Discussion of "Regularization of Wavelets Approximations" by A. Antoniadis and J. Fan

T. Tony Cai*

Department of Statistics

The Wharton School

University of Pennsylvania

Professors Antoniadis and Fan are to be congratulated for their valuable work on the penalized least-squares method for wavelet regression. In this very interesting article the authors present a general characterization of the penalized least-squares estimators for a class of additive smooth penalty functions. A main contribution of this paper is the unified and systematic derivation of the oracle inequalities and minimax properties for a whole class of wavelet estimators. An important advantage of the approach taken by the authors is its generality. As demonstrated nicely by the authors, the penalized least-squares method can be used successfully to treat both equispaced and non-equispaced samples. It can also be readily applied to other nonparametric function estimation problems using wavelets.

My remarks will primarily be confined to the discussion of extensions of the penalized least squares method presented in this paper.

As in the paper, I will use in the following discussion both a single index i and the more conventional double indices $(j,\ k)$ for wavelet coefficients.

*Supported in part by NSF Grant DMS-0072578.

1 Penalized least-squares

The general penalized least-squares can be written as

$$\ell(\boldsymbol{\theta}) = \|\mathbf{Y}_n - \mathbf{A}\boldsymbol{\theta}\|^2 + \lambda p(\boldsymbol{\theta}). \tag{1}$$

Assuming for the moment that the sample is equispaced and the sample size is a power of 2, then the matrix A is the inverse discrete wavelet transform W^{-1} and (1) can be equivalently written as

$$\ell(\boldsymbol{\theta}) = \|\mathbf{Z}_n - \boldsymbol{\theta}\|^2 + \lambda p(\boldsymbol{\theta}), \tag{2}$$

where $\mathbf{Z}_n = W\mathbf{Y}_n$ is the empirical wavelet coefficients. The estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ minimizes $\ell(\boldsymbol{\theta})$ in (2). The performance of the penalized least-squares estimator depends on the penalty $p(\boldsymbol{\theta})$ and the regularization parameter λ. Let us first consider the effect of the penalty function $p(\boldsymbol{\theta})$.

Among possible choices of $p(\boldsymbol{\theta})$, the additive penalty $p(\boldsymbol{\theta}) = \sum_i p(|\theta_i|)$ is the most intuitive choice. When an additive penalty is used, the minimization problem becomes separable. Minimizing (2) is equivalent to minimizing

$$\ell(\theta_i) = \|z_i - \theta_i\|^2 + \lambda p(|\theta_i|)$$

for each coordinate i. The resulting penalized least-squares estimator in this case is separable. That is, the estimate of any coordinate θ_i depends solely on the empirical wavelet coefficient z_i, not on any other coefficients z_j. This is intuitively appealing. However, separable estimators have their drawbacks. The difficulty arises through the need to guard against "false positive" about the presence of true significant coordinates (corresponding to irregularities of the regression function f), and introduced a logarithmic factor into both the thresholding constant λ and the convergence rate. As a result, the estimator is often oversmoothed. The problem is unavoidable for separable estimators, since decisions about individual terms are based on a relatively low level of information. See Cai (1999a, 2000b) and Hall et al. (1999).

Therefore there are true benefits to consider more general penalty function $p(\boldsymbol{\theta})$ in (2). One possibility is to use a blockwise additive penalty. First we divide the coefficients into

non-overlapping blocks and then define a blockwise additive penalty

$$p(\boldsymbol{\theta}) = \sum_b p(\theta_{(b)}) \tag{3}$$

where $\theta_{(b)}$ denotes the coefficient vector in the b-th block. For example, one can divide the coefficients into equal-length blocks of size L. Denote the indices in the b-th block by $(b) = \{(b-1)L+1, \ldots, bL\}$ and denote by $\theta_{(b)} = (\theta_{(b-1)L+1}, \ldots, \theta_{bL})$ and $z_{(b)} = (z_{(b-1)L+1}, \ldots, z_{bL})$ the true and empirical wavelet coefficient vectors in the b-th block, respectively. Let the penalty be

$$p(\theta_{(b)}) = \lambda I\{\theta_{(b)} \neq 0\}. \tag{4}$$

Then the solution to the penalized least-squares problem (2) is a block thresholding estimator:

$$\hat{\theta}_i = z_i I\{\|z_{(b)}\|_{\ell^2}^2 > \lambda\}, \quad \text{for } i \in (b).$$

Block thresholding estimators increase the estimation accuracy by pooling information about neighboring coefficients together to make a simultaneous thresholding decision. The estimators have been shown to enjoy some desirable properties for appropriately chosen block size L and thresholding constant λ. For example the estimators can attain the exact minimax convergence rate without a logarithmic penalty and enjoy good numerical performance (Hall et al. (1999) and Cai (1999b)).

