The Canadian Journal of Statistics
Vol. 28, No. ?, 2000, Pages ???-???
La revue canadienne de statistique

Penalized regression with model-based penalties

Nancy E. HECKMAN and J.O. RAMSAY
University of British Columbia and McGill University

Key words and phrases: nonparametric regression, penalized least squares, splines.
Mathematics subject classification codes (1991): primary 62G07, 65D07.

ABSTRACT

Nonparametric regression techniques such as spline smoothing and local fitting depend implicitly on a parametric model. For instance, the cubic smoothing spline estimate of a regression function μ based on observations t_i, Y_i is the minimizer of $\sum(Y_i - \mu(t_i))^2 + \lambda\int(\mu'')^2$. Since $\int(\mu'')^2$ is zero when μ is a line, the cubic smoothing spline estimate favors the parametric model $\mu(t) = \alpha_0 + \alpha_1 t$. Here we consider replacing $\int(\mu'')^2$ with the more general expression $\int(L\mu)^2$ where L is a linear differential operator with possibly nonconstant coefficients. The resulting estimate of μ performs well, particularly if $L\mu$ is small. We present a $O(n)$ algorithm for the computation of μ. This algorithm is applicable to a wide class of L's. We also suggest a method for the estimation of L. We study our estimates via simulation and apply them to several data sets.

RÉSUMÉ

Les techniques de régression non paramétrique telles que l'ajustement local ou le lissage au moyen de splines dépendent implicitement d'un modèle paramétrique sous-jacent. Dans l'estimation lisse par spline cubique d'une fonction de régression μ à partir d'observations t_i, par exemple, on choisit Y_i de façon à minimiser $\sum\{Y_i - \mu(t_i)\}^2 + \lambda\int(\mu'')^2$. Puisque le second terme s'annule lorsque μ est linéaire, l'estimateur favorise le modèle paramétrique $\mu(t) = \alpha_0 + \alpha_1 t$. Les auteurs proposent ici de remplacer le terme $\int(\mu'')^2$ par une expression plus générale, $\int(L\mu)^2$, faisant intervenir un opérateur différentiel linéaire L à coefficients éventuellement variables. Ils montrent que l'estimation de μ qui en résulte se comporte bien, particulièrement si $L\mu$ est petit. Ils fournissent un algorithme d'ordre $O(n)$ pour le calcul de μ qui est valable pour une classe assez large d'opérateurs. Ils suggèrent en outre une méthode d'estimation de L. Des simulations et des exemples d'application concrets illustrent le propos.

1. INTRODUCTION: THE FAVORED MODEL

Nonparametric approaches to estimating regression functions contrast with more traditional parametric models in offering greater flexibility in fitting the data. Although these smoothing methods seem not to be model-based, they implicitly use a pre-specified parametric model, and we demonstrate that the choice of this model can be important for fitting the data well.

For example, the cubic smoothing spline estimate of the regression function μ is the function that minimizes $\sum(Y_i - \mu(t_i))^2 + \lambda \int (\mu''(t))^2 \, dt$, where the Y_i's are the data and λ is a smoothing parameter. The first sum of squares term measures μ's fit to the data, and the second penalty term measures μ's closeness to a straight line since, if μ is a line, then the penalty term is equal to zero. Hence when λ is zero, the estimate of μ interpolates the data; when λ is large the estimate is close to the least squares regression line. Furthermore, if the Y_i's lie on a line then the estimate of μ will be that line, no matter what the value of λ. The parametric model $\mu(t) = \beta_0 + \beta_1 t$ may thus be considered the favored model for this penalty, or can be called hypersmooth in the sense defined by this penalty.

In local linear regression (Fan, 1992; Cleveland, 1979; Fan & Gijbels, 1996), the favored model is also linear. The estimate of $\mu(t)$ is $\hat{\mu}(t) = \hat{\alpha}_1$, where $\hat{\alpha}_1$ and $\hat{\alpha_2}$ minimize $\sum \left(Y_i - \alpha_1 - \alpha_2 t_i\right)^2 W(|t_i - t|/\lambda)$ for a known function W. Typically $W(u)$ is small if $|u|$ is large. Thus one assumes that, locally, μ is close to the parametric form $\mu(t) = \alpha_1 + \alpha_2 t$. The value of λ determines what values of t_i are considered local to t. If λ is large, then the estimate of μ is close to the least squares line. If the data Y_i lie on a line, then the estimate of μ will be that line.

The need for a wider class of favored functions is illustrated in Figure 1, showing two nonparametric regression estimates of the expected incidence of malignant melanomas in males between 1936 and 1972 (Andrews & Herzberg, 1985). In the sense defined in Section 3, each estimate uses 5.6 parameters. The estimate labeled CUB is a cubic smoothing spline, and the estimate labeled PER, an application of our methodology, is described in Section 5.1. Like the cubic smoothing spline estimate, the PER method is computationally fast and very flexible. However, it uses $\alpha_0 + \alpha_1 t + \alpha_3 \cos \omega t + \alpha_4 \sin \omega t$, with $\omega = 0.58$, as its favored parametric model, where ω was estimated from the data. As we can see, the CUB estimate picks up the linear trend, but completely misses the periodic component; but the PER estimate picks up both the linear and periodic trend.

FIGURE 1 HERE

The smoothing spline PER works better because an appropriate choice of penalty can lead to an estimator with low bias. Suppose that the data satisfy

$$Y_i = \mu(t_i) + \epsilon_i$$

where $a \le t_1 < t_2 < \cdots < t_n \le b$ and the ϵ_i's are independent with mean 0. The penalized weighted least squares estimate of μ minimizes

$$\sum d_i \big(Y_i - \mu(t_i)\big)^2 + \lambda P(\mu), \tag{1}$$

where the d_i's are positive known weights and P, the penalty, takes on non-negative values. If the t_i's are not distinct, (1) can be modified. For instance, suppose that $t_1 = t_2$ and that the d_i's are all equal to 1. Just as in ordinary least squares, we would replace $(Y_1 - \mu(t_1)^2 + (Y_2 - \mu(t_2)^2$ with $2(\bar{Y} - \mu(t_1))^2$, where $\bar{Y}$ is the average of Y_1 and Y_2. Throughout, we assume μ is in the Sobolev space $\mathcal{H}^m$, the set of all $\mu : [a, b] \to \Re$ with $\mu^{(j)}$ absolutely continuous, $j = 0, \cdots, m-1$, and $\int_a^b \left(\mu^{(m)}\right)^2 < \infty$. Now, for $\hat{\mu}_\lambda$ minimizing (1)

$$\sum d_i \Big[E\big(\hat{\mu}_\lambda(t_i)\big) - \mu(t_i)\Big]^2 \le \lambda P(\mu), \tag{2}$$

where μ is the true regression function (Wahba, 1990). Therefore, the more closely the true μ satisfies $P(\mu) = 0$, or is in the kernel of P, the smaller is the bias of $\hat{\mu}_\lambda$. Thus, the kernel of penalty P defines the favored parametric model. Of course, the estimate of μ need not follow the favored parametric model, and, in fact, can be quite far from it if the data require λ to be relatively small. That is, the choice of P does not determine a spline smoother's capacity to track the data. However, an appropriate choice of P can reduce the "number of parameters" required to track the data. This point is explained in more detail in Sections 3 and 6.

