Silvia Bianconcini

A Reproducing Kernel Perspective of Smoothing Spline Estimators

Quaderni di Dipartimento

Serie Ricerche 2008, n. 3

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Scienze Statistiche “Paolo Fortunati”

A Reproducing Kernel Perspective of Smoothing Spline Estimators

Silvia Bianconcini
Department of Statistics, University of Bologna
Via Belle Arti, 41 - 40126 Bologna, Italy
e-mail: silvia.bianconcini@unibo.it

Abstract: Spline functions have a long history as smoothers of noisy time series data, and several equivalent kernel representations have been proposed in terms of the Green's function solving the related boundary value problem. In this study we make use of the reproducing kernel property of the Green's function to obtain a hierarchy of time-invariant spline kernels of different order. The reproducing kernels give a good representation of smoothing splines for medium and long length filters, with a better performance of the asymmetric weights in terms of signal passing, noise suppression and revisions. Empirical comparisons of time-invariant filters are made with the classical non linear ones. The former are shown to loose part of their optimal properties when we fixed the length of the filter according to the noise to signal ratio as done in nonparametric seasonal adjustment procedures.

Keywords: equivalent kernels, nonparametric regression, Hilbert spaces, time series filtering, spectral properties.

1. Introduction

The origin of smoothing splines appears to lie in the work on graduating time series data by [Whittaker, 1923], but spline smoothing techniques were generally regarded as numerical analysis methods, mainly used in engineering, until extensive research by Grace Wahba demonstrated their utility for solving a host of statistical estimation problems. It has now become clear that smoothing splines, and their variants, provide extremely flexible data analysis tools. As a result, they have become quite popular and have found applications in such diverse areas as the analysis of growth data, medicine, remote sensing experiments and economics. [Schoenberg, 1946] was the first who introduces the word *spline* in connection with smooth, piecewise polynomial approximation. However, the ideas were already used in the aircraft, ship-building, and automobile industries. In the latter, the use of splines seems to have several independent beginnings. Credit is claimed on behalf of de Casteljau at Citroën, Pierre Bézier at Renault, and Birkhoff, Garabedian, and de Boor at General Motors (GM), all for work occurring in the very early 1960s or late 1950s. These numerical analysts found wonderful things to do with spline functions, because of their ease of handling in the computer coupled with their good approximation theoretic properties. Important references on splines from this view point are [Golomb and Weinberger, 1959], [De Boor and Lynch, 1966], [De Boor, 1978], [Schumaker, 1981], [Prenter, 1975], and the conference proceedings edited by [Greville, 1968] and [Schoenberg, 1964a].

Generalizations of the problem proposed by [Schoenberg, 1964a,b] were derived in [Kimeldorf and Wahba, 1971]. Historically that work is very close to the one of [Golomb and Weinberger, 1959] and [De Boor and Lynch, 1966], and later work on characterizing solutions to variational problems arising in smoothing has been made easier by the lemmas given there. In that paper, the authors demonstrated the connection between these variational problems and Bayes estimates, a problem that has its historical roots in the work of [Parzen, 1962, 1970]. The formulas in [Kimeldorf and Wahba, 1971] were not very well suited to the computing capabilities of the day and the work did not attract much attention from statisticians, being rejected by mainstream statistics journals as considered too "far out". The first spline paper in an important statistical review is that on histosplines by [Boneva and Stefanov, 1971], which lacks a certain rigor but certainly is of historical importance. In the later 1970s a number of things happened to propel splines to a popular niche in the statistics literature: computing power became available, which made the computation of

splines with large data sets feasible, and, later, inexpensive; a good data-based method for choosing became available, and most importantly, splines engaged the interest of a number of creative researchers. Simultaneously the work of [Duchon, 1977], [Meinguet, 1979], [Utreras, 1979], [Wahba and Wendelberger, 1980], and others on multivariate thin-plate splines led to the development of a practical multivariate smoothing method, which had few real competitors in the so-called "nonparametric curve smoothing" literature. There rapidly followed splines and vector splines on the sphere, partial splines and interaction splines, variational problems where the data are non-Gaussian and where the observation functionals are nonlinear, and where linear inequality constraints are known to hold. Along with these generalizations came improved numerical methods, publicly available efficient software, numerous results in good and optimal theoretical properties, confidence statements and diagnostics, and many interesting and important applications. The body of spline methods available and under development provided a rich family of estimation and model building techniques that have found use in many scientific disciplines. Recent key references on penalized and smoothing splines are [Eubank, 1988], [Wahba, 1990], [Green and Silverman, 1994], [Eilers and Marx, 1996], [Hastie, 1996], [Hastie et al., 2001], and [Ruppert et al., 2002]. In the mid-1990s, the attention was concentrated to analyze the connection between splines and another vibrant area of data analytic research, known as reproducing kernel methods. Even if [De Boor and Lynch, 1966] have already introduced the reproducing kernel methodology to solve smoothing spline estimation problems, the emergence of support vector machines, starting with [Boser et al., 1992], have elucidated the connection between these two sets of literature.

Reproducing kernel methods are performed within the functional analytic structure known as a Reproducing Kernel Hilbert Space (RKHS). An early RKHS reference is [Aronszajn, 1950] and contemporary summaries include [Wahba, 1990, 1999], [Evgeniou et al., 2000], and [Pearce and Wand, 2006]. The latter show how penalized splines are embedded in the class of reproducing kernel methods and help to connect these two bodies of research, envisaging that support vector machines and other kernel methods have the most to gain from this connection, particularly for the solution of classification and prediction problems.

In this study, we derive a reproducing kernel representation of smoothing splines. Under the assumption of equally spaced observations, we show how transform a smoothing spline into a kernel estimator with invariant local fitting and smoothing properties. This enables us to build easily an hierarchy of kernel

splines of different orders.

The paper is structured as follows. In Section 2, we provide a basis function representation of smoothing splines with particular emphasis on the cubic ones. Section 3 describes the equivalent kernel representation based on the Green's function, whose properties are here studied. A reproducing kernel perspective is given in Section 4, where RKHS are introduced and several Sobolev spaces illustrated. Section 5 deals with the problem of time series filtering. The theoretical properties of time-invariant smoothing splines are analyzed by means of spectral techniques, and we compared their performances using real life series. Finally, Section 6 gives the conclusions.

2. Smoothing Spline of order m

Let us suppose that observations are taken on a continuous random variable Y at n predetermined values of a continuous independent variable t. Let $\{(t_i, y_i), i = 1, 2, ..., n\}$ be the observed values of t and Y, assumed to be related by the regression model

$$y_i = \mu(t_i) + \epsilon_i \qquad i = 1, 2, ..., n \tag{1}$$

where the ϵ_i are zero mean, uncorrelated random variables with a common variance σ_ϵ^2, and $\mu(t_i)$ are values of some unknown function at the design points $t_1, t_2, ..., t_n$. We will assume that $0 \leq t_1 \leq ... \leq t_n \leq 1$. There is no loss of generality in making this assumption (see [Eubank, 1988])[1].

The determination of a suitable inferential methodology for the model (1) will hinge on the assumptions it is possible to make about μ. There are two different approaches of the regression analysis problem: parametric and nonparametric.

Parametric methods require very specific, quantitative information from the experimenter about the form of μ that places restrictions on what the data can tell us about the regression function. Such techniques are the most appropriate

[1]This is equivalent to assume that the t_i's are generated by a continuous, positive design density f_0 on $[0, 1]$ through a relationship such as

$$\int_0^1 f_0(t)dt = \frac{2i-1}{2n}, \quad i = 1, 2, ..., n$$

In words, this means that t_i is the $100\frac{(2i-1)}{2n}$-th percentile of the density f_0 . The canonical case of a uniform design corresponds to $f_0(t) = \mathbf{1}_{[0,1]}(t)$.

when the theory, past experiments and/or other sources are available, providing detailed knowledge about the process under study. In contrast, nonparametric regression techniques relay on the experimenter to supply only qualitative information about μ and let the data speak for themselves concerning the actual form of the regression curve. These methods are best suited for inference in situations where there is little or no prior information available about the regression curve. In this latter class lie smoothing spline techniques, which assume that μ belongs to the m-th order Sobolev space

$$W_2^m[0,1] = \left\{ \begin{array}{ll} \mu: & \mu^{(j)} \text{is absolutely continuous}, j = 1, 2, ..., m-1 \\ & \mu^{(m)} \in L^2[0,1] \end{array} \right\}$$

The space $W_2^m[0,1] \subset L^2[0,1]$, hence the properties of $L^2[0,1]$ functions will be applicable to the elements of $W_2^m[0,1]$.

Suppose that μ belongs to the space $W_2^m[0,1]$, a nonparametric regression estimator of μ can be approximated by a polynomial function of order $m-1$ as stated by the Taylor's theorem.

Theorem 1 (Taylor's theorem) *If $\mu \in W_2^m[0,1]$, then there exist coefficients $\theta_0, \theta_1, ..., \theta_{m-1}$ such that*

$$\mu(t) = \sum_{j=0}^{m-1} \theta_j t^j + \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} \mu^{(m)}(u) du$$

where

$$(t-u)_+^{m-1} = \left\{ \begin{array}{ll} (t-u)^{m-1}, & t \geq u; \\ 0, & t < u. \end{array} \right.$$

The Taylor's theorem suggests that if, for some positive integer m, the remainder term

$$Rem_m(t) = \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} \mu^{(m)}(u) du \tag{2}$$

is uniformly small, then we could write

$$y_i \cong \sum_{j=0}^{m-1} \theta_j t_i^j + \epsilon_i, \qquad i = 1, 2, ..., n \tag{3}$$

In other words, the data would follow an approximate polynomial regression

model. We could then estimate the polynomial coefficients by least squares or some other methods (see *e.g.* [Kendall et al., 1983]).

The Taylor's theorem arguments for the use of polynomial regression are tantamount to lump the remainder terms $Rem_m(t_1), ..., Rem_m(t_n)$ into the random error component of the model (1). If the remainders (2) at the t_i's are small relative to the random errors, polynomial regression (3) may work fairly well. If not, problems can arise. Since both the remainder and random errors are unknown there is no way to know whether or not the Taylor's theorem arguments are applicable to a specific choice of m that is made with any given data set. In view of the uncertainty about the magnitude of the remainder from a polynomial approximation of μ and of the random errors, it would seem natural to try to modify the polynomial regression estimator, attempting to compensate the possibility of large remainder terms. This line of reasoning leads to smoothing (and least squares) spline estimators for μ. Smoothing polynomial splines provide an alternative way of overcoming the limitations of a global polynomial model by adding polynomial pieces at given points, called *knots*, so that the polynomial sections are joined together ensuring that certain continuity properties are fulfilled.

There are several ways of representing a spline function, some of which are more amenable from the computational standpoint. The following one, known as truncated power representation, has the advantage of representing the spline as a multivariate regression model.

Definition 2 (Spline of order m) *A spline of order m with k knots at $\kappa_1, \kappa_2, ..., \kappa_k$ is any function of the form*

$$\mu(t) = \sum_{j=0}^{m-1} \theta_j t^j + \sum_{i=1}^{k} \eta_i (t - \kappa_i)_+^{m-1}, \qquad \forall t \in [0,1] \tag{4}$$

for some set of coefficients $\theta_0, \theta_1, ..., \theta_{m-1}, \eta_1, ..., \eta_k$.

