PERIODICA POLYTECHNICA SER. MECH. ENG. VOL. 41, NO. 1, PP. 69-78 (1997)

SMOOTHING SPLINE CHARACTERIZED BY CONFIDENCE BAND FOR SOLVING REGRESSION PROBLEMS[1]

Gábor HALÁSZ and László KULLMANN

Department of Hydraulic Machines
Technical University of Budapest
H-1521 Budapest, Hungary
Fax: + 3 61 463-3091
e-mail: halasz@vizgep.bme.hu

Received: February 3, 1996

Abstract

The quality of an approximating function using measured data may be characterized by the magnitude of difference between approximating and exact functional relationship. Smoothing spline has been used as approximating function. Based on measurements, a method is presented for determining a band around the approximating spline containing the graph of the exact functional relationship between physical variables on a calculable probability level.

Keywords: measured data, smoothing spline, confidence band.

1. Introduction

Measurements in engineering practice are mostly expected to reveal deterministic relations between the considered variables and to determine an approximate functional relationship based on the recorded values of variables. In the course of measurements, variables of the deterministic relationship are affected by random effects, allocating procedures for determining the functional relationship to the scope of regression analysis (see VINCZE, 1968) for details. The approach to problems of the involved methods often fails to meet demands of engineering practice, since mathematical conditions for making useful statistic statements cannot be technically provided.

A feature common to practical problems of measurement evaluation is the existence of a deterministic relation $y(x)$, for the sake of simplicity between the two variables examined, measured values of both variables being subject to measurement errors, and the mathematical form of relation $y(x)$ being unknown. Measurements endeavour to disclose theoretical re-

[1]Delivered at the plenary meeting of the Fluid and Heat Engineering Committee of the Hungarian Academy of Sciences.

gression $y(x)$ but to that , information from measurements is insufficient. Thereby determination of approximate regression $g(x)$ has to be made up with. Function class comprising $g(x)$ is arbitrarily selected, nevertheless, any information offered by theoretical background and instrumental examination of the tested process may be reckoned with, together with reasonable and practical aspects. All these features make neither the class of functions nor the function $g(x)$ itself unambiguous. Comparison between functions taken from different function classes, established by different methods, mostly relies on a variance-type expression given by RALSTON (1969) as:

$$D^2 = \sum_{i=1}^{n} \Big[\xi_i - g(x_i)\Big]^2,$$

where ξ_i is the measured function value, and $g(x_i)$ is the corresponding approximate ordinate. The approximate regression with the lower D^2 is considered to be the better. This is the well known method of least square. A decision based on this criterion has to be made with caution, minimum of D^2 being zero, accessible e.g. by an interpolation polynomial of the due power.

NYIRI (1991) gave a generalisation of the least square method based on Whittaker's idea for nonequidistant data where both variables are affected by errors.

Here a new aspect of qualifying approximate functions will be suggested. Self intended support of comparison is theoretical regression $y(x)$, and an adequate approximation is offered by $g(x)$ with a curve 'close' to that of $y(x)$ in the range of measurement. The method to be presented suits to determine a confidence band about a given approximate regression $g(x)$ including the curve of the unknown $y(x)$ at a calculable probability. In knowledge of the band size it may be decided whether $g(x)$ is acceptable or not to solve the given engineering problem.

Producing a cubic smoothing spline proved to be an appropriate procedure to determine function $g(x)$.

2. The Smoothing Spline

A smoothing spline $g(x)$ consists of cubic parabola arcs connected continuously to the second order at joints x_i. Third derivatives are, however, different on the two sides of the joint. SPÄTH (1978) suggests that this difference r_i be proportional to the difference between spline base point ordinate ξ_i, and spline ordinate g_i:

$$r_i = p_i(\xi_i - g_i),$$

where p_i is the smoothing parameter, either different or equal at spline section joints. Base points (x_i, ξ_i), parameter values p_i and boundary specification at spline ends are given. Coefficients of cubic polynomials describing spline sections are wanted. A cubic polynomial is determined by four constants, to be defined in many different ways. Here they will be chosen as follows: the four constants are ordinates g_i; and g_{i+1}, as well as numerical values of second derivatives g''_i and g''_{i+1}' at starting and end points x_i and x_{i+1} of spline section i, respectively. Defining the spline section i over the section of length $x_{i+1} - x_i = \Delta x_i$:

$$g(x) = g_i \frac{\Delta x_i - (x - x_i)}{\Delta x_i} + g_{i+1} \frac{x - x_i}{\Delta x_i} +$$

$$+ g''_i \frac{-(x - x_i)^3 + 3(x - x_i)^2 \Delta x_i - 2(x - x_i)\Delta x_i^2}{6\Delta x_i} +$$

$$+ g''_{i+1} \frac{(x - x_i)^3 - (x - x_i)\Delta x_i^2}{6\Delta x_i} \tag{2.1}$$

is the equation of the spline section $x_i \leq x \leq x_{i+1}$.

