Wavelets for regression smoothing

Wavelet And Multifractal Analysis 2004

Summer School

Corsica, July 19 - 31, 2004

Anestis Antoniadis

University Joseph Fourier

Outline

- Generalities on nonparametric regression
- Penalized regression splines
- Penalized wavelet estimation with quadratic penalties
- Other penalties and model selection
- Extensions

Generalities

Regression models describe the dependence of the response variable of interest Y on one or more predictor variables $\mathbf{X}$.
A basic analysis starts with a random sample of size n from the distribution of (X, Y) where the conditional mean and variance of the ith response Y_i are given by

$$\mathbb{E}(Y_i/X = x_i) = \eta(x_i) \quad \mathrm{Var}(Y_i/X = x_i) = \sigma^2(x_i) = \sigma^2.$$

The parameter σ is a scale parameter which is assumed to be constant in what follows.

Nonparametric methods

Nonparametric models are those models where the form of the conditional expectations is dictated by the data.

Two approaches:

- Fit simple parametric models locally, e.g., moving averages or local polynomial estimation.
- Fit a highly complex parametric model with a complexity penalty to prevent "overfitting". Roughly, "overfitting" means fitting the noise, not the signal, so that features are detected that would not be present in an independent replicate of the data.

Both approaches require a "smoothing parameter."

Spline based methods

We will focus on nonparametric penalized regression methods involving the use of basis functions and quadratic penalties, pointing out some basic principles they have in common with splines when fitting regular curves and wavelets when fitting less regular curves. More precisely, we will consider the estimation of $\eta(x)$ through the minimization of

$$Z_n(\eta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \eta(x_i))^2 + \lambda J(\eta), \qquad (1)$$

where $J(\cdot)$ is a roughness functional. The parameter λ controls the trade-off between lack of fit of η and roughness.

To estimate $\eta(\cdot)$ in a flexible manner we represent it as a linear combination of known basis functions $\{h_k,\ k = 1, \ldots, K\}$,

$$\eta(x) = \sum_{k=1}^{K} \beta_k h_k(x),$$

and then try to estimate the coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$.

Usually the number K of basis functions used in the representation of η should be large in order to give a fairly flexible way for approximating η.

Popular examples of such basis functions h_k are *wavelets* and *polynomial splines*.
A crucial problem is the choice of the number K of basis functions. A small K may result in a function space which is not flexible enough to capture the variability of the data, while a large number of basis functions may lead to serious overfitting.
Traditional ways of "smoothing" are through *basis selection* see e.g. Friedman and Silverman (1989), Friedman (1991) and Stone et al. (1997) or *regularization*.

For x given, let H be the matrix whose columns are $h_k(x_i)$, for $k = 1, \ldots, K$. We have $\eta(x) = H(x)\boldsymbol{\beta}$. Denote

$$L_{\mathbf{y}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - H\boldsymbol{\beta})^2 \, .$$

Then (**??**) becomes

$$Z_n(\boldsymbol{\beta}) = \frac{1}{n} L_{\mathbf{y}}(\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta}).$$

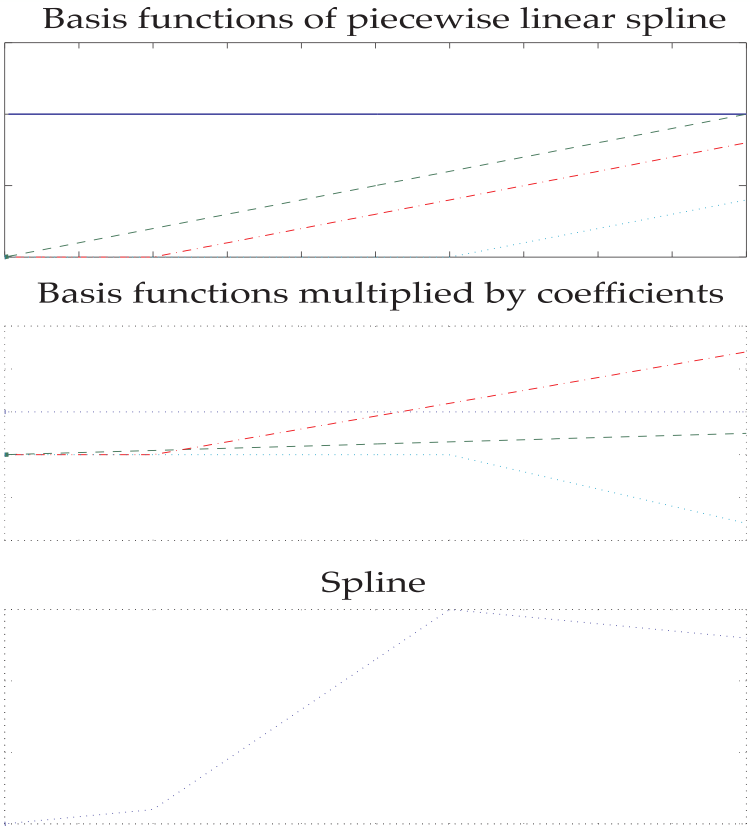
Power basis

Polynomial regression splines are continuous piecewise polynomial functions where the definition of the function changes at a collection of knot points, which we write as $t_1 < \cdots < t_K$.
Using the notation $z_+ = max(0, z)$, then, for an integer $p \geq 1$, the truncated power basis for polynomial of degree p regression splines with knots $t_1 < \cdots < t_K$ is

$$\{1, x, \ldots, x^p, (x - t_1)_+^p, \ldots, (x - t_K)_+^p\}.$$

$\eta(x, \boldsymbol{\beta}) = H(x)\boldsymbol{\beta}$ is a pth degree polynomial on each interval between knots and has $p - 1$ continuous derivatives everywhere. Refer to Eilers and Marx (1996) and Ruppert and Carroll (2000).

Example of Regression Splines



Basis functions of piecewise linear spline

Basis functions multiplied by coefficients

Spline

Illustration of a Truncated Power Basis construction.

