Estela Bee Dagum and Silvia Bianconcini

A Unified Probabilistic View of Nonparametric Predictors via Reproducing Kernel Hilber Spaces

Quaderni di Dipartimento

Serie Ricerche 2008, n. 6
ISSN 1973-9346

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Scienze Statistiche "Paolo Fortunati"

A unified probabilistic view of nonparametric predictors via reproducing kernel Hilbert spaces

Estela Bee Dagum and Silvia Bianconcini

Department of Statistics, University of Bologna

Via Belle Arti, 41 - 40126 Bologna, Italy

estela.beedagum@unibo.it, silvia.bianconcini@unibo.it

Abstract

We provide a common approach for studying several nonparametric estimators used for smoothing functional data. Linear filters based on different building assumptions are transformed into kernel functions via reproducing kernel Hilbert spaces. For each estimator, we identify a density function or second order kernel, from which a hierarchy of higher order estimators is derived. These are shown to give excellent representations for the currently applied symmetric filters. In particular, we derive equivalent kernels of smoothing splines in Sobolev and polynomial spaces. The asymmetric weights are obtained by adapting the kernel functions to the length of the various filters, and a theoretical and empirical comparison is made with the classical estimators used in real time analysis. The former are shown to be superior in terms of signal passing, noise suppression and speed of convergence to the symmetric filter.

Keywords: polynomial kernel regression, smoothing splines, functional spaces, spectral properties, revisions.

1. Introduction

The estimation of the nonstationary mean of a time series is of great interest in most scientific areas where observations are measured with error, such as economics, finance, biostatistics, health, hydrology, labor market. Based on the assumption that the time series can be decomposed as a sum of a signal plus an erratic component, signal extraction deals with the problem of finding the "best" estimate of the nonstationary mean given the observations corrupted by noise.

A common approach is to use nonparametric smoothing estimators which can be finite or infinite in length, under the main assumption that the signal is a smooth function of time. The main methods discussed here, based on different criteria of fitting and smoothing, are: (a) density functions, (b) local polynomial fitting, (c) graduation theory, and (d) smoothing spline regression.

A unified perspective for all of these different nonparametric estimators is provided by means of the Reproducing Kernel Hilbert Space (RKHS) methodology.

The main theory and systematic development of reproducing kernels and associated Hilbert spaces was laid out by Aronszajn (1950), who showed that the properties of RKHS are intimately bounded up with properties of nonnegative definite functions. Parzen (1959) was the first to apply these concepts to time series problems by means of a strictly parametric approach. From a nonparametric perspective, De Boor and Lynch (1966) used this methodology in the context of cubic spline approximation. Later, Kimeldorf and Wahba (1971) proved that minimum norm interpolations and smoothing problems with quadratic constraints imply an equivalent Gaussian stochastic process.

Recently, reproducing kernel methods have been prominent as a framework for penalized spline and quantile regression (see *e.g.* Wahba 1990), and in the support vector machine literature, as described in Wahba (1999), Evgeniou, Pontil, and Poggio (2000), Cristianini and Shawe-Taylor (2000), and Pearce and Wand (2006).

In this study we show how nonparametric estimators can be transformed into kernel functions of order two, that are probability densities, and from which corresponding hierarchy of estimators are derived. The density function provides the "initial weighting shape" from which the higher order kernels inherit their properties. This kernel representation enables the comparison of estimators based on different smoothing criteria, and has important consequences in the derivation of the asymmetric filters which can be applied to the most recent observations. In particular, those obtained by means of RKHS are shown to have superior properties from the view point of signal passing, noise suppression and revisions relative to the classical ones.

Section 2 describes the unified approach of RKHS to nonparametric estimation, and discusses the classical symmetric smoothers and their kernel representation. Section 3 illustrates, theoretically and empirically, the behavior at the boundaries of equivalent reproducing kernels, with particular emphasis on those corresponding to cubic smoothing splines. Finally, Section 4 gives the conclusions.

2. Nonparametric estimators in RKHS

Let $\{y_t, t = 1, 2, ..., N\}$ denote the time series, specified as follows

Assumption 1 *The time series $\{y_t, t = 1, 2, ..., N\}$ can be decomposed into the sum of a systematic component, called the signal (or nonstationary mean) $g(t)$, plus an erratic component u_t, called the noise, such that*

$$y_t = g(t) + u_t. \tag{1}$$

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

Assumption 2 *The time series $\{y_t, t = 1, 2, ..., N\}$ is a finite realization of a family of square Lebesgue integrable random variables,* i.e. $\int_T Y(t)^2 dt < \infty$, *$T \subseteq \mathbb{R}$. Hence, the random process $\{Y(t)\}_{t \in T}$ belongs to the space $L^2(T)$.*

$L^2(T)$ is a Hilbert space endowed with the inner product defined by

$$< Y(t), Y(s) >_{L^2(T)} = E(Y(t)Y(s)) = \int_T Y(t)Y(s)f_0(t)f_0(s)dtds \tag{2}$$

where $Y(t), Y(s) \in L^2(T)$, and f_0 is a probability density function which weights each observation to take into account its position in time. In the following, $L^2(T)$ will be indicated as $L^2(f_0)$.

If the time series is without seasonality or seasonally adjusted, the signal g represents the trend and cyclical components, usually referred to as trend-cycle when they are estimated jointly in rather short series, say 12 years or less. The determination of a suitable inferential methodology for model (1) will hinge on the assumptions made about g.

Assumption 3 *The signal g is a smooth function, that is g belongs to the Sobolev space $W_2^{p+1}(T) \subseteq L^2(f_0)$.*

$W_2^{p+1}(T)$ is the set of functions $g \in L^2(f_0)$, whose weak derivatives $g^{(k)}, k = 1, 2, ..., p+1$, in the sense of distributions, belong to $L^2(f_0)$ (Adams 1975).
Under Assumption 3, g can be locally approximated by a polynomial function of the time distance j between y_t and the neighboring observations y_{t-j}, such that

$$g_t(j) = a_0 + a_1 j + ... + a_p j^p + \varepsilon_t(j), \quad j = -m, ..., m \tag{3}$$

where $a_0, a_1, ..., a_p \in \mathbb{R}$, and ε_t is assumed to be purely random and mutually uncorrelated with u_t. Therefore, the analysis of the signal can be performed in the space $\mathbf{P}_p \subset L^2(f_0)$ of polynomials of degree at most p, being p a non-negative integer.

