watson regression estimator introduced in [27, 34], and denoted in the sequel by $\widehat{m}_{\mathrm{NW}}(\cdot)$.

There is another way of viewing this estimator. Indeed, it is easily seen that it is obtained via

$$\widehat{m}_{\mathrm{NW}}(x) = \underset{\beta}{\operatorname{argmin}} \sum_{i-1}^{n} (Y_i - \beta)^2 w_i(x) \ , \tag{3}$$

with $w_i(x)$ as in (2). So, the Nadaraya-Watson estimator is nothing but the result of fitting a constant function *locally* to the observed data, using a *weighted* least-squares method.

Another commonly-used kernel type estimator is the Gasser-Müller estimator, see [16], that is defined as

$$\widehat{m}_{\mathrm{GM}}(x) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(u-x)du \, Y_{(i)}, \tag{4}$$

where $Y_{(i)}$ is the Y-observation associated with the ordered $X_{(i)}$-observation, and where $s_i = \frac{X_{(i)}+X_{(i+1)}}{2}$ (with $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$) denotes the midpoint between two consecutive X-observations. Note that the sum of the weights equals

$$\sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(u-x)du = \int_{s_0}^{s_n} K_h(u-x)du = \int_{-\infty}^{+\infty} K_h(u-x)du = \int_{-\infty}^{+\infty} K(v)dv \ ,$$

by a change of variable, and that this is 1 when K is a probability density function. So, the Gasser-Müller estimator only differs from the Nadaraya-Watson estimator in the definition of the weights which for the Gasser-Müller estimator are $w_i(x) = \int_{s_{i-1}}^{s_i} K_h(u-x)du$. Both estimators are in fact the result of fitting locally a constant function via the least-squares method.

Generalizing this simple idea, one can also proceed with fitting locally to the observed data a polynomial of degree p using least-squares. In other words, suppose that the regression function m can be approximated locally by

$$\begin{aligned} m(z) &\approx m(x) + m'(x)(z-x) + \frac{m''(x)}{2!}(z-x)^2 + \cdots + \frac{m^{(p)}(x)}{p!}(z-x)^p \\ &= \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \\ &\equiv \sum_{j=0}^{p} \beta_j (z-x)^j \ , \end{aligned}$$

for z in a neighbourhood of x, by using Taylor's expansion. Using weighted least-squares regression we would then consider the minimization problem

$$\min_{\beta_0,\cdots,\beta_p} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \right\}^2 K_h(X_i - x) \ . \tag{5}$$

Denote by $\widehat{\beta}_j$, $j = 0, \cdots, p$, the solution to this minimization problem. Then it is easily seen from the definition of β_j that an estimator for $m^{(j)}(x)$ is given by $j!\widehat{\beta}_j$. So, this method allows to simultaneously estimating the regression function $m(x)$ as well as all derivatives up to order p. Estimation of the entire functions is done by solving the weighted least-squares problem (5) for a grid of points x.

These estimators are known as local polynomial estimators. Special cases are the local constant estimators ($p = 0$), such as the Nadaraya-Watson and Gasser-Müller estimators. Another particular case is that of local linear fitting ($p = 1$), in which the minimization problem reads as:

$$\min_{\beta_0,\beta_1} \sum_{i-1}^{n} \{Y_i - \beta_0 - \beta_1(X_i - x)\}^2 K_h(X_i - x) \ .$$

It is easily seen that the estimator for $m(x)$ resulting from this is given by

$$\widehat{m}_{\mathrm{LL}}(x) = \frac{\sum_{i=1}^{n} w_i(x) Y_i}{\sum_{j=1}^{n} w_j(x)} \ , \tag{6}$$

with $w_i(x) = K_h(X_i - x)\{S_{n,2} - (X_i - x)S_{n,1}\}$, where $S_{n,\ell} = \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(X_i - x)^\ell, \ell = 0, 1, 2$.

Local linear estimation has some advantages above local constant estimation (see [14]), and therefore we often prefer to use this estimator instead of local constant estimates.

Local polynomial fitting has been studied in detail in, for example, [14].

Spline methods

A different nonparametric estimation method consists of assuming that the unknown regression function can be approximated by a so-called spline function. A linear spline is a continuous function that is piecewise linear. A quadratic spline is a function that is piecewise quadratic with a continuous first derivative. A cubic spline is a function that is piecewise cubic with continuous first two derivatives. In general, a polynomial spline of degree d is a function that is piecewise polynomial of degree d and with continuous first $d-1$ derivatives. More formally, let $t_1, \cdots, t_J$ be a fixed sequence of 'knots' such that $-\infty < t_1 < t_2 < \cdots < t_J < +\infty$. Then, a polynomial spline of degree d is a $d-1$ times continuously differentiable function and such that the function on each of the intervals $(-\infty, t_1], [t_1, t_2], \cdots, [t_{J-1}, t_j], [t_J, +\infty)$ is a polynomial of degree d. The collection of all splines of degree d forms a linear space of dimension $J + d + 1$. This is easily seen, since on each subinterval the function is presented via $d + 1$ parameters (a polynomial of degree d), and at each knot point there are d constraints imposed on the function (continuity of the function itself and its first $d-1$ derivatives). This leads to $(J+1)\times(d+1) - J\times d = J+d+1$ number of free parameters. Possible sets of basis functions for this linear space of polynomial splines of degree

d are a power basis, consisting of functions $(x - t_j)_+^d, (j = 1, \cdots, J), 1, x, x^2, \cdots, x^d$, and the B-splines basis as defined by [2]. Here $(x - t_j)_+ = \max(x - t_j, 0)$ denotes the positive part of the function $x - t_j$.

Let $B_1, B_2, \cdots, B_{J+d+1}$ denote a basis for the linear space of polynomial splines of degree d. Then such a polynomial spline can be expressed as

$$s(x) = \sum_{j=1}^{J+d+1} \theta_j B_j(x) .$$

For a given sequence of knots, the *regression spline estimate of degree d* of m is then defined as the best spline approximation via least-squares, i.e.

$$\widehat{m}(x) = \sum_{j=1}^{J+d+1} \widehat{\theta}_j B_j(x) , \tag{7}$$

where $\widehat{\theta}_j$ are solutions of the least-squares problem:

$$\min_{\theta_1, \cdots, \theta_{J+d+1}} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{J+d+1} \theta_j B_j(X_i) \right\}^2 .$$

Note that for calculation of this estimator one needs to fix, in advance, the number of knot points as well as their positions, and one needs to choose a basis in the linear space.

A possibility to avoid the choice of knots is via the so-called *smoothing spline* method. For simplicity, we explain the method in case of cubic splines. The smoothing spline method then considers a penalized least-squares regression

$$\sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 + \lambda \int \{m''(x)\}^2 dx , \tag{8}$$

where $\lambda > 0$ is a smoothing parameter. The first term measures the goodness of the fit for a given function m. The second term is a penalty term for functions that are too rough (i.e. for which m is such that the second derivative function m'' has a quite unsmooth behaviour). The smoothing parameter balances between the two terms: if it is put to 0, the resulting minimizer of (8) will simply be the function that interpolates the data; with increasing λ one puts more importance into the smoothness property of the estimator; the extreme case that $\lambda = +\infty$ results in an estimator for which the penalty term equals 0 (to compensate for the exploding λ), resulting in a simple linear regression with $m(x) = \alpha + \beta x$. The smoothing spline estimator is obtained by minimizing (8) over the space of functions that are twice differentiable and for which $\int \{m''(x)\}^2 dx < \infty$.

Incorporating smoothing constraints: a general framework

Some of the nonparametric estimation methods are more natural for dealing directly with the constraint problem. For the regression spline estimation method (7), the idea is to construct a basis of monotone spline functions and to consider nonnegative linear combination of these basis functions. The construction of such monotone spline functions is discussed in [28].

