Adv. & Appl. in Stat. **3**(1) (2003), 15-32

AUTOMATIC CONTINUOUS SMOOTHING

ERHARD RESCHENHOFER and LAKDERE BENKHEROUF

Department of Statistics and Decision Support Systems
University of Vienna, A-1010 Vienna, Universitätsstr. 5, Austria

Department of Mathematics and Statistics
Sultan Qaboos University, Al-Khod 123, Sultanate of Oman

Abstract

This paper proposes an efficient method for automatic continuous smoothing, which is based on the determination of the trade-off between the goodness of fit and the danger of over-fitting. The usefulness of this method is demonstrated by applying it to economic time series. It is also shown how this method can be used for spectral density estimation. Algorithms are given for the automatic selection of the degree of smoothing as well as for the carrying out of the smoothing.

1. Introduction

Smoothing methods are indispensable tools for the analysis of time series. The first step in a conventional analysis is to remove the trend of the time series. This can be achieved either by differencing or by first estimating the trend and then subtracting it. To estimate the trend we may compute weighted averages of adjacent values (nonparametric smoothing) or fit suitable approximating functions (e.g., polynomials) to the data (parametric smoothing). Once the trend has been removed, the next step is to describe the dependence among the values at different times (serial dependence). Assuming that the detrended series is roughly

2000 Mathematics Subject Classification: 62G05, 62P20.

Key words and phrases: penalized least squares estimators, Hodrick-Prescott filter, model selection criteria, nonparametric spectral density estimation.

Received December 2, 2002

stationary and restricting attention to the first- and second-order moments, we can use the spectral density to describe the serial dependence. Estimates of the spectral density can be obtained by parametric or nonparametric smoothing of its sample analogue, the periodogram.

Both in case of nonparametric smoothing and in case of parametric smoothing, the choice of the degree of smoothing is crucial. In the former case, the degree of smoothing is determined by the number of values included in the weighted averages as well as the size of the weights and in the latter case by the form of the approximating function and the number of its parameters. Normally, the choice of the number of values included in the averages and the number of parameters of the approximating function, respectively, is far more important than the choice of the weights and the form of the approximating function, respectively. However, there are exceptions to this rule. For example, the widely used method of smoothing a time series $y_1, ..., y_n$ by minimizing

$$\sum_{t=1}^{n} (y_t - g_t)^2 + \lambda \sum_{t=3}^{n} ((g_t - g_{t-1}) - (g_{t-1} - g_{t-2}))^2 \qquad (1)$$

with respect to $g_1, ..., g_n$ yields a solution, where each g_t is a weighted average of all data. Here the degree of smoothing depends only on the weights, which in turn are determined by the parameter λ. The larger the value of λ, the smoother is the solution. If $g_t = a + bt$, the second sum in (1) vanishes. Thus the solution will approach a linear time trend as λ approaches infinity. This method has a long history of use, particularly in the actuarial sciences (see, e.g., Whittaker [16]). It was propagated by Hodrick and Prescott [5] as a useful tool for trend estimation and is therefore called *Hodrick-Prescott filter* by economists. To a certain extent the popularity of the Hodrick-Prescott filter is due to the fact that Hodrick and Prescott have proposed particular values of λ for various situations, e.g., $\lambda = 1{,}600$ for log quarterly data. In general, there are no clear rules for the choice of the degree of smoothing. In case of nonparametric smoothing, the degree of smoothing is typically chosen by subjective judgment of the investigator. In case of parametric smoothing,

the number of parameters of the fitted function is frequently chosen with the help of an "objective" decision rule like the Akaike information criterion (AIC; Akaike [2]). Unfortunately, the use of a particular decision rule does not really guarantee objectivity since we may obtain completely different results depending on which class of approximating functions (polynomials, rational functions, etc.) we use. Moreover, since there are not only many different classes of approximating functions but also many different decision rules (see, e.g., Schwarz [11], Sawa [10], Sugiura [13], Reschenhofer [9]) there is still the problem of how to select the decision rule.