The coefficients $a_0, a_1, ..., a_p$ of the polynomial nonstationary mean are estimated by projecting the observations in a neighborhood of y_t on the subspace $\mathbf{P}_p$, or equivalently by minimizing the weighted least square fitting criterion

$$\min_{g \in \mathbf{P}_p} \|y_{t+j} - \hat{g}_t(j)\|^2_{\mathbf{P}_p} = \int_{-m}^{m} (y_{t+j} - \hat{g}_t(j))^2 f_0(j) dj, \tag{4}$$

where $\| \cdot \|^2_{\mathbf{P}_p}$ denotes the $\mathbf{P}_p$-norm, and the positive real number m determines the neighborhood of t on which the deviation between y_{t+j} and $\hat{g}_t(j)$ is taken into account in the L^2-sense. For this reason, $2m+1$ is called the bandwidth. The weighting function f_0 depends on the distance between the target point t and each observation in the $2m+1$ points neighborhood (for $m+1 \leq t \leq N-m$).
The solution for $\hat{a}_0$ provides the estimate $\hat{g}_t(0)$, for which a general characterization and explicit representation can be provided by means of the RKHS methodology.

Definition 1 *Given a Hilbert space $\mathcal{H}$, the reproducing kernel R is a function*

$$\begin{aligned} R: T \times T &\rightarrow \mathbb{R} \\ (s,t) &\mapsto R(s,t), \end{aligned}$$

satisfying the following properties:

1. *$R(t,.) \in \mathcal{H}$, $\forall t \in T$;*
2. *$< g(.), R(t,.) >= g(t)$, $\forall t \in T$ and $g \in \mathcal{H}$.*

Condition 2. is called the *reproducing property*: the value of the function g at the point t is reproduced by the inner product of g with $R(t,.)$. $R(t,.)$ is called the *reproducing kernel* since

$$< R(t,.), R(.,s) >= R(t,s). \tag{5}$$

Corollary 1 *The space $\mathbf{P}_p$ is a reproducing kernel Hilbert space of polynomials on some domain T, that is there exists an element $R_p(t;.) \in \mathbf{P}_p$, such that*

$$P(t) =< P(.), R_p(t;.) > \quad \forall t \in T \quad \textit{and} \quad \forall P \in \mathbf{P}_p$$

The proof easily follows by the fact that any finite dimensional Hilbert space has a reproducing kernel (see for details Berlinet and Thomas-Agnan 2003).

Theorem 2 *Under the Assumptions 1, 2, and 3, the minimization problem (4) has a unique and explicit solution given by*

$$\hat{g}_t = \int_{-m}^{m} y_{t+j} K_{p+1}(j) dj \tag{6}$$

where K_{p+1} is a kernel function of order $p+1$.

Proof. By the projection theorem (see *e.g.* Priestley 1981), each element y_{t+j} of the Hilbert space $L^2(f_0)$ can be decomposed into the sum of its projection in a Hilbert subspace of $L^2(f_0)$, such as the space $\mathbf{P}_p$, plus its orthogonal complement as follows

$$y_{t+j} = \Pi_{P_p}[y_{t+j}] + \{y_{t+j} - \Pi_{P_p}[y_{t+j}]\} \tag{7}$$

where $\Pi_{P_p}[y_{t+j}]$ denotes the projection of the observations $y_{t+j}, j = -m, ..., m$, on $\mathbf{P}_p$. By orthogonality, for every $j \in T$

$$\hat{g}_t(0) = \hat{g}_t = \Pi_{P_p}[y_t] =< \Pi_{P_p}[y_{t+j}], R_p(j, 0) >=< y_{t+j}, R_p(j, 0) > . \quad (8)$$

Thus, $\hat{g}_t(0)$ is given by

$$\begin{aligned} \hat{g}_t(0) &= \int_{-m}^{m} \Pi_{P_p}[y_{t+j}] R_p(j, 0) f_0(j) dj & (9) \\ &= \int_{-m}^{m} y_{t+j} R_p(j, 0) f_0(j) dj & (10) \end{aligned}$$

where R_p is the reproducing kernel of the space $\mathbf{P}_p$. ■
Hence, the estimate $\hat{g}_t$ can be seen as a local weighted average of the observations, where the weights are derived by a kernel function K of order $p + 1$, where p is the degree of the fitted polynomial.

Definition 2 *Given $p \geq 2$, K_p is said to be of order p if*

$$\int_{-m}^{m} K_p(j) dj = 1, \quad \textit{and} \quad \int_{-m}^{m} j^i K_p(j) dj = 0, \quad (11)$$

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

The following result is fundamental (Berlinet 1993).

Corollary 3 *Kernels of order $(p+1)$, $p \geq 1$, can be written as products of the reproducing kernel $R_p(t, .)$ of the space $\mathbf{P}_p \subseteq L^2(f_0)$ and a density function f_0 with finite moments up to order $2p$. That is,*

$$K_{p+1}(t) = R_p(t, 0) f_0(t). \quad (12)$$

Remark 1 (Christoffel-Darboux Formula) *For any sequence $(P_i)_{0 \leq i \leq p}$ of $(p + 1)$ orthonormal polynomials in $L^2(f_0)$,*

$$R_p(t, 0) = \sum_{i=0}^{p} P_i(t) P_i(0). \quad (13)$$

Therefore, eq. (12) becomes

$$K_{p+1}(t) = \sum_{i=0}^{p} P_i(t)P_i(0)f_0(t). \tag{14}$$

Applied to real data, the kernel acts as a locally weighted average or linear filter that for each target point t gives the estimate

$$\hat{g}_t = \sum_{j=1}^{N} \kappa_{tj} y_j, \quad t = 1, 2, ..., N \tag{15}$$

where κ_{tj} denotes the weights to be applied to the observations y_j to get the estimate $\hat{g}_t$ for each point in time t. The weights κ_{tj} depend on the shape of the nonparametric estimator K_{p+1} and on the value of a bandwidth parameter b, such that

$$\kappa_{tj} = \frac{K_{p+1}\left(\frac{t-j}{b}\right)}{\sum_{i=1}^{N} K_{p+1}\left(\frac{t-i}{b}\right)}. \tag{16}$$

For any observed value y_t, a weighted average is computed and each weight is obtained as a function of the distance between the target point t and the $(t+j)$'s, $j = -m, ..., m$, close to the target point that belong to an interval whose amplitude is established by the bandwidth parameter b.

