International Journal of Statistics and Management System
Vol. 4 No. 1 (January-June, 2019)

Partioned spline estimators for growth curve estimation in wildlife studies*

William P. Brown†, Paul P. Eggermont‡, Vincent N. LaRiccia‡ and Roland R. Roth‡

Received: 03rd September 2018 Revised: 14th December 2018 Accepted: 10th March 2019

Abstract

Nonparametric regression estimators are often employed to estimate growth curves. However, more than one smoothing parameter may be required to estimate growth curves for some species, particularly those with distinct life-stages. This can be problematic, especially if confidence intervals about the mean function are also required. Here a straightforward method, based on the spline estimator, is proposed. First, the natural history of the species in question is used to determine a finite number of possible partitions of the growth curve. If needed the partitions are then adjusted so that they overlap. Next, a spline estimator is fit to each partition. Finally the individual estimators of the growth curve in each partition are blended together to form one coherent curve.

The resulting estimator remains a linear function of the data and converges to the true function at the same rate as the optimal single parameter spline estimator. Further, mild conditions on how the partitions have been selected, and on the blending step are given, which ensure that the associated fiducial confidence intervals are asymptotically valid. Finally, data driven methods for making all of the required decisions are given.

Key words and phrases: spline smoothing, confidence intervals, growth curves.

†*Mailing Address:* Department of Biology, Kutztown University, Kutztown, PA 19530, USA. *E–mail:* wpbrown23@yahoo.com

‡*Mailing Address:* Department of Food and Resource Economics, University of Delaware, 213 Townsend Hall Newark, DE 19716, USA. *E–mail:* eggermon@udel.edu, lariccia@udel.edu.

1 Introduction. The usefulness of spline estimators of the mean function in growth curve studies is well documented. See for example Gasser, Muller, Köhler, Molinari, and Prader [9], Muller [12], Aggery [2], Przybylski and Garcia-Berthou [15], Schillaci and Stallmann [17], and Brown, Eggermont, LaRiccia, and Roth [5]. In growth studies the dependent variable, Y equals some measure of the size of an individual, and the independent random variable, X equals the age of the individual. The variables are assumed to be related by the typical regression model

$$Y = f_0(X) + \varepsilon, \tag{1.1}$$

where $f_0(\cdot)$ is an unknown smooth function and ε is an independent error term. As in all regression problems the goal is to estimate f_0 based on a sample of n observations,

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Also, the associated confidence intervals for the $f_0(X_i)$'s are desired.

From (1.1) it will be assumed that

$$Y_i = f_0(X_i) + \varepsilon_i, \tag{1.2}$$

with

$$\varepsilon_1, \cdots \varepsilon_n, \ i.i.d. \tag{1.3}$$

$$E[\varepsilon_1] = 0, E[\varepsilon_1^2] = \sigma^2 \text{ and } E[\varepsilon_1^4] < \infty. \tag{1.4}$$

Conditions on the design, $X_1, \cdots, X_n$, and f_0 are given below.

For a given value of $\lambda > 0$, the m^{th} order spline estimator, f_λ, is defined as the unique solution to the problem

$$\min_{s.t.\ f \in\ W^{m,2}} n^{-1} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda^{2m} \int_A |f^m(x)|^2 \, dx, \tag{1.5}$$

where f^m denotes the m^{th} derivative of f, and

$$\begin{aligned} W^{m,2} \ = \ \Big\{ & f : f^i \text{ is absolutely continuous for } i = 0, 1, \cdots, m-1 \\ & \text{and } \int (f^m(x))^2 dx < \infty \Big\} \end{aligned} \tag{1.6}$$

with problem (1.5) in mind, we assume

$$f_0 \in W^{m,2} \tag{1.7}$$

As illustrated in the following example, in many studies involving wildlife the sampling scheme, as well as the nature of f_0, is such that the problem is multi-scaled. By multi-scaled problems we mean those regression problems for which a single smoothing parameter, λ, is not sufficient.

Example: The Wood Thrush (Hylocichla mustelina) is a Neotropical migrant that breeds in Delaware, USA. Growth data were collected incidentally from 1975 through 2002, inclusive, at the University of Delaware, USA Woods (UDW), an isolated forest fragment of approximately 16 ha located on the University Experimental farm in Newark, Delaware, USA. As part of a larger reproductive study (Brown and Roth [3, 4]), all known nestlings were uniquely color-banded and measured approximately seven to ten days after hatching. Season-long mist-netting efforts over the course of the study (Roth, Johnson, and Underwood [16]) provided repeat measurement data for many birds, including birds captured as fledglings and independent, hatching-year birds. At each capture, mass was measured to the nearest 0.5 g using a 50g or 100g Pesola spring balance. This resulted in 1,314 observations from 933 individuals (see Brown and Roth [4] for further details). In addition to these data, first reported in Brown and Roth [4], 330 additional observations are included here from nestlings aged 1 to 13 days after hatching. These data were collected from 1975 to 1978, inclusive. As most measurements after the initial banding of individuals as nestlings were collected incidentally over the course of the study, the data set is of the mixed (longitudinal and cross-sectional) type.

Figure 1 gives the distribution of ages (days after hatching), i.e. X_i, for all nestlings in this data set. Note they range from 1 to 92. Further, such a distribution of ages is somewhat typical of avian growth studies, with the bulk of the data being collected while individuals are accessible as nestlings and observations falling off dramatically after the birds are capable of flight.

