The Role of Pseudo Data for Robust Smoothing with Application to Wavelet Regression

Hee-Seok Oh
Department of Statistics
Seoul National University
Seoul, Korea
heeseok@stats.snu.ac.kr

Doug Nychka
Geophysical Statistics Project
National Center for Atmospheric Research
Boulder, CO 80307, U.S.A.
nychka@ucar.edu

Thomas C. M. Lee
Department of Statistics
Colorado State University
Fort Collins, CO 80523-1877, U.S.A.
tlee@stat.colostate.edu

September 6, 2005

Abstract

This paper proposes a robust curve and surface estimate based on M-type estimators and penalty based smoothing. This approach also includes an application to wavelet regression. The concept of pseudo data, a transformation of the robust additive model to one with bounded errors, is used to derive some theoretical properties and also motivate a computational algorithm. The resulting algorithm, termed the ES–algorithm, is computationally fast, simple to describe and easy to implement. It can be extended to other settings such as irregularly spaced data and image denoising. Moreover, results from a simulation study and real data examples demonstrate the promising empirical properties of the procedure.

Keywords: ES–algorithm, M–estimation, Pseudo data, Regularization, Penalized least squares, Robust smoothing, Wavelets.

Abbreviated Title: Robust Smoothing by Pseudo Data

1 Introduction

In this paper we study the problem of robust nonparametric smoothing. Our emphasis is on M–type penalized smoothing including wavelet thresholding methods. This class of estimators are minimizers of a criterion that balances fidelity to the data with smoothness in the estimate. Expressing the estimate as the solution to a variational problem is a flexible paradigm and accommodates both a frequentist and Bayesian interpretation. A major contribution of this paper is a robust penalized regression algorithm that is derived based on the concept of *pseudo data*. There are a variety of penalized smoothing estimators where the data enters the penalty through a Gaussian log likelihood; i.e., a least squares type penalty for goodness-of-fit. Moreover there are many instances where both asymptotic theory and efficient computational algorithms have been developed for such least squares smoothers. Pseudo data and its sample based analog are useful because they transform the robust smoothing problem into a more conventional least squares smoothing. We use pseudo data to develop some asymptotic theory for robust estimators that inherits the properties of a least squares analog. This provides a simple way to infer asymptotic properties of the robust smoother based on theory for the least squares case. We also use *empirical pseudo data* to develop an iterative algorithm based on a sequence of least squares smoothers. This takes advantage of existing algorithms for computing least squares estimators and so simplifies implementation of a robust estimator. We term this iterative algorithm the ES–algorithm.

One significant application of empirical pseudo data is to wavelet type estimators. Unlike most existing methods, this does not involve a complex, non–linear optimization and inherits the computationally efficient algorithms associated with wavelets. In contrast to other proposals for robust wavelet-type regression our procedure is simple to describe, straightforward to implement, and it can be easily extended to other settings, such as irregularly spaced data or higher dimension data such as images. Our approach is also appropriate for more conventional smoothers where the roughness penalty is based on an inner product, or prior covariance and in this later situation the estimators can be interpreted as robust Kriging estimators.

1.1 Previous work

Many robust smoothing procedures have been proposed in the literature. Previous works for M-type smoothing include Huber (1979), Cox (1983), Härdle and Gasser (1984), Silverman (1985), and Hall and Jones (1990). In particular, Cox (1983) introduces a theoretical construct for smoothing problems, pseudo data, and demonstrates that a nonlinear M–type smoothing spline can be approximated by a linear smoothing spline computed from such pseudo data. We extend Cox's theoretical results to include roughness penalties based on reproducing kernel norms. The new robust wavelet thresholding method proposed by this paper is also motivated by this pseudo data approach. As another approach, Cleveland (1979) proposes locally weighted scatter plot smoothing (LOWESS) which is a resistant method based on local polynomial fits.

The subject of robust wavelet thresholding has also been previously studied. Kovac and Silverman (2000) propose the following procedure. First a statistical test is applied to identify outliers. Then these outliers are removed from the data. The last step is to apply a thresholding procedure for irregularly spaced data to estimate the regression function. However, this method has the disadvantage of losing potential information by removing observations. More recently, Sardy et al. (2001) propose a robust M–type wavelet denoising procedure. This procedure is computationally intensive, as it involves solving a nontrivial nonlinear optimization problem. Finally, Averkamp and Houdré (2003) extend the minimax theory for wavelet thresholding to some broader classes of *known* symmetric heavy tail noises. This method, therefore, is not always applicable in practice as the noise distribution is usually unknown.

1.2 Outline of the paper

The rest of the paper is organized as follows. In Section 2, we first review robust penalized least squares and M–type smoothing. Then we present the concept of pseudo data, which organizes our statistical approach. The material in this section can be easily extended to other smoothers that can be formulated as a penalized least squares problem. Section 3 presents some theoretical results of the proposed method when the roughness penalty is derived from an inner product (or quadratic form). Section 4 follows with the introduction of empirical pseudo data and suggests an iterative algorithm for computing the robust estimator. To investigate the practical performance of the proposed method, in Section 5 the method is examined in a simulation study and real data examples. Section 6 extends the method to the irregularly spaced and two–dimensional settings. Lastly, concluding remarks are offered in Section 7.

2 Penalized M-type smoothing

Given n pairs of observations (x_i, y_i), $i = 1, \ldots, n$ we assume an additive model satisfying

$$y_i = g(x_i) + \epsilon_i, \tag{1}$$

where the ϵ_i's are independent and identically distributed random errors and g is an unknown smooth function of interest. The distribution of the errors can potentially be heavy tailed and motivates the need for a robust estimator. We will also assume that x_i are distinct and without loss of generality restricted to the unit interval. Although through most of the discussion will treat g as being one a dimensional function, the extension to higher dimensions is straightforward.

2.1 Penalized least squares smoothers

Given a nonnegative penalty function J, a penalized least square estimate $\hat{g}$ is the function that minimizes

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + J(f) \tag{2}$$

over f such that $J(f) < \infty$. One classical example of this approach is a cubic smoothing spline, where $J(f) = \lambda \int (f'')^2 dx$. In many cases although the penalized estimator is formally a minimizer over a function space, it has a solution that is finite dimensional. This includes both splines and geostatistical estimators. In addition wavelet estimators also fall into the category of having a finite dimensional representation although some wavelet penalty functions cannot be derived from an inner product.

Thus, to simplify the discussion we will assume a penalized estimator that can be expressed as a finite linear combination of n basis functions $\{\psi_l\}_{l=1}^n$. In this case, we will prescribe that the minimizer has the form $f(x) = \sum_{l=1}^n \theta_l \psi_l(x)$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ are basis coefficients. Now let $W_{i,l} = \psi_l(x_i)$ be a matrix of the basis functions evaluated at the observations and $\boldsymbol{f} = \{f(x_1), \ldots, f(x_n)\}$. Then $\boldsymbol{f} = W\boldsymbol{\theta}$ is the prediction of the function at the observations for a given vector of coefficients. In view of this linear relationship it is equivalent to parameterize this problem in term of the vector $\boldsymbol{f}$ – one can always recover the basis coefficients and evaluate the function at arbitrary points. In other terms given the prediction vector $\boldsymbol{f}$ one can recover the entire function by interpolating these values with the specified basis. Based on these points we will focus on the penalized least squares problem:

$$\min_{\boldsymbol{f} \in \Re^n} \sum_{i=1}^{n} (y_i - \boldsymbol{f}_i)^2 + p_\lambda(\boldsymbol{f}). \tag{3}$$

Here $p_\lambda(\cdot)$ denotes a general form of roughness penalty with a positive, free parameter λ. Under the assumption that $p_\lambda(\boldsymbol{f})$ is convex and differentiable the estimator can be characterized by its score equations. Let $\Psi(\boldsymbol{f})$ be the gradient of the penalized least squares criterion in (3). Then

$$\Psi(\boldsymbol{f})_i = -2(y_i - \boldsymbol{f}_i) + \frac{\partial p_\lambda(\boldsymbol{f})}{\partial \boldsymbol{f}_i} \tag{4}$$

and so the least squares, penalized estimator is given by $\Psi(\hat{\boldsymbol{g}}) = 0$.

