# Model selection in a stochastic setting.

T. Prvan* M. R. Osborne†

(Received 8 August 2003; revised 24 February 2004)

**Abstract**

The given data is a set of observations on functionals of a trajectory of a system of differential equations. The a priori information is that the system is a member of a parametric family of systems of increasing complexity. The problem is to use the data to identify the particular member of this family which generated the observed data. The method associates each candidate model with the analogue of a generalised smoothing spline fitted to the given data. The resulting values of the smoothing parameter as well as graphical inspection of fit provide a basis for model selection.

*Department of Statistics, Macquarie University, Sydney, NSW 2109, Australia.
mailto:tprvan@efs.mq.edu.au

†Institute of Mathematical Sciences, ANU, ACT, Australia.
mailto:mike.osborne@maths.anu.edu.au

See http://anziamj.austms.org.au/V45/CTAC2003/Prva/home.html for this article,  Published August 8, 2004. ISSN 1446-8735

# Contents



# 1 Introduction

The observed data is assumed to be of the form

$$y_i = \boldsymbol{h}^T \boldsymbol{x}(t_i) + \epsilon_i \,. \tag{1}$$

where $\boldsymbol{h}$ defines the "observation functional" and the observational error is denoted $\epsilon_i$. The observational errors $\epsilon_i$ are assumed to be independently and identically distributed according to a Normal distribution with zero mean and common variance $\sigma^2$ (denoted $\epsilon_i \sim N(0, \sigma^2)$). The system of differential equations that $\boldsymbol{x}(t)$ satisfies will be assumed to be linear:

$$\frac{d\boldsymbol{x}}{dt} = M(t, \boldsymbol{\beta})\boldsymbol{x} \,. \tag{2}$$

The vector $\boldsymbol{\beta}$, depending on the problem being considered, may contain unknown parameters that need to be estimated from the data or known parameters. If the parameters are assumed to be known or fixed $\boldsymbol{\beta}$ will be

suppressed. If there is no time component $t$ this will also be suppressed. The problem has a well defined solution and $\boldsymbol{h}$ must have the property to capture it. The vector $\boldsymbol{h}$ must satisfy the condition that the solution of the differential equation is identifiable.

Generalised smoothing splines will be used to obtain the model that best fits the data. A stochastic formulation of the generalised smoothing spline will be used which permits the use of the Kalman filter, followed by the Fixed-interval, Discrete-time smoother (RTS smoother) and then an application of the Interpolation smoother to obtain $\mathcal{E}\{\boldsymbol{x}(t) \mid y_1, \ldots, y_n\}$ and hence the point estimate at time $t$.

Section 2 outlines the stochastic formulation of smoothing splines and then introduces generalized smoothing splines. Section 3 gives computational details and Section 4 provides smoothness properties of the generalised smoothing spline. Some model selection examples will be given in Section 5.

## 2 Stochastic formulation of smoothing spline extended to generalised smoothing spline

We first introduce the smoothing spline and the stochastic differential equation that a smoothing spline solves.

Suppose that the data $(t_1, y_1), \ldots, (t_n, y_n)$ are observed and it is assumed that the data can be decomposed as the signal plus noise model

$$y_i = f(t_i) + \epsilon_i\,, \quad \epsilon_i \sim N(0, \sigma^2)\,, \quad i = 1, \ldots, n\,. \tag{3}$$

The signal $f(t)$ is unknown and we wish to approximate it. One way to do this is by fitting a smoothing spline to the data. A smoothing spline $f$ is the minimizer of

$$\sum_{i=1}^{n} (y_i - f(t_i))^2 + \mu \int_{t_1}^{t_n} \left(f^{(m)}(t)\right)^2 dt\,. \tag{4}$$

The resultant curve is a piecewise polynomial of degree $2m-1$ with $2m-2$ continuous derivatives.