The smoothing spline method can also incorporate the monotonicity constraint in a natural way, by simply minimizing the penalized sum of squares (8) over the space of *monotone* twice differentiable functions for which $\int \{m''(x)\}^2 dx < \infty$. See also [28, 29] for comments on the calculation and implementation issues for estimation of monotone regression functions using the regression spline method or the smoothing spline method.

Recently, in [25] it was shown that the essence of the above idea is not limited to methods that directly involve minimization over function spaces. The key issue is to try to view the methods as a projection of the data with respect to appropriate norms. This implies working with much larger normed vector spaces than is usually done.

As an illustration we explain how estimates of the form (3), such as for example the Nadaraya-Watson and Gasser-Müller estimators, can be viewed as projection methods.

As mentioned before minimization problem (3) needs to be solved for at least a grid of points x, leading to the estimated values of the regression function at these grid points. Consider the counting measure ν defined on the set of all grid points. Then this set of minimization problems is equivalent to finding

$$\widehat{m}_S = \underset{m}{\operatorname{argmin}} \int \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(x)\}^2 w_i(x)\nu(dx) \ , \tag{9}$$

where the integration is over the real line, and where the factor $1/n$ has no impact on the minimization problem. The problem remains equivalent if ν is any other measure: the integration with respect to $\nu(dx)$ has no effect because the minimum can be found for all x individually. Another natural choice for ν is the Lebesgue measure, $\nu(dx) = dx$, for which the integration corresponds to Lebesgue integration.

Having viewed the equivalence between minimization problem (3) for obtaining the estimated regression function, and minimization problem (9), the natural formulation of the constrained estimation problem becomes transparant. Indeed, instead of considering minimization in (9) over all functions m, one considers minimization over the set of functions satisfying the constraint and denoted by the set $\mathcal{C}$:

$$\widehat{m}_{S,C} = \underset{m\in\mathcal{C}}{\operatorname{argmin}} \int \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(x)\}^2 w_i(x)\nu(dx) \ , \tag{10}$$

In contrast to the unconstrained minimization problem (9), the measure ν has an effect since the minimizers at different points are linked through the constraints on the function m. To calculate the constrained estimator, it is common to choose ν to be the counting measure associated with the points at which one wants to calculate the constrained estimator. The Lebesgue measure would be a natural choice for theoretical analysis.

Note that the above constraint estimation method is valid for any type of constraint C, not only for monotonicity constraints.

We now explain how to view these estimates (9) and (10) as projection methods. We therefore introduce a normed vector space, denoted by $\mathcal{V}_S$ that contains the space of data vectors and the space of candidate regression functions. Within this vector space one identifies two subspaces: the subspace of the data vectors $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$, denoted by $\mathcal{V}_S^Y$; and the subspace of the candidate regression functions, denoted by $\mathcal{V}_S^m$. It will be shown hereafter that in the space $\mathcal{V}_S$, the simple smoothers of type (3) appear as projections of the data vector into the subspace $\mathcal{V}_S^m$. For simplicity of presentation, we consider

$$\mathcal{V}_S = \left\{ \mathbf{f} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} : \ f_i : I\!R \rightarrow I\!R, i = 1, \cdots, n \right\} .$$

The data vector $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ can be viewed as an element of this space $\mathcal{V}_S$ by taking $f_i(x) = Y_i$ for each i, i.e. by taking constant functions f_i. The subspace of such n-tuples of constant functions will be denoted by $\mathcal{V}_S^Y$. Consider a candidate regression function m. This can be viewed as an element of the space $\mathcal{V}_S$ by taking $f_i(x) = m(x)$ for all i, i.e. taking all f_i's equal to m. The subspace of n-tuples with identical entries will be called $\mathcal{V}_S^m$. When $w_i(x) \geq 0$, an inner product on $\mathcal{V}_S$ is defined by

$$\langle \mathbf{f}, \mathbf{g} \rangle = \int \frac{1}{n} \sum_{i=1}^{n} f_i(x) g_i(x) w_i(x) \nu(dx) \ ,$$

and the norm induced on $\mathcal{V}_S$ by

$$||\mathbf{f}||^2 = \int \frac{1}{n} \sum_{i=1}^{n} f_i^2(x) w_i(x) \nu(dx) \ .$$

Having introduced these concepts, it now appears that the unconstrained and constrained minimization problems, (9) and (10) respectively, can be written as,

$$\widehat{m}_S = \underset{\mathbf{m} \in \mathcal{V}_S^m}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{m}\|^2$$

and

$$\widehat{m}_S = \underset{\mathbf{m} \in \mathcal{C}_S^m}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{m}\|^2 , \tag{11}$$

respectively, where $\mathcal{C}_S^m \subset \mathcal{V}_S^m$ is the subset of n-tuples with identical entries that are constrained by being monotone in x.

Denote by $\widehat{\mathbf{m}}_S$ the projection of $\mathbf{Y}$ on the subspace $\mathcal{V}_S^m$. Then $\mathbf{Y}-\widehat{\mathbf{m}}_S$ is orthogonal to $\widehat{\mathbf{m}}_S - \mathbf{m}$ with respect to the inner product. Using a Pythagorean relationship one then has

$$\|\mathbf{Y} - \mathbf{m}\|^2 = \|\mathbf{Y} - \widehat{\mathbf{m}}_S\|^2 + \|\widehat{\mathbf{m}}_S - \mathbf{m}\|^2$$

and minimization problem (11) is then reduced to

$$\widehat{m}_S = \underset{\mathbf{m} \in \mathcal{C}_S^m}{\operatorname{argmin}} \|\widehat{\mathbf{m}}_S - \mathbf{m}\|^2 ,$$

and since

$$\begin{aligned} \|\widehat{\mathbf{m}}_S - \mathbf{m}\|^2 &= \int \frac{1}{n} \sum_{i=1}^{n} \left\{\widehat{m}_S(x) - m(x)\right\}^2 w_i(x)\nu(dx) \\ &= \int \left\{\widehat{m}_S(x) - m(x)\right\}^2 w(x)\nu(dx) , \end{aligned}$$

with $w(x) = \frac{1}{n}\sum_{i=1}^{n} w_i(x)$, this is equivalent to

$$\widehat{m}_S = \underset{m \in \mathcal{C}}{\operatorname{argmin}} \int \left\{\widehat{m}_S(x) - m(x)\right\}^2 w(x)\nu(dx) .$$

So, the projection of the data vector $\mathbf{Y}$ on the space $\mathcal{C}_S^m$ (in the enlarged space $\mathcal{V}_S^m$) is the same as the projection (in the space of ordinary functions) of the unconstrained smooth onto $\mathcal{C}$. This reduces the computations.

Local polynomial fitting can also be viewed as a projection method, but since it involves fitting locally with a polynomial of degree p (and not a constant function as in the previous cases) it involves normed vector spaces which are sets of $n(p+1)$ tuples of functions. For details see [25]. The latter work shows that many of the smoothing methods can in fact be viewed as a projection of the data, and hence this work offers a nice general projection framework for constrained smoothing.

Typically the estimates obtained by these methods often show 'kinks', essentially where the unconstrained estimator if non-monotone. See also more later.

Isotonic regression approach

This approach is among the first and most classical approaches. Denote by $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ the order statistics corresponding to the X_i-observations. As before let $Y_{(i)}$ denote the Y-observation associated to $X_{(i)}$. The regression function $m(x) = E(Y|X = x)$ is assumed to be nondecreasing. In [5] Brunk proposed the following nondecreasing estimator for m at the point $X_{(i)}$:

$$\widehat{m}_I(X_{(i)}) = \max_{s\leq i} \min_{t\geq i} \sum_{j=s}^{t} \frac{Y_{(j)}}{t-s+1} \ .$$

When linear interpolating all points $\{\widehat{m}_I(X_{(i)})\}$ one obtains the *isotonic least-squares regression estimator*. See also [1].