Figure 2a gives the scatter plot of the data along with the cubic spline estimator, with λ selected by generalized cross validation (GCV) for the original 1314 observations. For clarity, Figure 2b gives the scatter plot of $(X_i, \bar{Y}_i)$, the spline estimator, and the corresponding individual 95% confidence intervals.

Remark. Throughout the paper we are using fiducial asymptotic confidence intervals. That is, for a given value of x the confidence interval is given by

$$f_\lambda(x) \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(f_\lambda(x))}, \tag{1.8}$$

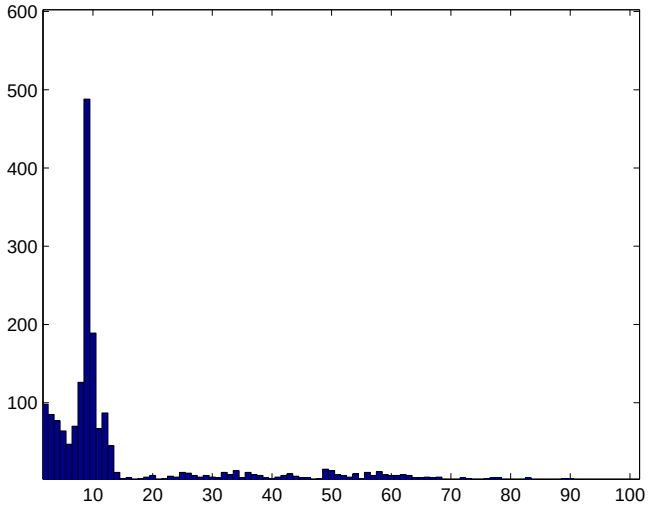

Figure 1: Histogram of the ages of the Wood Thrushes included in the study. The x-axis is age measured in days, and the y-axis is frequency.

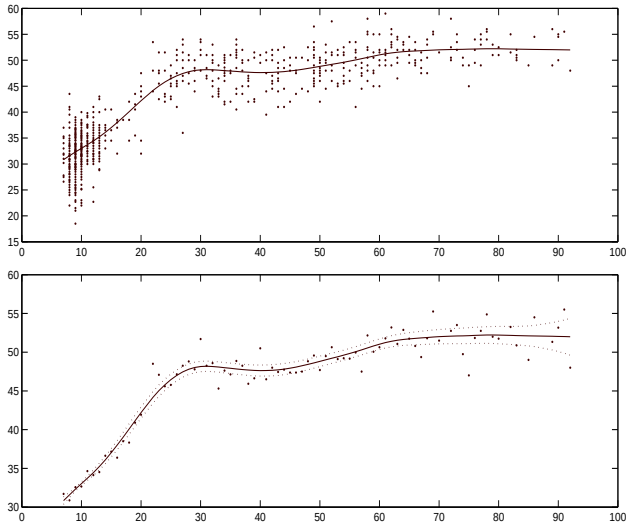

Figure 2: Part a is a scatter plot of the original Wood Thrush data, along with the spline estimator with λ selected by GCV. Part b is a scatter plot of $(X_i, \bar{Y}_i)$, along with the spline estimator and the individual 95% confidence intervals. The x-axis is age measured in days, and the y-axis is mass in grams.

where $\widehat{Var}(f_\lambda(x))$ is an estimate of $Var(f_\lambda(x))$ the asymptotic variance of $f_\lambda(x)$. Also we do not require deterministic design points, X_i. The proposed confidence intervals are conditional on the design. (For the specifics see Section 3.)

The smoothing spline estimator seemed to provide an excellent estimate of the mean model for the original Wood Thrush growth data. Two biologically important aspects of the estimator are the following. First, the predicted decrease in juvenile mass immediately after independence (i.e., 27-34 days after hatching). Secondly, the estimated value of adult mass, its stability over the adult ages, and the age at which adult mass was achieved are all as expected. See Brown, Eggermont, LaRiccia, and Roth [5] for more details.

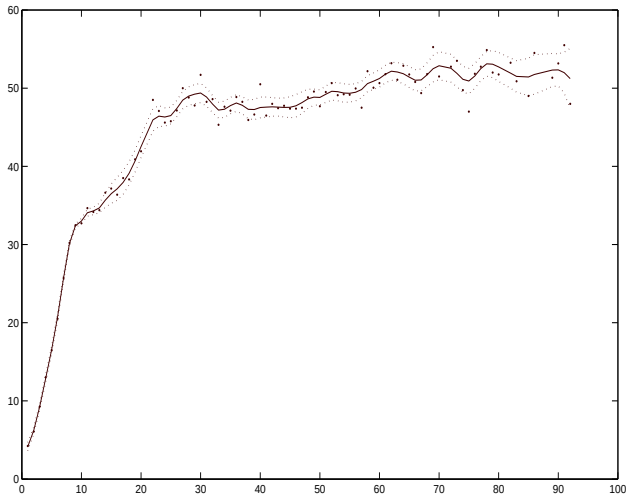

Figure 3: A scatter plot of $(X, \bar{Y}_i)$ for the complete data Wood Thrush data set, along with the spline estimator with λ selected by GCV and the associated 95% confidence intervals. The x-axis is age in days and the y-axis is mass in grams.