To illustrate the fact that we should not blindly trust any data-independent rule for the choice of the degree of smoothing, we look at the seasonally adjusted quarterly real US GDP from 1947.1 to 2001.4 (see Figure 1). The data were downloaded from the website of the St. Louis Fed's FRED® Database. Applying the Hodrick-Prescott filter with the recommended value of $\lambda = 1{,}600$ to this time series, we obtain a rather changeable trend estimate. Figure 1 suggests that the growth rate changes every five to ten years. Normally, fluctuations with such a small wavelength are rather considered as business cycles. Only very slowly changing functions like simple polynomials or exponential functions are considered as appropriate for the description of a trend. It seems that we have to use a much larger value of the smoothing parameter. Indeed, the trend estimate obtained with $\lambda = 100{,}000$ looks much steadier (see Figure 1). This curve resembles a broken linear trend. The slowdown in growth follows the oil-price shock in early 1973 (see also Perron [6, 7]). Clearly, if we increase the smoothing parameter further, we will end up with an unbroken linear trend. To decide which type of trend is most plausible in light of the data we would need an objective, data-driven rule for choosing λ.

In general, unless strong restrictions are imposed on the data generating process, it is hopeless to try to determine the true nature of the trend of a macroeconomic time series like the US GDP. In the end it comes down to a matter of taste, whether we prefer a model with a simple deterministic trend and long-term dependence, a model with a more complex deterministic trend and medium-term dependence, or a model

with a stochastic trend and short-term dependence. The nastiness of this modeling problem stems mainly from the fact that we have to deal with the trend and the serial dependence simultaneously. However, when we switch from quarterly data to annual data the situation improves immediately. The first differences of the annual log GDP exhibit no obvious serial dependence (Durlauf [3]). Moreover, since possible shifts in the mean are orders of magnitude smaller than the variability of the annual changes, the assumption of serial independence is meaningful regardless of whether or not there is a break in the trend.

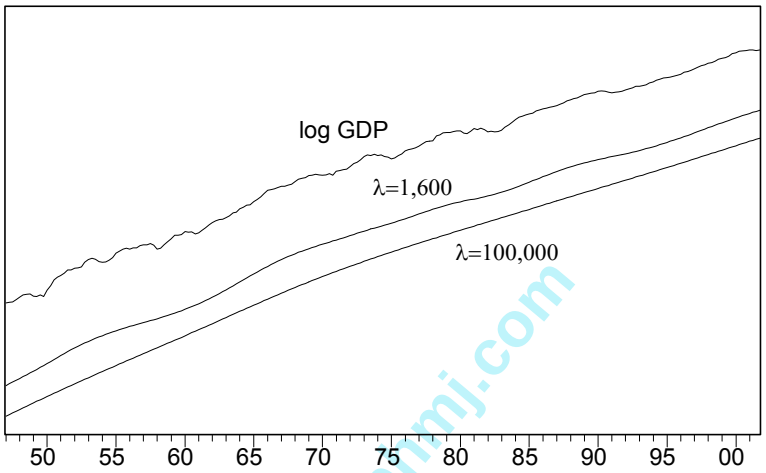

Figure 1. Smoothing the log quarterly GDP with the Hodrick-Prescott filter using two different values of the smoothing parameter λ

The Hodrick-Prescott filter has been designed for the smoothing of a series that exhibits a roughly linear increase. If we want to smooth the differenced log US GDP, we need another method. This method could then also be used for spectral density estimation. The next section describes the analogue of the Hodrick-Prescott filter for series exhibiting no clear trend and a method for the automatic determination of the smoothing parameter. All computational details are given in Appendices A and B. In Sections 3 and 4, the methodology is applied to real data. First the growth component of the log annual US GDP is estimated and

then the spectral density of the differenced log quarterly US GDP is estimated. Section 5 concludes.

2. Methods

Suppose $y_1, ..., y_n$ is a sequence of uncorrelated random variables with common variance σ^2. A linear smoother $g = Zy$ of the random vector $y = (y_1, ..., y_n)'$ is obtained by minimizing

$$\sum_{t=1}^{n} (y_t - g_t)^2 + \lambda \sum_{t=2}^{n} (g_t - g_{t-1})^2 \tag{2}$$

with respect to $g_1, ..., g_n$ (see Appendix A). The smaller the value of λ, the better is the fit and the greater is the danger of over-fitting. In the extreme case of $\lambda = 0$ there is no smoothing at all. Vice versa, the larger the value of λ, the smoother is the solution. If $g_1 = g_2 = \cdots = g_n$, the second sum in (2) vanishes, hence the solution will approach a constant function as λ approaches infinity.

