LARGER POSTERIOR MODE WAVELET THRESHOLDING AND APPLICATIONS

LUISA CUTILLO,
Department of Mathematics and Applications, University of Naples "Federico II",
Naples, Italy
Email: l.cutillo@dma.unina.it

YOON YOUNG JUNG,
Department of Statistics, Ewha Womans University,
Seoul, Korea
Email: yjung@isye.gatech.edu

FABRIZIO RUGGERI,
IMATI - CNR,
Milan, Italy
Email: fabrizio@mi.imati.cnr.it

and

BRANI VIDAKOVIC,
School of Industrial and Systems Engineering, Georgia Institute of Technology,
Atlanta, USA
Email: brani@isye.gatech.edu

July 1, 2005

Abstract

This paper explores the thresholding rules induced by a variation of the Bayesian MAP principle. The MAP rules are Bayes actions that maximize the posterior. The proposed rule is thresholding and always picks the mode of the posterior larger in absolute value, thus the name LPM. We demonstrate that the introduced shrinkage performs comparably to several popular shrinkage techniques. The exact risk properties of the thresholding rule are explored, as well. We provide extensive simulational analysis and apply the proposed methodology to real-life experimental data coming from the field of Atomic Force Microscopy.

1 INTRODUCTION

The Bayesian paradigm has become very popular in wavelet-based data processing. In addition to possible incorporation of prior information about the unknown signal by Bayesian models, the resulting Bayes rules are usually shrinkers. For example, in location models the Bayesian estimator shrinks toward the prior mean, usually zero. This shrinkage property holds in general, although examples of Bayes rules that expand can be constructed, see Vidakovic and Ruggeri (1999). The Bayes rules can be constructed to mimic the traditional wavelet thresholding rules: to shrink the large coefficients slightly and shrink the small coefficients heavily. Furthermore, most practicable Bayes rules should be easily computed by simulation or expressed in a closed form.

One statistical task in which the wavelets are successfully applied is recovery of an unknown signal $\mathbf{f}$ imbedded in Gaussian noise ϵ. Wavelet transform $\mathbf{W}$ is applied to noisy measurements $y_i = f_i + \epsilon_i, \;\; i = 1, \dots, n,$ or, in vector notation, $\mathbf{y} = \mathbf{f} + \epsilon$. The linearity of transformation $\mathbf{W}$ implies that the transformed vector $\mathbf{d} = \mathbf{W}(\mathbf{y})$ is the sum of the transformed signal $\theta = \mathbf{W}(\mathbf{f})$ and transformed noise $\eta = \mathbf{W}(\epsilon)$. Furthermore, orthogonality of $\mathbf{W}$ and normality of the noise vector ϵ imply that the transformed noise vector η is normal, as well.

Bayesian estimation is applied in the wavelet domain, i.e., after the data have been transformed. The wavelet coefficients can be modeled in totality, as a single vector, or individually, due to the decorrelating property of wavelet transforms. In this paper we model wavelet coefficients individually, i.e., elicit a model on a typical wavelet coefficient.

Thus, we drop the standard wavelet indexing and concentrate on the model: $d = \theta + \epsilon$. Bayesian methods are applied to estimate the location parameter θ, which will be, in the sequel, retained as the shrunk wavelet coefficient and back transformed to the data domain. Various Bayesian models have been proposed in the literature. Some models have been driven by empirical justifications, others by pure mathematical considerations; some models lead to simple and explicit rules, others require extensive Markov Chain Monte Carlo simulations. Reviews of some early Bayesian approaches can be found in Abramovich and Sapatinas (1999), Vidakovic (1998, 1999) and Ruggeri and Vidakovic (2005). Müller and Vidakovic (1999) provide an edited volume on various aspects of Bayesian modeling in the wavelet domain.

