ON THE SMOOTHING PARAMETER IN CASE OF DATA FROM MULTIPLE SOURCES

by
Breaz Nicoleta

Abstract. In this paper we focuse on data smoothing by spline function. The smoothing parameter λ that is involved in this kind of modeling is obtained here from the generalized cross validation (GCV) procedure. Also for data from two sources with different weights is already known a GCV formula for parameter λ given in a particular case when the smoothing function is a function on the circle. We extend this formula to a more general case and in the same time for more than two sources.

1. INTRODUCTION

We consider a regressional model written with *n* observational data

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

where $x_i \in [0,1]$, $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)' \sim N(0, \sigma^2 I)$, a gaussian *n* dimensional vector with zero mean and $\sigma^2 I$ matrix of covariances. About the regression function we just know the information that *f* is in some space W_m defined as

$$W_m = W_m[0,1] = \left\{ f : f, f', ..., f^{(m-1)} \text{ absolutely continuous, } f^{(m)} \in L_2 \right\}.$$

Then, we can speak about a spline smoothing problem. So we obtain an estimate of f by finding $f_\lambda \in W_m$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - f(x_i)\right]^2 + \lambda\int_0^1 \left(f^{(m)}(u)\right)^2 du, \quad \lambda > 0 . \qquad (1)$$

It is known that the solution of this problem is the natural polynomial spline of degree $2m-1$ with knots $x_i, i = \overline{1,n}$.

For the beginning we consider the smoothing parameter λ fixed. Then we can search a solution for (1) in a certain subspace of W_m, spaned by *n* appropriate chosen basis functions. According to [2] such basis functions are related to B-spline but here we are not interested in this. We just consider *f* of the form

$$f = \sum_{k=1}^{n} c_k B_k$$

with B_k basis functions.

Now we rewrite the problem in matriceal form using the following notations:

$$y = (y_1, y_2, ..., y_n)'$$

$$c = (c_1, c_2, ..., c_n)'$$

$$B = (B_{ij})_{\substack{1\le i\le n\\ 1\le j\le n}},\ B_{ij} = B_j(x_i).$$

Also, from [2] we know that we can write the seminorm $J(f) = \int_0^1 \left(f^{(m)}(u)\right)^2 du$ in matriceal form $c'\Sigma c$ for some matrix Σ. Then the problem is to find c to minimize $\|y - Bc\|^2 + \lambda c'\Sigma c$. The solution for this problem is given as $c = (B'B + n\lambda\Sigma)^{-1} B'y$. When the smoothing parameter λ is too small we obtain some function *f* which is close to data despite of its smoothness and when λ is too big we obtain some function f which is very smooth but is not sufficient close to data.

Among the methods which provide an optimal λ from the data are the(cross validation)CV and (general cross validation) GCV procedures described in [2].

According to CV method, λ is the minimizer of the expression

$$CV(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - f_\lambda^{[k]}(x_k)\right)^2$$

with $f_\lambda^{[k]}$ the spline estimate using all data but the k-th data point of *y*. As a generalization, GCV procedure use a more general function

$$GCV(\lambda) = \frac{\frac{1}{n}\|(I - A(\lambda))y\|^2}{\left[\frac{1}{n}Tr(I - A(\lambda))\right]^2}$$

where $A(\lambda)$ is the influence or hat matrix given by the relation

$$\hat{y} = \begin{pmatrix} f_\lambda(x_1) \\ \vdots \\ f_\lambda(x_n) \end{pmatrix} = A(\lambda)y\ .$$

From [2] we know that $A(\lambda)$ has the form $A(\lambda)=B(B'B+n\lambda\Sigma)^{-1}B'$. Also we remind that an estimate for σ^2 is given by $\hat{\sigma}^2=\frac{\|(I-A(\lambda))y\|^2}{Tr(I-A(\lambda))}$.

In the problem formulated here the data came from a single source. In [1], Feng Gao formulate this problem for two sources and give a similar GCV method for λ. Gao consider a particular case when f is a function on the circle.

In this paper we extend the results of Gao for more than two sources and in the same time we consider a more general case for domain of function f.

