THE CROSS-VALIDATION METHOD IN THE SMOOTHING SPLINE REGRESSION

by
Nicoleta Breaz

Abstract. One of the goals, in the context of nonparametric regression by smoothing spline functions, is to choose the optimal value for the smoothing parameter. In this paper, we deal with the cross validation method(CV), as a performance criteria for smoothing parameter selection. First, we implement a *CV*-based algorithm, in Matlab 6.5 medium and we apply it on a test function, in order to emphase the quality of the fitting by the *CV*-smoothing spline function. Then, we fit some real data with this kind of function.

1.Introduction

We consider the observational regression model,

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

with $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)' \sim N(0, \sigma^2 I)$ and the data, y_i, having the weights $w_i, w_i > 0$. If the plot of data presents some classical trend, as polynomial for example, we choose the parametric regression technique, else, we choose the nonparametric regression technique.

A parametric model is based on some assumed form of the regression function which depends on a finite number of many unknown parameters(see for example, the polynomial regression). In this case, the goal is to estimate these parameters.

By contrast, a nonparametric model doesn't make assumptions about the shape of the estimator but about the "quality" of the estimator. This quality refers to some general properties as smoothness, for example.

Moreover, if the data are noisy, it is more appropriate to find an estimator that is not very close to data but is sufficiently smooth (see[6]).

Such estimator will minimize, for example, the following expression:

$$\sum_{i=1}^{n} w_i (y_i - f(x_i))^2 + \lambda \int_a^b \left(f^{(m)}(x)\right)^2, \quad \lambda \geq 0, [x_{\min}, x_{\max}] \subseteq [a,b]. \qquad (1)$$

First part of this expression represents the goodness-of-fit to the data and the second part represents the smoothness of the estimator. The parameter λ, called smoothing parameter, controls the tradeoff between the closeness to the data and the

smoothness. If $\lambda = 0$, we obtain the interpolant to the data and if $\lambda \to \infty$, we obtain the straight line least squares approximation. Obviously a large value of λ leads to a smooth curve but not so close to data and a small value of λ leads to a rough curve that follows the data closely .

The expression (1) is often known, as penalized least squares criteria. If we search the solution of this variational problem in some appropriate space, we obtain the estimator called smoothing spline. The name "spline" comes from fact that the estimator is practically, a natural polynomial spline function, of $2m-1$ degree (see [3]).

A case of interest is the particular case, $m = 2$, when we obtain as an estimator, the natural cubic spline function. These functions are piecewise-cubic polynomial function, with continuous first and second derivatives, at the break points.

Although the smoothing spline appears in the context of the nonparametric regression, however, the estimator depends on a parameter λ, namely, the smoothing spline parameter. There are known several methods to select the smoothing parameters and among these, is cross validation method (CV).

2.The CV-smoothing parameter selection method

When we try to choose the optimal model to data we can use some performance criteria as a testing tool (see [1]). One of these performance criteria is based on a natural way to select that fitting and implicitly, that λ, which minimizes the expected prediction error,

$$PSE(\lambda) = E(y' - f_\lambda(x'))^2,$$

where x', y' are new data.

Since additional data are not usually available, an estimator of $PSE(\lambda)$ will be used instead of $PSE(\lambda)$. According to [1], one of such estimator is the (leaving-out-one) cross-validation function, given by

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i \left(y_i - f_\lambda^{(-i)}(x_i)\right)^2 ,$$

where $f_\lambda^{(-i)}$ is the smoothing spline estimator, fitted from all data, less the *i*-th data. The (leaving-out-one) cross validation method uses *n* learning samples, everyone with $n-1$ data, to obtain the estimators $f_\lambda^{(-i)}$, $i = \overline{1,n}$ and *n* test samples, with one data, in

order to validate the models. Since $CV(\lambda)$ is an estimator for $PSE(\lambda)$, a value of λ, that minimizes $CV(\lambda)$, represents an optimal choice for λ.

