
http://researchrepository.murdoch.edu.au/23145/

Copyright: © 2014 Elsevier B.V.

It is posted here for your personal use. No further distribution is permitted.
Performance criteria and discrimination of extreme undersmoothing in nonparametric regression

Mark A. Lukas

Mathematics and Statistics, Murdoch University, South Street, Murdoch, WA 6150, Australia

1. Introduction

A fundamental problem of data analysis is the estimation of a smooth function \( f: [a, b] \rightarrow \mathbb{R} \) from data \( y_i \) that follow the model:

\[
y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n,
\]

where the design points \( x_i \) satisfy \( a \leq x_1 < x_2 < \cdots < x_n \leq b \) and the \( \epsilon_i \) are i.i.d. \( N(0, \sigma^2) \) random errors. If a functional form for \( f \) is not known, it is appropriate to use a nonparametric regression method. Several such methods have been proposed and studied, including kernel smoothing, local polynomial smoothing, series estimators, regression splines and smoothing splines; see Eubank (1988), Eubank (1999), Green and Silverman (1994), Hart (1997), Wahba (1990), and Wand and Jones (1995) for discussion of these methods.

We will mainly consider smoothing spline estimators. The smoothing spline of degree \( 2m - 1 \) is defined as the function \( f_\lambda \) that minimizes

\[
n^{-1} \sum_{i=1}^{n} (y_i - h(x_i))^2 + \lambda \int_a^b (h^{(m)}(x))^2 \, dx
\]

The prediction error (average squared error) is the most commonly used performance criterion for the assessment of nonparametric regression estimators. However, there has been little investigation of the properties of the criterion itself. This paper shows that in certain situations the prediction error can be very misleading because it fails to discriminate an extreme undersmoothed estimate from a good estimate. For spline smoothing, we show, using asymptotic analysis and simulations, that there is poor discrimination of extreme undersmoothing in the following situations: small sample size or small error variance or a function with high curvature. To overcome this problem, we propose using the Sobolev error criterion. For spline smoothing, it is shown asymptotically and by simulations that the Sobolev error is significantly better than the prediction error in discriminating extreme undersmoothing. Similar results hold for other nonparametric regression estimators and for multivariate smoothing. For thin-plate smoothing splines, the prediction error’s poor discrimination of extreme undersmoothing becomes significantly worse with increasing dimension.
in the Sobolev space $W^{m,2}[a,b]$. Here $\lambda > 0$ is the smoothing parameter; a larger value of $\lambda$ yields a smoother estimate. The most popular smoothing splines are cubic smoothing splines, for which $m=2$.

To assess the quality of a spline estimator $f_\lambda$ and hence define the “best” choice of $\lambda$, one needs a suitable performance criterion (optimality criterion). The most commonly used performance measure is the prediction error (average squared error)

$$T(\lambda) = n^{-1} \sum_{i=1}^n (f(x_i) - f_\lambda(x_i))^2,$$

(1)

together with the prediction risk $ET(\lambda)$. A related measure is the (squared) $L_2$ error (integrated squared error)

$$l(\lambda) = \int_a^b (f(x) - f_\lambda(x))^2 \, dx$$

(2)

together with the associated risk $El(\lambda)$. Note that, for equally spaced points $x_i$, $T(\lambda)$ in (1) is a standard quadrature approximation of $l(\lambda)$ in (2), and so they can be expected to be fairly close.

Most of the extensive literature on the performance of smoothing spline estimators, including various parameter selection methods, uses the prediction error or $L_2$ error (or risk) as the only performance measure. This includes both asymptotic analyses as well as simulation studies; see, e.g., Li (1986) and Lee (2003). However, there has been much attention given to the question of whether the prediction error $L_2$ error is actually the most suitable performance criterion. One of the aims of this paper is to shed light on this question.

It was shown in Marron and Tsybakov (1995) that the use of the $L_2$ error (as well as the $L_1$ and $L_\infty$ errors) in nonparametric regression does not properly match what the eye can see. The authors argue that a better qualitative assessment of an estimated curve, when compared to a true curve, is obtained by considering the distance between the graphs of the curves (as sets of points in $\mathbb{R}^2$), rather than the vertical distance associated with a function space norm, and they propose several visual error criteria in this way. However, as noted in Marron and Tsybakov (1995), this approach is not suitable if the regression is to be used for prediction purposes.

In this paper, we take the more general view that the performance criterion used in nonparametric regression should make the regression suitable for estimation of $f(x)$ on the whole interval $[a,b]$ and for associated prediction, while, at the same time, the criterion should be consistent with our visual perception of the quality of the estimator.

We will focus on a specific weakness of the prediction error as a performance measure. Clearly, the prediction error is a discrete measure in that it depends on $f - f_\lambda$ only through its values at the discrete set of points $x_i$, but this is not its main weakness. The main problem with the prediction error (as well as the $L_2, L_1$ and $L_\infty$ errors) is that it is insensitive to deviations in the derivative and (linearized) curvature (and higher derivatives) of the spline $f_\lambda$. This situation is inconsistent with our visual perception of the quality of a fitted curve. Because $f$ is smooth, large deviations in the derivative or curvature of $f_\lambda$ from those of $f$ can be easily identified visually from the graphs of $f$ and $f_\lambda$.

To make this more concrete, suppose that the error is approximately given by $f(x) - f_\lambda(x) \approx c \cos kx$ for $x \in [0,1]$, where $c > 0$ is small and $k$ is a large integer satisfying $k \approx c^{-1}$. Assume that the $x_i$ are equally spaced in $[0,1]$ and $n$ is sufficiently large so that $T(\lambda) \approx l(\lambda)$. Then $T(\lambda) \approx l(\lambda) \approx c^2/2$, which is small even though $f - f_\lambda$ is very wiggly, with large integrated squared (linearized) curvature satisfying

$$\int_0^1 (f(x) - f_\lambda(x))^2 \, dx \geq K \approx k^2 \pi^2/2.$$

Therefore, the prediction error (and $L_2$ error) fails to detect that $f - f_\lambda$ is very wiggly, and yet this would be easily seen from its graph.

Clearly, the same conclusion holds if $f - f_\lambda \approx h$, where $h(x)$ is a short finite sum of the form

$$h(x) = \sum c_j \cos k_j \pi x + d_j \sin l_j \pi x$$

and $c_j > 0$ and $d_j > 0$ are small, with $k_j \geq c_j^{-1}$ and $l_j \geq d_j^{-1}$. Note that this assumption for $f - f_\lambda$ is quite plausible if $\lambda$ is very small, because, in this situation, $f - f_\lambda$ would be close to the spline that interpolates the errors $e_i$ at the design points $x_i$. It is likely that this error vector is of high frequency, measured, say, by the number of sign changes. Then, using the Demmler–Reinsch basis for the space of smoothing splines, $f - f_\lambda$ would be approximately equal to a finite sum involving the high frequency basis functions, and it is known that, for equally spaced $x_i$, the basis functions are approximately equal to trigonometric functions; see Culpin (1986), Eubank (1988, Sect. 5.3).

The above reasoning indicates that the prediction error may fail to detect when a spline estimate $f_\lambda$ is very wiggly. It will be shown that this does actually occur in practice. Moreover, the prediction error can be a misleading performance measure because it can fail to discriminate an extreme undersmoothed spline estimate from a good estimate. Section 2.1 presents simulation results that illustrate this property and we identify the situations where it can occur; these are: small sample size $n$ or small error variance $\sigma^2$ or a function $f$ with high curvature.

In Section 2.2, we define a measure of the prediction error’s capacity to discriminate extreme undersmoothing. This is defined as the probability that the value of the prediction error for the most extreme undersmoothed spline estimate, i.e., the interpolating spline, is relatively close to the minimum prediction risk (within a factor $D$). The larger the value of this probability, the more likely it is that the prediction error will fail to discriminate extreme undersmoothing. We investigate
the asymptotic behavior of this probability and show that, under suitable assumptions which are consistent with \( f \) belonging to a certain subset of \( W^{q,2}[a,b] \) for some \( q \geq 1 \), the probability goes to 0 quickly as \( n \to \infty \). However, the rate depends on \( a \) and a parameter that measures the roughness of \( f \); the rate is slower for a smaller value of \( a \) or a function \( f \) with higher integrated squared curvature. Simulation results in Section 4 show that this behavior is evident in practice and that, for realistic values of \( a \), the prediction error displays poor discrimination of extreme undersmoothing for a range of small values of \( n \).

To overcome this problem, it is reasonable to consider measuring the error \( f - f_\lambda \) with a norm that is sensitive to deviations in the derivative and curvature (and possibly higher derivatives), as well as the function. Assume that \( f \in W^{m,2}[a,b] \). Clearly, \( f_\lambda \in W \) by definition. (In fact \( f_\lambda \) is more regular than this if \( m \geq 2 \), since \( f_\lambda \in C^{2m-2}[a,b] \) Waheb, 1990.) Therefore, it is reasonable to measure the error \( f - f_\lambda \) using a Sobolev norm of order \( m \). We will consider a Sobolev norm involving only the highest \( m \)th derivative, and define the associated (squared) Sobolev error by

\[
W(\lambda) = \|f - f_\lambda\|^2_W = \int_a^b (f(x) - f_\lambda(x))^2 \, dG + \kappa \int_a^b (f^{(m)}(x) - f_\lambda^{(m)}(x))^2 \, dx,
\]

where \( G \) is the limiting cumulative distribution function for the design points \( (e.g., G(x) = (x-a)/(b-a) \) for equally spaced points) and \( \kappa = (b-a)^{m-1} \). This choice of the constant \( \kappa \) makes \( W(\lambda) \) independent of the length of the interval \( [a,b] \) under horizontal scaling of \( f - f_\lambda \) (Lukas et al., 2014). Although this Sobolev error involves only deviations in \( f_\lambda \) and its \( m \)th derivative, it is automatically sensitive to deviations in the derivatives of orders \( 1, \ldots, m-1 \) as well (Lukas et al., 2012).

As seen from the following, the Sobolev error criterion is suitable for estimation of \( f(x) \). Clearly, if \( W(\lambda) \) is small, then automatically the \( L_2(G) \) error defined by the first term on the right-hand side of (3) is also small. The Sobolev error can also be used to bound the error \( f - f_\lambda \) in the maximum norm (which is true regardless of the class of estimators used, so long as they belong to \( W \)). This follows (under suitable assumptions on \( G \)) from the Sobolev imbedding theorem (Adams, 1975), which implies that there is a constant \( C \) such that \( \max_x \lambda \|h(x)\| \leq C \|h\|_W \) for any \( h \in W \), in particular for \( h = f - f_\lambda \).

It follows that, if \( W(\lambda) \) is small enough, then \( f_\lambda \) is close to \( f \) in a uniform sense. Furthermore, the Sobolev criterion simultaneously facilitates uniform estimation of \( f \) and of its (classical) derivatives \( f^{(k)} \) of all orders \( k = 1, \ldots, m-1 \). This follows from the Sobolev inequality (Adams, 1975) \( \max_x \lambda \|h^{(k)}(x)\| \leq C \|h\|_W \) for some constant \( C \) and \( k = 1, \ldots, m-1 \), with \( h = f - f_\lambda \).

To see how the Sobolev error responds to high frequency errors in \( f_\lambda \), let \( [a,b] = [0,1] \), \( G(x) = x \) and \( f(x) - f_\lambda(x) = c \cos kx \) for small \( c > 0 \) and large integer \( k \geq c^{-1} \) as above. Then

\[
W(\lambda) \geq c^2 \frac{1}{2} + k^{2m-1} \frac{1}{2m^2} \]

which is certainly not small (and it is larger for larger \( m \)). Therefore, like our eyes, the Sobolev error can detect the high frequency errors.

