Discussion on Antoniadis and Fan "Regularization of Wavelet Approximations"

BRANI VIDAKOVIC[1]

Georgia Institute of Technology

Atlanta, GA 30332-0205

1 Introduction

Anestis Antoniadis and Janqing Fan deserve congratulations for a wonderful and illuminating paper.

Links among wavelet-based penalized function estimation, model selection, and now actively explored wavelet-shrinkage estimation, are intriguing and attracted attention of many researchers. Antoniadis and Fan provide numerous references. The nonlinear estimators resulting as optimal in the process of regularization, for some specific penalty functions, turn out to be the familiar hard- or soft-thresholding rules, or some of their sensible modifications. Simply speaking, the penalty function determines the estimation rule,

[1]The discussant is an Associate Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology, 765 Ferst Drive, Atlanta, GA 30332-0205. This research was supported in part by NSF Grant DMS-0072585 at Georgia Tech.

and in many cases, a practicable and ad-hoc shrinkage rule can be linked to a regularization process under a reasonable penalty function. The authors explore the nature of penalty functions resulting in thresholding-type rules. They also show, that for a large class of penalty functions, corresponding shrinkage estimators are adaptively minimax and have other good sampling properties.

My discussion will be directed toward the link of the regularization problem and Bayesian modeling and inference in the wavelet domain, which is only hinted by Antoniadis and Fan.

2 Bayes Wavelet Modeling

Any decision that is made about the the model, including an estimate, test, or a prediction, should take into account available prior information and possible costs of inaccurate actions. Bayesian decision theory is concerned with devising actions that minimize the average cost to the decision maker using coherently obtained posterior that incorporates both observations and the *a priori* information. Some of the benefits of Bayesian modeling in the wavelet domain are now well understood and a variety of methods, based on Bayes estimation of the "signal part" in an observed wavelet coefficient, are capable of incorporating particular information about unknown signal (smoothness, periodicity, and self-similarity, for instance).

It is now a standard practice in wavelet shrinkage to specify a location model on wavelet coefficients, elicit a prior on their locations (the signal part in wavelet coefficients), exhibit the Bayes estimator for the locations and, if resulting Bayes estimators are shrinkage, apply the inverse wavelet transformation to such estimators.

In considering this model-induced shrinkage the main concern is, of course, performance of induced shrinkage rules, measured by the realized mean square error, while the match between models and data in the wavelet domain is paid no special attention. It is

certainly desirable for selected models to describe our empirical observations well, for majority of signals and images. At the same time, the calculation of shrinkage rules should remain inexpensive. Our experience is that the realistic but complicated models, for which the rules are obtained by expensive simulations, are seldom accepted by the practitioners, despite their reportedly good performance. The two desirable goals *simplicity and reality* can be achieved simultaneously and Bayesian interpretation of regularization provides a way, which is the point of my discussion.

2.1 About Prior Selection

The authors consider a paradigmatic normal location model with known variance, in which a typical wavelet coefficient z is modelled as $\phi(z-\theta)$, where θ is the signal part. The choice of prior θ is often based on inspecting the empirical realizations of coefficients of the pure signals (noiseless data). Lack of intuition on links between function features and nature of wavelet coefficients and great variety of possible signals call for use of automatic priors.

Jim Berger and Peter Müller indicated in late 1993 [personal communication] that priors from the ϵ–contamination family are suitable for the signal part in the wavelet domain since the resulting Bayes rules are "close" in shape to standard thresholding rules. The point mass at zero, δ_0, in

$$\pi(\theta) = \epsilon\delta(0) + (1-\epsilon)\xi(\theta), \tag{1}$$

induces nonlinear shrinkage and models sparsity whereas $\xi(\theta)$ is a "spread" distribution that models wavelet coefficients with large energies (squared magnitudes). This spread distribution can be improper. Besides, adequate changes in ϵ provide a possibility of level-wise adaptive rules.

Various priors on the signal part have been proposed recently by many authors. Pa-

pers by Abramovich *et al.*(1998), Clyde *et al.* (1998), Chipman *et al.* (1997), Vidakovic (1998a), and many others propose priors with different degrees of intricacy, but in spirit similar to the Berger-Müller proposal (1). An overview can be found in Vidakovic (1998b).