P-splines

Let $\hat{\boldsymbol{\beta}}$ minimize

$$PSS(\boldsymbol{\beta}) = L_{\mathbf{y}}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{K} \beta_{p+j}^2$$

This penalized least-squares estimator is called a P-spline (term due to Eilers and Marx (1996)). Other quadratic penalties on $\boldsymbol{\beta}$ are possible.

- In particular, one can find B so that
$$\boldsymbol{\beta}^T B \boldsymbol{\beta} = \int_{\min x_i}^{\max x_i} (\ddot{\eta}(x, \boldsymbol{\beta}))^2 dx$$
- The choice of spline basis is unrestricted: Truncated power functions are not needed, though we find them convenient. A natural cubic spline basis could be used, so that cubic smoothing splines are a special case of P-splines.

Smoothing parameter

Here $\lambda > 0$ is a penalty or regularization weight. As $\lambda \rightarrow \infty$, the penalized spline converges to a pth degree polynomial fit. As $\lambda \rightarrow \infty$, the penalized spline converges to the OLS fitted spline.

The coefficient β_{p+j} is the jump in the pth derivative at the knot t_j. Thus, the penalty is on the $(p+1)$th derivative, interpreted as a generalized function.

P-splines – cont.

Let $\mathbf{y} = (y_1, \cdots, y_n)^T$ and $D = \text{diag}(0, \ldots, 0, 1, \ldots, 1)$ ($p + 1$ zeros and K ones). Then

$$PSS(\boldsymbol{\beta}) = \|\mathbf{y} - H\boldsymbol{\beta}\|^2 + \lambda \|D\boldsymbol{\beta}\|^2.$$

Since $\hat{\boldsymbol{\beta}}(\lambda)$ solves $\partial PSS(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = 0$ we have

$$\hat{\boldsymbol{\beta}}(\lambda) = \left(H^T H + \lambda D\right)^{-1} H^T \mathbf{y}.$$

(Ridge regression)

Cross Validation (CV)

Goal: Choose λ.
$\eta(x, \hat{\boldsymbol{\beta}}(\lambda))$ is the fitted spline with penalty weight λ.
$\eta_{-i}(x, \hat{\boldsymbol{\beta}}(\lambda))$ is the fitted spline with penalty weight λ and without using (x_i, y_i).

$$CV(\lambda) = \sum_{i=1}^{n} \left\{ y_i - \eta_{-i}(x_i, \hat{\boldsymbol{\beta}}(\lambda)) \right\}^2$$

Principle: Choose λ that minimizes $CV(\lambda)$.
Problem: CV is computationally intensive.

Solution: Use generalized cross validation (GCV), an approximation to CV.

Generalized cross validation

The Smoother or Hat Matrix

$$S(\lambda) := H(H^T H + \lambda D)^{-1} H^T \longrightarrow \hat{\mathbf{y}} = S(\lambda)\mathbf{y}$$

Degrees of freedom: $DF(\lambda) = \mathrm{tr}(S(\lambda))$.
GCV:

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}(y_i - \eta(x_i, \hat{\boldsymbol{\beta}}(\lambda)))^2}{(1 - DF(\lambda)/n)^2}.$$

So use $\hat{\lambda}$ that minimizes $GCV(\lambda)$ and define $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\lambda})$.

Penalized wavelet regression

We will now be concerned with developing a counterpart to the spline smoothing technique for the case of fitting less regular curves.

Several variants of this penalized approach have been suggested by several authors (Devore and Lucier (1992), Antoniadis (96), Amato and Vuza (97), Dechevsky and Penev (99).

The regularity of the curve is discussed in terms of the size of its norm in a Sobolev space with a relatively small value of the positive smoothness index.

We will also study a specific version of generalized full cross validation for the choice of the smoothing parameter.

Setup and assumptions

Considering the structure of the wavelet multiresolution, the design points x_i will be assumed to have the form $x_i = i\Delta$, with $\Delta = 2^{-N}$ for $i = 0, \ldots, n-1$, where $N = \log_2 n$, n being the sample size.

For simplicity of presentation we shall work with periodic wavelets on the interval on $[0, 1]$.

Functions ϕ and ψ will respectively denote the scaling function and the wavelet associated to a periodic q-regular multiresolution analysis of $L^2([0, 1[)$.

Penalized wavelet regression – cont.

We will assume that $\eta \in H^\nu([0,1[)$, where $H^\nu(H^\nu([0,1[)$ denotes the Sobolev space of order ν ($\nu > 1/2$).
Assume further that the scaling function ϕ is a coiflet of order L with $L > [\nu] + 1$.
The nested structure of a multiresolution analysis leads to an efficient tree-structured algorithm for the decomposition of functions in V_N for which the coefficients $\langle \eta, \phi_{N,k} \rangle$ are given.

When a function is given in sampled form there is no general method for deriving the coefficients $\langle \eta, \phi_{N,k} \rangle$. We will approximate the projection P_{V_N} by some operator Π_N in terms of the sampled values $\eta(\frac{k}{2^N})$.

Since the coiflets have L vanishing moments, we have: *The set of non zero coefficients $\alpha^{\{N\}}(k) = 2^{N/2}\langle \eta, \phi_{N,k}\rangle$ has a cardinality equivalent to $\mathcal{O}(n)$. Moreover, with $L > [\nu] + 1$, the following uniform (in $0 \leq k \leq 2^N - 1$) bound holds:*

$$|\alpha^{\{N\}}(k) - \eta(\frac{k}{2^N})| \leq C_1\, 2^{-N\nu}$$

where C_1 is a constant only depending on the coiflet ϕ.

One is therefore able to approximate the coefficients $\langle \eta, \phi_{N,k}\rangle$ with an error $\mathcal{O}(2^{-N/2}2^{-N\nu})$.

