CUBIC SPLINE ESTIMATORS OF THE PROBABILITY DENSITY FUNCTION

Căuş Vasile-Aurel
University of Oradea Faculty of Science Department of Mathematics and Informatics vcaus@uoradea.ro

The estimation of the probability density function, using spline functions, of a random variable X is studied. It is assumed that the variable is measured with an error having a normal distribution with known parameters.

Keywords: Cubic spline, probability density, least squares method.

Jel Classification: C13, C14

1. Introduction

We study the cubic spline function used to approximate the probability density function of a random variable. We obtain, using the least square method, an interpolating cubic spline having minimal quadratic oscillation in average.

Let the $i-th$ observation on X be denoted by $Y_i = X_i + \varepsilon_i$, where ε_i is the measurement error. It is assumed that $\{X_i\}$ and $\{\varepsilon_i\}$ are mutually independent and each is identically distributed. The model for a variable measured with error is:

$$Y_i = X_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Denote by $f_\varepsilon(x)$, $f_X(x)$ and $f_Y(x)$ the probability density functions for ε, X and Y. We suppose that the probability density functions are continuous.

Chen, Fuller and Breidt[379] proposed a methodology to approximate the probability density function of X by using cubic grafted polynomial.

One of the concluding remarks of mentioned paper is that the spline estimators are reasonable. The aim of this paper is to connect that affirmation with the result on cubic spline functions having minimal oscillation in average. Such functions are very useful when working with experimental data.

2. Cubic spline functions having minimal quadratic oscillation

Alexandru Mihai Bica[380] gave the following definition of quadratic oscillation in average:

Let $f:[a,b] \to \mathbf{R}$ be a continuous function such that $f(x_i) = Y_i, \forall\, i = \overline{0,n}$ and denote by $f_i, i = \overline{1,n}$ his restriction to the subinterval $x \in [x_{i-1}, x_i]$ of the division Δ_n. The functional $\rho(\cdot, \Delta_n, Y): C[a,b] \to \mathbf{R}$ defined by:

$$\rho(f, \Delta_n, Y) = \left(\sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} \left[f_i(x_i) - D_i(Y)(x) \right]^2 dx \right)^{\frac{1}{2}},$$

379 Cong Chen, Wayne A. Fuller, and F. Jay Breidt, “Spline estimators of the density function of a variable measured with error”, Communications in statistics, vol. 32, No. 1, 2003, p. 76.

380 Alexandru Mihai Bica, Mathematical models in biology governed by differential equations, PhD thesis, "Babes-Bolyai" University Cluj-Napoca, 2004.

where

$$D_i\left(Y\right) = Y_{i-1} + \frac{Y_i - Y_{i-1}}{h_i}\left(x_i - x_{i-1}\right), x \in \left[x_{i-1}, x_i\right], i = \overline{1,n}$$

is called the quadratic oscillation in average of the function f corresponding to the division Δ_n and the vector of values $\left\{Y_i\right\}$.

It is not difficult to see that, except $D_i\left(Y\right)$, any interpolation procedures present oscillation in the interior of the interval $x \in \left(x_{i-1}, x_i\right)$, $i = \overline{1,n}$

Consider the division Δ_n, the vector of values $\left\{Y_i\right\}$ and the cubic spline S, generated by the integration of the following two point boundary problem:

$$\begin{cases} S_i''\left(x\right) = \dfrac{1}{h_i}\left[M_i\left(x - x_{i-1}\right) + M_{i-1}\left(x_i - x\right)\right], \quad x \in \left[x_{i-1}, x_i\right] \\ S_i\left(x_{i-1}\right) = Y_{i-1}, \quad i = \overline{1,n} \\ S_i\left(x_i\right) = Y_i \end{cases}$$

where

$$M_i = S_i''\left(x_i\right), i = \overline{0,n}.$$

In Gheorghe Micula and Sanda Micula[381] is shown that the expression of S_i is:

$$S_i\left(x\right) = \frac{M_i\left(x - x_{i-1}\right)^3 + M_{i-1}\left(x_i - x\right)^3}{6h_i} + \left(Y_{i-1} - \frac{M_{i-1}h_i^2}{6}\right)\left(\frac{x_i - x}{h_i}\right) + \left(Y_i - \frac{M_i h_i^2}{6}\right)\left(\frac{x - x_{i-1}}{h_i}\right),$$

$x \in \left[x_{i-1}, x_i\right]$, $i = \overline{1,n}$

Vasile-Aurel Căuş[382] shown that, for any division Δ_n and any vector of values $\left\{Y_i\right\}$, there exist an unique pair $\left(\overline{M_0}, \overline{M_n}\right) \in \mathbf{R}^2$ for which the quadratic oscillation in average of the corresponding cubic spline is minimal.