In Section 3, we investigate how well the Sobolev error can discriminate extreme undersmoothed spline estimates from good estimates. A measure of the Sobolev error’s capacity to discriminate extreme undersmoothing is defined in a similar way to that for the prediction error. It is shown that, under suitable assumptions, this measure goes to 0 as \( n \to \infty \) at a rate that is significantly faster than for the corresponding measure for the prediction error. This means that, for all sufficiently large \( n \), the Sobolev error is better than the prediction error in discriminating extreme undersmoothing. Simulation results in Section 4 show that this is also true for small \( n \).

In our treatment of the Sobolev error, the (minimal) order of smoothness of \( f \) is taken to be equal to the order \( m \) of the roughness penalty for the smoothing spline. (This is very reasonable in the most common situation where \( m = 2 \).) One can imagine the general situation where all of these orders are possibly different, with say \( f \in W^{q,2}[a,b] \) and the Sobolev error of order \( r \leq \min(q,m) \). It is conjectured that also in this general situation, under suitable conditions, the Sobolev error is better than the prediction error in discriminating extreme undersmoothing.

In Section 5 it is shown that similar results about the prediction and Sobolev errors hold for other nonparametric regression estimators, in particular the Gasser–Müller kernel estimator. Section 6 considers multivariate smoothing by thin-plate smoothing splines. Theoretical and simulation results show that the prediction error has significantly worse discrimination of extreme undersmoothing compared to the univariate situation. The Sobolev error provides better discrimination of extreme undersmoothing.

The results of this paper have important implications for theory and simulation studies of parameter selection methods. In most simulation studies to date, the selection methods are assessed using the inefficiency with respect to the prediction error; see, e.g., Lee (2003) and Chen and Huang (2011). However, this approach can give an inaccurate assessment of a selection method for small values of \( n \) or \( a \), or for functions with high curvature. This is because the method might often select a value of \( \lambda \) that is far too small and yet this would not be adequately reflected in the corresponding inefficiency for the prediction error.

The Sobolev error does not have this problem and it makes a very reasonable performance criterion. Therefore, it is appropriate to develop and investigate parameter selection methods with this criterion in mind. For spline smoothing, it was shown in Lukas et al. (2012) and Lukas et al. (2014) using asymptotic analysis and simulations that, unlike generalized cross-validation (GCV), the robust GCV method and the modified GCV method perform well with respect to the Sobolev criterion.

Please cite this article as: Lukas, M.A., Performance criteria and discrimination of extreme undersmoothing in nonparametric regression. Journal of Statistical Planning and Inference (2014), http://dx.doi.org/10.1016/j.jspi.2014.05.006
2. The prediction error and discrimination of extreme undersmoothing

2.1. Illustrations of the problem

The most extreme undersmoothed spline estimate is \( f_{0^-} \), corresponding to \( \lambda \rightarrow 0^+ \). It is well known (Wahba, 1990) that \( f_{0^-} \) is the natural spline interpolating the noisy data, which is defined as the function in \( W^{m,2}(a,b) \) that minimizes \( \int_a^b (h^{(m)}(x))^2 \, dx \) subject to \( h(x_i) = y_i, i = 1, \ldots, n \). It can be expected that, for almost all realizations of the error, \( f_{0^-} \) would be a poor estimate of the underlying smooth function.

To illustrate this and see how the prediction error responds to undersmoothed spline estimates, we will consider the two functions \( f(x) = \sin(2\pi x)/(x+1/2) \) and \( f(x) = \sin(4\pi x)/(x+1/2) \) on \([0,1]\). The first of these functions is the same, after scaling onto \([0,1]\), as an example used in Kou and Efron (2002) and Lukas et al. (2012) for simulations of certain parameter selection criteria. Fig. 1(a) and (b) displays the graphs of the two functions (solid) together with particular data sets \((x_i,y_i), i = 1, \ldots, n\) (+), where \( n = 20, x_i = (i-1)/(n-1), y_i = f(x_i) + \varepsilon_i \), and \( \varepsilon_i \) are pseudo-random variates that are i.i.d. \( N(0, \sigma^2) \) with \( \sigma = 0.1 \). Also shown in each figure is the extreme undersmoothed spline estimate \( f_{0^-} \) (dotted) and a smooth spline estimate \( f_s \) (dashed). The corresponding errors \( f_{0^-} - f \) (dotted) and \( f_s - f \) (dashed) are plotted in Fig. 2(a) and (b).

![Fig. 1: Function (a) \( f(x) = \sin(2\pi x)/(x+1/2) \) and (b) \( f(x) = \sin(4\pi x)/(x+1/2) \), with data \((x_i,y_i)\) (+) for \( \sigma = 0.1 \) and \( n = 20 \), and \( f_{0^-} \) (dotted) and a smooth spline estimate \( f_s \) (dashed).](image)

![Fig. 2: Error \( f_{0^-} - f \) (dotted) and \( f_s - f \) (dashed) for function (a) \( f(x) = \sin(2\pi x)/(x+1/2) \) and (b) \( f(x) = \sin(4\pi x)/(x+1/2) \), where \( f_{0^-} \) and \( f_s \) are plotted in Fig. 1(a) and (b).](image)

![Fig. 3: Twenty replicates of \( T(\lambda) \) (dotted), \( ET(\lambda) \) (solid) and the horizontal line at 1.5 min \( ET \) (dashed) for (a) \( f(x) = \sin(2\pi x)/(x+1/2) \) and (b) \( f(x) = \sin(4\pi x)/(x+1/2) \) with \( \sigma = 0.1 \) and \( n = 20 \).](image)
Fig. 4. Twenty replicates of $T(\lambda)$ (dotted), $ET(\lambda)$ (solid) and the horizontal line at 1.5 min $[ET]$ (dashed) for $f(x) = \sin(2\pi x)/(x+1/2)$ with (a) $\sigma=0.1$ and $n=10$, and (b) $\sigma=0.01$ and $n=20$.

It is visually clear in both Fig. 1(a) and (b) that $f_{0^+}$ (which interpolates the data) is a very poor estimate of the function $f(x)$. It is far too wiggly and, furthermore, it would be nearly useless for estimating first or higher derivatives of $f(x)$. Based on our visual perception, the smoother spline estimate (dashed) in Fig. 1(a) and (b) is a much more acceptable estimate of the function. This is also clear from Fig. 2(a) and (b).

For the function $f(x) = \sin(2\pi x)/(x+1/2)$ in Fig. 1(a) and 20 replicates of the data, Fig. 3(a) shows the corresponding 20 replicates of the prediction error $T(\lambda)$ (dotted), together with the prediction risk $ET(\lambda)$ (solid) and a horizontal dashed line at 1.5 min$[ET]$, where min$[ET]=\min ET(\lambda)$. While there is substantial variability in the replicates of $T(\lambda)$, they all have a fairly similar shape in that the graph is steep for $\lambda$ above the minimizer and it is relatively flat below it.

For the lowest replicate of $T(\lambda)$ for small $\min T$ in Fig. 3(a), $T(10^{-10})$ has the same value ($\approx 0.0047$) as $T(\lambda)$ for $\lambda=4.2 \times 10^{-5}$. The corresponding data set is the one shown in Fig. 1(a). The spline $f_{j}$ for $\lambda=10^{-10}$ is essentially the same as the extreme undersmoothed $f_{0^+}$, and the spline $f_{j}$ for $\lambda=4.2 \times 10^{-5}$ is the preferred, smooth estimate in Fig. 1(a). Therefore, according to our visual perception, the prediction error fails to discriminate the extreme undersmoothing for $\lambda=10^{-10}$ and $\lambda=0^+$. A simple quantity that helps to diagnose this problem is the ratio $T(10^{-10})/min T$, where $min T=\min T(\lambda)$. If this is small (close to 1), then we can expect that the prediction error will fail to discriminate the extreme undersmoothing for $\lambda=10^{-10}$. For the replicate discussed above, $T(10^{-10})/min T \approx 3.0$, so this value is small enough. For the 20 replicates in Fig. 3(a), the ratio ranges from 2.2 to 4.3, with median 3.0.

Using the same 20 replicates of the error vector as in Fig. 3(a), Fig. 3(b) shows the corresponding replicates of $T(\lambda)$ for the function $f(x) = \sin(4\pi x)/(x+1/2)$. It has higher curvature, and it appears that the situation is even worse here, because the ratio $T(10^{-10})/min T$ is generally smaller than for Fig. 3(a); it ranges from 1.5 to 3.7, with median 1.8. For the lowest replicate of $T(\lambda)$ for small $\min T$ in Fig. 3(b), the ratio is 2.1 and the corresponding data set is shown in Fig. 1(b). The preferred, smooth estimate in Fig. 1(b) is the spline $f_{j}$ for $\lambda=4.4 \times 10^{-6}$ and, since this satisfies $T(\lambda)=T(10^{-10})$, the prediction error fails to discriminate the extreme undersmoothing for $\lambda=10^{-10}$ and $\lambda=0^+$.

For the first function $f(x) = \sin(2\pi x)/(x+1/2)$, Fig. 4(a) shows 20 replicates of $T(\lambda)$ when $n$ is decreased to 10 (and $\sigma=0.1$), while Fig. 4(b) shows 20 replicates of $T(\lambda)$ when $\sigma$ is decreased to 0.01 (and $n=20$). Comparing these figures with Fig. 3(a), it is clear that, in both situations, the ratio $T(10^{-10})/min T$ is generally closer to 1, so the discrimination of extreme undersmoothing is worse.

2.2. Asymptotic analysis of the prediction error

To analyze the behavior observed above, first note that the prediction error for $f_{0^+}$ is

$$T(0^+) = n^{-1} \sum_{i=1}^{n} (f_{0^+}(x_i) - f(x_i))^2 = n^{-1} \sum_{i=1}^{n} e_i^2.$$  \hfill (4)

If the errors $e_i$ are i.i.d. $N(0, \sigma^2)$, then $ET(0^+) = \sigma^2$ and $X_{\sigma} = N(n/\sigma^2)T(0^+)$ has a chi-square distribution with $n$ degrees of freedom. Using the moments $EX=\sigma^2$ and $E(X^2) = n(n+2)$, the variance of $T(0^+)$ is

$$\text{Var}(T(0^+)) = (\sigma^4/n)[E(X^2) - (EX)^2] = 2n^{-1}\sigma^4.$$ \hfill (5)

To compare $T(0^+)$ with $\min T$, we will use the known asymptotic expressions for $T(\lambda)$ and $ET(\lambda)$ (see, e.g., Wahba, 1990). We will assume that the errors are i.i.d. normal and we will make the same additional assumptions as in Cox (1984a), Nychka (1990) and Lukas et al. (2012). The assumptions are as follows.

Assumption A1. The random errors $e_i$ are i.i.d. $N(0, \sigma^2)$.

Assumption A2. The points $x_i$ satisfy the conditions of either Case A or Case B below, which cover deterministic and random design points, respectively.

Case A. Let $g_{\lambda}$ denote the empirical distribution function for the design points $\{x_i\}$. There is a distribution function $G$ such that $sup |g_{\lambda}(x) - G(x)| = O(1/n)$. If $\{x_i\}$ are equally spaced, then this holds with $G(x) = (x-a)/(b-a)$.
Case B. The set \( \{x_i\} \) is a random sample from a distribution with c.d.f. \( G \).

