J. Ital. Statist. Soc. (1997)
2, pp. 131-138

COMMENTS ON «WAVELETS IN STATISTICS: A REVIEW» BY A. ANTONIADIS

Jianqing Fan
University of North Carolina, Chapel Hill and University of California, Los Angeles

Summary

I would like to congratulate Professor Antoniadis for successfully outlining the current state-of-art of wavelet applications in statistics. Since wavelet techniques were introduced to statistics in the early 90's, the applications of wavelet techniques have mushroomed. There is a vast forest of wavelet theory and techniques in statistics and one can find himself easily lost in the jungle. The review by Antoniadis, ranging from linear wavelets to nonlinear wavelets, addressing both theoretical issues and practical relevance, gives in-depth coverage of the wavelet applications in statistics and guides one entering easily into the realm of wavelets.

1. Variable smoothing and spatial adaptation

Wavelets are a family of orthonormal bases having ability of representing functions that are local in both time and frequency domains (subjects to the contraints as in Heisenberg's uncertainty principle). These properties allow one to compress efficiently a wide array of classes of signals, reducing dimensionally from n highly-correlated dimensions to much smaller nearly-independent dimensions, without introducing large approximation errors. These classes of signals include piecewise smooth functions and functions with variable degrees of smoothness and functions with variable local frequencies. The adaptation properties of nonlinear wavelet methods to the Besov classes of functions are thoroughly studies by Donoho, Johnstone, Kerkyacharian and Picard [37, 38]. The adaptation properties of nonlinear wavelet methods to the functions with variable frequencies can be found in Fan, Hall, Martin and Patil [3]. The time and frequency localization of wavelet functions permits nonlinear wavelets methods to conduct automatically variable smoothing: different location uses a different value of smoothing parameter. This feature enables nonlinear wavelet estimators to recover functions with different degrees of smoothness and different local frequencies. Namely, nonlinear wavelet estimators possess spatial adaptation property.

As pointed out in Donoho, Johnstone, Kerkyacharian and Picard [37], linear estimators, including linear wavelet, kernel and spline methods, can not efficiently estimate functions with variable degrees of smoothness. A natural question is if the traditional methods can be modified to efficiently estimate the functions with variable degrees of smoothness. The answer is positive. To recover functions with different degrees of smoothness, variable bandwidth schemes have to be incorporated into kernel or local polynomial estimators, resulting in highly nonlinear estimators. For example, with the implementation of variable bandwidth as in Fan and Gijbels [2], it is shown via simulations that the resulting local linear estimator performs at least comparably with the nonlinear wavelet techniques. See also Fan, Hall, Martin and Patil [4] for the idea of using cross-validation to choose variable bandwidths. However, there is a premium for using the variable bandwidth method: The computational cost can be high. In a seminal work by Lepski, Mammen and Spokoiny [6], it is shown that with a variable bandwidth selection, the kernel regression smoother can also enjoy the optimal rates of convergence for Besov classes of functions in a similar fashion to the nonlinear wavelet estimators.

Nonlinear wavelets and variable bandwidth smoothing are no monopoly in adaptation to variable degrees of smoothness. When variable smoothing is incorporated in smoothing splines, the resulting estimator can also possess spatial adaptation property. See Luo and Wahba [9] for details.

2. Thresholding and subset selection

Thresholding rules have strong connections with model selection in the traditional linear models. Suppose that we have a linear model

$$Y = X\beta + \varepsilon.$$

Then the least-squares estimate is $\hat\beta = (X^TX)^{-1} X^TY$. Now suppose that the columns of X are orthonormal. Then, the least-square estimate in the full model is $\hat\beta = X^TY$, the orthogonal transform of the vector Y. Let $\hat\alpha_i$ be i^{th} smallest value of the vector $|\hat\beta|$. The stepwise deletion algorithm in the linear model is to delete a variable, one at a time, with the smallest absolute t- value. For the orthonormal design matrix, this corresponds to delete the variable with the smallest absolute value of the estimated coefficients. When a variable is deleted, the remaining variables are still orthonormal and the estimated coefficients remain unchanged. So, in the second step, the algorithm deletes the variable that has the second smallest estimated coefficient $\hat\beta$ in the full model. If the stepwise deletion is carried out m times, the remaining variables are those with the largest $n - m$ values of $|\hat\beta|$, namely

$$\left\{i: \left|\hat\beta_i\right| > \lambda\right\}, \quad \text{with } \alpha < \lambda < \alpha_{m+1}$$