There are two other ways to look at the above estimator. The points $\{m_I(X_{(i)})\}$ are the solution of a least-squares problem: find the monotone nondecreasing set of points $\{Z_1, \cdots, Z_n\}$ that minimizes

$$\sum_{i=1}^{n} (Z_i - Y_i)^2 \ .$$

A second equivalent definition of the isotonic regression estimator is as follows. Denote by $S_k = \sum_{j=1}^{k} Y_{(j)}$, $k = 1, \cdots, n$, and let $S_0 = 0$. The cumulative sum diagram

(CSD) of $\{Y_{(j)}\}$ is defined as $\{(k, S_k) : 0 \leq k \leq n\}$. Let G be the supremum of all convex functions on $[0, n]$ whose graphs lie below the CSD, i.e. G is the greatest convex minorant (GCM) of the CSD on $[0, n]$. Then an equivalent definition of the above estimator is:

$$\widehat{m}_I(X_{(i)}) = \text{the left-hand slope of } G \text{ at } i\ , \quad 1 \leq i \leq n\ .$$

We now provide an illustration of the construction of the isotonic regression estimate for a simple example of five observations. The following table shows the observed values $X_{(i)}$ and $Y_{(i)}$, $i = 1, \cdots, 5$, the calculated values for S_k, for $k = 1, \cdots, 5$, as well as the calculated values for $\widehat{m}_I(X_{(i)})$. The latter values were calculated using the cumulative sum diagram and the greatest convex minorant.

Table 1 About Here.

Figure 1 shows the observed data, together with the isotonic regression estimate. In the right-hand panel of Figure 1 we depict the cumulative sum diagram and the greatest convex minorant (in dashed line) for this example.

Figure 1 About Here.

Statistical properties of the isotonic regression estimator have been studied in [4, 5, 6], [22] and [19], among others. Some drawbacks of the isotonic regression estimator are: the linear interpolation part is rather arbitrary; often there are too many flat parts in the estimator, and the rate of convergence of the estimator is $n^{-1/3}$ (at points with positive slope $m'(x)$). The question is whether one could not obtain smoother estimators with a better rate of convergence.

In the paper [26] it was proposed to add a smoothing step (using an appropriate kernel) and to smooth the isotonic regression estimate: i.e. based on the "isotonised data" $(X_{(1)}, \widehat{m}_I(X_{(1)})), \cdots, (X_{(n)}, \widehat{m}_I(X_{(n)}))$ one obtains by smoothing with kernel K and bandwidth h the estimator:

$$\widehat{m}_{IS}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_{(i)} - x}{h}\right) Y_i^*}{\sum_{i=1}^{n} K\left(\frac{X_{(i)} - x}{h}\right)} \ ,$$

where, for simplicity of notation, we denoted $Y_i^* = \widehat{m}_I(X_{(i)})$, the pseudo data obtained after 'isotonisation' of the raw data.

Hence, the estimator $\widehat{m}_{IS}(\cdot)$ is obtained after two steps: (i) isotonisation of the data; and (ii) smoothing of the calculated pseudo data by a kernel estimator. Interchanging these two steps leads to smoothing the data first and then calculating the isotonic regression estimate of the smoothed data. In other words, obtain

$$\widehat{m}_{S}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \ ,$$

and then consider the pseudo data $(X_1, \widehat{m}_S(X_1)), \cdots, (X_n, \widehat{m}_S(X_n))$ and obtain the isotonic regression estimator based on these pseudo data, i.e.

$$\widehat{m}_{SI}(X_{(i)}) = \max_{s \le i} \min_{t \ge i} \sum_{j=s}^{t} \frac{Y_{(j)}^{*S}}{t - s + 1} \ , \tag{12}$$

where we denoted $Y_j^{*S} = \widehat{m}_S(X_{(j)})$, the pseudo data obtained after 'smoothing' of the raw data.

Clearly, the estimator $\widehat{m}_{SI}(x)$ is a nondecreasing function, whereas the estimator $\widehat{m}_{IS}(x)$ is not necessarily nondecreasing. In [23] the (asymptotic) properties of both estimators (in the case of fixed equidistant design) were studied. This study led to the following findings: the estimator $\widehat{m}_{IS}(x)$ has always a smaller variance and a larger bias than the estimator $\widehat{m}_{SI}(x)$; which one is better depends on the chosen kernel function; the two estimators are asymptotically first order equivalent; both estimators have rate of convergence $n^{-2/5}$.

The estimator (12) is in fact similar to the estimator defined in (10), as was mentioned in [25].

This and the previous estimation approach have in common that the produced constrained estimates are less 'smooth' than the usual unconstrained estimates. For example, typically the estimates 'flatten' out the unconstrained estimate at places where this estimate is non-monotone. As a consequence, the estimates often show 'kinks'.

Tilting method

This approach is among the more recent ones, and was motivated by the fact that previous methods produce constrained estimates that appear less 'smooth' as the unconstrained estimates.

Consider estimators that can be expressed in the form

$$\widehat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} A_i(x) Y_i = \sum_{i=1}^{n} \frac{1}{n} A_i(x) Y_i \ , \tag{13}$$

where $A_i(x)$ is a weight function depending on x, h, $X_1, \cdots, X_n$, but not on $Y_1, \cdots, Y_n$. Examples of such estimators are the Nadaraya-Watson estimator, the Gasser-Müller estimator and the local linear estimator, defined in (1) & (2), (4) and (6). The weight functions $A_i(x)$ for these three estimators are given in the following table.

Table 2 About Here.

A method for monotonizing such (unconstrained) estimates to obtain estimates that are monotone is the so-called tilting method introduced in [21]. The basic idea is to adjust the unconstrained estimate by *tilting* the empirical distribution so as to make the least possible change (with respect to a certain distance measure) subject to imposing the monotonicity constraint. Below we describe how this tilting is carried out.

Instead of considering an estimator such as in (13), one considers the generalization

$$\widehat{m}(x|p) = \sum_{i=1}^{n} p_i \, A_i(x) Y_i \ ,$$

where $p = (p_1, \cdots, p_n)$ is a probability distribution on the set $\{X_1, \cdots, X_n\}$. When this probability distribution is the uniform discrete distribution one has $p_{\text{unif}} \equiv \left(\frac{1}{n}, \cdots, \frac{1}{n}\right)$, and this leads to the unconstrained estimates.

There are several important remarks to be made regarding special merits of such a tilting method. First, the resulting constrained estimate will have the same smoothness as the unconstrained estimate since the smoothness of the estimate is essentially determined by the smoothness properties of the weight functions. Secondly, the method is applicable to general kernel methods, and only modifies them in regions where there is violence of the monotonicity constraint. Away from places where the unconstrained estimate is non-monotone, the constrained and unconstrained estimates are very close. Thirdly, the smoothing parameter h can be chosen prior to and separately from the monotonizing step, using the available methods in the unconstrained estimation problem. Fourthly, the method can also be used when one wants to impose strict monotonicity.

In order to impose that $\widehat{m}(\cdot|p)$ is increasing on an interval $\mathcal{I}$ one chooses $p = \widehat{p}$ to minimize a distance $D(p)$ from p to the uniform distribution subject to $\widehat{m}'(\cdot|p) \geq 0$ on the interval $\mathcal{I}$. If one wants to impose strict monotonicity one requires instead that $\widehat{m}'(\cdot|p) \geq \epsilon$ for some $\epsilon > 0$. Since $p = (p_1, \cdots, p_n)$ needs to be a distribution function the additional constraints $\sum_{i=1}^{n} p_i = 1$ and $\min p_i \geq 0$ are needed.

