The Canadian Journal of Statistics
Vol. 29, No. 1, 2001, Pages 37–50
La revue canadienne de statistique

Nonparametric smoothing using state space techniques

Patrick E. BROWN and Piet de JONG

Key words and phrases: Diffuse random vectors; favoured model; Kalman filter smoother; locally best invariant tests; penalized least squares; Reinsch algorithm; smoothations; spline smoothing.

MSC 2000: Primary 62G08, 62M10.

Abstract: The authors examine the equivalence between penalized least squares and state space smoothing using random vectors with infinite variance. They show that despite infinite variance, many time series techniques for estimation, significance testing, and diagnostics can be used. The Kalman filter can be used to fit penalized least squares models, computing the smoothed quantities and related values. Infinite variance is equivalent to differencing to stationarity, and to adding explanatory variables. The authors examine constructs called "smoothations" which they show to be fundamental in smoothing. Applications illustrate concepts and methods.

De l'emploi de techniques afférentes à l'espaces d'états aux fins de lissage non paramétrique

Résumé : Les auteurs examinent l'équivalence entre les moindres carrés pénalisés et le lissage de l'espace d'états au moyen de vecteurs aléatoires à variance infinie. Ils montrent que malgré le problème de variance infinie, plusieurs techniques de diagnostic, d'estimation et de test de signification propres aux chroniques restent valables. Le filtre de Kalman permet d'évaluer les modèles des moindres carrés pénalisés en fournissant entre autres des valeurs lissées. La variance infinie est équivalente à la différenciation à des fins de stationnarité et à l'ajout de variables explicatives. Les auteurs étudient en outre des quantités appelées "lissations," dont ils montrent l'importance pour le lissage. Des applications illustrent les méthodes et procédures décrites.

1. INTRODUCTION AND MOTIVATING EXAMPLES

Figures 1 and 2 display two data sets illustrating smoothing based on the state space model. The rough curve in each panel is the data (x_i, y_i), $i = 1, \ldots, n$. The smooth middle curve is the smoothed data, with upper and lower smooth curves defining confidence bands. The derivations of these curves, their properties and appropriateness, as well as associated diagnostics, are discussed in this article.

State space methods, including smoothing, are common in time series analysis (Anderson & Moore 1979; Harvey 1989). But neither of the data sets in Figure 1 or 2 are time series. The first panel displays $n = 150$ human measurements where y_i is the electrical voltage in the chewing muscle corresponding to a given force x_i exercised by the muscle. The relation appears non-linear and heteroscedastic. These data are analyzed with kernel smoothing in Müller (1988) and with a nonparametric quasi-likelihood method assuming a known functional relation in Chiou & Müller (1999). Figure 2 displays the "draft lottery" data set discussed in Fienberg (1971) consisting of $n = 366$ draw numbers y_i in the 1970 US draft lottery versus the actual birthday $x_i = i$. Notice that the smooth for the dental voltages is noticeably "kinked" while that for the draft lottery numbers moves, mainly down, by 12 discrete steps corresponding to different months. The steps suggest lack of randomness.

Nonparametric smoothing for non time series typically uses techniques such as penalized least squares (PLS), cubic splines (Silverman 1985; Green & Silverman 1994) L-splines, kernel methods or local polynomial smoothing. These methods have developed independently from time series methods. Several authors have pointed out the equivalence of PLS and smoothing based on the state space model (Wahba 1978; Weinert, Byrd & Sidhu 1980; Wecker & Ansley 1983; Kohn & Ansley 1989; Wahba 1990; Harvey & Stock 1994). This equivalence is displayed using stochastic differential equations. Ansley, Kohn & Wong (1993) and more recently Shively, Kohn & Wood (1999) apply state space techniques for spline regression. Although these approaches and derivations are general, elegant and have a firm statistical basis, they have failed to make the deserved impact.

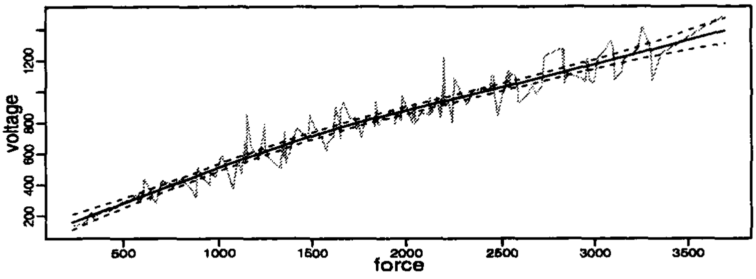

FIGURE 1: Dental data. The grey jagged line is the electrical voltage in chewing muscles versus force. The solid black line is the smooth with the dashed lines the associated confidence bands.

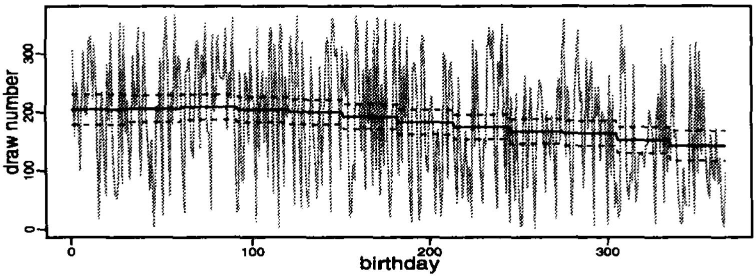

FIGURE 2: Draw number versus birthday number for the 1970 U.S. draft lottery. Step functions are the smooth and associated confidence bands.

One aim of this article to display the transparent and direct connection between the methods of time series analysis and state space smoothing to nonparametric smoothing, without resort to stochastic differential equations. For example, it is shown that cubic spline smoothing is implicitly based on an ARIMA(0,2,1) model with MA coefficient of $2 - \sqrt{3}$. Given the transparent connection between time series methods and nonparametric smoothing, it is easy to generalize time series methods to more general smoothing problems, and examples of this are given throughout this article.

Penalized least squares problems are typically solved using the algorithm due to Reinsch (1967). State space smoothing is efficiently carried out using the Kalman Filter Smoother (KFS), the modern form of which properly deals with "diffuseness" (Ansley & Kohn 1985; de Jong 1991) and streamlines the smoothing pass (de Jong 1988, 1989; Kohn & Ansley 1989). The KFS includes the Reinsch algorithm as a special case. The latter automatically handles "diffuseness" or "nonstationarity" by "differencing to stationarity". Diffuseness and differencing to stationarity are discussed in Sections 2 and 5. The Reinsch algorithm computes smooths via quantities here called "smoothations". The general properties and uses of smoothations are explored in Section 4.

Heckman & Ramsay (2000) express the "favoured" aspects of a model and fit using PLS constructs. By using the connection between PLS and state space smoothing, these constructs can be translated to statements about the underlying state space model. Different aspects of a PLS penalty matrix are made explicit by considering the statistical representation of the model. This is taken up in Section 6.