Why is this estimate a smoother? A common penalty is the quadratic form $p_\lambda(\boldsymbol{f}) = \lambda \boldsymbol{f}^T R \boldsymbol{f}$ where R is a nonnegative definite matrix derived from a reproducing kernel and $\lambda > 0$. More details on this choice will be given in the next section. However, with this form one can verify that $\hat{\boldsymbol{g}} = (I + \lambda R)^{-1}\boldsymbol{y}$. It is simple to show that $(I+\lambda R)^{-1}$ must have eigenvalues in [0,1] and so will be a smoothing matrix where λ, known as the smoothing parameter, controls the size of the eigenvalues. In addition various nonparametric estimates such as wavelet shrinkage can be also expressed in this form. For example, the soft-thresholding rule is equivalent to a penalty $p_\lambda(\boldsymbol{f}) = \lambda \sum_{i=1}^n |\theta_i|$ with $\boldsymbol{\theta} = W^{-1}\boldsymbol{f}$ and λ controls the threshold level. Better aligned with the goals of our work, Antoniadis and Fan (2001) consider a class of penalties that are differentiable and provide a range of estimators between the standard hard and soft threshold wavelet estimators.

2.2 Robust penalized smoothers

It is well known that the minimizer of (3) is sensitive to the presence of outliers (Simonoff, 1996). To avoid this problem when outliers may be encountered, one can replace the sum of squares in (3) by a robust M-type penalty. The robust estimate of $g(x)$, $\hat{\boldsymbol{g}}$, is the minimizer of

$$\sum_{i=1}^n \rho(y_i - \boldsymbol{f}_i) + p_\lambda(\boldsymbol{f}). \tag{5}$$

over $\boldsymbol{f} \in \Re^n$. The function $\rho(t)$ is typically convex and symmetric about zero, quadratic in the neighborhood of zero and increasing at a rate slower than t^2 for t large. The robust feature is that compared to squared errors $\rho(t)$ downweights extreme residuals. Given the convexity of ρ and a convex penalty p, the existence of a unique minimizer follows from Proposition 2.1 in Cox (1983). Under the assumption that both ρ and p are differentiable and convex, the robust estimator is characterized by its score equation. Taking the derivative of (5) with respect to $\boldsymbol{f}$ one obtains the gradient vector that we will denote as $\Phi(\boldsymbol{f})$ and a robust estimator will be the root of $\Phi(\hat{\boldsymbol{g}}) = 0$. Specifically, with $\eta = \rho'$ one obtains

$$\Phi(\boldsymbol{f})_i = -\eta(y_i - \boldsymbol{f}_i) + \frac{\partial p_\lambda(\boldsymbol{f})}{\partial \boldsymbol{f}_i}. \tag{6}$$

A common choice of ρ is the Huber loss function which is a continuous function constructed from quadratic and piecewise linear segments:

$$\rho(t) = \begin{cases} t^2 & \text{if } |t| \leq C \\ C(2|t| - C) & \text{otherwise.} \end{cases}$$

For theoretical results it is useful for ρ to have several derivatives and for this purpose one can use the the log of the *cosh* function because it is both smooth and is also qualitatively similar to the Huber function. Due to the nonlinear nature of ρ, the minimization of (5) is often difficult. Also the theoretical properties for this estimator are more difficult to analyze than penalized least squares. In next section, we introduce the idea of pseudo data that in turn leads to some theory to understand this estimator and motivates an efficient algorithm for computing the estimate.

2.3 Pseudo data

Let $\eta = \rho'$ be the (almost everywhere) derivative of ρ, and define *pseudo data* $\tilde{y}_i$ as

$$\tilde{y}_i = g(x_i) + \frac{\eta(\epsilon_i)}{2}. \tag{7}$$

In this discussion we assume that η will be bounded and $\eta(\epsilon_i)$ will have a finite variance. This will hold for the Huber ρ because ρ is constant beyond the range C. Of course in practice the pseudo data $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)^T$ cannot be calculated as it involves the unknown function. However, this transformation facilitates a theoretical analysis and leads to a practical computing algorithm for a robust wavelet smoother.

In the context of robust smoothing spline regression, this concept of pseudo data has been applied to derive limit results and rates of convergence, and to develop automatic smoothing parameter selection methods (Cox, 1983; Cantoni and Ronchetti, 2001; Oh and Nychka, 2005). The key here is that (7) defines an alternative additive model for g where the distribution of the random component is light tailed and so it fits into a standard smoothing problem. The path breaking result of Cox (1983) is that a robust cubic smoothing spline applied to the data $\boldsymbol{y}$ is asymptotically equivalent to a least squares spline applied to the pseudo data $\tilde{\boldsymbol{y}}$. The next subsection describes the class of roughness penalties that we use to generalize Cox's work.

2.3.1 Roughness penalties from reproducing kernels

For the theoretical analysis we will focus on a roughness penalty $J(f)$ that is derived from a reproducing kernel norm. Let $K(x, y)$ denote a symmetric, positive definite kernel. Following the functional analytic formalism for example in Aubin (2000) one can always define an inner product $\langle \cdot, \cdot \rangle$ on a Hilbert space of functions, $\mathcal{H}$ such that $K(x, \cdot) \in \mathcal{H}$ for all x and $\langle K(x, \cdot), K(x', \cdot) \rangle = K(x, x')$. The later expression is the well known "reproducing" property of the kernel with respect to the inner product and for this reason $\mathcal{H}$ is termed a reproducing kernel Hilbert space (RKHS). Given a RKHS we assume the penalty on functions given by $J(f) = \langle f, f \rangle \equiv \|f\|_{\mathcal{H}}$. It is also a well known result that the solution to minimizing (2) will be unique and can be represented using a finite set of basis functions with $\psi_i(x) = K(x, x_i)$ and from (3) $W_{i,j} = K(x_i, x_j)$. Switching to the finite parameterization in terms of $\boldsymbol{f}$, $p_\lambda(\boldsymbol{f}) = \lambda \boldsymbol{f}^T R \boldsymbol{f}$ where $R = W^{-1}$. This surprisingly simple form for the penalty in terms of the basis functions is a standard result for spline type estimators and is a direct consequence of the reproducing property of the kernel under the inner product.

There is also an equivalence between the penalized estimators using a reproducing kernel roughness penalty and geostatistical estimators. If one assumes that g is a mean zero Gaussian process with covariance function proportional to K and the errors have finite variance then $\hat{\boldsymbol{f}}$ obtained as the minimizer of (3) is also the best linear unbiased predictor of g – the standard Kriging estimator used in the analysis of spatial data. If one makes the additional assumption that the errors are normally distributed then $\hat{\boldsymbol{f}}$ is also the conditional expectation of g given the observations. Following this geostatistical interpretation, the smoothing parameter, λ is the ratio of the measurement error variance (or nugget variance) to the variogram sill parameter.

3 Some asymptotic theory

Here we present a theoretical result which states that, under certain conditions, an M-type robust smoother estimator defined as the minimizer of (5) can be well approximated by the minimizer of

$$\sum_{i=1}^{n} (\tilde{y}_i - \boldsymbol{f}_i)^2 + p_\lambda(\boldsymbol{f}),$$

where $\tilde{\boldsymbol{y}}$ is pseudo data defined in (7). This result can be considered as a generalization of Huber's asymptotic linearization of robust linear regression estimates (Huber, 1973), or an extension of Cox's asymptotic linearization of robust smoothing splines (Cox, 1983). For simplicity of analysis, we now focus on roughness penalties that arise from reproducing kernels and so are quadratic forms in $\boldsymbol{f}$. Besides splines and geostatistical estimators this also includes a linear version of wavelet estimators that have been studied by several authors (Antoniadis, 1996; Dechevsky and Penev, 1999; Angelini et al., 2000).

3.1 Theorem for asymptotic equivalence

We will use throughout the Euclidean norm: for $\boldsymbol{x} \in \Re^n$, $||\boldsymbol{x}||^2 = \sum_{i=1}^{n}(x_i^2)$ and the normalized version : $||\boldsymbol{x}||_n^2 = (1/n)||\boldsymbol{x}||^2$. The following five assumptions will be used for our main theoretical result.

Assumption 1. Under the additive model (1), $\{\epsilon_i\}$ are mean zero, independent and identically distributed random variables.

Assumption 2. $\eta \in C^2(-\infty,\infty)$ and (without loss of generality) η is normalized such that $E\{\eta(\epsilon_i)\} = 0$ and $E\{\eta^{'}(\epsilon_i)\}/2 = 1$.

Assumption 3. a) $\mathcal{H}$ is an RKHS on $[0,1]$ such that $g \in \mathcal{H}$. Let $\mathcal{C} = \{f \in \mathcal{H} : ||f||_{\mathcal{H}} \leq D\}$ and assume that $\mathcal{C}$ is compact with respect to L_2 norm.

b) Given the reproducing kernel K for $\mathcal{H}$, $p_\lambda(\boldsymbol{f}) = \lambda \boldsymbol{f}^T R \boldsymbol{f}$, where $R_{i,j}^{-1} = K(x_i, x_j)$.

Assumption 4. Let $\{\lambda_n\}$ be a sequence of smoothing parameters with corresponding penalized least squares estimators $\tilde{\boldsymbol{g}}_n$ based on pseudo data. Let $C_n = E||\tilde{\boldsymbol{g}}_n - \boldsymbol{g}||_n^2$ and assume that $C_n \to 0$ as $n \to \infty$.