3.Numerical experiments

In order to show how the CV method works, we implement in Matlab 6.5 medium, the following algorithm, based on CV:

CV-Algorithm

Step 1. Read the sample data $(x_i, y_i), i = \overline{1,n}$ and if is necessary, order and reweight the data, in respect with data sites, x_i. In stead of n_1 data, $(x_{n_1}, y_{n_{1i}})$, with weights $w_{n_{1i}}$, $i = \overline{1,n_1}$, we will have just one data, (x_{n_1}, y_{n_1}), with

$$y_{n_1} = \frac{\sum_{i=1}^{n_1} w_{n_{1i}} y_{n_{1i}}}{\sum_{i=1}^{n_1} w_{n_{1i}}}$$

and the weight,

$$w_{n_1} = \sum_{i=1}^{n_1} w_{n_{1i}} .$$

After this step, the data sites, x_i, must be strictly increasing and having the tail $n' \le n$.

Step 2. For each *i*, $i = \overline{1,n'}$, determine the cubic smoothing spline, $f_\lambda^{(-i)}$, based on leaving-out-one resampling method.

Step 3. Calculate the value of the function, $CV(\lambda)$. *STOP.*

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

Step 4. Calculate $CV(\lambda)$, for different values of λ.

The appropriate value of λ is λ_{CV}, with

$$CV(\lambda_{CV}) = \min_{\lambda} CV(\lambda).$$

If we set

$$\lambda = \frac{1-q}{q}, 0 < q < 1,$$

we can search λ, by searching q, over a grid on $[0,1]$. In this paper, we use a regular grid, with 1000 points.

Obviously, a large value for q leads to a small value for λ and consequently to a rough curve, closely to data points. By contrast, small values for q give large values for λ and smooth, but not closely to data, curves.

In that following, we will consider the test function,

$$f(x) = 3^x - 2^x + e^{-5x} + e^{-20\left(x-\frac{1}{2}\right)^2}$$

and the noisy data (x_i, y_i), with

$$x_i = \frac{i}{n},\ y_i = f(x_i) + \varepsilon_i,\ \varepsilon_i \in N(0;0,1),\ i = \overline{1,100}.$$

Here ε_i, $i = \overline{1,100}$, come from a random number generator simulating independently and identically distributed, random variables.

By running the algorithm presented above, for 100 replicates, we obtain an average value $q_{CV} = 0,9996$, that leads to $\lambda_{CV} = 4 \cdot 10^{-4}$.

The following three figures represent the plot of the data, the test function and the smoothing spline function, obtained for three different values of λ: a too large one, 0,0526 (spline 1), a too small one, 10^{-6} (spline 2), and the CV-value.

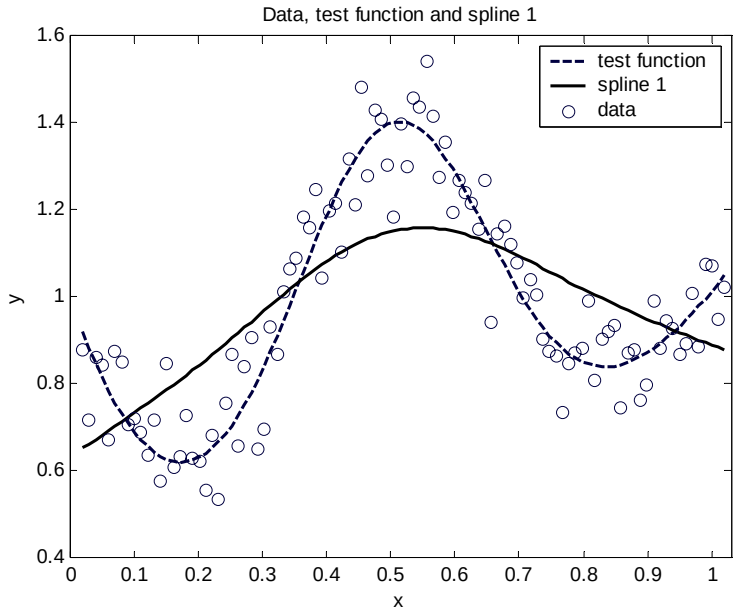
Data, test function and spline 1
test function
spline 1
data
y
x
1.6
1.4
1.2
1
0.8
0.6
0.4
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Fig. 1