In this paper we explore thresholding rules induced by a variation of the Bayesian MAP principle. MAP rules are Bayes actions that maximize the posterior. In all models considered in this paper the posterior is infinite at zero, i.e., zero is trivially the mode of the posterior. If no other modes exist, zero is the Bayes action. If the second, non-zero mode of the posterior exists, this mode is taken as the Bayes action. Such a rule is thresholding and always picks the mode larger in absolute value if such local mode exists. This is the motivation for the name LPM - Larger (in absolute value) Posterior Mode. We demonstrate that thresholding induced by replacing the wavelet coefficient with the larger posterior mode of the corresponding posterior, performs well compared to several popular shrinkage techniques.

The paper is organized as follows. In Section 2 the basic model is described, the LPM rule is derived, and the exact risk properties of the LPM rule are discussed. Section 3 discusses two models that generalize the model from Section 2 by relaxing the assumption of known variance. Derivations of LPM rules corresponding to these two models are deferred to the Appendix. Comprehensive simulations and comparisons are provided in Section 4. This section also contains discussion on the selection of hyperparameters and a real-life application of the introduced shrinkage. We conclude the paper with Section 5 in which some possible directions for future research are outlined.

2 LARGER POSTERIOR MODE (LPM) WAVELET THRESHOLDING

As is commonly done in Bayesian wavelet shrinkage, a Bayesian model is proposed on observed wavelet coefficients. Due to the decorrelation property of wavelet transforms, the coefficients are modeled individually and independently. In the exposition that follows, the double index jk representing scale/shift indices is omitted and a "typical" wavelet coefficient, d, is considered. Therefore, our model starts with

$$d = \theta + \epsilon, \tag{1}$$

where we are interested in the location θ corresponding to the signal part contained in the observation d.

Bayes rules under the squared error loss and regular models often result in shrinkage rules resembling thresholding rules, but they are never thresholding rules. In many applications rules of the thresholding type are preferable to smooth shrinkage rules. Examples include model selection, data compression, dimension reduction, and related statistical tasks in which it is desirable to replace by zero a majority of the processed coefficients.

This paper considers construction of bona fide thresholding rules via selection of a larger (in absolute value) posterior mode (LPM) in a properly set Bayesian model. The models considered in this paper produce posteriors with no more than two modes. The selected mode is either zero (a single mode – thus trivially the larger) or non-zero mode if the posterior is bimodal.

2.1 DERIVATION OF THE THRESHOLDING RULE

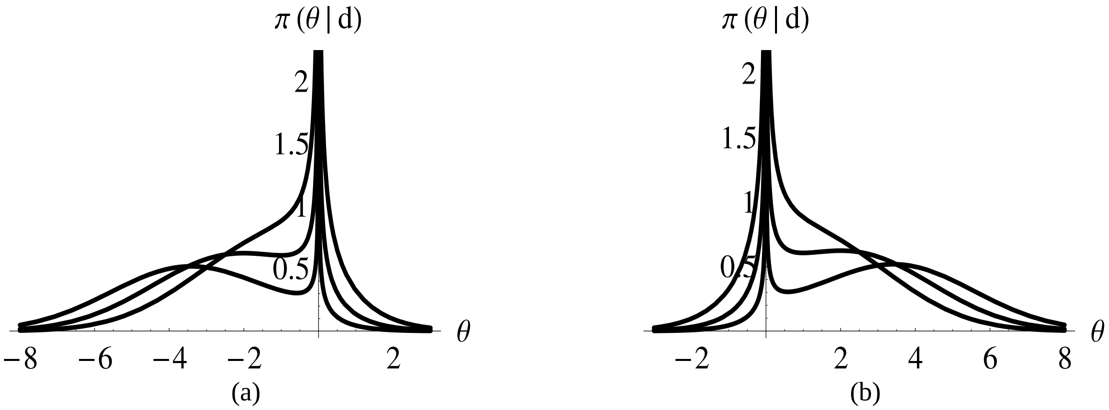
We consider several versions of the model. In the basic version, discussed in this Section, the variance of the noise is assumed known and a prior is elicited only on the unknown location. This version of the model can be found in Robert (2001) in the context of Bayesian credible intervals. In the generalized versions discussed in the following section, the variance of the noise is not assumed known and will be modeled by (i) inverse-gamma and (ii) exponential priors which are independent from the location parameter.

