A CLASSICAL SMOOTHING FILTER

This note describes the means of implementing a classical smoothing device which is used in a wide variety of applications ranging from industrial design to econometric time-series analysis. In industrial design, the device corresponds to the well-know Reinsch smoothing spline which has been used widely in automotive and aeronautical engineering and in naval architecture. In econometrics, the same device, which is used for trend estimation, has been called the Hordrick–Prescott filter.

Imagine that a set of T coordinates $(x_t, y_t); t = 0, \ldots, T-1$ are available and that it is required to interpolate a curve $\xi(t)$ through these points so as to describe a smooth trajectory which does not depart too far from the data. A criterion which balances the two conflicting objectives of smoothness and goodness of fit is to find a trend which minimises the function

$$S = \sum_{t=0}^{T-1}(y_t - \xi_t)^2 + \lambda \sum_{t=1}^{T-2}\left\{(\xi_{t+1} - \xi_t) - (\xi_t - \xi_{t-1})\right\}^2, \tag{1}$$

where the values $\xi_t = \xi(x_t)$ are the ordinates of the interpolated function at the points x_t. The first term of the criterion function is the sum of squares of the deviations of the curve from the points. On the understanding that the abscissae are equally spaced, the sum of squares from the second term, which comprises the centralised second differences of the sequence $\{\xi_t\}$, represents a measure of the overall curvature of the function $\xi(x)$. The purpose of the parameter λ is to strike a balance between the two aspects of the criterion which are liable to be in conflict.

A wide variety of curves may be fitted to the data points in fulfilment of the criterion of minimising the function S. In the case of the Reinsch smoothing spline, the function $\xi(x)$ is a compound curve made of short cubic segments which bridge the gaps between adjacent nodes (x_{t-1}, ξ_{t-1}), (x_t, ξ_t). The segments are subject to continuity conditions at the nodes which require that the first and second derivatives of adjacent segments are equal at the junctions.

The cubic spline is the mathematical analogue of an old-fashioned draftsman's tool which can be used to draw smooth curves. The draftsman's spline is a thin flexible piece of wood which is clamped to a series of pins placed along the path of the curve which has to be described. The pins to which the spline is clamped correspond to the data points through which we might interpolate a mathematical cubic spline. The cubic spline becomes a device for modelling a trend when, instead of passing through the data points, it is allowed, in the interests of smoothness to deviate from them. One can imagine a spline which is attached to the pins by springs. Then the degree of smoothing would be determined by the stiffness of the springs.

In smoothing a time series, there is no need to bridge the gaps between adjacent data points unless one wishes to envisage an underlying continuous process of which the points represent periodic observations. Therefore, in the sequel, we shall confine our attention to the determination of the ordinates ξ_t. We shall make two separate approaches to the problem; and we shall end by showing that they are, essentially, equivalent.

The Problem in Terms of Matrix Algebra

Define the vectors $y = [y_0, y_1, \ldots, y_{T-1}]'$, $\xi = [\xi_0, \xi_1, \ldots, \xi_{T-1}]'$, and let Q' be a matrix of order $(T-2) \times T$ which is the analogue of the centralised second difference operator. Also define the matrix $W = Q'Q$ of order $(T-2) \times (T-2)$. For specific examples of these matrices, one may consider the case where $T = 6$. Then

$$(2) \qquad Q' = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} 6 & -4 & 1 & 0 \\ -4 & 6 & -4 & 1 \\ 1 & -4 & 6 & -4 \\ 0 & 1 & -4 & 6 \end{bmatrix}.$$

With this notation, the criterion of (1) can be written as

$$(3) \qquad S = (y - \xi)'(y - \xi) + \lambda \xi' Q Q' \xi$$

Differentiating the criterion function S with respect to ξ and setting the result to zero for a minimum gives a first-order condition

$$(4) \qquad y - \xi + \lambda Q Q' \xi = 0,$$

from which it follows that

$$(5) \qquad y = (\lambda Q Q' + I)\xi$$

This equation may be solved to obtain the vector ξ of the estimated trend. Observe that, whereas the matrix QQ' is singular, the matrix $\lambda QQ' + I$ will nonsingular for all finite values of λ

There is a way of solving the equation which may have some advantage in terms of numerical stability when the value of λ is large. Consider premultiplying the equation (5) by Q'. Then we get

$$(6) \qquad \begin{aligned} d = Q'y &= (\lambda Q'Q + I)Q'\xi \\ &= (\lambda W + I)\delta \end{aligned}$$

where $d = Q'q$ is the vector $[d_1, d_2, \ldots, d_{T-2}]'$ of the second differences of the data and where $\delta = Q'\xi$ is the corresponding vector obtained by differencing ξ.

The square matrix $\lambda W + I = \Phi$ is of order $T - 2$ and, in view of its size and the sparseness of its nonzero elements, it should never be formed in practice. Only the generic elements of its rows should be stored in the computer. Nevertheless, the matrix represents a useful concept in describing the solution procedure.