Assumption 5. Let $A(\lambda) = (I + \lambda R)^{-1}$.

a) If $a_n = \max_j\{A(\lambda_n)_{j,j}\}$ then $a_n \to 0$ as $n \to \infty$.

b) There is an $M < \infty$ such that $tr(A(\lambda_n))/\lambda_n < M$ and $tr(A(\lambda_n))/n < M$ for all n.

Theorem 1

Under Assumptions 1 – 5,

$$||\tilde{\boldsymbol{g}} - \hat{\boldsymbol{g}}||_n/\sqrt{C_n} \longrightarrow 0$$

in probability as $n \to \infty$.

One interpretation of this result is that $\hat{\boldsymbol{g}}$ inherits the same asymptotic mean squared error properties as $\tilde{\boldsymbol{g}}$.

Theorem 1 can be proved by adapting the techniques in Theorem 3.1 of Cox (1983), which was tailored for cubic spline estimators. Cox's proof is itself an adaptation of original results of Huber (1973) on robust regression. We believe that original Cox's argument is incomplete, however, in that basic inequality must hold uniformly and we have modified the proof accordingly (Lemma 1). Unfortunately adding this important technical detail lengthens our argument.

3.2 Comment on assumptions

The assumptions 1-4 are standard for smoothing problems but are listed in full for completeness. The compactness condition in Assumption 3 may seem unusual but plays an important part in our analysis. However, based on classical embedding results for function spaces if $\mathcal{H}$ is contained in the space of continuous functions (Adams, 1975), then this assumption holds. Indeed, one usually assumes more smoothness than just continuity for $\mathcal{H}$ so Assumption 3 is not overly restrictive. Assumption 4 simply asserts the consistency of the least squares smoother based on pseudo data. As mentioned in the introduction it is not our intent to analyze penalized least squares smoothers. Rather, we focus on transferring the properties of the least squares case to the robust one.

Assumption 5 makes the most technical requirements of the smoother. The first condition, (5a) insures that the diagonal elements of the smoothing matrix $A(\lambda_n)$ converge to zero. This requirement can be justified for splines based on the equivalent kernel representations (Nychka, 1995 and Messer, 1991) and an assumption that the observation locations are approximately uniformly spaced. From a kernel smoothing perspective this assumption of the smoother matrix is also reasonable. The simplest kernel estimator would have diagonal elements of order $1/(nh_n)$ where h_n is the bandwidth and so require that $nh_n \to 0$ as $n \to \infty$. Assumption (5b) involves balancing the rates of the smoothing parameter with the effective degrees of freedom, $trA(\lambda_n)$ of the smoother. In particular based on the equivalent kernel theory in Nychka (1995) one expects that $trA(\lambda_n) \sim (\lambda_n/n)^{-1/2m}$ where m is the order of the spline and $\mathcal{H}$ is an mth order Sobolev space. The reader should be aware that the results of Nychka (1995) and Messer (1991) use a slightly different version of the smoothing parameter. Due to the normalization by $(1/n)$ in the sum of squares the "λ" appearing in these published works is equal to λ_n/n in this paper. With respect to condition (5b) we have, from Nychka (1995), $trA(\lambda_n)/\lambda_n \sim \lambda_n^{-(2m+1)/2m} n^{-1/2m}$ and so (5b) holds provided that $\lambda_n = O(n^{-2m/(2m+1)})$. This is the optimal rate one would choose when g is in $\mathcal{H}$ but has no additional smoothness. So from a heuristic analysis this condition does not constrain λ_n from achieving the optimal convergence rate with respect to mean squared error. If g has more smoothness than $\mathcal{H}$ then the optimal rate for λ_n is slower than $n^{-2m/(2m+1)}$ and so (5b) accommodates a wider range of convergence rates for the smoothing parameter that includes the optimal rate. As in the case of (5a) this condition has a simple analogy with kernel-type smoothers. Making the connection that the mth order spline will have an equivalent kernel with bandwidth $(\lambda_n/n)^{-1/2m}$ the asymptotic rates given above are easily deduced.

3.3 Proof of equivalence theorem

The proof has three basic parts. Finding uniform bounds on the score functions (4) and (6), applying a fixed point argument, and evaluating the pseudo data score function at the robust estimator. The basic idea behind the proof is to use the fact that the difference between the score functions for the robust estimator and the pseudo data, least squares estimator, $\Psi(\boldsymbol{f}) - \Phi(\boldsymbol{f})$ is small.

The first part of the proof is achieved by the content of Lemma 1 in the Appendix.

For the second part of the proof, define $\mathcal{C}$ with a radius so that $g \in \mathcal{C}$ and the sets $F_n = \{\boldsymbol{f} \in \mathcal{C} : ||\boldsymbol{f} - \boldsymbol{g}||_n \leq 4\sqrt{C_n}/\delta\}$. For convenience we abuse notation and the reader should interpret $\boldsymbol{f} \in \mathcal{C}$ to mean that the function interpolating the discrete set of function values $\boldsymbol{f}$ is contained in $\mathcal{C}$.

Now define the function from $\Re^n$ to $\Re^n$

$$U(\boldsymbol{x}) = \boldsymbol{x} - (1/2)A(\lambda)\Phi(\boldsymbol{x} + \boldsymbol{g}).$$

Given the existence of the fixed point $U(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}}$ then it is straight forward to verify that $\hat{\boldsymbol{x}} + \boldsymbol{g}$ is a solution to $\Phi(\hat{\boldsymbol{x}} + \boldsymbol{g}) = 0$. Thus, $\hat{\boldsymbol{x}} + \boldsymbol{g}$ is a robust estimator. We now apply the fixed point argument of Cox (1983) and Huber (1973). By Lemma 2 from the Appendix, for any $\delta > 0$ and n sufficiently large, U maps the compact, convex set, $F_n - \boldsymbol{g}$, unto itself with probability greater than $1 - \delta$. On this event, by Brouwer's fixed point theorem there must exist, at least one point $\hat{\boldsymbol{x}} \in F_n - \boldsymbol{g}$ such that $U(\hat{\boldsymbol{x}}) = \hat{\boldsymbol{x}}$. Thus a robust estimator must exist in this neighborhood of $\boldsymbol{g}$ with probability greater than $1 - \delta$.

The third step of the proof bounds the norm between the robust estimator and the estimator based on pseudo data. By definition, $\tilde{\boldsymbol{g}} = A(\lambda)\tilde{\boldsymbol{y}}$ and $\Phi(\hat{\boldsymbol{g}}) = 0$ using (4)

$$||A(\lambda)\{\Phi(\hat{\boldsymbol{g}}) - \Psi(\hat{\boldsymbol{g}})\}||_n = ||A(\lambda)\Psi(\hat{\boldsymbol{g}})||_n = ||(\hat{\boldsymbol{g}} - \tilde{\boldsymbol{g}})||_n.$$

¿From step 2 we have concluded that for n sufficiently large and with probability greater than $(1 - \delta)$, $\tilde{\boldsymbol{g}} \in F_n$. Now applying Lemma 1 with $B = 4\sqrt{C_n}/\delta$ (implying that $F \equiv F_n$) and $\epsilon < \delta^2/8$ we obtain the bound

$$||\hat{\boldsymbol{g}} - \tilde{\boldsymbol{g}}||_n \leq \left[2\epsilon/\delta + 2M\sqrt{C_n}/\delta^2\right]\sqrt{C_n} \leq \left[\delta/2 + 2M\sqrt{C_n}/\delta^2\right]\sqrt{C_n}$$

with probability greater than $1 - 2\delta$.

Finally choose n sufficiently large so the quantity in square brackets is less than δ. We have now shown that for any $\delta > 0$ there exists an $N < \infty$ such that for $n > N$ and with probability greater than $1-\delta$ there is a robust estimator in F_n and $||\hat{\boldsymbol{g}}-\tilde{\boldsymbol{g}}||_n \leq \delta\sqrt{C_n}$. These statements are equivalent to convergence in probability and the theorem now follows.

4 Empirical Pseudo Data

The concept of pseudo data suggests an equivalence between a robust estimator and an more conventional least squares method. We will exploit this connection to characterize robust wavelet shrinkage and motivate a practical algorithm. Ideally if one knew g one could form the pseudo data in (7) and apply a least squares estimator. Of course in practice g is unknown, so we consider instead a fixed point analogy to pseudo data. If $\hat{g}$ is an estimate of g, we form the empirical pseudo data (EPD)

$$z_i = \hat{g}(x_i) + \frac{\eta\{y_i - \hat{g}(x_i)\}}{2} \tag{8}$$

and consider the least squares penalized estimate based on $\boldsymbol{z}$. The ultimate estimate of interest is a fixed point where the estimate obtained from the EPD is the same as that used to construct it. The algorithm to achieve this is outlined below.