2. CHOOSING THE PENALTY: L-SPLINES

2.1. L: A LINEAR DIFFERENTIAL OPERATOR

Here we consider L-splines, defined by penalties that depend on linear differential operators. Let L be an mth order linear differential operator

$$L = D^m + \sum_{j=0}^{m-1} w_j D^j \tag{3}$$

where D^j denotes the jth derivative operator and the w_j's are continuous real-valued weight functions. Note that $L\mu$ is square integrable for all $\mu \in \mathcal{H}^m$. In (1), we take as penalty $P(\mu) = \int (L\mu)^2$ and thus seek μ to minimize

$$H(\mu, w_0, \cdots, w_{m-1}) = \sum d_i (Y_i - \mu(t_i))^2 + \lambda \int \left[L\mu(t)\right]^2 \, dt. \tag{4}$$

Suppose that the favored model is a linear combination of m basis functions $u_1, \cdots, u_m \in \mathcal{H}^m$ that are linearly independent and have m continuous derivatives. If the associated m by m Wronskian matrix function $\mathbf{W}(t)$ having elements

$$\mathbf{W}(t)_{ij} = D^{(j-1)} u_i(t) \tag{5}$$

is invertible for all $t \in [a, b]$, then one can find the weight functions w_j in (3) by solving $\mathbf{W}(t)\ \mathbf{w}(t) = -D^m \mathbf{u}(t)$ for $\mathbf{w}(t)$, where $\mathbf{w}(t) = (w_0(t), \cdots, w_{m-1}(t))'$ and $\mathbf{u}(t) = (u_1(t), \cdots, u_m(t))'$. See, for example, Coddington (1961).

Table 1 contains simple examples of differential operators and bases for the corresponding parametric families. The last two operators define the classes of differentiable positive functions and twice-differentiable monotone functions, respectively (Ramsay, 1998).

TABLE 1 HERE.

How does one choose a favored parametric model? For the melanoma incidence data in Figure 1, the model was chosen based on the belief that melanoma incidence is increasing with an overlaid periodic component due to sunspot activity. Many of the classical parametric models in nonlinear regression are based upon physical models, and any of these can be considered as the favored model in a penalized likelihood approach, so that parametric assumptions may be applied in a flexible way.

In this section we assume that the basis, $u_1, \cdots, u_m$, for the kernel of L is known. In Section 4, we discuss how to use the data to choose appropriate u_j's.

2.2. L-SPLINE ALGORITHMS AND APPLICATIONS

The idea of smoothing with a penalty defined by a linear differential operator L has a considerable history. Prior to their appearance in statistics, L-splines were much studied by numerical analysts as generalizations of polynomial splines. Schultz & Varga (1967) and Schumaker (1981) provide reviews of this early literature. Kimeldorf & Wahba (1970) proposed a slightly restricted form of L-spline for statistical smoothing. They also noted that the minimizer of (4) is the Bayes estimate under a Gaussian process prior on μ. This stochastic process connection was taken up in more detail by Wahba (1978) and Weinert, Byrd & Sidhu (1980).

Smoothing methods differ fundamentally from time-based regression approaches such as time series models and state-space methods. The latter attempt to estimate $\mu(t)$ using only data for $t_i \leq t$, while smoothing uses the entire data. A series of papers (Ansley & Kohn, 1985; de Jong, 1997; Kohn & Ansley, 1989, 1991, 1993; Kohn, Ansley & Wong, 1992; Wecker & Ansley, 1983) showed that a modified Kalman filter would produce the same estimates as a smoothing spline. The modified filter uses a diffuse initial state and is run both forward and backward through the t_i sequence. These authors also showed that this Kalman filter could be carried out in $O(n)$ operations.

Unfortunately, this valuable line of investigation has not had the attention in books on smoothing, or the impact on smoothing practice, that it merits. One reason may be that no publicly accessible software was, to our knowledge, ever made available for what turned out to be a rather complex algorithm. Moreover, their algorithms present special problems when the w_j's are non-constant, since the algorithms require the computation of matrix functions of the form

$$F(t,s) = \exp\big[\int_s^t A(v)\,dv\big], \quad U(t,s) = \int_s^t F(t,v) bb' F(t,v)'\,dv$$

where $A(v)$ is an m by m matrix and b is an m-vector (Ansley, Kohn, & Wong, 1993). Calculation of F involves m^2 integrals for each value of t and s. For nonconstant w_j's, these integrals must typically be calculated numerically. The further integration in U seems likely to make the calculations prohibitively expensive. However, in some cases the matrix F is sparse. By contrast, the technique that we describe, which is also $O(n)$, is essentially as fast for the nonconstant w_j case as for the constant.

Unequally spaced values of t_i may present implementation challenges for the Kalman filter method. However, our method is not effected by unequal argument spacing.

In defense of the Kalman-Filter approach, however, Kohn & Ansley (1993) found that, for the $L = D^2$ case, the approach is at least competitive with the widely available cubic smoothing spline method of Reinsch (1967, 1971) and Hutchison & de Hoog (1985). Moreover, the use of the Kalman Filter brings other useful things, such as confidence and prediction intervals in $O(n)$ calculations, and bandwidth choice by maximum likelihood. We hope that the state-space approach continues to evolve, and that these disparate strands in the literature converge to a unified presentation.

Finally, Wang & Brown (1996) use a modified spline to fit a family of curves, each of which has period of one day. They estimate a phase shift for each of the curves and then minimize (4) with $L = D^2 + (2\pi)^2 I$ over all functions μ with $\mu(0) = \mu(1)$ and $\mu'(0) = \mu'(1)$.

2.3. THEORETICAL BACKGROUND

The results in this section follow easily from the techniques of Wahba (1990).

For fixed w_j's, any μ that minimizes (4) must lie in a finite dimensional subspace of $\mathcal{H}^m$. This subspace is determined solely by L and the t_i's, and so is known. We wish to find $\mu \in \mathcal{H}^m$ that minimizes (4). We use the inner product on $\mathcal{H}^m$: $\langle f, g\rangle = \sum_{j=0}^{m-1} f^{(j)}(a)g^{(j)}(a) + \int(Lf)(Lg)$. By standard results of differential equation theory, whenever $L\mu \equiv 0$ and $\mu^{(j)}(a) = 0$, $j = 0, \cdots, m-1$ then $\mu \equiv 0$, and so $\langle , \rangle$ is indeed an inner product.

With this inner product, the linear functional $\mu \to \mu(t)$ is continuous and thus $\mathcal{H}^m$ is a reproducing kernel Hilbert space. Therefore there exists a bivariate function $k(\cdot, \cdot)$ called the reproducing kernel such that $k(s,t) = k(t,s)$, $k(s, \cdot) \in \mathcal{H}^m$ for all $s \in [a,b]$, and $\int(Lk(\cdot,t))(s)(Lf)(s)\ ds = f(t)$ for all $f \in \mathcal{H}^m$ with $D^j f(a) = 0$, $j = 0, \cdots, m-1$.

The reproducing kernel function k is expressible in terms of G, the Green's function associated with L. The Green's function is given by

$$G(s,w) = \left[\begin{array}{l} \sum_{i=1}^m u_i(s)u_i^*(w), \quad w \leq s \\ 0 \ \ \text{otherwise} \end{array} \right] \tag{6}$$

where $(u_1^*(w), \cdots, u_m^*(w))$ is the last row of the inverse of the Wronskian $\mathbf{W}(w)$. The function k is then

$$k(s,t) = \int_a^b G(s,w)G(t,w)\ dw. \tag{7}$$

For more details on Green's functions, see Coddington (1961).