Several nonparametric estimators have been developed in the literature for time series smoothing. One approach is based on least squares and includes: (a) kernel estimators, (b) local polynomial fitting, and (c) graduation theory. A second approach corresponds to smoothing spline regression. Smoothing splines introduce a roughness penalty term in the minimization problem (4), searching for an optimal solution between both fitting and smoothing of the data. This would require an adapted RKHS. However, we show in our paper how an equivalent kernel representation for the smoothing spline can be derived in the polynomial space $\mathbf{P}_p$.
We provide a unified perspective for all of these different nonparametric estimators, according to which they are transformed into kernel functions and grouped into hierarchies with the following property: each hierarchy is identified by a density f_0 and contains estimators of order 2, 3, 4,... which are products of orthonormal polynomials with f_0. The den-

sity function represents the second order kernel within the hierarchy, and provides the "initial weighting shape" from which the higher order kernels inherit their properties. Therefore, if f_0 is optimal according to a specific smoothing criteria, each kernel of the hierarchy inherits the optimality property at its own order. Kernel functions can be compared by considering smoothers of different order within the same hierarchy as well as kernels of the same order, but belonging to different hierarchies. Filters of any length, including the infinite ones, can be derived in the RKHS framework. Therefore, for every estimator we will identify the density function f_0 and the corresponding reproducing kernel.

2. 1. Polynomial kernel regression

Local kernel regression deals with the problem of fitting a polynomial trend to the observations $y_{t+j}, j = -m, ..., m$, the value of the fitted function at $j = 0$ being taken as the smoothed observation $\hat{g}_t$. Representing the weights assigned to the residuals from the local polynomial regression by a symmetric and nonnegative function K_0, the problem is the minimization of

$$\sum_{i=1}^{N} K_0\left(\frac{t-i}{b}\right)\left[y_t - a_0 - a_1(t-i) - ... - a_p(t-i)^p\right]^2, \tag{17}$$

where the parameter b determines the bandwidth of the weighting function, since $K_0(z) = 0$, if $\mid z \mid\geq 1$.

Kernel estimators, local polynomial regression smoothers and filters derived in the graduation theory differ in the degree of the fitted polynomial, in the shape of the weighting function, and the neighborhood of observations taken into account. We shall discuss here the most often applied nonstationary mean smoothers, namely the Gaussian kernel, the Loess estimator developed by Cleveland (1979), and the Henderson filter. To derive the corresponding kernel hierarchy by means of the RKHS methodology, the density corresponding to K_0 and its orthonormal polynomials have to be calculated.

Kernel estimates are obtained by locally fitting linear polynomial trends where $p = 1$, weighting the observations in a neighborhood of the target point t using a probability density function. The Gaussian kernel family is already well known in the literature as Gram-Charlier hierarchy, studied among others by Deheveuls (1977), Wand and Schucany (1990), and Granovsky and Muller (1991). The associated density is the standard Normal

$$f_{0G}(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$$

and the standard Hermite polynomials are those orthogonal respect to this function. Starting from f_{0G}, the third order estimator within the hierarchy, denoted by K_{3G}, is easily determined by multiplying the density with a combination of Hermite polynomials up to degree two

$$K_{3G}(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \times \left(\frac{3-t^2}{2}\right). \tag{18}$$

In the literature, methods for building kernels of order higher than two start from second order smoothers, but no natural link is given between different order estimators. On the other hand, hierarchies make clear these relationships: two kernels in the same hierarchy differ by a product of f_0 and a linear combination of orthonormal polynomials.

The LOESS estimator, originally called LOWESS (LOcally Weighted Scatterplot Smoother), is based on nearest neighbor weights and is applied in an iterative manner for robustification. It consists of locally fitting polynomials of degree p by means of weighted least squares, where the weighting function proposed by Cleveland (1979) is the tricube one

$$K_{0T}(t) = \left(1 - |t|^3\right)^3 I_{[-1,1]}(t), \tag{19}$$

where $I_{[-1,1]}(t)$ denotes the indicator function. The second main parameter along with p is the width of a neighborhood or amount of observations around the estimated point. As it increases, the estimated trend becomes smoother.

Dagum and Bianconcini (2006) derived the Loess kernel hierarchy based on the tricube density

$$f_{0T}(t) = \frac{70}{81} \left(1 - |t|^3\right)^3 \mathrm{I}_{[-1,1]}(t), \tag{20}$$

where $\frac{70}{81}$ represents the integration constant of K_{0T} on the support $[-1, 1]$.

The third order kernel is given by

$$K_{3T}(t) = \frac{70}{81}\left(1 - |t|^3\right)^3 \left(\frac{539}{293} - \frac{3719}{638}t^2\right) \tag{21}$$

and obtained via multiplication of f_{0T} by a linear combination of its orthonormal polynomials up to degree two. These latter are derived using a determinantal expression based on the moments of the density function f_{0T}, as shown in Dagum and Bianconcini (2006).

Henderson's starting point was the requirement that the filter should reproduce a cubic polynomial trend without distortion. Henderson proved that three alternative smoothing criteria give the same weight diagram, as shown explicitly by Kenny and Durbin (1982) and Gray and Thomson (1996): (1) minimization of the variance of the third differences of the series defined by the application of the moving average; (2) minimization of the sum of squares of the third differences of the coefficients of the moving average formula; (3) fitting a cubic polynomial by weighted least squares, where the weights are chosen as to minimize the sum of squares of their third differences, and given by

$$K_{0H}(j) \propto \{(m+1)^2 - j^2\}\{(m+2)^2 - j^2\}\{(m+3)^2 - j^2\} \tag{22}$$

Dagum and Bianconcini (2007) showed that the weight diagram of the Henderson smoother can be well-reproduce by two different density functions and corresponding orthonormal polynomials. These functions are the exact density f_{0H} derived by the penalty function K_{0H} (Bianconcini 2006), and the biweight one

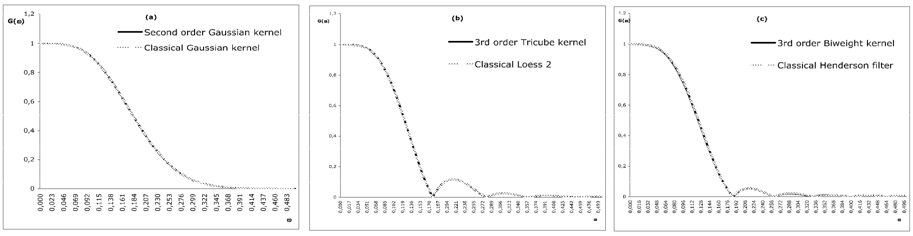
$$f_{0B}(t) = \frac{15}{16}(1-t^2)^2 I_{[-1,1]}(t). \tag{23}$$

These authors proved that the two density functions are very close one another, hence the former can be well-approximated by the latter. One of the main advantages of this approximation is that the biweight density function and the corresponding hierarchy do not need to be calculated any time that the length of the filter changes, as happens for f_{0H}. Furthermore, it belongs to the well-known Beta distribution family, and the corresponding orthonormal polynomials are the Jacobi ones, for which explicit expressions for computation are available. In the following we consider the biweight Henderson kernel hierarchy, whose third order estimator results

$$K_{3B}(t) = \frac{15}{16}(1 - |t|^2)^2 \times \left(\, \tfrac{7}{4} - \tfrac{21}{4}t^2 \, \right). \tag{24}$$

Figures 1(a), 1(b), and 1(c) show the gain functions of the classical and RKHS representations of the 13-term Gaussian kernel, Loess and Henderson symmetric filters. It is apparent the very close approximation of the RKHS representations with the classical ones.