What are appropriate distance measures to use in this context? Some suitable measures were introduced in [8], and are

$$D_\rho(p) = \frac{1}{\rho(1-\rho)} \left\{ n - \sum_{i=1}^{n} (np_i)^\rho \right\}$$

for $-\infty < \rho < +\infty$ and $\rho \neq 0, 1$. Further, there are the Kullback-Leibler divergence measures

$$D_0(p) = -\sum_{i=1}^{n} \log(np_i) \qquad \text{and} \qquad D_1(p) = \sum_{i=1}^{n} p_i \log(np_i)$$

The distance $D_{1/2}(p)$ corresponds to the Hellinger distance.

For all values of ρ a zero distance $D_\rho(p) = 0$ only appears if and only if $p = p_{\text{unif}}$. Furthermore, in case $0 \leq \rho \leq 1$, the measure $D_\rho(p)$ is not well defined, unless $p_i \geq 0, \quad \forall i$, or strictly positive in case $\rho = 0$.

Implementation of the tilting method for monotonizing an estimator is straightforward, and uses existing procedures of quadratic programming. Important findings are that the constrained estimator equals a constant multiple (close to 1) of the unconstrained estimator, in places that are not near dips or bumps. In dips or bumps on the other hand, the constrained estimator corrects for non-monotonicity, and returns, graphically seen, quickly to the unconstrained estimator on either side of those places. Varying ρ usually makes little difference for the resulting estimators.

Discussion and related estimation issues

We discussed three quite different approaches for obtaining monotone estimates. The key ideas behind the approaches can also be applied in other contexts than this of monotone regression.

It is clear from the presentation that the projection framework for constrained smoothing is not limited to the monotonicity constraint and to a specific smoothing method.

Techniques similar to the isotonic regression approach have been applied in different contexts, such as for estimating a monotone density function or convex functions. The tilting method or variants of it have been used to estimate monotone hazard functions and estimating unimodal densities.

All methods discussed here had as primary goal to produce monotone estimates (or more generally constrained estimates). Different objectives could be to find out which of the observed features of an estimate are 'really there', or to produce estimates that control the number of local extremes. Paper [7] has the former as main objective, whereas the recent work [9] deals with estimation of a regression function with emphasis on controlling the number of local extremes.

TESTING FOR MONOTONICITY

In case there is doubt whether the unknown regression function is monotone or not, but one suspects that monotonicity might be a sensible assumption, then there is a need for testing for monotonicity. The testing problem of interest is

$$\begin{aligned} H_0 : &\quad m \text{ is monotone} \\ H_1 : &\quad m \text{ is non-monotone} \end{aligned} \tag{14}$$

There are several approaches for obtaining testing procedures for this problem. We review the main approaches, and discuss also briefly their applicability in other testing problems.

Among the very first approaches to testing for monotonicity, there is the theoretical work by [31] who studies a test based on the greatest discrepancy of an estimate of the derivative of the regression function from zero. That work does not discuss a practical implementation.

Critical bandwidth approach

A way for obtaining testing procedures for problem (14) is based on ideas adopted from testing for multimodality of a density function. Therefore we first discuss this testing problem.

Testing for multimodality of a density function

Among the important characteristics of a density function f is its multimodality: is the density function unimodal, bimodal, trimodal, ...? In other words: how many modes does a density have?

Let $X_1, \cdots, X_n$ be an i.i.d. sample from X having density f. A test for testing

$$\begin{aligned} H_0 : &\quad \text{the number of modes of } f \text{ is } \leq k \\ H_1 : &\quad \text{the number of modes of } f \text{ is } > k \, , \end{aligned} \tag{15}$$

with k a nonnegative integer, was proposed in [32] using a kernel density estimate for the unknown density function,

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \, ,$$

where h is a bandwidth parameter and K is a kernel function, defined similarly as for nonparametric kernel estimation of the regression function.

An important observation is that, with a suitably chosen kernel function, the modality of a kernel density estimate is a monotone decreasing function of the smoothing parameter: if h increases from small values to large values, the modality of the function $\widehat{f}(\cdot)$ decreases, i.e. $\widehat{f}(\cdot)$ has less and less modes. Define now the 'critical' bandwidth as the smallest bandwidth that just forces the density estimate to have the modality of the null hypothesis. If the null hypothesis is not true, then the smallest bandwidth that forces the estimate to have no more than k modes typically will be a large bandwidth (to impose the smaller modality). If on the other hand the null hypothesis is true, this critical bandwidth will be relatively small. So, intuitively such a 'critical bandwidth' is a good statistic for distinguishing between H_0 and H_1 in (15): reject H_0 when the 'critical bandwidth' is too large (above a threshold value). This threshold value or critical value is determined by controlling the probability of type I error of the decision rule, i.e. the probability of rejection the null hypothesis when H_0 is true. Calculation of this probability requires knowledge of the distribution of the test statistic (the 'critical bandwidth'), but unfortunately this distribution is unknown. A way out is then to approximate the distribution of the test statistic by using a (smoothed) bootstrap procedure. See also [33, 23] and [15].

Critical bandwidth testing for monotonicity of a regression function

We now turn to testing problem (14). Consider a nonparametric kernel type estimator $\widehat{m}(\cdot)$ of the unknown regression function $m(\cdot)$, such as, for example, a Nadaraya-Watson estimator, a Gasser-Müller estimator or a local linear estimator. Suppose that one has the following important feature for this estimate: the monotonicity of the estimate $\widehat{m}(\cdot)$ is monotone in the bandwidth.

Figure 2 About Here.

If such a 'Monotonicity of Monotonicity' property holds, then by adopting Silverman's idea for testing for multimodality of a density function (see [32]), one would consider a 'critical' bandwidth: the bandwidth at which the estimated function changes from non-monotonic to monotonic (for all larger bandwidths). See Figure 2. If the null hypothesis of monotonicity is true, this critical bandwidth will be relatively small. If the monotonicity is false, the critical bandwidth will typically be large to force monotonicity. Again, to construct the precise testing procedure (by controlling the probability of type I error) one needs to know the distribution of the test statistic under the null hypothesis. The idea could then be to develop a smoothed bootstrap procedure that leads to an approximation of this distribution. So far, the basic ideas for the critical bandwidth approach for testing for monotonicity. We now describe the resulting testing procedure.

Consider observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ from the regression model

$$Y_i = m(X_i) + \varepsilon_i \ , \tag{16}$$

where ε_i denotes the error term. Let $\widehat{m}_{\mathrm{LL}}(\cdot; h)$ denote the local linear estimate of m, as defined in (6). The test statistic is then defined as

$$h_{\text{crit}} = \inf \left\{ h : \widehat{m}_{\mathrm{LL}}(x; h) \text{ is a monotone function } \right\} \ .$$

A threshold value or critical value for the test would then be a value c such that

$$\text{probability of type I error} = P \left\{ h_{\text{crit}} > c | H_0 \right\} = \alpha \ ,$$

where α, the significance level of the test, is small and fixed in advance. Among the typical choices is $\alpha = 0.05$.

Alternatively one might calculate the p-value of the test statistic. Denote by h_c the observed value of the test statistic. The p-value of the test statistic is then given by

$$p - \text{value} = P \left\{ h_{\text{crit}} > h_c | H_0 \right\} \ .$$

This p-value is then compared with α. If the p-value is less than α (which is equivalent to $h_c > c$) we would reject H_0. If the p-value is greater than or equal to α, one would not reject H_0.