For the choice of the continuous smoothing parameter λ we need an automatic decision method. Perhaps the most popular decision method is the AIC. Normally the AIC is used for model selection. Given a sample of size n, the AIC measures the discrepancy between the true model and an approximating model by minus two times the log likelihood plus two times the number of estimated parameters. The first term measures the goodness of fit and the second term serves to penalize over-fitting. The AIC values are calculated for all candidate models and the model with the smallest AIC value is chosen. At first glance it is hard to see how the AIC can be used in a nonparametric setting. Clearly, we cannot use the number of estimated parameters to penalize over-fitting in nonparametric smoothing because there are no estimated parameters. We rather have to go back to the rationale behind the AIC (or, more precisely, a precursor of the AIC; see Akaike [1]). Suppose that $y = \mu + u$ and

$y^* = \mu + u^*$ are independent samples from a distribution with mean vector μ and covariance matrix $\sigma^2 I$ and $g = Zy$ is an unbiased linear estimator for μ based on sample y only. Then it follows from

$$E(y - g)'(y - g) = E(u - Zu)'(u - Zu) = Eu'(I - Z)'(I - Z)u$$

$$= tr\{(I - Z' - Z + Z'Z)Euu'\} = \sigma^2\{n - 2tr(Z) + tr(Z'Z)\}$$

and

$$E(y^* - g)'(y^* - g) = E(u^* - Zu)'(u^* - Zu) = Eu^{*'}u^* + Eu'Z'Zu$$

$$= \sigma^2\{n + tr(Z'Z)\}$$

that the sum of squared residuals

$$S = (y - g)'(y - g)$$

is a biased estimator of the mean square prediction error

$$E(y^* - g)'(y^* - g).$$

An unbiased estimator is given by

$$S\{n - 2tr(Z) + tr(Z'Z)\}^{-1}\{n + tr(Z'Z)\}$$

$$= S(1 + 2tr(Z)\{n - 2tr(Z) + tr(Z'Z)\}^{-1}). \quad (3)$$

If g is obtained by projecting y onto the k-dimensional subspace generated by the columns of a design matrix X, i.e., $Z = X(X'X)^{-1}X'$, the unbiased estimator (3) becomes

$$S(1 + 2k(n - 2k + k)^{-1}) = S(1 + 2k(n - k)^{-1}). \quad (4)$$

The logarithm of expression (4) multiplied by n is approximately equal to

$$n \log(S) + 2k,$$

which differs from the AIC only by an additive constant (in the Gaussian

case). Analogously, n times the logarithm of (3) approximately equals

$$n \log(S) + 2tr(Z). \tag{5}$$

Thus we may select the continuous smoothing parameter λ in (2) by minimizing (5) with respect to λ. Of course, the trace of the smoothing matrix can be used to penalize over-fitting only in the case of a linear smoother (Hastie and Tibshirani [4], Wahba [15]). For the nonlinear case, Ye [17] developed the concept of generalized degrees of freedom (see also Shen and Ye [12]). Unfortunately, both the proper definition and the estimation of the generalized degrees of freedom are nontrivial tasks. But even in the comparatively simple linear case we have to calculate S and $tr(Z)$ repeatedly for many different values of the smoothing parameters. We therefore need algorithms for the efficient computation of these quantities. In particular, this is true in the case of a continuous smoothing parameter. For the special case where Z is obtained by minimization of (2), algorithms for the computation of S and $tr(Z)$ are given in Appendices A and B, respectively.

The plausibility of the derivation of selection criterion (5) is somewhat reduced by the fact that we had to assume that $Eg = \mu$. It is therefore reassuring to know (see Reschenhofer [8]) that model selection criteria derived under the assumption that the models under consideration are true models (Akaike [2], Sugiura [13]) very often select the same models as the corresponding criteria derived without this problematic assumption (Sawa [10], Reschenhofer [9]).

3. Trend Estimation

For the first differences of the log annual US GDP, there is no indication that criterion (5) reaches an absolute minimum at some finite value of λ. It keeps decreasing even for values as large as $\lambda = 10^{15}$. Figure 2 shows that already for much smaller values the estimated mean is practically constant. We conclude that there is little evidence of a structural break.