The penalty in (4) is only a particular choice and can be replaced by other penalty functions. The block size can be allowed to vary from level to level. It will be interesting to study the properties of estimators derived from this class of blockwise additive penalty functions using the same unified approach as in this paper.

1.1 From Equispaced to non-equispaced

Wavelet methods for non-equispaced data in the literature are mostly based on interpolation/approximation either in the original function domain or in the wavelet domain. The regularized Sobolev interpolators used in this paper can be regarded as approximation in the wavelet coefficient space. As indicated by Figure 1 in the paper, the regularization parameter s appears to be an important parameter. It will be interesting to find an empirical rule for choosing the value of s.

The approach of the regularized Sobolev interpolators has the advantage of naturally extending the results for equispaced data to non-equispaced data. For example, one can easily extend the penalized least-squares equation (4.3) in the paper to other penalty functions such as a blockwise additive penalty. It will result in a block thresholding estimator based on the synthetic data.

1.2 The choice of the regularization parameter λ

Besides the penalty function $p(\boldsymbol{\theta})$, the regularization parameter λ plays a crucial role in the performance of the resulting estimators. In fact, some of the results in this paper appear to indicate that the performance of the penalized least-squares estimators do not differ significantly for all the smooth penalty functions under consideration. This means that it might be even more important to choose λ in an optimal way than to choose the penalty function. Penalized least-squares has also been considered by Chambolle et al. (1998) in the context of image processing. The regularization parameter λ in Chambolle et al. (1998) is chosen empirically by minimizing an upper bound of the risk. Similar approach can be used here for the general class of smooth penalty functions. The advantage of this approach is that it is computationally simple and fast. Another possibility is to use cross-validation which is computationally more intense.

2 Other function estimation problems

The penalized least-squares approach can be applied to other statistical contexts. In particular it can be used in statistical inverse problems where one observes a function of the interest indirectly. For example, let us consider the estimation of the derivative of a regression function. Suppose we observe

$$y_i = f(i/n) + \epsilon_i, \;\; i = 1, \ldots, n(= 2^J), \; \epsilon_i \overset{iid}{\sim} N(0, \; \sigma^2),$$

and wish to estimate the derivative f' under the mean squared error $E\|\hat{f}' - f'\|_2^2$. Using the Vaguelet-Wavelet Decomposition approach (Abramovich, et al. (1998) and Cai (2000a)), one can first construct an appropriate estimator $\hat{f}$ of f and then use its derivative $\hat{f}'$ as an

estimator of f'. Suppose f has the standard orthonormal wavelet expansion:

$$f(x) = \sum_{k=1}^{2^{j_0}} \xi_{j,k}\phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \theta_{j,k}\psi_{j,k}(x),$$

where ϕ and ψ are the father and mother wavelets respectively. Let

$$\hat{f}(x) = \sum_{k=1}^{2^{j_0}} \hat{\xi}_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} \hat{\theta}_{j,k}\psi_{j,k}(x).$$

Then, under regularity conditions, the risk $E\|\hat{f}' - f'\|_2^2$ of $\hat{f}'$ as an estimator of f' equals

$$\begin{aligned} &E\left(\sum_{k=1}^{2^{j_0}} 2^{j_0}(\hat{\xi}_{j_0,k} - \xi_{j_0,k})\phi'_{j_0,k}(x) + \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} 2^j(\hat{\theta}_{j,k} - \theta_{j,k})\psi'_{j,k}(x) + \sum_{j=J}^{\infty}\sum_{k=1}^{2^j} 2^j\theta_{j,k}\psi'_{j,k}(x)\right)^2 \\ &\asymp 2^{2j_0}E\|\hat{\xi}_{j_0,\cdot} - \xi_{j_0,\cdot}\|_{\ell^2}^2 + \sum_{j=j_0}^{J-1} 2^{2j}E\|\hat{\theta}_{j,\cdot} - \theta_{j,\cdot}\|_{\ell^2}^2 + \sum_{j=J}^{\infty} 2^{2j}\|\theta_{j,\cdot}\|_{\ell^2}^2, \end{aligned}$$

where $\psi'_{j,k}(x) = 2^{j/2}\psi'(2^j x - k)$, $\theta_{j,\cdot} = (\theta_{j,1}, \ldots, \theta_{j,2^j})$, and $\phi'_{j_0,k}$, $\xi_{j_0,\cdot}$ and $\hat{\theta}_{j,\cdot}$ are defined analogously.