It is now natural to approximate $P_{V_N}\eta$ by

$$(\Pi_N \eta)(t) = \sum_{k=0}^{2^N-1} \eta(\frac{k}{2^N})\, \phi(2^N t - k)$$

Using such an approximation we have:

$$\|P_{V_N}\eta - \Pi_N \eta\|_\infty \leq \mathcal{O}(2^{-N\nu})$$

Observing that $\mathbb{E}(Y_k) = \eta(\frac{k}{2^N})$ justifies completely replacing the original data by the "raw" function

$$\hat{\eta}_N(t) = 2^{-N/2} \sum_{k=0}^{2^N-1} Y_k \phi_{N,k}(t)$$

For any $\eta \in H^{\nu}([0,1])$, $\Pi_N \eta$ can be expanded as

$$\Pi_N \eta = \sum_{k=0}^{2^{j_0}-1} a_{j_0,k} \phi_{j_0,k} + \sum_{j=j_0}^{N} \sum_{k=0}^{2^j-1} d_{j,k} \psi_{j,k}.$$

The orthogonal discrete wavelet transform on the interval, say W, applied on $\hat{\eta}_N$, identified by the vector $\mathbf{y}$ of observed values, gives a new sequence of numbers $\{\left(v_{j_0,k}\right)_{k=0,\ldots,2^{j_0}-1}, \left(w_{j,k}\right)_{j=j_0,\ldots,N;k=0,\ldots,2^{j_0}-1}\}$, say $\left((v_{j,k}),(w_{j,k})\right)$ for short, that are interpreted as the coefficients of the expansion of the function $\hat{\eta}_N$.

Under the noise model, noise contaminates all wavelet coefficients of $\Pi_N\eta$ equally.

Consequently, the empirical wavelet coefficients $\Big((v_{j,k}), (w_{j,k})\Big)$, obtained by applying the discrete orthogonal transform on the data vector $2^{-N/2}\mathbf{y}$, can be considered as a version of the coefficients $\Big((a_{j,k}), (d_{j,k})\Big)$ of $\Pi_N\eta$, contaminated by a similar white noise $\epsilon_{j,k}$ of variance $\sigma^2/2^N$.

The regularization problem is now given by

$$\inf_{\eta\in\mathcal{H}}\left\{\|\hat{\eta}_N-\eta\|_2^2+\lambda\|P_{V_{j_0}^{\perp}}\eta\|_{\mathcal{H}}^2\right\}.$$

and using the equivalent sequence norms and these expression, minimizing the penalized functional given above is equivalent to minimize the expression

$$\sum_{k=0}^{2^{j_0}-1}\left[v_{j_0,k}-a_{j_0,k}\right]^2+\sum_{j=j_0}^{N}\sum_{k=0}^{2^j-1}\left[(w_{j,k}-d_{j,k})^2+\lambda 2^{2\nu j}d_{j,k}^2\right],$$

where $w_{j,k}=0$ for $j>N$ and $k=0,\ldots,2^j-1$.

We thus have $\hat{a}_{j_0,k} = v_{j_0,k}, \quad k = 0, \ldots, 2^{j_0} - 1$ and

$$\hat{d}_{j,k} = \frac{w_{j,k}}{1 + \lambda 2^{2\nu j}}, \quad j \geq j_0, k = 0, \ldots, 2^j - 1,$$

that is: $\hat{\eta}_\lambda = \sum_{k=0}^{2^{j_0}-1} v_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{N} \sum_{k=0}^{2^j-1} \hat{d}_{j,k}\psi_{j,k}$, a tapered wavelet series estimator of η, i.e. $\hat{\eta}_\lambda$ may be viewed as the result of passing the “raw” wavelet series estimate $\hat{\eta}_N$ through a low pass filter controlled by the parameters λ and ν. We see that the resulting estimator is a linear estimator of shrinkage type. Shrinking is level dependent. Up to level j_0 there is no shrinking; shrinking of the wavelet coefficients is heavier at higher levels.

Asymptotics

Asymptotic minimax optimality theory (see Delyon and Juditsky (1996)) allow us to make the above choice for j_0 and N. It turns out that j_0 should satisfy an inequality of the type $(\tilde{C}n)^{1/(2\nu+1)} \leq 2^{j_0} \leq 2(\tilde{C}n)^{1/(2\nu+1)}$ with $\tilde{C} > 0$ suitably chosen. A choice $2^N = O(n/\log n)$ is motivated by the observation that for $\nu > 1/2$ (when H^ν contains only continuous functions), both $|d_{j,k}| \leq O(n^{-1/2})$ and $|w_{j,k} - d_{j,k}| = O(n^{-1/2})$ hold simultaneously. Hence, for any n there is no point in going to levels of j with $2^j \geq n$ since the order of the coefficient estimated is smaller than the order of the error of estimation. The logarithmic factor in the choice of N ensures better behaviour of the estimator for continuous and smooth functions.

Asymptotics–cont.

Let us introduce the mean square error $ER(\lambda) = \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}_\lambda(x_i) - \eta(x_i))^2\right\}$ (the quantity we are interested in controlling). We then have

If $\lambda = \mathcal{O}\left(n^{-2\nu/(2\nu+1)}\right)$ *then* $ER(\lambda) = \mathcal{O}\left(n^{-2\nu/(2\nu+1)}\right)$.

The estimator is linear in the observations. Asymptotic minimax optimality theory suggests that linear wavelet estimators may in general be outperformed by nonlinear ones when no a priori information is available for the curve. But for $\nu > 1/2$ this case is within the parameter range where linear estimators have the same asymptotic rates as the best non-linear ones.

Choosing λ

Essentially two methods:

- One is generalized cross-validation used by Amato and Vuza (97), Jansen *et al.* (1997), Dechevski and Penev (1999) who have chosen to work directly with a wavelet analogue of the cross-validation formula for smoothing splines, overcoming some “compatibility condition” which doesn’t hold for the wavelet case.
- Another method (Antoniadis (96)) completely avoids the compatibility problem but the price to be paid for this is that the estimator depends explicitly on the noise variance. To deal with this, an estimator of the noise variance is incorporated within the definition of the estimator.