The special value of this result is due to the fact that it has been obtained without specifying the number of possible breaks and their locations in advance. In contrast, conventional methods for the detection of structural breaks typically require some prior information. Moreover, the results obtained with these methods depend crucially on the prior information. For example, Zivot and Andrews [18] estimated a single breakpoint in several macroeconomic time series and found less evidence against a stochastic trend than Perron [6] who fixed the breakpoint at the 1973 oil-price shock. Of course, the situation gets worse as the number of possible breakpoints increases. In the case of multiple breaks the simultaneous estimation of both the number of breaks and their locations is computationally infeasible for medium and large sample sizes (see, e.g., Sullivan [14]). Moreover, the performance of such estimation methods would be considerably reduced if there were smooth transitions rather than sharp breaks.

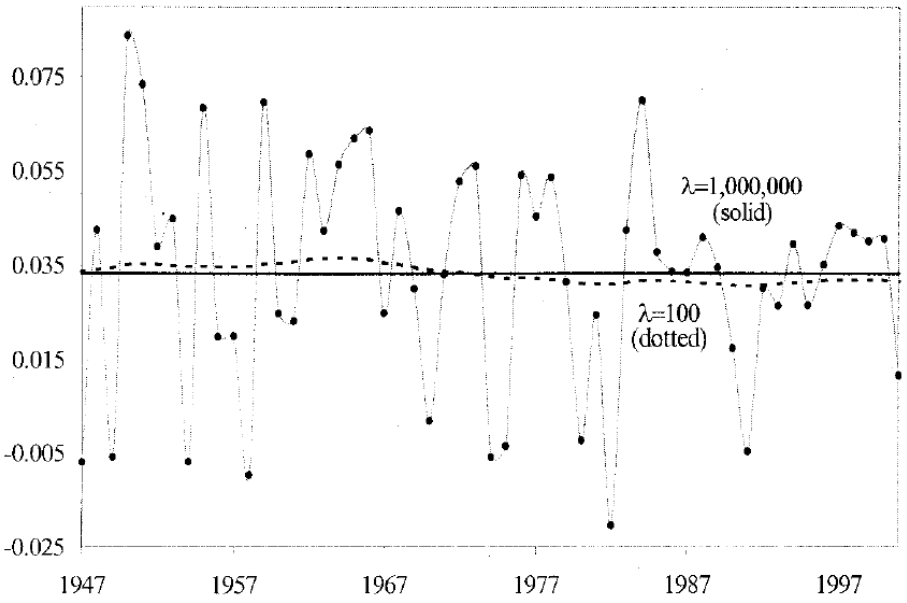

Figure 2. Smoothing the first differences of the log annual GDP with a penalized least squares method using two different values of the smoothing parameter λ

4. Spectral Density Estimation

We model the differenced log quarterly US GDP as a linear process $y_t = \sum_{j \in Z} \psi_j \varepsilon_{t-j}$ with absolutely summable impulse response coefficients ψ_j and i.i.d. innovations ε_t with $E\varepsilon_t = 0$ and $E\varepsilon_t^2 = \sigma^2$. Under these conditions, the process (y_t) has an absolutely summable autocovariance function

$$\gamma(k) = Cov(y_t, y_{t+k}) = Ey_t y_{t+k}, \quad k = 0, \pm 1, \pm 2, \pm 3, \ldots$$

and a continuous spectral density given by

$$f(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(k) \exp(-ik\omega) = \frac{1}{2\pi} \left\{ \gamma(0) + 2 \sum_{j=1}^{\infty} \gamma(k) \cos(k\omega) \right\}.$$

From n observations $y_1, \ldots, y_n$ we can estimate the autocovariances $\gamma(k)$ by the sample autocovariances $c(k)$, which are defined by

$$c(k) = \frac{1}{n} \sum_{j=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y})$$

for $0 \le k < n$ and $c(k) = c(-k)$ for $-n < k < 0$. Using n^{-1} rather than $(n-k)^{-1}$ or $(n-k-1)^{-1}$ in this definition ensures that $c(k)$ is non-negative definite. The periodogram of $y_1, \ldots, y_n$ is defined by

$$I(\omega) = \frac{1}{2\pi n} \left| \sum_{j=1}^{n} y_t \exp(-it\omega) \right|^2.$$