A minimizer of (4) exists and is of the form

$$\mu(\cdot) = \sum_{j=1}^m \alpha_j u_j(\cdot) + \sum_{j=1}^n \beta_j k(t_j, \cdot). \tag{8}$$

The optimal α_j's and β_j's can be found by minimizing

$$(\mathbf{Y} - \mathbf{U}\boldsymbol{\alpha} - \mathbf{K}\boldsymbol{\beta})'\mathbf{D}(\mathbf{Y} - \mathbf{U}\boldsymbol{\alpha} - \mathbf{K}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta} \tag{9}$$

where $\mathbf{Y} = (Y_1, \cdots, Y_n)'$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_m)'$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)'$, $\mathbf{U}_{ij} = u_j(t_i)$, $\mathbf{K}_{ij} = k(t_i, t_j)$, and $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ii} = d_i$.

Thus, one can find the minimizing μ by setting the derivative of (9) equal to zero and solving for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. However, the resulting linear equations can be extremely ill-conditioned, even for n as small as twenty. Moreover, the matrix calculations are slow, requiring in general $O(n^3)$ calculations.

In what follows, we shall only need the diagonal and first $2m$ off-diagonal elements of $\mathbf{K}$. We apply the results of Anselone & Laurent (1968) to the minimization problem. These authors considered L-spline smoothing in a rather abstract setting, showing that the problem can be solved with $n - m$ linear equations rather than n. Going from $O(n^3)$ to $O[(n-m)^3]$ does not offer much relief, but they offered a more detailed analysis of polynomial smoothing corresponding to $L = D^m$, and showed how to construct an $O(n)$ algorithm in this case. Our algorithm, described in detail in the Appendix, extends their approach by showing that a local application of the QR decomposition yields the minimizing μ in $O(n)$ calculations, and is numerically stable even when n is large (e.g. greater than 1000) and even when the w_j's in (3) are non-constant functions. In particular, we don't need to compute all of $\mathbf{K}$ - we only need the diagonal and first $2m$ off-diagonals of $\mathbf{K}$.

Splus modules *Pspline* (for $L = D^m$) and *Lspline* (for general L's) were developed by the second author and are available by anonymous *ftp* at *statlib.stat.cmu.edu* or from the web at *www.stat.cmu.edu*. The modules provide estimates of the regression function and its derivatives.

3. DEGREES OF FREEDOM AND CHOICE OF λ

To more meaningfully scale the value of the smoothing parameter λ, one can use df_λ, the effective number of parameters or degrees of freedom that $\hat{\mu}_\lambda$ uses in estimating μ. A common definition (Hastie & Tibshirani, 1990) is trace $\mathbf{S}_\lambda$ where $\mathbf{S}_\lambda$ is the hat matrix, not depending on $\mathbf{Y}$, and satisfying $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{S}_\lambda \mathbf{Y}$ for $\hat{\boldsymbol{\mu}}_\lambda = (\hat{\mu}_\lambda(t_1), \cdots, \hat{\mu}_\lambda(t_n))'$ and $\mathbf{Y} = (Y_1, \cdots, Y_n)'$. One can show that, for fixed $\mathbf{w}$, trace $\mathbf{S}_\lambda$ is a decreasing function of λ ranging from n when $\lambda = 0$ to m, the dimension of the kernel of L, as $\lambda \to \infty$. However, we have found that it is also important to take account of the number of parameters defining L that are estimated from the data, and consequently, we define the degrees of freedom to be

$$df_\lambda = \text{trace } \mathbf{S}_\lambda + \text{number of estimated parameters.} \tag{10}$$

The size of λ depends on what L is employed, and especially on m, but the effective number of parameters provides a basis for comparing fits based on different penalties. Figure 2 shows three penalized least squares estimates of melanoma incidence, each using ten effective parameters. The three estimates have penalties with $L = D^2$, D^3 and D^4. The values of λ are 4.35, 35.69, and 910.52, respectively.

FIGURE 2 HERE.

The effective number of parameters is also used when choosing λ by generalized cross-validation (see, for instance, Green & Silverman, 1994, or Eubank, 1988). The value of λ is chosen to minimize

$$\text{GCV}(\lambda) = \frac{nSSE(\lambda)}{(n - df_\lambda)^2} = \frac{n}{n - df_\lambda}\hat{\sigma}^2(\lambda) \tag{11}$$

where $SSE(\lambda) = (\mathbf{Y} - \hat{\boldsymbol{\mu}}_\lambda)'\mathbf{D}(\mathbf{Y} - \hat{\boldsymbol{\mu}}_\lambda)$ and $\hat{\sigma}^2(\lambda) = SSE(\lambda)/(n - df_\lambda)$. The theoretical justification of minimizing GCV is that GCV is a good measure of the predictive power of $\hat{\mu}_\lambda$. A heuristic justification of the use of GCV is as follows. Usually, the numerator of GCV is small (that is, $\hat{\mu}_\lambda$ is close to interpolating the data) when the denominator is small (when df_λ is close to n). Thus minimizing GCV means fitting the data well with few parameters.

We may also compare the fit achieved by the favored model to that achieved using a specific λ via the F-ratio:

$$\text{F}(\lambda) = \frac{\frac{SSE(\infty) - SSE(\lambda)}{df_\lambda - df_\infty}}{\frac{SSE(\lambda)}{df_\lambda}}. \tag{12}$$

Here $SSE(\infty)$ and df_∞ indicate, respectively, the error sum of squares and degrees of freedom for the favored model.

4. ESTIMATION OF THE PENALTY

In practice, we usually have a general idea of the favored functions u_j, but we may need to estimate a parameter vector θ before these functions, properly denoted as $u_j(t;\theta)$, can be completely specified. This was the case for the melanoma data, where the period ω was unknown. Note that any parameterization of the u_j's induces a parameterization of the coefficient functions, the w_j's, and vica versa.

The parameter θ can be estimated by using nonlinear least squares techniques to minimize $\sum_i (Y_i - \sum_j \alpha_j u_j(t_i;\theta))^2$ as a function of the α_j's and θ. One then uses the corresponding $\hat{w}_j$'s in (3) and (4) to estimate μ by $\hat{\mu}_\lambda$. The smoothing parameter can then be chosen by methods such as generalized cross-validation. For $L = D^4 + \omega^2 D^2$ with ω^2 unknown, we estimated the parameter ω by minimizing $\sum (Y_i - \alpha_1 - \alpha_2 t_i - \alpha_3 \cos \omega t_i - \alpha_4 \sin \omega t_i)^2$ as a function of ω and the α_j's.

We also explored other methods, such as minimizing the GCV criterion with respect to both λ and the parameters θ. However, our experience has been that GCV often has multiple local minima and exhibits rather complex behavior as a function of both λ and θ, and the preliminary nonlinear least squares analysis was found to be more reliable.

Of course, the nonlinear least squares method isn't appropriate if the data are far from the favored model. However, in this case there is no advantage to using the penalty associated with the favored model.

When the GCV criterion is used to select λ and possibly θ, we have tended to add the number of estimated parameters in θ to the degrees of freedom measure, as in equation (10). This has the practical problem that the degrees of freedom for error $n - df_\lambda$ in the denominator of GCV become negative as $\lambda \to 0$. However, when the data call for such a small amount of smoothing, we suspect that there is little to be gained over polynomial smoothing by our methods, so this situation should not prove troublesome in realistic situations.