This definition is equivalent to say that

(a) μ is a piecewise polynomial of order $m-1$ in each of the $(k-1)$ subinterval $[\kappa_i, \kappa_{i+1})$,

(b) μ has $m-2$ continuous derivatives, and

(c) μ has a discontinuous $(m-1)$th derivative with jumps at $\kappa_1, ..., \kappa_k$.

The set of functions $(t-\kappa_i)_+^{m-1}, i = 1, 2, ..., k$, defines what is usually called the truncated power basis of degree $(m-1)$. According to eq. (4), the spline is a linear combination of polynomial pieces; at each knot a new polynomial piece, starting off at zero, is added so that the derivatives at that point are continuous up to order $m-2$.

Let $S^m(\kappa_1, ..., \kappa_k)$ denote the space of all functions of the form (4). $S^m(\kappa_1, ..., \kappa_k)$ is a vector space in the sense that is closed under finite vector addition and scalar multiplication. Since the function $1, t, ..., t^{m-1}$, $(t-\kappa_1)_+^{m-1}, ..., (t-\kappa_k)_+^{m-1}$ are linearly independent, it follows that $S^m(\kappa_1, ..., \kappa_k)$ has dimension $m+k$.
In the sequel we shall assume that:

1. the observations are available at discrete points, $y_i, i = 1, 2, ..., n$, and
2. the knots are placed at the design points at which observations are made $(\kappa_i = t_i, i = 1, 2, ..., n)$.

We know that the behavior of polynomials fit to data tends to be erratic near the boundaries, and extrapolations can be dangerous. These problems are exacerbates with splines. The polynomial fit beyond the boundary knots behave even more wildly than the corresponding global polynomials in that region. This leads to *natural smoothing splines* which add additional constraints, ensuring that the function is of degree $(\frac{m}{2}-1)$ beyond the boundary knots.

Definition 3 (Natural spline of order m**)** *A spline function is a natural spline of order* m *with knots at* $t_1, t_2, ..., t_n$ *if in addition to properties (a), (b), and (c), it satisfies*

(d) μ *is a polynomial of order* $(\frac{m}{2}-1)$ *outside of* $[t_1, t_n]$.

The name natural spline stems from the fact that, as a result of (d), μ satisfies the natural boundary conditions

$$\mu^{(\frac{m}{2}+j)}(0) = \mu^{(\frac{m}{2}+j)}(1) = 0, \qquad j = 0, ..., \frac{m}{2}-1. \tag{5}$$

Let $NS^m(t_1, ..., t_n)$ denote the collection of all natural splines of order m with knots $t_1, t_2, ..., t_n$. Then $NS^m(t_1, ..., t_n)$ is a subspace of $S^m(t_1, ..., t_n)$ obtained by placing m (linear) restrictions arising from property (d) on the coefficients in eq. (4). In particular, to ensure that is a natural spline of order m

we must have

$$\theta_{\frac{m}{2}} = ... = \theta_{m-1} = 0 \tag{6}$$

in eq. (4) since it must be a polynomial of order $(m/2)$ for $t < t_1$. One may verify that $NS^m(t_1, ..., t_n)$ has dimension n.

Example 1. (Cubic smoothing splines) Consider the cubic spline model, which arises from setting $m = 4$ in eq. (4):

$$\mu(t) = \sum_{j=0}^{3} \theta_j t^j + \sum_{i=1}^{n} \eta_i (t - t_i)_+^3, \qquad \forall t \in [0, 1] \tag{7}$$

The original cubic spline model (7) has $4+n$ parameters. The natural boundary conditions (5) require that the second and the third derivatives are zero for $t \leq t_1$ and $t \geq t_n$. This implies to impose 4 restrictions (2 zeros and 2 linear) on the parameters of the cubic spline. In fact, the second and third derivatives are respectively

$$\mu''(t) = 2\theta_2 + 6\theta_3 t + 6 \sum_{i=1}^{n} \eta_i (t - t_i)_+,$$

$$\mu'''(t) = 6\theta_3 + 6 \sum_{i=1}^{n} \eta_i (t - t_i)_+^0.$$

For $\mu''(t)$ to be zero for $t \leq t_1$ outside it is required that $\theta_2 = \theta_3 = 0$, whereas for $t \geq t_n$ we also need $\sum \eta_i = 0$ and $\sum i\eta_i = 0$. On the other hand, $\mu'''(t) = 0$ for $t \leq t_1$ and $t \geq t_n$ if and only if $\theta_3 = 0$ and $\sum \eta_i = 0$.

[Lancaster and Salkauskas, 1986] showed that the following relations hold for natural cubic smoothing splines:

$$\boldsymbol{B}\gamma = \boldsymbol{D}\mathbf{y} \tag{8}$$

where $\boldsymbol{\gamma}$ and $\mathbf{y}$ are vectors defined as follows,

$$\mathbf{y}^T = [y_1 \quad y_2 \quad ... \quad y_n],$$

$$\boldsymbol{\gamma}^T = (\gamma_2, \gamma_3, ..., \gamma_{n-1}), \quad \gamma_i = \mu''(t_i)$$

and $\boldsymbol{B}$ and $\boldsymbol{D}$ are matrices of dimension $(n-2) \times (n-2)$, and $(n-2) \times n$

respectively, given by

$$\begin{bmatrix} \frac{1}{3}(t_3-t_1) & \frac{1}{6}(t_3-t_2) & 0 & \cdots & 0 \\ \frac{1}{6}(t_3-t_2) & \frac{1}{3}(t_4-t_2) & \frac{1}{6}(t_4-t_3) & \cdots & 0 \\ 0 & \cdots & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \frac{1}{6}(t_n-t_{n-1}) \\ 0 & \cdots & 0 & \frac{1}{6}(t_n-t_{n-1}) & \frac{1}{3}(t_{n+1}-t_{n-1}) \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{(t_2-t_1)} & -\frac{(t_3-t_1)}{(t_3-t_2)(t_2-t_1)} & \frac{1}{(t_3-t_2)} & \cdots & 0 \\ 0 & \frac{1}{(t_3-t_2)} & -\frac{(t_4-t_2)}{(t_4-t_3)(t_3-t_2)} & \frac{1}{(t_4-t_3)} & \cdots \\ 0 & \cdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \frac{1}{(t_n-t_{n-1})} & 0 \\ 0 & \cdots & \frac{1}{(t_n-t_{n-1})} & -\frac{(t_{n+1}-t_{n-1})}{(t_n-t_{n-1})(t_{n+1}-t_n)} & \frac{1}{(t_n-t_{n-1})} \end{bmatrix}$$

If the unknown smooth function μ has to be estimated on the basis of the n observations $y_i, i = 1, 2, ..., n$, the estimator $\hat{\mu}$ is given by the solution of the following optimization problem [Schoenberg, 1964b]

$$\min_{\mu \in W_2^m[0,1]} \left[\sum_{i=1}^{n} (y_i - \mu(t_i))^2 + \lambda \int_0^1 (\mu''(t))^2 dt \right] \tag{9}$$

The first term measures the closeness to the data, while the second one penalizes curvature in the function, and $\lambda > 0$ established a trade-off between the two. When $\lambda = 0$, $\hat{\mu}$ can be any function that interpolates the data, whereas if $\lambda = \infty$ the simple line fit, since no second derivative can be tolerated. These cases vary from very rough to very smooth, and the hope is that $\lambda \in (0, \infty)$ indexes an interesting class of functions in between.

The (unique) optimal solution $\hat{\mu}$ of problem (9) is a natural cubic spline with knots at points t_i. The estimated values are related to the observations y_i as follows [Wahba, 1990]

$$\hat{\boldsymbol{\mu}} = \mathbf{A}(\lambda)\mathbf{y} \tag{10}$$

where $\hat{\boldsymbol{\mu}}^T = (\hat{\mu}(t_1), \hat{\mu}(t_2), ..., \hat{\mu}(t_n))$ and $\mathbf{A}(\lambda)$ is the so called *influential matrix*. Therefore, each $\hat{\mu}(t_i)$ is a weighted linear combination of all the observed values, with weights given by the elements of the i-th row of $\mathbf{A}(\lambda)$. Clearly, these weights depend on the value of λ. λ is a parameter to be estimated, and the estimation is usually done with the Generalized Cross Validation (GCV) procedure which minimizes the mean square prediction error. On the other hand, λ can be assumed as given and each cubic spline predictor can be approximated with time invariant linear filters, as shown in [Dagum and Capitanio, 1999] which provided this explicit form of matrix $\mathbf{A}(\lambda)$:

$$\mathbf{A}(\lambda) = \left(I - \mathbf{D}^T \left(\frac{1}{\lambda}\mathbf{B} + \mathbf{D}\mathbf{D}^T \right)^{-1} \mathbf{D} \right). \tag{11}$$

3. The Equivalent Kernel Representation

Eq. (9) can be generalized to a natural smoothing spline estimator of order m, defined as the solution of the problem

$$\min_{\mu \in W_2^m[0,1]} \left[\sum_{i=1}^{n} f_0(t_i)(y_i - \mu(t_i))^2 + \lambda \|\mu^{(m)}\|^2 \right] \tag{12}$$

where $\|\cdot\|$ denotes the $L^2(0,1)$-norm, $f_0(t), t \in [0,1]$ is a probability density function, and λ is a positive smoothing parameter.
Different choices of $f_0(t), t \in [0,1]$, generally lead to different finite sample and asymptotic properties for the estimate $\hat{\mu}(t)$. Ideally, the selection of $f_0(t)$ may depend on the correlation structure of the data, but because it is usually unknown and may be difficult to estimate we do not have a general optimal $f_0(t)$. Intuitively, we can provide equal weight to each single observation that is shown to give satisfactory estimators (see, *e.g.* [Lin and Carroll, 2000]).

The criterion (12) is defined on an infinite-dimensional function space, that is the Sobolev space of functions for which the second term is defined.
Remarkably, it can be shown that eq. (12) has an explicit finite-dimensional, unique minimizer which is the natural smoothing spline of order m, $\hat{\mu}(t)$, with knots at the unique values $t_i, i = 1, 2, ..., n$. The result in eq. (10) can be extended to any positive integer m, introducing a symmetric function $S_\lambda(t,s)$ which belongs to $W_2^m[0,1]$ when either t or s is fixed, so that $\hat{\mu}(t)$ is given by

$$\hat{\mu}(t) = \sum_{i=1}^{n} S_\lambda(t, t_i) y_i f_0(t_i) \tag{13}$$

The explicit expression of $S_\lambda(t,s)$ is unknown. For the theoretical development of $\hat{\mu}(t)$, we will approximate $S_\lambda(t,s)$ by an equivalent kernel function whose explicit expression is available.
To do so, defined the $W_2^m[0,1]$-inner product as

$$\begin{aligned} < f, g >_{W_2^m[0,1]} &= < f, g >_{L^2(f_0)} + \lambda < f^{(m)}, g^{(m)} >_{L^2(0,1)} = \\ &= \int_0^1 f(t) g(t) f_0(t) dt + \lambda \int_0^1 f^{(m)}(t) g^{(m)}(t) dt \end{aligned} \tag{14}$$

we rewrite the minimization problem (12) as follows

$$\min_{\mu \in W_2^m[0,1]} \|\mu\|^2_{L^2(f_0)} - 2 < \mu, y >_{L^2(f_0)} + \lambda \|\mu^{(m)}\|^2_{L^2[0,1]} \tag{15}$$

Eq. (15) defines the following Euler conditions

$$\begin{aligned} \lambda \mu^{(2m)}(t) + f_0(t)\mu(t) &= f_0(t) y(t), \quad \forall t \in [0,1] \\ \mu^{(k)}(0) &= \mu^{(k)}(1) = 0, \quad k = m, m+1, ..., 2m-1 \end{aligned} \tag{16}$$

where $\mu^{(k)}$ denotes the k-th derivative of μ.