Therefore, the stepwise deletion algorithm leads to the hard thresholding rule. Since the wavelet transforms are orthonormal, the hard-thresholding estimator is the same as the least-squares estimator by using the stepwise deletion rule.

The soft-thresholding rule can be viewed similarly. Let us for a moment assume that we have n-dimensional Gaussian white noise model:

$$z = \theta + \varepsilon, \quad \text{with } \varepsilon \sim N\left(0, \sigma^2 I_n\right). \tag{2.1}$$

Suppose that the vector θ is sparse so that it can reasonably be modeled as an i.i.d. realization from a double exponential distribution with a scale parameter λ_l. Then, the Bayesian estimate is to find $\hat{\theta}$ that minimizes

$$\frac{1}{2}\sum_{i=1}^{n}(z_i - \theta_i)^2 + \lambda \sum_{i=1}^{n}|\theta_i| \tag{2.2}$$

where $\lambda = \sigma^2/\lambda_l$. Minimization of (2.2) is equivalent to minimizing (2.2) componentwise. The solution to the above problem yields the soft-thresholding rule.

$$\hat{\theta}_j = sgn(z_j)\left(|z_j| - \lambda\right)_+ .$$

This connection was observed by Donoho, Johnstone, Hoch and Stern [1] and forms the core of the lasso introduced by Tibshirani [10].

The minimization problem (2.1) is closely related to (11) in the review paper with $p = 1$. If the L_l-penalty in (2.1) is replaced by the weighted L_2-loss, then we obtain a shrinkage rule that is similar to equation (12) of the reviewed paper. In wavelet applications, one applies the above method to wavelet coefficients from resolution $J_0 + 1$ to $log_2\, n$. This results in the Donoho and Johnstone soft-shrinkage rule.

We would like to note a few other penalty functions. Consider the more general form of penalized least-squares:

$$\frac{1}{2}\sum_{i=1}^{n}(z_i - \theta_i)^2 + \lambda \sum_{i=1}^{n} p\left(|\theta_i|\right) \tag{2.3}$$

It can be shown that with the discontinuous penality $p_2(\theta) = |\theta| I\left(|\theta| \le \lambda\right) + \lambda / 2I\left(|\theta| > \lambda\right)$, which remains constant for large values of $|\theta|$, the resulting solution is the harding thresholding rule:

$$\hat{\theta}_j = |z_j| I\left(|z_j| > \lambda\right)$$

with the continuous penalty function $P_3(\theta) = \min(|\theta|, \lambda)$, the solution is a mixture of a soft and hard thresholding rule:

$$\hat{\theta}_j = \left(|z_j| - \lambda\right)_+ I\left\{|z_j| \le 1.5\lambda\right\} + |z_j| I\left\{|z_j| > 1.5\lambda\right\}.$$

When the continuous differentiable penaly-function

$$p_4'(\theta) = I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda), \quad for\,\theta > 0\; and\, a > 2$$

is used, the resulting solution is a piecewise linear thresholding:

$$\hat{\theta}_j = \begin{cases} \left(|z_j| - \lambda\right)_+ & \text{when } |z_j| \le 2\lambda \\ \dfrac{(a-1)z_j - a\lambda}{a-2} & \text{when } 2\lambda < |z_j| \le a\lambda \\ z_j & \text{when } |z_j| > a\lambda \end{cases}$$

This thresholding function is in the same spirit to that in Bruce and Gao [20]. The penality function, its derivative and the solution θ_j as a function of z_j are depicted in the following figure.