5. EXAMPLE OF L-SPLINE SMOOTHS

5.1. PERIODIC DATA WITH LINEAR TREND: MELANOMA DATA

To analyze the melanoma data, we use the fact that melanoma incidence is believed to be roughly periodic with a linear trend. Therefore, we choose the preferred parametric model $\mu(t) = \alpha_1 + \alpha_2 t + \alpha_3 \cos \omega t + \alpha_4 \sin \omega t$ and, with $d_i = 1$, (4) becomes

$$H(\mu, \omega) = \sum_i \left(Y_i - \mu(t_i)\right)^2 + \lambda \int \left[(D^4 + \omega^2 D^2)\mu\right]^2 . \qquad (13)$$

Our nonlinear least squares estimate of μ yielded $\hat{\omega} = 0.58$, corresponding to a period of 10.8 years. Using this value of ω in (13), minimizing with respect to μ and choosing λ by generalized cross-validation yields the estimator labelled PER in Figure 1. This estimate of μ corresponds to trace $\mathbf{S}_\lambda = 4.6$, for a total of 5.6 parameters after correcting for estimating ω.

Figure 3 contains our periodic penalty nonlinear least squares estimate (the PER estimate of Figure 1) along with three other estimates. The nonlinear least squares estimate (denoted NLS) is almost identical to the periodic penalty estimate. The other two estimates use a penalty with $L = D^4$, so that this penalty is of the same order as our periodic penalty. One estimate (labelled GCV) chooses λ by generalized cross-validation. The other estimate (labelled DF) chooses λ by forcing the number of parameters to be equal to 5.6, which is equal to the number of parameters used by our NLS estimate. We see that the DF estimate captures the linear trend in the data but completely misses the periodic component. The GCV estimate tracks the data very well. However, it uses 17 parameters, a very high number considering that there are only 37 data points.

FIGURE 3 HERE

The F-ratio function (12) is given in Figure 4. We see that the L-spline does not improve on the favored model for any of the values of λ that are plotted, while the polynomial spline, not surprisingly, improves significantly on the fit offered by its favored model over a wide range values.

FIGURE 4 HERE

Figure 3 suggests the use of a different favored model, namely $\mu(t) = \alpha_1 + \alpha_2 t + \alpha_3 t \cos \omega t + \alpha_4 t \sin \omega t$. Modification of the analysis for this favored model is straightforward: one simply needs to calculate the reproducing kernel k of Section 2.3. Note that this favored model yields a differential operator L with nonconstant coefficients w_j. To calculate the w_j's, see the comments just after equation (5).

An alternative, parametric analysis of these data might involve time series methods. This would be particularly useful for predicting melanoma incidence given past incidence, that is, for estimating the conditional expectation of melanoma incidence given the past. However, here our goal is to estimate the unconditional expected melanoma incidence. Even in this case, one might assume a time series model: $Y_i = \alpha_1 + \alpha_2 t_i + \alpha_3 \cos \omega t_i + \alpha_4 \sin \omega t_i + \epsilon_i$ where the ϵ_i's follow, say, an AR(p) model with p unknown. We carried out analysis with this model, by studying the residuals of the nonlinear least squares estimate of μ. From plots and from analysis of the residuals via Splus's *ar* function, we conclude that the residuals are indeed independent, that is, that $p = 0$.

The data analysis naturally leads us to the question "does the parametric model provide an adequate fit to the data?". One can answer this question using ad hoc tests based on the F-statistic or with chi-squared statistics, both of which are recommended for model-testing in generalized additive models (see Hastie & Tibshirani, 1990). Or one can use the tests proposed by Cox, Koh, Wahba & Yandell (1988) or Cox & Koh (1989). These are locally most powerful tests, under a Bayesian inspired alternative. Ramsay (1998) used a bootstrapping technique to test a fit for monotonicity.

5.2 US BIRTH RATE DURING WORLD WAR II

The monthly US birth rate for the period January 1940 to December 1948 is plotted in Figure 5. These data were discussed by Simonoff (1996) in the context of variable bandwidth smoothing. We see here three trends. The twice-yearly surge in births is due to enhanced opportunities for gestation afforded by the Christmas and summer holidays. The overall trend can realistically be viewed as exponential. Finally, there are the dramatic effects due to young parents' reactions to the prospect of being involved in World War II, to their enlistment, and to their demobilization. The values are counts and so one would expect that the variances would be increasing with the mean. Therefore we carry out a weighted smoothing with weights chosen assuming a Poisson distribution.

FIGURE 5 HERE

We propose the favored model defined by the three functions

$$\begin{aligned} u_1(t) &= \sin(\pi t/3) \\ u_2(t) &= \cos(\pi t/3) \\ u_3(t) &= \exp\{c_1 t + c_2 F[(t-a)/b]\} \end{aligned} \tag{14}$$

where $F(z) = z/(1+z^2/2)^2$, and time t is measured in months from January 1940. The first two functions accommodate sinusoidal effects with a period of six months, and the third is designed to pick up exponential trend plus the decline followed by the surge in birth rate. The parameters a and b define the location and spread of the enlistment/demobilization effect, and these plus c_1 and c_2 make four parameters to estimate. We estimated these by fitting u_3 alone to the data by nonlinear least squares. This fit is plotted as a dotted line in Figure 5.

These three functions define an order three L with non-constant coefficients. Following the procedure described in the Appendix, we first computed the determinant $|\mathbf{W}(t)|$ of the Wronskian matrix and confirmed that it was nowhere 0 over the interval $[0, 96]$. Then we computed the three adjoint functions, $u_1^*(t), u_2^*(t)$, and $u_3^*(t)$, followed by the matrix function $\mathbf{F}(t)$ defined in (20). The reproducing kernel values $k(t_i, t_j)$ were then computed from (19) by numerical integration. We augmented the degree of freedom measure trace $\mathbf{S}_\lambda$ by four in computing GCV. For comparison purposes, we also smoothed the data using an order 3 polynomial spline; that is, using the penalty $P(\mu) = \int (D^3\mu)^2$. This was fit using the *Pspline* module. For both fits, we accounted for nonstationarity of variance by smoothing with weights d_i proportional to $1/\hat{Y}_i$ where $\hat{Y}$ was defined by fitting the data with u_3 alone.

The minimum GCV estimators for both the L-spline and P-spline, shown in Figure 6, corresponded to about 55 degrees of freedom. These corresponded to global standard error estimates computed by $\hat{\sigma}^2 = \sum [Y_i - \hat{\mu}(t_i)]^2/(n - df_\lambda)$ of about $\hat{\sigma} = 24$, and the two fits are hardly distinguishable. But this is surely too close a fit, since for a count of 2500, we should expect a standard error of about 50, even ignoring other contributions to variance such as economic and climate factors influencing birth rate.

Figure 6 displays the L-spline and P-spline fits corresponding to 20 degrees of freedom. These correspond to the more reasonable standard error estimates of 73.4 and 76.3, respectively. While both offer reasonably good accounts of the large-scale effects in the data, the L-spline is obviously able to pick up the small-scale harmonic effects much better, while the P-spline fit seems slightly preferable for the birth rate spurt at the beginning of 1943. Also shown in the figure is the least squares fit, using the basis in (14).

FIGURE 6 HERE

Figure 7 plots the variance estimate $\hat{\sigma}^2$ as a function of degrees of freedom for the two types of smooths. We see that the L-spline has a substantial advantage over the P-spline fit up to about 20 degrees of freedom, after which the two methods offer similar levels of fit. This plot along with our experience suggests that the intelligence in the L-spline fit only pays off in terms of goodness of fit for small degrees of freedom, but for medium degrees of freedom, such as 20 in this problem, the L-spline fit may be better for qualitative reasons, and finally for large degrees of freedom both methods will give about the same results.