2.MAIN RESULTS

We consider the case when the data came from l sources with unknown weights as in

$$\begin{aligned}
y_{1i} &= f(x_{1i})+\varepsilon_{1i},\ i=\overline{1,N_1}\\
y_{2i} &= f(x_{2i})+\varepsilon_{2i},\ i=\overline{1,N_2}\\
&\dots\dots\dots\dots\dots\dots\\
y_{li} &= f(x_{li})+\varepsilon_{li},\ i=\overline{1,N_l},\\
N_1 &+ N_2+\ldots+N_l=n,
\end{aligned}$$

with residual gaussian vectors defined as

$$\begin{aligned}
\varepsilon_1 &= (\varepsilon_{11},\varepsilon_{12},\ldots,\varepsilon_{1N_1})' \sim N(0,\sigma_1^2 I)\\
\varepsilon_2 &= (\varepsilon_{21},\varepsilon_{22},\ldots,\varepsilon_{2N_2})' \sim N(0,\sigma_2^2 I)\\
&\dots\dots\dots\dots\dots\dots\\
\varepsilon_l &= (\varepsilon_{l1},\varepsilon_{l2},\ldots,\varepsilon_{lN_l})' \sim N(0,\sigma_l^2 I)
\end{aligned}$$

We assume that $\sigma_1^2,\sigma_2^2,\ldots,\sigma_l^2$ are unknown and $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_l$ are independent.

If we knew $\sigma_1^2,\sigma_2^2,\ldots,\sigma_l^2$, an estimate of function f is a solution of the variational problem given as

$$\min_{f \in W_m} \frac{1}{N_1 + N_2 + \ldots + N_l} \left[\frac{1}{\sigma_1^2} \sum_{i=1}^{N_1} \left(y_{1i} - f(x_{1i})\right)^2 + \frac{1}{\sigma_2^2} \sum_{i=1}^{N_2} \left(y_{2i} - f(x_{2i})\right)^2 + \ldots \right.$$
$$\left. + \frac{1}{\sigma_l^2} \sum_{i=1}^{N_l} \left(y_{li} - f(x_{li})\right)^2 \right] + \lambda J(f).$$

We search for $f \in W_m$ of the form $f = \sum_{k=1}^{n} c_k B_k$ and we use the notations

$$y_1 = \left(y_{11}, y_{12}, \ldots, y_{1N_1}\right)'$$
$$y_2 = \left(y_{21}, y_{22}, \ldots, y_{2N_2}\right)'$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$y_l = \left(y_{l1}, y_{l2}, \ldots, y_{lN_l}\right)'$$
$$c = \left(c_1, c_2, \ldots, c_n\right)'$$

$$B_1 = \left(B_{ij}\right)_{\substack{1 \le i \le N_1 \\ 1 \le j \le n}}, \ B_{ij} = B_j(x_{1i})$$
$$B_2 = \left(B_{ij}\right)_{\substack{1 \le i \le N_2 \\ 1 \le j \le n}}, \ B_{ij} = B_j(x_{2i})$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$B_l = \left(B_{ij}\right)_{\substack{1 \le i \le N_l \\ 1 \le j \le n}}, \ B_{ij} = B_j(x_{li})$$

Now we have *l* matrices *B*.
The model becomes

$$y_1 = B_1 c + \varepsilon_1$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$y_l = B_l c + \varepsilon_l$$
$$\varepsilon_i \sim N\left(0, \sigma_i^2 I\right), i = \overline{1,l}$$

and the variational problem is now equivalent to find c the minimizer of

$$\frac{1}{\theta \cdot n}\left(r_1 \left\| y_1 - B_1 c \right\|^2 + r_2 \left\| y_2 - B_2 c \right\|^2 + \ldots + r_l \left\| y_l - B_l c \right\|^2 + \alpha c' \Sigma c\right) \qquad (2)$$

with Σ some known matrix, θ a nuisance parameter, α a smoothing parameter and r_i the weighting parameters given as

$$\theta = \sigma_1 \sigma_2 ... \sigma_l$$

$$r_i = \frac{\sigma_1 \sigma_2 ... \sigma_{i-1} \sigma_{i+1} ... \sigma_l}{\sigma_i}, \ i = \overline{1,l} \ \left(r_1 r_2 ... r_l = 1 \right)$$

$$\alpha = \sigma_1 \sigma_2 ... \sigma_l \cdot \lambda \cdot n .$$

We can prove the following proposition:

Proposition 1. For fixed $r = \left(r_1, r_2, ..., r_l \right)$ and α the solution of the variational problem (2) is

$$\hat{c}_{r,\alpha} = \left(r_1 B_1' B_1 + r_2 B_2' B_2 + ... + r_l B_l' B_l + \alpha \Sigma \right)^{-1} \left(r_1 B_1' y_1 + r_2 B_2' y_2 + ... + r_l B_l' y_l \right) \quad (3)$$

Proof. We denote by E the expression that has to minimize so we have the condition

$$\frac{\partial E}{\partial c} = 0 .$$

This condition is equivalent with

$$- r_1 B_1' y_1 + r_1 B_1' B_1 c - ... - r_l B_l' y_l + r_l B_l' B_l c + \alpha \Sigma c = 0$$

and the solution of this matriceal equation is

$$c = \left(r_1 B_1' B_1 + r_2 B_2' B_2 + ... + r_l B_l' B_l + \alpha \Sigma \right)^{-1} \left(r_1 B_1' y_1 + r_2 B_2' y_2 + ... + r_l B_l' y_l \right).$$

According to formula (3) it is important to get some estimates for $r\left(r_1, r_2, ..., r_{l-1} \text{ and } r_l = \frac{1}{r_1 r_2 ... r_{l-1}} \right)$ and α. Using the similar methods with [1] we give a GCV function that we can use for estimating r and α in the same time.

We introduce the notations

$$\begin{aligned} y &= \left(y_1', y_2', ..., y_l' \right)' \\ B &= \left(B_1', B_2', ..., B_l' \right)' \end{aligned} \quad (4)$$

and the matrix

$$I(r)=\begin{pmatrix} \frac{1}{\sqrt{r_1}}I_{N_1} & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & \frac{1}{\sqrt{r_2}}I_{N_2} & 0 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & \frac{1}{\sqrt{r_{l-1}}}I_{N_{l-1}} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & \sqrt{r_1 r_2 \ldots r_{l-1}}\, I_{N_l} \end{pmatrix} \qquad (5)$$

Also we use the notations

$$\begin{aligned} M &= \left(r_1 B_1' B_1 + r_2 B_2' B_2 + \ldots + r_l B_l' B_l + \alpha\Sigma \right) \\ y^r &= I^{-1}(r)\cdot y \qquad\qquad (6) \\ B^r &= I^{-1}(r)\cdot B \end{aligned}$$

We can prove the following proposition:

Proposition 2. The influence matrix $A^r(r,\alpha)$ defined by

$$\hat{y}^r = A^r(r,\alpha)y^r \qquad (7)$$

has the form

$$A^r(r,\alpha)=\begin{pmatrix} r_1 B_1 M^{-1} B_1' & \sqrt{r_1 r_2}\, B_1 M^{-1} B_2' & \ldots & \sqrt{r_1 r_l}\, B_1 M^{-1} B_l' \\ \sqrt{r_1 r_2}\, B_2 M^{-1} B_1' & r_2 B_2 M^{-1} B_2' & \ldots & \sqrt{r_2 r_l}\, B_2 M^{-1} B_l' \\ \ldots & \ldots & \ldots & \ldots \\ \sqrt{r_l r_1}\, B_l M^{-1} B_1' & \sqrt{r_l r_2}\, B_l M^{-1} B_2' & \ldots & r_l B_l M^{-1} B_l' \end{pmatrix}$$

or

$$A^r(r,\alpha) = B^r M^{-1} {B^r}' .$$

Proof. The solution (3) of the variational problem (2) with the notations (4), (5), (6) can be written as

$$\hat{c}_{r,\alpha} = \left({B^r}' B^r + \alpha\Sigma \right)^{-1} {B^r}' y^r = M^{-1} {B^r}' y^r .$$

The condition (7) becomes

$$B^r \cdot \left[M^{-1} B^{r'} y^r \right] = A^r(r,\alpha) \cdot y^r$$

and further

$$A^r(r,\alpha) = B^r M^{-1} B^{r'} \, .$$

In order to get a GCV formula for r and α, first we construct for our case a CV-like formula.