The sum of squared observations can be decomposed into a sum of periodogram ordinates associated with the Fourier frequencies $\omega_j = 2\pi j/n$, i.e.,

$$\sum_{j=1}^{n} y_t^2 = \sum_{j \in F(n)} I(\omega_j),$$

where $F(n) = \{-[(n-1)/2], \ldots, [n/2]\}$. For any non-zero Fourier frequency

ω_j, we have

$$I(\omega_j) = \frac{1}{2\pi}\sum_{k=-n}^{n} c(k)\exp(-ik\omega_j) = \frac{1}{2\pi}\left\{c(0) + 2\frac{1}{2\pi}\sum_{k=1}^{n} c(k)\cos(k\omega_j)\right\}.$$

Because of the resemblance between $f(\omega_j)$ and $I(\omega_j)$ we may use $I(\omega)$ for the estimation of $f(\omega)$. If $0 < \nu_1 < \cdots < \nu_j < \pi$ and $f(\omega) > 0$ for all ω, then the random vector $(I(\nu_1), ..., I(\nu_j))'$ converges in distribution to a vector of independent and exponentially distributed random variables with means $f(\nu_1), ..., f(\nu_j)$. This implies that the periodogram is a very erratic estimator of the spectral density. Assuming that f is quite smooth, we can easily construct a superior estimator by smoothing the periodogram. Usually, we consider weighted averages of neighboring periodogram ordinates:

$$\hat{f}(\omega_j) = \sum_{k=-r}^{r} w_k I(\omega_{j+k}).$$

As the number of involved ordinates, r, increases, the variance of the estimator decreases, but at the same time the bias increases. In general, the choice of the smoothing parameter r is more critical than the choice of the weights w_k.

Alternative estimates of the spectral density can be obtained by modeling the process (y_t) as an autoregressive (AR) process

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

or an autoregressive moving average (ARMA) process

$$y_t - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

estimating the model parameters by maximum likelihood (ML), and finally plugging the ML estimates into the formulas for the spectral

densities implied by these processes, i.e.,

$$f(\omega) = \frac{\sigma^2}{2\pi}|1 - \phi_1 \exp(-i\omega) - \cdots - \phi_p \exp(-i\omega p)|^{-2} \tag{6}$$

and

$$f(\omega) = \frac{\sigma^2}{2\pi}|1 + \theta_1 \exp(-i\omega) + \cdots + \theta_q \exp(-i\omega p)|^2$$
$$|1 - \phi_1 \exp(-i\omega) - \cdots - \phi_p \exp(-i\omega p)|^{-2}, \tag{7}$$

respectively. Estimates of the model parameters can also be obtained by fitting the inverse trigonometric polynomial (6) and the trigonometric rational function (7), respectively, to the periodogram.

Finally, the smoothing procedure described in Section 2 can also be used to estimate the spectral density. However, since this methodology is based on additive errors rather than on multiplicative errors, it must be applied to the log periodogram. For the Fourier frequencies $\omega_j \in (0, \pi)$, the log periodogram can be written as

$$\log(I(\omega_j)) = \log(f(\omega_j)) + \log(v_j)$$
$$= [\log(f(\omega_j)) + E(\log(v_j))] + [\log(v_j) - E(\log(v_j))],$$

where the random variables v_j, $0 < j < n/2$, approximately are independent and have a standard exponential distribution. The expectation of the log of a standard exponential random variable equals minus Euler's constant ($\gamma = 0.57721...$), hence an estimate of the log spectral density can be obtained by adding γ to the smoothed log periodogram. Figure 3 shows various estimates of the spectral density function obtained with different values of the continuous smoothing parameter λ. Of particular interest for economists is whether or not there is a steep decline of the spectral density near frequency zero. This is due to the fact that the value of the spectral density at frequency zero is a widely used measure for the size of the stochastic trend component. The result obtained with the automatically chosen λ suggests that this component is relatively large.