Figure 3 gives the cubic spline estimator, with λ again selected by GCV, for the complete data set. While the main shape of the estimator has been kept, clearly for $X > 20$ the spline estimator is now under smoothed, and is not realistic. Note that the cause of the problem is a classical example of the problem discussed in Nychka [13], is two-fold, and likely to occur in many wildlife studies. First because of the ease of collection there is an inordinate amount of data collected on young birds, where the growth rate is the largest. This dictates that the GCV selected value of λ for the entire data set be quite small. On the

other hand for mature birds, for which $f_0'(x) \approx 0$, there are only a relatively small number of observations. These last two facts imply that for this section of the graph a larger value of λ would be better. A simple-minded fix would be to just increase the value of λ until the curve looks somewhat smooth. From a practical point of view the main problem with this approach is that now the bias of the estimator is larger than its variance, and hence the asymptotic confidence intervals are no longer valid. (See the discussion of confidence intervals, especially in Eubank [8, page 268] and Eggermont and LaRiccia [7, Chapter 23]).

In the literature there are at least three ways in which multi-scaled problems have been handled. The first approach, discussed in Abramovich and Steinberg [1], is to keep a single smoothing parameter but modify the penalty function to take into account the distribution of the X_i's, call it $h(x)$, and f_0^m, the m^{th} derivative of the true mean function. Specifically, the original spline minimization problem, (1.2), is replaced by a problem of the form

$$\min n^{-1} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda^{2m} \int_A \phi(\hat{h}(x), \hat{f}_0^{\,m}(x)) \, |f^m(x)|^2 \, dx, \tag{1.9}$$

where ϕ is a non-increasing function, and $\hat{h}$ and $\hat{f}_0^{\,m}$ are estimators of h and f_0^m. A similar approach was proposed by Pintone, Speckman, and Holmes [14]. Our problem with this approach is that in preliminary studies the resulting estimator seemed to be much more sensitive to the selection of ϕ, than to the data or f_0. A second approach is to have an individual λ for each value of X, or at least each X_i, see Cummins, Filloon and Nychka [6]. However, for most growth model problems f_0 is quite smooth. Thus this approach seems to be overkill.

The third general approach, and the one taken here, is to first split the X domain into a collection of subsets and then to determine the estimator individually on each subset and finally to smoothly "blend" these individual estimators together. For a general discussion of partitioning the X space in nonparametric regression see Györfi, Kohler, Krzyzak, and Walk [10]. The main problem associated with this approach is in determining both the number and then the actual subsets required in a data driven fashion. Also, the question of how to "blend" the individual estimates together must be addressed. However as discussed in Section 2, for many wildlife studies the biology of the animal being studied dictates a relatively small number of possible partitions which should be considered. Further, biological theory also presents an argument for a simple method for selecting the number of partitions to be used. Finally, note that for the proposed estimator the subsets are not

disjoint. The overlap in the subsets is then used to "blend" the estimators together in a fashion which makes asymptotic confidence intervals of the mean function possible.

2 The proposed estimator. In this section we develop the proposed estimator for fitting growth data in cases where the sampling is similar to that observed in the above example. The basic form of the proposed estimator is quite simple. Let A denote the domain of X. For an integer k define the subsets (a_i, b_i) for $i = 1, \ldots, k$ such that

$$A = \bigcup_{i=1}^{k} (a_i, b_i).$$

Now for each i let $\hat{f}_{\lambda_i,i}$ denote the cubic spline estimator, with λ selected by either Generalized Cross Validation (GCV), Generalized Maximum Likelihood (GML), or Mallows Cp, based only on those data points for which $X_j \in (a_i, b_i)$, and define

$$\begin{aligned} \hat{f}_i(x) &= \hat{f}_{\lambda_i,i}(x) \text{ for } x \in (a_i, b_i) \\ &= 0 \text{ otherwise.} \end{aligned}$$

Finally for $i = 1, \ldots, k$ let $\delta_i(x)$ be a non-negative weight function with

$$\sum_{i=1}^{k} \delta_i(x) = 1$$

for all values of x. Then for given values of k, (a_i, b_i), and $\delta_i(x)$ the estimator is defined by

$$\hat{f}(x) = \sum_{i=1}^{k} \delta_i(x) \hat{f}_i(x). \tag{2.1}$$

Clearly the problem is to determine a data driven method for selecting the values of k, (a_i, b_i), and $\delta_i(x)$, for which $\hat{f}$ converges at the proper rate, and, as is typical of fiducial confidence intervals for splines, away from the boundaries, i.e.,

$$X \in \left(\lambda_1 \ln\left(\frac{1}{\lambda_1}\right), 1 - \lambda_k \ln\left(\frac{1}{\lambda_k}\right) \right)$$

the confidence intervals are asymptotically valid.

One of the main keys to the approach taken here is that the biology (science) of the animal (system) being studied typically provides one with a small number of natural overlapping partitions of A, which can be used to determine the sets (a_i, b_i). As an example

Table 1: The four developmental periods of juvenile Wood Thrushes, represented in days since hatching. The upper age range (92 days) of adult-sized, HY birds is arbitrary and indicates the oldest known nestling caught during the breeding season.

Developmental period	Beginning	End	Ending age range
Nestling	1	13	12–15
Fledgling	13	30	27–34
Independence	30	41	37–44
Adult-sized HY	41	92	—

for Wood Thrushes, the first year of their life can be broken up into the four periods; nestlings, fledglings, independent fledglings, and adult-sized birds (see Table 1). Note that the age range of each development period was a conservative estimate of the age (days after hatching) that the life stage begins and ends (Brown and Roth [4]). For example, under normal circumstances nestlings fledge at 12-15 days old, but commonly fledge at 13 days of age. Based on these four developmental periods, only eight cases with a varying number of partitions need to be considered for modeling growth (Table 2). These range from case 1, with the entire range of growth data considered as one partition, to case 8, with four temporally overlapping partitions (nestlings, fledglings, independent fledglings, and HY birds). The other cases contain all possible combinations of developmental periods that include the entire age range of the data (Table 2).