For each y_i belonging to $L^2(f_0)$, it can be shown that the solution to the boundary value problem (16) exists and is unique, if the corresponding homogenous problem only admits the null solution (see *e.g.* [Mathews and Walker, 1979], and [Gyorfy et al., 2002]). In particular, the solution is determined by the unique Green's function $G_\lambda(t,s)$, such that

$$\hat{\mu}(t) = \int_0^1 G_\lambda(t,s) y(s) f_0(s) ds = < G_\lambda(t,s), y(s) >_{L^2(f_0)} \tag{17}$$

In the smoothing spline literature, the equivalent kernel of the smoothing spline estimator $S_\lambda(t,s)$ is usually obtained by approximating the Green's function $G_\lambda(t,s)$; see *e.g.* [Speckman, 1981], [Cox, 1984a,b], [Silverman, 1984], [Messer, 1991], [Messer and Goldstein, 1993], [Nychka, 1995], and [Chang et al., 2001].

For the case of uniform design density, [Cox, 1984a] computed the Green's function for eq. (16) with periodic boundary conditions by means of Fourier

series, and then fixed the natural boundary conditions (for $m = 2$). [Messer and Goldstein, 1993] determined the Green's function for eq. (16) on the line by means of Fourier transform methods, and then fixed the natural boundary conditions on the finite interval.

On the other hand, for "arbitrary" smooth design densities f_0, [Nychka, 1995] for $m = 1$, [Chang et al., 2001] and [Abramowich and Grinshtein, 1999] for $m = 2$, used the Wentzel-Brillouin-Kramers (WBK) method, although only the latter explicitly mention it. The WKB method applies to the boundary value problem

$$\begin{aligned} \lambda\mu^{(2m)}(t) + f_0(t)\mu(t) &= f_0(t)y(t), \quad \forall t \in \mathbb{R} \\ \mu^{(k)}(t) \to \infty \quad \text{for} \quad t &\to \pm\infty, k = m, m+1, ..., 2m-1 \end{aligned} \tag{18}$$

and deals with the asymptotic behavior of the solution as $\lambda \to 0$.

There are three aspects to take into account in the equivalent kernel set-up: (1) the accuracy of the Green's function as an approximation to the original smoothing spline estimator, (2) the properties of the $G_\lambda(t, s)$ estimator of the regression function, and (3) the convolution kernel like properties of the Green's function.

Concerning the point (1), in the literature several authors have identified approximations of the spline weight function. [Silverman, 1984]'s kernel representation provides an excellent intuition about how a spline estimate weights the data relative to fairly arbitrary distribution of the observation points. [Messer, 1991]'s Fourier analysis gives a high order approximation to the spline estimator for all $m \geq 2$, when $\{t_i\}$ are equally spaced. An extension to the case of unequally spaced observations is given by [Nychka, 1995].

To evaluate the (2) properties of the $G_\lambda(t, s)$ estimator of the regression function, we note that the exact form for the Green's function will depend in a complicated manner on both f_0 and λ. In addition, $G_\lambda(t, s)$ is not a convolution kernel and has a different shape depending on the distance of t and s from the endpoints. However, suppose for the moment that a simple expression for $G_\lambda(t, s)$ is available. Under the assumption of uniform design density, one might consider the approximations

$$\begin{aligned} E[\hat{\mu}(t)] = E\left[\frac{1}{n}\sum_{j=1}^{n} S_\lambda(t, t_j)y_j\right] &= \frac{1}{n}\sum_{j=1}^{n} S_\lambda(t, t_j)\mu(t_j) \\ \approx \int_0^1 S_\lambda(t, s)\mu(s)f_0(s)ds &\approx \int_0^1 G_\lambda(t, s)\mu(s)f_0(s)ds \end{aligned} \tag{19}$$

and

$$Var[\hat{\mu}(t)] \approx \frac{\sigma_\varepsilon^2}{n} \int_0^1 [G_\lambda(t,s)]^2 f_0(s) ds \tag{20}$$

In order to study eq. (19) and eq. (20), it turns out that is not necessary to know the exact form of $G_\lambda(t,s)$. Under suitable restrictions on the rate that λ converges to zero, if μ has $2m$ continuous derivatives, then it is reasonable to expect

$$E[\hat{\mu}(t)] - \mu(t) \approx \frac{(-1)^{m-1}\lambda}{f_0(t)} \mu^{(2m)}(t), \tag{21}$$

and

$$Var[\hat{\mu}(t)] \approx \frac{\sigma_\varepsilon^2 C_m}{n f_0(t)} \left(\frac{f_0(t)}{\lambda} \right)^{1/2m} \tag{22}$$

for t interior of $[0,1]$. Here C_m is a constant depending only on the order of the spline. Now set $\rho(t) = (\lambda / f_0(t))^{1/2m}$, one obtains that

$$E[\hat{\mu}(t) - \mu(t)]^2 \approx \rho(t)^{4m} [\mu^{(2m)}(t)]^2 + \frac{\sigma^2 C_m}{n \rho(t) f_0(t)} \tag{23}$$

In this form, $\rho(t)$ can be interpreted as a variable bandwidth and the accuracy of $\hat{\mu}(t)$ is comparable to a $2m$-th order kernel estimator. Based on the work of [Fan, 1992, 1993], the pointwise mean square error is comparable to that of locally weighted regression estimators. If we wanted to achieve a constant bias or mean square error across t, we would have to consider not only the curvature of μ, $\mu^{(2m)}(t)$, but also the local density of the observations $f_0(t)$. This discussion is only relevant to points t in the interior of $[0,1]$. The bias of a spline estimate at the boundary may exhibit slower convergence rates, depending on the derivatives of μ at the endpoints. This effect has been identified in [Rice and Rosenblatt, 1983] and is also well established for kernel estimators.

To study (3) the kernel like properties of the Green's function, we note that $G_\lambda(t,s)$ is not a convolution kernel, but it has quite analogous properties [Eggermont and LaRiccia, 2005]. In fact, there exist positive constants c, γ and δ such that for all $\lambda > 0$,

$$\begin{aligned} \sup_{t\in[0,1]} \|G_\lambda(t,\cdot)\|_\infty &\leq c\lambda^{-1} \\ \sup_{t\in[0,1]} \|G_\lambda(t,\cdot)\|_1 &\leq c \\ \sup_{t\in[0,1]} \|G_\lambda(t,\cdot)\|_{BV} &\leq c\lambda^{-1} \end{aligned} \tag{24}$$

and for every $t, s \in [0,1]$

$$|G_\lambda(t,s)| \leq \gamma\lambda^{-1}\exp(-\delta\lambda^{-1}|t-s|). \tag{25}$$

In eq. (24), $\|\cdot\|_p$ denotes the standard norm on $L^p(0,1)$, for $1 \leq p \leq \infty$, and $\|\cdot\|_{BV}$ denotes the seminorm on the space of functions (no equivalence classes) of bounded variation on $[0,1]$. Note that convolution kernels have these properties, except for the exponential decay (but obviously, a convolution kernel decays like a L^1 function). Furthermore, as for kernel estimators, the behavior of the Green's function can be analyzed as a function of λ. Particularly, [Eggermont and LaRiccia, 2005] proved that there exists a constant c such that $\forall\lambda > 0, \theta \in [0,1]$ and $\forall p, 1 < p < \infty$

$$\sup_{s\in[0,1]} \|G_\lambda(\cdot,s) - G_\theta(\cdot,s)\|_p \leq c\lambda^{1+1/p}\left|1-\frac{\lambda}{\theta}\right|. \tag{26}$$

Example 2. (Green's function of the cubic smoothing spline) Following [Chang et al., 2001], we consider the cubic smoothing spline problem for fixed time designs. The boundary value problem is characterized by the fourth order differential equation

$$\begin{aligned} \lambda\mu^{(4)}(t) + f_0(t)\mu(t) &= f_0(t)y(t), \quad \forall t \in (0,1) \\ \mu^{(k)}(0) &= \mu^{(k)}(1) = 0, \quad k = 2,3. \end{aligned} \tag{27}$$

Let $G_\lambda(s,t)$ be the Green's function associated with eq. (27). Then, the estimate $\hat{\mu}(t)$ satisfies

$$\hat{\mu}(t) = \int_0^1 G_\lambda(t,s)y(s)f_0(s)ds, \quad \forall t \in [0,1] \tag{28}$$

The Green's function $G_\lambda(s,t)$ is explicitly obtained as the solution of

$$\lambda \frac{\partial}{\partial t^4} G_\lambda(s,t) + G_\lambda(s,t) = \begin{cases} 0 & \text{for} \quad t \neq s \\ 1 & \text{for} \quad t = s \end{cases} \tag{29}$$

subject to the following conditions:

(a) $G_\lambda(s,t) = G_\lambda(t,s) = G_\lambda(1-t, 1-s)$;

(b) $\frac{\partial^v}{\partial t^v} G_\lambda(0,t) = \frac{\partial^v}{\partial t^v} G_\lambda(1,t) = 0, \quad$ for $v = 2, 3$;

(c) $\frac{\partial^3}{\partial t^3} G_\lambda(t,s)|_{s=t^-} = -\frac{\partial^3}{\partial t^3} G_\lambda(t,s)|_{s=t^+} = \frac{1}{\lambda}$.

[Chang et al., 2001] derive explicitly $G_\lambda(t,s)$ when f_0 is the uniform density function. Letting $\gamma = \int_0^1 (f_0(s))^{1/4} ds$ and $\Gamma(t) = \gamma^{-1} \int_0^1 (f_0(s))^{1/4} ds$, they define

$$H_\lambda(t,s) = H^U_{\lambda/\gamma^4}(\Gamma(t), \Gamma(s)) \Gamma^{(1)}(s) (f_0(s))^{-1} \tag{30}$$

to be the equivalent kernel of $S_\lambda(t,s)$, equal to

$$H^U_\lambda(t,s) = \frac{\lambda^{-1/4}}{2\sqrt{2}} \left[\sin\left(\frac{\lambda^{-1/4}}{\sqrt{2}} |t-s| \right) + \cos\left(\frac{\lambda^{-1/4}}{\sqrt{2}} |t-s| \right) \right] \exp\left(-\frac{\lambda^{-1/4}}{\sqrt{2}} |t-s| \right) \tag{31}$$

when $f_0(\cdot)$ is the uniform density.
In general, $H_\lambda(t,s)$ is not the only equivalent kernel that could be considered. Another possibility is to use the one suggested by [Messer and Goldstein, 1993], even if it has not a sharp exponential bound as that given in eq. (31).

An important outcome proved by [Chang et al., 2001] is that $H^U_\lambda(t,s)$ is the dominating term of $G^U_\lambda(t,s)$, as stated in the following lemma.