GFCV for λ

Denote $\hat{\eta}_\lambda = \sum_{k=0}^{2^{j_0}-1} \hat{a}_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{N} \sum_{k=0}^{2^j-1} \hat{d}_{j,k}\psi_{j,k}$, and $\hat{\eta}_\lambda^{(-i)} = \sum_{k=0}^{2^{j_0}-1} \hat{a}_{j_0,k,(-i)}\phi_{j_0,k} + \sum_{j=j_0}^{N} \sum_{k=0}^{2^j-1} \hat{d}_{j,k,(-i)}\psi_{j,k}$, with $\hat{a}_{j_0,k,(-i)} = (n-1)^{-1} \sum_{\ell=1,\ell\neq i}^{n} \phi_{j_0,k}(x_\ell)y_\ell$ and $\hat{d}_{j,k,(-i)} = (n-1)^{-1} \sum_{\ell=1,\ell\neq i}^{n} \frac{1}{1+\lambda 2^{2\nu j}}\psi_{j,k}(x_\ell)y_\ell$.
Then, paralleling the P-splines case the typical cross validation functional to be minimized with respect to λ is

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\eta}_\lambda^{(-i)}(x_i))^2.$$

Compatibility condition

Reduction in computation for minimizing CV is achieved if the "compatibility condition" holds. First define $\hat{y}_{(-i)}(\lambda) = \hat{\eta}_\lambda(x_i; y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$.
Compatibility means that
$\hat{y}_{(-i)}(\lambda) = \hat{\eta}_\lambda(x_i; y_1, \ldots, y_{i-1}, \hat{y}_{(-i)}(\lambda), y_{i+1}, \ldots, y_n)$
holds.
Under the above condition the cross-validation functional can be expressed in terms of the ordinary residuals:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\eta}_\lambda(x_i))^2/(1 - h_{ii}(\lambda))^2,\ h_{ii}(\lambda) = \frac{\partial \hat{\eta}_\lambda}{\partial y_i}(x_i)$$

Unfortunately, for a shrinking type estimator the compat-

Compatibility condition–cont.

The idea is to opt for another alternative by changing the cross validation criterion itself. Note that standard cross validation is defined entirely with respect to samples of size $(n-1)$. One can adjust it for samples of size n as suggested by Bunke et al. (1993). In their approach, the value of y_i is replaced by $\hat{\eta}_\lambda(x_i)$ instead of leaving it out in defining the prediction of the i-th design point. The resulting functional is called the FCV (full cross validation) functional :

$$FCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{\eta}_\lambda(x_i))^2.$$

Compatibility condition–cont.

It turns out that under the condition of linearity only, one gets in terms of the ordinary residuals:

$$FCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\eta}_\lambda(x_i))^2 \cdot (1 + h_{ii}(\lambda))^2,$$

where

$$h_{ii}(\lambda) = \frac{1}{n}\sum_k \phi_{j_0,k}^2(x_i) + \sum_{j=j_0}^{N}\sum_k \frac{1}{1+\lambda 2^{2\nu j}}\psi_{j,k}(x_i)^2$$

Following the same idea, as for GCV one defines the generalized FCV (GFCV) by replacing $h_{ii}(\lambda)$ by $n^{-1}\sum_{i=1}^{n} h_{ii}(\lambda)$.

Asymptotics again

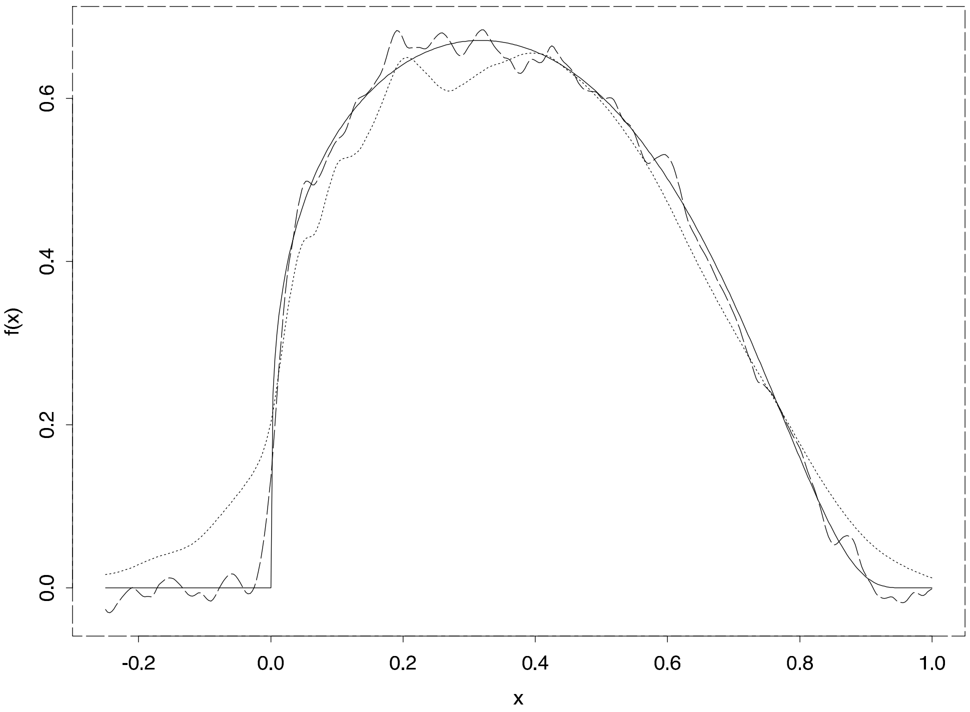
Denote $\hat{\lambda} = \text{argmin}_{\lambda} \mathbb{E}(GFCV(\lambda))$ and $\lambda^{\star} = \text{argmin}_{\lambda} ER(\lambda)$. One can then show that

If $\eta \in B^{\nu'}_{2,2} \cap B^{\nu_1}_{\infty,\infty}$ where $0 < \nu_1 \leq \nu' < \nu < r$, if $j_0 = O(1)$, $2^N = O(n/\log n)$, then for n large enough, $\hat{\lambda}$ and λ^* exist and $\frac{ER(\hat{\lambda})}{ER(\lambda^{\star})} \downarrow 1$ as $n \to \infty$. Moreover,

$$\mathbb{E}(GFCV(\lambda)) \simeq ER(\lambda) + \sigma^2$$

holds for large n, in a neighborhood of λ^*, uniformly in $\lambda \geq 0$.