Obviously, continuity of the first derivative over x_i is provided by

$$\frac{\Delta x_{i-1}}{6} g''_{i-1} + \frac{\Delta x_{i-1} + \Delta x_i}{3} g''_i + \frac{\Delta x_i}{6} g''_{i+1} -$$

$$- \frac{1}{\Delta x_{i-1}} g_{i-1} + \left(\frac{1}{\Delta x_{i-1}} + \frac{1}{\Delta x_i}\right) g_i - \frac{1}{\Delta x_i} g_{i+1} = 0. \tag{2.2}$$

A boundary condition may be zero spline curvature at the end point, or a given value for the first derivative, e.g. g'_1 and g'_n at the first and last spline points, respectively. The boundary equations in the equation system for spline constants are

$$\frac{\Delta x_1}{3} g''_1 + \frac{\Delta x_1}{6} g''_2 + \frac{1}{\Delta x_1} g_1 - \frac{1}{\Delta x_1} g_2 = -g'_1 \tag{2.3}$$

and

$$\frac{\Delta x_{n-1}}{6} g''_{n-1} + \frac{\Delta x_{n-1}}{3} g''_n + \frac{1}{\Delta x_{n-1}} g_{n-1} + \frac{1}{\Delta x_{n-1}} g_n = g'_n. \tag{2.4}$$

Boundary and joint conditions are united in an equation system

$$A\mathbf{g}'' + B\mathbf{g} = \mathbf{b} \tag{2.5}$$

where both A and B (B is symmetrical) are tridiagonal matrices, with coefficients taken from *Eqs.* (2.2), (2.3), and (2.4):

$$\mathbf{g}''^T = (g_1'', g_2'', \ldots, g_n''),$$
$$\mathbf{g}^T = (g_1, g_2, \ldots, g_n),$$
$$\mathbf{b}^T = (-g_1', 0, \ldots, g_n').$$

The next step will be to compute differences r_i of third derivatives at the joint. From (2. 1)

$$g'''(x) = g'''(x_i) = -\frac{1}{\Delta x_i} g_i'' + \frac{1}{\Delta x_i} g_{i+1}''$$

constant throughout section i. Thereby:

$$r_i = g'''(x_i) - g'''(x_{i-1}) = \frac{1}{\Delta x_{i-1}} g_{i-1}'' - \left(\frac{1}{\Delta x_{i-1}} + \frac{1}{\Delta x_i}\right) g_i'' + \frac{1}{\Delta x_i} g_{i+1}''$$

that is, $\mathbf{r} = -Bg''$; B is a symmetric matrix. Let D be the diagonal matrix of smoothing parameters $p_1, \ldots, p_n$ while ξ the vector of given spline base point ordinates $\xi^T = (\xi_1, \xi_2, \ldots, \xi_n)$. Now, smoothing condition may be written as:

$$-Bg'' = D(\xi - \mathbf{g})$$

or

$$-B\mathbf{g}'' + D\mathbf{g} = D\xi. \tag{2.6}$$

Solution of *Eqs.* (2.5) and (2.6) for $\mathbf{g}$ and $\mathbf{g}$" for given ξ, D, A, B, b is made in two steps. From *Eqs.* (2.6) multiplied by the inverse of D:

$$\mathbf{g} = D^{-1}D\xi + D^{-1}B\mathbf{g}'' = \xi + D^{-1}B\mathbf{g}'' \tag{2.7}$$

substituted into (2.5):

$$A\mathbf{g}'' + B\xi + BD^{-1}B\mathbf{g}'' = \mathbf{b},$$

that is

$$(A + BD^{-1}B)\mathbf{g}'' = \mathbf{b} - B\xi,$$

where $A + BD^{-1}B$ is a pentadiagonal matrix allowing a rapid determination of $\mathbf{g}$". At last, $\mathbf{g}$ is obtained from (2.7). Properly assigning subscripts to band matrix elements keeps the storage needs low, proportional to n. For zero curvature at the end point, the procedure is similar but the equation contains fewer unknowns.