Lemma 4 *Suppose that $G^U_\lambda(t,s)$ is the Green's function of the differential equation (27) with $f_0(t) = \mathbf{1}_{[0,1]}(t)$. When $\lambda \to 0$, the solution $G^U_\lambda(t,s)$ of eq. (29) is given by*

$$G^U_\lambda(t,s) = H^U_\lambda(t,s) \left\{ 1 + O\left[\exp\left(-\frac{\lambda^{-1/4}}{\sqrt{2}} \right) \right] \right\} \tag{32}$$

where $H_\lambda^U(t,s)$ is defined in eq. (31).

Most of the asymptotic theory for splines is based on showing that the correction term is negligible.

4. The Reproducing Kernel Hilbert Space Approach

The derivation of the Green's function corresponding to a smoothing spline of order m requires the solution of a $2m \times 2m$ system of linear equations for each value of λ. A simplification is provided by studying the function $G_\lambda(t,s)$ as the reproducing kernel of the Sobolev space $W_2^m(T)$, where T is an open subset of $\mathbb{R}$.

A RKHS is a Hilbert space characterized by a kernel that reproduces, via an inner product, every function of the space or, equivalently, a Hilbert space of real valued functions with the property that every point evaluation functional is bounded and linear.

Smoothing splines and reproducing kernel methods are two areas of data analytic research emerge in the mid-1990s, although the essential ideas have been around for much longer. Fundamental references in the smoothing splines literature are [Whittaker, 1923], [Schoenberg, 1946, 1964b], [Wahba, 1990], and more recently [Eilers and Marx, 1996], [Hastie, 1996], and [Ruppert et al., 2002]. On the other hand, an early reference on RKHS theory is [Aronszajn, 1950], and contemporaneous summaries include [Wahba, 1999], [Evgeniou et al., 2000], and [Cristianini and Shawe-Taylor, 2000].

Reproducing kernel methods have become prominent in the nonparametric regression literature as a framework for the smoothing spline methodology, as summarized by [Wahba, 1990]. However, the adoption of these ideas by the machine learning community has widened the scope of reproducing kernel methods quite considerably, in particular for the solution of classification and prediction problems. Kernel based on smoothing splines offer the opportunity to incorporate some principles more straightforwardly than commonly used kernels (see *e.g.* [Pearce and Wand, 2006]).

Each specific application usually requires the use of an adapted RKHS. Let $\{(t_i, y_i), 1 \leq i \leq n\}$ be the dataset, $L(\cdot,\cdot)$ be a loss function and $\lambda > 0$ be a smoothing parameter. The estimate $\hat{\mu}(t)$ within $W_2^m(T)$, with respect to L and

λ , is the solution to

$$\min_{\mu \in W_2^m(T)} \left\{ \sum_{i=1}^{n} L(y_i, \mu(t_i)) + \lambda \|\mu^{(m)}\|^2_{L^2(T)} \right\}. \tag{33}$$

The minimization problem (33) is directly related to the W_2^m-norm, obtained by the seminorm

$$\|\mu^{(m)}\|^2_{L^2(T)} = \int_T \left(\mu^{(m)}(t)\right)^2 dt, \tag{34}$$

determined by the roughness penalty term. Different (topologically equivalent) W_2^m-norms can be constructed taking into account several loss functions. For continuous $y(t_i)$, examples of loss functions are

$$L(a,b) = \begin{cases} (a-b)^2 & \text{(squared error loss)} \\ (|a-b| - \epsilon)_+ & (\epsilon - \text{insensitive loss for some } \epsilon > 0) \end{cases} \tag{35}$$

For $y_i \in \{-1, 1\}$, as arises in two-category classification, examples are

$$L(a,b) = \begin{cases} \log(1 + \exp(-ab))^2 & \text{(Bernoulli log-likelihood)} \\ (1-ab)_+ & \text{(hinge loss)} \end{cases}$$

Some of them are very well-documented in the literature, but there are cases where the expression of the kernel is not available for the nonparametric estimation of functions under shape restrictions.

The reproducing kernel representation of smoothing splines is not unique, but depends on the specific norm we consider. As stated by [Gu and Wahba, 1992]: *¨The norm and the reproducing kernel in a RKHS determine each other uniquely, but like other duals in mathematical structures, the interpretability, and the availability of an explicit form for one part is often at the expenses of the same for the other part¨.*

4. 1. The Spaces and Their Norms

Let T denote an open subset of $\mathbb{R}$, and $W_2^m(T)$ the classical Sobolev Space, *i.e.* the set of functions μ of $L^2(T)$ whose weak derivatives $\mu^{(k)}$, $k = 1, 2, ..., m$,

in the sense of generalized functions[2], belong to $L^2(T)$ ([Adams, 1975]). The classical norms for these spaces are

$$\|\mu\|^2 = \sum_{j=0}^{m} \int_{t\in T} \left(\mu^{(j)}(t)\right)^2 dt \tag{36}$$

but, for our particular application, the following ones seem more appropriate:

$$\|\mu\|^2 = \int \mu(t)^2 dt + \lambda \int_{t\in T} \left(\mu^{(m)}(t)\right)^2 dt. \tag{37}$$

They are simpler to interpret as a weighted sum of the L^2-norms of μ (*square error loss function*) and its m-th derivative $\mu^{(m)}$ (*roughness penalty term*), the parameter λ regulating the balance. Eq. (36) and (37) are topologically equivalent by virtue of the Sobolev inequalities [Agmon, 1965], which can be applied to the cases of the real line or of a bounded open interval.

We now recall some facts about the reproducing kernel Hilbert space theory. Let D_t^k be the derivative functional of order k at the design point t, that is

$$D_t^k(\mu) = \mu^{(k)}(t), \quad \forall t \in T, \forall \mu \in W_2^m(T).$$

If D_t^k is continuous, by the Riesz representation theorem, there exists a representer d_t^k in $W_2^m(T)$ of D_t^k , in the sense that

$$D_t^k(\mu) =< \mu, d_t^k >, \qquad \forall t \in T, \forall \mu \in W_2^m(T).$$

A *reproducing kernel Hilbert space* is a Hilbert space in which the evaluation operators D_t^0 are continuous functionals for all t in T. The function

[2]Let $D(\mathbb{R})$ be the space of infinitely differentiable functions ϕ with compact support, known as *test functions*. Defining the space of *Schwartz distributions* or *generalized functions*, $D'(\mathbb{R})$, to be the space of all continuous linear functionals (the *topological dual space*) of $D(\mathbb{R})$, the *derivative in the sense of distributions* or *weak derivative* of order p, D^p, is a linear operator from $D(\mathbb{R})$ to $D'(\mathbb{R})$ satisfying the formula (obtained by p integration by parts):

$$\int_{\mathbb{R}} D^p f(s)\phi(s)d\lambda(s) = (-1)^{-p} \int_{\mathbb{R}} f(s)D^p\phi(s)d\lambda(s)$$

In other words, given the distribution f, its *distributional derivative of order* k is defined by

$$< f^{(k)}, \phi >= -(-1)^k < f, \phi^{(k)} >, \quad \forall \phi \in D(\mathbb{R})$$

$K(t,s) = d_t^0(s)$ is known as the *reproducing kernel* of the space, and the reader is referred to [Aronszajn, 1950] for more extensive properties.

The Sobolev spaces $W_2^m(T)$ is a RKHS, that is the functionals D_t^k are well defined and continuous, if and only if $m - k > 1/2$ [Wahba and Wendelberger, 1980].

In this setting, our aim is to provide an invariant kernel representation for a smoothing spline of general order m by means of the reproducing kernel (*i.e.* Green's function) $K_{m,\lambda}(t,s)$ of $W_2^m(T)$. Considering the norm (37), this is possible only for the case of the real line. On the other hand, when $T = (a,b)$ or under different norms, the reproducing kernel will vary with the design points, hence it is not appropriate to study the smoothing splines as linear filters.

4. 2. An Hierarchy of Kernel Spline Estimators

When $T = \mathbb{R}$, the space $W_2^m(\mathbb{R})$ falls into the family of Beppo-Levi spaces described in [Thomas-Agnan, 1991]. It follows from the results of such paper that the reproducing kernel is translation invariant, and can be written with a slight abuse of notation as

$$K_{m,\lambda}(t,s) = K_{m,\lambda}(t-s).$$

It is given by

$$\Im K_{m,\lambda}(\omega) = \frac{1}{1 + \left(2\pi\frac{\omega}{\lambda}\right)^{2m}} \tag{38}$$

where $\Im$ denotes the Fourier transform [3], as defined in [Thomas-Agnan, 1991]. Even though the proof can be found in this reference, it is interesting to outline it here in this very simple example. Under the norm (37), by definition, $K_{m,\lambda}$ satisfies $\forall \mu \in W_2^m(\mathbb{R})$

$$\int_{-\infty}^{\infty} \mu(t) K_{m,\lambda}(t,s)dt + \lambda \int_{-\infty}^{\infty} \mu^{(m)}(t) \frac{\partial^m}{\partial t^m} K_{m,\lambda}(t,s)dt = \\ =< \mu(t), K_{m,\lambda}(t,s) >_{W_2^m(\mathbb{R})} = \mu(s). \tag{39}$$

[3] The Fourier transform in $L^2(\mathbb{R})$ may be defined as follows

$$\Im f(\omega) = \int_{-\infty}^{\infty} \exp(-2\pi i\omega s) f(s) d\lambda(s), \quad \forall f \in L^2(\mathbb{R})$$

.

Using the Parseval identity [4] in these two integrals, and the Fourier inversion formula [5] in the right hand side, one easily concludes that the function $K_{m,\lambda}(t,s)$ is solution to the following equation

$$\Im K_{m,\lambda}(\omega,s) + (\lambda 2\pi\omega)^{2m}\Im K_{m,\lambda}(\omega,s) = \exp(-2\pi i\omega s). \tag{40}$$

This first shows that $\Im K_{m,\lambda}(\omega,s) = \exp(-2\pi i\omega s)\Im K_{m,\lambda}(\omega,0)$, and therefore the kernel is translation invariant, and that $\Im K_{m,\lambda}(\omega,0)$ is given by eq. (38). From the formula (38) and the properties of Fourier transform, one concludes that $K_{m,\lambda}$ can be expressed in terms of $K_{m,1}$ by

$$K_{m,\lambda}(t) = \frac{1}{\lambda}K_{m,1}\left(\frac{t}{\lambda}\right). \tag{41}$$

$K_{m,1}$ is a kernel which is familiar to the nonparametric statisticians since it is the asymptotically equivalent kernel to smoothing spline of order m. The theory for this equivalence can be found in [Silverman, 1984] as well as the analytic expression of this kernel for $\lambda = 1$, and $m = 1$ or 2. [Thomas-Agnan, 1991] found a formula for $K_{m,1}$ for general m by contour integration. The result is stated in the following proposition.

Proposition 5

$$K_{m,1}(t) = \sum_{k=0}^{m-1} \frac{\exp\left(-|t|e^{i\frac{\pi}{2m}+k\frac{\pi}{m}-\frac{\pi}{2}}\right)}{2me^{(2m-1)\left(i\frac{\pi}{2m}+i\frac{k\pi}{m}\right)}}. \tag{42}$$

[4] The Parseval's formula says that for f and g in $L^2(\mathbb{R})$

$$\int_{-\infty}^{\infty} \Im f(\omega)g(\omega)d\lambda(\omega) = \int_{-\infty}^{\infty} \Im g(\omega)f(\omega)d\lambda(\omega).$$

Since the Fourier transform defines an automorphism of $D(\mathbb{R})$, the Parseval formula provides a way of extending it into an automorphism of $D'(\mathbb{R})$.