An example



Simulated fit

True curve: solid line; the penalized estimator: dashed line; soft thresholded estimator: dotted line. $n = 1024$, $j_0 = 1$, $N = 6$, $\sigma^2 = 0.2$.

Other penalties

The wavelet estimation procedure described before is controlled by a quadratic penalty and as such produces linear estimates that have good rates for smooth functions only. Moreover, it does not take into account that most functions have usually a sparse wavelet representation.

In order to deal with such problems a variety of other type of penalties have been proposed in the recent literature, see for example Solo (1998), McCoy (1999), Moulin and Liu (1999), Belge *et al.* (2000), Antoniadis and Fan (2001) to cite only a few.

- In Solo (1998), the penalized least-squares with an L_1 penalty is modified to a weighted least squares in order to deal with correlated noise and an iterative algorithm is discussed for its solution. The choice of the regularization parameter is not discussed.
- By analogy to smoothing splines, E. J. McCoy (1999) uses a penalty function which simultaneously penalizes the residual sum of squares and the second derivative of the estimator at the design points. For a given regularization parameter, the solution of the resulting optimization problem is found using simulated annealing, but there is no suggestion on a possible method of chosing the smoothing parameter in her work.

- In Moulin and Liu (1999), the soft and hard thresholded estimators appear as MAP estimators in the context of Bayesian estimation under zero-one loss, with generalized Gaussian densities serving as a prior distribution for the wavelet coefficients (see also Leoporini & Pesquet (1998)).
- A similar approach is also used by Belge (2000) in the context of wavelet domain image restoration. The smoothing parameter in Belge (2000) is selected by the *L*-curve criterion (see Hansen and O'Leary (1993)). It is however known (see Vogel (1996)) that such a criterion can sometimes lead to nonconvergent solutions when the function to be recovered presents some irregularities.

Setup

Assume again that the design points are $t_i = i/2^N$ for $i = 0, \ldots, n-1$ and $N = \log_2 n$.

Let $\boldsymbol{\eta}$ be the underlying regression function collected at all dyadic points $\{i/2^J,\ i = 0, \ldots, 2^J - 1\}$.

Apply the Wavelet Transform on $\boldsymbol{\eta}$: $\boldsymbol{\theta} = W\boldsymbol{\eta}$ and $\boldsymbol{\eta} = W^T\boldsymbol{\theta}$, to get an

Overparametrized linear model:

$$\mathbf{Y}_n = W^T\boldsymbol{\theta} + \boldsymbol{\epsilon}.$$

The Wavelet basis on which $\boldsymbol{\eta}$ is projected is chosen by fixing the resolution N. The estimate of $\boldsymbol{\theta}$ and therefore of $\boldsymbol{\eta}$ is recovered by penalized least-squares

$$2^{-1}\|\mathbf{Y}_n - W^T\boldsymbol{\theta}\|^2 + \sum_{i\in I_N} p_\lambda(|\theta_i|)$$

Since W is an orthonormal matrix this is equivalent in minimizing componentwize

$$2^{-1}\sum_{i=1}^{n}(z_i - \theta_i)^2 + \lambda \sum_{i\geq i_0} p(|\theta_i|),$$

where z_i is the i^{th} row of $\mathbf{z} = W\mathbf{Y}_n$.

General types of Penalties

Several penalty functions have been used in the literature.

- The L_2 penalty $p(\theta) = |\theta|^2$ yields a ridge type regression.
- The L_1 penalty $p(\theta) = |\theta|$ results in soft thresholding rule introduced by Bickel (1983) and used by Donoho and Johnstone (1994) in the wavelet setting.
- The penalty $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$, leads to the hard-thresholding rule (Antoniadis, 1997).
- The mixture penalty $p_\lambda(|\theta|) = \lambda \min(|\theta|, \lambda)$.
- More generally, the L_q $(0 \leq q \leq 1)$ penalty leads to bridge regression (see Frank and Friedman (1993)).

Conditions on p

Usually, the penalty function p is chosen to be symmetric and increasing on $[0, +\infty)$. Furthermore, p can be convex or non-convex, smooth or non-smooth.
In the wavelet setting, Antoniadis and Fan (2001) provide some insights into how to choose a penalty function. A good penalty function should result in

- *unbiasedness*,
- *sparsity* ,
- *stability*.

Examples

Penalty function	Convexity	Smoothness at 0	Authors
$p(\beta) = \lvert\beta\rvert$	yes	$p'(0^+) = 1$	(Rudin 1992)
$p(\beta) = \lvert\beta\rvert^\alpha, \;\; \alpha \in (0,1)$	no	$p'(0^+) = \infty$	(Saquib 1998)
$p(\beta) = \alpha\lvert\beta\rvert/(1+\alpha\lvert\beta\rvert)$	no	$p'(0^+) = \alpha$	(Geman 92, 95)
$p(0) = 0, \; p(\beta) = 1, \forall\beta \neq 0$	no	discontinuous	Leclerc 1989
$p(\beta) = \lvert\beta\rvert^\alpha, \; \alpha > 1$	yes	yes	Bouman 1993
$p(\beta) = \alpha\beta^2/(1+\alpha\beta^2)$	no	yes	McClure 1987
$p(\beta) = \min\{\alpha\beta^2, 1\}$	no	yes	Geman 1984
$p(\beta) = \sqrt{\alpha+\beta^2}$	yes	yes	Vogel 1987
$p(\beta) = \log(\cosh(\alpha\beta))$	yes	yes	Green 1990
$p(\beta) = \begin{cases} \beta^2/2 & \text{if} \quad \lvert\beta\rvert \leq \alpha, \\ \alpha\lvert\beta\rvert - \alpha^2/2 & \text{if} \quad \lvert\beta\rvert > \alpha. \end{cases}$	yes	yes	Huber 1990

Examples of penalty functions

Discussion

Some neccessary conditions for unbiasedness, sparsity and stability have been derived by Antoniadis and Fan (2001).