Wecker and Ansley [5] presented a stochastic formulation of the smoothing spline utilising a result by Wahba [4]. She showed that a polynomial smoothing spline is the solution to the stochastic differential equation

$$\frac{d^m x}{dt^m} = \sigma\sqrt{\lambda}\frac{d\omega}{dt}\,. \tag{5}$$

where $\omega(t)$ is a Wiener process (see for example Billingsley [1]) with unit dispersion parameter, $\lambda = 1/\mu$ and $\boldsymbol{x}(t_1) = [x(t_1), \ldots, x^{(m-1)}(t_1)]^T$ has a diffuse prior (that is, $\boldsymbol{x}(t_1) \sim N(\mathbf{0}, \gamma^2 I_m)$ and $\gamma^2 \to \infty$). The solution is

$$f(t) = \lim_{\gamma^2\to\infty} x(t \mid n)\,,$$

where $x(t \mid n)$ is the expected value of $x(t)$ conditioned on the data $y_1, \ldots, y_n$; that is, $\mathcal{E}\{x(t) \mid y_1, y_2, \ldots, y_n\}$. The stochastic differential equation (5) can be written in the matrix companion form

$$\frac{d\boldsymbol{x}}{dt} = \begin{pmatrix} \mathbf{0}_{m-1} & I_{m-1} \\ 0 & \mathbf{0}_{m-1}^T \end{pmatrix}\boldsymbol{x} + \sigma\sqrt{\lambda}\boldsymbol{e}_m\frac{d\omega}{dt}\,, \tag{6}$$

and the notation $\boldsymbol{e}_j$ is used to denote an $m$-vector having all zeros except for a 1 in the $j$th position. Corresponding to assumption (3) we write the observation equation as

$$y_i = \boldsymbol{e}_1^T\boldsymbol{x}(t_i) + \epsilon_i\,.$$

To seek generalised smoothing splines, as defined in Osborne and Prvan [2], we generalise the stochastic differential equation associated with smoothing spline to

$$\frac{d\boldsymbol{x}}{dt} = M(t,\boldsymbol{\beta})\boldsymbol{x} + \sigma\sqrt{\lambda}\boldsymbol{b}\frac{d\omega}{dt}\,, \tag{7}$$

where $M(t,\boldsymbol{\beta}) : R^m \to R^m$ and $\omega$ is a Wiener process with unit dispersion parameter. The initial conditions are the same as for the polynomial

smoothing spline and the point estimates are $\mathcal{E}\{\boldsymbol{h}^T\boldsymbol{x}(t) \mid y_1, \ldots, y_n\}$. The corresponding observation equation is

$$y_i = \boldsymbol{h}^T\boldsymbol{x}(t_i) + \epsilon_i \,. \tag{8}$$

We now obtain the state space formulation of the generalised smoothing spline. Let $X(t, t_1)$ be the fundamental matrix solution of the associated homogeneous differential equation, that is,

$$\frac{dX}{dt} = M(t, \boldsymbol{\beta})X \,, \quad X(t_1, t_1) = I_m \,.$$

The solution to the stochastic differential equation (7) satisfying initial condition $\boldsymbol{x}(t_1) = \boldsymbol{x}_1$ is

$$\boldsymbol{x}(t) = X(t, t_1)\boldsymbol{x}_1 + \sigma\sqrt{\lambda}\int_{t_1}^{t} X(t, s)\boldsymbol{b}\frac{d\omega}{ds}\, ds \,.$$

This solution can be written in the form of a recursion as

$$\boldsymbol{x}_i = X_i\boldsymbol{x}_{i-1} + \sigma\sqrt{\lambda}\boldsymbol{u}_i \,, \tag{9}$$

with $\boldsymbol{x}_i = \boldsymbol{x}(t_i)$, $X_i = X(t_i, t_{i-1})$ and $\boldsymbol{u}_i = \boldsymbol{u}(t_i, t_{i-1})$ where

$$\begin{aligned}
\boldsymbol{u}_i &= \int_{t_{i-1}}^{t_i} X(t_i, s)\boldsymbol{b}\frac{d\omega}{ds}\, ds \,, \\
\boldsymbol{u}_i &\sim N(0, R(t_i, t_{i-1})) \,, \\
R(t_i, t_{i-1}) &= \int_{t_{i-1}}^{t_i} X(t_i, s)\boldsymbol{b}\boldsymbol{b}^T X(t_i, s)^T\, ds \,.
\end{aligned}$$