[5] The Fourier inversion formula is given by

$$\Im\Im f(\omega) = f(-\omega).$$

Fourier transform and differentiation in $D'(\mathbb{R})$ satisfy the following identity

$$\Im(D^m f)(\omega) = (2\pi i\omega)^m \Im f(\omega).$$

Proof. To compute the integral $\int_{-\infty}^{\infty} \frac{\exp(2\pi i\omega s}{1+(2\pi\omega)^{2m}} d\omega$, for $s \geq 0$, integrate on the boundary of the upper half disc of the complex plane $\{|z| \leq R, Im(z) \geq 0\}$, and let R tend to ∞. The poles on the upper half plane are $\frac{1}{2\pi} \exp\left(i\frac{\pi}{m} + i\frac{k\pi}{m}\right)$, for $k = 0, ..., m-1$. The integral is then equal to the product of $2\pi i$ by the sum of the residues of the integrand at these poles, which yields eq. (42). ■

The most frequent case of application, which are $m = 1, 2$, and 3, are given explicitly in this corollary.

Corollary 6

$$
\begin{aligned}
K_{1,1}(t) &= \frac{1}{2} \exp\left(-|t|\right) \\
K_{2,1}(t) &= \frac{1}{2} e^{-\frac{|t|}{\sqrt{2}}} \sin\left(|t|\frac{\sqrt{2}}{2} + \frac{\pi}{4}\right) \\
K_{3,1}(t) &= \frac{1}{6}\left\{e^{-|t|} + 2e^{-\frac{|t|}{2}} \sin\left(|t|\frac{\sqrt{3}}{2} + \frac{\pi}{6}\right)\right\}
\end{aligned}
$$

The kernel hierarchy representation of the spline estimator given in eq. (42) enables to derive attractive features from the properties of kernel functions:
(a) the kernels in eq. (42) deform smoothly near the boundary in such a way as to correct for boundary bias;
(b) the kernel can be evaluated by a simple scaling operation on a fixed function;
(c) the scaling function $K_{m,1}(t)$ is available in closed form for all orders of kernel, and for estimating any derivative. It is a sum of exponentially damped trigonometric polynomials.

The key idea is to exploit a certain symmetry in the construction of the Green's function, approximating it by the kernel $K_{m,1}(t)$ which retains the asymptotic properties of $G_\lambda(t, s)$ and allows λ to enter as a scaling parameter. On the other hand, these kernel estimate will inherit many of the properties of the corresponding spline estimate, as shown in [Messer and Goldstein, 1993]. These authors proved that the spline equivalent kernel $K_{m,1}(t)$ of order m can be viewed as containing an interior translation-invariant component which is of order $2m$ at any fixed interior point.

Other reproducing kernel representations of smoothing splines can be derived either by restricting the parametric set T to a bounded interval (a, b) or by considering more general $W_2^m(T)$-norms. In both cases, we cannot find an hierarchy of convolution kernel estimators, as given in eq. (42).

When we restrict our attention to a bounded interval (a, b), the kernel representation of a smoothing spline of order m cannot be derived within the RKHS framework, but we make use of the Green's function property of the equivalent kernel. Therefore, there are not computational gains with respect to what described in section 3.

For general variational problems, other family of norms can be considered. The most frequent ones encountered for the space $W_2^m(T)$ are obtained by completing a seminorm of the form $\int_T (Jf)(t)^2 dt$ into a norm, where J is a linear differential operator of order m. This is usually achieved by choosing a linear operator (of boundary conditions) B from $W_2^m(T)$ to $\mathbb{R}^m$. [Schumaker, 1981] gives some guidance for the initial value problem, and [Dalzell and Ramsay, 1993] for arbitrary B, but none provide an invariant kernel representation. An important subclass of these problems has been solved by means of a different use of the reproducing kernel methodology, as described in the learning machine literature. A brief and simplified introduction to these classes of models can be found in [Wahba, 1990], [Girosi et al., 1995], [Evgeniou et al., 2000], and [Hastie et al., 2001]. These authors show that the solution of a general variational problem is finite dimensional and it can be expressed in terms of the reproducing kernel $K(t, s)$ of the space $W_2^m(T)$

$$\hat{\mu}(t) = \sum_{i=1}^{n} \alpha_i K(t, t_i). \tag{43}$$

One of the main attractive of this formulation is a Bayesian interpretation of such models, in which μ can be viewed as a realization of a zero-mean stationary Gaussian process, with prior covariance function K. However, we do not enter in the details of such an approach, reminding the reader to the references given above. This use of the RKHS methodology does not fit our need to find an invariant general solution for smoothing spline of order m, as given by eq. (42).

5. Smoothing Splines in Time Series Filtering

Spline functions have a long history as smoothers of noisy time series data. Empirical applications can be found in several studies, among others [Poirer, 1973], [Buse and Lim, 1977], [Smith, 1979], [Capitanio, 1996], [Dagum and Capitanio, 1999], [Moshelov and Raveh, 1997], and [Kitagawa and Gersch, 1996].

A basic assumption in time series analysis is that the input series $\{y_t, t = 1, 2, ..., n\}$ can be decomposed into the sum of a systematic component, called the *signal* or nonstationary mean $\mu(t)$, plus an erratic component ε_t, called the *noise*, such that

$$y_t = \mu(t) + \epsilon_t. \tag{44}$$

The noise ϵ_t is assumed to be either a white noise, $WN(0, \sigma_\epsilon^2)$, or more generally to follow a stationary and invertible AutoRegressive Moving Average (ARMA) process.

If the input series is seasonally adjusted or without seasonality, the signal μ represents the trend and cyclical components, usually referred to as trend-cycle for they are estimated jointly. The trend-cycle can be deterministic or stochastic, and have a global or local representation. If μ is differentiable, using the Taylor-series expansion it can be represented *locally* by a polynomial of degree p of the time distance j, between y_t and the neighboring observations y_{t+j}. Hence, given ϵ_t for some time point t, it is possible to find a local polynomial trend estimator

$$\mu_t(j) = a_0 + a_1 j + ... + a_p j^p + e_t(j), \quad j = -h, ..., h \tag{45}$$

where $a_0, a_1, ..., a_p \in \mathbb{R}$ and e_t is assumed to be purely random and mutually uncorrelated with ϵ_t.

The coefficients $a_0, a_1, ..., a_p$ can be estimated by ordinary or weighted least squares or by summation formulae. The solution for $\hat{a}_0$ provides the trend-cycle estimate $\hat{\mu}_t(0)$, which equivalently consists in a weighted average applied in a moving manner [Kendall et al., 1983]. Once a (symmetric) span $2h + 1$ of the neighborhood has been selected, the w_j's for the observations corresponding to points falling out of the neighborhood of any target point are null or approximately null, such that the estimates of the $n - 2h$ central observations are obtained by applying $2h+1$ symmetric weights to the observations neighboring the target point. The missing estimates for the first and last h observations can be obtained by applying asymmetric moving averages of variable length to the first and last h observations respectively, *i.e.*

$$\hat{\mu}_t = \sum_{j=-h}^{h} w_j y_{t-j}, \quad t = h + 1, ..., n - h \quad \text{(central observations)}, \tag{46}$$

$$\hat{\mu}_p = \sum_{r=1}^{h_p} w_r y_r, \quad p = 1, ..., h \quad \text{(initial observations)},$$

$$\hat{\mu}_q = \sum_{z=1}^{h_q} w_z y_{n+1-z}, \quad q = n-h+1, ..., n \quad \text{(final observations)},$$

where $2h+1$ is the length of the time invariant symmetric linear filter and h_p and h_q are the time-varying lengths of the asymmetric filters.
Using the backshift operator B, such that $By_t = y_{t-1}$, eq. (46) can be written as

$$\hat{\mu}_t = \sum_{j=-h}^{h} w_j B^j y_t = W(B) y_t, \tag{47}$$

where $W(B)$ is a linear nonparametric estimator. The nonparametric estimator $W(B)$ is said to be of order p if

$$\sum_{j=-h}^{h} w_j = 1, \tag{48}$$

$$\sum_{j=-h}^{h} j^i w_j = 0, \tag{49}$$

for some $i = 1, 2, ..., p \geq 2$. In other words, it will reproduce a polynomial trend of degree $p-1$ without distortion.

Several nonparametric estimators have been developed, based on different assumptions of smoother building. [Gray and Thomson, 1996a,b] used the same criteria of fitting and smoothing of spline functions to develop a family of local trend linear filters. These authors show that their filters are a generalization of other widely applied smoothers due to [Henderson, 1916]. Within the context of short-term trend estimation for current economic analysis, [Dagum and Capitanio, 1997, 1998] have compared the 13-term Henderson (H13) filter with Cubic Smoothing Splines (CSS). Their results indicated that, for certain fixed values of the smoothing parameter λ, the trend-cycle estimates from CSS were better than those from the H13 on the basis of: (a) number of false turning points in the final estimate of the trend-cycle and (b) time lag to detect a "true" turning

point. Furthermore, [Dagum and Capitanio, 1999] showed how approximate the asymmetric cubic spline predictors by means of time-invariant linear filters, which are symmetric for middle observations and asymmetric for end points. These authors analyzed the main properties of the influential matrix $\mathbf{A}(\lambda)$ for the CSS, using the solution given in eq. (11). They found that, based on a large number of numerical evaluations, as λ decreases (favoring fitting versus smoothing), the nonzero values tend to concentrate along the main diagonal, and to reproduce the same pattern on each row. Furthermore, for fixed λ, the elements of $\mathbf{A}(\lambda)$ do not change for a number of observations $n \geq 30$.

A different time-invariant representation of smoothing splines has been introduced in the previous section by eq. (43). The main advantage of the reproducing kernel formulation is given by the fact that we are able to obtain a time-invariant linear approximation for a smoothing spline of general order m, not only for the cubic one as done by [Dagum and Capitanio, 1999]. Furthermore, an important outcome of the RKHS theory is that smoothing splines can be grouped into a hierarchy identified by the Laplace density f_0, and containing second and higher order estimators which are the products of trigonometric polynomials with f_0.
Here, we want to evaluate the goodness of the reproducing Kernel (KER) formulation with respect to the "classical" Cubic Smoothing Splines (CSS) and to the Linear Approximation (LA) provided by [Dagum and Capitanio, 1999]. The theoretical properties of time-invariant smoothing splines are analyzed by means of spectral techniques, and we compare their performances relative to the classical smoothers using real life series.

5. 1. Theoretical Properties of the Smoothing Spline Hierarchy

The weight system of the spline kernel hierarchy is directly obtained by the kernel functions given in the Corollary 6 by specifying:

(a) the order $2m$ of the kernel, and

(b) the bandwidth parameter λ.