Figure 1: Gain functions of the classical and RKHS representations of symmetric 13-term (a) Gaussian kernel, (b) Loess, and (c) Henderson filters



2. 2. Smoothing spline regression

A different formulation is required for smoothing splines which search for an optimal solution between both fitting and smoothing of the data, under the assumption that the signal follows locally a polynomial of degree p.

Extending previous work by Whittaker (1923) and Whittaker and Robinson (1924), Schoenberg (1964) showed that a natural smoothing spline estimator of order $p+1$ arises by minimizing the loss function

$$\min_{g \in W_2^{p+1}(T)} \|y_t - \hat{g}_t\|^2_{W_2^{p+1}} = \int_T (y_t - \hat{g}_t)^2 \, dt + \lambda \int_T \left(g^{(p+1)}(t) \right)^2 dt \tag{25}$$

where $\| \cdot \|_{W_2^{p+1}}$ denotes the W_2^{p+1}-norm, and the parameter λ regulates the balance between the goodness of fit and the smoothness of the curve. As $\lambda \to 0$ the solution approaches an interpolating spline, whereas as

$\lambda \to \infty$, the solution tends to the least squares line.
Eq. (25) equivalently defines the boundary value problem

$$\begin{aligned} \lambda g^{(2p+2)}(t) + g(t) &= y(t), \quad \forall t \in T \qquad (26) \\ g^{(k)}(0) &= g^{(k)}(1) = 0, \quad k = p+1, p+2, ..., 2p+1, \end{aligned}$$

which has a unique solution if the corresponding homogenous system only admits the null solution (see *e.g.* Gyorfy, Kohler, Krzyzak, and Walk 2002). In particular, it is determined by the unique Green's function $G_\lambda(t,s)$, such that

$$\hat{g}(t) = \int_T G_\lambda(t,s)y(s)ds =< G_\lambda(t,s), y(s) >_{L^2(T)} \qquad (27)$$

The derivation of $G_\lambda(t,s)$ corresponding to a smoothing spline of order $p+1$ requires the solution of a $(2p+2) \times (2p+2)$ system of linear equations for each value of λ. A simplification is provided by studying $G_\lambda(t,s)$ as the reproducing kernel $R_{p+1,\lambda}(t,s)$ of the Sobolev space $W_2^{p+1}(T)$, where T is an open subset of $\mathbb{R}$. When $T = \mathbb{R}$, the space $W_2^{p+1}(\mathbb{R})$ falls into the family of Beppo-Levi spaces described in Thomas-Agnan (1991). The corresponding reproducing kernel is translation invariant, and can be expressed in terms of $R_{p+1,1}$ as follows

$$R_{p+1,\lambda}(t) = \frac{1}{\lambda} R_{p+1,1}\left(\frac{t}{\lambda}\right). \qquad (28)$$

A general formula for $R_{p+1,1}$ is provided in Proposition 4 (see Thomas-Agnan 1991).

Proposition 4

$$R_{p+1,1}(t) = \sum_{k=0}^{p} \frac{\exp\left(-|t| e^{i\frac{\pi}{2p+2} + k\frac{\pi}{p+1} - \frac{\pi}{2}}\right)}{2(p+1)e^{(2p+1)\left(i\frac{\pi}{2p+2} + i\frac{k\pi}{p+1}\right)}}, \quad p = 1, 2, 3... \qquad (29)$$

Proposition 4 describes an equivalent kernel hierarchy for smoothing splines of order $p+1$, $p = 1, 2, ...$ It is identified by the standard Laplace density $R_{1,1}$, obtained as second order estimator by selecting $p = 1$. Higher order kernels are derived by multiplying $R_{2,1}$ by a combination of trigonometric polynomials which take into account for the roughness penalty term in (25). The third order smoother $R_{3,1}$ is familiar to the nonparametric statisticians since it is the asymptotically equivalent kernel to cubic smoothing spline derived by Silverman (1984). When the neighborhood of points for the estimation is small, as in most socioeconomic cases, eq. (29) gives a poor approximation of the classical cubic smoothing spline. The classical cubic smoothing spline is often represented by the *influential matrix* $\boldsymbol{A}(\lambda)$ (Wahba 1990, Green and Silverman 1994), which relates the estimated values $\hat{\boldsymbol{g}}$ to the observations $\boldsymbol{y}$ as follows

$$\hat{\boldsymbol{g}} = \boldsymbol{A}(\lambda)\boldsymbol{y} \tag{30}$$

where $\hat{\boldsymbol{g}}' = (\hat{g}_1, \hat{g}_2, ..., \hat{g}_N)$, and $\boldsymbol{y}' = (y_1, y_2, ..., y_N)$. Each $\hat{g}_t$ is a weighted linear combination of all the observed values y_t, with weights given by the elements of the t-th row of $\boldsymbol{A}(\lambda)$. In this study, we assume λ as given, and approximate each cubic spline predictor with time invariant linear filters. This enables us to analyze the properties of the estimators looking at the corresponding transfer functions.

To obtain a reproducing kernel representation of smoothing splines coherent with that derived for local kernel regression estimators, we have to find a density function f_0 according to which higher order kernels are obtained via multiplication of f_0 with corresponding orthonormal polynomials. This density has to be taken into account for the regularized term $\lambda \int_T (g^{p+1}(u))^2 du$ in eq. (25), in view of deriving the spline estimates as solution of the weighted least squares minimization problem (4).
Starting from the results of Proposition 4, we first consider the standard Laplace density multiplied by the corresponding orthonormal polynomials. To evaluate the goodness of this approximation we calculate the Euclidean distance Δ between the classical cubic smoothing spline CSS, and the corresponding third order kernel within each hierarchy K, both in terms of weights and gain functions

$$\Delta_{weights} = \sqrt{\sum_{j=-m}^{m} \mid \kappa_{CSS}(j) - \kappa_K(j) \mid^2} \qquad \Delta_{gain} = \sqrt{\sum_{\omega=0}^{1/2} \mid G_{CSS}(\omega) - G_K(\omega) \mid^2},$$

where ω denotes the frequency in cycles per unit of time and $G(\omega)$ is the gain of the filter. For illustrative purposes, we compute these measures for filter spans generally applied to monthly time series, that is 9, 13 and 23 terms, even if filter of any length can be considered. Table 1 shows that the third order standard Laplace kernel presents large discrepancies for each span indicating that asymptotically it does not approach to the CSS. On the other hand, the equivalent kernel representation derived in the Sobolev space $R_{3,1}$ provides the worst approximation for the shortest filter length, whereas it has a better performance as the span increases. It follows that, given the poor performance of the standard Laplace estimators, another density function need to be identified.