In both cases, it is assumed that \( G \in C^\infty[a, b] \) and \( G \) is strictly positive on \([a, b]\).

It is well known that the spline estimator is a linear smoother. Let \( A=\lambda^d \) denote the smoothing matrix defined by
\[
 f_\lambda = A y, \quad \text{where} \quad y=(y_1, \ldots, y_n)^T \quad \text{and} \quad f_\lambda = (f_\lambda(x_1), \ldots, f_\lambda(x_n))^T.
\]
Then \( f_\lambda = Af \), where \( f = (f(x_1), \ldots, f(x_n))^T \). Note that
\[
 T(\lambda) = n^{-1} \| f - f_\lambda \|^2, \quad \text{where} \quad \| \cdot \| \text{ is the Euclidean norm.}
\]

Let \( \alpha_n \) be a sequence for which \( \alpha_n \sim n^{-4m/5}\log(n) \) for case A, and \( \alpha_n \sim n^{-2m/5}\log(n) \) for Case B. Here and later \( \alpha_n \sim Cn^p \) means that \( C_i \alpha_n \leq C_2 \alpha_n \) for some positive constants \( C_i \) and \( C_2 \).

Assumption A3. There exist constants \( p \in \{1/m, 2\} \) and \( c = c(p) > 0 \) such that, as \( n \to \infty \), the squared bias satisfies
\[
 b^2(\lambda) = n^{-1} \| (I-A)f \|^2 = c \| f \| (1+o(1)),
\]
uniformly for \( \lambda \in [\alpha_n, \infty) \).

Remark 1. Assumptions A3 can be thought of as a smoothness assumption in which the parameter \( p \) depends on the regularity of \( f \). From known results about the bias (Cox, 1984a; Nychka, 1990), it can be expected that, under certain conditions, (6) holds for \( f \) in a subset of a fractional order Sobolev space \( W^{m,2}[a, b] \) (including certain boundary conditions when \( m \geq m+1/2 \)) with the corresponding \( p=s \) for \( 1/m \leq s < 2 \) and \( p=2 \) for \( s \geq 2 \). The parameter \( \alpha \) is related to the smoothness of \( f \). In particular, for equally spaced points \( x_j = (2i-1)/(2m) \), \( f \in W^{m,2}[0, 1] \) and \( f \) satisfies the natural boundary conditions \( f^{(0)}(0) = f^{(0)}(1) = 0, \ j = m, \ldots, 2m-1 \), then (6) holds with \( p=2 \) and \( c = \int_0^1 f^{(m)}(x)^2 dx \) \( (\text{Speckman, 1981}) \). If \( m=2 \) here, then, using the Poincaré inequality: \( \int_0^1 h^2 \; dx \leq \int_0^1 (h')^2 \; dx \) if \( h(0) = 0 \), with \( h=f^s \) and then with \( h=f^{(3)} \) (since \( f^{(0)}(0)=f^{(3)}(0)=0 \)), we have \( \int_0^1 f^{(2m)}(x)^2 dx \leq c \). So, for \( m=1 \) and \( m=2 \), \( c \) is large for a function with high integrated squared curvature. This conclusion also applies for larger \( m \) if \( f \) satisfies additional zero boundary conditions.

Under the above assumptions, it can be shown (Cox, 1984a; Lukas et al., 2012; Nychka, 1990) that, as \( n \to \infty \),
\[
 ET(\lambda) = (c_p \alpha + \sigma^2 n^{-1/2} \lambda^{-1/2+1}(1+o(1)),
\]
where \( \alpha = \pi^{-1} \int_0^1 G(\lambda(x)) \1/(2m) \; dx \) and
\[
l_2 = \int_0^\infty (1 + \lambda^2)^{-1/2 \lambda} \; dx = (\Gamma(1/2m))^{-1} (\Gamma(2-1/2m)) (2m),
\]
uniformly for \( \lambda \in [\alpha_n, \infty) \). In addition, the prediction error satisfies \( \sup_{\lambda \geq \alpha_n} |T(\lambda)/ET(\lambda) - 1| = o_p(1) \) \( (\text{Nychka, 1990}) \). In the important situation of cubic splines with equally spaced points \( x_0 \) on \( [0, 1] \), we have \( m=2, \; \alpha = \pi^{-1} \) and \( l_2 = 5/6 \).

Let \( \lambda_{ET} \) denote the minimizer of \( ET(\lambda) \) for \( \lambda \in [\alpha_n, \infty) \). Minimizing the right-hand side of (7) gives the known estimate \( (\text{Nychka, 1990}; \text{Wahba, 1990}) \)
\[
\lambda_{ET} = \left( \frac{a \alpha^2}{2mpcn} \right)^{(2m/2mp+1)} (1+o(1)).
\]
Therefore, the minimum value of \( ET(\lambda) \) for \( \lambda \in [\alpha_n, \infty) \) satisfies
\[
 \min(ET) = ET(\lambda_{ET}) = c(2mp+1) \left( \frac{a \alpha^2}{2mpcn} \right)^{(2m/2mp+1)} (1+o(1)).
\]

Because \( T(\lambda) \) is a random function, it is of interest to know how its variability depends on \( \lambda \). In the important case of cubic splines with equally spaced points, the next result shows that, for large \( n \), the variance of \( T(\lambda) \) decreases as \( \lambda \) increases up to \( \lambda_{ET} \). This behavior is evident in Fig. 3(a) and (b), even though \( n=20 \) is not large.

Theorem 1. Suppose that assumptions A1, A2 and A3 hold. In addition, let \( m=2 \) and assume that the points \( x_i \) are equally spaced. Then, as \( n \to \infty \), the variance \( \text{Var}(T(\lambda)) \) is a decreasing function of \( \lambda \) for \( \lambda \in [\alpha_n, \lambda_{ET}] \). Furthermore, there exists a constant \( C \) such that, as \( n \to \infty \),
\[
 \frac{\text{Var}(T(\lambda_{ET}))}{\text{Var}(T(0^+))} \leq \frac{C}{\lambdabar^2} \left( \frac{\lambdabar^4}{4p+1} \right) \to 0.
\]

Proof. See the Appendix.

This result helps to explain why sometimes \( T(\lambda) \) does not discriminate extreme undersmoothing. If \( ET(\lambda) \) is fairly flat for \( \lambda \leq \lambda_{ET} \), then the larger variance of \( T(\lambda) \) as \( \lambda \to 0^+ \) will sometimes result in a value of \( T(0^+) \) that is close to \( \min T \).

We will use \( T(0^+) \) and \( \min(ET) \) to define an informative measure of the prediction error’s capacity to discriminate extreme undersmoothed estimates. Although, for finite \( n \), \( \min T \) will be different from \( \min(ET) \), it is nevertheless reasonable to use \( \min(ET) \) for our purpose. This is because we can regard a value \( \lambda = \lambda_{ET} \) defined by \( T(\lambda) = \min(ET) \) (assuming it exists) to be a reasonable choice of the smoothing parameter, and it would be desirable to be able to distinguish between the choices \( \lambda = 0^+ \) and \( \lambda = \lambda_{ET} \). Consequently, the appropriate question is: how often is \( T(0^+) \) close to \( T(\lambda_{ET}) = \min(ET) \)? To
measure this, we will consider the probability
\[ P_{T_{0}+}=\Pr(T(0^+)/\min(E_T) \leq D), \]  
(10)

where \( D \geq 1 \) is a constant to be chosen. Clearly, a smaller value of \( D \) is associated with a tighter comparison of \( T(0^+) \) and \( \min(E_T) \). In Fig. 3(a) and (b), and Fig. 4(a) and (b), the comparisons can be seen for \( D=1.5 \).

For our analysis, we will approximate the probability in (10) by using the asymptotic estimate of \( \min(E_T) \) in (9) and substituting the chi-square random variable \( X=(n/\sigma^2)T(0^+) \) to obtain
\[ P_{T_{0}+} \approx \tilde{P}_{T_{0}+}=\Pr(X \leq x(n, \sigma)), \]  
(11)

where
\[ x(n, \sigma) = Dc(2mp+1)(n/\sigma^2)\left(\frac{a_2\sigma^2}{2mp}\right)^{2mp/(2mp+1)}. \]  
(12)

In the remainder of this section, we will investigate how the probability \( \tilde{P}_{T_{0}+} \) in (11) depends on \( \sigma \) and, separately, on \( n \).

Because \( X \) is independent of \( \sigma \), and \( \sigma \) has a negative exponent on the right-hand side of (12), it is clear that, for each \( n \), the probability \( \tilde{P}_{T_{0}+} \) increases monotonically as \( \sigma \) is decreased, and \( \tilde{P}_{T_{0}+} \to 1 \) as \( \sigma \to 0 \). The asymptotic rate can be identified by using the known cumulative distribution function for \( X \), namely
\[ \Pr(X \leq x) = F(x; n) = \gamma(n/2, x/2)/\Gamma(n/2), \]
where \( \gamma(a, z) = \int_0^z e^{-t}t^{a-1} \, dt \) is the lower incomplete gamma function and \( \Gamma(a, z) = \int_0^\infty e^{-t}t^{a-1} \, dt \) is the gamma function. The following result, proved in the Appendix, is an easy consequence of an asymptotic expansion of \( \gamma(a, z) \) as \( z \to \infty \) (Abramowitz and Stegun, 1984, 6.5.32).

**Theorem 2.** For each fixed \( n \), as \( \sigma \to 0 \), it holds that
\[ 1 - \tilde{P}_{T_{0}+} \sim \frac{x(n, \sigma)/2^{n/2} - 1 - e^{-x(n, \sigma)/2}}{\Gamma(n/2)}, \]
(13)

where \( x(n, \sigma) = K(n)\sigma^{-2/(2mp+1)} \to \infty \) and \( K(n) \) is defined by (12).

It is clear from Theorem 2, due to the exponential factor in (13), that \( \tilde{P}_{T_{0}+} \) increases rapidly to 1 as \( \sigma \to 0 \).

The behavior of the probability \( \tilde{P}_{T_{0}+} \) as \( n \to \infty \) is more difficult to determine. This is because in (11), both \( X \) and \( x(n, \sigma) \) depend on \( n \). However, by using two appropriate asymptotic expansions, we can obtain the following result:

**Theorem 3.** Write \( x(n, \sigma) \) in (12) as \( x(n, \sigma) = Cn^\beta \), where \( \beta = 1/(2mp+1) \) and
\[ C = Dc(2mp+1)(1/\sigma^2)\left(\frac{a_2\sigma^2}{2mp}\right)^{2mp/(2mp+1)}. \]  
(14)

As \( n \to \infty \), it holds that
\[ \tilde{P}_{T_{0}+} \sim \frac{(\pi n)^{-1/2}e^{-n/2}}{1 - Cn^\beta} \to 0, \]
(15)

where
\[ r = (1 - \beta)\ln n - (1 + \ln C) \to \infty \]
and \( Cn^\beta \to 0 \). Also, as \( n \to \infty \), it holds that
\[ \tilde{P}_{T_{0}+} \sim \frac{(1/2)e^{-c_1\tau n/2}}{1 - Cn^\beta} + (\pi n)^{-1/2}e^{-n/2}\left(\frac{1}{1 - Cn^\beta} - (2r)^{-1/2}\right). \]
(17)

where \( \text{erfc}(x) = 2\pi^{-1/2} \int_x^\infty e^{-t^2} \, dt \) is the complementary error function and \( \tau = r + Cn^\beta \to \infty \).