The ES–algorithm

1. Given an initial estimate $\hat{\boldsymbol{g}}^0$.
2. Loop on L until convergence.

   E-Step: (**E**valuation of EPD) Use (8) to form empirical pseudo data $\boldsymbol{z}^L$ based on $\hat{\boldsymbol{g}}^L$.

   S-Step: (**S**moothing of EPD) Obtain the least squares penalized estimator $\hat{\boldsymbol{g}}^{L+1}$ based on $\boldsymbol{z}^L$.

Assume that p_λ is convex and differentiable and for the moment that λ is fixed. Assume that the algorithm converges and let $\hat{\boldsymbol{g}}^\infty$ denote the estimator at convergence. This estimator will satisfy

$$-2(\boldsymbol{z}_k^\infty - \hat{\boldsymbol{g}}_k^\infty) + \left.\frac{\partial p_\lambda(\boldsymbol{f})}{\partial \boldsymbol{f}_k}\right|_{\boldsymbol{f}=\hat{\boldsymbol{g}}^\infty} = 0. \tag{9}$$

Examining (9) and noting that $(\boldsymbol{z}_k^\infty - \hat{\boldsymbol{g}}_k^\infty) = (1/2)\eta(y_k - \hat{\boldsymbol{g}_k^\infty})$ we see that this is the same system of equations that characterize the robust estimator in (6). Thus $\hat{\boldsymbol{g}}^\infty$ is in fact the solution to the robust penalized smoothing problem in (5).

We choose to implement this algorithm in a more general context. In the S-Step we also include the adaptive choice of the smoothing parameter, λ. For example, in the S-Step one could use a wavelet thresholding scheme to determine λ. If this more complicated algorithm converges, the final estimate has the following interpretation: form empirical pseudo data based on $\hat{\boldsymbol{g}}^{\infty}$ the least squares, and possibly adaptive estimator applied to the EPD will again be equal to $\hat{\boldsymbol{g}}^{\infty}$. An illustration of this use of the algorithm is given in the next section.

4.1 Implementation for robust wavelet regression

Although this discussion is phrased in terms of penalized smoothers in the S-Step, the above algorithm may have more practical generality. What is required in the S-Step is to obtain the least squares estimator and this may be done without resorting to the explicit penalized minimization. This feature can be illustrated by robust wavelet regression. Let W be a discrete and orthogonal wavelet transform matrix. A key step in classical wavelet regression is to estimate the true wavelet coefficients $\boldsymbol{\theta} = W\boldsymbol{f}$ by thresholding the empirical wavelet coefficients $\boldsymbol{d} = W\boldsymbol{y}$. It is known that such (non-robust) wavelet thresholding estimators are special cases of the penalized least squares estimators discussed in the previous section (e.g., see Antoniadis and Fan, 2001). That is, a thresholded estimate for $\boldsymbol{\theta}$ can be obtained as the minimizer of

$$\|\boldsymbol{d} - \boldsymbol{\theta}\|^2 + p_\lambda(\boldsymbol{\theta})$$

for some suitably chosen penalty function $p_\lambda(\boldsymbol{\theta})$ with penalty parameter λ. Given a penalty function $p(\cdot)$ which is a nonnegative, nondecreasing and differentiable function $(0, \infty)$, the solution to the minimization of the above problem exists and is unique (Antoniadis and Fan, 2001). This motivates the following ES–algorithm for performing robust wavelet thresholding. In the S-Step one applies a soft thresholding operation to the Lth iterative empirical wavelet coefficients $\boldsymbol{d}^L = W\boldsymbol{z}^L$ obtaining the Lth iterative estimate $\hat{\boldsymbol{\theta}}^L$ for $\boldsymbol{\theta}$ and $\hat{f}^L = W^T\hat{\boldsymbol{\theta}}^L$.

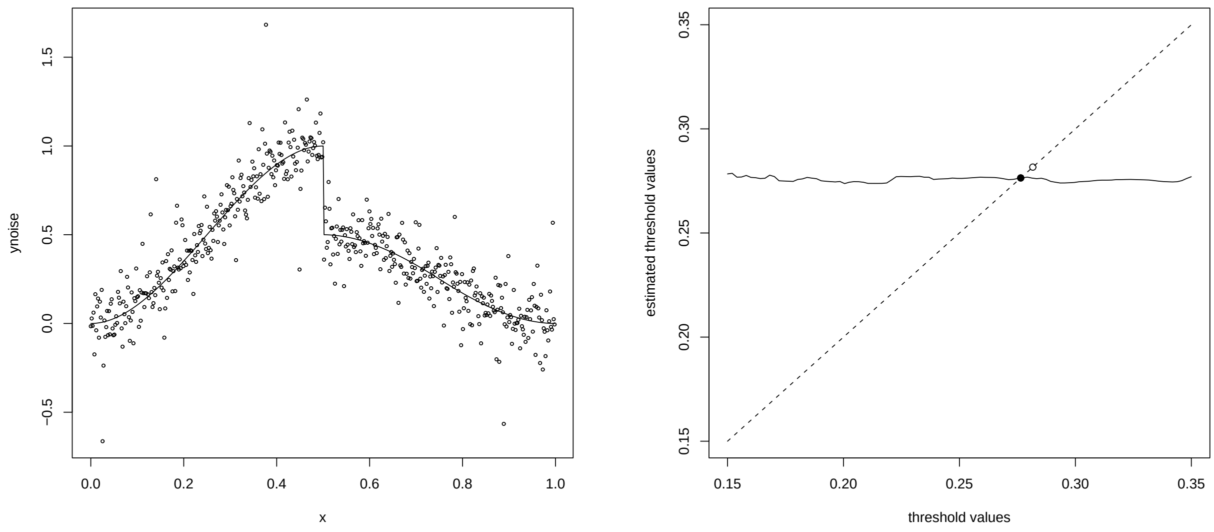
Note that the application of a soft thresholding operation in the S-Step of the above algorithm is equivalent to minimizing the following L_1-type penalized least squares criterion:

$$\mathcal{L}(\boldsymbol{\theta}) = \|\tilde{\boldsymbol{y}} - W^T\boldsymbol{\theta}\|^2 + \lambda \sum_{i=1}^{n} |\theta_i|.$$

This $\mathcal{L}(\boldsymbol{\theta})$ criterion has been studied extensively by Alliney and Ruzinsky (1994). In our case the matrix W^T has full rank and $\mathcal{L}(\boldsymbol{\theta})$ is a strictly convex functional of $\boldsymbol{\theta}$, it so follows that $\mathcal{L}(\boldsymbol{\theta})$ has a unique minimum. Moreover, by applying Theorem 4 of Alliney and Ruzinsky (1994), we conclude that the above ES–algorithm at convergence will also be the minimizer of $\mathcal{L}(\boldsymbol{\theta})$.

The choice of λ for the wavelet problem is equivalent to choosing a threshold and here we discuss the more practical variation when the threshold is determined adaptively from the data.

Figure 1: The choice of λ. In the ES-algorithm for wavelet regression, the circle denotes a starting threshold and the solid circle is the threshold converged after a few iteration.



We propose at each step of the S-Step to estimate the threshold based on the current version of the EPD. At convergence, as mentioned above, one obtains a fixed point where the adaptive estimator based on the EPD is the same as the estimator used to construct the EPD.

Our computational approach given this goal is efficient and we illustrate this concept by contrasting with a grid search over the threshold. Given the data set plotted in the left panel and a grid of λ over the interval $(0.15, 0.35)$ for each value of λ we have computed the robust estimator. We use the ES–algorithm to do this, of course, but that detail is not important. Each value of the smoothing parameter corresponds to an equivalent threshold and these are the values along the horizontal axis of the right panel in Figure 1. Moreover, for each of these estimators we formed the EPD and estimated a threshold based on the empirical Bayes method of Johnstone and Silverman (2005). These estimated thresholds are plotted on the vertical axis in Figure 1. Note that there is less variability among the estimated thresholds. Where the 45 degree line intersects this trace one obtains a threshold that is a fixed point. The algorithm outlined above where the threshold is estimated with each iteration of the S-Step will converge to this fixed point (solid circle) and so avoids the grid search illustrated in the right panel. Of course the algorithm may not converge in which case one might have to resort to a grid search to find the fixed point. One intriguing aspect of this example is that the estimated threshold based on EPD is not overly sensitive to the exact choice of robust estimator. In this example there are a range of robust estimators that will generate EPD that will give good threshold estimates.