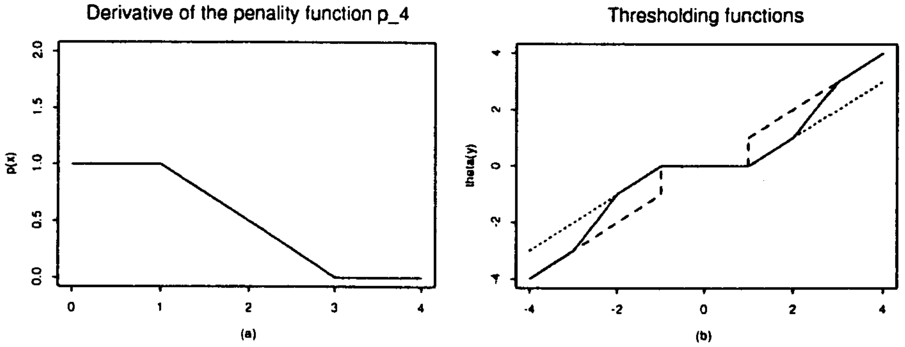

Fig. 1 – (a) Derivative of penalty function p_4; (b) Thresholding functions. Solid line – piecewise linear thresholding, dotted line – soft-thresholding, dashed line – hard thresholding.

3. Robustness and likelihood based models

Because of the localization nature of wavelets transform, the wavelets coefficients are seriously affected by outliers, particularly for those at high-resolution levels. These contaminated coefficients are usually large and they can not be eliminated by thresholding rules. Hence, the wavelet estimators are highly affected by outliners.

It is easier and more interpretable to model directly the unknown function. Let us assume that the collected data (X_i, Y_i) are independent. Conditioning on X_i, Y_i has a density $f_i\big(g(X_i), Y_i\big)$. Model (9) in the review article is a special case of this likelihood modeling with f_i being a normal density. For simplicity, we consider the uniform density case with $x_i = i/n$. Let W be the orthonormal matrix corresponding to a wavelet transform. Let $\theta = Wg$ be the wavelet transform of the vector $g = \big(g(1/n), \cdots, g(n/n)\big)^T$. Then, $g(i/n) = \theta^T w_i$ where w_i is the i^{th} column of W. The penalized likelihood function can be written as

$$\sum_{i=1}^{n} \ell_i\left(\theta^T w_i, y_i\right) - \lambda \sum_{i=m}^{n} |\theta_i|, \tag{3.1}$$

for some thresholding parameter λ. As noted in the last section, when ℓ_i is the normal likelihood, the resulting estimator is the Donoho and Johnstone soft-thrinkage estimator. Thus, the penalized likelihood estimator is an extension of the wavelet shrinkage estimator. It also admits Bayesian interpretation as in (2.2).

When $\ell_i(g, y) = \rho(y - g)$, then (3.1) becomes

$$\sum_{i=1}^{n} \rho\left(y_i - \theta^T w_i\right) - \lambda \sum_{i=m}^{n} |\theta_i|. \tag{3.2}$$

If an outlier-resistant loss function such as the L_1-loss or more generally Huber's psi-function (see Huber [8]) is used, the resulting wavelet estimator is robust.

We now close this section by introducing an iterative algorithm to compute the estimator defined by (3.1). Let us assume that $\ell(t, y)$ are continuous. Suppose that we are given the initial value θ_0 that is close to the minimizer of (3.1). Then, (3.1) can locally be approximated by a quadratic function:

$$\ell(\theta_0) + (\theta - \theta_0)^T \nabla \ell(\theta_0) + \frac{1}{2}(\theta - \theta_0)^T \nabla^2 \ell(\theta_0)(\theta - \theta_0) - \lambda \sum_{i=m}^{n} \theta_i^2 / |\theta_{i0}|, \tag{3.3}$$

where

$$\ell(\theta_0) = \sum_{i=1}^{n} \ell_i\left(\theta_0^T w_i, y_i\right), \quad \nabla \ell(\theta_0) = \sum_{i=1}^{n} \ell_i'\left(\theta_0^T w_i, y_i\right) w_i.$$

and

$$\nabla^2 \ell(\theta_0) = \sum_{i=1}^{n} \ell_i''\left(\theta_0^T w_i, y_i\right) w_i w_i^T.$$