**Proof.** See the Appendix.

Clearly, due to the exponential factor in the asymptotic estimate (15), \( \tilde{P}_{T_{0}+} \) decays very quickly as \( n \to \infty \). However, from (16), the decay rate is affected by the values of the parameters \( C \) and \( \beta \); larger values of \( C \) and \( \beta \) give a slower decay. Therefore, because \( \sigma \) appears in \( C \) in (14) with a negative exponent, a smaller value of \( \sigma \) will give a slower decay.

It is clear from (14) that, for a fixed value of \( p \) (say \( p=2 \)), \( C \) increases with \( c \), so a larger value of \( c \) (a function with higher curvature) will give a slower decay. However, we cannot determine the effect of the parameter \( c \) because \( C \) depends on both \( p \) and \( c = c(p) \), and the behavior of \( c(p) \) is unknown.

Numerical experiments with (15) show that it is not very accurate for small values of \( n \), especially if \( x(n, \sigma) = Cn^\beta \) is fairly close to \( n \). The asymptotic estimate in (17) is much more accurate in these situations. This is because it is derived from an asymptotic expansion of \( F(x/2, n/2) \) that is accurate for a large range of \( x \). Because of the parameter \( \tau \) in both the erfc and exponential terms in (17), it is clear that the decay rate of this estimate is affected by the values of \( C \) and \( \beta \) in the same way as...
discussed above, so a smaller value of \( \sigma \) or a larger value of \( c \) will give a slower decay as \( n \) increases, even for quite small values of \( n \).

To illustrate the behavior of the probability \( \hat{P}_{10^{-1}} \), we let \( D = 1.5 \) and consider the same examples as in Section 2.1, i.e., the functions \( f(x) = \sin(2\pi x)/(x+1/2) \) and \( f(x) = \sin(4\pi x)/(x+1/2) \) in Fig. 1(a) and (b), respectively, and cubic smoothing splines \( m = 2 \) with equally spaced points \( x_i = (i-1)/(n-1), \ i = 1, \ldots, n \). The probability \( \hat{P}_{10^{-1}} \) (which is defined using the asymptotic behavior of \( ET(\lambda) \)) is approximated by using estimates of the parameters \( p \) and \( c \) in (6) obtained from the plot of \( b_2(\lambda) \) for \( n = 100 \). The estimates are \( p = 2 \) and \( c = 2.8 \times 10^5 \) for \( f(x) = \sin(2\pi x)/(x+1/2) \) and \( p = 2 \) and \( c = 1.4 \times 10^{10} \) for \( f(x) = \sin(4\pi x)/(x+1/2) \). Note that, while \( n = 100 \) is not very large, we use this value to see if the asymptotic behavior for \( \hat{P}_{10^{-1}} \) derived above is evident for \( n \) near 100. The estimate \( p = 2 \) is consistent with the results in Remark 1, as is the fact that the estimates of \( c \) are large. This is because \( \int_0^1 f^{(4)}(x)^2 \, dx \) is large (equal to \( 1.1 \times 10^8 \) for \( f(x) = \sin(2\pi x)/(x+1/2) \) and \( 4.1 \times 10^8 \) for \( f(x) = \sin(4\pi x)/(x+1/2) \)) and \( c \) in (6) will be even larger due to boundary effects.

For these two functions \( f(x) \), Fig. 5(a) and (b), respectively, show the approximations of \( \hat{P}_{10^{-1}} \) plotted against \( n \in [5, 120] \) for \( \sigma = 0.1 \) (solid), \( \sigma = 0.01 \) (dashed) and \( \sigma = 0.001 \) (dash-dot). The plots are consistent with the analytic results for \( \hat{P}_{10^{-1}} \) derived above. Since the graphs shift to the right substantially as \( \sigma \) is decreased, the approximation of \( \hat{P}_{10^{-1}} \) increases rapidly as \( \sigma \) decreases. In addition, for each value of \( \sigma \), the approximation of \( \hat{P}_{10^{-1}} \) decays rapidly to 0 as \( n \) increases. The decay is less rapid for a smaller value of \( \sigma \) or for a larger value of \( c \), both of which define a larger value of the constant \( c \) in (14). For the particular values of \( \sigma \) and \( c \) used in Fig. 5(a) and (b), \( C \) ranges from 12.2 to 40.7.

It should be noted that the quantities plotted in Fig. 5(a) and (b) are approximations of \( \hat{P}_{10^{-1}} \), involving the asymptotic form for \( ET(\lambda) \), and so cannot be expected to be accurate estimates of \( \hat{P}_{10^{-1}} \) for small \( n \). However, it will be seen from simulations in Section 4 with the same examples that the behavior inferred from Fig. 5(a) and (b) also holds for \( \Pr(T(0^+) \leq D \min_T) \) for small \( n \). This means that the capacity of the prediction error to discriminate extreme undersmoothing is worse for either small \( n \), small \( \sigma \) or a function with high curvature. These findings are consistent with the observations made in Section 2.1.

3. The Sobolev error and discrimination of extreme undersmoothing

3.1. Illustrations of the behavior

To illustrate the behavior of the Sobolev error \( W(\lambda) \) in (3), we will use \( m = 2 \) and the same two functions as in Section 2.1, namely \( f(x) = \sin(2\pi x)/(x+1/2) \) and \( f(x) = \sin(4\pi x)/(x+1/2) \) on \([0, 1]\). As discussed in Section 1 and also shown in Fig. 2(a) and (b), for very small \( \lambda \), the error \( f - \hat{f} \) is very wiggly. Because the second term of \( W(\lambda) \) is the integrated squared curvature
of \(f_\lambda - f\) and the first term is approximately equal to \(T(\lambda)\), the Sobolev error \(W(\lambda)\) has much higher values than the prediction error \(T(\lambda)\) for \(\lambda\) near 0. As \(\lambda\) increases, \(f_\lambda - f\) becomes less wiggly so the second term of \(W(\lambda)\) decreases considerably. This has the effect that the graph of \(W(\lambda)\) is not as flat as the graph of \(T(\lambda)\) for \(\lambda\) below the minimizer.

This property can be seen clearly in Fig. 5(a) and (b) which show graphs of \(W(\lambda)\) (dotted) for the same twenty replicates of the error vector as were used for the graphs of \(T(\lambda)\) in Fig. 3(a) and (b), and also \(EW(\lambda)\) (solid) which was estimated to graphical accuracy by the mean of 100 replicates of \(W(\lambda)\). In both Fig. 6(a) and (b), for all twenty replicates, \(W(10^{-10}) \approx W(0^+)\) is significantly larger than the minimum value of \(W(\lambda)\), so the Sobolev error is able to discriminate the extreme undersmoothing that occurs when \(\lambda = 0^+\).

Recall that Fig. 1(a) illustrates, for one of the twenty replicates, how the prediction error fails to discriminate the extreme undersmoothing that occurs when \(\lambda = 0^+\). For the Sobolev error, because in Fig. 6(a), \(W(0^+)\) is larger than \(\lim_{\lambda \to \infty} W(\lambda)\) for each replicate, there is no (slightly or even greatly) oversmoothed spline \(f_\lambda\) for which \(W(\lambda)\) has the same value as \(W(0^+)\). It is well known that, for any data set, as \(\lambda \to \infty\), the cubic spline \(f_\lambda\) approaches the least squares regression line for the data. Therefore, the value of \(W(0^+)\) for the extreme undersmoothed spline \(f_0\), in Fig. 1(a) is significantly larger than the Sobolev error of the least squares regression line for the data shown.

Fig. 6(a) and (b) indicate that, as for the prediction error, the Sobolev error is more likely to have difficulty discriminating extreme undersmoothing for a function with higher curvature. The same is true for small values of \(n\) and small values of \(\sigma\), as illustrated in Fig. 7(a) and (b), respectively. However, by comparing Figs. 3(a), (b), 4(a), (b) with Figs. 6(a), (b), 7(a), (b), which respectively use the same replicates, it is clear that the Sobolev error is better able to discriminate extreme undersmoothing than the prediction error in all the situations.

We remark that, for computational simplicity, \(W(\lambda)\) in Figs. 6 and 7 was calculated using the approximation

\[
W(\lambda) \approx \| f_\lambda - f \|_{W^2}^2 \approx \| f_{int} - f \|_{W^2}^2, \tag{18}
\]

where \(f_{int}\) is the natural cubic spline interpolating \(f(x)\) at \(x_1, \ldots, x_n\). Because each function \(f(x)\) is smooth, this approximation is quite accurate even for \(n\) as small as 10.

### 3.2. Asymptotic analysis of the Sobolev error

It is well known that the smoothing matrix \(A(\lambda)\) has a diagonalization \(A(\lambda) = U \text{diag}(\lambda_{1:n}) U^T\), where \(U\) is orthogonal and independent of \(\lambda\), and \(\lambda_i = 1/(1 + \tau_i)\), \(i = 1, \ldots, n\), for a certain nondecreasing sequence \([\tau_i]\), with \(\tau_i = 0\) for \(i = 1, \ldots, m\). This spectral decomposition will be used in an approximation of \(W(\lambda)\).

As discussed in Lukas et al. (2012), \(W(\lambda)\) can be approximated by

\[
W(\lambda) \approx \| f_\lambda - f \|_{W^2}^2 \approx \overline{W}(\lambda) = \| f_\lambda - f_{int} \|_{W^2}^2, \tag{19}
\]

where \(f_{int}\) is the natural spline interpolating \(f(x)\) at the points \(x_i\), and \(\| \cdot \|_{W^2}^2\) is the partially discretized squared Sobolev norm given by

\[
\| h \|_{W^2}^2 = n^{-1} \| h \|^2 + \kappa \| h^{(m)} \|_{L_2}^2,
\]

where \(\kappa = (b-a)^{m+1}\). Clearly, \(\overline{W}(\lambda) = T(\lambda) + \kappa \| f_\lambda^{(m)} - f_{int}^{(m)} \|_{L_2}^2\). Using the same reasoning as in the proof of Lemma 2 in Lukas et al. (2012), the term \(\| f_\lambda^{(m)} - f_{int}^{(m)} \|_{L_2}^2\) in \(W(\lambda)\) can be expressed in terms of the spectral decomposition of \(A(\lambda)\)

\[
\| f_\lambda^{(m)} - f_{int}^{(m)} \|_{L_2}^2 = n^{-1} \sum_{i = m+1}^{n} \tau_i |a_{ij}(U^T y_i - (U^T f))|^2. \tag{20}
\]

Therefore, since \(a_{ij} = 1\) when \(\lambda = 0\), the Sobolev error for the extreme undersmoothed spline estimate \(f_{0^-}\) can be written as

\[
W(0^+) \approx W(0^+) = n^{-1} \| f_{0^-} - f \|^2 + \kappa \| f_{0^-}^{(m)} - f_{int}^{(m)} \|_{L_2}^2
\]
analogous to Assumption A3. It is known (Utreras, 1987, 1988) that, under the assumption made on \(b\), there are constants \(c_1\) and \(c_2\) independent of \(i\) and \(n\) such that \(c_1(i-m)^{2m} \leq \tau_i \leq c_2(i-m)^{2m}\) for all \(i = m+1, \ldots, n\). Therefore, using the form of a known asymptotic expression (Speckman, 1981) which includes the dependence on \(G\), there is a constant \(d_1 > 0\) such that

\[
\tau_i \geq d_1(i-m)^{2m} \left( \int_a^b \left( G(x) \right)^{1/(2m)} dx \right)^{-2m}
\]

for all \(i > m\). In particular, for cubic splines and equally spaced \(x_i\) on \([0,1]\) \((G(x) = x)\), the bound (23) holds with \(d_1 = 1 - \pi^2/18\) (Kou, 2003). Using (23) in (22) yields

\[
\mathbb{E} \mathbb{W}(0^+) \leq \sigma^2 + \sigma^2 n^{-1} \sum_{i=m+1}^{n} (\pi(i-m))^2m,
\]

where \(d = d_1 \kappa \int (G(x))^{1/(2m)} dx -2m\). Note that \(d \geq d_1\), with equality for equally spaced \(x_i\) on \([a,b]\), since

\[
\kappa \int_a^b (G(x))^{1/(2m)} dx \geq \left( \int_0^1 \frac{dG}{dt}(a+t(b-a)) dt \right)^{-1} \geq (G(b)-G(a))^{-1} = 1,
\]

where the inequality follows from “Holder’s inequality for \(0 < p < 1\)” (Hewitt and Stromberg, 1965, p. 191) (with \(p = 1/(2m)\) and the second function identically equal to 1). From (24) it is clear that \(\mathbb{E} \mathbb{W}(0^+) \approx \sigma^2\) for all sufficiently large \(n\). For cubic splines and equally spaced \(x_i\), the lower bound in Kou (2003, Cor. 2.2) gives

\[
\mathbb{E} \mathbb{W}(0^+) \leq \sigma^2 + \sigma^2 n^{-1} \sum_{i=3}^{n} (\pi(i-2))^4 \left(1 - (\pi(i-2))^2/(18n^2)\right)
\]

and a graph shows that the right-hand side of (25) is greater than \(6n^4\sigma^2\) for \(n \geq 10\).