FIGURE 7 HERE

We acknowledge that both the melanoma and birthrate data may be better approached by time series methods; our intention here is only to compare various smoothing techniques.

6. SIMULATIONS

Through simulations, we can investigate more precisely whether the appropriate choice of penalty can substantially increase the accuracy of estimation. We used regression functions that are similar to the favored model $\mu(t) = \alpha_1 + \alpha_2 t + \alpha_3 \cos\omega t + \alpha_4 \sin\omega t$. We compared the following estimates of the regression function.

NLS : nonlinear least squares estimate based on the parametric default model;

PER.NLS : penalized least squares estimate based on the appropriate penalty as in (4) with $L = D^4 + \omega^2 D^2$; the parameter ω^2 used in the penalty is that calculated in *NLS* and λ is chosen by GCV (this is method 1 of Section 4);

D^4-***GCV*** : penalized least squares estimate based on the penalty with $L = D^4$; λ is chosen by GCV; the estimates are smoothing splines of degree 7;

D^4-***df*** : penalized least squares estimate based on the penalty with $L = D^4$; λ is chosen so that the number of parameters used is the same as in *PER.NLS*, that is, we use number of parameters equal to the trace of the hat matrix plus one, for the estimation of ω; the estimates are smoothing splines of degree 7.

In the three penalized estimates, the weights in (4) were $d_i = 1$.

We generated 400 random functions, μ_j, $j = 1, \cdots, 400$, as described in Section 6.1 below. For each μ_j we created 200 data sets. The kth data set contained 101 observations of the form

$$Y_{ki} = \mu_j(i/100) + \sigma\epsilon_{ki}, \quad i = 0, \cdots, 100 \tag{15}$$

where the ϵ_{ki}'s are independent standard normal random variables and σ is either 5 or 20. From this dataset we calculated $\hat{\mu}_{kj}$.

We studied the bias and variance of our estimates as follows. From the 200 estimates we calculated the average estimate

$$\bar{\hat{\mu}}_j(t) = \frac{1}{200}\sum_{k=1}^{200} \hat{\mu}_{kj}(t),$$

the average squared bias

$$\mathrm{BIAS}_j^2 = \frac{1}{97}\sum_{i=2}^{98}\left[\bar{\hat{\mu}}_j(i/100) - \mu_j(i/100)\right]^2,$$

and the average variance

$$\mathrm{VAR}_j = \frac{1}{97 \times 200}\sum_{k=1}^{200}\sum_{i=2}^{98}(\hat{\mu}_{kj}(i/100) - \bar{\hat{\mu}}_j(i/100))^2.$$

We've eliminated the t_i's near 0 and 1 in order to focus our attention on accuracy of estimation away from the boundary of the data. Results summing over all values were qualitatively similar to those presented here. Note that

$$\mathrm{BIAS}_j^2 + \mathrm{VAR}_j = \frac{1}{97 \times 200} \sum_{k=1}^{200} \sum_{i=2}^{98} (\hat{\mu}_{kj}(i/100) - \mu_j(i/100))^2 \equiv \mathrm{MSE}_j.$$

For the penalty with $L = D^4 + \omega^2 D^2$, the favored model is spanned by the functions 1, t, $\cos \omega t$, and $\sin \omega t$.

Figure 8 shows one of the curves used along with a data set generated from this curve by adding normal regression errors with standard deviation equal to 5.

FIGURE 8 HERE.

6.1. RANDOM FUNCTION GENERATION

We want to generate functions that are somewhat periodic, to illustrate the importance of the choice of penalty. However, we don't want exactly periodic functions, since then a parametric procedure would be most appropriate.

There are undoubtedly many ways to generate roughly periodic functions. Here, we use a somewhat ad hoc method. We'd like μ to be a random process with a period that varies randomly in time, but is approximately equal to ω. Specifically we'd like random processes $\mathrm{w}(\cdot)$ and $\mu(\cdot)$ with $L\mu \equiv 0$ where $(L\mu)(t) = \mu^{(4)}(t) + \mathrm{w}^2(t)\mu^{(2)}(t)$ and with

$$\mathrm{E}(\mathrm{w}(t)) \equiv \omega. \tag{16}$$

However, for $\mathrm{w}(\cdot)$ not constant, there is no general closed form solution of the differential equation $L\mu = 0$. Therefore, we replace $L\mu = 0$ by

$$\mathrm{E}(\mathrm{L}\mu) \equiv 0. \tag{17}$$

Let $\epsilon(\cdot)$ be white noise with standard deviation σ_ϵ, $b(t) = \int_0^t \epsilon(s)\, ds$ be Brownian motion, and $B(t) = \int_0^t b(s)\, ds$. Then (16) and (17) are satisfied for $\mathrm{w}(t) = \omega + b(t)$ and $\mu(t) = C\left[D^{-2} \cos\left(\omega t + B(t)\right)\right]$, C an arbitrary constant. Finally, we generate a random μ by

$$\mu(t) = C\left[\alpha_0 + \alpha_1 t + D^{-2}\cos(\omega t + B(t))\right] \tag{18}$$

where α_0 and α_1 are independent normal random variables, independent of $B(\cdot)$, with mean zero and standard deviation σ_α. To approximate the process $B(\cdot)$, we first approximate $b(\cdot)$ using linear interpolation of partial sums of independent normal deviates. We then numerically integrate $b(\cdot)$ by the trapezoidal rule. We also approximate $D^{-2}\cos(\omega t + B(t))$ by numerical integration. Figure 9 contains four of the 1000 curves generated using $\omega = 8\pi$, $\sigma_\epsilon = 50$, $\sigma_\alpha = 0.01$, and $C = 10{,}000$, the values used in our simulations. The values of these parameters were chosen in order to generate curves that were close to the favored parametric model, but not too close, and the value of C gave MSE values of a reasonable order of magnitude. The solid curve in the figure is calculated from (18) with $B(t) \equiv 0$, that is, $\mu(t) = 10{,}000\ D^{-2}\ \cos(\omega t) = 10{,}000\ (1 - \cos(\omega t))/\omega^2$.

FIGURE 9 HERE.

6.2. SIMULATION RESULTS

Results are given in Tables 2 ($\sigma = 5$) and 3 ($\sigma = 20$). We define the number of parameters used for *PER.NLS* as the trace of the hat matrix plus one, thus allowing for the estimation of ω. Our results show that the *PER.NLS* estimate has small mean squared error and uses a reasonable number of parameters. Since the distribution of the mean squared error is sometimes long-tailed, we have also included summary statistics for the logarithm base 10 of the MSE.

In terms of MSE, *PER.NLS* is by far the best method when $\sigma = 5$ and, when $\sigma = 20$, it is comparable to the best method *NLS*. *PER.NLS* tends to use a few more parameters than D^4-GCV.

The MSE of D^4-*df* is larger than that of *PER.NLS*, since D-*df* is forced to use the same number of parameters as *PER.NLS* but with the "incorrect" penalty.

We can study how much of the MSE is due to BIAS^2 and how much is due to VAR. As might be expected, the bias of *PER.NLS* is small compared to the variance, since *PER.NLS* uses a favored model that is appropriate for our regression function. When $\sigma = 5$, D^4-GCV has the ratio closed to *PER.NLS*. This might be expected, since these two estimators choose the smoothing parameter λ by generalized cross-validation which, roughly, minimizes the MSE by choosing a value of λ which gives the appropriate balance between the squared bias and the variance. However, the ratios are further apart when $\sigma = 20$. From plots of the estimates, it seems that the high noise level forces D^4-GCV to try to fit something close to a line. This would result in a higher ratio of squared bias to variance. This observation is supported by the fact that, when $\sigma = 20$, D^4-GCV uses a small number of parameters. So evidently, D^4-GCV attributes the periodic component of μ to noise.