In general, the selection of λ is a crucial task and there is yet no universally accepted approach for this choice. In smoothing spline problems, the trade-off parameter λ is known as hyperparameter in the Bayesian terminology and it has the interpretation of a noise to signal ratio: the larger the λ the smoother the

trend-cycle. The estimation of λ was first done using Ordinary Cross Validation (OCV). OCV consisted of deleting one observation and solving the optimization problem with a trial value of λ, computing the difference between the predicted value and the deleted observation, accumulating the sums of squares of these differences as one runs through each of the data points in turn, and finally choosing the λ for which the accumulated sum is the smallest. This procedure was improved by [Craven and Wahba, 1979] who developed the Generalized Cross Validation (GCV) method available in most computer packages. The GCV estimate of λ is obtained by minimizing

$$V(\lambda) = \frac{(1/n)\left|(I - \mathbf{A}(\lambda))\mathbf{y}\right|^2}{\left[(1/n)tr(I - \mathbf{A}(\lambda)\right]^2} \tag{50}$$

where $\mathbf{A}(\lambda)$ is the influential matrix given in eq. (10), and its trace represents the "degrees of freedom for the signal" and so, eq. (50) can be interpreted as minimizing the standardized sum of squares of the residuals.

In time series linear filtering, the selection of the bandwidth is directly determined by fixing the length of the filter. This latter can be selected according to some criteria, generally the noise to signal (I/C) ratio as done in the non-parametric seasonal adjustment package X11ARIMA. The I/C ratio measures the size of the irregular component in the series; the greater it is, the higher the order of the moving average selected. In order to calculate this ratio, a first decomposition of the Seasonally Adjusted (SA) series is computed using a 13-term Henderson filter. The six "lost" points at the beginning and end of the series are ignored. Hence, we have trend-cycle C and irregular I. We then calculate, for both the C and the I series, the average absolute monthly growth rate (multiplicative model) or the absolute monthly change (additive model), written $\overline{C}$ and $\overline{I}$. Thus, we have:

$$\overline{C} = \frac{1}{n-1}\sum_{t=2}^{n} |C_t/(\text{ or } -)C_{t-1}| \tag{51}$$

$$\overline{I} = \frac{1}{n-1}\sum_{t=2}^{n} |I_t/(\text{ or } -)I_{t-1}| \tag{52}$$

In the X11ARIMA method, the I/C ratio is then computed and:

(a) if it is smaller than 1, a 9-term Henderson moving average is selected;

(b) if the ratio is smaller than 3.5 but greater than 1, a 13-term Henderson filter is chosen;

(c) otherwise, we select a 23-term moving average.

Once the length of the filter has been selected, the bandwidth parameter is chosen to ensure that the 99% of the area under the kernel curve is covered. Such a percentage takes into account the heavy tails of the Laplace density. Hence, we compute the integral

$$\int_{-t}^{t} K(x)dx = 0,99$$

where K is the kernel function under investigation (second, third or fifth order), and $[-t, t]$ denote the symmetric interval at which corresponds a covered area equal to 99%. If we fix the length of the symmetric filter equal to $2h + 1$, then $t = h/\lambda$ and hence $\lambda = h/t$. The number of decimals corresponding to the value $K(t)$ will define the number of decimals in our weights.

Filters of any length, including infinite ones, can be derived in the RKHS framework. In this study, we will consider filters of length 9, 13, and 23 terms in according to those selected for the Henderson filters in the smoothing of monthly series by means of the X11ARIMA procedure.

The properties of linear filters can be studied by analyzing their frequency response functions defined by

$$H(\omega) = \sum_{j=-h}^{h} w_j e^{i\omega j}, \quad 0 \leq \omega \leq 1/2 \tag{53}$$

where w_j are the weights of the filter and ω is the frequency in cycles per unit of time. In general, the frequency response functions can be expressed in polar form as follow,

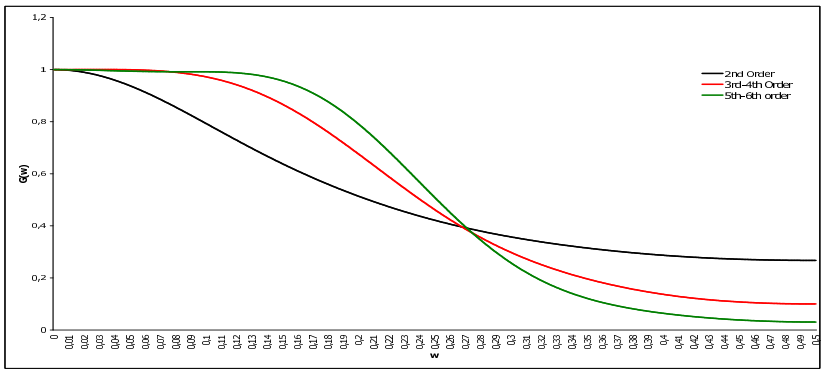
$$H(\omega) = G(\omega)e^{i\phi(\omega)} \tag{54}$$

where $G(\omega)$ is called the gain of the filter and $\phi(\omega)$ is called the phase shift of the filter and is usually expressed in radians. The expression (54) shows that if the input function is a sinusoidal variation of unit amplitude and constant phase shift $\psi(\omega)$, the output function will also be sinusoidal but of amplitude $G(\omega)$ and phase shift $\psi(\omega) + \phi(\omega)$. The gain and phase shift vary with ω. For symmetric filters the phase shift is 0 or $\pm\pi$, and for asymmetric filter takes

values between $\pm\pi$ at those frequencies where the gain function is not zero. For a better interpretation the phase shifts are often given in months instead of radians, that is $\phi(\omega)/2\pi\omega$ for $\omega \neq 0$.

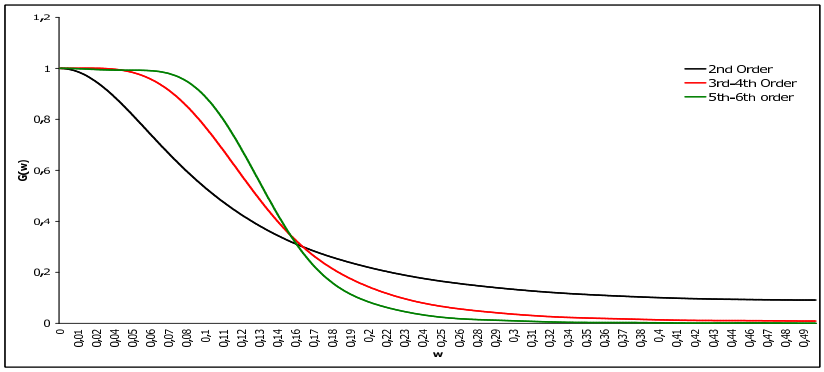
Figure 1 shows the gain functions of the symmetric 13-term filters within the spline hierarchy (similar results are obtained for the 9-term kernels).

Figure 1: Gain functions of symmetric 13-term spline kernels



The three filters pass a lot of power at all the frequencies, in particular at the highest ones associated to the noise, with the worst performance for the second order kernel. This is due to the fact that when we restrict the spline to be a time invariant filter of short length, part of the smoothing properties of such estimators are no longer optimal (see *e.g.* [Dagum and Luati, 2004]). The bandwidth parameter selected to ensure filter of 13 terms are equal to 0.875, 0.667 and 0.633 for the second order, third/fourth one, and fifth/sixth order kernel, respectively. Since the bandwidth is directly related to the smoothing parameter λ which appears in the minimization problem (12), those values tend to give more importance to the fitting part respect to the smoothing one. From the view point of signal passing and noise suppression, the 23-term symmetric filters present better properties, as illustrated in Figure 2. In this case, the bandwidths selected are equal to 1.604 for the second order kernel, to 1.224 for the third/fourth order one, and 1.162 for the fifth/sixth order filter. The higher order kernels perform better in terms of trend-cycle estimators, than the second order one, and furthermore, the third order filter tends to suppress more power at the frequency $\omega = 0.10$, related to cycle of 10 months, often interpreted as false turning points.

Figure 2: Gain functions of symmetric 23-term spline kernels



The third order kernel within the hierarchy provides a new representation of the CSS in general and of its linear approximation studied by [Dagum and Capitanio, 1999]. This has important consequences in the derivation of the asymmetric filters, in particular for that corresponding to the last point, which is the most important in current economic analysis.

Based on the results of the section 2 (eq. 10), we obtain, for central values, the following 13-term LA filter:

$$y_t = \sum_{j=-6}^{6} \mathbf{A}(\lambda)_j y_{t+j} \tag{55}$$

where for $\lambda = 0.11$, the symmetric weights are:

$$\left[\begin{array}{ccccccc} 0.0005 & 0.0011 & -0.0037 & -0.0212 & -0.0062 & 0.2303 & \mathbf{0.5984} \end{array}\right].$$

For the last observation we have

$$y_t = \sum_{j=-6}^{0} \mathbf{A}(\lambda)_j y_{t+j} \tag{56}$$

where for $\lambda = 0.11$, the asymmetric weights are:

$$\left[\begin{array}{ccccccc} 0.0004 & 0.0019 & -0.0004 & -0.0231 & -0.0482 & 0.1564 & \mathbf{0.9132} \end{array}\right].$$

A comparison is then performed with the 13-term third order kernel $K_{2,1}$ within

the hierarchy, whose weights for the central observations are obtained as follows

$$w_j = \frac{K_{2,1}(j/\lambda)}{\sum_{i=-6}^{6} K_{2,1}(i/\lambda)}, \quad j = -6, ..., 6 \tag{57}$$

and given by

$$\left[\begin{array}{ccccccc} 0.0010 & -0.0007 & -0.0102 & -0.0227 & 0.0210 & 0.2485 & \mathbf{0.5265} \end{array} \right].$$

The last point asymmetric weights are derived by $K_{2,1}$ adapted to the length of the filter, that is

$$w_j = \frac{K_{2,1}(j/\lambda)}{\sum_{i=-6}^{0} K_{2,1}(i/\lambda)}, \quad j = -6, ..., 0 \tag{58}$$

hence, equal to

$$\left[\begin{array}{ccccccc} 0.0013 & -0.0010 & -0.0134 & -0.0298 & 0.0275 & 0.3255 & \mathbf{0.6898} \end{array} \right].$$

Figure 3 shows that the gain of the symmetric kernel performs similarly to that of LA, but the former passes less noise than the latter. On the other hand, there is a strong better performance for the last point asymmetric kernel, that does not amplify the gain power has done by LA, as illustrated in Figure 4.
For both the filters the phase shift is less than one month (Figure 5). We do not show here the results for filters of length 9 and 23 terms, since the conclusions drawn are similar.