To calculate the p-value one of course also needs to know the distribution of the test statistic h_{crit}. We approximate this unknown distribution via a smoothed bootstrap procedure. Such a procedure consists of resampling from the original data B times, and to approximate the p-value via

$$\frac{1}{B}\#\left\{b = 1, \cdots, B : h_c^{*b} > h_c\right\} ,$$

where h_c^{*b} denotes the observed value for the critical bandwidth based on the bth resampling from the original data. This implies that for each resampling one needs to search for the smallest bandwidth that makes the estimate monotone. These calculations can be reduced by observing that

$$h_c^{*b} > h_c \iff \widehat{m}^{*b}(x; h_c) \text{ is non-monotone} ,$$

where $\widehat{m}^{*b}$ is the regression estimate based on the bth resample. Hence a shortcut for the above counting leads to the following approximation for the p-value:

$$\frac{1}{B}\#\left\{b = 1, \cdots, B : \widehat{m}^{*b}(x; h_c) \text{ is non-monotone }\right\} .$$

The outline of the testing proposal is then

STEP 1: Find the critical bandwidth h_c, say, which is the smallest h such that $\widehat{m}(x; h_c)$ is monotone;

STEP 2: Construct good estimates $\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n$ of the residual errors ε_i in (16);

STEP 3: Generate a bootstrap sample $\widehat{\varepsilon}_1^*, ..., \widehat{\varepsilon}_n^*$ from $\widehat{\varepsilon}_1, ..., \widehat{\varepsilon}_n$ and hence a bootstrap dataset $Y_i^* = \widehat{m}(X_i; h_c) + \widehat{\varepsilon}_i^*$, $i = 1, ..., n$;

STEP 4: Apply $\widehat{m}$ using h_c to $\{(X_i, Y_i^*), i = 1, ..., n\}$ and observe whether or not the result is monotone;

STEP 5: Repeat Steps 1 and 4 a large number of times, constructing the p–value by determining the proportion of non–monotone results at Step 4.

The intuition behind this specific proposal, and in particular Step 3, is that the p-value is a probability under H_0, and hence resampling is done using a monotone regression estimate.

There are some crucial issues though that remain to be studied: does the 'monotonicity of monotonicity' property hold?; how to construct a valid bootstrap procedure for assessing the distribution of the test statistic under the null hypothesis? We now discuss briefly these important issues, which are related to Steps 1 and 2 of the above outline.

Checking the 'monotonicity of monotonicity' property of kernel type estimators is not an easy task. For the Gasser-Müller estimator it has been shown that this holds. See [3]. The property does however not transfer to the local linear regression estimator. Via simulations one has been able though to observe that departures from monotonicity of monotonicity (i.e. occasions when monotonicity of the regression estimate holds for some bandwidth h but not for some $h_1 > h$) are extremely unusual, and also shortlived: as an example— out of 90 000 simulations this phenomenon occurred on only 2 occasions. Therefore, working with the local linear estimator leads to an acceptable testing procedure.

The second remaining issue, is how to bootstrap from the residuals $\varepsilon_i = Y_i - m(X_i)$? Assume for simplicity that the error terms ε_i's are i.i.d. with mean zero ($E(\varepsilon_i) = 0$) and variance σ^2 ($\mathrm{Var}(\varepsilon_i) = \sigma^2$). Then, the 'best' estimation of ε_i is equivalent to best estimation of m itself : minimizing $\sum_{i=1}^{n} (\widehat{\varepsilon}_i - \varepsilon_i)^2$ is equivalent to choosing $\widehat{m}$ to minimizing $\sum_{i=1}^{n} (\widehat{m}(X_i) - m(X_i))^2$.

These remarks then lead to the more detailed Steps 1 and 2 of the smoothed bootstrap procedure:

STEP 1: Find the critical bandwidth h_c which is the smallest h such that $\widehat{m}(x; h_c)$ is monotone, regardless of whether $\widehat{m}(x; h_1)$ may be non–monotone for some $h_1 > h_c$;

STEP 2: For $i = 1, ..., n$, calculate $\widehat{\varepsilon}_i = Y_i - \widehat{m}(X_i; h_0)$ where h_0 is the bandwidth selected by, for example, the method provided in [30].

Running gradients approach

The critical bandwidth approach for testing for monotonicity suffers from some difficulties: in cases where the true regression function is flat, or nearly flat, the critical bandwidth approach has low power. To understand this intuitively, consider m nondecreasing within a compact interval I, but flat on a subinterval J of I. Assume normal errors, regularly-spaced design points, and estimate m by a local linear estimator, using a compactly supported kernel. Then, as n increases, the critical bandwidth h_{crit} converges to a positive constant (depending on the support of the kernel and the length of J). However, the same is true if, at another location, the curve $m(x)$ has a sufficiently small dip. Consequently, the existence of a flat part can effectively mask a small dip at another place.

A first possible remedy for this drawback could be to permit the bandwidth to have a degree of local variation. In the context of density estimation this is the approach followed by [15]. An alternative approach is a procedure based on running gradients introduced by [20]. Consider regression model (16) with x_i regularly-spaced points or conditioned values of r.v.'s. From now on we assume that the x_i-values have been ordered. In the paper [20] the authors want to test the null hypothesis

$$H_0 : m \text{ is nondecreasing on the interval } I \ . \tag{17}$$

The idea is to consider subsequent groups of consecutive points, to fit straight lines through these points via least-squares, and to obtain the estimated slopes for all groups of points. If the regression function in nondecreasing then it is expected that there are no large negative values of slopes. Formally, with r and s fixed integers, satisfying $0 \leq r \leq s-2 \leq n-2$, one minimizes

$$S(a, b|r, s) = \sum_{i=r+1}^{s} \{Y_i - (a + bx_i)\}^2 \ , \tag{18}$$

to obtain the least-squares estimates

$$(\widehat{a}, \widehat{b}) = \operatorname*{argmin}_{(a,b)} S(a, b|r, s) \ .$$

Figure 3 contains a simple illustration of the idea for the same observations as in Figure 1, with $s-r = 3$. There are three groups of each three consecutive observations. In this example the estimated gradients were -0.605, -0.789 and 0.875 respectively.

Figure 3 About Here.

Since the variance of the x_i-observations can be quite different for different groups of $s-r$ points, this should be taken into account when looking at the slope estimates. The variance of a group of $s-r$ points is estimated by

$$Q(r, s)^2 = \sum_{i=r+1}^{s} \left\{ x_i - \frac{1}{s-r} \sum_{j=r+1}^{s} x_j \right\}^2 \ ,$$

and one can show that

$$\operatorname{Var}\left\{\widehat{b}(r, s) Q(r, s)\right\} = \sigma^2 \qquad \forall (r, s) \ ,$$

which means that $Q(r, s)$ is a kind of standardizing factor.

The test statistic considered is then

$$T_\ell = \max_{0 \le r \le s-\ell \le n-\ell} \left\{ -\widehat{b}(r,s) Q(r,s) \right\} ,$$

with ℓ an integer such that $2 \le \ell \le n$. The null hypothesis is then rejected if the value of T_ℓ is too large.

The main question is then how to assess the distribution of the test statistic under the null hypothesis. In other words: how to *calibrate* the test? In [20] it is proposed to calibrate the test for the most difficult nondecreasing form of m for which to test, namely m a constant function. When the regression function is constant, the model is $Y_i = \text{constant } + \varepsilon_i$, and hence minimizing (18) is equivalent to minimizing

$$S^c(a,b|r,s) = \sum_{i=r+1}^{s} \{\varepsilon_i - (a + bx_i)\}^2 .$$

The distribution of this quantity only depends on the distribution of the errors. One then distinguishes two cases: the case of normal errors ($\varepsilon_i \sim \mathrm{N}(0;\sigma^2)$) and the case of an unknown error distribution. When the errors are assumed to be normal, then the distribution of the statistic T_ℓ can be obtained analytically (at least theoretically). If this is too involved, one can simply approximate this distribution by simulating from a normal distribution, and as such approximating the distribution of T_ℓ via Monte-Carlo simulations. See [20] for more details and a discussion on how to deal with the second case.

