NUMERICAL EXPERIMENTS WITH LEAST SQUARES SPLINE ESTIMATORS IN A PARAMETRIC REGRESSION MODEL

by
Nicoleta Breaz

Abstract. In this paper, we discuss two aspects regarding the least squares estimators in a parametric regression model. First, we make a numerical experiment that emphasizes the goodness of the least squares spline fit with respect to the polynomial fit. For the second goal of the paper, we implement a CV-based knots selection algorithm in Matlab 6.5 environment and we apply it on some test function, in order to analyse how the CV method works, in the knots selection for the least squares spline estimators.

1. Introduction

We consider the regression model,

$$Y = f(X) + \varepsilon ,$$

with *X*, *Y*, two random variables and ε, the error term. After a sampling, we obtain the observational model,

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

and we suppose that $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)' \sim N(0, \sigma^2 I)$.

If we assume that f has some parametric form, well defined except from a finit number of unknown coefficients, then we are in the parametric regression setting.

One of the most used regression models is the polynomial regression, that is

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + ... + \alpha_q x_i^q + \varepsilon_i .$$

It is well known that such a model can be estimated with the least squares method, after a reduction to a multiple linear model, with the explicative variables, $X, X^2,..., X^q$. But before make this estimation it is necessary to establish the form of the regression function, or more precisely, in this case, the polynomial's degree. An ordinary solution to this problem is to estimate the model, for different values of q and then, to compare these models, by performing a regression analysis, either by graphical comparison and/or by quantitative methods (see [7], [8], [9]).

As an alternative, there exists some data driven selection methods that give the appropriate value for q (see [4] and [8]). One of such methods is the cross validation method (CV). Based on this method, the optimal degree *q* will be that which minimizes the cross validation function,

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

where $f_q^{(-i)}$ is the regression polynomial, fitted from all data, less the *i*-th data. A leaving-out-one resampling method is used here. We obtain the fitted models, $f_q^{(-i)}$, $\mathrm{i} = \overline{1,\mathrm{n}}$, from *n* learning samples (each one, with *n-1* data), then we validate these models by other *n* test samples, formed with the one-left-out data.

We already discussed in [2] and [3], how the CV method works, in the case of degree selection, for the polynomial regression and in the case of smoothing parameter selection, for the smoothing spline regression, respectively.

Also, in [2], we proposed a combined algorithm for degree selection, in the case of the polynomial regression. This algorithm is based either on the CV method and on the regression analysis techniques.

Degree selection algorithm

Step 1. Simulate a sample data, $(x_i, y_i), \mathrm{i} = \overline{1,\mathrm{n}}$ and if it is necessary, order and weight the data, with respect to the data sites, x_i.

Step 2. For these simulated data, find a sample with *p* replicate values of q_{CV} and the related distribution.

Step 3. Retain the mode of the distribution, q_{CV}^1 and also, the mode of the remaining values, q_{CV}^2.

Step 4. If the fit with the q_{CV}^{1} degree polynomial is validated by both graphical and quantitative regression analysis, *Stop.*

Else, follow the next step.

Step 5. Make the regression comparative analysis for the cases q_{CV}^{1}, q_{CV}^{2}, $q_{CV}^{1}-1$, $q_{CV}^{1}+1$ and establish the optimal fitting. *Stop.*

However, on wide data range, the polynomial regression works only as a local approximation method, leading to the necesity of finding a more flexible model. A solution to this problem could be the statement of a polynomial regression problem, on each subinterval from a partition of the data range, dealing with a switching regression model, that bears structural changes in some data points.

In this way, the polynomial spline functions, with their great flexibility, arise as good estimators in such cases. We remind here that a polynomial spline function of *m* degree, defined with respect to a mesh, $\Delta : -\infty \le a < t_1 < t_2 < ... < t_N < b \le \infty$, is a piecewise polynomial function (the breaks or the function knots are $t_k, \mathrm{k} = \overline{1, N}$), with the pieces joined at the knots, so that the function has *m-1* continuous derivatives (see [5] and [6]).