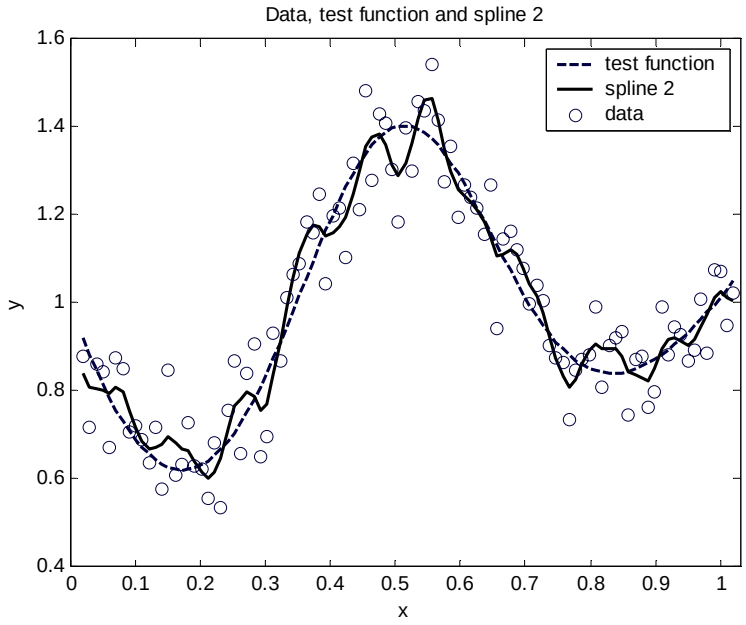
Data, test function and spline 2
test function
spline 2
data
y
x
1.6
1.4
1.2
1
0.8
0.6
0.4
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Fig. 2

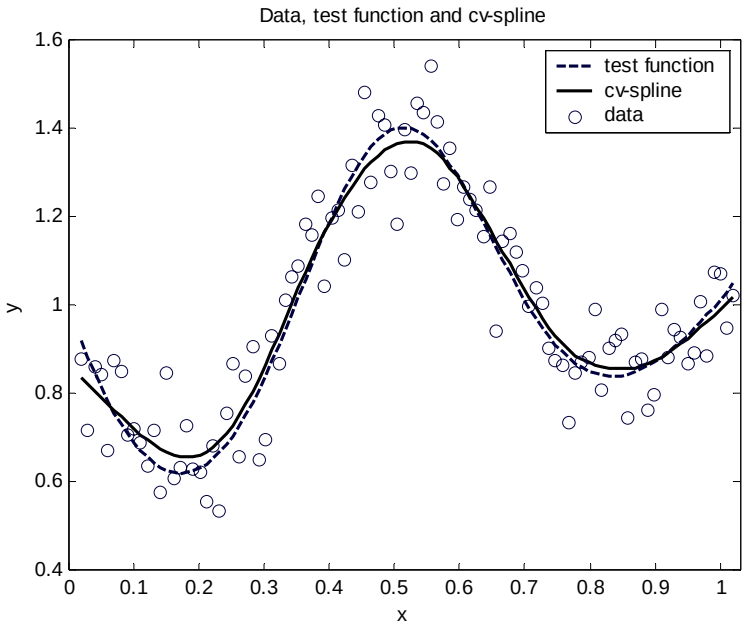

Fig. 3

We observe that, for a too large value of λ, the estimated curve is not so close to data but is more smooth than real curve and for a too small value of λ, the estimated curve is not so smooth but is closer to data than the real curve is.

By contrast, the CV-value of λ gives us an optimal estimator. In this case, the estimated curve is more like the real one, not too close to data and not too smooth.