Figure 3: Gain functions of symmetric 13-term CSS and third order spline kernel

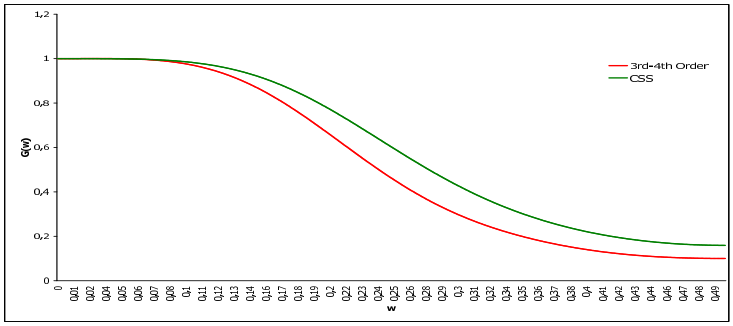


Figure 4: Gain functions of (last point) asymmetric 7-term CSS and third order spline kernel

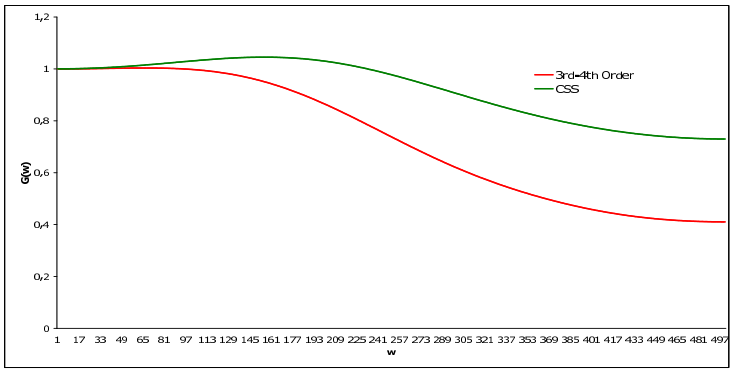
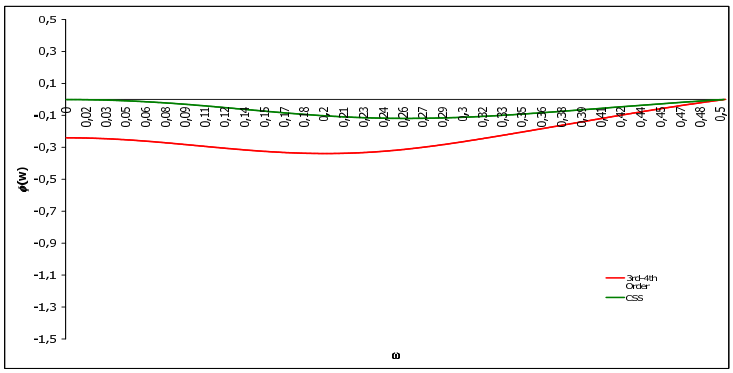


Figure 5: Phase shift functions of (last point) asymmetric 7-term CSS and third order spline kernel



5. 2. An Empirical Application

The comparison of the trend-cycle estimates obtained with Cubic Smoothing Splines (CSS), the Linear Approximation (LA) due to [Dagum and Capitanio, 1999], and the KErnel Representation (KER) is done as follows:

(1) the input series for the three types of estimators is a seasonally adjusted series which has been modified by replacing all extreme values with zero weights. The identification and replacement of extreme values is done with the default option of X11ARIMA which defines as extreme value with zero weight any irregular falling outside $\pm 2.5\sigma$.
(2) The CSS trend-cycles are estimated using the package "pspline" in the R software based on [Heckman and Ramsay, 1996]. We required to select automatically the number of knots and to estimate the corresponding smoothing parameter λ using the generalized cross-validation.
(3) We looked at the value of the I/C ratio given by the X11ARIMA method which would have been used to determine the appropriate length of the time invariant linear approximation and kernel representation of the CSS, in according to the range used by the procedure to select the Henderson filter.
(4) The comparison among these three types of trend-cycle estimators is based on measures of fidelity and smoothness, as suggested by [Gray and Thomson, 1996a]. Fidelity is more commonly known as Mean Square Error (MSE), calculated as an average of the squared differences (residuals) between the observed and the estimated series. For comparative purposes we need an adimensional measure, and thus, the MSE is standardized by the observations. That is,

$$MSE = \frac{1}{n-2h} \sum_{t=h+1}^{n-h} \left(\frac{y_t - \hat{y}_t}{y_t} \right)^2 \tag{59}$$

where y_t denotes the observed value at time t, $\hat{y}_t$ the corresponding estimate (applying the $2h+1$-term symmetric filter), and n is the length of the series.

Smoothness is measured by the sum of squares of the third differences of the estimated values, still divided by the observed data in view of standardizing Q, that is

$$Q = \sum_{t=4}^{n} \left(\frac{\Delta^3 \hat{y}_t}{y_t} \right)^2 . \tag{60}$$

The smaller Q, the closer $\Delta^3 \hat{y}_t$ is to zero, and the closer the estimated curve $\hat{y}_t$ is to a second-order polynomial in t.
In general, there is an inverse relationship between the MSE and Q. The smaller MSE, the higher is Q. This is due to the fact that Q depends on the variability of the final output whereas the MSE depends on the residual vari-

ance. Minimizing MSE ensures the trend estimate is in some sense close to the "true" value, whereas minimizing Q ensures that the fitted trend polynomial is close to a smooth polynomial of degree 2. A smoother is considered optimal if eliminates all the noise power without modifying the signal, this is an ideal case. In smoother construction, there is always a compromise between signal passing and noise suppression. Hence, if a smoother leaves too much noise in $\hat{y}_t$, the value of Q will be high. On the other hand, a filter which removes all the noise can at the same time suppress part of the signal and then will give a small value of Q. For these reasons, it is important to consider the two measures simultaneously, to evaluate how optimal is the performance of a smoother.

The trend-cycle estimators were applied to a sample of twenty Italian economic indicators, but to illustrate how the CSS functions respond to the variability of the data and compare with the two approximations, we have selected three typical cases from our sample.

The three Italian economic indicators are the Index of Industrial Production of Energy (IIPE), Orders of Durable Goods (ODG), and Total Exportations (TE). These are monthly series that cover the periods January 1990 - December 2006 for IIPE and ODG, and January 1991 - December 2006 for TE. Their corresponding seasonally adjusted data are further modified by extreme values as described in (1) above, and the CSS, LA and KER trend-cycles are estimated.

Case 1. Index of Industrial Production of Energy

To estimate the CSS trend of the IIPE series we select the generalized cross-validation technique for the estimation of the parameter λ, using the package "pspline" in the software R, and we obtained an estimated value equal to 0.41.

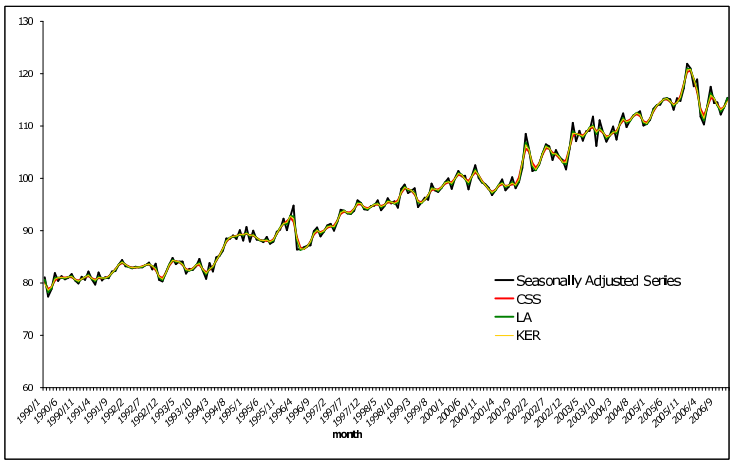
We then evaluate the variability of this series and found that the I/C ratio given by X11ARIMA was 2.27 indicating that this method would have chosen the 13-term filter for the estimation of the trend-cycle. Therefore, we obtained the 13-term linear approximation of the CSS by choosing λ equal to 0.11, and a bandwidth parameter equal to 0.667 to obtain a third order kernel of the same length.

Figure 6 shows analogous patterns for the trend-cycle estimates, given the similarity of the parameters selected by the three estimators.

The fitting and smoothing measures confirmed the graphical analysis. The MSE is more or less the same for the three estimators, equal to 0.000008 for the CSS, to 0.000004 for LA, and equal to 0.000006 for KER. On the other hand, the smallest value of the Q measure is obtained for the CSS (0.015), whereas

the kernel has a better performance relative to the LA with a Q value equal to 0.036 for the former and 0.066 for the latter.

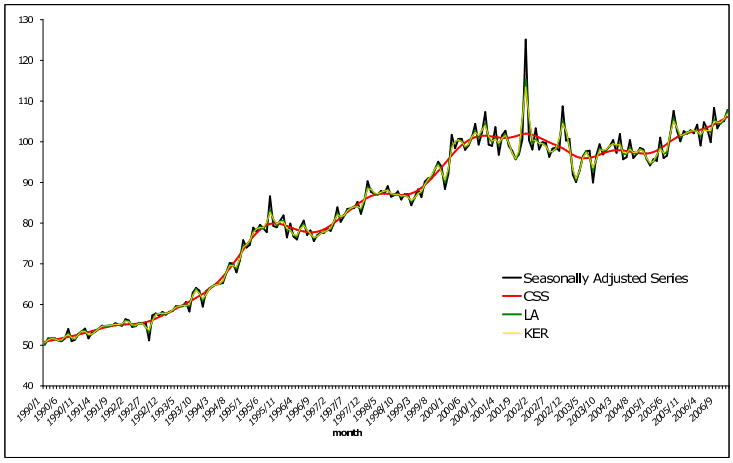
Figure 6: Seasonally adjusted IIPE series and its trend-cycle estimates



Case 2. Orders of Durable Goods

For this series, the I/C ratio is equal to 2.39, indicating that the default option of the X11ARIMA still selects the 13-term filter. On the other hand, the GCV selection of λ is 118.80, hence in the minimization problem (12) more importance will be given to the smoothing part relative to the fitting one.
Figure 7 clearly shows that the LA and KER provide similar estimates, that tend to undersmooth the data; whereas the smoothest trend-cycle estimates are obtained for the CSS.
The MSE is larger for the CSS (0.0008), confirming that this filter interpolates less the data points, whereas the LA and KER present similar results, with the fidelity measure equal to 0.0002 for the former and 0.0003 for the latter. On the other hand, the CSS presents the smallest value of Q, equal to 0.00001, followed by the KER, with a value of 0.16804, and the LA has the worst performance with $Q = 0.32136$.

Figure 7: Seasonally adjusted ODG series and its trend-cycle estimates

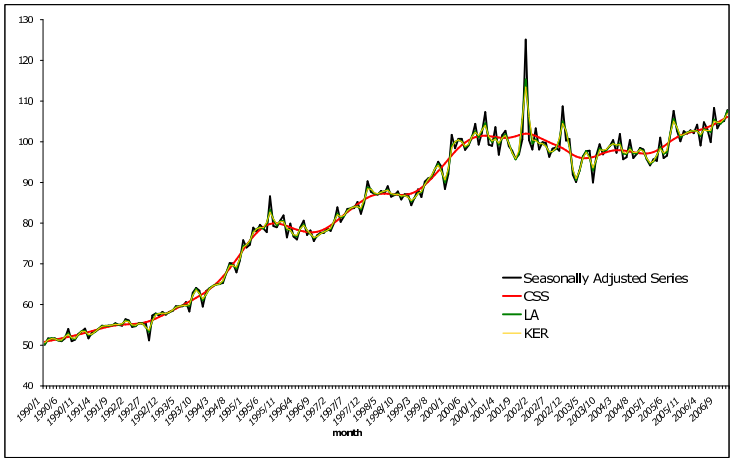


Case 3. Total Exportations

The automatic CSS for the Italian total exportations series produces a smoother trend than LA and KER, as for the ODG series. This is due to the fact that the selected parameters are strongly different. In particular, the GCV estimate is 15.39, whereas the I/C ratio for the series is equal to 2.27, indicating the appropriateness of the 13-term filter. Therefore, the selected smoothing parameters are 0.11 and 0.667 for the LA and KER, respectively. This implies that the LA and KER give very similar and poor results from a trend-cycle perspective (see Figure 8). A better compromise is performed by CSS as confirmed by the fitting measure, equal to 0.0002 for both LA and KER, and to 0.0005 for CSS, and by the smoothing measure Q, that is 0.0002 for CSS, 0.1053 for KER, and 0.2078 for LA.