To estimate \(\mathbb{W}(\lambda)\) for general \(\lambda\), we will use Assumptions A1 and A2 in Section 2.2, and the next assumption that is analogous to Assumption A3.

Let \(b^2(\lambda)\) be the squared bias with respect to the Sobolev semi-norm, defined as

\[
b^2(\lambda) = \kappa \|E_{\lambda} f^m - f^m\|_2^2.
\]

**Assumption A4.** Either part (a) or part (b) holds.

(a) There are constants \(p \in (1, 2)\), \(c\) and \(c_1\) such that, as \(n \to \infty\),

\[
b^2(\lambda) = c\lambda^p (1 + o(1))\quad \text{and} \quad b^2(\lambda) = c_1\lambda^2 (1 + o(1)),
\]

uniformly for \(\lambda \in [a_n, \infty)\). (This is the same assumption as in Section 4 of Lukas et al., 2012.)

(b) There are constants \(c\) and \(c_1\) such that, as \(n \to \infty\),

\[
b^2(\lambda) = c\lambda^2 (1 + o(1))\quad \text{and} \quad b^2(\lambda) = c_1\lambda^2 (1 + o(1)),
\]

uniformly for \(\lambda \in [a_n, \infty)\).

For Assumption A4(a) to hold, it is necessary (Cox, 1988; Lukas, 1993) that \(f\) has regularity between that corresponding to \(\lambda m^2[a,b]\) and \(\lambda m^2[a,b]\). Assumption A4(a) is associated with \(f\) having smoothness such that \(S_p = \sum_{i=0}^{n-1} \lambda^2 t_i f^2(\lambda)^{-1}\) is bounded (independent of \(n\), while A4(b) is associated with \(f\) having higher smoothness such that \(S_3 = \sum_{i=0}^{n-1} \lambda^2 t_i f^2(\lambda)^{-1}\) is bounded. The association for A4(b) follows since from (20),

\[
b^2(\lambda) = \kappa n^{-1} \sum_{i=m+1}^{n} \lambda^2 t_i (f^2(\lambda)^{-1})^2 \leq \lambda^2 \kappa n^{-1} \sum_{i=m+1}^{n} t_i^2 (f^2(\lambda)^{-1})^2 = \lambda^2 \kappa S_3
\]

and \(b^2(\lambda) = n^{-1} (1 - \lambda f^2)^2 = n^{-1} \sum_{i=0}^{n-1} \lambda^2 t_i f^2(\lambda)^{-1} (\lambda f_i)^2 \leq \lambda^2 \kappa_1 n^{-1} S_3\). The association for A4(a) follows since

\[
b^2(\lambda) = \kappa n^{-1} \sum_{i=m+1}^{n} \lambda^2 t_i f^2(\lambda)^{-1} (\lambda f_i)^2 \leq \lambda^2 \kappa_1 n^{-1} S_3
\]

and, similarly, \(b^2(\lambda) \leq \lambda^2 \kappa P_2\).

We will also assume the following about \(\mathbb{W}(\lambda)\), where \(\mathbb{W}(\lambda)\) is defined in (19).

---

Please cite this article as: Lukas, M.A., Performance criteria and discrimination of extreme undersmoothing in nonparametric regression. Journal of Statistical Planning and Inference (2014), http://dx.doi.org/10.1016/j.jspi.2014.05.006
Assumption A5. As \( n \to \infty \), \( E(W(\lambda)) = E(W(\lambda)(1 + o(1))) \), uniformly for \( \lambda \in [a_n, \infty) \).

Under assumptions A1, A2, A4 and A5, from the proof of Theorem 3 in Lukas et al. (2012), as \( n \to \infty \), the expected Sobolev error can be estimated as

\[
E(W(\lambda)) = \begin{cases} (c_1 + \omega_1) \alpha^2 (1 + o(1)) \lambda^{-1} \alpha^{-1-1/(2m-1)}(1 + o(1)) & \text{for A4(a)}, \\ (c_1 + \omega_1) \alpha^2 (1 + o(1)) & \text{for A4(b)}, \\ \end{cases}
\]

uniformly for \( \lambda \in [a_n, \infty) \). In addition, the minimizer \( \lambda_{E_\lambda} \) of \( E(W(\lambda)) \) for \( \lambda \in [a_n, \infty) \) is

\[
\lambda_{E_\lambda} = \left\{ \begin{array}{ll} \frac{al_1 - l_2}{2m - 1} & \text{for A4(a)}, \\ \frac{al_1 - l_2}{4m(c + c_1)n} & \text{for A4(b)} \end{array} \right.
\]

and the minimum value of \( E(W(\lambda)) \) for \( \lambda \in [a_n, \infty) \) can be expressed as

\[
\min[E(W(\lambda))] = \left\{ \begin{array}{ll} \frac{al_1 - l_2}{2m - 1} \left( (1 + o(1)) \right) & \text{for A4(a)}, \\ \frac{al_1 - l_2}{4m(c + c_1)n} \left( (1 + o(1)) \right) & \text{for A4(b)} \end{array} \right.
\]

It is clear that \( E(W(0^+)) \) and \( \min[E(W)] \) have very different behaviors depending on \( n \). From (24), \( E(W(0^+)) \) increases quickly with \( n \) (unlike \( ET(0^+) \)), while, from (26), \( \min[E(W)] \rightarrow 0 \) as \( n \to \infty \). This is already an indication that \( W(\lambda) \) will perform better than \( T(\lambda) \) in discriminating extreme undersmoothed spline estimates.

To measure the capacity of the Sobolev error to discriminate extreme undersmoothed estimates, we will consider the probabilities

\[
P_{W(0^+)} := P(W(0^+)/\min[E(W)] \leq D) \quad \text{and} \quad \tilde{P}_{W(0^+)} := P(W(0^+)/\min[E(W)] \leq D),
\]

where \( D \) is a constant to be chosen and \( \min[E(W)] \) is the asymptotic estimate of \( \min[E(W)] \) on the right-hand side of (26). The measure will be bounded and compared to the corresponding measure for the prediction error.

From (21) and since \( \|\varepsilon\|^2 = \|UW\|^2 \), we have

\[
W(0^+) = n^{-1} \sum_{i=1}^{n} (U^T\varepsilon_i)^2 + \kappa n^{-1} \sum_{i=m+1}^{n} T(U^T\varepsilon_i)^2 \geq n^{-1} \sum_{i=1}^{m} (U^T\varepsilon_i)^2 + \kappa \sum_{i=m+1}^{n} (U^T\varepsilon_i)^2.
\]

Because \( (U^T\varepsilon) \) are i.i.d. \( N(0, \sigma^2) \), the right-hand side of (28) is a weighted sum of i.i.d. chi-square variables. Using (28) in the measure \( P_{W(0^+)} \) in (27) yields

\[
P_{W(0^+)} \leq 
\sum_{i=1}^{m} Y_i + \sum_{i=m+1}^{n} \kappa \tau i \left( n/\sigma^2 \right) \|\varepsilon\|^2 \min[E(W)]
\]

where \( Y_i = \left( U^T\varepsilon \right)_i / \sigma^2, i = 1, \ldots, n \), are i.i.d. chi-square random variables with one degree of freedom. By using a stochastic inequality for weighted sums of independent chi-square variables (Okamoto, 1960; Yu, 2011), we obtain the following result:

**Theorem 4.** Suppose that assumptions A1, A2, A4 and A5 hold. Then, as \( n \to \infty \),

\[
P_{W(0^+)} \leq P(T(0^+)/\|\varepsilon\|^2) \leq \tilde{H} \sigma^{-2} \gamma_n,
\]

where \( q = 2m/(2m + 1) \) for A4(a) and \( q = 8m^2/((4m + 1)(6m + 1)) \) for A4(b), and

\[
\gamma_n = n^{q/2} \left( d^{-1}(e/\pi)^{2m} n^{-2m(2m + 1)} \right)
\]

and \( \tilde{H} = H(1 + o(1)) \), with

\[
H = \begin{cases} \frac{c_1}{c_2} \frac{1}{(2m + 1)} & \text{for A4(a)}, \\ \frac{c_1}{c_2} \frac{1}{4m+1} & \text{for A4(b)} \end{cases}
\]

Furthermore, for any \( \epsilon > 0 \), as \( n \to \infty \),

\[
\tilde{P}_{W(0^+)} \leq \exp \left\{ - \left( n \left( 2m - q \right) n \ln(n) - \frac{1}{2} \ln \left( d^{-1}(e/\pi)^{2m} \sigma^{-2} \right) \right) \right\}(1 + o(1)),
\]

where \( \tilde{H} = H(1+\epsilon) \).
Proof. See the Appendix.

This result indicates that the Sobolev error has better discrimination of extreme undersmoothing than the prediction error. From (30) and (31), since \( n^{-2/3} \to 0 \) as \( n \to \infty \), we can expect \( P_{\hat{T}_n^{0+}} \) to be smaller than \( P_{\hat{T}_n^{0-}} \) for all \( n \) large enough that approximately \( H \sigma^{-2} n^{-2/3} \sigma^2 < 1 \). From (33), since \( q < 1 \) for \( A_4(a) \) and \( q < 1/3 \) for \( A_4(b) \), clearly \( P_{\hat{T}_n^{0+}} \to P_{\hat{T}_n^{0-}} \) converges to 0 rapidly as \( n \to \infty \). Note that the rate of decay of the bound in (33) is affected by the size of \( H \) and \( \sigma \); clearly, the rate is faster for a smaller value of \( H \) and a larger value of \( \sigma \).

4. Simulations

We consider the same two functions \( f(x) = \sin(2\pi x)/(x+1/2) \) and \( f(x) = \sin(4\pi x)/(x+1/2) \) as in Sections 2.1 and 3.1, with equally spaced design points \( x_i = (i-1)/(n-1), i = 1, \ldots, n \), i.i.d. \( N(0, \sigma^2) \) errors and cubic smoothing splines. In this section, to assess the prediction error’s capacity to discriminate extreme undersmoothing, we will compare \( T(0^+) \) with \( \min T \). The value of \( T(0^+) \) is approximated by \( T(10^{-10}) \) and \( \min T \) is calculated by minimizing \( T(\lambda) \) on the grid with logarithmic steps of 0.05. The simulations were implemented in MATLAB.