For given $\lambda$ a forward pass of the Kalman filter, backward pass of the RTS Smoother and interpolation smoother are implemented on the *state space formulation* (8) and (9) to obtain $\boldsymbol{x}(t \mid n)$ and hence the generalised smoothing spline and its first $m-1$ derivatives evaluated at $t$. The smoothing

parameter $\lambda$ is usually chosen by generalised cross validation or maximum likelihood, for more details refer to Osborne and Prvan [2, 3] and references contained therein. In Osborne and Prvan [2] it was shown that the effect of $\gamma$ becomes asymptotically negligible by the $m$th step of the Kalman Filter when $\gamma$ is large.

The diffuse prior can be dealt with explicitly by setting $\gamma$ sufficiently large. For details on how to deal implicitly with the diffuse prior refer to Wecker and Ansley [5].

By varying $\boldsymbol{h}$ and $\boldsymbol{b}$ we come up with classes of generalised smoothing splines having different smoothness properties. The smoothness result in Osborne and Prvan [2] will still hold since it was developed for general $M(t, \boldsymbol{\beta})$. The form of the vector $\boldsymbol{h}$ depends on the observed data. The smoothness properties depend on the interaction properties between $\boldsymbol{h}$ and $\boldsymbol{b}$. As $\boldsymbol{h}$ is fixed by the form of the data designer questions come down to making an appropriate choice of $\boldsymbol{b}$.

# 3 Computation

We initialise the Kalman Filter with $\boldsymbol{x}_{1|0} = \boldsymbol{0}$ and $S_{1|0} = \gamma^2 I_m$ where $\gamma$ is chosen to be large. The effects of $\gamma$ are asymptotically negligible by step $m$ of the Kalman filter. The Kalman filter recursions are, for $i = 2, \ldots, n$:

$$
\begin{aligned}
\boldsymbol{x}_{i|i-1} &= X(t_i, t_{i-1})\boldsymbol{x}_{i-1|i-1}\,, \\
S_{i|i-1} &= X(t_i, t_{i-1})S_{i-1|i-1}X(t_i, t_{i-1})^T + \lambda\sigma^2 R(t_i, t_{i-1})\,, \\
d_i &= \boldsymbol{h}^T S_{i|i-1}\boldsymbol{h} + \sigma^2\,, \\
\varsigma_i &= y_i - \boldsymbol{h}^T\boldsymbol{x}_{i|i-1}\,, \\
\boldsymbol{x}_{i|i} &= \boldsymbol{x}_{i|i-1} + S_{i|i-1}\boldsymbol{h}\varsigma_i\,, \\
S_{i|i} &= S_{i|i-1} - S_{i|i-1}\boldsymbol{h}d_i^{-1}\boldsymbol{h}^T S_{i|i-1}\,.
\end{aligned}
$$

The quantities $\boldsymbol{x}_{n|n}$ and $S_{n|n}$ obtained from the forward pass of the Kalman filter initialise the RTS Smoother, for $j = n, \ldots, 1$:

$$\begin{aligned} A_j &= S_{j|j} X(t_{j+1}, t_j)^T S_{j+1|j}^{-1} , \\ \boldsymbol{x}_{j|n} &= \boldsymbol{x}_{j|j} + A_j(\boldsymbol{x}_{j+1|n} - \boldsymbol{x}_{j+1|j}) , \\ S_{j|n} &= S_{j|j} + A_j(S_{j+1|n} - S_{j+1|j}) A_j^T . \end{aligned}$$

Notice that the smoothed covariance is an end product in itself and does not enter the recursion for the smoothed state vectors.