In the ES–algorithm no particular smoothing parameter selection rule is required in the S-Step. Although we have had success with the empirical Bayes procedure of Johnstone and Silverman (2005), one could consider other thresholding methods to obtain $\hat{\boldsymbol{\theta}}^L$. In addition, the above algorithm can be easily extended to other settings such as higher dimension and irregularly spaced data; see Section 6. Lastly, the above algorithm converges very quickly based on our numerical experiments reported in the next section.

5 Practical performance

5.1 Simulation study

A simulation study was conducted to evaluate the practical performance of the ES–algorithm for robust wavelet regression. The experimental setup was essentially the same as in Sardy et al. (2001). The four test functions introduced by Donoho and Johnstone (1994) were used: *blocks*, *bumps*, *heavisine* and *doppler*. In this simulation study these functions were normalized so that $\int\{g(x)-\bar{g}\}^2 dx = 7^2$ where $\bar{g} = \int g(x)dx$. Altogether two different sample sizes, $N = 1024$ and $N = 4096$, and three different types of noise were considered:

- G: standard Gaussian noise $N(0,1)$,
- C: contaminated Gaussian Mixture $0.9N(0.1) + 0.1N(0,4^2)$, and
- T: t-distribution with three degrees of freedom.

For each combination of test function and noise, 100 samples were generated. Then for each generated sample, three regression estimates were obtained by applying, respectively,

- `EBayes`: the empirical Bayes thresholding method of Johnstone and Silverman (2005),
- `REBayes`: the proposed robust procedure with `EBayes` as the thresholding method, and
- `Med3`: a moving median filter with size 3.

Table 1: Averaged MSE values of various curve estimates for test functions *blocks* and *bumps*. Numbers in parentheses are estimated standard errors. Averaged MSE values for *RBPur* are reported from Sardy et al. (2001) and have not been computed and are noted in italics. Since these MSE values are averaged from a smaller number of repetitions, their estimated standard errors are of different orders.

		blocks			bumps		
N	method	G	C	T	G	C	T
1024	`EBayes`	27.7 (0.275)	111 (1.76)	135 (8.01)	36.1 (0.337)	128 (1.93)	158 (12.6)
	`REBayes`	27.0 (0.286)	41.1 (0.633)	47.8 (0.688)	117 (0.602)	139 (0.911)	142 (1.19)
	RBPur	*77 (2.31)*	*103 (3.09)*	*114 (3.42)*	*190 (5.70)*	*220 (6.60)*	*240 (7.20)*
	`Med3`	46.4 (0.352)	66.5 (0.698)	76.8 (0.793)	134 (0.582)	160 (0.917)	167 (1.18)
4096	`EBayes`	11.0 (0.0901)	90.0 (1.04)	124 (8.52)	13.4 (0.107)	94.6 (0.940)	111 (3.69)
	`REBayes`	13.6 (0.135)	22.7 (0.319)	26.2 (0.376)	20.8 (0.199)	30.7 (0.357)	35.5 (0.397)
	RBPur	*33 (0.990)*	*44 (1.32)*	*49 (1.47)*	*24 (0.720)*	*35 (1.05)*	*38 (1.14)*
	`Med3`	45.1 (0.150)	63.5 (0.312)	73.4 (0.390)	49.7 (0.168)	69.4 (0.331)	79.6 (0.360)

The cutoff C for the Huber function was taken as $C = \kappa\hat{\sigma}$, where $\hat{\sigma}$ is a robust estimate of the noise variance and κ is a positive constant usually chosen as $\kappa = 1.345$ (Huber, 1981). However, we shall follow the suggestion of Sardy et al. (2001) and use $\kappa = 2.0$. Also, estimates obtained from `Med3` were used as the initial regression estimates for `REBayes`.

The mean–squared–error $\text{MSE} = \frac{1}{N}\sum\{g(x_i) - \hat{g}(x_i)\}^2$ was then calculated for each regression estimate. The averaged MSEs and their estimated standard errors are listed in Tables 1 and 2. Also listed in these tables are the corresponding MSE values of the robust basis pursuit (*RBPur*) procedure of Sardy et al. (2001). These MSE values were taken from Table 1 of Sardy et al. (2001).

Table 2: Similar to Table 1 but for test functions *heavisine* and *doppler*.

		heavisine			doppler		
N	method	G	C	T	G	C	T
1024	`EBayes`	7.84 (0.154)	84.7 (1.87)	102 (6.61)	19.1 (0.190)	103 (1.71)	134 (8.15)
	`REBayes`	12.1 (0.337)	19.5 (0.636)	20.8 (0.703)	28.7 (0.381)	40.1 (0.784)	46.1 (0.786)
	RBPur	*17 (0.510)*	*27 (0.810)*	*33 (0.990)*	*34 (1.02)*	*49 (1.47)*	*58 (1.74)*
	`Med3`	46.0 (0.288)	64.0 (0.695)	72.6 (0.695)	53.6 (0.274)	74.4 (0.620)	85.9 (0.789)
4096	`EBayes`	3.15 (0.0450)	77.9 (0.953)	89.1 (2.73)	6.23 (0.0760)	85.1 (1.02)	96.8 (3.27)
	`REBayes`	6.69 (0.183)	12.9 (0.311)	15.7 (0.392)	10.4 (0.175)	18.2 (0.334)	21.1 (0.344)
	RBPur	*7 (0.210)*	*11 (0.330)*	*13 (0.390)*	*11 (0.330)*	*17 (0.510)*	*20 (0.600)*
	`Med3`	44.7 (0.140)	62.2 (0.356)	72.5 (0.333)	45.2 (0.142)	63.5 (0.323)	73.5 (0.322)

The following major observations can be made. The proposed procedure `REBayes` outperformed *RBPur* for most cases. Except for *bumps* with $N = 1024$, `REBayes` greatly improved the performance of `EBayes` for non–Gaussian noise. The poor performance of `REBayes` on *bumps* with $N = 1024$ is due to the fact that the tall narrow features in *bumps* are hard to distinguish from outliers when the sample size is not large enough.

5.2 Real data

Both the `EBayes` and `REBayes` procedures were applied to the glint data set analyzed by Sardy et al. (2001). The data are radar glint observations from a target captured at $N = 512$ angles. The data together with the regression estimates are displayed in Figure 2. It can be seen that `REBayes` is resistant to the adverse effects of outliers.

6 Extensions to other settings

6.1 Irregularly spaced data

Methods for wavelet thresholding with irregularly spaced data have been studied by different authors Kovac and Silverman (2000), Antoniadis and Fan (2001) and more recently Lee and Meng (2005). The proposed algorithm can be coupled with any of these methods to perform irregularly spaced data robust wavelet thresholding. The main modification is to use any of these irregularly spaced wavelet thresholding method in the S-Step of the ES–algorithm. Displayed in Figure 3(a) is the Canadian age and income data set studied for example by Ruppert et al. (2003), together with a regression estimate obtained by the irregularly spaced wavelet thresholding scheme proposed by Kovac and Silverman (2000). This same thresholding scheme was also paired with the proposed algorithm to obtain a robust regression estimate; see Figure 3(b). By comparing the two plots, one can see that the robust fitting was less sensitive to outlying data.

6.2 Robust image denoising

Extension to robust image denoising is straightforward. The only modification required is to replace the 1-D wavelet smoothing step in the algorithm by a 2-D method. Displayed in Figure 4(a) is the image data `lennon` available from the `WaveThresh` package of Nason (1998), and Figure 4(b) displays a degraded version. This degraded image was obtained in the following way. Firstly the original image was scaled so that the sum of the squares of all the pixel intensities is 1. Next 90%

Figure 2: Glint data set with EBayes and REBayes estimates.