To be coherent with the results of Proposition 4, we consider the log $F_{2m_1,2m_2}$ distribution family, introduced by Prentice (1976), that presents the standard Normal $(m_1, m_2 \rightarrow 0)$ and Laplace $(m_1, m_2 \rightarrow \infty)$ as limiting cases. A lot of densities belongs to this class of distributions, among other the logistic $(m_1 = m_2 = 1)$, and the exponential $(m_1 \neq 0, m_2 \rightarrow 0$. We concentrate on the former given its strong connection with the standard Laplace and other widely applied density functions, as recently shown by Lin and Hu (2007). These authors modified and extended previous work by Mudholkar and George (1978), providing a characterization of the logistic density in terms of sample median and Laplace distribution, as well as in terms of the smallest order statistics and exponential density (see also Jones 2006). Furthermore, George and Ojo (1980), and George, El-Saidi, and Singh (1986) studied the logistic distribution as approximation of Student's t functions. The density is defined as follows

$$f_{0L}(t) = \left(\frac{1}{4\beta}\right)^{-1} sech^2\left[\frac{1}{2}\left(\frac{t-\alpha}{\beta}\right)\right] \quad t \in (-\infty, \infty) \tag{31}$$

where α is the mean, set equal to 0, and $\beta > 0$ is the dispersion parameter. Table 1 shows the results for the third order kernel within the hierarchy derived by f_{0L}, where the dispersion parameter has been set equal to 0.2. It is given by

$$K_{3L}(t) = \frac{5}{4} sech^2\left(\frac{5}{2}t\right)\left(\frac{21}{16} - \frac{2085}{878}t^2\right), \tag{32}$$

where $sech$ is the hyperbolic secant function. This estimator really closely approximates the CSS, with a superior performance to the other filters for all the spans.

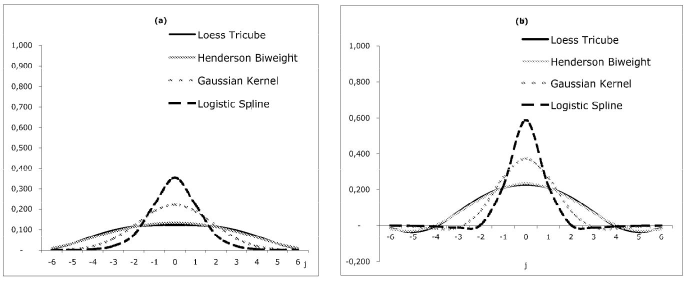
The logistic kernel representation of smoothing splines enables to compare directly all the hierarchies we have derived. Figure 2 (a) shows the density functions or second order kernels within each family, whereas Figure 2 (b) illustrates the third order smoothers. For space reasons, we consider the 13-term filters, since it is the most often applied length in trend-cycle estimation of monthly time series, but similar results are derived for 9-

Table 1: 2-norm distances between classical and reproducing kernel smoothing splines

3rd order kernel	**Filter length**					
	9		*13*		*23*	
	weights	gain function	weights	gain function	weights	gain function
Sobolev space $R_{3,1}$	0.144	3.213	0.428	1.925	0.482	1.709
Standard Laplace	0.049	1.088	0.608	3.863	0.583	2.916
Logistic $\beta = 0.2$	0.021	0.480	0.271	0.588	0.260	0.471

and 23-term smoothers.

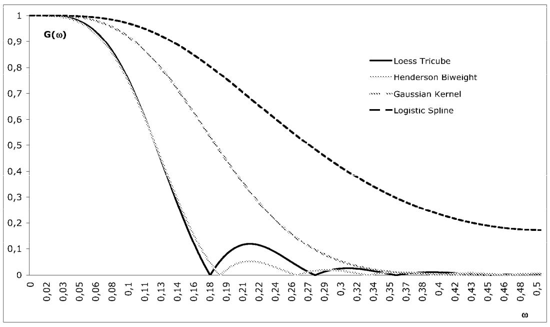
Figure 2: 13-term (a) second and (b) third order kernels



These hierarchies reproduce and describe several temporal dynamics by estimating polynomial trends of different degrees, that solve several minimization problem. This is shown in Figure 3, where we illustrate the gain functions of 13-term third order kernels.

Loess and Henderson kernels present similar properties in terms of trend-cycle estimation. They eliminate a large amount of noise and pass all the power associated to the signal frequency band, $0 < \omega \leq 0.06$. However, they do not suppress power at the frequency $\omega = 0.10$, corresponding to cycles of 10 months, often interpreted as false turning points. On the other hand, the third order Gaussian and spline kernels, when reduced to 13 terms, loose part of their optimality prediction properties, leaving untouched the signal but passing a lot of noise including a large number of unwanted ripples.

Figure 3: Gain functions of symmetric 13-term third order kernels



3. Boundary behavior

The kernels derived by means of the RKHS methodology provide a new and unified way to represent nonparametric estimators used currently. The third order kernels in the tricube, biweight, and logistic hierarchies are equivalent kernels of the classical Loess of degree 2 (Loess 2), Henderson filter, and cubic smoothing spline, respectively. No comparisons can be made for the third order Gaussian estimator which is already a kernel function, and for which no counterpart exists in the literature.

The reproducing kernel representation has important consequences in the derivation of the corresponding asymmetric smoothers, which are of crucial importance in current analysis where the aim is to obtain the estimate of the nonstationary mean for the most recent observations. In the RKHS approach, the asymmetric smoothers are derived by adapting the kernel functions to the length of the last m asymmetric filters, such that

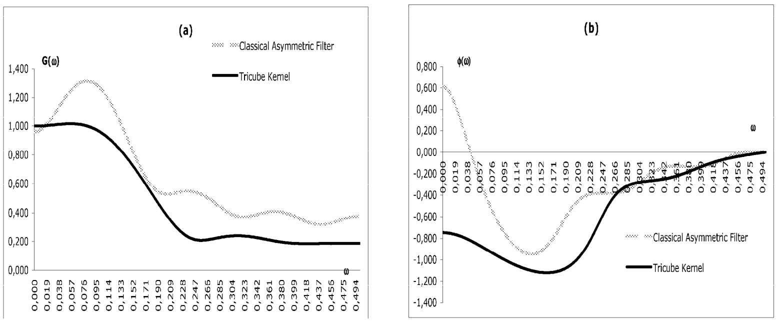
$$\kappa_{t,j} = \frac{K_3\left(\frac{j}{b}\right)}{\sum_{i=-m}^{q} K_3\left(\frac{i}{b}\right)} \qquad j = -m, ..., q$$

where j denotes the distance to the target point t ($t = N - m + 1, ..., N$), b is the bandwidth parameter selected in view of ensuring a symmetric filter of length $2m + 1$, and $m + q + 1$ is the asymmetric filter length.