The interpolation smoother for $t_{i-1} \le t < t_i$ is

$$\begin{aligned} \boldsymbol{x}(t \mid n) &= X(t, t_{i-1}) \boldsymbol{x}_{i-1|i-1} + A(t_i, t)(\boldsymbol{x}_{i|n} - \boldsymbol{x}_{i|i-1}) , \\ S(t \mid n) &= \Omega(t, t_{i-1}) + X(t, t_{i-1}) S_{i-1|i-1} X(t, t_{i-1})^T \\ &\quad - A(t_i, t)(S_{i|i-1} - S_{i|n}) A(t_i, t)^T , \end{aligned}$$

where

$$A(t_i, t) = \{X(t, t_{i_1}) S_{i-1|i-1} X(t_i, t_{i-1})^T + \Gamma(t_i, t)\} S_{i|i-1}^{-1}$$

and

$$\Gamma(t_i, t) = \lambda \sigma^2 R(t, t_i) X(t_i, t)^T .$$

Output from the Kalman filter permits the log likelihood to be written

$$L = -\sum_{i=1}^{n} \frac{1}{2} \left\{ \frac{\varsigma_i^2}{\sigma^2 + \boldsymbol{h}^T S_{i|i-1} \boldsymbol{h}} + \log(\sigma^2 + \boldsymbol{h}^T S_{i|i-1} \boldsymbol{h}) \right\} + \text{const} .$$

If dealing with the diffuse prior explicitly we would drop the first $m-1$ terms of the log likelihood and maximize this partial log likelihood over $\lambda$ to obtain the maximum likelihood estimate of the smoothing parameter for fixed parameters in the model being considered. We then choose the model whose fixed parameter values result in the smallest residual sum of squares because the response surface for the partial log likelihood is not unimodal. Recall that the residual sum of squares is the sum of the square of the differences between the observed value and fitted value at the data points.

# 4 Smoothness properties

The smoothness of the state vector estimate $\boldsymbol{x}(t \mid n)$ depends on the choice of $\boldsymbol{b}$. The smoothness of the point estimate also depends on $\boldsymbol{b}\boldsymbol{b}^T X(t_i, t)^T$. Now

$$\frac{d\boldsymbol{x}(t \mid n)}{dt} = M(t, \boldsymbol{\beta})\boldsymbol{x}(t \mid n) + \boldsymbol{b}\boldsymbol{b}^T X(t_i, t)^T S_{i|i-1}^{-1}(\boldsymbol{x}_{i|n} - \boldsymbol{x}_{i|i-1})\,. \tag{10}$$

The only points at which the state vector can fail to have continuity is at the points $t_{i-1}$. The term $\boldsymbol{b}\boldsymbol{b}^T X(t_i, t)^T$in (10) is of interest. Looking at the jump in the first derivative at the $i$th data point, after some manipulation, we get

$$\frac{d\boldsymbol{x}(t_i^+ \mid n)}{dt} - \frac{d\boldsymbol{x}(t_i^- \mid n)}{dt} = \boldsymbol{b}^T\boldsymbol{h}\left\{\frac{\varsigma_i}{\sigma^2 + \boldsymbol{h}^T S_{i|i-1}\boldsymbol{h}} - \frac{1}{\sigma^2}h^T(\boldsymbol{x}_{i|n} - \boldsymbol{x}_{i|i})\right\}\,.$$

The jump vanishes provided $\boldsymbol{b}^T h = 0$. The extension of this result to higher derivatives is given below.