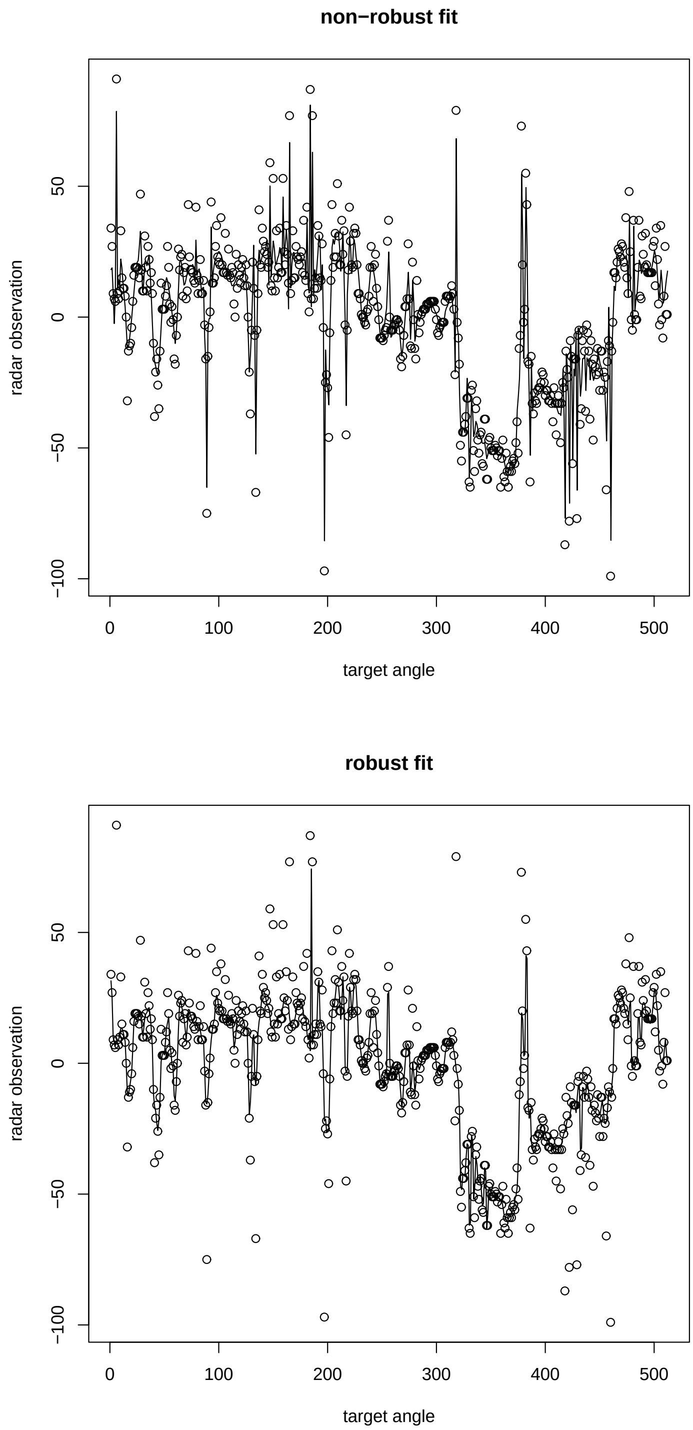
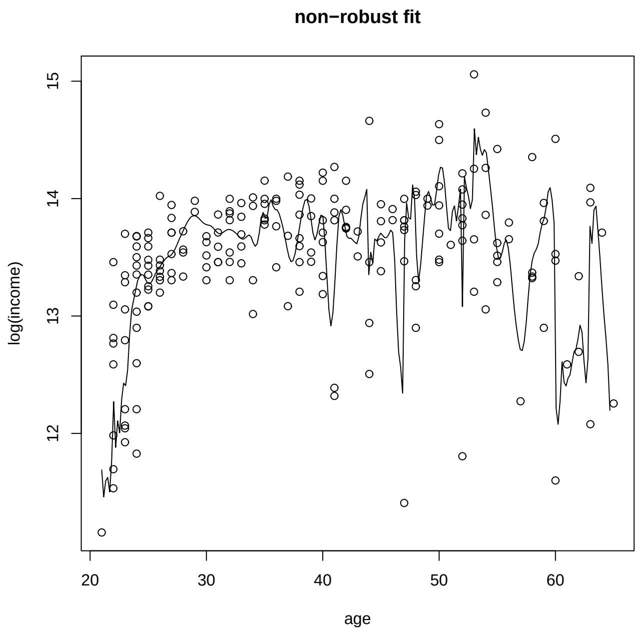
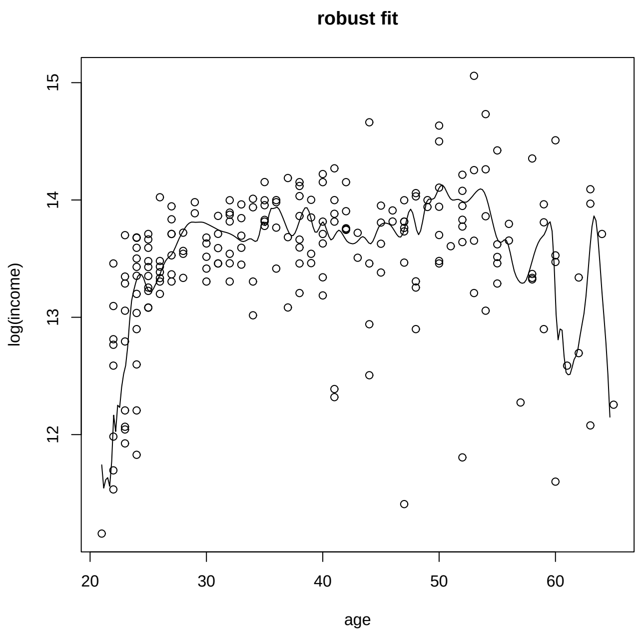


Figure 3: Robust thresholding for irregularly spaced data.

(a)



(b)



of the pixel intensities were contaminated with independent $N(0,(1/7)^2)$ noise while independent $N(0,(4/7)^2)$ outliers contaminated the remaining 10% of the image. We used this degraded image to study several approaches for denoising these data illustrated in Figure 4. The non–robust (Figure 4(c)) and robust (Figure 4(d)) wavelet thresholding were applied to this degraded image. For simplicity the hard thresholding rule with thresholding value $0.3\sigma\sqrt{2\log N}$ was used with $\sigma = 1/7$, N is the total number of pixels. The use of multiplier 0.3 was suggested by Kovac and Silverman (2000). Certainly from a visual point of view the robust method is a substantial improvement in the image quality.

Our theoretical results in Section 3 were developed for penalized linear estimators where the basis could have multi-resolution structure such as wavelets but the penalty is required to be a quadratic form on the basis coefficients. These are the "linear" wavelet estimators proposed by Antoniadis (1996) and we are interested in their performance for the `lennon` test image. The plot in Figure 4(e) is the denoised image using a linear wavelet estimator and Figure 4(f) is the corresponding robust linear estimate. For both estimates the amount of smoothing was chosen by AIC. While the robust version seems to yield a better denoising both estimators are inferior to the robust estimator that exploits thresholding.

7 Conclusions

In this paper a new method for robust smoothing is proposed and is motivated by the introduction of *pseudo data*. This method is computationally fast, easy to implement, and straightforward to extend to other settings. Results from numerical experiments and real examples suggest the method possesses promising empirical properties and gives comparable or superior mean squared error performance compared to more complicated methods. Moreover, the method has been successfully applied to handle irregularly spaced and image data.

Some theoretical properties of the method were extended from a specific result for cubic splines to a more general family of smoothers and we believe that this by itself is an important step. In approaching this problem we have found it useful to focus on the transfer of asymptotic properties from the least squares estimators to robust ones and not be overly concerned about establishing properties of the least squares estimators themselves. For spline estimators such a task is beyond the scope of a single paper. However, we also acknowledge that the theory lags our practical implementation in terms of smoothing parameter selection and the form of estimator. Although

Figure 4: Robust Image denoising.

(a) original

(b) noisy image

(c) non-robust reconstruction

(d) robust reconstruction

(e) non-robust linear reconstruction

(f) robust linear reconstruction

we show that pseudo data and a least squares smoother is equivalent to a robust smoother, what is really needed is the justification that EPD will yield an asymptotically equivalent alternative for smoothing parameter selection. Recent work has developed some promising connections between leave-one-out cross-validation for the robust estimator and cross validation applied to pseudo data (Oh and Nychka, 2005) and of our empirical results lead us to believe that EPD methods will give good estimates of λ.

The denoising example with the `lennon` image clearly delineates the superiority of a thresholded smoother compared to one with a quadratic penalty; i.e., linear smoother. We believe that our theory can be extended to the threshold case by additional Taylor series arguments. The reader should note that the key uniform bound in Lemma 1 is actually independent of the penalty and so holds with little modification.

In summary we believe that a pseudo data approach to robust smoothing is a valuable concept for transferring least squares smoothing techniques to outlier resistant ones. Although there are still many open theoretical questions concerning the ES–algorithm it appears to be a very practical method that is easily implemented in standard statistical software.

8 Appendix

The following three Lemmas assume the setting of Assumptions 1–5.

Lemma 0 *Consistency of roughness penalty*

Let $\tilde{\boldsymbol{g}}$ be a least squares penalized estimate of $\boldsymbol{g}$ with penalty function $\lambda J(f)$ and smoothing matrix $A(\lambda_n)$. Then

a) $E(J(\tilde{\boldsymbol{g}})) < J(\boldsymbol{g}) + \frac{2\sigma^2 tr[A(\lambda_n)]}{\lambda_n}$.

b) $\frac{nC_n}{\lambda_n} < J(\boldsymbol{g}) + \frac{2\sigma^2 tr[A(\lambda_n)]}{\lambda_n}$.