2. SMOOTHING METHODS AND MODELS

Different approaches to smoothing are connected as discussed in the following subsections. These connections are partly known and scattered in the literature and it is worthwhile to put them on a unified and practical footing.

2.1. *Penalized least squares.*

A Penalized Least Squares (PLS) smooth is a function which minimizes the PLS criterion. If $\phi(\partial)s(x)$ is a linear combination of derivatives of $s(x)$, where the linear combination may depend on x, then the PLS criterion is (Silverman 1985; Green & Silverman 1994)

$$\sum_{i=1}^{n}\{y_i - s(x_i)\}^2 + \lambda^2 \int_{-\infty}^{\infty} \{\phi(\partial)s(x)\}^2\, dx. \tag{1}$$

The second term ensures that the minimizer $s(x)$ of the PLS criterion is smooth in the sense of $\phi(\partial)s(x) \approx 0$, with smoothness increasingly important as λ becomes large.

It is well known that the minimizer $s(x)$ of (1) is an L-spline, and the derivatives of L-splines can be computed as a linear combination of differences of $s(x)$. Given this, (1) can be written in matrix form. Let y and s be the vectors with components y_i and $s(x_i)$ respectively, $i = 1, \ldots, n$. Then (1) simplifies to

$$(y-s)'(y-s) + \lambda^2 s'D'R^{-1}Ds, \tag{2}$$

where D is a differencing matrix and R serves to "correct" the differences and produce a vector of derivatives. Green & Silverman (1994) show the forms of D and R when $\phi(\partial)s(x)$ is the second derivative. Minimization criteria of the form (2) were originally considered by Whittaker (1923). Akaike (1980) considers the minimization of (2) in a Bayesian, equal spacing time series setting. In time series and when $R = I$, the resulting smoothing computations are known as the Hodrick–Prescott filter (Harvey 1989).

Differentiation (2) with respect to s and equating to zero yields

$$\hat{s} \equiv (I + \lambda^2 D'R^{-1}D)^{-1}y = y - \lambda^2 D'(R + \lambda^2 DD')^{-1}Dy, \tag{3}$$

where the equality follows from matrix inversion identities and holds for general R and D. The final expression is the basis for the algorithm of Reinsch (1967), which exploits the banded structure of R and DD'. The final term on the right in (3) is the smoothing error $y - \hat{s}$. Smoothing errors divided by λ^2 are here called "smoothations" and their appropriate generalization is considered in Section 4.

2.2. *The signal plus noise model.*

Penalized least squares is equivalent to linear prediction assuming a signal plus noise model.

DEFINITION 1. *The signal plus noise model is $y = s + \lambda\eta$, where λ is a scalar and s and η are random vectors with $\eta \sim (0, \sigma^2 I)$ and* $\mathrm{cov}\,(s, \eta) = 0$.

This formulation is closely related to one given by Silverman (1985) except that here s is assumed random. Note that $\mathrm{cov}\,(s)$ is left unspecified which fits in with subsequent developments where $\mathrm{cov}\,(s)$ diverges in which case s is said to be diffuse. Diffuse random vectors occur implicitly or explicitly in many aspects of smoothing.

DEFINITION 2. *A random vector s is diffuse if $\{\mathrm{cov}\,(s)\}^{-1}$ is singular. It is fully diffuse if $\{\mathrm{cov}\,(s)\}^{-1} = 0$ and partially diffuse otherwise, in which case $\{\mathrm{cov}\,(s)\}^{-1}$ is of the form $\sigma^{-2}D'R^{-1}D$, where D is of full row rank and R is nonsingular.*

Least squares, in essence, deals with transformations $X\beta$ of fully diffuse random vectors β, where X is of full column rank. In this case, neither $\mathrm{cov}\,(X\beta)$ nor $\{\mathrm{cov}\,(X\beta)\}^{-1}$ is well defined. This is taken up in Subsection 4.3 and Section 6.

THEOREM 1. *If y is generated by the signal plus noise model, then*

$$\begin{aligned} \mathrm{E}(s \mid y) &= \mathrm{E}(s) + [I + \lambda^2\sigma^2 \{\mathrm{cov}(s)\}^{-1}]^{-1}\{y - \mathrm{E}(s)\}, \\ \mathrm{cov}(s \mid y) &\equiv \mathrm{cov}\{s - \mathrm{E}(s \mid y)\} = \lambda^2\sigma^2[I + \lambda^2\sigma^2 \{\mathrm{cov}(s)\}^{-1}]^{-1}. \end{aligned}$$

Further if s is diffuse, $\{\mathrm{cov}(s)\}^{-1} = \sigma^{-2}D'R^{-1}D$, and $D\mathrm{E}(s) = 0$, then

$$\mathrm{E}(s \mid y) = \hat{s}, \quad \mathrm{cov}(s \mid y) = \lambda^2\sigma^2\{I - \lambda^2 D'(R + \lambda^2 DD')^{-1}D\},$$

where $\hat{s}$ is the PLS solution (3).

Throughout this paper, the conditional expectation and covariance notation is used in the Best Linear Unbiased Prediction (BLUP) sense. Proofs are given in the Appendix. Note that $\mathrm{E}(s \mid y)$ is invariant to σ^2.

Thus PLS smoothing is acting as if y is generated by a signal plus noise model with a diffuse signal and computing $\mathrm{E}(s \mid y)$. Robinson (1991) discusses the equivalence between BLUP and Kalman filtering, and in the discussion after Robinson's paper, T. Speed points out the connection between BLUP and PLS.

THEOREM 2. *Suppose s is partially diffuse, $\{\mathrm{cov}(s)\}^{-1} = \sigma^{-2}D'R^{-1}D$, and X is of full column rank with columns that span the null space of D. Then*

$$\mathrm{cov}(s) = \lim_{\kappa \to \infty} \sigma^2(\Sigma + \kappa XX'), \quad \mathrm{cov}(Ds) = \sigma^2 R = \sigma^2 D\Sigma D',$$

and $DX = 0$.

The above result is useful in a variety of smoothing contexts as illustrated below.

2.3. Cubic spline model.

Cubic spline smoothing is implicitly based on a rigid and elementary time series model. It is achieved if in (1), $\phi(\partial)s(x)$ is the second derivative of $s(x)$. Suppose the ordering variable is equally spaced: $x_i = i$, $i = 1, \ldots, n$. Further put $\theta \equiv 2 - \sqrt{3}$ and consider sequentially uncorrelated random error terms $(\eta_i, \xi_i) \sim (0, \sigma^2 I)$ with

$$y_i = s_i + \lambda\eta_i, \quad \Delta^2 s_{i+1} \equiv s_{i+1} - 2s_i + s_{i-1} = \xi_i + \theta\xi_{i-1}, \tag{4}$$

where $(s_1, \Delta s_2)$ is assumed fully diffuse. Then $\mathrm{E}(s \mid y)$ is equivalent to the cubic spline smooth. It is natural to question the appropriateness of the second differences $\Delta^2 s_i$, the MA(1) model for these differences and the specific choice of $\theta \equiv 2 - \sqrt{3}$ for the MA(1) coefficient.