First, corresponding to (10), we estimate the probability \( \Pr(T(0^+) \leq 1.5 \min T) \) for \( n = 6, 8, \ldots, 120 \) and \( \sigma = 0.1, 0.01 \), using 1000 replicates of the error vector. The estimated probabilities are plotted against \( n \) in Fig. 8(a) for \( f(x) = \sin(2\pi x)/(x+1/2) \) and in 8(b) for \( f(x) = \sin(4\pi x)/(x+1/2) \). It is clear that the behavior of the probability (as it depends on \( n \) and \( \sigma \)) observed in these figures is similar to that in Fig. 5(a) and (b).

We also present the results by plotting the median and interquartile range of \( \min T/T(0^+) \) using a simplified box plot. This representation has the advantage that there is no specification of the factor \( D \) (as in \( D = 1.5 \)), and so we can get an idea of the probabilities \( \Pr(T(0^+) \leq D \min T) \) for different values of \( D \). (Note that \( \min T/T(0^+) \) is used rather than \( T(0^+)/\min T \) so that the values lie in [0,1].) Fig. 9(a) and (b) show the plots of the medians (and) against \( n \) for \( \sigma = 0.1, 0.01 \) and 0.001, corresponding to Fig. 8(a) and (b), respectively. To see the connection, from Fig. 9(a), the median of \( \min T/T(0^+) \) is roughly 2/3 when \( n = 10 \) so \( \Pr(T(0^+) \leq 1.5 \min T) \) is approximately 0.5 when \( n = 10 \), which is consistent with Fig. 8(a). As another case, Fig. 9(a) shows that, for \( f(x) = \sin(2\pi x)/(x+1/2) \) and \( \sigma = 0.001 \), we would need to take \( n \geq 40 \) (approximately) to be 75% certain that \( \min T/T(0^+) \leq 0.8 \). Similarly, Fig. 9(b) shows that, for \( f(x) = \sin(4\pi x)/(x+1/2) \) and \( \sigma = 0.001 \), we would need to take \( n \geq 58 \) (approximately) to be 75% certain that \( \min T/T(0^+) \leq 0.8 \).

For the same examples, we will assess the Sobolev error’s capacity to discriminate extreme undersmoothing by comparing \( W(0^+) \) with \( \min W \). For computational simplicity, \( W(\lambda) \) was calculated using the approximation (18).

![Fig. 8](image-url) Simulation estimates of \( \Pr(T(0^+) \leq 1.5 \min T) \) for (a) \( f(x) = \sin(2\pi x)/(x+1/2) \) and (b) \( f(x) = \sin(4\pi x)/(x+1/2) \), plotted against \( n \) for \( \sigma = 0.1 \) (bottom), \( \sigma = 0.01 \) (middle) and \( \sigma = 0.001 \) (top), using 1000 replicates for each point.

![Fig. 9](image-url) The median (and) of computed values of \( \min T/T(0^+) \) with upper and lower quartiles (spanned by line segments) for (a) \( f(x) = \sin(2\pi x)/(x+1/2) \) and (b) \( f(x) = \sin(4\pi x)/(x+1/2) \), plotted against \( n \) for \( \sigma = 0.1 \) (bottom), \( \sigma = 0.01 \) (middle) and \( \sigma = 0.001 \) (top), using 1000 replicates for each point.
Fig. 10. Simulation estimates of \( \Pr(\text{W}(0^+) \leq 1.5 \text{ min } \text{W}) \) for (a) \( f(x) = \sin(2\pi x)/(x+1/2) \) and (b) \( f(x) = \sin(4\pi x)/(x+1/2) \), plotted against \( n \) for \( \sigma = 0.1 \) (bottom), \( \sigma = 0.01 \) (middle) and \( \sigma = 0.001 \) (top), using 1000 replicates for each point.

Fig. 11. The median (⋅) of computed values of \( \min W/\text{W}(0^+) \) with upper and lower quartiles (spanned by line segments) for (a) \( f(x) = \sin(2\pi x)/(x+1/2) \) and (b) \( f(x) = \sin(4\pi x)/(x+1/2) \), plotted against \( n \) for \( \sigma = 0.1 \) (bottom), \( \sigma = 0.01 \) (middle) and \( \sigma = 0.001 \) (top), using 1000 replicates for each point.

We estimate the probability \( \Pr(\text{W}(0^+) \leq 1.5 \text{ min } \text{W}) \) for \( n = 6, 8, \ldots, 120 \) and \( \sigma = 0.1, 0.01 \) and 0.001, using the same 1000 replicates of the error vector as used for the prediction error. The estimated probabilities are plotted against \( n \) in Fig. 10(a) for \( f(x) = \sin(2\pi x)/(x+1/2) \) and in 10(b) for \( f(x) = \sin(4\pi x)/(x+1/2) \). Clearly, as \( n \) increases, the probabilities decay significantly faster to 0 than those for the prediction error in Fig. 8(a) and (b), consistent with Theorem 4.

Fig. 11(a) and (b) shows the median (⋅) and interquartile range of \( \min W/\text{W}(0^+) \) plotted against \( n \) for \( \sigma = 0.1, 0.01 \) and 0.001. Clearly, the median decreases with \( n \) significantly faster than the corresponding median for the prediction error in Fig. 9(a) and (b). As a particular comparison, for \( f(x) = \sin(2\pi x)/(x+1/2) \) and \( \sigma = 0.001 \), we would only need to take \( n \geq 26 \) to be 75% certain that \( \min W/\text{W}(0^+) \leq 0.8 \) (not \( n \geq 40 \) as found for the prediction error).

5. Extensions to other estimators

While the results in the previous sections about the prediction and Sobolev errors have been developed for smoothing spline estimators, similar results also hold for other nonparametric regression estimators.

In particular, for the prediction error, consider the Gasser–Müller kernel estimator, with boundary correction kernels, and the local linear estimator. Both of these estimators interpolate the data as the bandwidth \( h \to 0^+ \), so \( T(0^+) = \sum_i \epsilon_i^2 \) (as for smoothing splines). In addition, for both estimators, if \( f \in C^2[a, b] \), the asymptotically optimal prediction risk is (Gasser and Müller, 1979; Eubank, 1999; Fan, 1992)

\[
\min \{ET\} = C \left( \int_a^b (f'(x))^2 \, dx \right)^{1/5} (\sigma^2/n)^{4/5}(1+o(1)),
\]

where \( C \) depends only on the kernel. Clearly, (34) is of the same form as (9) with \( m=2 \) and \( p=1 \). Therefore, very similar asymptotic results about the prediction error’s capacity to discriminate extreme undersmoothing hold for these methods.

To intuitively see that the Sobolev error should be better than the prediction error in the discrimination of extreme undersmoothing for the Gasser–Müller estimator, let \([a, b] = [0, 1]\) and \( G(x) = x \), and consider the Fourier series estimator of the form

\[
g_n(x) = a_0(y) + 2 \sum_{j=1}^{n-1} a_j(\lambda) x_j(y) \cos(\pi \lambda x),
\]

This estimator is also consistent and asymptotically normal.
where
\[ a_j(y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i \int_{-\infty}^{\infty} \cos(\pi j x) \, dx \]

with \( s_0 = 0, s_1 = \frac{(x_i + x_{i+1})}{2} \) and \( s_n = 1 \), and \( a_j(\lambda) \) is a taper sequence depending on a smoothing parameter \( \lambda \). As discussed in Hart (1997, Sec. 2.4), this estimator is essentially the same as a Gasser–Müller estimator with kernel \( K \) if the taper satisfies \( a_j(\lambda) = \phi_K(ax) \), where \( \phi_K \) is the characteristic function of the density function \( K \). For the case of extreme undersmoothing (i.e., as \( \lambda \to 0^+ \)), we have \( a_0(0^+) = 1 \) for all \( j \). Assume that
\[ f(x) = a_0(f) + 2 \sum_{j=1}^{n-1} a_j(f) \cos(\pi j x), \]

where the right-hand side is a truncated cosine series with Fourier coefficients replaced by quadrature approximations. Then, using the orthogonality of the cosine functions (and also the sine functions if \( m \) is odd), it is easy to show that
\[ \| f - g_{0\lambda} \|_2^2 \approx (a_0(e))^2 + 2 \sum_{j=1}^{n-1} (a_j(e))^2 + 2 \sum_{j=1}^{n-1} (\sigma_j)^2 (a_j(e))^2. \]

For equally spaced points \( x_i = \frac{(2i-1)}{2} (2m) \), using the midpoint rule, we have \( a_j(e) \approx n^{-1/2} v_j \), where \( v_j = [\cos(\pi (2j-1)/2m)]_{j=1}^{2m} \). Hence \( (a_0(e))^2 + 2 \sum_{j=1}^{n-1} (a_j(e))^2 \approx n^{-1} \| e \|^2 \) since the vectors \( n^{-1/2} v_0 \) and \( n^{-1/2} v_j \), \( j = 1, \ldots, n-1 \), are orthogonal. Then
\[ \| f - g_{0\lambda} \|_2^2 \approx n^{-1} \| e \|^2 + n \sum_{j=1}^{n-1} (\sigma_j)^2 (n^{-1/2} v_j)^2, \]

which is very similar to Eq. (21) for spline smoothing, since \( r_i \) behaves like \( (\pi (i-m))^m \). Therefore, the distribution of \( f - g_{0\lambda} \) will be similar to that of \( W(0^+) \) for the smoothing spline.

It is shown in Gasser and Müller (1984) that, under reasonable assumptions, if \( f \in C^{m+2}[a, b] \) and if the kernel \( K \) of the Gasser–Müller estimator \( g_0 \) satisfies certain moment conditions, such that \( K^{(m)} \) is of order \( (m, m+2) \), and, if appropriate boundary kernels are used, then the integrated mean squared error for \( g_0 \) satisfies
\[ E \| g_0^{(m)} - f^{(m)} \|_2^2 \sim C_1 h^4 \int_a^b (f^{(m+2)}(x))^2 \, dx + C_2 (\sigma^2/n) h^{-2(m+1)} \]

as \( n \to \infty \), where the constants \( C_1 \) and \( C_2 \) depend only on \( K \). In this situation, \( K \) is a second order kernel for the estimation of \( f \), and, with appropriate boundary kernels, the integrated mean squared error for \( g_0 \) satisfies
\[ E \| g_0 - f \|_2^2 \sim D_1 h^4 \int_a^b (f^2(x))^2 \, dx + D_2 (\sigma^2/n) h^{-1} \]

where the constants \( D_1 \) and \( D_2 \) depend only on \( K \). Adding the right-hand sides of (36) and (37), and minimizing, shows that the (squared) Sobolev risk \( EW(h) \) has the asymptotically optimal rate \( n^{-4/(2m+2)} \) (which is the same as the optimal rate for (36)).

As for smoothing splines, from (35), \( E \| f - g_{0\lambda} \|_2^2 \) increases quickly with \( n \), while \( \min(\text{EW}) \to 0 \) as \( n \to \infty \). Moreover, we can expect that a result similar to Theorem 4 will hold, showing that the Sobolev error has significantly better capacity than the prediction error to discriminate undersmoothing for kernel estimators \( g_0 \).