Consider the model

$$\begin{array}{rcl} d|\theta & \sim & \mathcal{N}(\theta, \sigma^2), \\ \theta|\tau^2 & \sim & \mathcal{N}(0, \tau^2), \\ \tau^2 & \sim & (\tau^2)^{-k},\ k > 0, \end{array} \tag{2}$$

where the variance σ^2 is assumed known and in practice estimated from the data and plugged in the model. We seek a MAP solution, i.e., an estimator of θ that (locally) maximizes the posterior, $\pi(\theta|d)$. To find the extrema of the posterior on θ we note that the posterior is proportional to the joint distribution of d,θ and τ^2, so the value of θ maximizing the joint distribution maximizes the posterior, as well. The joint distribution is proportional to

$$\begin{aligned}
p(d,\theta) &= \int p(d|\theta)p(\theta|\tau^2)p(\tau^2)d\tau^2 \\
&= \int \frac{1}{\sqrt{2\pi}\sigma}e^{-(d-\theta)^2/(2\sigma^2)}\frac{1}{\sqrt{2\pi\tau^2}}e^{-\theta^2/(2\tau^2)}\frac{1}{(\tau^2)^k}d\tau^2 \\
&= \frac{1}{2\pi\sigma}e^{-(d-\theta)^2/(2\sigma^2)}\int(\tau^2)^{-(k+1/2)}e^{-\theta^2/(2\tau^2)}d\tau^2 \\
&= \frac{1}{2\pi\sigma}e^{-(d-\theta)^2/(2\sigma^2)}\int y^{(k-1/2)-1}e^{-\theta^2 y/2}dy \\
&= \frac{1}{2\pi\sigma}e^{-(d-\theta)^2/(2\sigma^2)}\frac{\Gamma(k-1/2)}{(\theta^2/2)^{k-1/2}}, k > 1/2
\end{aligned}$$



(a)

(b)

Figure 1: Posterior distribution for $k = 3/4$ and $\sigma^2 = 2^2$; (a) $d = -4, -3, -2$; (b) $d = 2, 3, 4$. The unimodal density graphs in panels (a) and (b) correspond to $k = -2, 2$, respectively.

This leads to posterior

$$p(\theta|d) \propto p(d,\theta) \propto e^{-(d-\theta)^2/(2\sigma^2)}\ |\theta|^{-2k+1}. \tag{3}$$

Figure 1 (a,b) depicts the posterior distribution for $k = 3/4$, $\sigma^2 = 2^2$, and various values of d. Note that if d is small in absolute value compared to σ^2, the posterior is unimodal with (infinite) mode at zero. For $|d|$ large, the posterior is bimodal with non-zero mode sharing the same sign as the observation d.

The logarithm of the posterior is proportional to

$$\ell = \log p(\theta|d) \propto -\frac{(d-\theta)^2}{2\sigma^2} + (1-2k)\log\theta,$$

and has extrema at the solutions of a quadratic equation,

$$\theta^2 - d\theta + \sigma^2(2k-1) = 0,$$

$$\theta_{1,2} = \frac{d \pm \sqrt{d^2 - 4\sigma^2(2k-1)}}{2}.$$

The roots $\theta_{1,2}$ are real if and only if $d^2 \geq 4\sigma^2(2k-1)$, i.e., if $|d| \geq 2\sigma\sqrt{2k-1} = \lambda$. If this condition is not satisfied, then the likelihood is decreasing in $|\theta|$ and the MAP is given by $\hat{\theta} = 0$.

The value of the posterior at zero is infinite, thus zero is always a mode of the posterior. When this is the only mode, the resulting rule takes value zero. If the second, non-zero mode exists, then this mode is taken as the Bayes action.