To show the equivalence, we observe that from (4) each component of s is a linear combination of the components of $\xi = (\xi_{-1}, \xi_0, \ldots, \xi_{n-1})$ and the fully diffuse $\alpha_1 = (s_1, \Delta s_2)$. If D is the familiar $(n-2) \times n$ double differencing matrix

$$D = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & 1 & -2 & 1 \end{pmatrix},$$

then $DX = 0$. From (4), the covariance matrix of Ds is $\sigma^2 R$, which is tri-diagonal with $\sigma^2(1 + \theta^2)$ on the diagonal and $\sigma^2\theta$ on the first off-diagonals.The smooth is invariant to σ^2, which can then be scaled to yield $\sigma^2(1 + \theta^2) = 2/3$ and $\sigma^2\theta = 1/6$. These are the forms of D and R for cubic spline smoothing (3) given equally spaced data (Green & Silverman 1994). Solving the two equations shows $\theta = 2 \pm \sqrt{3}$. Similar arguments show that general L-spline smoothing is equivalent to smoothing based on a more general ARIMA model (Mazzi 1997).

The results hold for non-equally spaced data. Envisage an x-grid with spacing determined by the number of significant digits in x_i. Then (4) is assumed to operate on this grid with there being many "missing" y_i values. These missing values do not affect $\mathrm{E}\,(s \mid y)$.

Not counting the scale σ, which does not affect the smooth, the cubic spline model appears to have a single unknown parameter λ. But two other parameters are estimated implicitly: the initial conditions $\alpha_1 = (s_1, \Delta s_2)$. In effect α_1 is estimated via Generalized Least Squares (GLS) with the estimate factored into the smooth. The GLS estimation of α_1 is equivalent to treating it as fully diffuse and calculating $\mathrm{E}\,(\alpha_1 \mid y)$.

As $\lambda \to 0$, it is supposed that increasingly precise measurements are made on the signal s_i. Conversely, as $\lambda \to \infty$ the state noise ξ_i is assumed to be relatively zero and hence $\Delta^2 s_i = 0$, or, in other words, the s_i lie on a straight line with intercept and slope $\alpha_1 = (s_1, \Delta s_2)$. All straight lines have the same penalty, zero, and in this limiting case the data y are used as the "tie breaker" to decide on the exact form of the straight line. Clearly, this is the least squares line. For cubic spline smoothing, the straight line is the "favoured model". This is further considered in Section 6.

2.4. State space models.

A state space form of the cubic spline model (4), defined in terms of the state vector $\alpha_i \equiv (s_i, \Delta s_{i+1})$ and error components $\varepsilon_i \equiv (\eta_i, \xi_i) \sim (0, \sigma^2 I)$, is

$$y_i = (\,1 \quad 0\,)\,\alpha_i + (\,\lambda \quad 0\,)\,\varepsilon_i, \quad \alpha_{i+1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \alpha_i + \begin{pmatrix} 0 & 1 \\ 0 & 1+\theta \end{pmatrix} \varepsilon_i, \tag{5}$$

where $\theta = 2 - \sqrt{3}$. The equations in (5) are special cases of the flexible state space form

$$y_i = Z_i \alpha_i + G_i \varepsilon_i, \quad \alpha_{i+1} = T_i \alpha_i + H_i \varepsilon_i, \tag{6}$$

where $i = 1, \ldots, n$ and $\varepsilon_i \sim (0, \sigma^2 I)$. The equations in (6) are called the measurement and state equation, respectively. The Z_i and G_i permit varying observation patterns and heteroscedasticity, while T_i and H_i can model varying spacing. The state equation spells out the correlational structure of adjacent, and by implication, all states α_i. The signal is $s_i \equiv Z_i \alpha_i$ and, as in (5), measurement and state error are often uncorrelated: $\mathrm{cov}\,(G_i \varepsilon_i, H_i \varepsilon_i) = \sigma^2 G_i H_i' = 0$.

The state space model (6) is completed with an appropriate assumption about α_1. Backsubstituting out the α_i in (6) for $i = 2, \ldots, n$ shows that each component of y is linear in α_1 and the error components $\varepsilon_i \sim (0, \sigma^2 I)$. Thus the only potential source of diffuseness is α_1.

With the state space model (6), the $\mathrm{E}\,(s_i \mid y)$s and related quantities such as $\mathrm{E}\,(\alpha_i \mid y)$ and $\mathrm{E}\,(\varepsilon_i \mid y)$ and associated error covariance matrices are computed with the KFS which includes the Reinsch algorithm as a special case. The KFS is a generalization, since PLS smoothing is acting as if the signal is generated by a specific state space model. An interesting feature of the Reinsch algorithm is that the data y are automatically "differenced to stationarity" as discussed in Section 5.

2.5. Least squares error minimization.

Both PLS and state space smoothing can be viewed as smoothing based on ordinary least squares subject to an observational constraint. The next theorem states this formally and is closely related to similar results appearing in the literature (Schoenberg 1964; Kimeldorf & Wahba 1971; Wahba 1978; Akaike 1980).

THEOREM 3. *Suppose $y = G\varepsilon$, where y and G are given and G is of full row rank. Then the minimum of $\sigma^{-2}\varepsilon'\varepsilon$ subject to $y = G\varepsilon$ is $y'\,\{\mathrm{cov}\,(y)\}^{-1}\,y$ achieved at $\varepsilon = E(\varepsilon \mid y)$, computed* *assuming* $\varepsilon \sim (0, \sigma^2 I)$. *Further if* $s = Z\varepsilon$ *and* $\operatorname{cov}\{(y, s)\}$ *has the same rank as* $\operatorname{cov}(\varepsilon)$, *then the minimization is equivalent to minimization over* s *of*

$$\{y - \mathrm{E}(y \mid s)\}' \{\operatorname{cov}(y \mid s)\}^{-1} \{y - \mathrm{E}(y \mid s)\} + s' \{\operatorname{cov}(s)\}^{-1} s,$$

in that the same minimum is achieved at $s = \mathrm{E}(s \mid y) = Z\varepsilon$.

The smooth $\mathrm{E}(s \mid y)$ depends on G which, in effect, spells out the covariance structure of the observations. The specification $y = G\varepsilon$ is attained with the state space model if the state equation in (6) holds for $i = 1$ with $\alpha_0 = 0$. The minimizing errors $\mathrm{E}(\varepsilon_i \mid y)$ and $\operatorname{cov}(\varepsilon_i \mid y)$ can be computed with the KFS provided G has a state space structure.