Given the appropriate boundary points of the partition the question of how to connect the different estimators needs to be addressed. Remember that for $i = 2, \ldots, k$

$$a_i < b_{i-1}.$$

Now for the non-overlapping parts of A, the estimator is defined in terms of a single $\hat{f}_i$. That is for $x \in (a_1, a_2)$ we take

$$\hat{f}(x) = \hat{f}_1(x), \tag{2.2a}$$

for $i = 2, \ldots, k-1$ and $x \in (b_{i-1}, a_{i+1})$

$$\hat{f}\ (x) = \hat{f}_i(x), \tag{2.2b}$$

Table 2: The eight possible cases to be considered in modeling partitioned growth data based on developmental periods of the Wood Thrush (see Table 1). The number of partitions in each cases is indicated by k, a_i represents the beginning of each partition, and b_i represents when each partition ends, in days since hatching.

Case	k	$(a_i\ , b_i\ )$			
1	1	(1,92)			
2	2	(1,15),	(12,92)		
3	2	(1,34),	(27,92)		
4	2	(1,44),	(37,92)		
5	3	(1,15),	(12,44),	(37,92)	
6	3	(1,15),	(12,34),	(27,92)	
7	3	(1,34),	(27,41),	(37,92)	
8	4	(1,15),	(12,34),	(27,44),	(37,92)

and for $x \in (b_{k-1}, b_k)$

$$\hat{f}(x) = \hat{f}_k(x). \tag{2.2c}$$

On the intervals (a_i, b_{i-1}) the estimator is taken to be nothing more than a weighted average of $\hat{f}_i$ and $\hat{f}_{i-1}$, with the weights being a function of how far one is away from the corresponding boundary. Specifically let $w_i(x)$ be a non-increasing function with $0 \leq w_i(x) \leq 1$, then for all $x \in (a_i, b_{i-1})$

$$\hat{f}(x) = w_i(x)\hat{f}_{i-1}(x) + (1 - w_i(x))\hat{f}_i(x). \tag{2.2d}$$

3 Asymptotic Distribution. In this section the asymptotic distribution of $\hat{f}(x)$, for a fixed partition is determined. Also without loss of generality, we shall assume $A = (0, 1)$. The estimator as defined by (2.2) is nothing more than a linear function of k separate spline estimators. Thus as long as the a_i, b_i, and w_i 's are selected so that the overall estimator, $\hat{f}$, does not depend on the individual $\hat{f}_i$ near their boundaries, asymptotic confidence intervals are easily determined. Further since for all n only a small number of possible subsets are being considered, the following simple restrictions on the a_i, b_i, and w_i are adequate for the

asymptotic confidence intervals to be valid.

The specific assumptions required are the following: For $i = 1, \ldots, k$ let n_i denote the number of observation for which $X_j \in [a_i, b_i]$. Then we assume that for all $i = 1, \ldots, k$

$$n_i \to \infty \text{ and } \frac{n_i}{\max(n_1, \cdots, n_k)} \to r_j > 0. \tag{3.1}$$

Next define $\delta_1 > 0$ and $\delta_2 > 0$. Now for all values of i let a_i, b_i, and w_i be selected so that

$$w_i(x) \text{ is continuous on } [a_i, b_i] \tag{3.2}$$

$$|a_i - b_{i-1}| > \delta_1, \tag{3.3}$$

and

$$\text{for all } x \in (a_i, a_i + \delta_2), \qquad w_i(x) = 0 \tag{3.4}$$

are satisfied.

Now with respect to the design we assume

$$X_1, X_2, \ldots \text{ are } \quad i.i.d., \tag{3.5}$$

having a probability density, w, with respect to Lebesgue measure on (0,1), and that, for positive constants w_1 and w_2

$$w_1 \le w(x) \le w_2 \text{ for all } x \in (0, 1). \tag{3.6}$$

Finally, with respect to the $\lambda_i \equiv \lambda_{n,i}$ selected for each partition we assume that there exists a deterministic sequence $\gamma_{n,i}$ for which

$$\frac{\lambda_{n,i}}{\gamma_{n,i}} - 1 = o_p((\log n)^{-1}) \quad \text{ and } \quad \gamma_{n,i} \approx n^{-1/(2m+1)}, \tag{3.7}$$

for $n = 1, \ldots, k$. Remember the goal is to develop confidence intervals conditioned on the design. Thus let $X_n = (X_1, \cdots, X_n)^T$, $\gamma_n = (\gamma_{n1}, \cdots, \gamma_{nk})^T$, and $\lambda_n = (\lambda_{n1}, \cdots, \lambda_{nk})^T$. Also define

$$\gamma_{\max} = \max\{\gamma_1, \cdots, \gamma_k\} \textit{ and } n_{\min} = \min\{n_1, \cdots, n_k\}.$$

Now for a fixed smoothing parameter,γ_n, denote the conditional variance of the estimator by

$$Var(\hat{f}(x)|X_n, \gamma_n),$$

and let

$$sd(\hat{f}(x)) = (\mathrm{Var}(\hat{f}(x)|X_n, \gamma_n)_{\gamma_n=\lambda_n})^{1/2}.$$

Finally for each x define the pivot variable

$$PV(x) = \frac{\hat{f}(x) - f_0(x)}{sd(\hat{f}(x))}, \tag{3.8}$$

upon which the confidence interval is based. We then have the following theorem.