Now we denote by $c_{r,\alpha}^{[k]}$, the spline estimate of *c*, using all but the k-th data point of *y*; also we denote by y_k the k-th data point and by B^k and $B^{k,r}$ the k-th row of *B* and B^r respectively.

Then the (cross validation) CV function for our model would be

$$CV(r,\alpha) = \frac{1}{n} \sum_{k=1}^{n} \left(y_k - B^k c_{r,\alpha}^{[k]} \right)^2$$

Next we prove the corresponding leaving-out-one lemma for our case that is:

Lemma 3. Let be $h_{r,\alpha}[k,z]$ the solution to the variational problem

$$\min_c \frac{1}{\theta \cdot n} \left(r_1 \| y_1 - B_1 c \|^2 + \ldots + r_l \| y_l - B_l c \|^2 + \alpha c' \Sigma c \right)$$

with the k-th data point y_k^r replaced by *z*. Then we have

$$h_{r,\alpha}\left[k, B^{k,r} c_{r,\alpha}^{[k]} \right] = c_{r,\alpha}^{[k]}$$

Proof. We have

$$\left(B^{k,r} c_{r,\alpha}^{[k]} - B^{k,r} c_{r,\alpha}^{[k]} \right)^2 + \sum_{\substack{i=1 \\ i \neq k}}^{n} \left(y_i^r - B^{i,r} c_{r,\alpha}^{[k]} \right)^2 + \alpha \left(c_{r,\alpha}^{[k]} \right)' \Sigma c_{r,\alpha}^{[k]} =$$

$$= \sum_{\substack{i=1 \\ i \neq k}}^{n} \left(y_i^r - B^{i,r} c_{r,\alpha}^{[k]} \right)^2 + \alpha \left(c_{r,\alpha}^{[k]} \right)' \Sigma c_{r,\alpha}^{[k]} \le$$

$$\leq \sum_{\substack{i=1 \\ i\neq k}}^{n}\left(y_i^r - B^{i,r}c\right)^2 + \alpha c'\Sigma c \leq$$

$$\leq \left(B^{k,r}c_{r,\alpha}^{[k]} - B^{k,r}c\right)^2 + \sum_{\substack{i=1 \\ i\neq k}}^{n}\left(y_i^r - B_i^r c\right)^2 + \alpha c'\Sigma c,\ (\forall)c \in R^n$$

Based on the leaving-out-one lemma now we can prove the following theorem:

Theorem 4. We have the identity:

$$CV(r,\alpha) = \frac{1}{n}\sum_{k=1}^{n}\frac{\left(I_k^{-1}(r)\cdot\left[I - A^r(r,\alpha)\right]I^{-1}(r)y\right)^2}{\left(I_k^{-1}(r)\cdot\left[I - A^r(r,\alpha)\right]\left(I_k^{-1}(r)\right)'\right)^2}$$

with $I_k^{-1}(r)$, the k-th row of the matrix $I^{-1}(r)$.

Proof. We consider the following identity

$$y_k - B^k c_{r,\alpha}^{[k]} = y_k - B^k c_{r,\alpha}^{[k]}\cdot\frac{y_k^r - B^{k,r}c_{r,\alpha}}{y_k^r - B^{k,r}c_{r,\alpha}}. \tag{8}$$

Further we have

$$y_k - B^k c_{r,\alpha}^{[k]} = \frac{y_k^r - B^{k,r}c_{r,\alpha}}{\dfrac{\sqrt{r_{s_k}}\cdot y_k - \sqrt{r_{s_k}}B^k c_{r,\alpha}}{y_k - B^k c_{r,\alpha}^{[k]}}}$$

where

$$s_k = \begin{cases} 1 \quad \text{if} \quad k \in \{1,2,\dots,N_1\} \\ 2 \quad \text{if} \quad k \in \{N_1+1,\dots,N_2\} \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ l \quad \text{if} \quad k \in \{N_{l-1}+1,\dots,N_l\} \end{cases}.$$

Moreover, we can write

$$y_k - B^k c_{r,\alpha}^{[k]} = \frac{y_k^r - B^{k,r}c_{r,\alpha}}{\sqrt{r_{s_k}} - \dfrac{B^{k,r}c_{r,\alpha} - B^{k,r}c_{r,\alpha}^{[k]}}{y_k - B^k c_{r,\alpha}^{[k]}}} \tag{9}$$