2.6. *Kernel smoothing.*

State space smoothing implicitly defines a smoothing kernel. Thus each $\mathrm{E}(s_i \mid y)$ is a linear combination of the components of y and the coefficients define the kernel. The kernel varies with i and automatically takes account of end effects. With the state space smoothing, the shape and bandwidth of the kernel are dictated by the model and parameter values. Specifying the kernel via a model and parameters seems more appealing than specifying the kernel directly. A model-based kernel may be computed for each i with the KFS algorithm but the kernel values are of limited use.

3. APPLICATIONS

3.1. *Dental voltages.*

Assume initially that the pressure x_i is equally spaced. A tentative model is the local linear trend model (Harvey 1989) couched in terms of "levels" and "slopes" μ_i and ν_i and parameters λ and ψ:

$$y_i = \mu_i + \lambda\eta_i, \quad \mu_{i+1} = \mu_i + \nu_i + \xi_i, \quad \nu_{i+1} = \nu_i + \psi\omega_i,$$

where $(\eta_i, \xi_i, \omega_i) \sim (0, \sigma^2 I)$. If $\omega_i \equiv \xi_i$, then a single error at each i is distributed across the level and slope using the weights 1 and ψ. The further constraint $\psi = 1 + \theta = 3 - \sqrt{3}$ corresponds to cubic spline smoothing. All three models approach the straight line model $y_i = \mu_0 + i\nu_0 + \lambda\eta_i$ as both $|\lambda|$ and $|\lambda/\psi|$ become large. When $\psi = 0$, the model is a random walk with fixed drift $\nu_0 = \cdots = \nu_n$, and observed with measurement noise.

The state space form of the above model for equally spaced data is

$$y_i = (1 \quad 0)\alpha_i + (\lambda \quad 0 \quad 0)\varepsilon_i, \quad \alpha_{i+1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\alpha_i + \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & \psi \end{pmatrix}\varepsilon_i. \tag{7}$$

The force variable x is integer-valued but not equally spaced. However the observed y_i can be regarded as a unit spaced series with many missing observations. If $h_i \equiv x_{i+1} - x_i$, then the state equation is iterated over the h_i steps implying $T_i = T^{h_i}$, where T is as in (7) and H_i is appropriately defined by iterating the state equation h_i times. If $h_i = 0$, then there is more than one measurement at x_i and $T_i = I$ and $H_i = 0$.

Maximizing the log-likelihood assuming normal error terms leads to maximum likelihood estimates $\lambda = 250$, $\psi = 0.0085$, $\sigma = 0.33$ and hence $\lambda\sigma = 82.83$. The corresponding smooth and associated normal based 95% confidence intervals are displayed in Figure 1. Imposing the cubic spline restriction yields $\lambda = 19508$, $\sigma = 0.0043$ and $\lambda\sigma = 83.00$, with practically the same value of the log-likelihood. The two models are not nested, precluding a formal significance test, but both $\lambda\sigma$ and the log-likelihood suggest the cubic spline restriction is reasonable.

3.2. *Lottery data.*

For these data, $x_i = i$, the birthday number, which are equally spaced from $i = 1$ to $n = 366$. In the lottery, birthdates had been inserted into capsules and poured into a box, one month at a time. Capsules for later months were put in last and remained near the top because of inadequate mixing. The monthly average draw number is thus suspected to vary, but within months the average is constant. Assume $y_i = s_i + \lambda\eta_i$ with $s_{i+1} = s_i$ unless i refers to the last day of a month, in which case $s_{i+1} = s_i + \xi_i$ with $(\eta_i, \xi_i) \sim (0, \sigma^2 I)$. Thus monthly averages are assumed to vary according to a random walk. These data are not normal. Nevertheless, the normal based log-likelihood was optimized with respect to the parameters yielding $\lambda = 8.23$ and $\sigma = 12.51$.

The smooth corresponding to these values of the parameters is displayed in Figure 2. Although not smooth in the traditional sense, the generally downward stepping estimate $\mathrm{E}(s_i \mid y)$ is a more cogent representation of the suspected nonrandomness than, for example, a smooth that varies from day to day. Further, the above analysis admits significance testing of the perceived differences in the s_i for different months. Thus $\mathrm{E}(s_1 - s_n \mid y)$ is estimated to be 61.89 and the estimates of $\mathrm{E}(s_1 \mid y)$ and $\mathrm{E}(s_n \mid y)$ have estimated standard deviations of 12.90 and 12.93, respectively. Assuming the two estimates are uncorrelated, the estimated standard deviation of the difference 61.89 is 18.26, suggesting a statistically significant difference.

4. SMOOTHATIONS

Smoothations occur in all aspects of smoothing. In terms of (3), the vector of smoothations is $D'(R + \lambda^2 DD')^{-1}Dy$. Thus with the Reinsch algorithm the smoothations are computed from Dy and multiplied by λ^2 to obtain the smoothing error, which is subtracted from y to obtain the actual smooth. A general statistical definition of smoothations is as follows.

DEFINITION 3. *If $y \sim (\mu, \sigma^2\Sigma)$, then the smoothations of y are the components of $u \equiv \Sigma^{-1}(y - \mu)$.*

Standardized smoothations $u_i/\sqrt{\mathrm{cov}(u_i)}$ corresponding to the dental data and fit, as discussed in Section 3, are displayed in Figure 3. Note that the definition covers the case where y is diffuse; i.e., $\{\mathrm{cov}(y)\}^{-1}$ is singular. Further, since $\mathrm{cov}(u) = \sigma^2\Sigma^{-1}$ then $M_i \equiv \sigma^{-2}\mathrm{cov}(u_i)$ is diagonal block i of Σ^{-1}. Both u_i and M_i are calculated with the KFS. Further properties and characterizations of smoothations are explored in the following subsections.

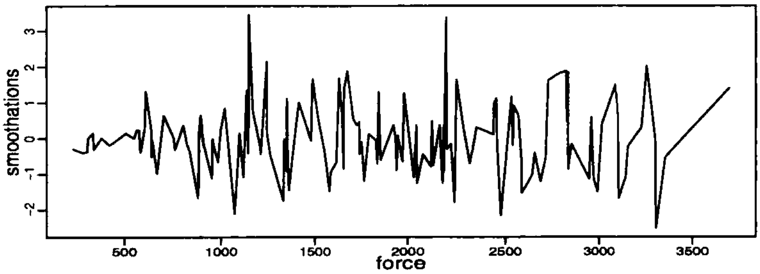

FIGURE 3: Standardized smoothations for the dental data.

