Diss. ETH Nr. 8920

'Parametric' Smoothing Quality in Nonparametric Regression: Shape Control by Penalizing Inflection Points

DISSERTATION
submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY (ETH)
ZURICH
for the degree of
Doctor of Mathematics

presented by
Martin Beat Mächler
Dipl. Math. ETH
born on February 21, 1959
citizen of Thalwil, ZH

Accepted on the recommendation of
Prof. Dr. F. Hampel, examiner
Prof. Dr. H. Künsch, co-examiner

1989

Frank Hampel

Abstract

The present work of applied statistics was motivated by the lack of data smoothing procedures in data analysis which would produce curves as smooth as a graphically skilled human normally draws: The usual non-parametric regression estimators such as smoothing splines or kernel estimators are quite useful for the approximation of many smooth functions. But these approximating curves often show many little wiggles which do not appear to be necessary for a good description of the data. Each wiggle corresponds to one or two superfluous inflection points. A "perfect" smoother would produce as few inflection points as necessary for low bias.

To the given data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we want to fit a function $f(x)$ which should be as smooth as possible in the above sense. We consider the underlying model

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n\ ,$$

where the random erros ε_i are assumed to have mean zero and, in general also to be independent.

The "Maximum Likelihood" (*'m.l.'*) method is well-known to be an asymptotically optimal method of estimation under rather weak assumptions. Here, for the case of independent errors, a nonparametric m.l. estimation would be equivalent to the choice of *any interpolating* function f – which is definitely not what we aim at. In the case of independent $\varepsilon_i \sim H_i(0)$, the negative log-likelihood which is to be *minimized* can be written as $\sum_{i=1}^{n} \rho_i(y_i - f(x_i))$.

The method of *"Maximum Penalized Likelihood"* arises from restricting the m.l. method to a certain class of smooth functions f which means that they are required to have 'roughness' $R[f] \leq C$. In contrast to the smoothing splines approach where the roughness is chosen as $R[f] = \int_{x_1}^{x_n} f''^2\ dx$, we choose our criterion more carefully considering what the 'real' roughness of a curve means to the eye, namely inflection points and, to lesser degree, inflection points of higher derivatives f', etc. Or, expressed slightly differently, *change* of curvature which we approximate by f'''/f''. In this approach these inflections points 'of higher order' are taken into account explicitly only by the generalization "$\nu > 2$" (chapt. 3).

The Maximum *Penalized* Likelihood criterion given by

$$\min_{f \in C^m[x_1, x_n]} \quad \sum_{i=1}^{n} \rho(y_i - f(x_i)) \quad + \quad \lambda \int_{x_1}^{x_n} \left(\frac{f'''(t)}{f''(t)} \right)^2 dt,$$

favours functions f with low mean "standardized *change* of curvature"
$\kappa'/\kappa = f'''/f'' - 3f'f''/(1 + f'^2) \approx f'''/f''$, where $\kappa(x)$ is the curvature of f at x.

If f has the inflection points $w_1, w_2, .., w_{n_w}$ then the criterion's integrand can be represented as $(f'''/f'')(x) = \sum_{j=1}^{n_w} \frac{1}{x-w_j} + h_f'(x)$, where the function h_f' is f'''/f'', diminished by the "poles", and is therefore continuous.

The penalty term contains n_w times infinity which leads to disjoint classes of functions, each with a fixed number of inflection points. In each class, the penalty, modified to $\int (h_f'(t))^2 \, dt$, still penalizes the change of curvature and prevents more than n_w inflection points.

It is proved that this variational problem has a solution for any positive smoothing parameter, under very weak assumptions on ρ.

In order to solve the above variational problem and a quite interesting generalization of it, the corresponding Euler-Lagrange differential equation is considered. It gives rise to a non-standard boundary condition problem where some 'boundary' conditions arise at each data point x_i. The differential equation problem is numerically quite delicate, i.e., it is badly conditioned which means a certain instability on small changes of the initial conditions. Therefore, a generalized "multiple shooting" method is developed. It leads to a Newton method to find a high-dimensional zero, and uses an even larger system of differential equations to provide the derivatives for Newton.

Finally, a minimization to choose the inflection points w_j is needed, and some considerations are necessary, how to look for a minimum in the space of the w_j.

To initially estimate the inflection points and provide a decent starting value for the above Newton method, we investigate a "pre-smoothing" procedure. Optimal discrete weight estimators for derivatives are developed and are seen to be equivalent to local polynomial regression which is done by orthogonal polynomials. A generalized 'Horner' algorithm is used for efficient and accurate evaluation of the function and derivatives in this polynomial basis.

Zusammenfassung

Diese Arbeit im Umfeld der angewandten Statistik entstand, weil in der Datenanalyse Glättungsverfahren fehlen, die ebenso glatte Kurven liefern würden, wie sie ein zeichen-technisch geschulter Mensch von Hand normalerweise zeichnet: Die üblichen nicht-parametrischen Regressionsverfahren wie glättende Splines oder Kern-Schätzer sind ganz nützlich für die Approximation vieler glatter Funktionen. Aber diese glättenden Kurven zeigen häufig viele kleine Wellen, die überflüssig scheinen für eine gute Datenbeschreibung. Jede dieser Wellen entspricht einem oder zwei unnötigen Wendepunkten. Ein perfekter Glätter wird (unter anderem) so wenig Wendepunkte wie (für kleinen Bias) nötig entstehen lassen.