**Result 1** *The first $k$ derivatives of $\boldsymbol{x}(t \mid n)$ are continuous only if*

$$\boldsymbol{b}^T P_j(M(t, \boldsymbol{\beta}))^T\boldsymbol{h} = 0\,, \quad j = 0, 1, \ldots, k-1\,, \tag{11}$$

*where, for $i = 1, 2, \ldots,$*

$$P_0(M(t, \boldsymbol{\beta})) = I_m\,, \quad P_i(M(t, \boldsymbol{\beta})) = \frac{dP_{i-1}}{dt} - M(t, \boldsymbol{\beta})P_{i-1}\,.$$

If $M(t, \boldsymbol{\beta})$ is a constant matrix, say $M$, then these $P_i(M(t, \boldsymbol{\beta}))$ simplify to $(-1)^i M^i$.

# 5 Model selection examples

Consider the two following systems of differential equations:

$$\frac{dA}{dt} = -k_1 A\,, \quad \frac{dB}{dt} = k_1 A - k_2 B\,, \quad \frac{dC}{dt} = k_2 B\,; \tag{12}$$

and

$$\frac{dA}{dt} = -kA\,, \quad \frac{dB}{dt} = kA\,. \tag{13}$$

In both cases the sum of the solutions will add up to a constant. The two systems could be competing models for (say) a simple chemical reaction. Say $A > B > C$ against $A > C$ with the parameters specifying the reaction rates in the two cases.

The systems of differential equations can be rewritten as $\frac{d\boldsymbol{x}}{dt} = M\boldsymbol{x}$ where

$$M = \begin{pmatrix} -k_1 & 0 & 0 \\ k_1 & -k_2 & 0 \\ 0 & k_2 & 0 \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} -k & 0 \\ k & 0 \end{pmatrix} \quad \text{respectively.}$$

## 5.1 Implications of choice of $\boldsymbol{h}$ and $\boldsymbol{b}$

For the simpler model (13) $\boldsymbol{h}$ has to have a nontrivial component of $\boldsymbol{e}_2$ to be identifiable. In a similar manner the more complicated model (12) $\boldsymbol{h}$ must have a nontrivial component of $\boldsymbol{e}_3$ to be identifiable. We must have a component of the resultant species to be identifiable.

Suppose that we fit a generalised smoothing spline to the data using the simpler model (13) with $\boldsymbol{h} = \boldsymbol{e}_2$ and $\boldsymbol{b} = \boldsymbol{e}_1$. Using (11) to determine the continuity properties of the generalised smoothing spline fit to the data we have that

$$\boldsymbol{h}^T\boldsymbol{b} = \boldsymbol{e}_2^T\boldsymbol{e}_1 = 0\,,$$

$$\boldsymbol{h}^T M \boldsymbol{b} = \boldsymbol{e}_1^T \begin{pmatrix} -k & 0 \\ k & 0 \end{pmatrix} \boldsymbol{e}_2 = k\,.$$

The resultant curve fitted has one continuous derivative. If we had used $\boldsymbol{h} = \boldsymbol{e}_2$ and $\boldsymbol{b} = \boldsymbol{e}_2$ instead the resultant curve would have no continuous derivatives.

In a similar manner we could show that fitting a generalised smoothing spline to the data using the other model (12) with $\boldsymbol{h} = \boldsymbol{e}_3$ and $\boldsymbol{b} = \boldsymbol{e}_1$ results in a fitted curve with two continuous derivatives. The state covariance matrix involves both $k_1$ and $k_2$ and has full rank. If we use $\boldsymbol{h} = \boldsymbol{e}_3$ and $\boldsymbol{b} = \boldsymbol{e}_2$ instead, the resultant curve has one continuous derivative. The state covariance matrix involves only $k_2$ and is semi positive definite.

If the best fit occurs when the smoothing parameter is small and the resultant fit is smooth, then this is evidence that the model fits well. A small smoothing parameter $\lambda$ in (7) results in the stochastic forcing term being negligible and the differential equation (2) being a plausible model. Unfortunately, the definition of small depends on the data at hand which is why as a precaution a plot of the fit with data superimposed should be inspected as well.