TABLES 2-3 HERE.

APPENDIX: THE $O(n)$ ALGORITHM

Let k, $\mathbf{U}$, $\mathbf{K}$, and $\mathbf{D}$ be as in Section 2.3 and $\hat{\boldsymbol{\mu}}_\lambda = (\hat{\mu}_\lambda(t_1), \cdots, \hat{\mu}_\lambda(t_n))'$. In what follows, we shall only need the diagonal and first $2m$ off-diagonals of $\mathbf{K}$.

It can be shown that that, for some $\mathbf{c} \in \Re^{n-m}$, $\hat{\boldsymbol{\mu}}_\lambda$ satisfies

$$\hat{\boldsymbol{\mu}}_\lambda = \mathbf{Y} - \lambda \mathbf{Q}\mathbf{c}$$

where

$$(\mathbf{Q}'\mathbf{DKDQ} + \lambda \mathbf{Q}'\mathbf{DQ})\mathbf{c} = \mathbf{Q}'\mathbf{DY}$$

and $\mathbf{Q}$ is any n by $n-m$ matrix of full column rank with $\mathbf{Q}'\mathbf{DU} = 0$.

If the matrices $\mathbf{Q}'\mathbf{DKDQ}$ and $\mathbf{Q}$ are banded, then we can use these equations to solve for $\hat{\boldsymbol{\mu}}_\lambda$ in $O(n)$ calculations. In fact, if $\mathbf{Q}_{k+i,k} = 0$ for $i \neq 0, \cdots, m$, then we can show, by the properties of $\mathbf{K}$, that $[\mathbf{Q}'\mathbf{DKDQ}]_{k,k+i} = 0$ for $|i| > m$, as follows. Using (6) and (7), write

$$\begin{aligned} \mathbf{K}_{ij} \quad &= \textstyle\int G(t_i, w) G(t_j, w)\, dw \\ &\equiv \textstyle\sum_{r,s} \mathbf{U}_{ir} \mathbf{U}_{js} \mathbf{F}_{rs}(\min\{t_i, t_j\}) \end{aligned} \tag{19}$$

where

$$\mathbf{F}_{rs}(t) = \int_a^t u_r^*(w) u_s^*(w)\, dw. \tag{20}$$

Since $\mathbf{Q}'\mathbf{DKDQ}$ is symmetric, it suffices to show that $[\mathbf{Q}'\mathbf{DKDQ}]_{k\ell} = 0$ for $k - \ell > m$.

$$\begin{aligned} [\mathbf{Q}'\mathbf{DKDQ}]_{k\ell} &= \sum_{i,j=1}^{n} [\mathbf{DQ}]_{ik} \mathbf{K}_{ij} [\mathbf{DQ}]_{j\ell} \\ &= \sum_{i,j=0}^{m} [\mathbf{DQ}]_{k+i,k} \mathbf{K}_{k+i,\ell+j} [\mathbf{DQ}]_{\ell+j,\ell}. \end{aligned}$$

Since $k - \ell > m \geq j - i$ whenever $0 \leq i, j \leq m$, in the above summation we have $k + i > \ell + j$. So

$$\begin{aligned} [\mathbf{Q}'\mathbf{DKDQ}]_{k\ell} &= \sum_{i,j=0}^{m} [\mathbf{DQ}]_{k+i,k} \sum_{r,s=1}^{m} \mathbf{U}_{k+i,r} \mathbf{U}_{\ell+j,s} \mathbf{F}_{r,s}(t_{\ell+j}) [\mathbf{DQ}]_{\ell+j,\ell} \\ &= \sum_{j=0}^{m} \sum_{r,s=1}^{m} \mathbf{F}_{r,s}(t_{\ell+j}) \mathbf{U}_{\ell+j,s} [\mathbf{DQ}]_{\ell+j,\ell} \sum_{i=0}^{m} [\mathbf{DQ}]_{k+i,k} \mathbf{U}_{k+i,r}. \end{aligned}$$

But $\sum_{i=0}^{m} [\mathbf{DQ}]_{k+i,k} \mathbf{U}_{k+i,r} = [\mathbf{Q}'\mathbf{DU}]_{kr} = 0$.

To find a banded $\mathbf{Q}$ with $\mathbf{Q}'\mathbf{DU} = 0$, set $\mathbf{Q}_{ij} = 0$ for $i \neq j, j+1, \cdots, j+m$. Thus the remaining elements of $\mathbf{Q}$ must satisfy $\sum_{\ell=0}^{m} \mathbf{Q}_{i+\ell,i} d_{i+\ell} u_j(t_{i+\ell}) = 0$ for all $j = 1, \cdots, m$. For each i, the values $\mathbf{Q}_{i+\ell,i}$, $\ell = 0, \cdots, m$, are easily found by a QR-decomposition of $(\mathbf{DU})_i$, the $m+1$ by m matrix with ℓjth entry equal to $d_{i+\ell} u_j(t_{i+\ell})$. Write the decomposition as $(\mathbf{DU})_i = \mathbf{Q}_i \mathbf{R}_i$. Then the required values are in the $(m+1)$st column of $\mathbf{Q}_i$.

Once the coefficients $\mathbf{c}$ are in hand, one can go on to compute the spline function $\hat{\mu}_\lambda$ or its derivative at any other set of argument values. First, note that $\mathbf{Q}'\mathbf{D}(\hat{\boldsymbol{\mu}}_\lambda - \mathbf{KDQc}) = 0$, so there exists $\mathbf{d} \in \Re^m$ with $\hat{\boldsymbol{\mu}}_\lambda - \mathbf{KDQc} = \mathbf{Ud}$, and $\mathbf{d}$ can readily be computed. Now let $\mathbf{U}^{(\ell)}$ and $\mathbf{K}^{(\ell)}$ contain the values $\mathbf{U}_{ij}^{(\ell)} = D^\ell u_j(\cdot)$ and

$\mathbf{K}_{ij}^{(\ell)} = D^\ell k(t_j, \cdot)$ evaluated at t_i, where the argument values t_i may now be different in value and number from those in the data. Then the ℓth derivative evaluated at these values is

$$\hat{\boldsymbol{\mu}}_\lambda^\ell = \mathbf{U}^{(\ell)}\mathbf{d} + \mathbf{K}^{(\ell)}\mathbf{DQc}\ .$$

The matrix $\mathbf{Q}'$ can be thought of as a generalized divided difference matrix associated with L, since $\mathbf{Q}'\mathbf{D}(f(t_1), \cdots, f(t_n))' = 0$ for the original set of t_i's whenever f is in the kernel of L. For instance, suppose that $\mathbf{D}$ is the identity matrix and that the t_i's are evenly spaced with $t_{i+1} - t_i = \Delta$. If $m = 2$ and $L = D^2$, then $u_1(t) = 1$ and $u_2(t) = t$ span the kernel of L and we can take $\mathbf{Q}'_{ii} = \mathbf{Q}'_{i,i+2} = 1$, $\mathbf{Q}'_{i,i+1} = -2$, and $\mathbf{Q}'_{ij} = 0$ otherwise. Thus $\mathbf{Q}'$ is the usual second divided difference matrix. If $m = 2$ and $L = D^2 + \omega^2 I$, then $u_1(t) = \cos\omega t$ and $u_2(t) = \sin\omega t$ span the kernel of L. We can take $\mathbf{Q}'_{ii} = \mathbf{Q}'_{i,i+2} = 1$, $\mathbf{Q}'_{i,i+1} = -2\cos\omega\Delta$ and $\mathbf{Q}'_{ij} = 0$ otherwise.