3. Estimation of the Confidence Band

Introducing notations in *Fig. 1* $y(x)$ and $g(x)$ are the wanted theoretical, and the known approximate regressions, respectively. Both are assumed to be differentiable in measurement interval $[a, b]$. Random variable η is interpreted in the same interval; its realizations are values of the independent variable to be adjusted in measurements. In theoretical regression $\zeta = y(\eta)$ is assigned to the variable η hence an error-free measurement would result in a sample for random vector variable $(\eta, y(\eta))$. Measurement results for both independent and dependent variables are, however, affected, by measurement errors μ and ν, resp., of normal distribution, of zero expected value. Assume μ and ν, as well as μ and η to be mutually independent.

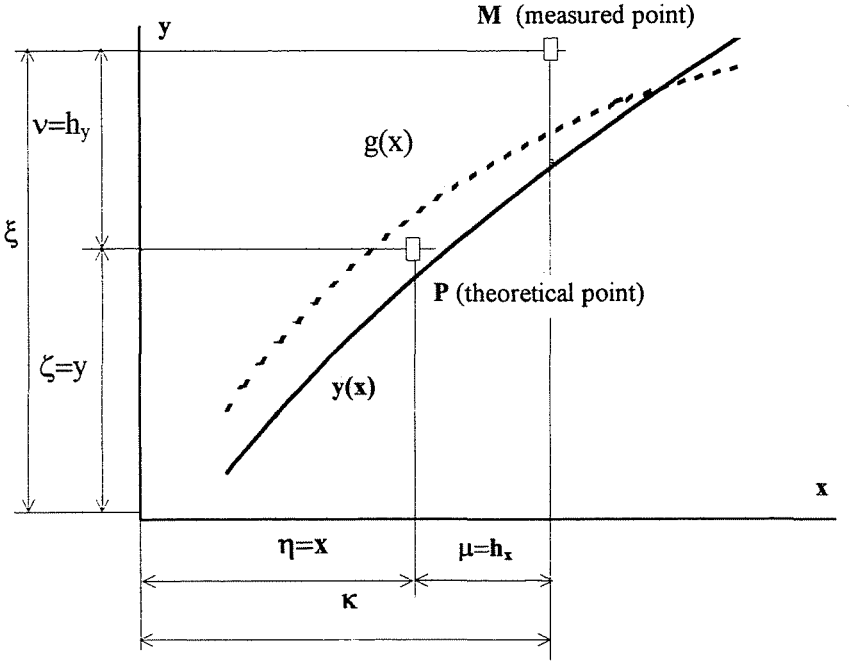

Fig. 1. Theoretical and approximate regression

Accordingly, coordinates of measurement point M in *Fig. 1* have been given by variables $\kappa = \eta + \mu$ and $\xi = \zeta + \nu$. At last, let us introduce random variable δ and difference function $s(x)$ such as :

$$\delta + g(\eta + \nu) = y(\eta) + \nu, \tag{3.1}$$

$$s(x) = y(x) - g(x). \tag{3.2}$$

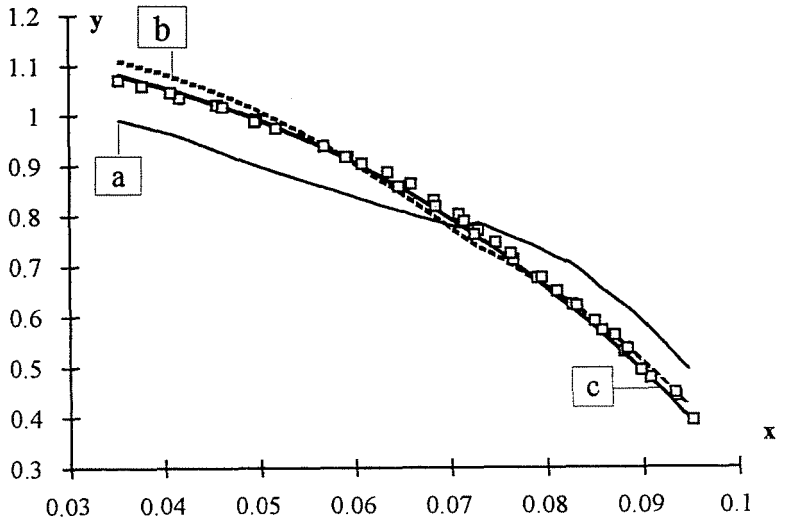

Fig. 2. Measured points and smoothing spline

Difference function $s(x)$ is assumed to be decomposable to a finite number of invertible sections in interval $[a, b]$. Expressing δ from (3.1) – and approximating $g(\eta + \mu)$ by its first order Taylor polynomial:

$$\delta \cong s(\eta) + \nu - \mu g'(\eta).$$

Let us determine the conditional expected value of δ under condition $B = \{s(\eta) = m\}$. In conformity with assumptions made for variables μ, ν and η:

$$M(\delta|B) = m. \tag{3.3}$$

Equality (3.3) points to the significance of variable δ : its conditional expected value equals ordinate m of difference function $s(x)$. Our endeavour is just to assess maximum and minimum of conditional expected value (3.3) (hence, of $s(x)$) while x proceeds along interval $[a, b]$.