Signs of differences

The idea of tracing back possible non-monotonicity by following gradient estimates of subsequent groups of consecutive points is quite appealing. However, there are even simpler ways to trace back possible non-monotonicity. For simplicity of explanation, suppose that one looks at the signs of all subsequent differences $Y_{i+1} - Y_i$, for $i = 1, \cdots, n-1$. If the function m is nondecreasing, then one expects the signs of these differences to be mostly positive. If however m departs from being a non-decreasing function, then the signs of these differences should show some tendency towards negativity. This is the basic idea behind the testing procedures proposed in [17]. We now describe these procedures in more detail. Consider the hypothesis

$$\begin{aligned} H_0: &\quad m \text{ is nondecreasing on the interval } I = (x_1, x_n) \\ H_1: &\quad m \text{ is decreasing on a subinterval } J \text{ of } I \\ H_0': &\quad m \text{ is constant on } \; I \; . \end{aligned}$$

Define the signs of the differences between (not necessarily consecutive) observations Y_{i+k} and Y_i:

$$S_{i,k} = \operatorname{sgn}(Y_{i+k} - Y_i)\,, \quad i = 1, \cdots, n-k\;.$$

Under H_0 or H_0', the signs S_{i+k} have no particular tendency towards negativity, but under H_1 they are more likely to be negative than positive on J. There are various ways of constructing tests based on these signs. We discuss tests based on counting the signs (leading to the so-called 'count' statistic) and tests based on runs of the signs (resulting in the so-called 'run' statistic).

Consider a given string of signs

$$\mathcal{S}_r = (S_{j+1,k}, \cdots, S_{j+r,k})$$

and let $N(\mathcal{S}_r)$ denote the number of +1's in the sequence (or alternatively, the number of -1's in the sequence). On can show that under H_0, $N(\mathcal{S}_r) \geq \frac{1}{2}r + O(r^{1/2})$ with high probability, and that $N(\mathcal{S}_r)$ will take smaller values under H_1. The count test statistic is defined as

$$T_{\text{count}} = T_{\text{count}}(k, r_1, r_2) = \max_{r_1 \leq r \leq r_2} \; \max_{0 \leq j \leq n-r-k} r^{-1/2} \left[\frac{1}{2}r - N(\mathcal{S}_r)\right]\;,$$

and will tend to be larger under H_1 than under H_0.

To define the run statistic, consider again a string of signs $\mathcal{S}_r = (S_{j+1,k}, \cdots, S_{j+r,k})$ and let $R(j,k)$ denote the supremum of integers r such that $S_{j+1,r}, \cdots, S_{j+r,k}$ is a sequence solely of -1's and put

$$T_{\text{run}} = T_{\text{run}}(k) = \max_{0 \le j \le n-k} R(j,k) \ .$$

Then, again T_{run} tends to be larger under H_1 than under H_0.

As in all testing procedures an important issue is how to assess the distribution of the test statistic under the null hypothesis. As we will see below, this is particularly easy within these testing procedures due to the fact that they are simply based on signs.

Let T denote either T_{count} or T_{run}. Then we reject H_0 if T is too large. The calibration of the test is done by considering the most difficult situation, i.e. that of a constant function m. Under H_0', the differences reduce to

$$Y_{i+k} - Y_i = \varepsilon_{i+k} - \varepsilon_i \ ,$$

and this fact facilitates further work a lot. Indeed, suppose that the errors ε_i's are i.i.d. from a continuous distribution function F. Then the distribution of the sign of the difference $\varepsilon_{i+k} - \varepsilon_i$ does not depend on F since

$$\varepsilon_{i+k} - \varepsilon_i > 0 \quad \Longleftrightarrow \quad \varepsilon_{i+k} > \varepsilon_i \quad \Longleftrightarrow \quad F(\varepsilon_{i+k}) > F(\varepsilon_i)$$

with $F(\varepsilon_i) = U_i$ a sequence of independent Uniform distributed r.v's.

Therefore, the distribution of the signs of $\varepsilon_{i+k} - \varepsilon_i$ can be obtained from the distribution of signs of differences of independent uniformly distributed random variables. Instead of trying to determine the exact distributions of the count or run statistic, one can also rely on the approximations of these distributions, obtained via Monte-Carlo simulations from any arbitrary continuous distribution function F. More precisely, denote $\mathcal{Y} = \{Y_1, \cdots, Y_n\}$ and let $\mathcal{Y}^* = \{Y_1^*, \cdots, Y_n^*\}$ be a set of independent r.v.'s with an arbitrary continuous distribution. Let T^* be the version of T arising from replacing $\mathcal{Y}$ by $\mathcal{Y}^*$. We then conduct B simulations and get an (approximate) critical value from the obtained B values of T^*.

There are some specific merits of these test statistics that are worth mentioning. First of all the methods are simple, since they are purely based on looking at signs and differences of Y_i observations, and require simple Monte-Carlo simulations for calibrating the testing procedures. In addition, they provide exact tests for monotonicity under the 'most difficult' null hypothesis, and the tests are thus guaranteed conservative. The theoretical properties that were shown in [17] revealed also that the 'count' test has asymptotic power 1 against local alternatives which are at least a distance $n^{1/2}k^{-1}(\log n)^{1/2}$ away from a constant function, and the 'run' test has asymptotic power 1 against local alternatives which are at least a distance $nk^{-1}(\log n)^{-1}$ away from a constant function. So, theoretically the count test (asymptotically) outperforms the run test (for local alternatives). From computational point of view though the run test takes less computation time.

Method based on local linear ranks

Another way for following the overall structure of the Y-observations is by assigning ranks to the Y-observations within a block of Y-observations, and to define local linear rank statistics based on these. This idea has been exploited in the paper [11] in a more general regression context. For given integers (j, k), with $0 \leq j < k \leq n$, let $R_{jk}(i)$ denote the rank of Y_i among the $k - j$ observations $Y_{j+1}, \cdots, Y_k$, i.e.

$$R_{jk}(i) = \sum_{h=j+1}^{k} \mathbf{1}\{Y_h \leq Y_i\} \ .$$

For the block of $k - j$ observations $Y_{j+1}, \cdots, Y_k$ one defines the linear rank statistic

$$T_{jk} = \sum_{i=j+1}^{k} \beta\left(\frac{i-j}{k-j+1}\right) q\left(\frac{R_{jk}(i)}{k-j+1}\right) \ ,$$

with β and q score functions, defined on $]0,1[$, such that $\beta(1-u) = -\beta(u)$ and $q(1-u) = -q(u)$ for $0 < u < 1$, and β and q are non-decreasing on $]0,1[$. For example $\beta(u) = q(u) = 2u - 1$ is an appropriate choice, corresponding to a Wilcoxon score function. Other linear rank statistics are obtained with other choices of score functions. For all possible blocks of Y-observations one considers these local linear rank statistics. A multiscale test statistic is then defined as

$$T = \max_{0 \leq j < k \leq n} \left(s_{k-j} |T_{jk}| - c_{k-j} \right) ,$$

with s_{k-j} and c_{k-j} suitably chosen normalizing constants. The statistic T is used for finding intervals on which a monotone trend in the data is significant.