4.1. *Smoothing error.*

If y is generated by the signal plus noise model, then smoothations are proportional to smoothing errors (De Jong 1988), i.e., $u_i = \lambda^{-2}\{y_i - \mathrm{E}(s_i \mid y)\}$, since

$$\begin{aligned} \mathrm{E}(s \mid y) &= \mathrm{E}(s) + \mathrm{cov}(s, y)\{\mathrm{cov}(y)\}^{-1}\{y - \mathrm{E}(y)\} = \mathrm{E}(s) + \sigma^{-2}\mathrm{cov}(s)\,u \\ &= \mathrm{E}\,s + \sigma^{-2}\{\mathrm{cov}(y) - \sigma^2\lambda^2 I\}u = y - \lambda^2 u. \end{aligned}$$

It also follows that $\mathrm{cov}(s_i \mid y) = \lambda^4\sigma^2 M_i$ and hence $u_i/(\sigma\sqrt{M_i})$ is the standardized smoothing error. The standardized smoothing errors reduce to the quantities suggested by Eubank (1985) in the context of PLS for the detection of responses which do not conform to the fit. The present definition of smoothation or smoothing error, however, is not tied to any particular model.

Figure 3 shows that for the dental data, two smoothing errors are significantly large and positive, warranting perhaps further investigation.

4.2. Deletion or jackknife residuals.

Smoothations are the standardized information in each observation, beyond that contained in all other observations (de Jong 1988). Thus smoothations are analogous to the innovations produced by the Kalman filter. Innovations are the new information in each observation, beyond that provided by the previous observations.

To explain this feature, put y^i as y excluding y_i. Then each scalar or vector component u_i of u has the characterization

$$y_i - \mathrm{E}\,(y_i \mid y^i) = M_i^{-1} u_i, \quad \mathrm{cov}\,(y_i \mid y^i) = \sigma^2 M_i^{-1}.$$

This follows from partitioned matrix inversion. Thus smoothations are standardized jackknife or interpolation residuals. Under the signal plus noise model, $\mathrm{E}\,(y_i \mid y^i) = \mathrm{E}\,(s_i \mid y^i)$ and hence, combining with the results of Subsection 4.1,

$$y_i - \mathrm{E}\,(s_i \mid y) = \lambda^2 M_i \{y_i - \mathrm{E}\,(s_i \mid y^i)\}.$$

Deletion residuals can be re-expressed in terms of the "hat" matrix discussed in Subsection 4.6 which is the formula for deletion residuals given in Green & Silverman (1994). However the present formulation and result applies to any state space setup.

4.3. Generalized least squares residuals.

The smoothations of y are identical to the smoothations of the GLS residuals, which result from subtracting the estimated diffuse component from y. This result generalizes the well known result from regression which states that studentized jackknife residuals are equal to the least squares residuals. To see this, suppose $\mathrm{cov}\,(y) = \sigma^2(\kappa X X' + \lambda^2 I)$ with $\kappa \to \infty$. Then $u = (\kappa X X' + \lambda^2 I)^{-1}\{y - \mathrm{E}\,(y)\}$ converges to

$$\lambda^{-2} D'(DD')^{-1} D\{y - \mathrm{E}\,(y)\} = \lambda^{-2}\{I - X(X'X)^{-1}X'\}y = \lambda^{-2}(y - X\beta),$$

provided $D\mathrm{E}\,(y) = 0$ and where $\beta = (X'X)^{-1}X'y$ and the third equality follows from the lemma in the Appendix. If $X = 0$, then $u = \lambda^{-2}y$; if $X = I$, then $u = 0$. More generally if $\mathrm{cov}\,(y) = \sigma^2\,(\Sigma + \kappa X X')$, then using the transformation $\Sigma^{-1/2}y$ shows $u = \Sigma^{-1}(y - X\beta)$, where now $\beta = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$, the GLS estimate.

For the dental data $\beta = (160.1, 0.470)$, the estimated level and slope at the lowest pressure. The X matrix indicates how the initial level and slope manifest themselves in each observation, and the generalized least squares residuals are the residuals after removing the estimated effects of these initial conditions. For the lottery data $\beta = 204.78$, the estimated average lottery number for the month of January. A properly randomized drawing would imply an average of 183 and it is interesting to note the estimated standard deviation associated with 204.78 is 12.93 indicating the estimate is 1.68 estimated standard deviations away from 183, suggesting a statistically significant departure from randomness. The GLS residuals are the estimated deviations from the expected level based on the January estimate.

4.4. Fitted derivatives.

Suppose the cubic spline model (4) and put u^{Dy} as the smoothations computed from Dy, the second differenced data. Then u^{Dy} is the vector of second derivatives of the fitted smooth at the observed x_i. This stresses that although x is measured discretely and (6) is stated discretely, the discrete grid is arbitrarily fine and the signal is defined continuously.

To show the result, we will consider the signal plus noise model with $D\mathrm{E}\,(s) = 0$. Then

$$Ds = D\mathrm{E}\,(s\,|\,y) = D\{\mathrm{cov}\,(s)\}u = \mathrm{cov}\,(Ds)\,u^{Dy} = \sigma^2 R u^{Dy},$$

where, from Theorem 4, $u = D'u^{Dy}$. For a fitted cubic spline (Green & Silverman 1994) $Ds = Rs''$, where s'' is the vector of second derivatives of $s(x)$ evaluated at the x_i. The result follows, since R is nonsingular. It also follows that $u = D's''$.

4.5. Lagrange multipliers.

Smoothations are the values of relaxing each observational constraint $y = G\varepsilon$ in the implicit constrained least squares minimization of Theorem 3. Consider the state space model (6) and suppose the state equation holds for $i = 0$ with $\alpha_0 = 0$. Then from Subsection 2.5, $\mathrm{E}\,(\varepsilon\,|\,y)$ is that value of ε which minimizes $\varepsilon'\varepsilon$ subject to $y = G\varepsilon$. Using Lagrangian multipliers then, at the optimum the vector of Lagrange multipliers is $(GG')^{-1}y = u$. Thus a large smoothation indicates the corresponding observation is very binding in the least squares minimization.

4.6. Leverages.

The variance of the smoothation, $\mathrm{cov}\,(u_i) = \sigma^2 M_i$ is closely related to the well-known regression concept of leverage (Kohn & Ansley 1989). For the signal plus noise model, the leverage of y_i on the smooth at x_i is defined as

$$\frac{\partial}{\partial y_i'}\mathrm{E}\,(s_i\,|\,y) = I - \lambda^2 M_i.$$

The equality follows since

$$\begin{aligned}\mathrm{E}\,(s\,|\,y) &= \mathrm{E}\,(s) + \mathrm{cov}\,(s,y)\,\{\mathrm{cov}\,(y)\}^{-1}\,\{y - \mathrm{E}\,(y)\},\\ \frac{\partial}{\partial y'}\mathrm{E}\,(s\,|\,y) &= \sigma^{-2}\mathrm{cov}\,(s,y)\,\Sigma^{-1} \;=\; I - \lambda^2\Sigma^{-1},\end{aligned}$$

and M_i is diagonal block i of Σ^{-1}. The matrix $\partial\mathrm{E}\,(s\,|\,y)/\partial y'$ is often called the "hat" matrix and hence leverages are diagonals of the hat matrix. Each row of the hat matrix defines a smoothing kernel, indicating the weight associated with each observation in each smoothed value. All such kernels are easily computed with the KFS and for arbitrary state space models.