The geometric interpretation of the quadratic oscillation in average and of his minimal property is that in the plane there exist a set of points between the graph of $S\left(x; \overline{M_0}, \overline{M_n}\right)$ and the polygonal line joining the points $x \in \left(x_i, Y_i\right)$, $i = \overline{1,n}$. If we rotate this set around the Ox axis we obtain a body having minimal volume.

3. Estimate the probability density function

381 Gheorghe Micula, Sanda Micula, Handbook of splines, Kluwer Acad. Publ., vol 462, 1999.

382 Vasile-Aurel Căuş, "Minimal quadratic oscillation for cubic splines", Journal of Computational Analysis and Applications, vol. 9, 2007, pp. 89-90.

In Fuller[383] and Kooperberg and Stone[384] we find the idea of using spline functions linear at the ends.
To connect this idea to the result in paragraph 2 we have to rewrite the problem generating the spline function as follows:

$$\begin{cases} S_i''(x) = \dfrac{1}{h_i}\left[M_i(x - x_{i-1}) + M_{i-1}(x_i - x)\right], \quad x \in [x_{i-1}, x_i] \\ S_i(x_{i-1}) = Y_{i-1}, \quad i = \overline{2, n-1} \\ S_i(x_i) = Y_i \end{cases}$$

$$\left(\overline{M_0}, \overline{M_n}\right) = (0,0)$$

The last condition is obviously restrictive and, in this case, the result obtained by Vasile-Aurel Căuş[385] is not applicable.
Another possibility is to reformulate the problem in the following manner:

$$\begin{cases} S_i''(x) = \dfrac{1}{h_i}\left[M_i(x - x_{i-1}) + M_{i-1}(x_i - x)\right], \quad x \in [x_{i-1}, x_i] \\ S_i(x_{i-1}) = Y_{i-1}, \quad i = \overline{1, n} \\ S_i(x_i) = Y_i \\ S_0(x) = S_1'(x_0)(x - x_0) + S_1(x_0) \\ S_{n+1}(x) = S_n'(x_n)(x - x_n) + S_n(x_n) \end{cases}$$

The solution of this problem can be optimized to satisfy both conditions of minimal quadratic oscillation and linearity at the ends.
Moreover Vasile-Aurel Căuş[386] shown that, if the probability density function f_X and the derivative f_X' satisfies a Lipschitz condition with constant L then the following error estimation hold:

$$\left\| f_X - S \right\|_C \le \left(L + \max\left\{ \left| \overline{M_i} \right| : i = \overline{0, n} \right\} \right) \cdot \max\left\{ h_i^2 : i = \overline{1, n} \right\}$$

where $h_i = x_i - x_{i-1}, i = \overline{1, n}$

383 Wayne A. Fuller, Introduction to statistical time series, Wiley, New York, 1996.
384 Charles Kooperberg, Charles Stone, “A study of logspline density estimation”, Computational statistic and data analysis, vol. 12, pp. 327-347, 1991.
385 Op. cit. pp.89-90
386 Vasile-Aurel Căuş, “Minimal quadratic oscillation for cubic splines”, Journal of Computational Analysis and Applications, vol. 9, 2007, p. 91.

4. Conclusion

A spline function having the minimal quadratic oscillation in average was considered for the estimation of the probability density function.

The geometric interpretation is that, if we rotate, around the Ox axis, the polygonal line joining the points $x \in (x_i, Y_i)$, $i = \overline{1,n}$ and the geometrical image of spline function we obtain a body having minimal volume.

References

1. Alexandru Mihai Bica, *Mathematical models in biology governed by differential equations*, PhD thesis, "Babes-Bolyai" University Cluj-Napoca, 2004.
2. Alexandru Mihai Bica, Vasile-Aurel Căuş, Ioan Fechete, Sorin Mureşan, “Application of the Cauchy-Buniakovski-Schwarz’s inequality to an optimal property for cubic splines”, *Journal of Computational Analysis and Applications*, vol. 9, pp. 43-53, 2007.
3. Carl de Boor, “A practical guide to splines”, *Apllied Math. Sciences*, vol. 27, New York, Heidelberg, Berlin, Springer-Verlag, 1978.
4. Vasile-Aurel Căuş, “Minimal quadratic oscillation for cubic splines”, *Journal of Computational Analysis and Applications*, vol. 9, pp. 85-92, 2007.
5. Cong Chen, Wayne A. Fuller, and F. Jay Breidt, “Spline estimators of the density function of a variable measured with error”, *Communications in statistics*, vol. 32, No. 1, pp. 73-86, 2003.
6. Wayne A. Fuller, *Introduction to statistical time series*, Wiley, New York, 1996.
7. Charles Kooperberg, Charles Stone, “A study of logspline density estimation”, *Computational statistic and data analysis*, vol. 12, pp. 327-347, 1991.
8. Gheorghe Micula, Sanda Micula, *Handbook of splines*, Kluwer Acad. Publ., vol 462, 1999.