Ignoring higher order approximation errors, this problem is in principle equivalent to the following normal mean problem. Given

$$z_{j,k} = \theta_{j,k} + \epsilon_{j,k}, \quad \epsilon_{j,k} \overset{iid}{\sim} N(0,\ n^{-1}\sigma^2),$$

with $k = 1, ..., 2^j,\ j = j_0, \ldots, J-1$, and $2^J = n$, we wish to estimate $\boldsymbol{\theta}$ under the risk

$$R(\widehat{\boldsymbol{\theta}},\ \boldsymbol{\theta}) = \sum_{j=j_0}^{J-1}\sum_{k=1}^{2^j} 2^{2j}E(\hat{\theta}_{j,k} - \theta_{j,k})^2.$$

The minimax convergence rate over a Besov ball $B_{p,q}^r(C)$ in this case is $n^{2(r-1)/(1+2r)}$. The penalized least squares approach can be used to construct near-optimal estimators of $\boldsymbol{\theta}$. For example, using an additive penalty $p_\lambda(\boldsymbol{\theta})$ with $p_0 = (6\log n)^{1/2}$ as defined in Theorem 1 (instead of the standard choice $p_0 = (2\log n)^{1/2}$), the resulting estimator is adaptively within a logarithmic factor of the minimax risk over a wide range of Besov balls $B_{p,q}^r(C)$, assuming that the additive penalty p in (2) satisfies the conditions of Lemma 1. The blockwise additive penalty (4) can also be used here to obtain an exact rate-optimal block thresholding estimator of $\boldsymbol{\theta}$.

This problem can also be reformulated as a more general penalized weighted-least-squares problem. By renormalizing $z_{j,k}$, $\theta_{j,k}$ and $\epsilon_{j,k}$, one can equivalently consider the following heteroscedastic normal mean problem. Given

$$z^*_{j,k} = \theta^*_{j,k} + \epsilon^*_{j,k}, \quad \epsilon^*_{j,k} \overset{iid}{\sim} N(0,\ 2^{2j} n^{-1} \sigma^2),$$

with $k = 1, ..., 2^j,\ j = j_0, \ldots, J-1$, and $2^J = n$, we wish to estimate $\boldsymbol{\theta}^*$ under the conventional mean squared error $E\|\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|^2_{\ell^2}$. In this case, estimators of $\boldsymbol{\theta}^*$ can be constructed by solving a penalized weighted-least-squares problem. The estimator $\widehat{\boldsymbol{\theta}}^*$ minimizes

$$\ell(\boldsymbol{\theta}^*) = \sum_{j=j_0}^{J-1} \sum_k 2^{-2j} (z^*_{j,k} - \theta^*_{j,k})^2 + \lambda p(\boldsymbol{\theta}^*). \tag{5}$$

The weights here are inverse proportional to the variances of $z^*_{j,k}$. For general inverse problems the weights 2^{-2j} in (5) are replaced by a^{-2j} for some $a > 0$.

3 Conclusion

Penalized least-squares provides a unified framework for many seemingly different wavelet thresholding rules in different nonparametric function estimation contexts. This general framework enables us to systematically study a large class of wavelet estimators simultaneously. Clearly, there is much work ahead of us. This paper provides us a means for generating new ideas that can guide our future research in this area. I thank the authors for their clear and imaginative work.

References

[1] Abramovich, F. & Silverman, B.W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115-129.

[2] Cai, T. (1999a). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27**, 898-924.

[3] Cai, T. (1999b). On block thresholding in wavelet regression: adaptivity, block Size, and threshold level. Technical Report, Department of Statistics, Purdue University.

[4] Cai, T. (2000a). On adaptive estimation of a derivative and other related linear inverse problems. *J. Stat. Planning and Inf.*, to appear.

[5] Cai, T. (2000b). On adaptability and information-pooling in nonparametric function estimation. Technical Report, Department of Statistics, University of Pennsylvania.

[6] Chambolle, A., DeVore, R., Lee, N. & Lucier, B. (1998). Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. on Image Processing* **70**, 319-335.

[7] Hall, P., Kerkyacharian, G. & Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9**, 33-50.