6. Extensions to multivariate smoothing

The results of Sections 2–4 can also be extended to multivariate smoothing. In particular, consider the model
\[ y_i = f(x_i) + e_i, \quad i = 1, \ldots, n, \]
where \( x_i \in \Omega \subset \mathbb{R}^d \) are i.i.d. \( N(0, \sigma^2) \), and \( f \in \mathbb{H}^m(\Omega) \) for \( m > d/2 \). A popular estimator is the thin-plate smoothing spline \( f_\lambda \) of order \( m \), which is defined as the minimizer of
\[ n^{-1} \sum_{i=1}^{n} (y_i - h(x_i))^2 + \lambda |h|_m^2 \]

where
\[ |h|_m^2 = \sum_{m, n} \int_{\Omega} (D^n h)^2 \, dx \]

and \( D^n = \partial^n/(\partial x_1^{a_1} \cdots \partial x_d^{a_d}) \) and \( |a| = a_1 + \cdots + a_d \). It is well known that \( f_\lambda \) approaches the thin-plate interpolating spline as \( \lambda \to 0 \), so the prediction error satisfies \( T(0^+) = \sum |a|^2 \) and \( ET(0^+) = \sigma^2 \), as for the univariate smoothing spline. It is also known (Cox, 1984b; Utteras, 1987) that, under suitable prediction conditions, the asymptotically optimal rate satisfies
\[ \min(ET) = O((\sigma^2/n)^{m/(2m+d)}), \]

which, for \( d = 1 \), is consistent with (9), since \( p = 1 \). Assume that, as \( n \to \infty \), \( \min(ET) \sim k_{(\sigma^2/n)^{m/(2m+d)}} \) for some constant \( k_{(\sigma^2/n)}^{m/(2m+d)} \). Then the same analysis used in Section 2.2 to find the asymptotic behavior of the measure \( \hat{P}_{T(0^+)} \) can be applied to the
corresponding measure here, yielding estimates as in Theorem 3 but with \( \beta = d/(2m + d) \). It is interesting that, for larger \( d \), \( \beta \) is larger and the measure goes to zero more slowly. This indicates that the prediction error is worse in discriminating extreme undersmoothing for \( d \geq 2 \) than we observed for \( d = 1 \) in Sections 2 and 4. The analysis also indicates that, for any value of \( d \), the discrimination becomes worse for smaller \( s \).

To illustrate these points, we present some results of simulations on fitting a surface to data \( y_i = f(x_i) + \epsilon_i \), \( i = 1, \ldots, n \). The design points \( x_i \) are the grid points \( (i-1)(j-1)/(n_1 n_2), i = 1, \ldots, n_1, j = 1, \ldots, n_2 \) with \( n_1^2 = n \), the errors are i.i.d. \( N(0, \sigma^2) \) variates, and the function \( f \) is defined by \( f(x_1, x_2) = \sin(2\pi(x_1 + x_2)/(x_1 + x_2 + 1/2)) \), which is an extension of the univariate function used in previous sections. A thin-plate smoothing spline of order 2 is used to estimate the function. The computations were done using the Fortran package GCVPACK (Bates et al., 1987) with the MATLAB interface gcvpackmat due to X. Xie, available from the home page http://pages.stat.wisc.edu/~xie/.

Fig. 12(a) shows a particular data set for \( n = 14^2 = 196 \) and \( \sigma = 0.1 \), together with the extreme undersmoothed spline \( f_{\lambda} \) for \( \lambda = 10^{-10} \), which is essentially the same as the interpolating spline \( f_0 \). Fig. 12(b) shows the same data set together with

![Fig. 12](image-url)

**Fig. 12.** Data \((x_i, y_i)\) (o) for function \( f(x) = \sin(2\pi(x_1 + x_2)/(x_1 + x_2 + 1/2)) \), \( \sigma = 0.1 \) and \( n = 196 \), together with (a) \( f_0 \) and (b) a smooth thin-plate spline estimate \( f_{\lambda} \).

Fig. 13. (a) Simulation estimates of \( \Pr(T(0^+) \leq 1.5 \min T) \) and (b) the median (+) and interquartile range of computed values of \( \min T/T(0^+) \) for \( f(x) = \sin(2\pi(x_1 + x_2)/(x_1 + x_2 + 1/2)) \) plotted against \( n \) for \( \sigma = 0.1 \) (bottom), \( \sigma = 0.01 \) (middle) and \( \sigma = 0.001 \) (top), using 1000 replicates for each point.

![Fig. 13](image-url)

**Fig. 13.** (a) Simulation estimates of \( \Pr(T(0^+) \leq 1.5 \min T) \) and (b) the median (+) and interquartile range of computed values of \( \min T/T(0^+) \) for \( f(x) = \sin(2\pi(x_1 + x_2)/(x_1 + x_2 + 1/2)) \) plotted against \( n \) for \( \sigma = 0.1 \) (bottom), \( \sigma = 0.01 \) (middle) and \( \sigma = 0.001 \) (top), using 1000 replicates for each point.
a much smoother spline \( f_s \) for which \( T(\lambda) = T(10^{-10}) \). Although this spline is slightly oversmoothing with respect to the prediction error, it is visually a much better fit than the extreme undersmoothed spline in Fig. 12(a). Clearly, the prediction error fails to discriminate the extreme undersmoothing here.

Fig. 13(a) shows, for each \( n \in \{4^2, 5^2, \ldots, 14^2\} \) and \( \sigma = 0.1, 0.01 \) and 0.001, an estimate of the measure \( \Pr(T(0^+)) < 1.5 \min T \) based on 1000 replicates of the error vector. Clearly, for \( \sigma = 0.1 \), the measure decays in a similar way to that in Fig. 8(a) except that it is significantly slower; in Fig. 13(a), the measure becomes close to 0 (to graphical accuracy) when \( n = 121 \) whereas in Fig. 8(a) it occurs when \( n = 26 \). From Fig. 13(a), for \( \sigma = 0.01 \) and \( \sigma = 0.001 \), the measure has not even begun to decrease from 1 for this range of \( n \). These plots are consistent with the asymptotic results discussed above.

For the same replicates, Fig. 13(b) shows, for each value of \( \sigma \), the median and interquartile range of \( \min T / T(0^+) \) plotted against \( n \). Consistent with Fig. 13(a), the median decreases with \( n \) but slower than in the corresponding plot in Fig. 9(a). Note that, for \( n = 196 \) and \( \sigma = 0.1 \), the median of 0.34 (corresponding to \( D = 2.9 \)) is large enough that there are cases where the prediction error clearly fails to discriminate extreme undersmoothing. Clearly, the prediction error is visually a much better fit than the extreme undersmoothed spline in Fig. 12(a). Clearly, the prediction error clearly fails to discriminate extreme undersmoothing.

Define the Sobolev error to be

\[
\begin{align*}
W(\lambda) &= \int (f_s - f)^2 \, dG + \|f_s - f\|^2_{m}, \\
&= n^{-1} \sum_{i=1}^{n} \tau_i (U^T \varepsilon)^2 + \sum_{m=1}^{M} C^{2m/d} \|U^T \varepsilon\|^2_i,
\end{align*}
\]

where \( G \) is the limiting distribution function for the design points. As in the univariate situation, for thin-plate splines, the Sobolev error provides significantly better discrimination of extreme undersmoothing than the prediction error. To see this, let \( W(\lambda) = T(\lambda) + \|f_s - f\|^2_{m} \), where \( f_s \) is the thin-plate spline interpolating the function \( f \) at the design points. Assume that \( f_s \) is sufficiently close to \( f \) so that \( \|f_s - f\|^2_{m} \) is small compared to \( E[f_0 - f_{\lambda}]^2 \). Then, since \( E[f_0 - f_{\lambda}] \), the term \( \|f_s - f\|^2_{m} \) in (41) is well approximated by \( |f_s - f|^2_{m} \) as \( \lambda \to 0 \), and the latter expression satisfies (Uterras, 1988)

\[
|f_s - f|^2_{m} = (n\lambda)^{-1} e^{T} (A(\lambda) - A^2(\lambda)) e
\]

where \( A(\lambda) \) is the smoothing matrix and the \( \tau_i \) are the associated eigenvalues with orthogonal matrix \( U \) of eigenvectors. Here \( \tau_i = \lambda_i \), for \( i = 1, \ldots, M \), where \( M = \left( \frac{d}{m} \right) - 1 \), and, under suitable conditions on \( \Omega \), the non-zero eigenvalues satisfy \( \tau_i \approx \frac{i^2}{m} \) for \( i \geq M + 1 \) as \( n \to \infty \) (Uterras, 1988). Therefore, as in (28),

\[
W(0^+) \approx W(0) \approx n^{-1} \sum_{i=1}^{M} (U^T \varepsilon)^2 + n^{-1} \sum_{m=1}^{M} C^{2m/d} \|U^T \varepsilon\|^2_i
\]

for some constant \( C > 0 \), and the right-hand side of (43) is a weighted sum of i.i.d. chi-square variables. Since \( \|U^T \varepsilon\|^2_i \) is increasing, \( EW(0^+) = \sigma^2 = ET(\lambda^*) \) for all sufficiently large \( n \).

Let \( P_{W(0^+)} \), \( \Pr(W(0^+) / \min EW \leq 0) \), be known (Cox, 1984b) that, for \( m > 3d/2 \), if \( \varepsilon \in V_{2m,2} \) and \( \sigma \) satisfies certain “natural” boundary conditions, then \( \min ET = O(n)^{2m/(4m + d)} \) and \( \min EW = O(n)^{2m/(4m + d)} \) as \( n \to \infty \). Then, assuming that \( \min ET \approx (\sigma^2 / n)^{2m/(4m + d)} \), we have that \( \min EW \leq H(n)^{1/2} \min ET \), where \( q = 2m/(4m + d) \) and \( H \) is a constant. Combining this with (43) and using a stochastic inequality as in the proof of Theorem 4 yields

\[
P_{W(0^*)} \leq \Pr \left( T(0^+) / \min ET \leq DH(n)^{1/2} \prod_{m=1}^{M} C^{2m/d} - 1/n \right)
\]

Estimation of the right-hand side of (44) (as in the proof of Theorem 4) shows that, asymptotically, the Sobolev error performs significantly better than the prediction error in discriminating extreme undersmoothing by thin-plate splines.

Under suitable conditions, similar results hold for multivariate smoothing splines that are defined by (38) and (39) except that the integration is over a bounded domain \( \Omega \). This is because the associated eigenvalues have the same behavior and \( ET(\lambda) \) and \( EW(\lambda) \) behave in the same way as above (Cox, 1984b; Uterras, 1987, 1988).