- unbiasedness $\leftrightarrow \dot{p}(|\theta|) = 0$ for large $|\theta|$
- sparsity $\leftarrow |\theta| + \lambda \dot{p}(|\theta|) \geq 0$
- stability $\leftrightarrow \text{argmin}\{|\beta| + \lambda \dot{p}(|\theta|)\} = 0$

From the above, a penalty satisfying the conditions on sparsity and stability must be non-smooth at 0. In the two extremes, the hard thresholding rule is discontinuous (instable), while the soft thresholding rule shifts the estimator by an amount of λ even when $|z_i|$ stands way out of noise level, which creates unnecessary bias when θ is large.

SCAD penalty

To ameliorate these drawbacks, we may use the SCAD penalty introduced by Fan (1999),

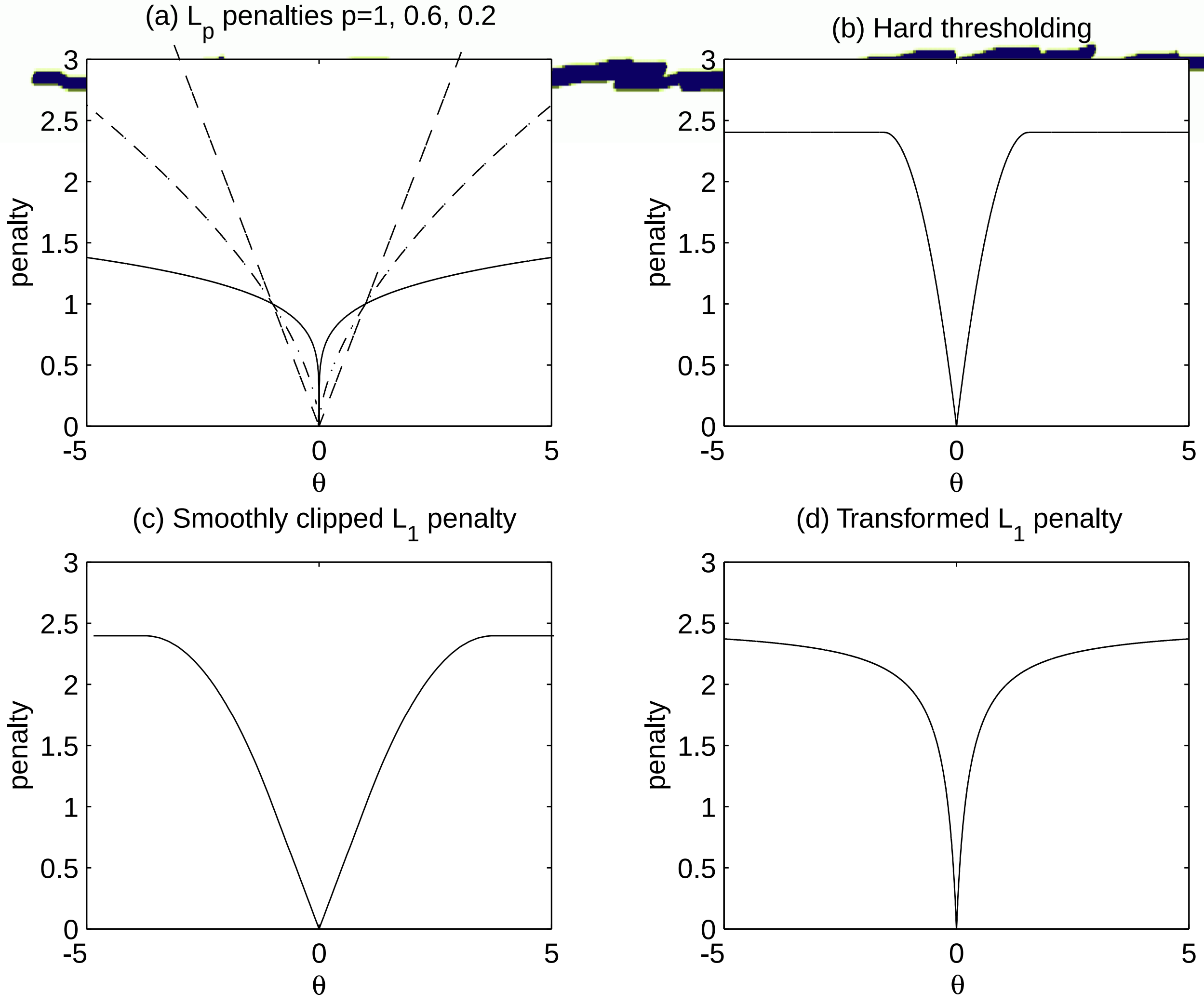
$$p'_\lambda(\theta) = I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda),$$

for $\theta > 0$ and $a > 2$ (usually $a = 3.7$), wich leads to

$$\hat{\theta}_j = \begin{cases} \text{sgn}(z_j)(|z_j| - \lambda)_+ & \text{when } |z_j| \leq 2\lambda \\ \frac{(a-1)z_j - a\lambda\text{sgn}(z_j)}{a-2} & \text{when } 2\lambda < |z_j| \leq a\lambda \\ z_j & \text{when } |z_j| > a\lambda \end{cases}$$

