CHOOSING A SMOOTHING PARAMETER FOR A CURVE FITTING BY MINIMIZING THE EXPECTED PREDICTION ERROR

by
Cristian Marinoiu

Abstrac.*The value of the smoothing parameter for a curve fitting can be chosen by minimizing the expected prediction error. In this paper we describe the methods to obtain this value and provide the results obtained in a computer application.*

Introduction

Let X and Y be two random variables and (x_i, y_i) $i=1, 2, \ldots, n$ a sample available from the unknown joint distribution of these variables.

We assume that y_i and x_i are related by the general regression model

$$y_i=f(x_i)+\varepsilon_i \quad i=1, 2, \ldots, n, \tag{1}$$

where:

$f(\cdot)$ is an unknown regression function,

ε_i, $i=1, 2, \ldots, n$ are errors with zero mean, $\operatorname{cov}(\varepsilon_i, \varepsilon_j)=0$ for $i \neq j$ and $cov(\varepsilon_i, \varepsilon_i)=\sigma^2$, the unknown common variance of the errors.

There are two important classes of models: the parametric regression models and the nonparametric regression models.

In the parametric regression model the form of the function f is assumed to be known, except for a finite number of many unknown parameters. By contrast, in the nonparametric regression model there are not assumptions about the shape of the function f; the f function must only satisfy some general properties related, for example, by differentiability or integrability.

The smoothing spline estimator of the regression function appears in the context of the nonparametric regression model. This estimator is obtained by imposing in the model two natural requirements: the estimator of f needs to be close to the data and at the same time smooth enough.

The first requirement is equivalent to minimizing the standard measure of goodness-of-fit to the data, namely

$$\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 .$$

Assuming a convenient properties for function f, a natural measure of smoothness is

$$\int \left(f(x)^{(m)}\right)^2 dx ,$$

where $f(x)^{(m)}$ is the *m* order derivative. Therefore, an optimal estimator that satisfies both requirements could be [1]:

$$(2) \qquad \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 + \lambda \int \left(f(x)^{(m)}\right)^2 dx , \quad \lambda > 0$$

A large value of the smoothing parameter λ leads to a smooth curve; a small value of λ favors a rough curve that follows the data points closely.

The estimator $\hat{f}$ obtained by minimizing the relation (2) is called *the smoothing spline estimator* of the regression function.

Under convenient hypotheses, for fixed λ, $(0< \lambda <\infty)$ there is a unique minimizer of (2) (see Theorem 5.3, [1]) that is a natural spline function.

Observation. *The cubic spline* is a special case of piecewise – polynomial function with continuous first and second derivatives at knots. *The natural cubic splines* are important because the discontinuity in the knots x_i, $i=1, 2, \ldots, n$ are invisible to the human eye [2]. They are obtained as a particular case, by setting $m=2$. The curve shown in the last section (Figure 1) is a natural cubic spline.

The selection of the smoothing parameter

There are several methods used to choose the smoothing parameter λ.

A simple way is to consider random values for λ and see the resulted graph of the function $\hat{f}_\lambda$. One selects λ providing a convenient fit to the data and at the same time a satisfactory smoothness.

Another natural way is to select λ optimal, minimizing the *expected prediction error* [3]

$$pse(\lambda) = E\left(y' - f_\lambda(x')\right)^2 ,$$

where (x', y') refers to new data drawn from the distribution *F*. Since additional data are not usually available, an estimator $\hat{pse}(\lambda)$ of $pse(\lambda)$ will be used instead of $pse(\lambda)$. The cross-validation (CV) and the generalized croos-validation (GCV) are common means to estimate $pse(\lambda)$. Particularly, the leave-one-out cross validation is convenient in computing.

Setting λ to the arbitrary value, an estimator of $pse(\lambda)$ can be obtained in a standard manner by using the bootstrap method [3], as follows:

Step 1. Read the sample data $z_i = (x_i, y_i)$, $i = 1, 2, ..., n$.

Step 2. Generate B (B integer positive) bootstrap samples $z^{*1}, z^{*2}, ..., z^{*b}, ..., z^{*B}$, where $z^{*b} = \left(z_1^*, z_2^*, ..., z_n^*\right)^b$, $b = 1, 2, ..., B$, $z_i^* = \left(x_i^*, y_i^*\right)$, $i = 1, 2, ..., n$. The data $z_1^*, z_2^*, ..., z_n^*$ are obtained by drawing with replacement from the population $z_1, z_2, ..., z_n$.

Step 3. For each bootstrap sample z^{*b} calculate:

- the corresponding cubic spline $\hat{f}_\lambda^{*b}$
- $pse^{*b}(\lambda) = \sum_{i=1}^{n} \left(y_i - f_\lambda^{*b}(x_i)\right)^2 / n$

Step 4. Obtain an estimator of $pse(\lambda)$ as $pse(\lambda) = \left(\sum_{b=1}^{B} pse^{*b}(\lambda)\right) / B$.

STOP.

In order to obtain λ for witch $pse(\lambda)$ is minimum, the following adequate step must be added:

Step 5. Calculate $\hat{pse}(\lambda)$ over different values $\lambda_i, \lambda_i = 1, 2, ..., p$, p arbitrarily chosen. The desired value of λ is λ_j, where $pse(\lambda_j) = \min_{i=1,n} \hat{pse}(\lambda_i)$.