The quadratic minimization problem (3.3) yields the solution

$$\theta_1 = \theta_0 - \left\{\nabla^2 \ell(\theta_0) - 2\lambda\Sigma(\theta_0)\right\}^{-1}\left\{\nabla \ell(\theta_0) - 2\lambda\, sgn(\theta_0)\right\}, \tag{3.4}$$

where

$$\Sigma(\theta_0) = diag\left(0, \cdots 0, \left|\theta_{m+1,0}\right|^{-1}, \cdots, \left|\theta_{n,0}\right|^{-1}\right) \text{ and}$$

$$sgn(\theta_0) = \left(0, \cdots, 0, sgn(\theta_{m+1,0}), \ldots, sgn(\theta_{n,0}), \cdots, sgn(\theta_{n,0})\right)^T.$$

A drawback of the above algorithm is that once a coefficient is shrunk to zero, it will remain zero. The benefit is that it reduces a lot of computation burden. A reasonable initial value θ_0 is to use the soft-threholded wavelet coefficients. This would shrink many coefficients to zero, resulting in a much smaller dimension of minimization problem.

The estimator (3.4) can be regarded as a one-step procedure to the constrained likelihood problem (3.1). Like in parametric case, with good initial value θ_0, the one-step procedure θ_1 can be as efficient as the fully iterative MLE. Now, regarding θ_1 as a good initial value, the next iteration can also be regarded as a one-step procedure and the resulting estimator can still be as efficient as the fully iterative MLE. Therefore, estimators defined by (3.4) after a few iterations can always be regarded as an one-step estimator, which is expected to be as efficient as the fully iterative method as long as the initial estimator is good enough. In this sense, one does not have to iterate (3.4) until it converges.

When the L_1-loss is used in (3.2), one can not directly approximate it by a quadratic equation as (3.3). However, it can be approximated as

$$\sum_{i=1}^{n} \left(y_i - \theta^T w_i\right)^2 / \left|y_i - \theta_0^T w_i\right| + \lambda \sum_{i=m}^{n} \theta_i^2 / \left|\theta_{i0}\right|.$$

From this quadratic approximation, an iterative algorithm can easily be obtained:

$$\theta_1 = \left\{WR(\theta_0)W^T + \lambda\Sigma(\theta_0)\right\}^{-1} WR(\theta_0)Y,$$

where $R(\theta_0) = diag\left(|r_1|^{-1}, \cdots, |r_n|^{-1}\right)$ with $r_i = \left|y_i - \theta_0^T w_i\right|.$

In the penalized likelihood (3.1), one can also use the quadratic penalty if the prior distribution of θ_i is Gaussian instead of double exponential. This leads to the following minimization problem:

$$-\sum_{i=1}^{n} \ell_i\left(\theta^T w_i, y_i\right) + \lambda \sum_{i=m}^{n} \delta_i \theta_i^2 \tag{3.5}$$

for some given δ_i. Note that (3.5) can also be regarded as a constrained MLE with parameter space

$$\left\{\theta: \sum_{i=m}^{n} \delta_i \theta_i^2 \leq Constant\right\},$$

which imposes some smoothness constraints on the underlying function. As in (3.4), the solution to the minimization problem (3.5) can be obtained via the following iterative algorithm:

$$\theta_1 = \theta_0 - \left\{\nabla^2 \ell(\theta_0) - 2\lambda\Sigma_0\right\}^{-1}\left\{\nabla \ell(\theta_0) - 2\lambda\Sigma_0\theta_0\right\},$$

where $\Sigma_0 = diag(\delta_1, \cdots, \delta_n)$.