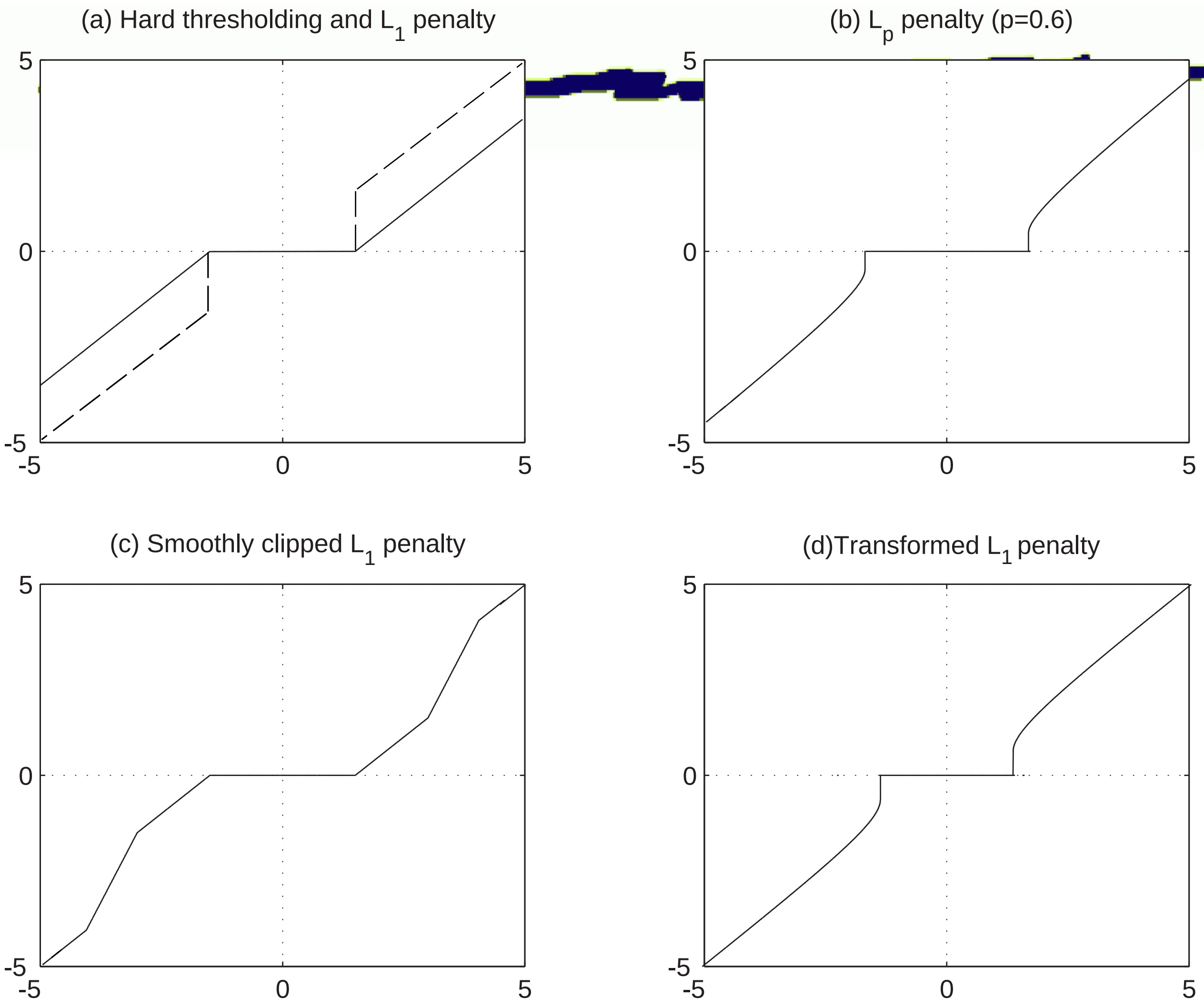
which is concave on $[0, \infty)$ and does not intend to over penalize large $|\theta|$.

Illustration



Plot of penalty functions

Thresholding functions



Plot of thresholding functions

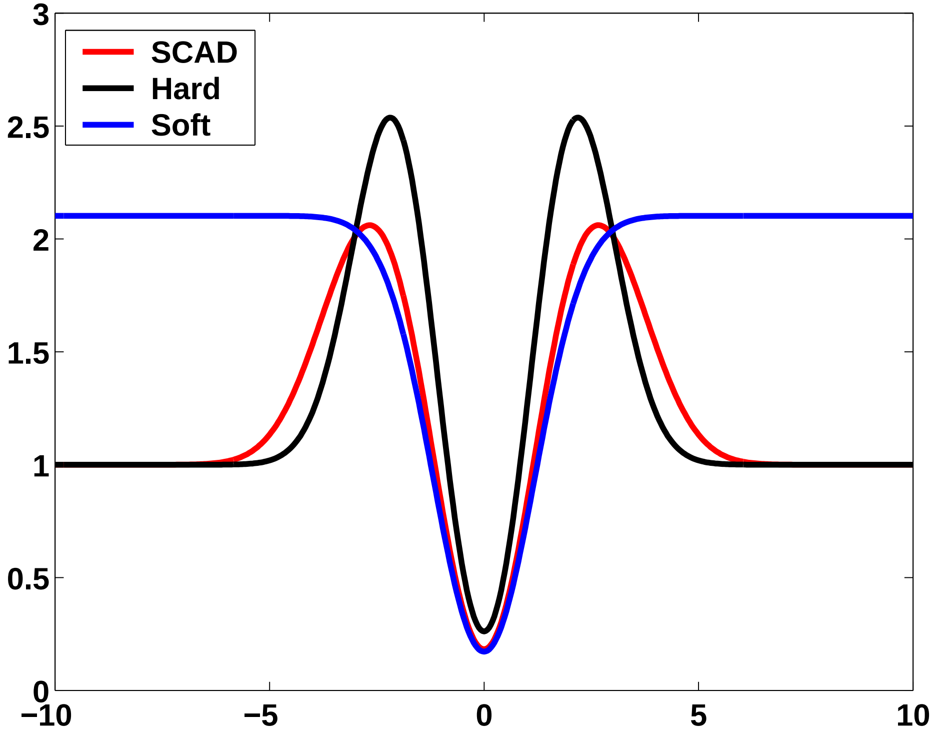
Performance of thresholding rules

To compare the thresholding rules we apply the tool of exact risk analysis. The closed forms for the L^2 risk functions $R(\theta, \lambda, \sigma^2) = \mathbb{E}(\hat{\theta} - \theta)^2$, for the hard and soft-thresholding rules have been derived by Donoho and Johnstone (1994). It is easy to show that

$$R(\theta, \lambda, \sigma^2) = \sigma^2 R(\theta/\sigma, \lambda/\sigma, 1)$$

For simplicity we denote by $R(\theta, \lambda)$ the risk function for $\sigma = 1$. If interested, you can find the expressions of the risk for most of the penalties reviewed above in Antoniadis and Fan (2001) when $Z \sim N(\theta, 1)$.