For this purpose:

- we determine density function $f(w|m)$ of variable δ under condition B; $f(w|m)$ is the weighted sum of normal density functions of expected value m as has been proved by one of the authors, HALÁSZ (1986).
- we determine density function $k(m)$ of variable $s(\eta)$ according to REZA (1966).
- we write relationship between conditional density functions as done by RÉNYI (1954):

$$f(w|m) = \frac{f(w)k(w|m)}{k(m)} \tag{3.4}$$

from which, after rearranging and integration, density function $f(w)$ may be received:

$$f(w) = \int_{-\infty}^{\infty} f(w|m)k(m)dm.$$

Thereby it has been proved that density function $f(w)$ of variable δ is a compound of normal density functions $f(w|m)$ weighted by $k(m)$.

Expected value m of the component density function $f(w|m)$ is exactly the ordinate m of difference function $s(x)$. Decomposing compound $f(w)$ to its components, selecting components with maximum and minimum expected values (T_1 and T_2), gives the size of the confidence band, namely, for

$$T_1 \leq s(x) \leq T_2 \qquad (p)$$

it is

$$g(x) + T_1 \leq y(x) \leq g(x) + T_2 \qquad (p).$$

Computation of significance level p relies on equality

$$p = P(T_1 \leq s(\eta) < T_2) = \int_{T_1}^{T_2} k(m)dm. \tag{3.5}$$

Theoretically there is nothing against numerical determination of weight function $k(m)$ but it requires full decomposition of mixed density function $f(w)$, while determination of T_1 and T_2 requires only estimation of expected values of extreme components. Therefore, approximation given in the dissertation of HALÁSZ (1986) is often satisfactory:

$$p \approx \int_{T_1}^{T_2} f(w)dw \tag{3.6}$$

to be estimated according to empirical density function approximating $f(w)$.

4. Applications of the Method

To apply the computing method described above requires absorption and much computing. Its effective practical application is assisted by a user-

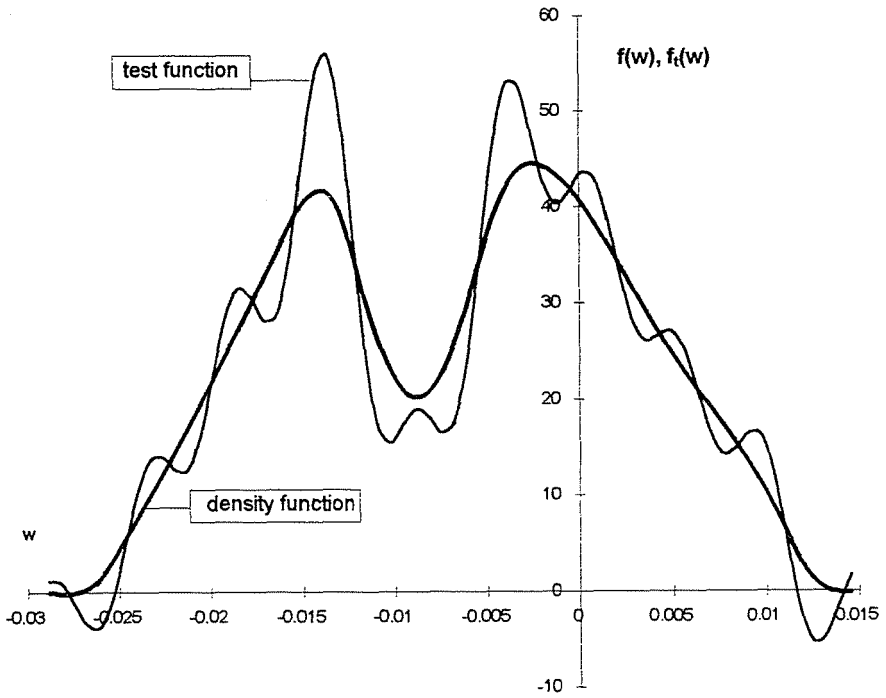

Fig. 3. Density and test function

friendly, interactive software. That is why a code has been developed at the Department of Hydraulic Machines TUB, with essential functions:

a) Input of measured data from a data file or keyboard, data storage, graphic display of data.