Thus, if we use the truncated power functions (see [5]), then we can write for a parametric spline regression model, the form

$$Y = \sum_{k=0}^{m} \alpha_k X^k + \sum_{k=1}^{N} \beta_k (X - t_k)_+^m + \varepsilon .$$

This model can be estimated with the least squares method, after a reduction to a multiple linear model, namely,

$$Y = \alpha_0 + \alpha_1 Z_1 + ... + \alpha_m Z_m + \beta_1 U_1 + ... + \beta_N U_N ,$$

with the explicative variables given by

$$Z_k = X^k, \mathrm{k} = \overline{1,\mathrm{m}},$$
$$\mathrm{U_k} = (X - t_k)_+^m, \mathrm{k} = \overline{1, \mathrm{N}}.$$

The resulted estimator of the regression function is called the least squares spline estimator.

Here, likewise in the polynomial regression, we have to estimate first, the degree of function, the number of knots and the location of its.

Selection of the degree *m* is usually made as a result of a graphical analysis of data and selection of the knots number, *N*, depends on the desired amount of flexibility. Also, there are several elementary rules for selecting the knots. Thus, for a linear spline, the knots are placed at the points where the data exhibit a change in slope, and for a quadratic or cubic spline, the knots are placed near local maxima, minima, or inflection points in the data.

Besides these settings, there are also data driven methods for selecting these parameters (see [1] and [4]). One of such methods and also, a comparison between polynomial and spline regression, follow in the next sections.

2. CV-based best polynomial fitting

For all numerical experiments which will be made in this paper, we consider as a test function, a piecewise function, $f:[-20,20]\to\mathbb{R}$, formed by six pieces of cubics, having the breaks –10, -1, 0, 2, 5 and the coefficients given in the following table:

Table 1

Monom Interval	x^3	x^2	x^1	x^0
$[-20,-10)$	3.5729	131.67	2288.8	11802
$[-10,-1)$	-8.4291	-228.39	-1311.8	-200.45
$[-1,0)$	317.64	749.83	-333.55	125.62
$[0,2)$	-190.84	749.83	-333.55	125.62
$[2,5)$	55.591	-728.76	2623.7	-1845.8
$[5,20]$	-6.9802	209.8	-2069.2	5975.6

Based on this function, we simulate the data (x_i, y_i), where $x_i, i=\overline{1,50}$, are the knots of a uniform partition of the interval $[-20,20]$, and y_i's are given by

$$y_i = f(x_i) + \varepsilon_i, i = \overline{1,50},$$

where ε_i's come from a random number generator simulating independently and identically distributed, $\varepsilon_i \sim N(0{,}500)$, random variables.

For these data, we start with a polynomial fitting. In order to establish the appropriate degree, we have used the degree selection algorithm, reminded in the introductory section of this paper. For 100 replicates of the experiment, the following distribution of the degree q_{CV} is obtained:

$$q_{CV} : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 0 & 0 & 1 & 1 & 85 & 13 \end{pmatrix}.$$

Now, applying the second part of the algorithm, we will reduce the problem of the optimal degree polynomial fitting to a comparison with the regression analysis techniques, between the mode of the distribution, $q_{CV} = 6$, and the cases $q_{CV} = 5$ and $q_{CV} = 7$. We will not present here the graphical comparison, but we mention that the curves for $q_{CV} = 6$ and $q_{CV} = 7$ have almost the same behaviour with respect to the data, being more appropriate than the curve for $q_{CV} = 5$. The quantitative analysis also recommends the case $q_{CV} = 6$, in spite of the other cases. For example the adjusted squared *R* has the values

$$\overline{R}^2(q = 5) = 0{,}940,$$
$$\overline{R}^2(q = 6) = 0{,}946,$$
$$\overline{R}^2(q = 7) = 0{,}947.$$

Hence, we state that the CV-based best polynomial estimator with respect to these data is the sixth degree polynomial.

Anyway, the 95%-confidence bounds presented below show that the accuracy of the best polynomial estimation is not very good (see the magnitude of the confidence interval for the coefficients a_6, a_5, a_4):

a_6 = 40,27 (-426,6; 507,1),

a_5 = 77,32 (-2,081; 156,7),

a_4 = 14,53 (0,911; 28,16),

a_3 = -1,034 (-1,786; -0,2822),

a_2 = -0,1757 (-0,2656; -0,08582),

a_1 = 0,002335 (0,000746; 0,003924),

a_0 = 0,0001942 (3,79e-005; 0,0003506).