Risk functions under quadratic loss



Risk functions of three thresholding rules

To make the scale roughly comparable, we took $\lambda = 2$ for the hard-thresholding rule and adjusted the values of λ for the other rules so that their estimated values are the same when $\theta = 3$.

Choosing λ

The performance of the penalized least-squares estimator depends on the regularization parameter λ. Again a convenient way to get a data based estimate of λ is by using generalized cross validation. However the minimization problem we are dealing with is not quadratic and it is not obvious how such a method may be applied.

Tibshirani (1996) proposed a GCV-type criterion for choosing the tuning parameter for the LASSO through a ridge estimate approximation. Of course, this approximation ignores some variability in the estimation process but the simulation study in Tibshirani (1996) suggests that it is a useful approximation.

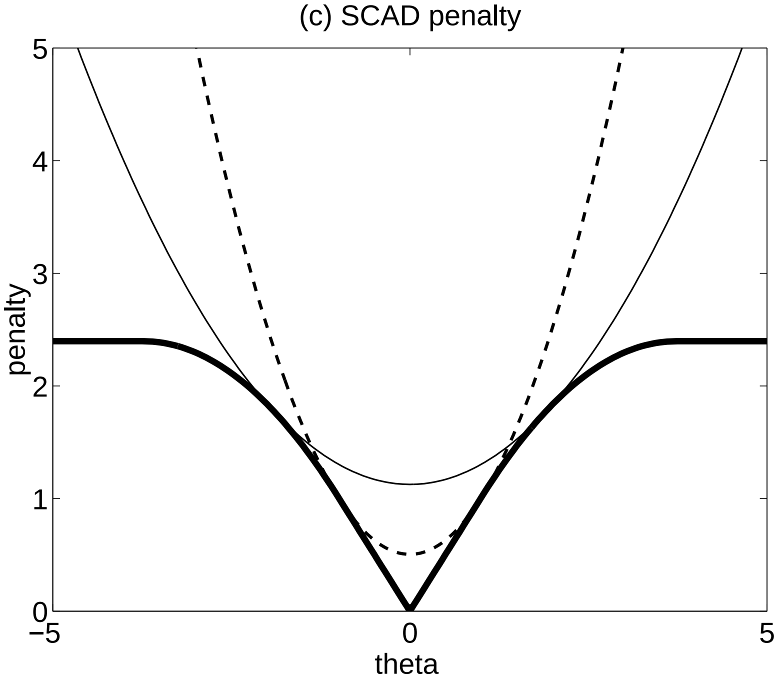
Choosing λ–cont.

The smoothly clipped L_1 penalty p_λ is not differentiable. However it can be locally approximated by a quadratic function as follows. Let θ_0 be a given initial value that is close to the solution of the minimization problem. When $|\theta_0| > 0$, we use

$$p_\lambda(|\theta|) \simeq p_\lambda(|\theta_0|) + \frac{1}{2}\dot{p}_\lambda(|\theta_0|)(\theta^2 - \theta_0^2), \quad \theta \simeq \theta_0$$

and 0 if $\theta_0 = 0$. The figure that follows shows the above approximations for two different values of θ_0.

Quadratic approximation



The penalty and its quadratic approximations

Setting the $\Sigma_\lambda(\boldsymbol{\theta}) = \text{diag}(\dot{p}_\lambda(\theta_i); i \in I_N)$ the solution can be found iteratively computing a ridge regression problem at each iteration. At convergence, one then uses the usual GCV criterion for ridge regression type problems.

Asymptotic properties

Assume that the signal η is in a Besov ball. Because of simple characterization of this space via the wavelet coefficients of its members, the Besov space ball $B^{\nu}_{p,q}(C)$ can be defined as

$$B^{\nu}_{p,q} = \left\{ f \in L_p : \sum_j \left(2^{j(\nu+1/2-1/p)} \|\boldsymbol{\theta}_{j\cdot}\|_p \right)^q < C \right\},$$

where $\boldsymbol{\theta}_{j\cdot}$ is the vector of wavelet coefficients at the resolution level j. Here, ν indicates the degree of smoothness of the underlying signal η.

Note that the wavelet coefficients $\boldsymbol{\theta}$ in the definition of the Besov space are continuous wavelet coefficients. They are approximately a factor of $n^{1/2}$ larger than the discrete wavelet coefficients $W\boldsymbol{\eta}$. This is equivalent to assume that the noise level is of order $1/n$.

Suppose that the penalty function $p_\lambda(\cdot)$ is a nonnegative, nondecreasing and differentiable function in $(0, \infty)$. Further, assume that the function $-\theta - \dot{p}_\lambda(\theta)$ is strictly unimodal on $(0, \infty)$ and that $p'_\lambda(0+) > 0$. Assume also that $\nu + 1/2 - 1/p > 0$. Then, the maximum risk of the penalized least-squares estimator $\hat{\eta}_\lambda$ over the Besov ball $B^\nu_{p,q}(C)$ is of rate $O(n^{-2\nu/(2\nu+1)} \log n)$ when the universal thresholding $\sqrt{2n^{-1} \log n}$ is used.

Some references

Eilers, P.H.C., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). Statistical Sciences, 11, 89–121.

Antoniadis, A. (1997), “Wavelets in Statistics: A Review” (with discussion), *Italian Jour. Statist.*, **6**, 97–144.

Antoniadis, A. and Fan, J. (2001),“Regularization of Wavelets Approximations”, with discussion, J. Amer. Statist. Assoc., 96, No 455, 939-963.

Tibshirani, R. (1995), “Regression shrinkage and selection via the lasso”, *J. Roy. Statist. Soc. Ser. B*, **57**, 267–288.

Vogel, C. R. (1996), “Non-convergence of the L-curve regularization parameter selection method”, *Inverse Problems*, **12**, 535–547.

The overheads of this lecture are available at:
http://www-lmc.imag.fr/lmc-sms/Anestis.Antoniadis/

Software in Matlab implementing theses methods (and many more) has been developed by Antoniadis, Bigot and Sapatinas and is available at
http://www-lmc.imag.fr/SMS/software/wavden

Thank you for your attention.