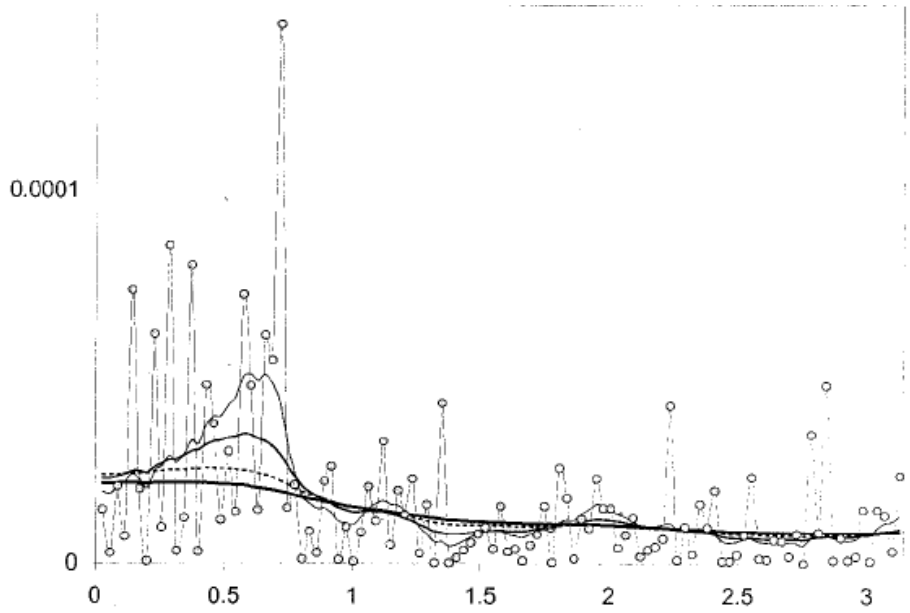

Figure 3. The periodogram of the differenced log quarterly US GDP is plotted together with estimates of the spectral density that have been obtained by smoothing the log periodogram with different values of the smoothing parameter ($\lambda = 10$: fine, $\lambda = 50$: medium, $\lambda = 500$: fat, automatic choice : dotted)

5. Concluding Remarks

In nonparametric trend estimation and nonparametric spectral density estimation, applied workers usually determine the degree of smoothing in a subjective way. This is mainly due to the fact that standard programs for time series analysis typically compute values of the AIC (and other decision criteria) for parametric models only, but have no built-in methods for choosing the optimal degree of nonparametric smoothing. Thus there is obviously a need for simple and computationally efficient automatic smoothing methods that can be implemented also by persons with moderate programming skills. The methods described in this paper meet these requirements.

References

[1] H. Akaike, Fitting autoregressive models for prediction, Ann. Inst. Stat. Math. 21 (1969), 243-247.

[2] H. Akaike, Information theory and an extension of the maximum likelihood principle, Second International Symposium on Information Theory, B. N. Petrov and F. Csaki, eds., Akademia Kiado, Budapest, 1973, pp. 267-281.

[3] S. N. Durlauf, Spectral based testing of the martingale hypothesis, J. Econometrics 50 (1991), 355-376.

[4] T. J. Hastie and R. J. Tibshirani, Generalized Additive Models, Chapman and Hall, London, 1990.

[5] R. J. Hodrick and E. C. Prescott, Postwar U.S. business cycles: an empirical investigation, J. Money, Credit, and Banking 29 (1997), 1-16.

[6] P. Perron, The great crash, the oil price shock, and the unit root hypothesis, Econometrica 57 (1989), 1361-1401.

[7] P. Perron, Further evidence on breaking trend functions in macroeconomic variables, J. Econometrics 80 (1997), 355-385.

[8] E. Reschenhofer, Some remarks on the entropy maximization principle, Estadística 49-51 (1997-99), 131-165.

[9] E. Reschenhofer, Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification, Econometric Theory 15 (1999), 377-387.

[10] T. Sawa, Information criteria for discriminating among alternative regression models, Econometrica 46 (1978), 1273-1291.

[11] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978), 461-464.

[12] X. Shen and J. Ye, Adaptive model selection, J. Amer. Stat. Assoc. 97 (2002), 210-221.

[13] N. Sugiura, Further analysis of the data by Akaike's information criterion and the finite corrections, Comm. Stat. A 7 (1978), 13-26.

[14] J. H. Sullivan, Estimating the locations of multiple change points in the mean, Comput. Stat. 17 (2002), 289-296.