These boundary kernels only satisfy the condition $\int_{-m}^{q} K_3(t)dt = 1$, meaning that the estimator will reproduce without distortion only a constant on the asymmetric support. Given the small number of points generally available in the boundaries, this condition can be considered sufficient in view of obtaining a better performance of the kernel in terms of trade-off between fitting and smoothing. Our main object is to analyze the goodness of the reproducing kernel representations versus the classical last point asymmetric filters.

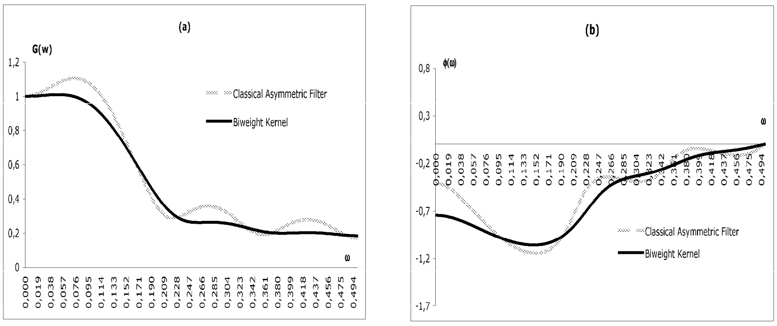
Cleveland (1979) showed that, in the middle of the series, Loess acts as a symmetric moving average with window length $2m + 1$. At the end of the series, its window length remains $2m + 1$, rather than decreasing to $m + 1$ as in the case of the most widely applied asymmetric concurrent trend-cycle estimators. As discussed by Gray and Thomson (1996b), this implies a heavier than expected smoothing at the ends of the series respect to the middle, and represents a drawback, particularly for economic time series where turning points are important to identify. As shown in Figure 4 (a), the last point Loess asymmetric kernel exhibits a gain function with better properties of signal passing and noise suppression relative to the classical one. This implies smaller filter revisions as new data are added to the series. The phase shifts for both filters (Figure 4 (b)) are smaller than one month in the signal frequency band usually defined as $0 < \omega \leq 0.055$.

Figure 4: (a) Gain and (b) phaseshift functions of the asymmetric (end point) weights of the third order tricube kernel and the classical Loess 2



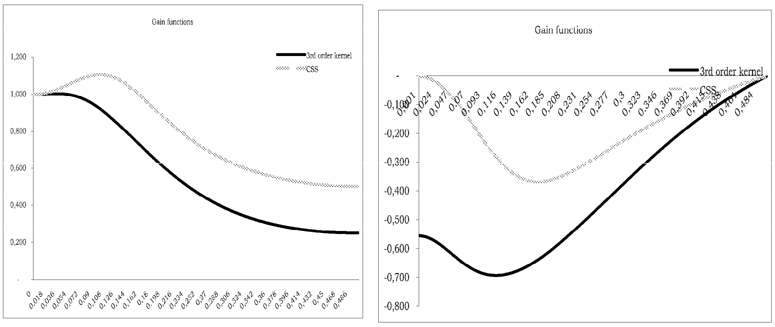
Similar conclusions can be derived for the last point asymmetric Henderson filter, as developed by Musgrave (1964). They are based on the minimization of the mean squared revision between the final estimates (obtained by the application of the symmetric filter) and the preliminary ones (obtained by the application of an asymmetric filter) subject to the constraint that the sum of the weights is equal to one (see *e.g.* Doherty 1992). The assumption made is that at the end of the series, the seasonally adjusted values follow a linear trend-cycle plus a purely random irregular ε_t, such that $\varepsilon_t \sim IID(0, \sigma^2)$.

Figure 5: (a) Gain and (b) phaseshift functions of the asymmetric (end point) weights of the third order biweight kernel and the classical Henderson filter



The asymmetric filters for the natural cubic smoothing splines are obtained by adding additional constraints, ensuring that the function is of degree 1 beyond the boundary knots. In this study, the asymmetric classical splines are obtained by fixing the λ parameter to have a $2m+1$-term symmetric filter, and then selecting the last m rows of the influential matrix $\boldsymbol{A}(\lambda)$. We illustrate the results for the last point asymmetric weights corresponding to a 23 term symmetric filter because the spline gives very poor results for short lengths such as 13 or less. Figure 6(a) shows that the asymmetric kernel exhibits a gain function with better properties of signal passing and noise suppression, without implying larger phase shifts. These latter are smaller than one month for both filters (see Figure 6(b)).

Figure 6: Gain Functions of the Asymmetric (End Point) Weights of the Third Order Spline Kernel and the CSS Smoother



3. 1. Empirical evaluation

This section examines how the third order logistic kernel performs on real data in comparison with the classical cubic smoothing spline. For empirical applications of the Loess and Henderson kernels, we refer the reader to Dagum and Bianconcini (2006, 2007).

We apply the last point filters to a set of 50 time series taken from the Hydman's time series library (http://www-personal.buseco.monash.edu.au/ hyndman/TSDL/). These series are related to different fields (finance, labor market, production, sales, transports, meteorology, hydrology, physics, and health), and are all seasonally adjusted. The periods selected vary to sufficiently cover the various lengths published for these series.

We want to study how the kernel and splines respond to the variability of the data. For each series, we apply the Generalized Cross Validation (GCV) criteria (Craven and Wahba 1979) to estimate the λ parameter, and consequently the length of the filters to be applied by considering the number of non-null elements in a central row of the matrix $\boldsymbol{A}(\lambda)$. The kernel estimator of the same length is then calculated. The smoothing 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 output. In our sample, λ ranges from a minimum of 0.013, at which corresponds a filter length equal to 7 terms, to a maximum of 15, which corresponds to a 43-term smoother. This enables us to analyze the pattern of the two estimators on series characterized by

different degrees of variability.
The comparison is based on the relative filter revisions between the final symmetric filter F and the last point asymmetric filter P, that is,

$$R_t = \frac{F_t - P_t}{F_t}, \quad t = 1, 2, ..., N \tag{33}$$

For each series, we calculate the ratio $\frac{MSE(R^K)}{MSE(R^C)}$ between the Mean Square Error (MSE) of the revisions corresponding to the third order logistic kernel R^K and to the classical cubic spline R^C.