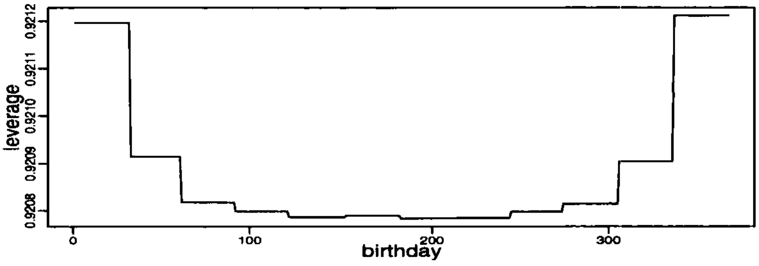

FIGURE 4: Leverages for the lottery data smooth.

Figure 4 plots the leverages for the lottery smooth. As expected, leverages are smallest in the middle of the x range, since there is more information on both sides of the point to pin down the smooth. This implies that the kernels, for different i, have larger bandwidths when i is near the middle rather than at either extreme. Leverages are constant in each month, since days within each month have the same influence on the fit for that month.

5. DIFFERENCING TO STATIONARITY

Differencing to stationarity is well known in time series analysis (Box, Jenkins & Reinsel 1994). It is implicit with the Reinsch algorithm, and the Reinsch development suggests an extension of the idea to series with, what may be regarded as, missing observations. Thus differencing to stationarity does not break down with missing observations and, insofar as smoothing is concerned, there is no loss of information.

THEOREM 4. *Suppose* $\{\mathrm{cov}\,(y)\}^{-1} = \sigma^{-2}D'R^{-1}D$. *Then* $u = D'u^{Dy}$ *where* u^{Dy} *is the vector of smoothations computed from* Dy. *If additionally,* y *is generated by the signal plus noise model with* $D\mathrm{E}\,(s) = 0$, *then* $\mathrm{E}\,(s \mid y) = y - \lambda^2 D'u^{Dy}$.

The indirectly computed smoothations $u = D'u^{Dy}$ can be used to construct smoothed estimates for any other state space quantities. For $i = n, \ldots, 1$ put $r_{i-1} = Z_i'u_i + T_i'r_i$ with $r_n = 0$. Then (de Jong 1988; Kohn & Ansley 1989) $\mathrm{E}\,(\varepsilon_i \mid y) = G_i'u_i + H_i'r_i$, from which all other smoothed state space quantities may be defined since $\mathrm{E}\,(\alpha_{i+1} \mid y) = T_i\mathrm{E}\,(\alpha_i \mid y) + H_i\mathrm{E}\,(\varepsilon_i \mid y)$.

6. PARAMETRIC AND NONPARAMETRIC SMOOTH COMPONENTS

Diffuse components yield parametric components in a smooth, and every smooth discussed above can be decomposed into parametric and nonparametric components. Parametric components correspond to the explanatory variables in a regression, while nonparametric components relate to the sequential covariance structure induced by the model.

Consider the signal plus noise model from Definition 1. Using the notation of Section 4.3,

$$\begin{aligned}\mathrm{E}\,(s \mid y) &= y - \lambda^2 u = y - \Sigma^{-1}(y - X\hat{\beta}) \\ &= \Sigma^{-1}X\hat{\beta} + (I - \Sigma^{-1})y = X\hat{\beta} + (I - \Sigma^{-1})(y - X\hat{\beta}),\end{aligned}$$

where $\hat{\beta}$ is the GLS estimate of the diffuse components. Thus if $\Sigma = I$, meaning the only source of randomness in the signal is diffuse, then $\mathrm{E}\,(s \mid y)$ is the least squares predictor arrived at by treating the fully diffuse components as fixed but unknown.

More generally, let s be generated by the state space model

$$\hat{s} \equiv \mathrm{E}\,(s \mid y) = \left(\frac{\partial s^{\beta}}{\partial \beta'}\right)\hat{\beta} + \hat{s}^0,$$

where s^{β} is $\mathrm{E}\,(s \mid y)$ evaluated at β. The first term on the right contains the parametric components of the smooth, while $\hat{s}^0$ is the nonparametric component. The KFS computes all components.

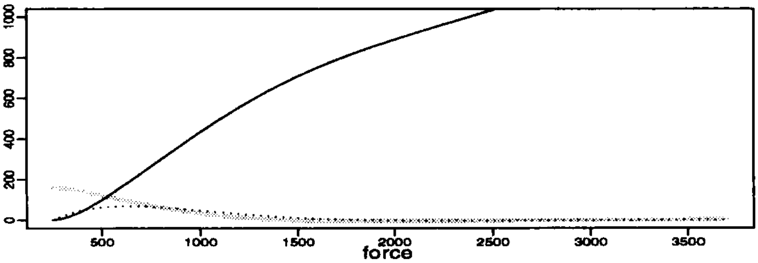

FIGURE 5: Parametric and nonparametric fit components for dental smooth.

Figure 5 displays the parametric and nonparametric components of the fit for the dental smooth. Each component has been scaled by the corresponding component of β. The local linear trend model for the dental data model has two diffuse components, the initial level and slope, and hence there are two parametric components in the fit. These are the lower two curves that rapidly cut out to zero in the first panel of Figure 5. Thus the parametric components play a limited role in the smooth, confined to the initial stretch of the data. The nonparametric component s^0 is the curve rising to the right. The sequential covariance structure induced by the model thus largely dictates the overall fit.

The parametric component is closely related to what Heckman & Ramsay (2000) define as the "favoured" model in relation to the PLS problem (1). The favoured model for s is the set of s as y varies and $\lambda \to \infty$. As λ in (1) increases, smoothness becomes the overriding concern and s has zero penalty and $Ds = 0$. Thus the favoured model is the null space of D or, equivalently, the range of X. The fit of the favoured model is $X\beta$, where $\beta = (X'X)^{-1}X'y$. In practice, λ is not arbitrarily large, permitting deviation from the favoured model $X\beta$.