3. Least squares spline fitting versus polynomial fitting

In order to emphasize the quality of the spline estimator, with respect to that of the polynomial estimator, we have fitted the same data with a cubic spline function (of third degree), having as knots of multiplicity 1, the breaks of the test function, namely, -10, -1, 0, 2, 5 and the end knots, -20, 20, of multiplicity 4. In this regard, for illustrate the goodness of spline fitting against the best degree polynomial fitting, we have plotted in the following figure, these two fittings, together with the test function and the data.

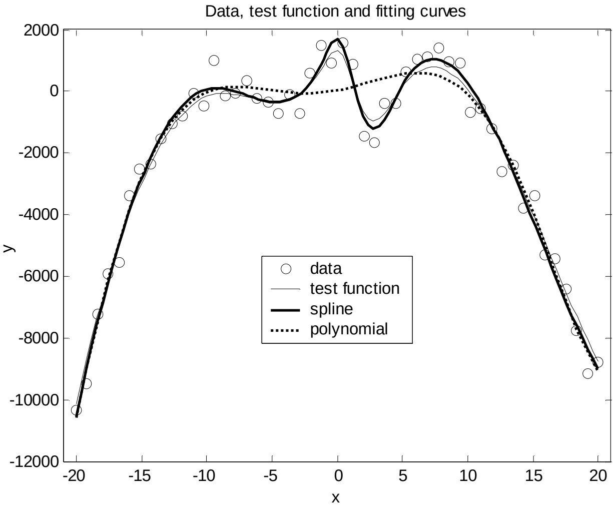

Figure 1

We can observe that, due to its flexibility, the spline curve comes more close to the data than the polynomial curve does. Moreover, the spline curve reconstructs well enough the test function, while the polynomial curve doesn't.

4. CV-based knots selection in the least squares spline regression

Simillary with the polynomial case, we propose here a data driven method for the selection of the least squares spline knots, based on a CV-like function,

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

In this expression, $f_{\lambda_N}^{(-i)}$ is the least squares spline estimator, with the knots given by $\lambda_N = \{t_1, t_2, ..., t_N\}$, obtained by leaving out the *i*-th data from the sample. The optimal set of knots, $\hat{\lambda}_{\hat{N}}$, will be derived by minimizing, for fixed *N*, the expression $CV(\lambda_N)$, thus resulting the estimator $\hat{\lambda}_N$ and then , by obtaining $\hat{\lambda}_{\hat{N}}$, with $\hat{N}$ the minimizer of $CV(\hat{\lambda}_N)$. A weighted version of this function (the generalized cross validation function, see [4]) was already used in this regard.

In order to show how this method works, we implemented the following CV-based algorithm in the Matlab 6.5 environment.

Knots selection algorithm

Step 1. Read the data (x_i, y_i), $i=\overline{1,n}$.

Step 2. For various sets of knots, determine the least squares fitting spline, $f_{\lambda_N}^{(-i)}$.

Step 3. For the cases considered in the previous step, calculate $CV(\lambda_N)$.

Step 4. Determine λ_N^{CV}, for which
$CV(\lambda_N^{CV}) = \min_{\lambda_N} CV(\lambda_N)$.

Stop.

We run this algorithm on the same simulated data as in the previous sections. Starting with the information contained in the shape of the observed data curve (see figure 1), we choose *m*=3 as the degree of the spline estimator. As competitors, we propose the following four sets of knots:

S$_1$={-20, -20, -20, -20, -10, -1, 0, 2, 5, 20, 20, 20, 20},

S_2={-20, -10, -1, 0, 2, 5, 20},
S_3={-20, -20, -20, -20, -10, -1, 0, 5, 20, 20, 20, 20},
S_4={-20, -20, -20, -20, -11, -2, 0, 1, 4, 20, 20, 20, 20}.

The first set is that which was used in the previous section. With respect to the first set, the second set gives up to the multiplicity of the end knots and the third to the knot t=2. The last set is choosed according to the elementary rools, reminded in the introductory section. Thus, we place the knots near local minima, maxima or inflection points in the data. Certainly, our expectations on the optimal knots are for the first set, this set being actualy formed with the breaks of the test function.

In what follows, we will see if the proposed CV-based algorithm can selects the optimal knots. For 100 replicates of the algorithm, with different seeds of the random number generator, we obtain the following mean values of the cross validation function, $CV(\lambda_N)$, corresponding to the four sets of knots:

314914,756972,
13638912,940932,
620668,166187,
389896,095517.