ACKNOWLEDGEMENTS

The authors thank the editor, associate editor and referees for helpful comments that let to the improvement of the manuscript. This research is supported by the Natural Sciences and Engineering Research Council of Canada through grants A7969 and A320 respectively.

REFERENCES

Andrews, D. F. & Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, New York: Springer Verlag.

Anselone, P. M., & Laurent, P. J. (1968). A general method for the construction of interpolating or smoothing spline functions. *Numer. Math.*, 12, 66–82.

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

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

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74, 829–836.

Coddington, E. A. (1961). *An Introduction to Ordinary Differential Equations.* New Jersey: Prentice-Hall.

Cox, D. D. & Koh, E. (1989). A smoothing spline based test of model adequacy in polynomial regression. *Ann. Inst. Statist. Math.*, 41, 383–400.

Cox, D. D., Koh, E., Wahba, G., & Yandell, B. S. (1988). Testing the (parametric) null hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.*, 16, 113–119.

de Jong, P. (1997). Fixed interval smoothing. Manuscript.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.*. 87, 998–1004.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman and Hall, London.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman Hall, London.

Hutchison, M. F. & de Hoog, F. R. (1985). Smoothing noisy data with spline functions. *Num. Math.*, 47, 99–106.

Kimeldorf, G. S. & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41, 495–502.

Kohn, R. & Ansley, C. (1989) A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika*, 76, 65-69.

Kohn, R. & Ansley, C. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Statist. Comp. and Sim.*, 86, 1042–150.

Kohn, R. & Ansley, C. (1993). Accuracy and efficiency of alternative spline smoothing algorithms. *J. Statist. Comp. and Sim.*, 46, 1–18.

Kohn, R., Ansley, C., & Wong, C. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika*, 79, 335–346.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *J. Roy. Statist. Soc. Ser. B,* 365–375.

Reinsch, C. (1967). Smoothing by spline functions. *Num. Math.*, 10, 177–183.

Reinsch, C. (1971). Smoothing by spline functions II. *Num. Math.*, 16, 451–454.

Schultz, M. H. & Varga, R. S. (1967). *L*-splines. *Num. Math.,* 10, 345–369.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory.* John Wiley and Sons, New York.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer, New York.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B*, 40, 364–372.

Wahba, G. (1990). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia.

Wang, Y. & Brown, M. B. (1996). A flexible model for human circadian rhythms. *Biometrics*, 52, 588–596.

Wecker, W. E. & Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Amer. Statist. Assoc.,* 78, 81–89.

Weinert, H. L. Byrd, R. H. & Sidhu, G. S. (1980). A stochastic framework for recursive computation of spline functions: Part II, smoothing splines. *J. Optim. Theory and App.,* 2, 255–268.

Table 1: Examples of differential operators and bases for the corresponding parametric families.

Operator L	Parametric Family for $\ker L$
D^2	$\{1, t\}$
D^4	$\{1, t, t^2, t^3\}$
$D^2 + \gamma D$, $\gamma \neq 0$	$\{1, \exp(-\gamma t)\}$
$D^4 + \omega^2 D^2$, $\omega \neq 0$	$\{1, t, \cos \omega t, \sin \omega t\}$
$(D^2 - \gamma D)(D^2 + \omega^2 D), \gamma, \omega \neq 0$	$\{1, \exp \gamma t, \cos \omega t, \sin \omega t\}$
$D - w(\cdot)I$, $w(t) \neq 0$	$\{\exp(\int_0^t w(u)\ du)\}$
$D^2 - w(\cdot)D$, $w(t) \neq 0$	$\{1, \int_0^t [\exp(\int_0^u w(v)\ dv)]\ du\}$

Table 2a

MSE: $\sigma = 5$								
	MSE			Log_{10} MSE		Total	Total	Ratio
	Mean	SE	Median	Mean	SE	Bias^2	Variance	Bias^2/Var
NLS	6.57	0.01	6.52	0.816	0.001	6.3	0.3	20.9
PER.NLS	2.42	0.04	2.17	0.33	0.01	0.3	2.1	0.1
D^4-*GCV*	21.4	0.5	33.3	1.10	0.02	12.5	8.9	1.4
D^4-*df*	16.9	0.3	16.7	1.12	0.01	11.5	5.3	2.2

Table 2b

DF: $\sigma = 5$				
	Mean	SE	Median	Range
PER.NLS	9.5	0.1	8.8	6.0 – 34.2
D^4-*GCV*	8.2	0.1	4.6	4.0 – 28.0

Results are based on 200×400 simulated data sets using 400 random curves, with regression error $\sigma = 5$. SE = standard deviation/$\sqrt{1000}$.

Table 3a

MSE: $\sigma = 20$								
	MSE			Log_{10} MSE		Total	Total	Ratio
	Mean	SE	Median	Mean	SE	Bias^2	Variance	Bias^2/Var
NLS	38.1	0.8	28.9	1.495	0.009	6.2	31.9	0.19
PER.NLS	40.0	0.8	33.3	1.508	0.010	1.56	38.5	0.04
D^4-*GCV*	48.9	0.4	46.2	1.679	0.003	31.0	17.9	1.73
D^4-*df*	51.9	0.5	48.1	1.699	0.004	20.3	31.6	0.64

Table 3b

DF: $\sigma = 20$				
	Mean	SE	Median	Range
PER.NLS	7.54	0.14	5.74	5.0 – 44.8
D^4-*GCV*	4.48	0.04	4.00	4.0 – 14.7

Results are based on 200×400 data sets, using 400 random curves, with regression error $\sigma = 20$. SE = standard deviation/$\sqrt{1000}$.

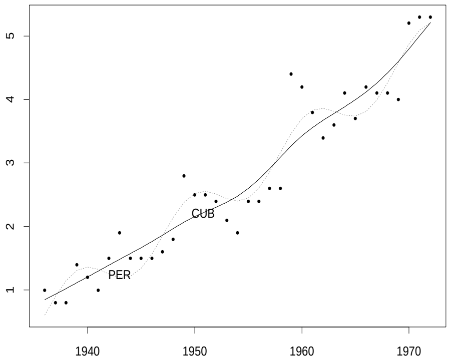

Figure 1: Male Melanoma Data. Estimates are the solid line, which is CUB, the cubic smoothing spline estimate, and the dashed line, which is PER, the estimate defined in Section 5.1. Each estimate uses 5.6 parameters.

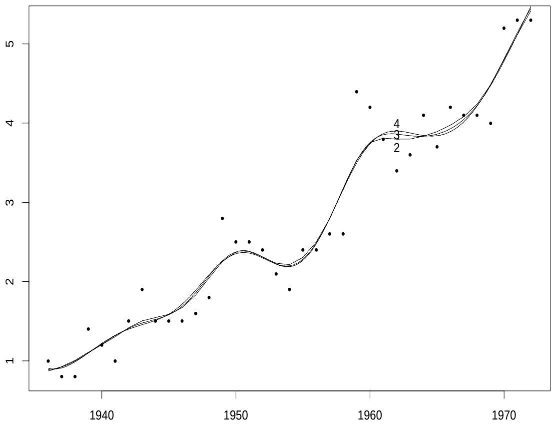

Figure 2: Estimates of melanoma incidence using penalties with L equal to D^2 (labelled 2), D^3 (labelled 3), and D^4 (labelled 4). All estimates use 10 parameters.