Theorem 3.1. *Assume (1.1) through (1.4), and (3.1) through (3.7). Then for*

$$x \in \left(\gamma_{n1} \ln\left(\frac{1}{\gamma_{n1}}\right), 1 - \gamma_{nk} \ln\left(\frac{1}{\gamma_{nk}}\right)\right)$$

there exists a sequence of zero mean Gaussian processes, $Z_n(x)$, with $Var(Z_n(x)) = 1$ so that

$$PV(x) = Z_n(x) + n_{n,\lambda_n}(x),$$

with

$$||n_{n,\lambda_n}(x)|| = O_p((n_{\min}\gamma_{\min})^{-1/2}\gamma_{\max}^{-1/2} \log n_{\min} + n_{min}^{-1/4}\gamma_{\max}^{-1/2})$$

Proof. For any partition, say $[a_i, b_i]$. Let

$$D_i(x, \lambda_i) = \hat{f}_i(x) - E[\hat{f}_i(x)|X_n]$$

Now Theorem 23.2.3 of Eggermont and LaRiccia [7] gives that under the assumptions there exists a sequence of zero mean Gaussian processes, say G_{n,γ_i}, with the same covariance kernel as $\hat{f}_i(x)$ so that

$$D_i(x, \lambda_i) = G_{n,\gamma_i}(x) + n_{i,\lambda_i}(x),$$

with

$$||n_{n,\lambda_i}||_\infty = O_p((n_i\gamma_i^{3/2})^{-1} \log n_i + n_i^{-3/4}\gamma_i^{-1})$$

Now define

$$D(x, \lambda_n) = \hat{f}(x) - E[\hat{f}(x)|X_n].$$

Since for all x, D is just a linear combination of 2 of the D_i's, there exists another sequence of zero mean Gaussian process, say G_{γ_n}, so that

$$D(x, \lambda_n) = G_{\gamma_n}(x) + \xi_{\lambda_n}(x),$$

with

$$||\xi_{\lambda_n}||_\infty = O_p((n_{min}\gamma_{max}^{3/2})^{-1}\log n_{min} + n_{min}^{-3/4}\gamma_{max}^{-1}),$$

Thus

$$PV(x) = \frac{G_{\gamma_n}(x)}{sd(\hat{f}(x))} + \frac{\xi_{\lambda_n}(x)}{sd(\hat{f}(x))} + \frac{E[\hat{f}(x)|x_n] - f_0(x)}{sd(\hat{f(x)})}$$

Clearly

$$||\frac{\xi_{\lambda_n}(x)}{sd(\hat{f}(x))}||_\infty = O((n_{m,n}\gamma_{max})^{-1/2}\gamma_{max}^{-1/2}\log n_{min} + n_{min}^{-1/4}\gamma_{max}^{-1/2}),$$

and we are done if the bias to standard deviation ratio term can be ignored. Now it is well know that under the above conditions (see, for example. Corollary 22.8.14 of Eggermont and LaRiccia [7]) that for of the individual spline estimators there exists a k so that

$$\sup_{x\in B_i(k)} |E[\hat{f}_i(x)|X_n] - f_o(x)| = O(\lambda_i^{2m} + \lambda_i^{-1/2}(n_i\lambda_i)^{-1}\log(n_i))$$

and

$$Var(f_i|X_n,\gamma)_{|\gamma=\lambda_i} = (n_i\gamma_i)^{-1},$$

where

$$B_i(k) = \{x|x \in (a_i + k\lambda_i\log(\frac{1}{\lambda_i}), b_i - k\lambda_i\log(\frac{1}{\lambda_i})\}$$

Now since $\gamma_i = O_p(n_i^{-1/(2m+1)})$, for $x \in B_i(k)$

$$\frac{|E[\hat{f}_i(x)|X_n] - f_0|}{sd(\hat{f}_i(x))} \to 0.$$

Finally, (3.4) establishes that $\hat{f}(x)$ does not depend upon $f_i(x)$ for $x \notin B_i(k)$, and we are done.

The final ingredient required to be able to determine confidence intervals, is an estimate of $Var(\hat{f}(x)|X_n)$.

At the design points this is simple. For each segment let $H_i(\lambda_i)$ denote the hat matrix associated with the spline estimator on the interval $[a_i, b_i]$. Then letting

$$\hat{Y} = (\hat{f}(x_1), \cdots, \hat{f}(x_n))^T,$$

$$\hat{Y} = R_j Y,$$

where R_j is the $n \times n$ block diagonal matrix made up of weighted averages of the $H_i(\lambda_i)$'s. Thus $Var(\hat{Y}|X_n) = \sigma^2 R_j R_j^T$.