Den gegebenen Datenpunkten $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ wollen wir eine Funktion $f(x)$ anpassen, die im obigen Sinne so glatt wie möglich sein soll. Wir betrachten das zugrunde liegende Modell

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n \ ,$$

wobei wir annehmen, dass die zufälligen Fehler ε_i Erwartungswert Null haben und unabhängig sind.

Die "Maximum Likelihood" (*'ML'*) Methode ist allgemein bekannt als asymptotisch optimale Schätzung unter ziemlich schwachen Voraussetzungen. Hier, im Falle von unabhängigen Fehlern, wäre eine nicht-parametrische ML-Schätzung äquivalent zur Wahl *irgendeiner interpolierenden* Funktion f – was wir ganz sicher nicht wollen. Falls $\varepsilon_i \sim H_i(0)$ (unabhängig), kann die negative Log-Likelihood, die *minimiert* werden muss, geschrieben werden als $\sum_{i=1}^{n} \rho_i(y_i - f(x_i))$.

Die Methode der *"Maximum Penalized Likelihood"* entsteht durch Beschränkung der ML-Methode auf eine gewisse Klasse von glatten Funktionen f, d.h. dass diese die "Rauhheit" ('roughness') $R[f] \leq C$ haben müssen. Im Gegensatz zum auf die glättenden Splines führenden Ansatz, bei dem die Rauhheit als $R[f] = \int_{x_1}^{x_n} f''^2 \, dx$ gemessen wird, achten wir bei der Wahl unseres Kriterium mehr darauf, was für das Auge Rauhheit wirklich bedeutet, nämlich Wendepunkte und — in kleinerem Masse — Wendepunkte höherer Ableitungen. Oder, mit etwas anderer Betonung, Krümmungs*änderung*, die wir approximieren durch f'''/f''. Die erwähnten Wendepunkte höherer Ordnung werden bei unserem Ansatz nur dann direkt berücksichtigt, wenn die Verallgemeinerung $\nu > 2$ betrachtet wird (siehe Kapitel 3).

Das "Maximum *Penalized* Likelihood" Kriterium, welches gegeben ist durch

$$\min_{f \in C^m[x_1, x_n]} \quad \sum_{i=1}^{n} \rho(y_i - f(x_i)) \quad + \quad \lambda \int_{x_1}^{x_n} \left(\frac{f'''(t)}{f''(t)} \right)^2 dt,$$

begünstigt Funktionen f mit kleiner "relativer Krümmungs*änderung*" $\kappa'/\kappa = f'''/f'' - 3f'f''/(1+f'^2) \approx f'''/f''$, wobei $\kappa(x)$ die Krümmung von f an der Stelle x bezeichnet.

Wenn wir für f die Wendepunkte $w_1, w_2, .., w_{n_w}$ annehmen, kann der Integrand des Kriteriums dargestellt werden als $(f'''/f'')(x) = \sum_{j=1}^{n_w} \frac{1}{x-w_j} + h_f'(x)$, wobei die Funktion h_f' gleich 'f'''/f'' minus die Pole', und damit stetig ist. Der Bestrafungsterm ('penalty') enthält dann n_w mal unendlich. Dies führt zu disjunkten Funktionen-Klassen, wobei in jeder Klasse die Anzahl Wendepunkte fest ist, der modifizierte "Penalty", $\int (h_f'(t))^2\, dt$, immer noch Krümmungsänderungen bestraft und verhindert, dass mehr als n_w Wendepunkte entstehen.

Es wird bewiesen, dass eine Lösung dieses Variationsproblems existiert für alle positiven Werte des Glättungsparameters, unter sehr schwachen Voraussetzungen an ρ.

Um dieses Variationsproblem und eine interessante Verallgemeinerung davon zu lösen, betrachten wir die zugehörige Euler-Lagrange Differerentialgleichung. Dies führt auf ein nicht-standard Randwert-Problem, bei dem gewisse Randbedingungen bei jedem Datenpunkt x_i auftreten. Dieses Randwertproblem ist numerisch ziemlich heikel, d.h. es ist schlecht konditioniert. Dies bedeutet eine Instabilität gegenüber kleinen Änderungen der Anfangsbedingungen. Deshalb wurde eine verallgemeinerte "Multiple Shooting" - Methode entwickelt. Sie enthält einen Newton-Algorithmus, um eine hochdimensionale Nullstelle zu finden. Dabei wird für die Ableitungsmatrix ein noch grösseres System von Differentialgleichungen benötigt.

Schliesslich muss über die Wendepunkte w_j minimiert werden, und es sind gewisse Überlegungen angebracht, in welcher Richtung nach einem Minimum im Raum der w_j gesucht werden soll.

Um eine erste Schätzung der Wendepunkte und einen guten Start'wert' für die obige Newton-Methode zu erhalten, untersuchen wir einen "Vor-Glätter". Wir entwickeln optimale Schätzer für Ableitungen auf der Basis von diskreten Gewichten. Wir zeigen, dass diese äquivalent sind zu lokaler Regression mit Polynomen, wobei *orthogonale* Polynome verwendet werden sollen. Wir beschreiben eine Verallgemeinerung des Horner-Verfahrens zur effizienten und numerisch stabilen Berechnung orthogonaler Polynome.