The above algorithms involve solving equation of form:

$$(WR_1W + R_2)^{-1}a \tag{3.6}$$

for given diagonal matrix R_1 and R_2 with nonnegative elements. A fast algorithm for computing such a vector is needed. One possible way is to use the following iterative algorithm. Let $b = (WR_1W + R_2)^{-1}a$. Then,

$$(WR_1W + R_2)b - a = 0.$$

This suggests the following iterative algorithm for finding b:

$$b_{i+1} = b_i - (\lambda I_n + R_2)^{-1}\,(W^T R_1 W b_i + R_2 b_i - a), \tag{3.7}$$

for some given value of $\lambda > 0$. The operations on the right hand side of equation (3.7) is easy to compute: Sine R_2 is a diagonal matrix, one can explicitly compute the inverse matrix $(\lambda I_n + R_2)^{-1}$. The vector $W^T R_1 W b$ can be computed by discrete wavelet transform and the inverse wavelet transform of the transformed vector multiplied with the diagonal matrix R_1. The effectiveness for this algorithm remains to be seen.

4. Applications to functional data analysis

With advantage of modern technology, data can easily be collected in a form of curves $\{X_i(t_j)\}$ $(i = 1, \cdots, n; j = 1, \cdots, T)$ – the i^{th} observation at time t_j. Such a kind of data are called functional data. For details on functional data analyses, see Ramsey and Silverman [7]. We outline here how wavelets can be used for comparing two sets of curves. Details can be found in Fan and Lin [5].

Suppose that we have two sets of functional data $\{X_i(t_j)\}$ and $\{Y_i(t_j)\}$, collecting at equip-spaced time point t_j. We are interested in testing if the mean curves are the same or not. If the data are only collected at one time point, then the above problem is the standard two-sample t-test problem. We assume that each observed curve is the true mean curve contaminated with stationary stochastic noise. The approach of Fan and Lin [5] can be outlined as follows.

Firstly, apply Fourier transform to each observed curve and obtain the transformed data. The Fourier transform converts stationary errors into nearly independent Gaussian errors. Secondly, compute the two-sample t-test statistic at each coordinate of the transformed data. This basically tests if the two groups have the same mean at each given frequency. The resulting t-test statistics from a T-dimensional vector. When n is reasonably large, this t-vector follows basically the Gaussian model. Our original problem becomes to test if the mean vector is zero or not. Thirdly, apply the wavelet threshold tests in Fan [42] to obtain an overall test-statistic. The role of wavelet thresholding can be regarded as to select powerful coordinates to test. From this, an overall P-value can be obtained.

REFERENCES

[1] Donoho, D. L., Johnstone, I. M., Hock, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object (with discussions). *Jour. Roy. Statist. Soc. B*, **54**, 41-81.

[2] Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statist. Soc. B*, **57**, 371-394.

[3] Fan, J., Hall, P., Martin, M. and Patil, P. (1999). Adaptation to high spatial inhomogeneity using wavelet methods. *Statistica Sinica*, 9, 85-102.

[4] Fan, J., Hall, P., Martin, M. and Patil, P. (1996). On the local smoothing of nonparametric curve estimators. *J. Amer. Statist. Assoc.*, **91**, 258-266.

[5] Fan, J.and Lin, S. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.*, 93, 1007-1021.

[6] Lepski, O. V., Mammen, E., Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, **25**, 929-947.

[7] Ramsay, J. O. and Silverman, B. W. (1997). *The analysis of Functional Data*. Springer-Verlag, New York.

[8] Huber, P. (1981). *Robust estimation*. New York.

[9] Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *Jour. Ameri. Statist. Assoc.*, **92**, 107-116.

[10] Tibshrani, R. (1996). Regression shrinkage and selection via lasso. *Jour. Roy. Statist. Soc. B.*, **58**, 267-288.