An alternative test statistic would be to restrict to pairs (j,k) for which $k - j \leq \ell$ with ℓ a pre-specified integer in the set $\{1, \cdots, n\}$:

$$T = \max_{1 \leq d \leq \ell} \left(\max_{j,k \in \{0,1,\cdots,n\} : k-j=d} (s_d |T_{jk}| - c_d) \right) .$$

The calibration of the test is done by refering to observations from a uniform distribution. This approach has been introduced and studied by [11]. See also [13].

Method based on measuring concordance

Yet another approach can be found in [18]. Supposing that the regression function m is continuously differentiable, the hypothesis H_0 in (17) can be written as

$$H_0 : m'(x) \geq 0 \qquad \text{for all } x \in I . \tag{19}$$

For i.i.d. observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ from a regression model $Y = m(X) + \varepsilon$, with X and ε independent random variables, this means intuitively that for a given $x \in I$, $m'(x) \geq 0$ if X and Y are concordant for values close to x. This was the motivation in [18] to consider local versions of Kendall's tau statistic, defined as

$$U_n(x) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \text{sign}(Y_j - Y_i)\text{sign}(X_i - X_j) K_h(X_i - x) K_h(X_j - x) ,$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ is a rescaled version of a nonnegative, symmetric continuous kernel K supported in $[-1,1]$ and twice continuously differentiable in $(-1,1)$. The sequence $h = h_n > 0$ is a bandwidth sequence, converging to zero as n tends to infinity. Appropriate functionals of the quantities $U_n(x)$ lead to a class of test statistics. An example of such a functional is $\sup\{\frac{U_n(x)}{c_n(x)} : x \in \mathcal{I}\}$, where $c_n(x)$ are normalizing constants. The limiting distributing of the test statistics are established in [18].

A question that arises with this method is how to choose the bandwidth parameter.

Spline-based method

In [10] regression model (16) is considered but the errors are assumed to come from a strictly stationary sequence of real random variables with zero mean. An α-mixing condition is assumed on this strictly stationary sequence. For estimating the regression function, a least-squares regression spline estimate is used. When testing for monotonicity, the paper [10] relies on (19), or equivalently on, $\sup_x\{-m'(x)\} \leq 0$. The method uses quadratic splines, since a quadratic spline has a continuous first order derivative. The test statistic is based on minus the first derivative of the least quadratic spline estimate evaluated at all knot points t_j, and looks at the maximum of normalized versions of these estimated minus first derivatives.

Discussion and related problems

The above overview shows that there are various approaches to construct testing procedures for testing for monotonicity. Among the major approaches are:

- The critical bandwidth approach.

- Local approaches that are tracing the local structure of the curve. Among these are procedures based on running gradients, signs of differences of Y-observations and linear rank statistics based on ranks of groups of Y-observations.

- Approaches based on local association measures between X and Y, such as a local version of Kendall's tau statistic.

Apart from these there are other approaches, such as looking at the supremum of minus the first derivative of m. Other possibilities could be to base a test statistic on a measure of the distance between an unconstrained and a constrained (i.e. monotonic) estimate of the regression function. Commonly-used distance measures are L_2-distances, supremum-distances, etc.

Another interesting issue would be to obtain confidence bands for monotone regression functions. See the recent work by [12] and references therein.

Interesting related testing problems are: testing for a monotone hazard function or testing whether a regression function is convex. Some references to these related problems are provided at the end of this contribution.

Acknowledgements. The author's work was prepared under financial support from 'Projet d'Actions de Recherche Concertées', No. 98/03–217 of the Belgian government and from the IAP research network nr P5/24 of the Belgian Federal Office for Scientific, Technical and Cultural Affairs.

Irène Gijbels

Institut de Statistique
Université catholique de Louvain
Voie du Roman Pays 20
B-1348 Louvain-la-Neuve, Belgium

gijbels@stat.ucl.ac.be

References

[1] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, New York.

[2] de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.

[3] Bowman, A.W., Jones, M.C. and Gijbels, I. (1998). Testing monotonicity of regression. *Jour. Computat. Graph. Statist.*, **7**, 489–500.

[4] Brunk, H.D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, **26**, 607–616.

[5] Brunk, H.D. (1958). On the estimation of parameters restricted by inequalities. *Ann. Math. Statist.*, **29**, 437–454.

[6] Brunk, H.D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference*, Puri, M.L. (ed), pp. 177–195. Cambridge University Press, Cambridge, UK.

[7] Chaudhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves. *Jour. Amer. Statist. Assoc.*, **94**, 807–823.

[8] Cressie, N.A.C. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests.*Jour. Royal Statist. Soc., Ser. B*, **46**, 440–464.

[9] Davies, P.L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.*, **29**, 1–65.

[10] Diack, C. A. T. (2000). Testing the shape of a regression curve. *Unpublished manuscript.*

[11] Dümbgen, L. (2002). Application of local rank tests to nonparametric regression. *Nonpar. Statist.*, **14**, 511–537.

[12] Dümbgen, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli*, **9**, 423–451.

[13] Dümbgen, L. and Spokoiny, V.G. (2001). Multiscale testing of qualitative hypothesis. *Ann. Statist.*, **29**, 124–152.

[14] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

[15] Fisher, N.I., Mammen, E. and Marron, J.S. (1994). Testing for multimodality. *Comp. Statist. Data Anal.* **18**, 499–512.

[16] Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics, **757**, 23–68. Springer-Verlag, New York.

[17] Gijbels, I., Hall, P., Jones, M.C. and Koch, I. (2000). Tests for monotonicity of a regression mean with guaranteed level. *Biometrika*, **87**, 663–673.

[18] Ghosal, S., Sen, A. and van der Vaart, A. (2000). Testing monotonicity of regression. *Ann. Statist.*, **28**, 1054–1082.

[19] Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* **Vol II** (L. M. Le Cam and R. A. Olshen, eds), 539–555.

[20] Hall, P., and Heckman, N. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.*, **28**, 20–39.

[21] Hall, P., and Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Ann. Statist.*, **29**, 624–647.

[22] Hanson, D.L., Pledger, G. and Wirght , F.T. (1973). On consistency in monotonic regression. *Ann. Statist.*, **1**, 401–421.

[23] Mammen, E. (1991a). Estimating a smooth monotone regression function. *Ann. Statist.*, **19**, 724–740.

[24] Mammen, E., Marron, J.S. and Fisher, N.I. (1992). Some asymptotics for multimodality tests based on kernel estimates. *Probab. Theory Rel. Fields*, **91**, 115–32.

[25] Mammen, E., Marron, J.S., Turlach, B.A. and Wand, M.P. (2001). A general framework for constrained smoothing. *Statist. Science*, **16**, 232–248.

[26] Mukerjee, H. (1988). Monotone nonparametric regression. *Ann. Statist.*, **16**, 741–750.

[27] Nadaraya, E.A. (1964). On estimating regression. *Theory Prob. Appl.*, **9**, 141–142.

[28] Ramsay, J.O. (1988). Monotone regression splines in action. *Statist. Science*, **3**, 425–461.

[29] Ramsay, J.O. (1998). Estimating smooth monotone functions. *Jour. Royal Statist. Soc., Ser. B*, **60**, 365–375.

[30] Ruppert, D., Sheather, S.J. & Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *Jour. Amer. Statist. Assoc.*, **90**, 1257–1270.

[31] Schlee, W. (1982) Nonparametric tests of the monotony and convexity of regression. In *Nonparametric Statistical Inference, vol. II* (eds: B.V. Gnedenko, M.L. Puri & I. Vincze), Amsterdam: North–Holland, 823–36.

[32] Silverman, B.W. (1981). Using kernel density estimates to investigate multimodality. *Jour. Royal Statist. Soc., Ser. B*, **43**, 97–99.