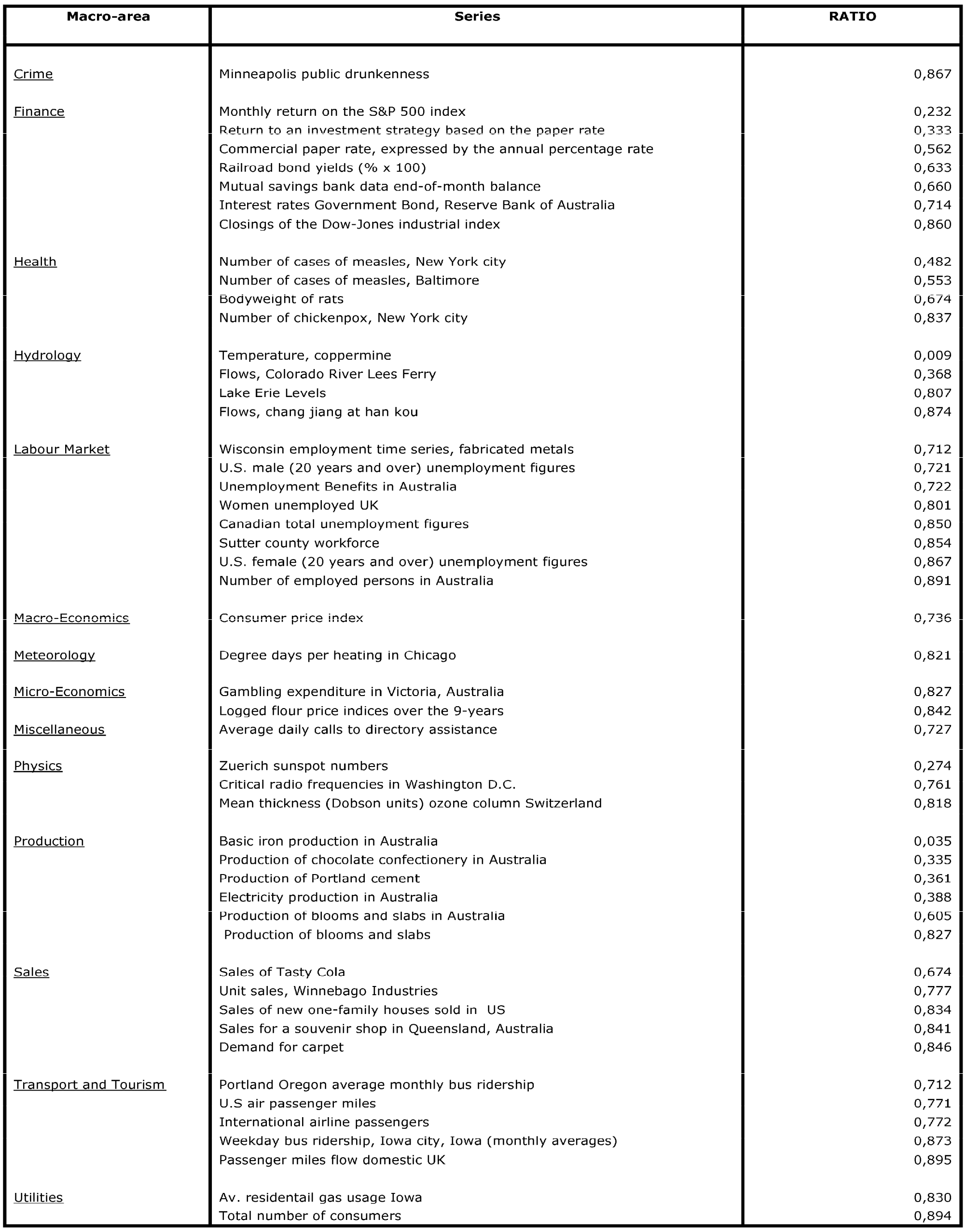
The results illustrated in Figure 7 indicate that the ratio is always smaller than one, showing that the logistic kernel last point predictor introduces smaller revisions than the classical one. This implies that the kernel estimates will be more reliable and efficient than the ones obtained by the application of the classical cubic spline. In particular, in the 38% of the sample the ratio is less than 0.7 and in general it is never greater than 0.895.

4. Concluding remarks

We have introduced a kernel representation of several nonstationary mean predictors by means of the Reproducing Kernel Hilbert Space (RKHS) methodology. This approach encompasses, from a probabilistic point of view, several nonparametric linear estimators developed in the literature for smoothing functional data. In particular, we have shown how an equivalent kernel representation for the smoothing spline can be derived in the polynomial space $\mathbf{P}_p$. Hence, we provide a unified perspective, according to which every nonparametric estimator can be transformed into a kernel function and grouped into an hierarchy with the following property: the hierarchy is identified by a density f_0 and contains estimators of order 2, 3, 4,... which are products of orthonormal polynomials with f_0.

Comparisons can be performed between smoothers of different order within the same hierarchy as well as kernel of the same order but belonging to different hierarchies. The asymmetric weights of the kernels are derived by adapting the third order functions to the length of the last

Figure 7: MSE revision ratio between classical and kernel last point splines



Macro-area	Series	RATIO
Crime	Minneapolis public drunkenness	0,867
Finance	Monthly return on the S&P 500 index	0,232
	Return to an investment strategy based on the paper rate	0,333
	Commercial paper rate, expressed by the annual percentage rate	0,562
	Railroad bond yields (% x 100)	0,633
	Mutual savings bank data end-of-month balance	0,660
	Interest rates Government Bond, Reserve Bank of Australia	0,714
	Closings of the Dow-Jones industrial index	0,860
Health	Number of cases of measles, New York city	0,482
	Number of cases of measles, Baltimore	0,553
	Bodyweight of rats	0,674
	Number of chickenpox, New York city	0,837
Hydrology	Temperature, coppermine	0,009
	Flows, Colorado River Lees Ferry	0,368
	Lake Erie Levels	0,807
	Flows, chang jiang at han kou	0,874
Labour Market	Wisconsin employment time series, fabricated metals	0,712
	U.S. male (20 years and over) unemployment figures	0,721
	Unemployment Benefits in Australia	0,722
	Women unemployed UK	0,801
	Canadian total unemployment figures	0,850
	Sutter county workforce	0,854
	U.S. female (20 years and over) unemployment figures	0,867
	Number of employed persons in Australia	0,891
Macro-Economics	Consumer price index	0,736
Meteorology	Degree days per heating in Chicago	0,821
Micro-Economics	Gambling expenditure in Victoria, Australia	0,827
	Logged flour price indices over the 9-years	0,842
Miscellaneous	Average daily calls to directory assistance	0,727
Physics	Zuerich sunspot numbers	0,274
	Critical radio frequencies in Washington D.C.	0,761
	Mean thickness (Dobson units) ozone column Switzerland	0,818
Production	Basic iron production in Australia	0,035
	Production of chocolate confectionery in Australia	0,335
	Production of Portland cement	0,361
	Electricity production in Australia	0,388
	Production of blooms and slabs in Australia	0,605
	Production of blooms and slabs	0,827
Sales	Sales of Tasty Cola	0,674
	Unit sales, Winnebago Industries	0,777
	Sales of new one-family houses sold in US	0,834
	Sales for a souvenir shop in Queensland, Australia	0,841
	Demand for carpet	0,846
Transport and Tourism	Portland Oregon average monthly bus ridership	0,712
	U.S air passenger miles	0,771
	International airline passengers	0,772
	Weekday bus ridership, Iowa city, Iowa (monthly averages)	0,873
	Passenger miles flow domestic UK	0,895
Utilities	Av. residentail gas usage Iowa	0,830
	Total number of consumers	0,894