By the leaving out one lemma we can write that

$$\frac{B^{k,r}c_{r,\alpha}-B^{k,r}c_{r,\alpha}^{[k]}}{y_k-B^k c_{r,\alpha}^{[k]}}=\frac{B^{k,r}h_{r,\alpha}\left[k,B^{k,r}c_{r,\alpha}\right]-B^{k,r}h_{r,\alpha}\left[k,B^{k,r}c_{r,\alpha}^{[k]}\right]}{y_k-B^k c_{r,\alpha}^{[k]}}=$$

$$=\frac{B^{k,r}h_{r,\alpha}\left[k,y_k^r\right]-B^{k,r}h_{r,\alpha}\left[k,B^{k,r}c_{r,\alpha}^{[k]}\right]}{\dfrac{y_k^r-B^{k,r}c_{r,\alpha}^{[k]}}{\sqrt{r_{s_k}}}}$$

Because $B^{k,r}c_{r,\alpha}$ is linear in each data point, we can replace the divided difference below by a derivative. So we can write that

$$\frac{B^{k,r}c_{r,\alpha}-B^{k,r}c_{r,\alpha}^{[k]}}{y_k-B^k c_{r,\alpha}^{[k]}}=\frac{\partial B^{k,r}c_{r,\alpha}}{\partial y_k^r}\cdot\sqrt{r_{s_k}}\,. \qquad (10)$$

Moreover, we have

$$\frac{\partial B^{k,r}c_{r,\alpha}}{\partial y_k^r}=a_{kk}\,, \qquad (11)$$

the kk-th entry from $A^r(r,\alpha)$ matrix.

So, according to (8), (9), (10) and (11) we can write

$$y_k-B^k c_{r,\alpha}^{[k]}=\frac{y_k^r-B^{k,r}c_{r,\alpha}}{(1-a_{kk})\cdot I_{kk}^{-1}(r)}\,.$$

Moreover,

$$y_k^r-B^{k,r}c_{r,\alpha}=y_k^r-A^{k,r}(r,\alpha)\cdot y^r=\left[I_k-A^{k,r}(r,\alpha)\right]\cdot I^{-1}(r)y$$

where $A^{k,r}(r,\alpha), I_k$ is a k-th row of matrix $A^r(r,\alpha)$ and I respectively.

So we have

$$y_k-B^k c_{r,\alpha}^{[k]}=\frac{\left[I_k-A^{k,r}(r,\alpha)\right]\cdot I^{-1}(r)y}{\left(1-a_{kk}^r\right)\cdot I_{kk}^{-1}(r)}\,.$$

After multiplying the numerator and the denominator by $I_{kk}^{-1}(r)$ we obtain the results.

In the same manner as in [1] or [2] we generate the GCV function as follows. We replace the denominator $I_k^{-1}(r)\left(I-A^r(r,\alpha)\right)I_k^{-1}(r)'$ by the average

$$\frac{1}{n}\sum_{k=1}^{n} I_k^{-1}(r)\left(I - A^r(r,\alpha)\right)I_k^{-1}(r)' = \frac{1}{n}Tr\left[I^{-1}(r)\left(I - A^r(r,\alpha)\right)\cdot I^{-1}(r)'\right]$$

and we have a GCV-like function

$$GCV(r,\alpha) = \frac{\frac{1}{n}\left\|I^{-1}(r)\left[I - A^r(r,\alpha)\right]I^{-1}(r)y\right\|^2}{\left[\frac{1}{n}Tr\left(I^{-1}(r)\left(I - A^r(r,\alpha)\right)\cdot I^{-1}(r)'\right)\right]^2}$$

This function is a generalization for the $GCV(\lambda)$ function from [2] (one source) and the $GCV(r,\alpha)$ function from [1] (two sources).

REFERENCES

[1] Gao Feng, *On combining data from multiple sources with unknown relative weights*, Technical Report, no. 894/1993, University of Wisconsin, Madison
[2] Wahba Grace, *Spline models for observational data*, SIAM, Philadelphia, CBMS-NSF Regional Conference, Series in Applied Mathematics, vol. 59.

Author:
Nicoleta Breaz, "1 Decembrie 1918" University of Alba lulia, Romania., Computer Science and Mathematics Department, e-mail: nbreaz@uab.ro