## 5.2 Simulation

Suppose we observe $C$ contaminated by noise for the system of $n$ differential equations

$$y_i = \frac{k_2}{k_1 - k_2} e^{-k_1 t_i} - \frac{k_1}{k_1 - k_2} e^{-k_2 t_i} + 1 + \epsilon_i\,, \quad i = 1, \ldots, n\,.$$

Using $n = 101$, data was simulated for various choices of $k_1$ and $k_2$ with a significant noise component (roughly one tenth of the range of the signal in the interval considered). We then first fitted the simpler system of equations (13) to the data with

$$M = \begin{pmatrix} -k & 0 \\ k & 0 \end{pmatrix}\,, \quad \boldsymbol{h} = \boldsymbol{e}_2 \quad \text{and} \quad \boldsymbol{b} = \boldsymbol{e}_1\,.$$

TABLE 1: Partial log likelihood values for a selection of $k$ using optimal $\lambda$.

| $k$ | $\lambda$ | $f$ | RSS | $k$ | $\lambda$ | $f$ | RSS |
|---|---|---|---|---|---|---|---|
| 1.0 | $10^{-6}$ | −28.3233 | 1.1288 | 0.1 | $10^{-6}$ | −27.7779 | 2.1588 |
| 0.9 | $10^{-6}$ | −28.2479 | 1.0206 | 0.01 | 97.45 | −23.9686 | 1.9985 |
| 0.8 | $10^{-6}$ | −28.2031 | 0.9496 | 0.001 | $1.08210^4$ | −19.4425 | 1.9831 |
| 0.7 | $10^{-6}$ | −28.1933 | 0.9248 | 0.0001 | $10^6$ | −14.8537 | 2.0210 |
| 0.6 | $10^{-6}$ | −28.2201 | 0.9552 | 0.00001 | $10^{-6}$ | −10.7087 | 2.6229 |
| 0.5 | $10^{-6}$ | −28.2792 | 1.0496 | 0.000001 | $10^6$ | −14.0190 | 7.5735 |
| 0.4 | $10^{-6}$ | −28.3548 | 1.2151 | | | | |
| 0.3 | $10^{-6}$ | −28.4059 | 1.4858 | | | | |
| 0.2 | $10^{-6}$ | −28.3279 | 1.7720 | | | | |

The $\boldsymbol{b}$ has been chosen to ensure maximum continuity properties for the generalised smoothing spline.

For the choice $k_1 = 1$ and $k_2 = 2$ simulated data the resultant fits for two choices of $k$ are given in Figure 1 along with the underlying signal. As well, values of the partial log likelihood for the optimal smoothing parameter are given in the Table 1 for a series of values of $k$ in the simpler model (13) fitted to the simulated data using a generalised smoothing spline. A local best fit to the underlying signal occurs for $k = 0.7$ and $\lambda = 10^{-6}$ which corresponds to the residual sum of squares being smallest (RSS=0.9248). Further investigation reveals a larger partial log likelihood for $k = 0.00001$ and $\lambda = 10^{-6}$ but it has a much larger residual sum of squares (RSS = 2.6229). This fit lies close to a linear fit to the data. Except for the beginning of the data the simpler model (13) fitted well for $k = 0.7$. See in Figure 1 that for $k = 0.00001$ the simpler model does not fit well at all. For the two parameter generalised