asymmetric filters. We showed the performance for the classical cubic smoothing splines and the corresponding reproducing kernel. Applied to a set of 50 real series, we computed a measure of revision for the last point filters. They show how the revisions are systematically smaller for the kernel representation than for the classical cubic spline. These results conform to their respective gain functions.

References

Adams, R. (1975): *Sobolev Spaces.* Academic Press, Inc, Harcourt Brace Jovanovich Publishers.

Aronszajn, N. (1950): "Theory of Reproducing Kernels," *Transaction of the AMS*, 68, 337–404.

Berlinet, A. (1993): "Hierarchies of Higher Order Kernels," *Probability Theory and Related Fields*, 94, 489–504.

Berlinet, A., and C. Thomas-Agnan (2003): *Reproducing Kernel Hilber Spaces in Probability and Statistics.* Kluwer Academic Publishers.

Bianconcini, S. (2006): *Trend-Cycle Estimation in Reproducing Kernel Hilbert Spaces.* Ph.D. Thesis, Department of Statistics, University of Bologna.

Cleveland, W. (1979): "Robust Locally Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

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

Cristianini, N., and J. Shawe-Taylor (2000): *An Introduction To Support Vector Machines And Other Kernel-Based Learning Methods.* Cambridge University Press.

Dagum, E., and S. Bianconcini (2006): "Local Polynomial Trend-Cycle Predictors in Reproducing Kernel Hilbert Spaces for Current Economic Analysis," *Anales de Economia Aplicada*, pp. 1–22.

——— (2007): "The Henderson Smoother in Reproducing Kernel Hilbert Space," *Journal of Business and Economic Statistics*, forthcoming.

De Boor, C., and R. Lynch (1966): "On Splines and Their Minimum Properties," *Journal of Math. Mech.*, 15, 953–969.

Deheveuls, P. (1977): "Estimation Nonparametrique de la Densite par Histogrammes Generalises," *Reveu de Statistique Applique*, 25, 5–42.

Doherty, M. (1992): "The Surrogate Henderson Filters in X-11," *Working Paper, Statistics New Zealand, Wellington, New Zealand.*

Evgeniou, T., M. Pontil, and T. Poggio (2000): "Regularization Networks And Support Vector Machines," *Advanced in Computational Mathematics*, 13, 1–50.

George, E., and M. Ojo (1980): "On a generalization of the logistic distribution," *Annals of the Institute of Statistical Mathematics*, 32.

George, E. O., M. El-Saidi, and K. Singh (1986): "A generalized logistic approximation of the Student t distribution," *Communications in Statistics B Simulation Computing*, 15(1), 261277.

Granovsky, B., and H. Muller (1991): "Optimizing Kernel Method: a Unifying Variational Principle," *International Statistical Review*, 59, 373–388.

Gray, A., and P. Thomson (1996): "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.

——— (1996b): "On a family of moving-average trend filters for the ends of series," *Proocedings of the Business and Bconomic Statistics Section*, American Statistical Association Annual Meeting, Chicago.

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

Gyorfy, L., M. Kohler, A. Krzyzak, and A. Walk (2002): *A Distribution-Free Theory of Nonparametric Regression.* New York: Springer-Verlag.

Jones, M. (2006): "The logistic and the log F distributions," *Handbook of logistic distribution, N. Balakrishan ed.*, 2nd ed., New York:Dekker.

Kenny, P., and J. Durbin (1982): "Local Trend Estimation and Seasonal Adjustment of Economic and Social Time Series," *Journal of the Royal Statistical Society B*, 145, 1–41.

Kimeldorf, G., and G. Wahba (1971): "Splines Functions and Stochastic Processes," *Sankhya, Series A*, 32, 173–180.

Lin, G., and C. Hu (2007): "On the characterization of the logistic distribution," *Journal of Statistical Planning and Inference*, in press.

Mudholkar, G., and E. George (1978): "A remark on the shape of the logistic distribution," *Biometrika*, 65.

Musgrave, J. (1964): "A Set of End Weights to End all End Weights, Working paper," *US Bureau of the Census, Washington.*

Parzen, E. (1959): *Statistical Inference on Time Series by Hilbert Space Methods.* Technical Report No. 53, Statistics Department, Stanford University, Stanford, CA.

Pearce, N., and M. Wand (2006): "Penalized Splines And Reproducing Kernel Methods," *The American Statistician*, 60(3).

Prentice, R. (1976): "A generalization of probit and logit methods for dose response curves," *Biometrics*, 32, 761768.

Priestley, M. (1981): *Spectral Analysis and Time Series.* Probability and Mathematical Statistics, Academic Press.

Schoenberg, I. (1964): "Monosplines and Quadrature Formulae," *in Theory and Applications of Spline Functions*, ed. T. Greville, Madison, WI: University of Wisconsin Press.

Silverman, B. (1984): "Spline Smoothing: The Equivalent Kernel Method," *Annals of Statistics*, 12, 898–916.

Thomas-Agnan, C. (1991): "Splines Functions and Stochastic Filtering," *Annals of Statistics*, pp. 1512–1527.

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

——— (1999): *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.

WAND, M., AND W. SCHUCANY (1990): "Gaussian-Based Kernels," *Canadian Journal of Statistics*, 18, 197–204.

WHITTAKER, E. (1923): "On a New Method of Graduation," *Proceedings of the Edinburgh Mathematical Association*, 78, 81–89.

WHITTAKER, E., AND G. ROBINSON (1924): *Calculus of Observations: a Treasure on Numerical Calculations.* Blackie and Son, London.