Interesting automatic (objective) priors have been proposed as well. Berger and Pericchi (1996) demonstrate that in the context of Bayesian model selection, in testing that the signal part is 0, Jeffreys' prior is:

$$\pi(\theta,\sigma)=\frac{1}{\sigma}\left[\epsilon\delta(0)+(1-\epsilon)\frac{1}{\pi\sigma(1+\theta^2/\sigma^2)}\right],$$

while the intrinsic prior is

$$\pi(\theta,\sigma)=\frac{1}{\sigma}\left[\epsilon\delta(0)+(1-\epsilon)\frac{1-\exp(-\theta^2/\sigma^2)}{2\sqrt{\pi}[\theta^2/\sigma]}\right].$$

The shrinkage rules, involving Bayes factors, in both cases can have simple approximations.

3 MAP-Principle

All information in Bayesian inference is contained in the posterior and posterior location measures (mean, median, mode) are standard Bayes rules for the location parameters. Typically, it is more difficult to exhibit the mean or median of a posterior, then the value at which the posterior is maximized, a posterior mode. This is because for the mean or median, an exact expression for the posterior is needed. MAP rules that maximize the posterior maximize, at the same time, the product of the likelihood and prior, and are typically shrinkage rules.

Given an observation z, the posterior distribution of θ is proportional to

$$\pi(\theta|z)\propto\phi(z-\theta)\cdot\pi(\theta). \tag{2}$$

Let $s(\theta) = -\log \pi(\theta)$ be the score of the prior. Notice that the posterior is is maximized at the same argument at which

$$s(\theta) - \log \phi(z - \theta) = \frac{1}{2\sigma^2}(z - \theta)^2 + s(\theta) \tag{3}$$

is minimized. If $s(\theta)$ is strictly convex and differentiable, the minimizer of (3) is a solution $\hat{\theta}$ of

$$s'(\theta) + \frac{1}{\sigma^2}(\theta - z) = 0.$$

One finds,

$$\hat{\theta} = h^{-1}(z), \quad h(u) = u + \sigma^2 s'(u). \tag{4}$$

Generally, the inversion in (4) may not be analytically feasible, but solution bay be achieved via an approximate sequence of invertible functions. Several examples of prior distributions on θ for which an analytical maximization is possible and authors provide some examples. Some additional solvable cases can be found in Fan (1997), Hyvärinen (1998), and Wang (1999).

For example, if $\pi(\theta) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|\theta|}$, then $s'(\theta) = \sqrt{2}\,\text{sign}(\theta)$, and $\hat{\theta}(d) = \text{sign}(d)\ \max(0, |z| - \sqrt{2}\sigma^2)$.

For

$$\pi(\theta) \propto e^{-a\theta^2/2 - b|\theta|},\ a, b > 0,$$

i.e., if $s'(\theta) = a\,\theta + b\,\text{sign}(\theta)$, the MAP rule is

$$\hat{\theta}(d) = \frac{1}{1 + \sigma^2 a}\text{sign}(d)\ \max(0, |d| - b\,\sigma^2).$$

If π is a "supergaussian" probability density,

$$\pi(\theta) \propto \left[\sqrt{\alpha(\alpha+1)} + \left|\frac{\theta}{b}\right| \right]^{\alpha+3},$$

the corresponding MAP rule is

$$\hat{\theta}(d) = \text{sign}(d) \max\left(0, \frac{|d| - ab}{2} + \frac{1}{2}\sqrt{(|d| + ab)^2 - 4\sigma^2(\alpha+3)}\right), \tag{5}$$

where $a = \sqrt{\alpha(\alpha+1)/2}$, and $\hat{\theta}(d)$ is set to 0 if the square root in (5) is imaginary.