This conditional covariance matrix can then be estimated by

$$\hat{V}(\hat{Y}|X_n) = s^2(R_j R_j^T),$$

with s^2 any of the consistent nonparametric estimators of σ^2. Putting all the above together we get the asymptotically valid individual confidence interval at X_i,

$$\hat{Y_i} \pm z_{x/2} s \sqrt{(R_j R_j^T)_{ii}}$$

□

4 Selecting the case. The final step in defining the estimator is to decide between the different cases. We consider three possible selection criteria, each of which, in a given application, might be the preferred approach. To keep notation to a minimum let

$$f_i \text{ for } i = 1, \ldots, M$$

denote the estimators associated with the M different cases, and denote the true mean function by f_0. Before developing the procedures we first discuss why different procedures are considered. Typically in the regression setting one wants to have a procedure which minimizes some guess at

$$SE(j) = \frac{1}{n} \sum_{i=1}^{n} (f_j(X_i) - f_0(X_i))^2. \tag{4.1}$$

However, since $SE(j) \rightarrow_p \int_A w(x)(f_j(x) - f_0(x))^2 dx$, for experiments with a sampling scheme like the above example, such a procedure will lead to a solution which basically ignores what happens for $X > 10$. Since in this example one is equally interested in the adequacy of the estimator over the entire range of X, it is probably more useful to try and minimize a weighted squared error

$$WSE(j) = \frac{1}{n} \sum_{i=1}^{n} v_i (f_j(X_i) - f_0(X_i))^2, \tag{4.2}$$

where the weights, v_i, are selected so that

$$WSE(j) \rightarrow_p \int_A (f_j(x) - f(x))^2 dx.$$

One often cited complaint with selecting estimators to minimize guesses at (4.1) or (4.2) is that they under smooth. (see the example in Section 5.) Here remember that for each segment λ_i was selected by GCV, and hence to minimize an approximation to (4.1). Further, the biologists believe that while f_0 may not be monotone, it should be relatively smooth. Hence, it seems reasonable to select that case which minimizes

$$\text{Wiggle(j)} = \text{the number of sign changes in } f_j^1. \tag{4.3}$$

Now on to the guesses for (4.1) through (4.3). Since the estimator is a linear function of Y, with, more or less, the same structure as the usual spline estimator, one can easily develop a GCV (actually a zero trace, Li [11]) like procedure for either (4.1) or (4.2). To see what is involved let us develop the zero trace procedure for (4.2). To start this off we consider a Mallows type correction to the most naive approximation to $WSE(j)$. Namely $\|W(I-R_j)Y\|^2$, where $W = diag(\upsilon_1, \ldots, \upsilon_n)$, and

$$(f_j(X_1) \cdots f_j(X_n))^T = R_j Y.$$

For this note that

$$WSE(j) = \|W(Y_0 - R_j Y\|^2,$$

and hence

$$E[WSE(j)] = \|W(I-R_j)Y_0\|^2 + \sigma^2 trace(R_j^T W^2 R_j).$$

However

$$\begin{aligned} E[\|W(I-R_j)Y\|^2] &= \\ &\quad \|W(I-R_j)Y_0\|^2 + \sigma^2 trace((I-R_j^T)W^2(I-R_j)) \\ &= \|W(I-R_j)Y_0\|^2 + \sigma^2\{trace(R_j^T W^2 R_j) \\ &\quad + trace(W^2) - 2\ trace(R_j W^2). \end{aligned}$$

Thus if σ^2 were known the Mallows type procedure would minimize

$$\|W(I-R_j)Y\|^2 + 2\sigma^2 trace(R_j W^2) - \sigma^2\ trace(W^2). \tag{4.4}$$

Note that since $trace(W^2)$ does not depend upon j it may be dropped. Next to get the zero trace type approximation to $WSE(j)$ we replace R_j in (4.4) by $\alpha I + (1-\alpha)R_j$ for arbitrary $\alpha \in [0,1]$. This gives the Mallows(α) criteria

$$(1-\alpha)^2 \|W(I-R_j)Y\|^2 + 2\sigma^2(trace\left\{(I-(1-\alpha)(I-R_j))W^2\right\} - \sigma^2\ trace(W^2)$$

which may be re-written as

$$(1-\alpha)^2 \left\|W(I-R_j)Y\right\|^2 + 2\sigma^2 \left\{trace(W^2) - (1-\alpha)trace[(I-R_j)W^2]\right\},$$

where the term $\sigma^2\ trace(W^2)$ has been dropped. For the zero trace procedure the goal is to find that value of α for which

$$trace(W^2) - (1-\alpha)trace[(I-R_j)W^2] = 0.$$

Clearly this is given by

$$(1-\alpha) = \frac{trace(W^2)}{trace[(I-R_j)W^2]}.$$

Using this value of α in Mallows(α) then gives the zero trace type procedure:

Select that value of j which solves the problem

$$\min \frac{[trace(W^2)]^2 \left\|W(I-R_j)Y\right\|^2}{(trace[(I-R_j)W^2])^2}. \tag{4.5}$$

We shall denote the solution by $j(WGCV)$, since problem (4.5) is reminiscent of the GCV problem.

With regard to approximating $WSE(j)$ the same sort of argument leads to the following procedure. Select that value of j which solves

$$\min \frac{\left\|(I-R_j)Y\right\|^2}{(trace[(I-R_j)])^2}. \tag{4.6}$$

For this case the similarity to the GCV problem is obvious, and hence we shall denote the solution by $j(GCV)$. Finally as the proxy for Wiggle(j) we take

$$Wig(j) = \sum_{i=2}^{n-1} \left\{1 - I[f_j(X_i) = median(f_j(X_{i-1}), f_j(X_i), f_j(X_{i+1}))]\right\}, \tag{4.7}$$

where $I\{a=b\} = 1$, if $a = b$, and 0 otherwise.