[15] G. Wahba, Spline Models for Observational Data, SIAM, Philadelphia, 1990.

[16] E. T. Whittaker, On a new method of graduations, Proc. Edinburgh Math. Soc. 41 (1923), 63-75.

[17] J. Ye, On measuring and correcting the effects of data mining and model selection, J. Amer. Stat. Assoc. 93 (1998), 120-131.

[18] E. Zivot and W. K. Andrews, Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis, J. Business and Economic Stat. 10 (1992), 25-44.

Appendix A

To find the vector $g = (g_1, ..., g_n)'$ that minimizes

$$S(g \mid y) = \sum_{t=1}^{n} (g_t - y_t)^2 + \lambda \sum_{t=1}^{n-1} (g_{t+1} - g_t)^2,$$

we differentiate $S(g \mid y)$ partially with respect to $g_1, ..., g_n$ and set the derivatives equal to zero. We obtain the system of first-order conditions

$$\begin{aligned}
&(g_1 - y_1) && -\lambda(g_2 - g_1) && && = 0\\
&(g_2 - y_2) && +\lambda[(g_2 - g_1) && -(g_3 - g_2)] && = 0\\
&\quad\vdots && \quad\vdots && \quad\vdots && \vdots\\
&(g_{n-1} - y_{n-1}) && +\lambda[(g_{n-1} - y_{n-1}) && -(g_n - g_{n-1})] && = 0\\
&(g_n - y_n) && +\lambda(g_n - g_{n-1}) && && = 0.
\end{aligned}$$

In matrix notation the system can be written as $y = \Lambda g$, where

$$\Lambda = \begin{pmatrix}
1+\lambda & -\lambda & 0 & 0 & \cdots & 0 & 0 & 0\\
-\lambda & 1+2\lambda & -\lambda & 0 & \cdots & 0 & 0 & 0\\
0 & -\lambda & 1+2\lambda & -\lambda & \cdots & 0 & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\
0 & 0 & 0 & 0 & \cdots & -\lambda & 1+2\lambda & -\lambda\\
0 & 0 & 0 & 0 & \cdots & 0 & -\lambda & 1+\lambda
\end{pmatrix}.$$

The matrix Λ is positive definite because it is a diagonally dominant matrix, i.e., $\Lambda_{ii} > 0$ and $\Lambda_{ii} \geq \sum_{j \neq i} \Lambda_{ij}$ for all i. Hence, the Hessian matrix

$$\left(\frac{\partial^2 S(g \mid y)}{\partial g_i \partial g_j} \right) = 2\Lambda$$

is also positive definite.

A simple numerical method to find the solution $g = \Lambda^{-1}y$ is based on the cumulative first-order conditions

$$g_2 = g_1 + \frac{1}{\lambda}(g_1 - y_1),$$

$$g_3 = g_2 + \frac{1}{\lambda}[(g_1 - y_1) + (g_2 - y_2)],$$

$$\vdots$$

$$g_n = g_{n-1} + \frac{1}{\lambda}\sum_{t=1}^{n-1}(g_t - y_t),$$

$$\sum_{t=1}^{n} g_t = \sum_{t=1}^{n} y_t.$$

Obviously, g_2 is increasing in g_1, g_3 is increasing in g_1 and g_2, and so on. Hence, g_2, g_3, ..., g_n are increasing in g_1. Therefore, all we have to do is to select an appropriate starting value, e.g., $g_1(0) = y_1$, and to change this value iteratively until $|\sum y_t - \sum g_t(k)| \leq \varepsilon$. It must be increased if $\sum g_t(k) < \sum y_t - \varepsilon$ and decreased if $\sum g_t(k) > \sum y_t + \varepsilon$. Here $g_t(k)$ is the value of g_t at iteration k.