Appendix

Proof of Theorem 1. Under the given assumptions, the estimates (7) and (8) of \( ET(\lambda) \) and \( \lambda ET \) hold (Lukas et al., 2012, Sec. 3). It is well known that the smoothing matrix \( A = A(\lambda) \) can be diagonalized as \( A = U \text{diag}(\sigma_{ij}) U^T \), where \( U \) is orthogonal and \( a_{ij} = 1/(1 + \lambda e_i) \) for some nondecreasing sequence \( \lambda_i \), with \( \lambda_1 = \lambda_2 = 0 \). Let \( b_{ij} = 1 - a_{ij} \). Then the prediction error can be written as

\[
T(\lambda) = n^{-1} \|f - Ay\|^2 = n^{-1} \sum_{i} b_{ii} (U^T f_i - a_{ii} (U^T \varepsilon_i))^2.
\]
Let \( \tilde{T}(y) = n^{-1} \sum_{i,j} -2a_{ij}b_{ij}(U^Tf_j)(U^Te_i) + a_{ii}^2(U^Te_i)^2 \) which, from the expansion of the right-hand side of (45), is the random part of \( \tilde{T}(y) \). Then

\[
E(T^2(y)) = n^{-2} \sum_{i,j} -2a_{ij}b_{ij}(U^Tf_j)(U^Te_i) + a_{ii}^2(U^Te_i)^2
\]

\[
= 4n^{-2} \sum_{i,j} a_{ij}^2b_{ij}^2(U^Tf_j)^2(U^Te_i)^2 + n^{-2} \sum_{i,j} a_{ij}^2a_{ii}^2(U^Te_i)^2(U^Te_i)^2 + n^{-2} \sum_{i,j} a_{ij}^2E(U^Te_i)^4
\]

\[
= 4\sigma^2n^{-2} \sum_{i,j} a_{ij}^2b_{ij}^2(U^Tf_j)^2 + \sigma^4n^{-2} \sum_{i,j} a_{ij}^2a_{ii}^2 + 3\sigma^4n^{-2} \sum_{i} a_{ii}^4
\]

where we have used the fact that \( E(U^Te_i)(U^Te_i) = 0 \) for \( i \neq j \). \( E(U^Te_i)^2 = \sigma^2 \), \( E(U^Te_i)^4 = 3\sigma^4 \). Clearly

\[
E(T(y)) = n^{-1} \sum_{i} a_{ii}^2,
\]

so the variance of \( T(y) \) is

\[
Var(T(y)) = E(T^2(y)) - (E(T(y))^2 = 4\sigma^2n^{-2} \sum_{i,j} a_{ij}^2b_{ij}^2(U^Tf_j)^2 + 2\sigma^4n^{-2} \sum_{i} a_{ii}^4,
\]

The derivative of \( Var(T(y)) \) can be written as

\[
\frac{d}{d\lambda}(Var(T(y))) = 8\sigma^2n^{-2}\lambda^{-1} \sum_{i,j} a_{ij}^2b_{ij}^2(U^Tf_j)^2 - 16\sigma^2n^{-2}\lambda^{-2} \sum_{i,j} a_{ij}^2b_{ij}^3 - 8\sigma^2n^{-2}\lambda^{-1} \sum_{i} a_{ii}^3b_{ij}
\]

The first term on the right-hand side of (47) can be expressed in terms of the squared bias \( b^2(\lambda) \) and, therefore, it can be estimated as

\[
8\sigma^2n^{-2}\lambda^{-1} \sum_{i} b_{ii}^2(U^Tf_j)^2 = 8\sigma^2n^{-2}\lambda^{-1} b^2(\lambda) = 8\sigma^2n^{-2}\lambda^{-1} (1 + o(1)).
\]

Using Theorem 2.3 in Kou (2003), the second term on the right-hand side of (47) can be estimated as

\[
8\sigma^4n^{-2}\lambda^{-2} \sum_{i} a_{ii}^3b_{ij} = 8\sigma^4n^{-2}(4\pi)^{-1}B(4 - 1/4, 1 + 1/4)\lambda^{-1}1/4 (1 + o(1)),
\]

where \( B \) is the beta function \( B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y) \). (Note that the parameter \( \lambda \) used in Kou (2003) corresponds to the value \( n\lambda \) here.) From the estimates in (48) and (49), it follows that \( (d/d\lambda)(Var(T(y))) < 0 \) if

\[
8\sigma^2n^{-2}\lambda^{-1} (1 + o(1)) < 8\sigma^4n^{-2}(4\pi)^{-1}B(4 - 1/4, 1 + 1/4)\lambda^{-1}1/4 (1 + o(1)),
\]

which holds if \( \lambda < \lambda_0 \approx \lambda_0 \approx (\pi^2n^{-1}1/4(4\pi + 1)). \)

For the second part of the theorem, from (46) and (5), and using Theorem 2.3 in Kou (2003) and (8), we obtain, as \( n \to \infty \),

\[
\frac{Var(T(n\lambda_0))}{Var(T(0^+))} < (2/\sigma^2)b^2(\lambda_0) + n^{-1} \sum_{i} a_{ii}^4
\]

\[
= (2/\sigma^2)b^2(\lambda_0) + n^{-1}(4\pi)^{-1}B(4 - 1/4, 1 + 1/4)\lambda_0^{-1}1/4 (1 + o(1))
\]

\[
= \frac{C}{\sigma^2} \left( \frac{\pi^2}{2} \right)^{1/4} \frac{1}{(1 + o(1))},
\]

where \( C = 2C^2x^{4}(4\pi + 1) + C_1C_2^{-1/4} (4\pi + 1) \) with \( C_1 = (4\pi)^{-1}B(4 - 1/4, 1 + 1/4) = \sqrt{2}/512 \) and \( C_2 = 5/24\pi \). This completes the proof.

**Proof of Theorem.** It is known (Abramowitz and Stegun, 1984, 6.5.32) that \( \gamma(a, z) \) satisfies the asymptotic expansion

\[
1 - \frac{\gamma(a, z)}{\Gamma(a)} \sim z^{a-1}e^{-z} \left( \frac{1}{1 + (a-1)(a-2)} \right)
\]

for fixed \( a \) as the real variable \( z \to \infty \). Substituting \( a = n/2 \) and \( z = x(n, z) \) yields the result.

**Proof of Theorem 3.** To derive (15), we use the convergent and asymptotic expansion (Ferreira et al., 2005, Section 3)

\[
\gamma(a + 1, z) = e^{-z}z^{a+1} \sum_{k=0}^{\infty} c_k(a)\Phi_k(z, a),
\]

where \( c_0 = 1, c_1 = 0 \) and \( c_{k+1}(a) = [c_k(a) - ac_k(a)]/(k + 1) \) for \( k \geq 1 \), and \( \Phi_0(z, a) = (1 - e^{-a})/(a - z) \) and \( \Phi_k(z, a) = [e^{-a} - k\Phi_{k-1}(z, a)]/(z - a) \) for \( k \geq 1 \). The expansion is asymptotic provided that (for real variables \( a \) and \( z \) \( a \to \infty \) and \( z \to \infty \) with \( z < a \) and \( a^{1/2}/(a-z) \to 0 \) (so \( z \) and \( a \) have to be sufficiently far apart). The expansion will be used with \( a + 1 = n/2 \) and
\[ z = x(n, \sigma)/2 = (C/2)n^\beta, \] which is valid, since \( z < a \) for all sufficiently large \( n \) and
\[ \delta = \frac{a}{a-z} = \frac{(n/2 - 1)^{1/2}}{n/2 - 1 - (C/2)n^\beta} \sim \sqrt{2n^{-1/2} - 1 - Cn^\beta} \to 0 \] as \( n \to \infty \). Using the first term of the sum in (50) and the asymptotic estimate (DLMF, 2011, 5.11.7)
\[ I(a+1) \sim \sqrt{2\pi e^{-a}a^{1/2}}, \]
it follows that
\[ F(x; n) = \frac{\gamma(n/2, x/2)}{I(n/2)} = (1/2) \text{erfc} \left[ -\eta(a/2)^{1/2} \right] - R_\eta(a), \]
where \( a = n/2 \) and \( \eta = \text{sign}(\lambda - 1)[2(\lambda - 1 - \ln \lambda)]^{1/2} \) for \( \lambda = (x/2)/a = x/n \). Here
\[ R_\eta(a) \sim (2\pi^{-1/2})^{-1/2} \exp \left[ - \left( 1 - \beta \right) \ln n - (1 + \ln C) n/2 \right] \]
as \( a \to \infty \), where \( c_0(\eta) = 1/(\lambda - 1) - 1/\eta \) and the \( c_k(\eta) \) are defined by a known recurrence relation. This expansion is valid for all \( x \geq 0 \) (including at the transition point \( x = n \)) and it can be applied when both \( a \to \infty \) and \( \lambda \to \infty \) (Temme, 1979). For \( x = x(n, \sigma) \), clearly \( \lambda = Cn^\beta \to 0 \) as \( n \to \infty \) and \( -\eta(a/2)^{1/2} = (\pi n/2)^{1/2} \), where \( \pi = (1 - \beta) \ln n - (1 + \ln C + Cn^\beta) \). Then, using the first term of (53) in (52) yields (17).

**Proof of Theorem 4.** From (29), a stochastic inequality for weighted sums of i.i.d. chi-square variables (Okamoto, 1960; Yu, 2011) yields
\[ P_{\mathcal{W}(a)} \leq \text{Pr} \left( \frac{\prod_{i=1}^{n} Y_i + \sum_{i=m+1}^{n} \kappa Y_i}{Y_i} \leq y \right) \leq \text{Pr} \left( \left( \prod_{i=m+1}^{n} \kappa Y_i \right)^{1/n} \sum_{i=1}^{n} Y_i \leq y \right) \]
where \( y = (n/\sigma^2)D \min(\mathcal{E}W) \) and \( X = \sum_{i=1}^{n} Y_i \) is a chi-square variable with \( n \) degrees of freedom. Using (4), it follows from (54) that
\[ P_{\mathcal{W}(a)} \leq \text{Pr} \left( \left( T(1/n) \min(\mathcal{E}W) \right) \leq D \left( \min(\mathcal{E}W)/\min(\mathcal{E}W) \right) \left( \prod_{i=m+1}^{n} \kappa Y_i \right)^{-1/n} \right), \]
The product \( \prod_{i=1}^{n} \kappa Y_i \) in (55) can be bounded below using (23). Then, substituting the asymptotic estimates for \( \min(\mathcal{E}W) \) and \( n \) yields (30) into (55) and identifying the factors involving \( \sigma \) and \( n \) yields (30)–(32). The asymptotic estimate of \( r_n \) in (31) follows easily from Stirling’s formula \( k! \sim (2\pi k)^{1/2}(k/e)^k \).

By the same reasoning, we also have \( P_{\mathcal{W}(a)} \leq \text{Pr}(X(n, \sigma) \leq C(n, \sigma)) \), \( C(n, \sigma) = \mathcal{C}n^p \). Here \( C = CH_\sigma - 2q = 1(\pi/2)^{2m}(1 + \epsilon) \), where \( C \) is defined in (14), and \( \beta = -2m + q + \bar{\beta} \), where \( \beta = 1/(2m + 1) \) as in Theorem 3. The probability \( \text{Pr}(X \leq X(n, \sigma)) = F(\mathcal{X}(n, \sigma), n) \) can be estimated using the same asymptotic expansion as obtained (15). With \( a + n = n/2 \) and \( \mathcal{X} = X(n, \sigma) \), clearly \( a/2(1/2 - n) \sim \delta_n \) as \( n \to \infty \), where \( \delta_n \) is defined in (51), since \( \mathcal{P} < 1 \). Hence \( F(\mathcal{X}(n, \sigma), n) \) has the same asymptotic behavior as in (15), but with \( C \) and \( \mathcal{P} \) in place of \( C \) and \( \beta \), respectively. Substituting the asymptotic estimates into \( F(\mathcal{X}(n, \sigma), n) \) and simplifying yields the bound in (33).

**References**


**Please cite this article as:** Lukas, M.A., Performance criteria and discrimination of extreme undersmoothing in nonparametric regression. Journal of Statistical Planning and Inference (2014), http://dx.doi.org/10.1016/j.jspi.2014.05.006

Please cite this article as: Lukas, M.A., Performance criteria and discrimination of extreme undersmoothing in nonparametric regression. Journal of Statistical Planning and Inference (2014), http://dx.doi.org/10.1016/j.jspi.2014.05.006