5 The wood thrush data re-visited. In this section the estimation procedure of Section 2 through 4 is applied to the Wood Thrush data. The first step is to make a concrete selection for w_i. Here we took a simple linear function defined on a restricted subset of (a_i, b_{i-1}). That is let $L_i = a_i + \delta$, and $U_i = b_{i-1} - \delta$, and for $x \in (L_i, U_i)$ set

$$w_i(x) = 2 * \frac{(x-c_i)}{(U_i - L_i)},$$

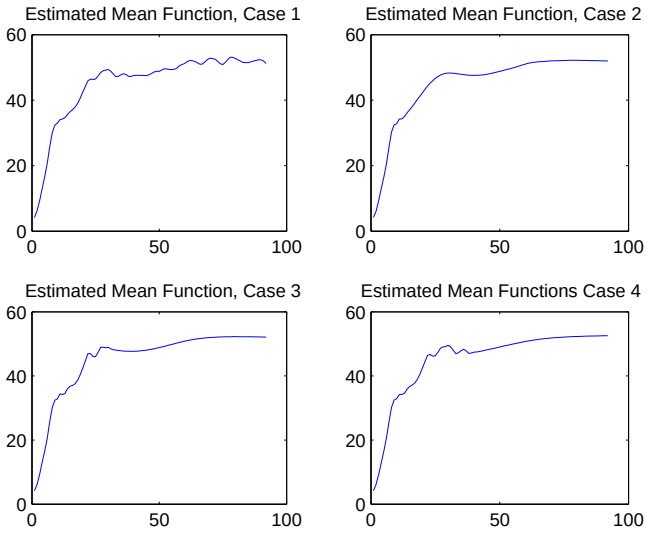

Figure 4: The blended spline estimator for cases 1 through 4. The x-axis is age in days, and the y-axis is mass in grams.

and equal to zero otherwise. In the above $c_i = 0.5 * (U_i + L_i)$.

Remark. Somewhat surprisingly we found the estimator to be quite insensitive to the specific selection of w_i or δ. However if this worries you, just add other possible choices of δ to the list of cases to be considered.

Now to the results. Figures 4 and 5 give the graphs of the estimators for the different cases considered, while Table 4 gives the corresponding values of the different measures.

Finally Figure 6 gives the graph of the estimator for Case 2, which minimizes $WIG(j)$. Note that the estimate has basically the same shape as the spline estimator based only on the original data, with an added crink in the curve at X near 13. Interestingly this is right at the stage when a bird would be getting ready to fledge, and there is some biology to suggest such a small drop in the growth curve might happen here. However, this hypothesis needs further investigation and the crink might be nothing more than a statistical glitch.

To further illustrate the procedure Table 5.2 gives the values of the three selection criteria for the Wood Thrush data when only observations with $X > 6$ are considered. (This is the case where a single λ provided an excellent fit.)

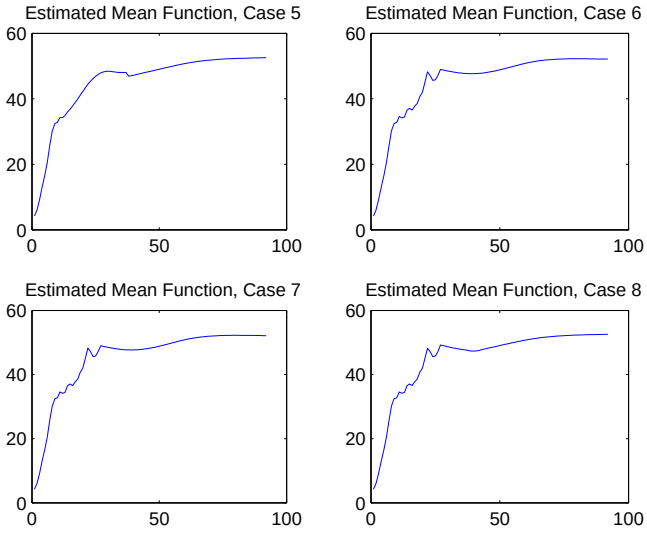

Figure 5: The blended spline estimator for cases 5 through 8. The x-axis is age in days, and the y-axis is mass in grams.

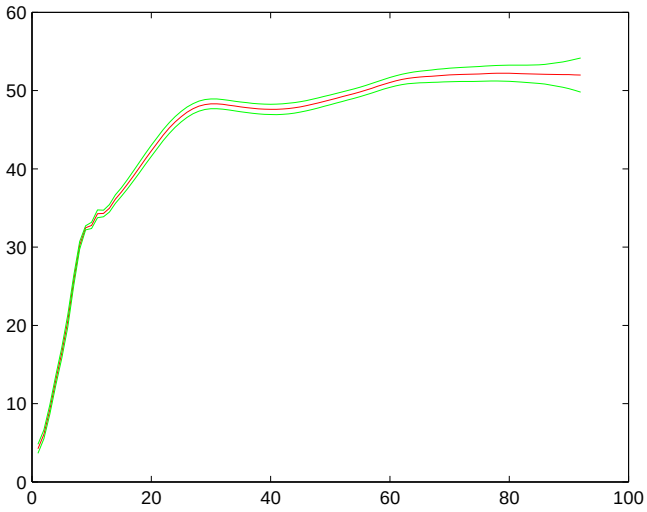

Figure 6: The selected blended spline estimator along with the associated 95% confidence intervals. The x-axis is age in days, and the y-axis is mass in grams.