Proof: By the definition of the penalized estimate as the minimizer,

$$\sum_{i=1}^{n}(y_i - \tilde{g}_i)^2 + \lambda_n J(\tilde{\boldsymbol{g}}) < \sum_{i=1}^{n}(e_i)^2 + \lambda_n J(\boldsymbol{g}),$$

and in matrix notation

$$||(I - A(\lambda_n))\boldsymbol{y}||^2 + \lambda_n \tilde{\boldsymbol{g}}^T R \tilde{\boldsymbol{g}} < ||\boldsymbol{e}||^2 + \lambda_n J(\boldsymbol{g}).$$

Taking expected values of both sides and simplifying

$$\lambda_n E\tilde{\boldsymbol{g}}^T R\tilde{\boldsymbol{g}} + ||(I - A(\lambda_n))\boldsymbol{g}||^2 + \sigma^2 tr[A^2(\lambda_n)] < \lambda_n J(\boldsymbol{g}) + 2\sigma^2 tr[A(\lambda_n)],$$

where $\sigma^2 = Var(e_i)$. Part a) follows by omitting the second and third positive terms on the LHS of this inequality and dividing by λ_n. Part b) follows by omitting the first term in the LHS, dividing both sides by λ_n and noting that the sum of the remaining second and third terms is equal to nC_n.

Lemma 1 *Score function approximation*

Let $\mathcal{C}$ be defined as in Assumption 3 such that $\boldsymbol{g} \in \mathcal{C}$, $M = \sup \eta''$, and for $B > 0$ let $F = \{\boldsymbol{f} \in \mathcal{C} : ||\boldsymbol{f} - \boldsymbol{g}||_n < B\}$. For any $\epsilon > 0$ there is an N such that for $n > N$

a)

$$P\left(\sup_{f\in F} ||A(\lambda)\left(\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f})\right)|| > \epsilon B + (M/2)B^2\right) < \epsilon.$$

b)

$$P\left(\sup_{f\in F} ||A(\lambda)^{1/2}\left(\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f})\right)|| > \epsilon B + (M/2)B^2\right) < \epsilon.$$

Proof:

By the definition of pseudo data and Taylor's theorem with remainder about $\boldsymbol{g}_i$,

$$\begin{aligned}
\Psi(\boldsymbol{f})_i - \Phi(\boldsymbol{f})_i &= -2(\tilde{y}_i - \boldsymbol{f}_i) + \eta(y_i - \boldsymbol{f}_i) \\
&= -2\left\{\boldsymbol{g}_i + \frac{\eta(\epsilon_i)}{2} - \boldsymbol{f}_i\right\} + \eta(\boldsymbol{g}_i + \epsilon_i - \boldsymbol{f}_i) \\
&= -\left[\{\eta'(\epsilon_i) - 2\}(\boldsymbol{f}_i - \boldsymbol{g}_i)\right] + \left[\frac{1}{2}\eta''(\epsilon_i + a_i)(\boldsymbol{f}_i - \boldsymbol{g}_i)^2\right] \\
&= \boldsymbol{u}_{1,i} + \boldsymbol{u}_{2,i},
\end{aligned}$$

where $0 \leq |a_i| \leq |\boldsymbol{f}_i - \boldsymbol{g}_i|$, $\boldsymbol{u}_{1,i}$ is equal to the first bracketed term and $\boldsymbol{u}_{2,i}$ is equal to the second bracketed term. Setting $T_1(\boldsymbol{f}) = ||A(\lambda)\boldsymbol{u}_1||_n$ and $T_2(\boldsymbol{f}) = ||A(\lambda)\boldsymbol{u}_2||_n$ by the triangle inequality

$$||A(\lambda)\left(\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f})\right)||_n \leq T_1(\boldsymbol{f}) + T_2(\boldsymbol{f}). \tag{10}$$

We will find a uniform bound in probability for $T_1(\boldsymbol{f})$.

$$E\{T_1^2(\boldsymbol{f})\} = E||A(\lambda)\boldsymbol{u}_1||_n^2 = \frac{1}{n^2}E(\boldsymbol{u}_1^T A^2(\lambda)\boldsymbol{u}_1) = \frac{\gamma}{n^2}\sum_{i=1}^n A^2(\lambda)_{i,i}(\boldsymbol{f}_i - \boldsymbol{g}_i)^2$$

with $\gamma = 4E\{(\eta'(\epsilon_i) - 1)^2\}$. Because R is nonnegative definite the eigenvalues of $A(\lambda)$ will be less than or equal to one, and it follows that $A^2(\lambda)_{i,i} < A(\lambda)_{i,i}$. Thus from the equation above we have

$$E\{T_1^2(\boldsymbol{f})\} < \frac{\gamma}{n}\sum_{i=1}^n A(\lambda)_{i,i}(\boldsymbol{f}_i - \boldsymbol{g}_i)^2 < \gamma a_n||\boldsymbol{f} - \boldsymbol{g}||_n$$

and it follows that $E\{T_1(\boldsymbol{f})\} \leq \sqrt{\gamma a_n}||\boldsymbol{f} - \boldsymbol{g}||_n$.

Based on this estimate we now find a uniform bound on $T_1(\boldsymbol{f})$. Consider a union of open balls defined by the L_2 norm with radius r that covers $\mathcal{C}$. By Assumption 3 there is a finite open cover of $N(r)$ balls denoted by $\{\mathcal{B}_\nu\}$ and with centers $\{\boldsymbol{f}_\nu - \boldsymbol{g}\}$ whose union contains F

$$\sup_{\boldsymbol{f}\in F} |T_1(\boldsymbol{f})| \leq \sup_\nu |T_1(\boldsymbol{f}_\nu)| + \max_\nu \left[\sup_{f\in\mathcal{B}_\nu} |T_1(\boldsymbol{f}) - T_1(\boldsymbol{f}_\nu)| \right]. \tag{11}$$

Considering the first term in (11), by Markov's inequality, the bound on $E[T_1(\boldsymbol{f})]$ and Bonferroni's inequality

$$\begin{aligned} P\left(\sup_\nu T_1(\boldsymbol{f}_\nu) < B\epsilon/2\right) &> 1 - \sum_\nu P\left[T_1(\boldsymbol{f}_\nu) > B\epsilon/2\right] \\ &> 1 - \sum_\nu \frac{2E[T_1(\boldsymbol{f}_\nu)]}{\epsilon B} \\ &> 1 - \frac{2N(r)\sqrt{\gamma a_n}}{\epsilon}. \end{aligned}$$

We now bound the second term on the RHS of (11). Given $A(\lambda)$ has eigenvalues bounded by one and η' is bounded by M_1, $|T_1(\boldsymbol{f}) - T_1(\boldsymbol{f}_\nu)| < (M_1+1)||\boldsymbol{f} - \boldsymbol{f}_\nu||_n$. Moreover, by the construction of the open cover we now have $\max_\nu[\sup_{f\in\mathcal{B}_\nu} |T_1(\boldsymbol{f}) - T_1(\boldsymbol{f}_\nu)|] < (M_1+1)rB$.

Choose r such that $(M_1+1)r < \epsilon$, and choose n sufficiently large such that $N(r)\sqrt{a_n}/\epsilon < \epsilon$. It now follows that

$$P\left[\sup_{\boldsymbol{f}\in F} T_1(\boldsymbol{f}) < B\epsilon\right] > 1 - \epsilon. \tag{12}$$

To complete the proof, we now derive uniform bound for $T_2(\boldsymbol{f})$, $T_2(\boldsymbol{f}) = ||A(\lambda)\boldsymbol{u}_2||_n < ||\boldsymbol{u}_2||_n$. By the assumption that η'' is uniformly bounded by M,

$$||\boldsymbol{u}_2||_n^2 \leq (M/2)^2 \frac{1}{n^2} \sum_{i=1}^n (\boldsymbol{f}_i - \boldsymbol{g}_i)^4 \leq (M/2)^2 ||\boldsymbol{f} - \boldsymbol{g}||_n^4.$$

Thus $\sup_{\boldsymbol{f}\in F} T_2(\boldsymbol{f}) \leq (M/2)B^2$. Combining this bound with (12) it now follows that for $\epsilon > 0$ and n sufficiently large

$$\sup_{\boldsymbol{f}\in F} T_1(\boldsymbol{f}) + T_2(\boldsymbol{f}) < \epsilon B + (M/2)B^2$$

with probability greater than $1-\epsilon$. Based on the bound in (10) part a) lemma now follows.

The proof of part b) follows the same arguments for part a) because at the point of bounding $A(\lambda)^2$ we use the inequality $A^2(\lambda)_{i,i} < A(\lambda)_{i,i}$. Thus, one could replace A^2 by A and all the steps are still valid.