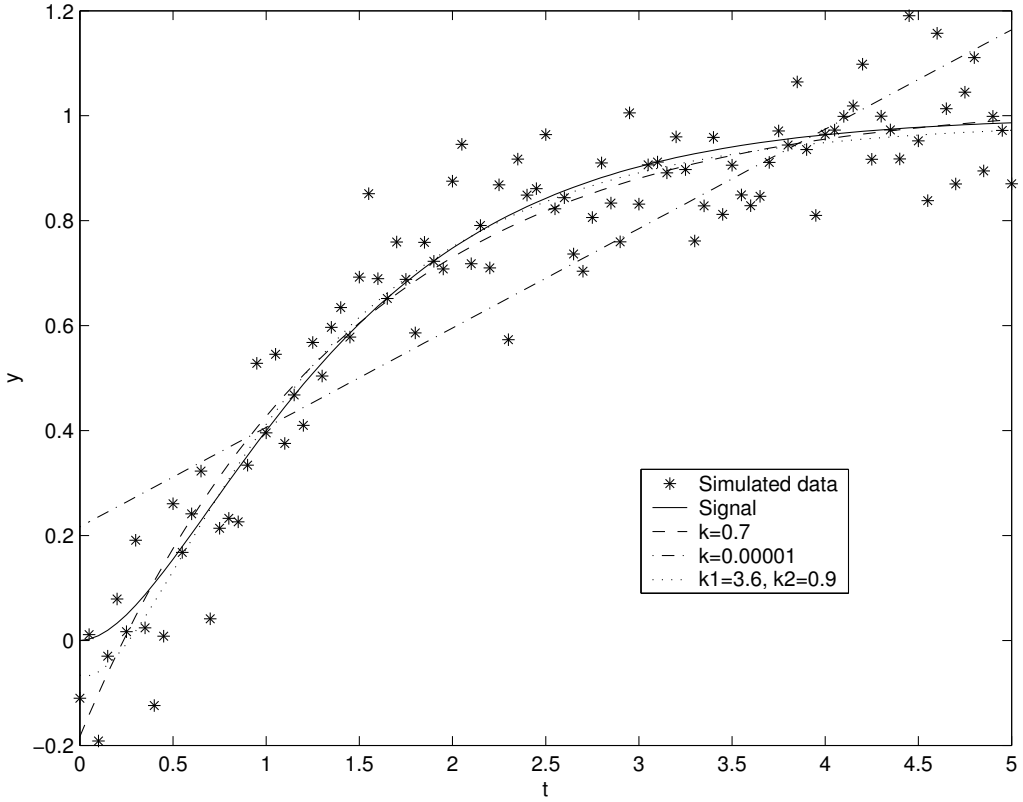


FIGURE 1: Generalised smoothing spline fits for simpler model for two choices of $k$ when the underlying signal has $k_1 = 1$ and $k_2 = 2$. Solid line is true underlying signal. Asterisks are the simulated data. Dotted line is two parameter smoothing spline fitted to data ($k_1 = 3.6$ and $k_2 = 0.9$).

smoothing spline model fitted to the data, using

$$M = \begin{pmatrix} -k_1 & 0 & 0 \\ k_1 & -k_2 & 0 \\ 0 & k_2 & 0 \end{pmatrix}, \quad \boldsymbol{h} = \boldsymbol{e}_3 \quad \text{and} \quad \boldsymbol{b} = \boldsymbol{e}_1 ,$$

we get $\lambda = 10^{-6}$ for $k_1 = 3.6$ and $k_2 = 0.9$ giving the smallest RSS over all values of $k_1$ and $k_2$ considered where the smoothing parameter is chosen to maximize the partial log likelihood. The resultant fit is better than that for $k = 0.7$ in the one parameter smoothing spline model (RSS=0.87359). This fit is also plotted on Figure 1. The beginning of the data is captured better.

# References


[1] P. Billingsley. *Probability and Measure.* Wiley, 1979. C790

[2] M. R. Osborne and T. Prvan. On algorithms for generalised smoothing splines. *J. Austral. Math. Soc. Ser. B*, 29:322–241, 1988. C790, C792

[3] M. R. Osborne and T. Prvan. Smoothness and conditioning in generalised smoothing spline calculations. *J. Austral. Math. Soc. Ser. B*, 30:43–56, 1988. C792

[4] G. Wahba, Improper priors, spline smoothing and the problem of guarding against model errors in regression, *J. R. Statist. Assoc.*, 40:364–372, 1978. C790

[5] W. Wecker and C. F. Ansley. Signal extraction approach to nonlinear regression and spline smoothing. *J. Amer. Statist. Assoc.*, 78:81–89, 1983. C790, C792