Observation. Setting $\lambda = \frac{(1-q)}{q}$, $0 < q < 1$ we can calculate $\hat{pse}(\lambda)$ over a grid values of q, $0 < q < 1$.

An application in petrochemistry

A set of data available from [4] is listed below (Table 1). They represent the observed values for gas productivity (η_B) and feedstock flow *(Gmp)* during 15 days in the cracking process.

Gas productivity	**Feedstock flow**

(η_B)	*(Gmp)*
183.40	52.3
183.10	52.8
184.30	52.8
189.70	51.4
183.80	52.4
182.40	52.1
182.60	52.8
183.70	52.2
182.80	52.8
182.71	51.8
187.90	52.3
191.10	52.0
184.60	53.0
182.20	51.3
182.72	52.7

Table 1 – The sample data for gas productivity (η_B) and feedstock flow *(Gmp)*.

Our aim is to find an optimal value of the smoothing parameter q in order to obtain a convenient cubic spline function

$$\hat{f}_{\lambda} \quad \left(\lambda = \frac{(1-q)}{q}\right)$$

from the model:

$$(\eta_B)_i = f_{\lambda}(Gmp_i) + \varepsilon_i,\ i = 1,2,...,15$$

To achieve this, the algorithm presented in the above section has been implemented in *Matlab 5.3*. This version offers a satisfactory function (called *csaps*) to obtain the natural cubic spline function $\hat{f}_q$ for a provided value q, $0 < q < 1$. This was an essential motivation to choose *Matlab* medium to implement the algorithm.

The optimal value of λ for the data from Table 1 was obtained as $\lambda = 0.11$ for *B=100*. The Figure 1 shows the obtained graph for the data listed in Table 1, at this value of λ.

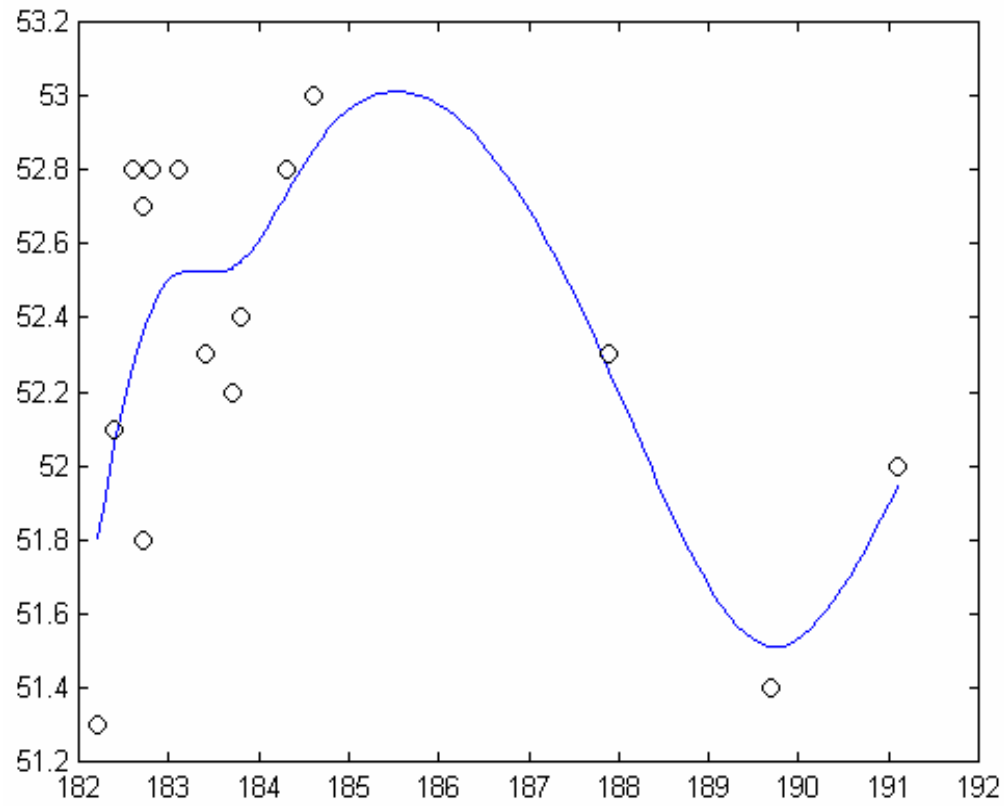

Figure 1 – The optimal natural cubic spline based on the data listed in Table 1.

References

1. Eubank, R.L. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York and Basel, 1988.
2. Hastie Tr., Tibshirani R., Friedman J., *The elements of statistical learning*, Springer-Verlag, New York, 2001.
3. Efron B. Tibshirani R. *An introduction to the bootstrap*, Chapman & Hall, New York, 1993.

4. Patrascioiu Cr., Marinoiu Cr., *Solutii numerice pentru modelarea statistica a procesului de cracare catalitica*, Revista Informatica Economica, Vol. III, nr.10, Inforec, ASE, Bucuresti, 1999.

Assoc. prof. Cristian Marinoiu, Petroleum and Gas University from Ploiesti