Appendix B

Pre-multiplying both sides of the equation $y = \Lambda g$ by

$$B = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix},$$

we get $By = Ag$ and $g = A^{-1}By$, where

$$A = B\Lambda = \begin{pmatrix} 1+\lambda & -\lambda & 0 & \cdots & 0 & 0 \\ 1 & 1+\lambda & -\lambda & \cdots & 0 & 0 \\ 1 & 1 & 1+\lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1+\lambda & -\lambda \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{pmatrix}.$$

Writing A as a partitioned matrix of the form

$$A = \begin{pmatrix} R & (0, ..., 0, -\lambda)' \\ (1, ..., 1) & 1 \end{pmatrix}$$

we obtain

$$A^{-1} = \begin{pmatrix} R^{-1} + R^{-1}(0, ..., 0, -\lambda)'\delta(1, ..., 1)R^{-1} & -R^{-1}(0, ..., 0, -\lambda)'\delta \\ -\delta(1, ..., 1)R^{-1} & \delta \end{pmatrix}$$

$$= \begin{pmatrix} Q - \lambda\delta Q_{n-1}s' & \lambda\delta Q_{n-1} \\ -\delta s' & \delta \end{pmatrix}$$

$$= \begin{pmatrix} q_{1,1} - \lambda\delta q_{1,n-1}s_1 & \cdots & q_{1,n-1} - \lambda\delta q_{1,n-1}s_{q-1} & \lambda\delta q_{1,n-1} \\ \vdots & & \vdots & \vdots \\ q_{n-1,1} - \lambda\delta q_{n-1,n-1}s_1 & \cdots & q_{n-1,n-1} - \lambda\delta q_{n-1,n-1}s_{n-1} & \lambda\delta q_{n-1,n-1} \\ -\delta s_1 & \cdots & -\delta s_{n-1} & \delta \end{pmatrix},$$

where $Q = R^{-1}$, $Q_k = (q_{1,k}, ..., q_{n-1,n-1})'$ is the k-th column of Q, s_k is the sum of all elements in Q_k, $s' = (s_1, ..., s_{n-1})$, and $\delta = \{1 - (1, ..., 1)Q(0, ..., 0, -\lambda)'\}^{-1} = (1 + \lambda s_{n-1})^{-1}$.

Thus,

$$tr(A^{-1}B) = \sum_{j=1}^{n-1} \{q_{j,j} + \cdots + q_{j,n-1} - \lambda\delta q_{j,n-1}(s_j + \cdots + s_{n-1})\}$$

$$+ \lambda\delta \sum_{j=1}^{n-1} q_{j,n-1} + \delta$$

$$= \xi - \lambda\delta \sum_{j=1}^{n-1} q_{j,n-1}(s_j + \cdots + s_{n-1}) + \lambda\delta s_{n-1} + \delta,$$

where

$$\xi = \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} q_{jk}.$$

All quantities occurring in the final expression for $tr(A^{-1}B)$ can be calculated recursively. For the derivation of the recursion formulas, we introduce the $n \times n$ matrix

$$R^* = \begin{pmatrix} R & (0, \ldots, 0, -\lambda)' \\ (1, \ldots, 1) & 1 + \lambda \end{pmatrix}.$$

The inverse of R^* takes the form

$$Q^* = \begin{pmatrix} q_{1,1} - \lambda\delta^* q_{1,n-1} s_1 & \cdots & q_{1,n-1} - \lambda\delta^* q_{1,n-1} s_{q-1} & \lambda\delta^* q_{1,n-1} \\ \vdots & & \vdots & \vdots \\ q_{n-1,1} - \lambda\delta^* q_{n-1,n-1} s_1 & \cdots & q_{n-1,n-1} - \lambda\delta^* q_{n-1,n-1} s_{n-1} & \lambda\delta^* q_{n-1,n-1} \\ -\delta^* s_1 & \cdots & -\delta^* s_{n-1} & \delta^* \end{pmatrix},$$

where

$$\delta^* = \{1 + \lambda - (1, \ldots, 1)Q(0, \ldots, 0, -\lambda)'\}^{-1} = (1 + \lambda + \lambda s_{n-1})^{-1}.$$

Thus,

$$q^*_{j,n} = \lambda\delta^* q_{j,n-1}, \quad j = 1, \ldots, n-1, \quad q^*_{n,n} = \delta^*,$$

$$s^*_j = s_j(1 - s^*_n), \quad j = 1, \ldots, n-1, \quad s^*_n = \lambda\delta^* s_{n-1} + \delta^*,$$

$$\xi^* = \xi - \lambda\delta^* \sum_{j=1}^{n-1} \sum_{k=j}^{n-1} q_{j,n-1} s_k + s^*_n.$$

■