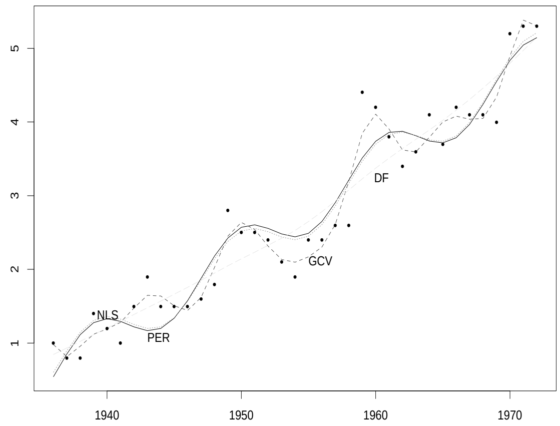

Figure 3: Male Melanoma Data. Estimates are the parametric nonlinear least squares estimate (the dashed line denoted NLS), the penalized periodic smoothing spline (the solid line denoted PER) with period estimated by parametric non-linear least squares and λ chosen by GCV, the penalized estimate using $L = D^4$ (denoted GCV) with λ chosen by GCV and the penalized estimate using $L = D^4$ (denoted DF) and λ chosen so that the number of parameters is equal to that used by PER.

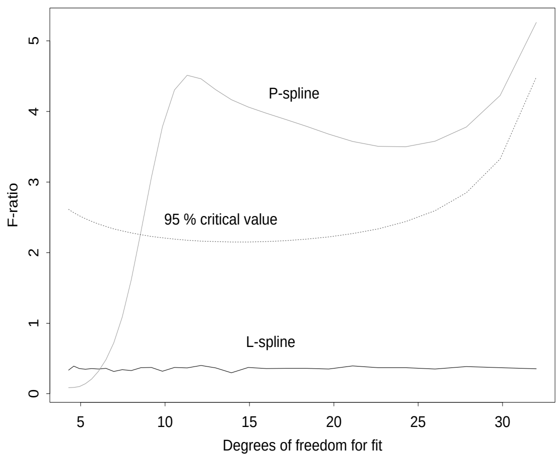

Figure 4: The F-ratio functions (12) for the male melanoma data for the L-spline and order 4 polynomial spline.

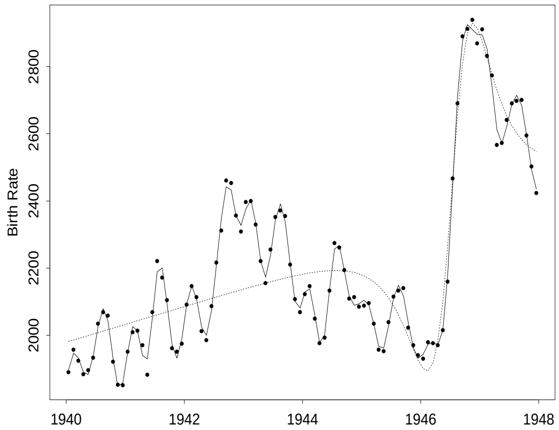

Figure 5: The points show the US monthly birthrate around World War II. The solid curve is the L-spline smooth corresponding to the minimum GCV value using as favored functions (14), and the dotted line indicates the nonlinear least squares fit achieved by using favored function $u_3(t)$ alone.

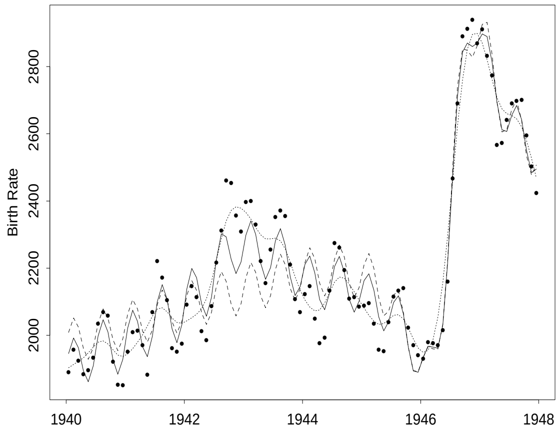

Figure 6: The solid curve is the L-spline smooth of the birth rate data corresponding to 20 degrees of freedom, the dashed-dotted line is the least-squares fit using only the favored functions (14), and the dotted line is the order 3 polynomial spline fit, also using 20 degrees of freedom.

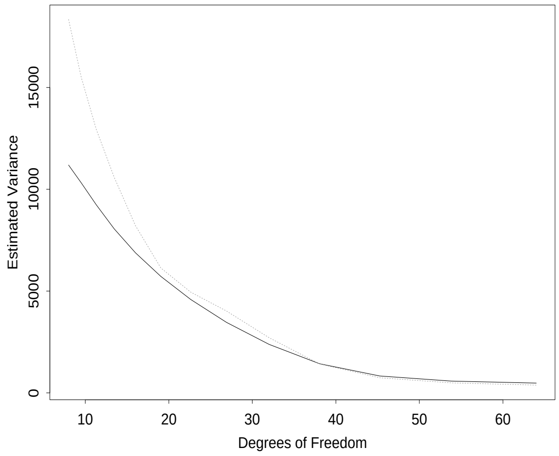

Figure 7: The estimate variance of estimate $\hat{\sigma}^2 = SSE/(n - df_\lambda)$ plotted as a function of degrees of freedom df_λ for the fit. The solid line is for the L-spline, and the dotted for the order 3 polynomial spline.

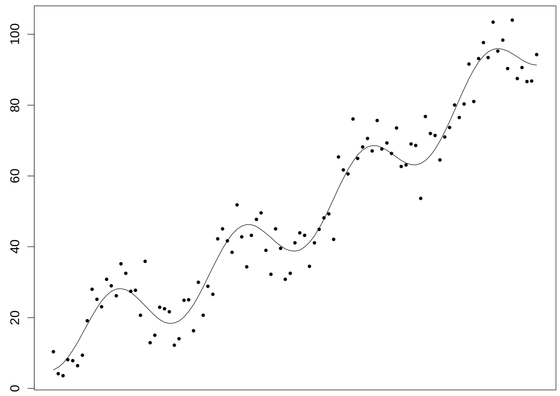

Figure 8: Data set generated from a random curve, with standard deviation of regression error equal to 5.

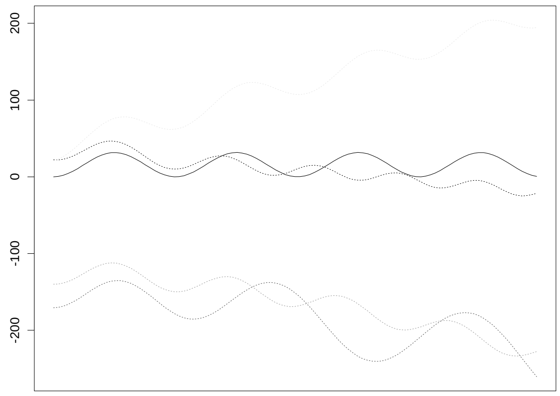

Figure 9: Four randomly generated functions (dashed) with “true” curve (solid).