If the signal is generated by a state space model, then $\lambda \to \infty$ implies the state equation errors are relatively zero and the state equations imply a deterministic evolution of the signal. Alternatively as $\lambda \to \infty$, then σ becomes relatively zero and, if $\{\mathrm{cov}\,(s)\}^{-1} = \sigma^2 D'R^{-1}D$, then $\mathrm{cov}\,(s)$ is divergent in all the diffuse directions defined by X, where $DX = 0$, and relatively zero in all other directions. Thus linear combinations of the components of s are either fully diffuse or, in effect, known. Thus $y = \mathrm{E}\,(s) + X\beta + \eta$, where $\eta \sim (0, \sigma^2 I)$, β is fully diffuse and X is of full and maximal column rank such that $DX = 0$. The fit of the favoured model is thus

$$\mathrm{E}\,(s) + X(X'X)^{-1}X'\{y - \mathrm{E}\,(s)\} = y - D'(DD')^{-1}D\{y - \mathrm{E}\,(s)\},$$

which reduces to the least squares fit $y - D'(DD')^{-1}Dy = X(X'X)^{-1}X'y$ if $D\mathrm{E}\,(s) = 0$. Note that $D'(DD')^{-1}Dy$ are the smoothations computed from Dy assuming $\sigma^2 R = \mathrm{cov}\,(Ds) = 0$.

For example with the dental data model, setting the state errors to zero yields a straight line for the signal and hence a straight line model is favoured. For the draft lottery model, zero state errors imply the favoured model is the constant mean model.

THEOREM 5. *For the signal plus noise model, if z is such that $Dz = 0$, then* $\mathrm{E}\,(s \mid y + z) = \mathrm{E}\,(s \mid y) + z$. *Alternatively for any* β, $\mathrm{E}\,(s \mid y + X\beta) = \mathrm{E}\,(s \mid y) + X\beta$.

7. TESTING FOR OVER OR UNDER SMOOTHING

Smoothations are important for testing over or under smoothing. Suppose the model (6) has been used to smooth the data but in fact the data have an extra source $\delta_i \sim (0, q\sigma^2 I)$, with

$$y_i = Z_i\alpha_i + G_i\varepsilon_i + X_i\delta_i, \quad \alpha_{i+1} = T_i\alpha_i + H_i\varepsilon_i + W_i\delta_i. \tag{8}$$

If $X_i \neq 0$, but $W_i = 0$, then the smoothing has not allowed for sufficient measurement error and there has been too much stress on fidelity to the data. This is called undersmoothing. If $X_i = 0$ but $W_i \neq 0$, then the signal is subject to more variation than originally allowed for and hence there is oversmoothing. Undersmoothing or oversmoothing may occur in patches. This is reflected in the appropriate choice of the X_i and W_i for different i.

If $\sigma^2\Sigma$ is the covariance matrix of y assuming $q = 0$, then given (8), the LBI statistic for testing $q = 0$ versus $q > 0$ is (King & Hillier 1985)

$$\sigma^{-2}y'\Sigma^{-1}\left\{\frac{\partial\,\mathrm{cov}\,(y)}{\partial q}\right\}\Sigma^{-1}y \Big/ y'\Sigma^{-1}y = \sigma^{-2}u'\left\{\frac{\partial\,\mathrm{cov}\,(y)}{\partial q}\right\}u \Big/ y'u.$$

If (8) holds, then

$$\mathrm{cov}\,(y) = \sigma^2\big[\Sigma + q\{W(1)W'(1) + W(2)W'(2) + \cdots + W(n)W'(n)\}\big]. \tag{9}$$

where $W(1), \ldots, W(n)$ are the matrices implied by (8) and specify how each δ_i manifests itself in y. Performing the differentiation shows

$$\begin{aligned}\sigma^{-2}u'\left\{\frac{\partial \operatorname{cov}(y)}{\partial q}\right\}u &= u'\{W(1)W'(1) + \cdots + W(n)W'(n)\}u \\ &= \sum_{i=1}^{n}(X_i'u_i + W_i'r_i)'(X_i'u_i + W_i'r_i),\end{aligned}$$

where the smoothations are computed assuming $q = 0$, r_i is as in Section 5, and the last equality follows from de Jong & Penzer (1998). Thus the LBI statistic is readily evaluated from the smoothations given the X_i and W_i in (8). These results generalize those in Harvey & Streibel (1997) which are set in the context of time series tests against stochastic levels, slopes and seasonals. These authors also discuss the distributions of the test statistics in these cases.

To illustrate, we will consider the lottery data set and model. To test constancy within each month, we put $W_i = 1$ whenever i does not refer to the last day of the month. The numerator of the LBI statistic is thus $\sum_i r_i^2$, where the i sum runs over days which are not the end of a month.

Testing for oversmoothing or undersmoothing can also proceed along the following lines. If $(X_i', W_i')'$ in (8) is zero except at a given i, then GLS estimate of δ and its covariance matrix is, for each given i, directly available from the KFS output (de Jong & Penzer 1998). GLS estimation is equivalent to treating δ as diffuse. Structuring $(X_i', W_i')'$ appropriately corresponds to the imposition of different "shocks" on the smooth, such as level or slope shocks.

APPENDIX: PROOFS

The proofs use the following Lemma.

LEMMA. *Suppose X is of full column rank and $DX = 0$, where D is of full and maximal row rank. Then (D', X) is nonsingular and $X(X'X)^{-1}X' + D'(DD')^{-1}D = I$.*

To prove the Lemma, we begin by noting that (D', X) is nonsingular, since the columns of X span the null space of D. Postmultiplying the left-hand side of the lemma equation by $(D', X)'$ yields $(D', X)'$, which implies the equality.

Theorem 1 is proved by noting that

$$\begin{aligned}\mathrm{E}(s\,|\,y) &= \mathrm{E}(s) + \operatorname{cov}(s, y)\{\operatorname{cov}(y)\}^{-1}\{y - \mathrm{E}(y)\} \\ &= \mathrm{E}(s) + \operatorname{cov}(s)\{\operatorname{cov}(s) + \lambda^2\sigma^2 I\}^{-1}\{y - \mathrm{E}(s)\},\end{aligned}$$

which rearranges to the required result.

To prove Theorem 2, we observe from the Lemma that (D', X) is nonsingular. Also $X'XX'X$ is nonsingular and

$$\begin{aligned}(\Sigma + \kappa XX')^{-1} &= (D', X)\begin{pmatrix} R & D\Sigma X' \\ X'\Sigma D' & X'\Sigma X + \kappa X'XX'X\end{pmatrix}^{-1}(D', X)' \\ &\longrightarrow (D', X)\begin{pmatrix} R^{-1} & 0 \\ 0 & 0\end{pmatrix}(D', X)' = D'R^{-1}D = \sigma^{-2}\{\operatorname{cov}(s)\}^{-1},\end{aligned}$$

where the limit as $\kappa \to \infty$ follows from partitioned matrix inversion.