Lemma 2 *Bounds on score mapping*

Let $U(\boldsymbol{x}) = \boldsymbol{x} - (1/2)A(\lambda)\Phi(\boldsymbol{x}+\boldsymbol{g})$ and $\delta > 0$. There is an N such that for all $\boldsymbol{x} \in F_n - \boldsymbol{g}$, $U(\boldsymbol{x}) \in F_n - \boldsymbol{g}$ for $n > N$ and with probability greater than $1-\delta$.

Proof: For any $\boldsymbol{f} \in \Re^n$ set $\boldsymbol{x} = \boldsymbol{f} - \boldsymbol{g}$, then

$$
\begin{aligned}
U(\boldsymbol{x}) &= -(1/2)A(\lambda)\Phi(\boldsymbol{f}) - (\boldsymbol{f}-\boldsymbol{g}) \\
&= -(1/2)A(\lambda)\{\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f})\} + [(1/2)A(\lambda)\Psi(\boldsymbol{f}) - (\boldsymbol{f}-\boldsymbol{g})]\,.
\end{aligned}
$$

Based on (4) and Assumptions 3 and 5, the term in square brackets simplifies to $\boldsymbol{g} - \tilde{\boldsymbol{g}}$. Thus

$$||U(\boldsymbol{x})||_n \le ||A(\lambda)(\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f}))||_n + ||\tilde{\boldsymbol{g}} - \boldsymbol{g}||_n. \tag{13}$$

We now apply Lemma 1-a) with $B = 4\sqrt{C_n}/\delta$ and $\epsilon < \delta/2$ to the first term on the RHS of (13) and apply Markov's inequality to the second term. With these bounds we have

$$\sup_{\boldsymbol{x}\in(F_n-\boldsymbol{g})} ||U(\boldsymbol{x})||_n \le \{1/4 + (1/2)(M/2)\sqrt{C_n} + 1/2\}\frac{4\sqrt{C_n}}{\delta}$$

with probability greater than $1-\delta$.

Now choose n sufficiently large so that the quantity in braces is strictly less than one. Thus $||U(\boldsymbol{x})||_n$ will be less than $\frac{4\sqrt{C_n}}{\delta}$ for all $\boldsymbol{x} \in (F_n - \boldsymbol{g})$ with probability greater than $1-\delta$.

The proof will be completed upon showing that $U(\boldsymbol{x})+\boldsymbol{g}$ is also contained in $\mathcal{C}$ for all $x \in F_n - \boldsymbol{g}$ with high probability. Specifically we must establish that $J(U(\boldsymbol{x})+\boldsymbol{g})$ for all $\boldsymbol{x} \in F_n - \boldsymbol{g}$ is less that D from Assumption 3 with high probability. Setting $\boldsymbol{f} = \boldsymbol{x} + \boldsymbol{g}$

$$
\begin{aligned}
J(U(\boldsymbol{x})+g)^{1/2} &= ||R^{1/2}(U(\boldsymbol{x})+g)|| \\
&\le (1/2)||R^{1/2}A(\lambda_n)(\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f}))|| + ||R^{1/2}\tilde{\boldsymbol{g}}|| \\
&= (1/2)||R^{1/2}A(\lambda_n)(\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f}))|| + J(\tilde{\boldsymbol{g}})^{1/2}.
\end{aligned} \tag{14}
$$

Now using the fact that $||R^{1/2}A(\lambda_n|| < (1/\sqrt{\lambda_n})||A(\lambda_n)^{1/2}||$,

$$||R^{1/2}A(\lambda_n)(\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f}))|| < (1/\sqrt{\lambda_n})||A(\lambda_n)^{1/2}(\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f}))||.$$

Applying Lemma 1-b) with $B = 4\sqrt{C_n}/\delta$ and $\epsilon = \delta/2$ and rearranging terms,

$$\sup_{\boldsymbol{f}\in F_n-\boldsymbol{g}} \sqrt{(n/\lambda_n)}||A(\lambda_n)^{1/2}(\Phi(\boldsymbol{f}) - \Psi(\boldsymbol{f}))||_n < \sqrt{(nC_n/\lambda_n)}(1/2 + (M/4)\sqrt{C_n})\sqrt{C_n} \tag{15}$$

with probability greater than $1 - \delta/2$. Moreover, in view of Assumption 5-b) and Lemma 0 for fixed δ there is an $\Omega_1 < \infty$ that is independent of n and bounds (15).

Now from Markovs inequality $P[J(\tilde{\boldsymbol{g}}) < 2EJ(\tilde{\boldsymbol{g}})/\delta] > 1 - \delta/2$ and combining this result with Lemma 0 it follows that

$$J(\tilde{\boldsymbol{g}}) < (2/\delta)\left\{J(\boldsymbol{g})/\delta + 2\sigma^2 tr[A(\lambda_n)])/\lambda\right\} \tag{16}$$

with probability greater than $1 - \delta/2$. For fixed δ by Assumption 5-b), there is an $\Omega_2 < \infty$ that is independent of n and bounds (16). Now combine the bounds on the two terms originating from (14)

$$\sup_{\boldsymbol{x} \in F_n - \boldsymbol{g}} J(U(\boldsymbol{x}) + \boldsymbol{g}) < \Omega_1^2 + \Omega_2$$

with probability greater than $1 - \delta$ and for n sufficiently large.

It follows now that $U(\boldsymbol{f}) + \boldsymbol{g}$ is contained in F_n with probability greater than $1 - 2\delta$.

Acknowledgment

This work was supported in part by U.S. National Science Foundation DMS-0203901, and by Korea Research Foundation Grant funded by Korea Government(MOEHRD, Basic Research Promotion Fund) (KRF-2005-003-C00032).

References

Adams, R. (1975) *Sobolev Spaces*, Academic Press, New York.

Alliney, S. and Ruzinsky, S.A. (1994) An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation. *IEEE Transactions on Signal Processing*, **42**, 618–627.

Angelini, C., De Canditiis, D. and Leblanc, F. (2003) Wavelet regression estimation in non parametric mixed effect models. *Journal of Multivariate Analysis*, **85**, 267–291.

Antoniadis, A. (1996) Smoothing noisy data with tapered Coiflets series. *Scandinavian Journal of Statistics*, **23**, 313–330.

Antoniadis, A. and Fan, J. (2001) Regularization of wavelet approximations. *Journal of the American Statistical Association*, **96**, 939–962.

Aubin, J.-P. (2000) *Applied Functional Analysis*, 2nd Edition, Wiley, New York.

Averkamp, R. and Houdré (2003) Wavelets thresholding for non necessarily Gaussian noise: Idealism. *The Annals of Statistics*, **31**, 110–151.

Cantoni, E. and Ronchetti, E. (2001) Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, **11**, 141–146.

Cleveland, W.S. (1979) Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cox, D.D. (1983) Asymptotics for M-type smoothing splines. *The Annals of Statistics*, **11**, 530–551.

Dechevsky, L.T. and Penev, S.I. (1999) Weak penalized least squares wavelet regression estimation, Technical Report S99-1, Department of Statistics, School of Mathematics, University of New South Wales, Australia.

Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Hall, P. and Jones, M.C. (1990) Adaptive M-estimation in nonparametric regression. *The Annals of Statistics*, **18**, 1712–28.

Härdle, W. and Gasser, T. (1984) Robust non-parametric function fitting. *Journal of the Royal Statistical Society Ser.* B, **46**, 42–51.

Huber, P.J. (1973) Robust regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, **1**, 799–821.

Huber, P.J. (1979) Robust smoothing. In Launer, R. L. and Wilkinson, G. N. (Eds.), *Robustness in Statistics*. Academic Press, New York, 33–48.

Huber, P.J. (1981) *Robust Statistics*, John Wiley & Sons, NY.

Johnstone, I.M. and Silverman, B.W. (2005) Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics*, **33**, 1700–1752.

Kovac, A. and Silverman, B.W. (2000) Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association*, **95**, 172–183.

Lee, T.C.M. and Meng, X-L. (2005) A self-consistent wavelet method for denoising images with missing pixels. *Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing*, **II**, 41–44.

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

Nason, G.P. (1998) *WaveThresh3 Software*. Department of Mathematics, University of Bristol, Bristol, UK.

Nychka, D.W. (1995) Splines as local smoothers. *The Annals of Statistics*, **23**, 1175–1197.

Oh, H-S. and Nychka, D.W. (2005) Smoothing spline regression by robust cross-validation. Manuscript.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) *Semiparametric Regression.* Cambridge University Press, Cambridge.

Sardy, S., Tseng, P. and Bruce, A. (2001) Robust wavelet denoising. *IEEE Transactions on Signal Processing*, **49**, 1146–1152.

Silverman, B.W. (1985) Some aspects of spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society Ser.* B, **47**, 1–52.

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