4.An application to real data

We will consider the same data as in [2], namely the observed values for the gas productivity, x_i and the feedstock flow, y_i, during 15 days, in the cracking process.

In that paper, the cubic smoothing spline was obtained also from CV method, but using the bootstrap method, for resampling. The optimal value for λ was 0,11.

For the same data, we apply our algorihtm presented here and we obtain the optimal value for λ, $\lambda_{CV} = 0{,}0102$.

In the following figure, we plot the data, our CV-estimated smoothing spline(cv-spline 1) and the 0,11-estimated smoothing spline, from [2](cv-spline 2).

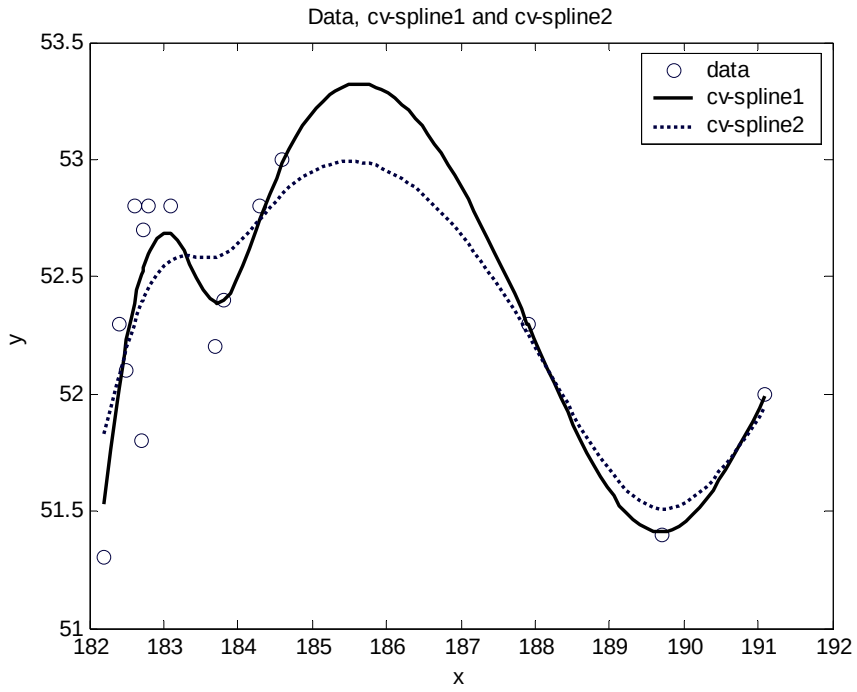

Fig. 4

It can be observed that our curve is more closely to data and 0,11-curve is more smooth.

But obviuosly, at this point, we cannot say that the estimator presented here is better than estimator from [2], but just that the estimator presented here is more close to data. For choosing one method, inspite the other, we must know more about the real process.

For example, if one knows that he is interested more in goodness of fit than in smoothness, he will choose the estimator with $\lambda = 0{,}0102$.

As a conclusion, if we know something prior about the "quantum" of the goodness of fit, or about the "quantum" of the smoothness, we can impose that the related term from (1) does not exceed an assumed tolerance.

References

1.Eubank R. L. - Nonparametric Regression and Spline Smoothing-Second Edition, Marcel Dekker, Inc., New York , Basel, 1999

2.Marinoiu C.-Choosing a smoothing parameter for a curve fitting by minimizing the expected prediction error, AUA, No.5/2003

3.Micula G.-Funcţii spline şi aplicaţii, Ed. Tehnică, Bucureşti, 1978
4.Micula G., Micula S.-Handbook of Splines, Kluwer Academic Publishers, Dordrecht/Boston/London, 1999
5.Stapleton J.H.- Linear Statistical Models. A Willey-Interscience Publications, Series in Probability and Statistics, New York, 1995
6.Wahba G.-Spline Models for Observational Data, SIAM Publications, Philadelphia, 1990

Author:
Nicoleta Breaz, „1 Decembrie 1918" University of Alba Iulia, Romania,
nbreaz@uab.ro