Similar cases were present in other series of the sample, but if we use the smoothing parameter determined by GCV in the two time-invariant representations, we will obtain symmetric filters of much longer lengths than those selected by the I/C ratio.

Figure 8: Seasonally adjusted ODG series and its trend-cycle estimates



In current economic analysis, long filters are not useful, since they are made of a large number of asymmetric weights that generate revisions and phase shifts making difficult the detection of true turning points. On the other hand, with the constraint of all the filters to be of a fixed length, their statistical properties, and those of the belonging hierarchy, are no longer necessarily optimal as in the case when the optimal smoothing parameter is chosen. Hence, the statistical properties of the fixed length smoothers have to be studied within the context of their respective weighting systems.

6. Conclusions

In this study we derived a kernel representation of smoothing splines by means of the Reproducing Kernel Hilbert Space (RKHS) methodology. We made use of the reproducing kernel property of the Green's function which solves the spline minimization problem.

We showed that the third order kernel is quite close to the time-invariant representation of the cubic smoothing spline derived by [Dagum and Capitanio, 1999], with a better performance of the former in terms of signal passing and noise suppression. The kernel representation has a computing advantage, in the sense that it can be derived for every smoothing spline of general order m,

whereas the linear approximation provided by [Dagum and Capitanio, 1999] is valid only for the cubic case. The symmetric weights of the kernel representation and those of the linear approximation are closer as the span of the filter increases, and we considered those lengths most often applied to monthly data. Furthermore, there are important consequences in the derivation of the asymmetric filters, in particular for that corresponding to the last point that is the most important in current economic analysis.

Applied to real time series, the kernel representation, whose length is selected according to the I/C ratio, performs worse than a non linear cubic smoothing spline with smoothing parameter determined by GCV. The use of the GCV value in the time-invariant representations is not a solution, since we will obtain symmetric kernels of too long lengths. This implies a larger number of asymmetric weights that generate revisions and phase shifts making difficult the detection of true turning points. Hence, the statistical properties of the fixed length smoothers need to be firstly studied within the context of their respective weighting systems.

References

F. Abramowich and V. Grinshtein. Derivation of equivalent kernels for general spline smoothing: a systematic approach. *Bernoulli*, 5:359–379, 1999.

R.A. Adams. *Sobolev spaces*. Academic press, Inc, Harcourt Brace Jovanovich publishers, 1975.

S. Agmon. *Lectures on elliptic boundary value problems*. D. Van Nostrand, Princenton NJ, 1965.

N. Aronszajn. Theory of reproducing kernels. *Transaction of the AMS*, 68: 337–404, 1950.

D. Boneva, L. Kendall and I. Stefanov. Splines transformations. *Journal of Royal Statistical Society, Ser. A*, 33:1–70, 1971.

E.B Boser, I.M. Gyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. *in Proceedings of the 5th annual ACM workshop on computational learning theory*, ed. D. Haussler, New York: ACM Press:144–152, 1992.

A. Buse and L. Lim. Cubic splines as a special case of restricted least squares. *Journal of american statistical association*, 72:64–68, 1977.

A. Capitanio. Un metodo non parametrico per l'analisi della dinamica della temperatura basale. *Statistica*, LVI, 2:189–200, 1996.

C. Chang, J. Rice, and C. Wu. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of American Statistical Society*, 96:605–619, 2001.

D.D. Cox. Asymptotics of m-type smoothing splines. *Annals of statistics*, 11: 530–551, 1984a.

D.D. Cox. Multivariate smoothing spline functions. *SIAM journal of numerical analysis*, 21:789–813, 1984b.

P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerical mathematics*, 31:377–403, 1979.

N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press, 2000.

E.B. Dagum and A. Capitanio. New results on trend-cycle estimation and turning point detection. *Proceedings of the business and economic statistics section*, pages 223 – 228, 1997.

E.B. Dagum and A. Capitanio. Smoothing methods for short-term trend analysis: cubic splines and henderson filters. *Statistica*, LVIII(1), 1998.

E.B. Dagum and A. Capitanio. Cubic spline spectral properties for short term trend-cycle estimation. *Proceedings of the business and economic section of the American Statistical Association*, pages 100–105, 1999.

E.B. Dagum and A. Luati. Relationship between local and global nonparametric estimators measures of fitting and smoothing. *Studies in Nonlinear Dymanics and Econometrics*, Volume 8.2:No 17, 2004.

D. Dalzell and J. O. Ramsay. Computing reproducing kernels with arbitrary boundary constraints. *SIAM Journal of Scientific Computing*, 14:511–518, 1993.

C. De Boor. *A practical guide to splines.* Springer-Verlag, New York, 1978.

C. De Boor and R. Lynch. On splines and their minimum properties. *J. Math. Mech.*, 15:953–969, 1966.

J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. *in Constructive theory of functions of several variables*, Springer-Verlag, Berlin:85–100, 1977.

P.P.B. Eggermont and V.N. LaRiccia. Equivalent kernels for smoothing splines. *Unpublished Manuscript*, pages 1–28, 2005.

P.H.C. Eilers and B.D. Marx. Flexible smoothing with b-splines and penalties (with discussion). *Statistical science*, 1996.

R.L. Eubank. *Spline smoothing and nonparametric regression*. New York: Marcel Dekker, 1988.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advanced in computational mathematics*, 13:1–50, 2000.

J. Fan. Design-adaptive nonparametric regression. *JASA*, 87:998–1004, 1992.

J. Fan. Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, 21:196–216, 1993.

R. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7:219–269, 1995.

M. Golomb and H. Weinberger. Optimal approximation and error bounds. *Proc. Symp. on Numerical Approximation,*, R. Langer eds.(University of Wisconsin Press, Madison, WI), 1959.

A. Gray and P. Thomson. Design of moving-average trend filters using fidelit and smoothness criteria. *in Time series analysis in memory of E.J. Hannan*, P.M. Robinson and M. Rosenblatt eds:205–219, 1996a.

A. Gray and P. Thomson. On a family of moving-average trend filters for the ends of series. *Proocedings of the business and economic statistics section*, American statistical association annual meeting, Chicago, 1996b.

P.J. Green and B.W. Silverman. *Nonparametric regression and generalized linear models*. London: Chapman and Hall, 1994.

T. Greville. *Theory and application of spline functions*. University of Wisconsin Press, Madison, WI, 1968.

C. Gu and G. Wahba. Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. *SIAM J. Sci. Statist. Comput.*, 12:383–398, 1992.

L. Gyorfy, M. Kohler, A. Krzyzak, and A. Walk. *A distribution-free theory of nonparametric regression*. New York: Springer-Verlag, 2002.

T. Hastie. Pseudosplines. *Journal of royal statistical society, series B*, 1996.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. New York: Springer-Verlag, 2001.

N. Heckman and J. O. Ramsay. Spline smoothing with model based penalties. *Unpublished manuscript*, McGill University, 1996.

R. Henderson. Note on graduation by adjusted average. *Transactions of the actuarial society of America*, 17:43–48, 1916.

M. G. Kendall, A. Stuart, and J.K. Ord. *The Advanced Theory of Statistics, Vol. 3*. C. Griffin, 1983.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline fucntions. *J. Math. Anal. Appl.*, 33:82–95, 1971.

G. Kitagawa and W. Gersch. *Smoothness priors analysis of time series*, volume Lecture notes in statistics, 116. New York: Springer-Verlag, 1996.

P. Lancaster and K. Salkauskas. *Curve and surface fitting: an introduction*. Academic press, London, 1986.

X. Lin and R.J. Carroll. Semiparametric regression for clustered data with a nonparametric cluster-level component. *citeseer.ist.psu.edu/article/lin00semiparametric.html*, 2000.

J. Mathews and R.L. Walker. Mathematical methods of physics. 1979.

J. Meinguet. Multivariate interpolation at arbitrary points made simple. *J. Appl. Math. Phys. (ZAMP)*, 30:292–304, 1979.

K. Messer. A comparison of spline estimate to its equivalent kernel estimate. *Annals of Statistics*, 19:817–829, 1991.

K. Messer and L. Goldstein. A new class of kernels for nonparametric curve estimation. *Annals of Statistics*, 21:179–196, 1993.

G. Moshelov and A. Raveh. On trend estimation of time series: a simple linear programmng approach. *Journal of the operational research society*, 1997.

D. Nychka. Splines as local smoothers. *Annals of Statistics*, 23:1175–1197, 1995.

E. Parzen. An approach to time series analysis. *Annals of mathematical statistics*, 32:951–989, 1962.

E. Parzen. Statistical inferences on time series by rkhs methods. *in Proc. 12th biennial seminar, R. Pyke ed.*, canadiam mathematical congress, Montreal, Canada:1–37, 1970.

N.D. Pearce and M.P. Wand. Penalized splines and reproducing kernel methods. *The american statistician*, 60(3), 2006.

D.J. Poirer. Piecewise regression using cubic splines. *Journal of the American Statistical Association*, 68:515–524, 1973.

P. Prenter. *Splines and variational methods*. John Wiley, New York, 1975.

J. Rice and M. Rosenblatt. Smoothing splines: regression, derivatives and deconvolution. *Annals of statistics*, 11:141–156, 1983.

D. Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression*. New York: Cambridge university press, 2002.

I. Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. *Quart. Appl. Math*, 4:45–99, 1946.

I. Schoenberg. Monosplines and quadrature formulae. *in Theory and applications of spline functions*, ed. T. Greville, Madison, WI: University of Wisconsin press., 1964a.

I.J. Schoenberg. Spline functions and the problems of graduation. *Proceedings of the National Academy of Sciences, USA*, 52:947–950, 1964b.

L. Schumaker. *Spline functions*. John Wiley, New York, 1981.

B. Silverman. Spline smoothing: the equivalent kernel method. *Annals of statistics*, 12:898–916, 1984.

P.L. Smith. Splines as a useful and convenient statistical tool. *The American Statistician*, 33:57–62, 1979.

P.L. Speckman. The asymptotic integrated mean square error fro smoothing noisy data by splines. *Manuscript, University of Oregon*, 1981.

C. Thomas-Agnan. Splines functions and stochastic filtering. *Annals of statistics*, pages 1512–1527, 1991.

F. Utreras. Cross-validation techniques for smoothing spline functions in one or two dimensions. *in Smoothing techniques for cureve estimation, T. Gasser and M. Rosenblatt, eds.*, Springer-Verlag, Heidelberg:196–231, 1979.

G. Wahba. *Spline models for observational data*. Philadelphia: SIAM, 1990.

G. Wahba. *Support vector machine, reproducing kernel Hilbert spaces, and randomized GACV*. in Advanced in kernel methods: support vector learning, eds B. Scholkopf, C. Burges, and A. Smola, Cambridge, MA: MIT press, 1999.

G. Wahba and J. Wendelberger. Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly weather review*, pages 1122 – 1143, 1980.

E.T. Whittaker. On a new method of graduation. *Proceedings of the Edinburg mathematical association*, 78:81–89, 1923.