Consider therefore the Cholesky decomposition $\Phi = MM'$, wherein M represents a lower-triangular matrix of three nonzero diagonals bands. With $\lambda \geq 0$, the matrix $\lambda W + I = \Phi$ is manifestly positive definite. Therefore the Cholesky decomposition is always available and the diagonal elements of M are guaranteed to be real positive numbers. Consider writing the equation (6) as

$$(7) \qquad d = MM'\delta = Mq \qquad \text{with} \quad q = M'\delta.$$

To solve this equation, we first obtain the vector q from $d = Mq$. Here the tth row takes the form of

(8) $$\begin{bmatrix} \mu_{t,t-2} & \mu_{t,t-1} & \mu_{t,t} \end{bmatrix} \begin{bmatrix} q_{t-2} \\ q_{t-1} \\ q_t \end{bmatrix} = d_t.$$

Rearranging this gives

(9) $$q_t = \frac{1}{\mu_{t,t}}\{d_t - \mu_{t,t-1}q_{t-1} - \mu_{t,t-2}q_{t-2}\},$$

which shows that the equation can be solved by a simple recursion beginning with $q_1 = d_1/\mu_{0,1}$ and $q_2 = \{d_1 - \mu_{1,1}d_1\}/\mu_{1,1}$. Once the elements of vector q have been computed, those of the vector δ may be calculated by a similar recursion based on the equation $q = M'\delta$. This second recursion, which notionally works backward in time. generates the elements of the vector δ, in the order $\delta_{T-2}, \delta_{T-2}, \ldots, \delta_1$.

Once the elements of the differenced vector $\delta = Q'\xi$ have been found, those of the trend vector ξ can be recovered from the equation

(10) $$\xi = y + \lambda Q\delta$$

whch comes directly from the first-order condition of (4).

The Problem in Terms of Linear Filtering

Let $y(t)$ denote the time series from which the trend is to be extracted. Let L denote the lag operator, which has the effect that $Ly(t) = y(t-1)$, and let $F = L^{-1}$ denote the forwards-shift operator such that $Fy(t) = y(t+1)$. If $\nabla = I - L$ and $\Delta = F - I$ denote, respectively, the backwards difference operator and the forwards difference operator, then $\nabla\Delta = F - 2I + L$ denotes the centralised version of the operator which produces the second difference of a series.

The problem which we can now envisage is that of estimating a trend series $\xi(t)$ such that the series

(11) $$\{y(t) - \xi(t)\}^2 + \lambda\{\nabla\Delta\xi(t)\}^2$$

is minimised for every value of t. For a fixed value of t, we must minimise the quantity

(12) $$\begin{aligned}\{y_t - \xi_t\}^2 &+ \lambda\{\xi_{t+2} - 2\xi_{t+1} + \xi_t\}^2 \\ &+ \lambda\{\xi_{t+1} - 2\xi_t + \xi_{t-1}\}^2 \\ &+ \lambda\{\xi_t - 2\xi_{t-1} + \xi_{t-2}\}^2.\end{aligned}$$

Differentiating this with respect to ξ_t and setting the result to zero gives a first-order condition for minimisation. The latter indicates that the sequence $\xi(t)$ must obey the condition

(13) $$\{\lambda F^2 - 4\lambda F + (6\lambda + 1)I - 4\lambda L + \lambda L^2\}\xi(t) = y(t).$$

The operator in equation (10) defines a symmetric two-sided filter which looks equally backwards and forwards in time. The filter can be written in the form of

$$\begin{aligned} \varphi(L) &= \{\lambda F^2 - 4\lambda F + (6\lambda + 1)I - 4\lambda L + \lambda L^2\} \\ &= (\mu_0 + \mu_1 F + \mu_2 F^2)(\mu_0 + \mu_1 L + \mu_2 L^2) \\ &= \mu(F)\mu(L). \end{aligned} \tag{14}$$

Notice that the coefficients of the filter polynomial correspond to the nonzero elements of a row of the matrix $\Phi = \lambda W + I$ of equation (6). The factorisation of the polynomial $\varphi(L)$ into its forwards and backwards components can be effected using the methods which are applied in finding the parameters of a moving-average process from its autocovariances.

The filter $\varphi(L) = \mu(F)\mu(L)$ might be applied directly to a process $\xi(t)$ to generate $y(t) = \varphi(L)\xi(t)$. However it cannot be used to generate $\xi(t)$ from $y(t)$ by a direct recursion based on the equation

$$\xi(t) = \frac{1}{\lambda}\{y(t) + 4\lambda\xi(t-1) - (6\lambda + 1)\xi(t-2) + 4\xi(t-3) - \lambda\xi(t-4)\} \tag{15}$$

which is a rearrangement of (13). The reason is that this difference equation is unstable. The roots of the operator $\varphi(L)$ consist of the roots of $\mu(L)$ and their reciprocals which are the roots of $\mu(F)$. If the roots of $\mu(L)$ lie outside the unit circle, in fulfilment of the condition for the stability of an ordinary difference equation, then the roots of $\mu(F)$ will assume values inside the unit circle which violate the stability condition.

In order to generate the values of the trend sequence $\xi(t)$ from those of the sequence $y(t)$, it is necessary to pursue two separate recursions. The first recursion, which runs forwards in time, finds the values of the sequence $q(t) = \mu^{-1}(L)y(t)$ via the equation

$$q(t) = \frac{1}{\mu_0}\{y(t) - \mu_1 q(t-1) - \mu_2 q(t-2)\}. \tag{16}$$

This process of generating $q(t)$ is the analogue of the first stage of the solution of the equations under (7). The second recursion, which runs backwards in time, finds the values of the sequence $\xi(t) = \mu^{-1}(F)q(t)$ via the equation

$$\xi(t) = \frac{1}{\mu_0}\{q(t) - \mu_1\xi(t+1) - \mu_2\xi(t+2)\}. \tag{17}$$

This process is the analogue of the second stage of the solution of the equations under (7).