[33] Silverman, B.W. (1983). Some properties of a test for multimodality based on kernel density estimates. In *Probability, Statistics and Analysis* (eds: J.F.C. Kingman & G.E.H. Reuter), Cambridge: Cambridge University Press, 248–259.

[34] Watson, G.S. (1964). Smooth regression analysis. *Sankhyā* Ser. A, **26**, 359–372.

Further Reading

Delecroix, M., Simioni, M. and Thomas-Agnan, C. (1996). Functional estimation under shape constraints. *Nonparam. Statist.*, **6**, 69–89.

Delecroix, M. and Thomas-Agnan, C. (2000). Spline and kernel smoothing under shape restrictions. In *Smoothing and Regression: Approaches, Computation and Application* (M. Schimek ed.), 109–134. Wiley, New York.

Gijbels, I. and Heckman, N. (2003). Nonparametric testing for a monotone hazard function via normalized spacings. *Nonparam. Statist.*, to appear.

Grenander, U. (1956). On the theory of mortality measurement, Part II. *Skand. Akt.*, **39**, 125–153.

Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2001a). A canonical process for estimation of convex functions: the "invelope" of integrated Brownian motion $+t^4$. *Ann. Statist.*, **29**, 1620–1652.

Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2001b). Estimation of convex functions: characterizations and asymptotic theory. *Ann. Statist.*, **29**, 1653–1698.

Hall, P. and Huang, L.-S. (2002). Unimodal density estimation using kernel methods.

Statistica Sinica, **12**, 965–990.

Hall, P., Huang, L.-S., Gifford, J.A. and Gijbels, I. (2001). Nonparametric estimation of hazard rate under the constraint of monotonicity. *Jour. Computat. Graph. Statist.*, **10**, 592–614.

Hall, P. and Van Keilegom, I. (2002). Testing for monotone increasing hazard rate. *Manuscript.*

Mammen, E. (1991b). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.*, **19**, 741–759.

Mammen, E. and Thomas-Agnan, C. (1999). Smoothing splines and shape restrictions. *Scand. Jour. Statist.*, **26**, 239–252.

Related Entries: Constrained estimation, Nonparametric estimation, nonparametric testing methods.

Table 1: *An example of calculating the isotonic regression estimate.*

i	$X_{(i)}$	$Y_{(i)}$	$S_i = \sum_{j=1}^{i} Y_{(j)}$	$\widehat{m}_I(X_{(i)})$
1	1	-2	-2	-2
2	3	5	3	0.5
3	6	-4	-1	0.5
4	8	2	1	2
5	10	3	4	3

Table 2: *Weight function $A_i(x)$ for some estimators.*

NW estimator	GM estimator	LL estimator
$\dfrac{nK_h(X_i - x)}{\sum_{j=1}^{n} K_h(X_j - x)}$	$n\int_{s_{i-1}}^{s_i} K_h(u - x)du$	$\dfrac{nw_i(x)}{\sum_{j=1}^{n} w_j(x)} = \dfrac{S_{n,2} - (X_i - x)S_{n,1}}{h\left[S_{n,0}S_{n,2} - S_{n,1}^2\right]} nK_h(X_i - x)$

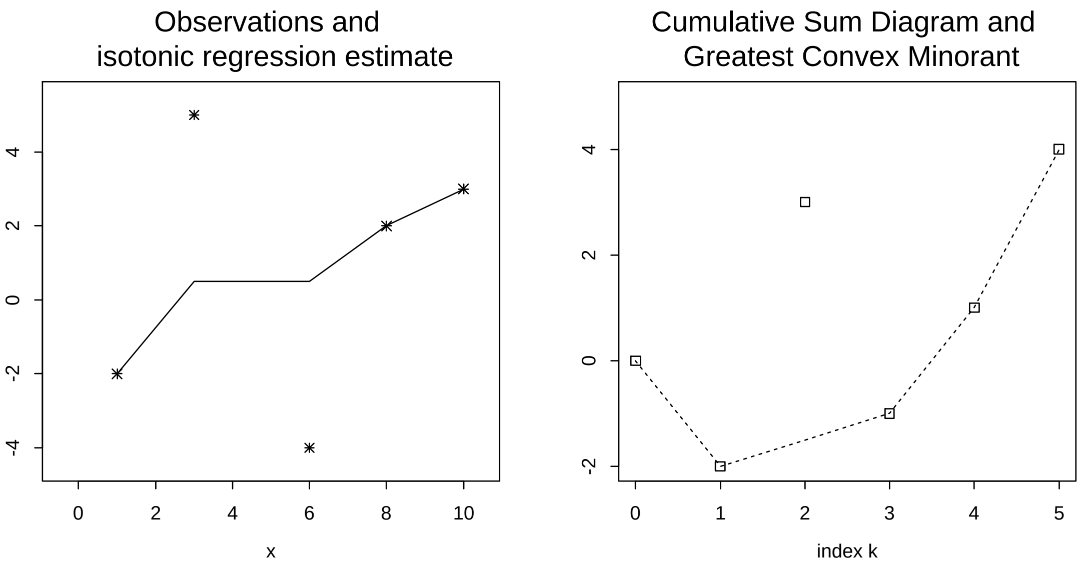

Figure 1: *Illustration of the construction of the isotonic regression estimate.*

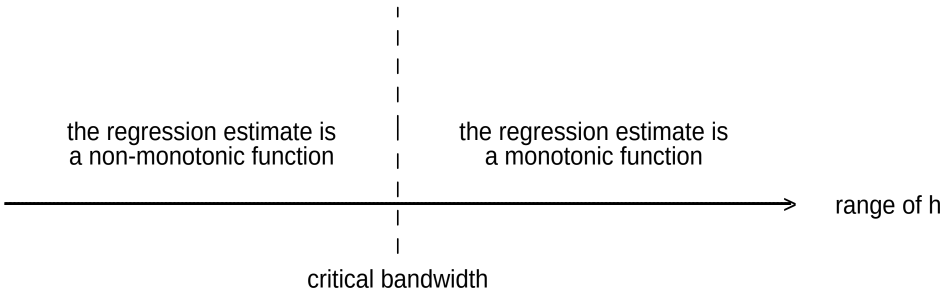

Figure 2: *The 'Monotonicity of Monotonicity' property for a regression estimate.*

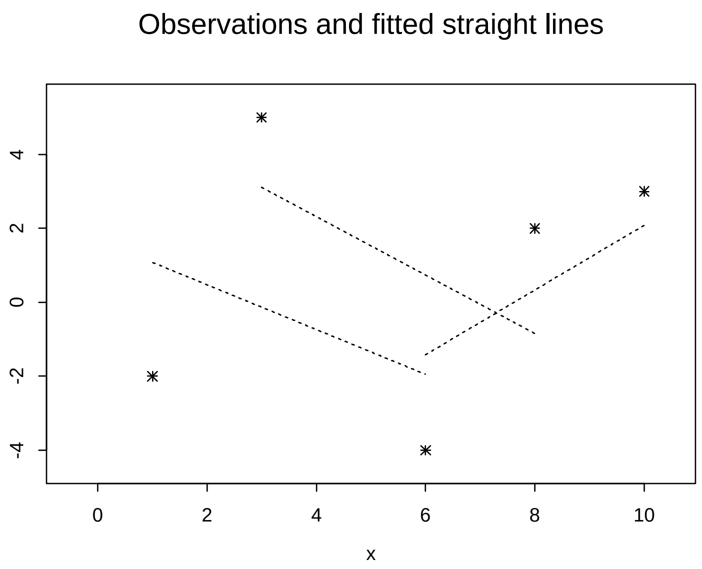

Figure 3: *Illustration of running gradients for the five observations of Figure 1, and with* $s - r = 3$.