b) Construction of approximate regression by means of smoothing spline. Purposeful selection and adjustment of the weight function for the spline is supported by a menu system and graphic display on screen.

c) Computation of the sample for the variable δ, for establishing empirical distribution and density function; that is estimation of $f(w)$. For the next step to decompose the density function – it is insufficient to construct the empirical density function as a usual step function, but the density function has to be continuously approximated. Therefore, a smoothing function has been fitted to points of the empirical distribution function (purposefully applying again the smoothing spline procedure), and differentiated to yield an approximation of density function $f(w)$.

d) Decomposition of the compound density function and estimation of expected values of extreme components were made by the decomposition method of MEDGYESSY et al. (1968). Screen display of the test function helped direct reading of the desired expected values. At last, significance level has been determined by approximation (3.6).

e) Diagrams representing steps and output of the band estima tion method may be displayed on a plotter or line printer. Our program package produces a file readable for a graphical program system, its facilities may be utilized in constructing diagrams.

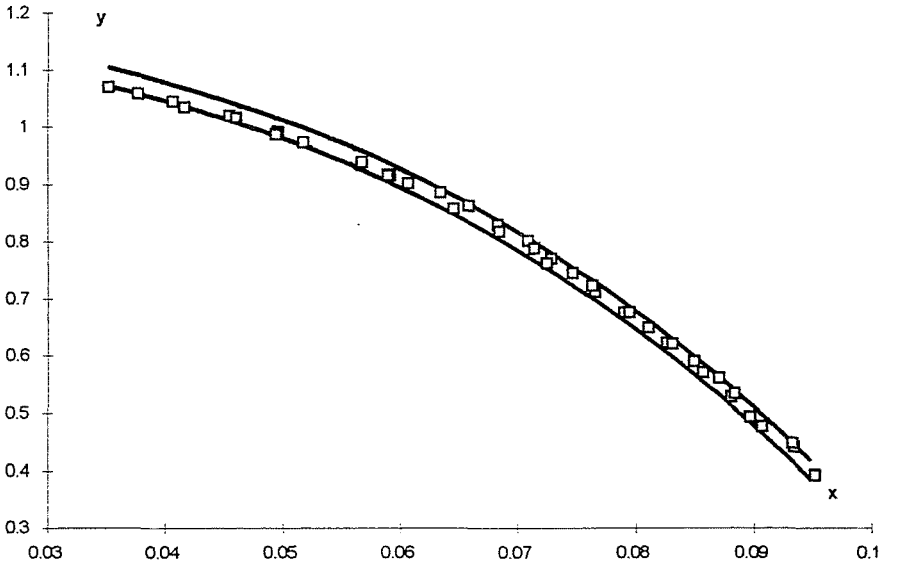

Fig. 4. Confidence band (96.4%)

The application of the program package is exemplified in *Fig. 2*, smoothing splines a, b, c together with measured points are shown. Curve a, constructed by the first approximation of spline weight function rather deviates from measurement points, but as a result of continuous refinement, smoothing spline c approximates measured results well. For this approximate function, empirical density function of variable δ and the test function have been constructed (see *Fig. 3*). *Fig. 4* shows diagram of the confidence band around the smoothing spline, as well as the estimated significance level.

References

HALÁSZ, G. (1986): Methods of Measurement Evaluation and Error Estimation for Fluid Mechanical Problems (in Hungarian). Candidate's Thesis, Budapest.

MEDGYESSY, P. – VARGA, L. (1968): Improved Method for Numerical Decomposition of compound Gaussian Functions (in Hungarian). Publications of Section III. Hungarian Academy of Sciences , No. 18, pp. 31–39.

NYIRI, A. (1991): Method for Smoothing both Variables of a Function. (in Hungarian). GÉP Vol. XLIII,: No. 7-8-9. pp. 217–220.

RALSTON, A. (1965): A First Course in Numerical Analysis. McGraw-Hill Inc. New-York.

RÉNYI, A. (1954): Probability Theory (in Hungarian), Tankönyvkiadó, Budapest.

REZA, F. M. (1966): An Introduction to Information Theory. McGraw-Hill Book Co.

SPÄTH, H. (1978): Spline-Algorithmen zur Konstruktion glatter Kurven und Flächen. Oldenburg, München-Wien.

VINCE, I. (1968): Mathematical Statistics with Engineering Applications. (in Hungarian) Műszaki Kiadó, Budapest.