To prove Theorem 3, we write $\nu = \varepsilon - \mathrm{E}(\varepsilon\,|\,y)$. Then for some nonsingular matrix B, $B\varepsilon = (y', \nu')'$ and

$$\varepsilon\{\operatorname{cov}(\varepsilon)\}^{-1}\varepsilon = (y', \nu')\{\operatorname{cov}(B\varepsilon)\}^{-1}(y', \nu')'.$$

Now $\mathrm{cov}\,(B\varepsilon)$ is block diagonal with diagonal blocks $\mathrm{cov}\,(y)$ and $\mathrm{cov}\,(\varepsilon \mid y)$. Hence

$$\varepsilon' \{\mathrm{cov}\,(\varepsilon)\}^{-1} \varepsilon = y' \{\mathrm{cov}\,(y)\}^{-1} y + \{\varepsilon - \mathrm{E}\,(\varepsilon \mid y)\}' \{\mathrm{cov}\,(\varepsilon \mid y)\}^{-1} \{\varepsilon - \mathrm{E}\,(\varepsilon \mid y)\}' x$$

which is minimized when $\varepsilon = \mathrm{E}\,(\varepsilon \mid y)$ with minimum $y' \{\mathrm{cov}\,(y)\}^{-1} y$. Further, suppose s is as specified. Then for some nonsingular matrix A, $\varepsilon = A(y', s')'$ and by a similar argument as above,

$$\varepsilon' \{\mathrm{cov}\,(\varepsilon)\}^{-1} \varepsilon = s' \{\mathrm{cov}\,(s)\}^{-1} f + \{y - \mathrm{E}\,(y \mid f)\}' \{\mathrm{cov}\,(y \mid f)\}^{-1} \{y - \mathrm{E}\,(y \mid f)\},$$

which is minimized when $\varepsilon = \mathrm{E}\,(\varepsilon \mid y) = A\{y', \mathrm{E}\,(s' \mid y)\}'$ or equivalently $f = \mathrm{E}\,(s \mid y)$.

We prove Theorem 4 by noting that from Theorem 2 we have

$$u = \{\mathrm{cov}\,(y)\}^{-1} y = D' \{\mathrm{cov}\,(Dy)\}^{-1} Dy = D' u^{Dy},$$

while from Section 4.1, $\mathrm{E}\,(s \mid y) = y - \lambda^2 u = y - \lambda^2 D' u^{Dy}$. Finally, to prove Theorem 5, we observe that if $Dz = 0$, then for some β, $z = X\beta$ and

$$\mathrm{E}\,(s \mid y + z) = y + z - \lambda^2 D' u^{D(y+X\beta)} = y + z - \lambda^2 D' u^{Dy} = \mathrm{E}\,(s \mid y) + z.$$

REFERENCES

H. Akaike (1980). Seasonal adjustment by a Bayesian modelling. *Journal of Time Series Analysis*, 1, 1–13.

B. D. O. Anderson & J. B. Moore (1979). *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ.

C. F. Ansley & R. Kohn (1985). Estimation, filtering and smoothing in state space models with incompletely specified initial conditions. *The Annals of Statistics*, 13, 1286–1316.

C. F. Ansley, R. Kohn & C.-M. Wong (1993). Nonparametric spline regression with prior information. *Biometrika*, 80, 75–88.

G. E. P. Box, G. M. Jenkins & G. C. Reinsel (1994). *Time Series Analysis: Forecasting and Control*, Third Edition. Prentice-Hall, Englewood Cliffs, NJ.

J. Chiou & H. Müller (1999). Nonparametric quasi-likelihood. *The Annals of Statistics*, 27, 36–64.

P. de Jong (1988). A cross-validation filter for time series models. *Biometrika*, 75, 594–600.

P. de Jong (1989). Smoothing and interpolation with the state-space model. *Journal of the American Statistical Association*, 84, 1085–1088.

P. de Jong (1991). The diffuse Kalman filter. *The Annals of Statistics*, 19, 1073–1083.

P. de Jong & J. R. Penzer (1998). Diagnosing shocks in time series. *Journal of the American Statistical Association*, 93, 796–806.

R. L. Eubank (1985). Diagnostics for smoothing splines. *Journal of the Royal Statistical Society Series B*, 47, 332–341.

S. E. Fienberg (1971). Randomization and social affairs: the 1970 draft lottery. *Science*, 171, 255–261.

P. G. Green & B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.

A. C. Harvey (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

A. C. Harvey & J. H. Stock (1994). Estimation, smoothing, interpolation and distribution for structural time-series models in continuous time. In *Models, Methods and Applications of Econometrics* (P. C. B. Phillips, Ed.). Basil Blackwell, Oxford, pp. 55–70.

A. C. Harvey & M. Streibel (1997). *Testing for Nonstationary Unobserved Components*. Mimeo.

N. E. Heckman & J. O. Ramsay (2000). Penalized regression with model-based penalties. *The Canadian Journal of Statistics*, 28, 241–258.

G. S. Kimeldorf & G. Wahba (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33, 82–95.

M. L. King & G. H. Hillier (1985). Locally best invariant tests of the error covariance matrix in linear regression models. *Journal of the Royal Statistical Society Series B*, 47, 98–102.

R. Kohn & C. F. Ansley (1989). A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika*, 76, 65–79.

S. Mazzi (1997). *The Model Based Approach to Smoothing*. Unpublished doctoral thesis, Department of Statistics, The University of British Columbia, Vancouver, BC, Canada.

H. Müller (1988). Nonparametric regression analysis for longitudinal data. *Lecture Notes in Statistics*, Springer-Verlag, 46, 36–38.

C. Reinsch (1967). Smoothing by spline functions. *Numerical Mathematics*, 10, 117–83.

G. K. Robinson (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6, 15–51.

I. Schoenberg (1964). Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences USA*, 52, 947–950.

S. Shively, R. Kohn & S. Wood (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, 94, 777–794.

B. W. Silverman (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society Series B*, 47, 1–52.

G. Wahba (1978). Improper priors, spline smoothing, and the problems of guarding against model errors in regression. *Journal of the Royal Statistical Society Series B*, 40, 364–372.

G. Wahba (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

W. Wecker & C. F. Ansley (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, 78, 81–9.

H. Weinert, R. Byrd & G. Sidhu (1980). A stochastic framework for recursive computation of spline functions: Part II: Smoothing splines. *Journal of Optimization Theory and Applications*, 30, 255–268.

E. T. Whittaker (1923). A new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41, 63–75.

Received 3 November 1999
Accepted 14 July 2000

Patrick E. BROWN: p.e.brown@lancaster.ac.uk
Department of Mathematics and Statistics
Lancaster University, Lancaster LA1 4YF, England, UK

Piet de JONG: piet.dejong@commerce.ubc.ca
Faculty of Commerce and Business Administration
University of British Columbia, Vancouver
British Columbia, Canada V6T 1Z2