Table 3:

Case	$WIG(j)$	WSE	$SE(j)$
1	19	7.4	3.8
2	3	6.9	3.4
3	9	5.8	3.1
4	6	6.1	3.0
5	4	6.8	3.2
6	9	6.5	3.5
7	10	6.0	3.0
8	8	6.7	3.5

Table 4:

Case	$WSE(j)$	$WIG(j)$	$SE(j)$
1	2.663	4	2.270
2	2.665	4	2.103
3	2.538	7	3.803
4	2.542	2	2.273
5	2.514	6	1.814
6	2.774	6	1.800
7	2.539	7	3.922
8	2.768	6	1.839

Figure 7 gives the estimator for Case 4, the minimizer of $WIG(j)$. Note that this estimator is more or less equivalent to the estimator for case 1, given in Figure 2.

Remark. We have considered other data sets also. In all cases minimizing either $WSE(j)$ or $SE(j)$ resulted in the selection of one of the less smooth estimators. Taking into account that the estimator based on minimizing any of the three measures, over a finite number of sets, will converge at the optimal rate, the Wood Thrush researchers have found that minimizing $WIG(j)$ provides the most satisfactory solution.

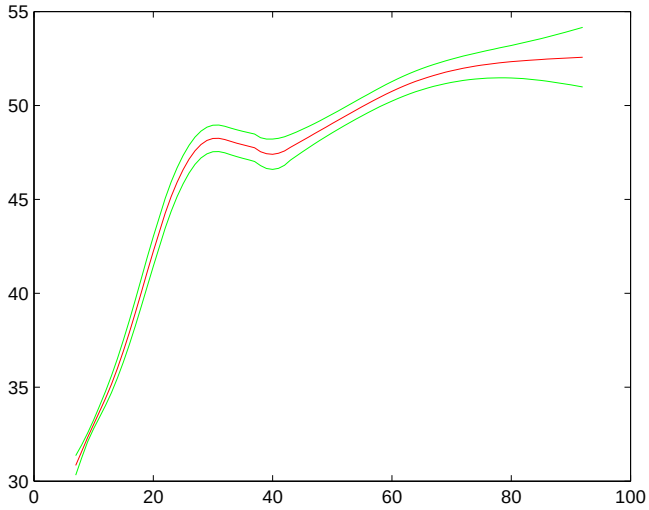

Figure 7: The Wig(j) selected blended spline estimator along with the 95% confidence interval for the original Wood Thrush data set. The x-axis is age in days, and the y-axis is mass in grams.

Acknowledgement. The authors would like to thank the referee who caught some easily fixed but important errors in the original draft.

REFERENCES

[1] Abramovich, F. and Steinberg, D. M. (1996). Improved inference in nonparametric regression using L–smoothing splines. *Journal of Statistical Planning and Inference.* **49** 327–341.

[2] Aggrey, S. E. (2002). Comparison of three nonlinear and spline regression models for describing chicken growth curves. *Poultry Science* **81** 1782–1788.

[3] Brown, W. P. and Roth, R. R. (2002). Temporal patterns of fitness and survival in the Wood Thrush. *Ecology* **83** 221–232.

[4] Brown, W. P. and Roth, R. R. (2004). Juvenile survival and r ecruitment of Wood Thrushes *Hylocichla mustelina* in a forest fragment. *Journal of Avian Biology.* **35** 316–326.

[5] Brown, W. P.; Eggermont, P.; LaRiccia, V. and Roth, R. R. (2007). Are parametric models suitable for estimating avian growth rates? *Journal of Avian Biology* **38** 495–506.

[6] Cummins, D. J.; Filloon, T. M. and Nychka, D. (2001). Confidence intervals for nonparametric curve estimates: Toward more uniform pointwise convergence. *Journal of American Statistical Association.* **96** 233–246.

[7] Eggermont, P. P. B, and LaRiccia, V. N. (2007). *Maximum Penalized Likelihood Estimation. Volume II: Regression.* Spring-Verlay, New York (in preparation).

[8] Eubank, R. (1999). *Non-parametric Regression and Spline Smoothing. 2nd Edition.* Marcel Dekker, New York.

[9] Gasser, T.; Müller, H-G., Köhler, W.; Molinari, L. and Prader, A. (1984). Nonparametric analysis of growth curves. *Ann. Statist.* **12** 210–294.

[10] Györfi, L.; Kohler, M.; Krzyzak, A. and Walk, H. (2002). *A distribution Free Theory of Nonparametric Regression.* Springer-Verlag, New York.

[11] Li, K. C. (1986). Asymptotic optimality of C and generalized cross validation in ridge regression with applications to spline smoothing. *Ann. Statist. 14* 1101–1102

[12] Müller, H. G. (1988). Nonparametric Regression Analysis of Longitudinal Data. Springer- Verlag, New York.

[13] Nychka, D. (1995). Splines as local smoothers. *Ann. Statist.* **23** 1175–119.

[14] Pintone, A.; Speckman, P. and Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika* **93** 113–125.

[15] Przybylski, M. and Garcia-Berthou, E. (2004). Age and growth of European bitterling (*Rhodeus sericeus*) in the Wiepz-Krzna Canal, Poland. *Ecohydrology and Hydrobiology.* **4** 207–213.

[16] Roth, R. R.; Johnson, M. K. and Underwood, T. J. (1996). Wood Thrush (*Hylocichla mustelina*). In "The Birds of North America, No. 246" (A. Poole and F. Gill, Eds.). *Academy of Natural Sciences*, Philadelphia, and American Ornithologists' Union, Washington, D.C.

[17] Schillaci, M. A. and Stallmann, R. R. (2005). Ontogeny and sexual dimorphism in booted macaques (*Macaca ochreata*). *Journal of the Zoological Society of London.* **267** 19–29.