Leporini and Pesquet (1998) explore cases for which the prior is an exponential power distribution [$\mathcal{EPD}(\alpha, \beta)$]. If the noise also has an $\mathcal{EPD}(a, b)$ distribution with $0 < \beta < b \leq 1$, this MAP solution is a hard-thresholding rule. If $0 < \beta \leq 1 < b$ then the resulting MAP rule is

$$\hat{\theta}(d) = d - \left(\frac{\beta a^b}{b\alpha^\beta}\right)^{1/(b-1)} |d|^{(\beta-1)/(b-1)} + o(|d|^{(\beta-1)/(b-1)}).$$

The same authors consider also the Cauchy noise and explore properties of the resulting rules. When the priors are hierarchical (mixtures) Leporini, Pesquet, and Krim (1999) demonstrated that the MAP solution can be degenerated and suggested Maximum Generalized Marginal Likelihood method. Some related derivations can be found in Chambolle et al. (1998) and Leporini and Pesquet (1999).

4 Penalties of Antoniadis and Fan in the MAP context

What are the common properties of priors linked to some penalty functions considered by Antoniadis and Fan? It is interesting that the priors look like histograms of "typical" wavelet coefficients, corresponding to noiseless signals and images. Such empirical densities exhibit sharp, "double exponential-like" peaks around zero and fairly flat tails.

On the other hand, shapes of the priors are in the spirit with standardly used modeling family (1) where the point mass at zero is softened by a peak at zero. The tail parts are in some of the examples improper (flat).

As an illustration we consider the priors corresponding to (AF 2.6), (AF 2.8), (AF 2.11), and the penalty suggested in Fan (1997), $p_\lambda(\theta) = |\theta|\mathbf{1}(|\theta| < \lambda) - \lambda/2\mathbf{1}(|\theta| \geq \lambda)$. They are

$$\begin{aligned}
\pi(\theta) &\propto e^{-\lambda \cdot \min(|\theta|, \lambda)}, \\
\pi(\theta) &\propto e^{-\lambda^2 + (|\theta| - \lambda)^2 \mathbf{1}(|\theta| < \lambda)}, \\
\pi(\theta) &\propto e^{-\lambda b|\theta|(1 + b|\theta|)^{-1}}, \text{ and} \\
\pi(\theta) &\propto e^{-|\theta|\mathbf{1}(|\theta| < \lambda) - \lambda/2\mathbf{1}(|\theta| \geq \lambda)}.
\end{aligned}$$

and are depicted in Figure 1a-d.

In conclusion I point out to some benefits of the MAP-point-of-view on regularized wavelet estimation:

- Honest statistical models whose marginals well match the observations,
- Possible incorporation of prior information, and
- Use of Bayesian machinery to exhibit solutions in cases when simple, closed form solutions are impossible.

Finally, I thank the Editor for his kind invitation to discuss this important paper.

References

[1] Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc. Ser. B*, **60**, 725–749.

[2] Berger, J. and Pericchi, L. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *J. Amer. Statist. Assoc.*,**91**, 109–122.