It can be observed that the first set of knots, S_1 (the breaks of the test function), followed by the sets S_4, S_3, S_2 (in this order), is that selected by the CV method. As the following plots show, this hierarchy is validated also by the graphical comparison of these four spline fits. For the clarity of the pictures sake, we plot in the first figure, the cases S_3, S_2 and in the second figure, the cases S_1, S_4. In both figures, we also plot the data and the test function.

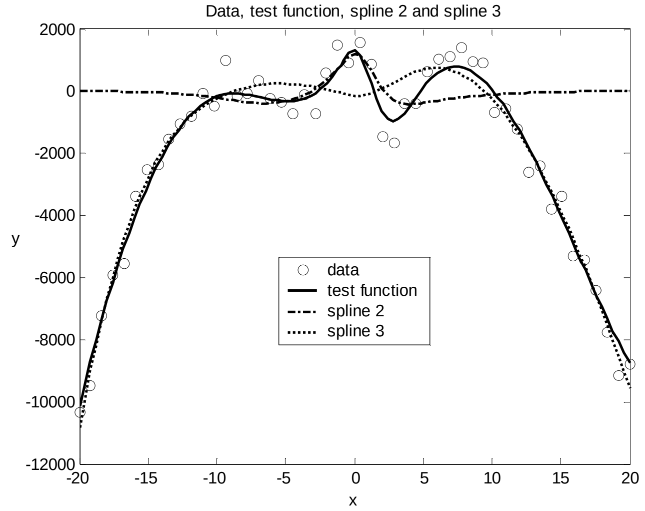

Figure 2

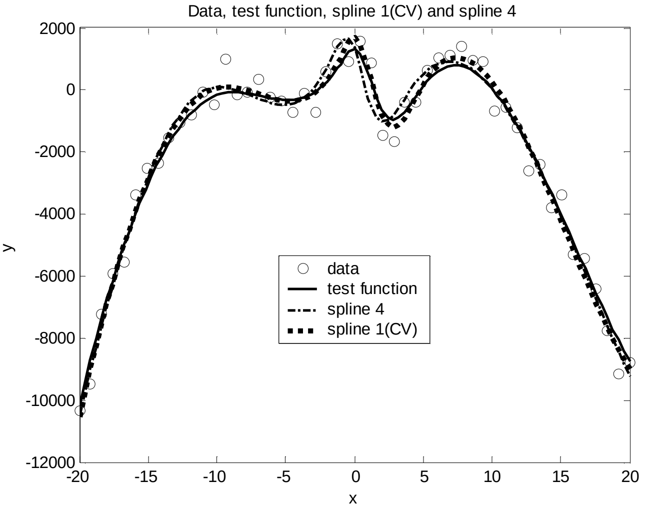

Figure 3

We can see that the curves from the first figure fit he data poorly, while the curves from the last figure, indicate good fits. Anyway, the better is the curve corresponding to the set S_1. Therefore, the cross validation method selects the appropriate set of knots in the least squares spline regression.

References

[1].de Boor C., A Practical Guide to Splines, Springer-Verlag, New York, 1978
[2].Breaz N., The cross-validation method in the polynomial regression, Acta Universitatis Apulensis, Mathematics-Informatics, Proc. of Int. Conf. on Theory and Appl. of Math. and Inf., Alba Iulia, no. 7, Part B, 67-76, 2004
[3].Breaz N., The cross-validation method in the smoothing spline regression, Acta Universitatis Apulensis, Mathematics-Informatics, Proc. of Int. Conf. on Theory and Appl. of Math. and Inf., Alba Iulia, no. 7, Part B, 77-84, 2004
[4].Eubank R.L., Nonparametrc Regression and Spline Smoothing-Second Edition, Marcel Dekker, Inc., New York, Basel, 1999
[5].Micula Gh., Funcţii spline şi aplicaţii, Editura Tehnică, Bucureşti, 1978
[6].Micula Gh., Micula S., Handbook of Splines, Kluwer Academic Publishers, Dordrecht-Boston-London, 1999
[7].Saporta G., Probabilités, analyse des données et statistique, Editions Technip, Paris, 1990
[8].Stapleton J.H., Linear Statistical Models, John Wiley & Sons, New York-Chichester-Brisbane, 1995
[9].Tassi Ph., Méthodes statistiques, 2^{e} édition, Economica, Paris, 1989

Author:
Nicoleta Breaz – “1 Decembrie 1918” University of Alba Iulia, E-mail address: nbreaz@uab.ro