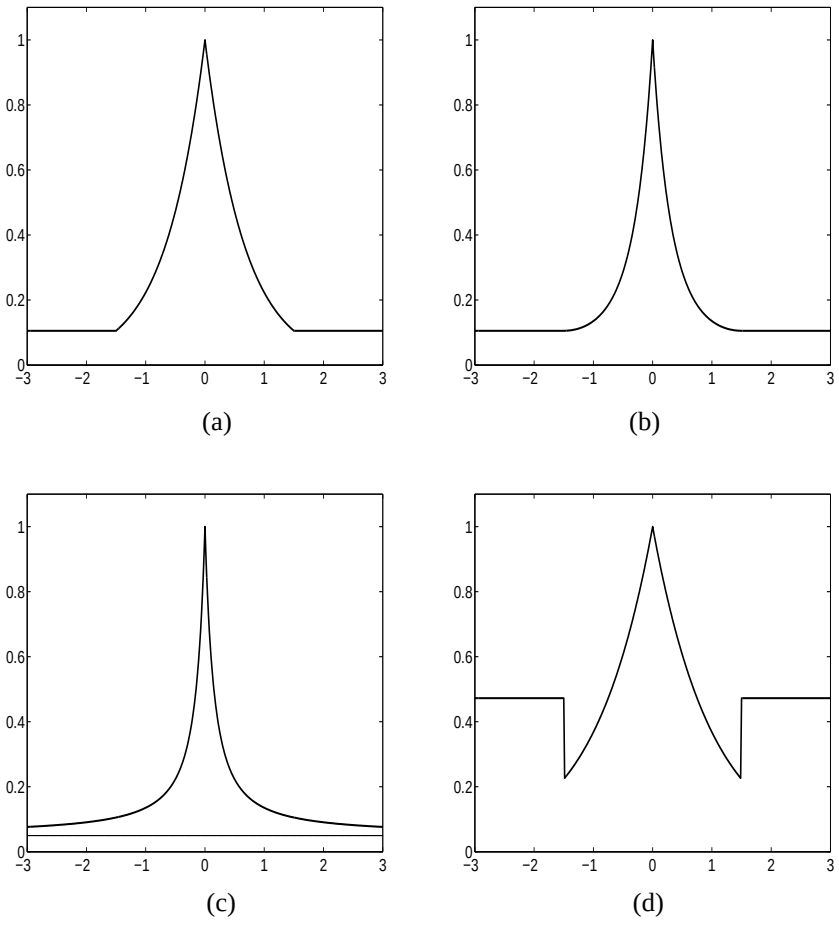

Figure 1: The MAP priors to the penalties (AF 2.6), (AF 2.8), (AF 2.11), and penalty from Fan (1997). In all cases $\lambda = 1.5$ and for the prior in panel (d), $b = 2$.

[3] Chambolle, A., DeVore, R. A., N-Y. Lee N-Y., and Lucier, B. J. (1998). Nonlinear Wavelet Image Processing: Variational Problems, Compression, and Noise Removal through Wavelet Shrinkage, *IEEE Trans. Image Process.*,**7**, 319–355.

[4] Chipman, H., Kolaczyk, E., and McCulloch, R. (1997). Adaptive Bayesian Wavelet Shrinkage. *J. Amer. Statist. Assoc.*,**92**, 1413–1421.

[5] Clyde, M. A., Parmigiani, G., and Vidakovic, B. (1998). Multiple Shrinkage and Subset Selection in Wavelets. *Biometrika,* **85**, 391–402.

[6] Fan, J. (1997), Comment on "Wavelets in Statistics: A Review" by A. Antoniadis, *Italian Jour. Statist.*,**6**, 97–144.

[7] Hyvärinen, A. (1998). Sparse Code Shrinkage: Denoising of Nongaussian Data by Maximum Likelihood Estimation, Technical Report A51, Helsinki University of Technology, Finland.

[8] Leporini, D., and Pesquet, J.-C. (1998). Wavelet Thresholding for a Wide Class of Noise Distributions, EUSIPCO'98, Rhodes, Greece, September 1998, 993–996.

[9] Leporini, D., and Pesquet, J.-C. (1999). Bayesian wavelet denoising: Besov priors and non-gaussian noises, *Signal Processing,* **81**, 55–67.

[10] Leporini, D., Pesquet, J.-C., Krim, H. (1999). Best Basis Representations with Prior Statistical Models, In: *Bayesian Inference In Wavelet Based Models*, Editors P. Müller and B. Vidakovic, Lecture Notes in Statistics,**141**, 109–113, Springer Verlag, New York.

[11] Pesquet, J.-C., Krim, H., Leporini, D., and Hamman, E. (1996). Bayesian approach to best basis selection, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 7-10 May, Atlanta, GA. **5**, 2634–2637.

[12] Vidakovic, B. (1998a). Nonlinear Wavelet Shrinkage With Bayes Rules and Bayes Factors, *Journal of the American Statistical Association,* **93**, 441, 173–179.

[13] Vidakovic, B. (1998b). Wavelet-based nonparametric Bayes methods. In: Practical Nonparametric and Semiparametric Bayesian Statistics. Editors D. Dey, P. Müller and D. Sinha, Lecture Notes in Statistics **133** , 133–155, Springer-Verlag, New York.

[14] Wang, Y. (1999). An Overview of Wavelet Regularization, In: *Bayesian Inference In Wavelet Based Models*, Editors P. Müller and B. Vidakovic, Lecture Notes in Statistics,**141**, 109–113, Springer Verlag, New York.