We assume, WLOG, $d > 0$. Since $k > 1/2$, $\sqrt{d^2 - 4\sigma^2(2k-1)} < d$ and both roots are positive and smaller than d, we have shrinkage. Then the LPM is $\frac{d+\sqrt{d^2-4\sigma^2(2k-1)}}{2}$, since the posterior is decreasing from 0 to smaller root, increasing between the two roots and decreasing after the larger root. For arbitrary d, and $\lambda = 2\sigma\sqrt{2k-1}$, the LPM rule is

$$\hat{\theta} = \frac{d + \text{ sign}(d)\,\sqrt{d^2 - 4\sigma^2(2k-1)}}{2}\mathbf{1}(|d| \geq \lambda). \tag{4}$$

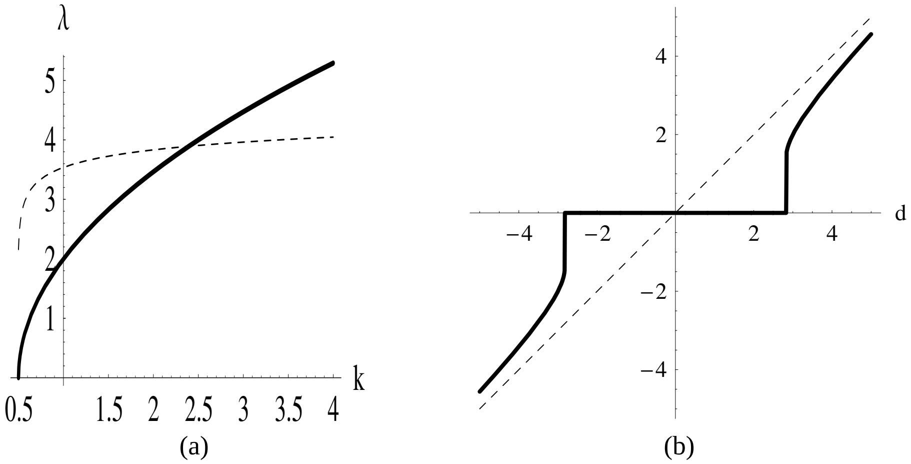


Figure 2: (a) Influence on the threshold λ by power parameter k; (b) LPM thresholding rule.

Figure 2 (a) compares values of threshold λ to properly scaled universal threshold (Donoho and Johnstone, 1994). In both cases the variance $\sigma^2 = 1$. The dotted line represents the values of universal threshold rescaled by $n = (k - 1/2) \cdot 2^{10}$. This sample size n is selected only for comparison reasons. As depicted in Figure 2 (b), the thresholding rule looks like a compromise between hard and soft thresholding, The rule generally remains close to d for intermediate and large values of d.

Note that the posterior (3) is proper (integrable at 0) if and only if $2k - 1 < 1$, i.e., when $k < 1$. The existence of a finite second mode does not require the posterior to be proper and we will consider all $k > 1/2$.

Remark. If the square root in (4) is approximated by Taylor expansion of the first order, $(1-u)^\alpha \approx 1 - \alpha u,$ the LPM rule mimics James-Stein estimator,

$$\hat{\theta} \approx \left(1 - \frac{\sigma^2(2k-1)}{d^2}\right)_+ d,$$

which is considered extensively in the wavelet shrinkage literature.

2.2 EXACT RISK PROPERTIES OF LPM RULES

The exact risk analysis of any proposed shrinkage rule has received considerable attention in the wavelet literature since it allows for comparison of different wavelet-based smoothing methods. When the rule is given in a simple form, the exact risk analysis can be carried out explicitly. For instance, Donoho and Johnstone (1994) and Bruce and Gao (1996) provide exact risk analyses for hard and soft thresholding under squared error loss. Gao and Bruce (1997) give a rationale for introducing the "firm" or "semi-soft" thresholding utilizing exact risk arguments. The goal of exact risk analysis is to explore robustness in risk, bias, and variance when the model parameters and hyper-parameters change.

For our model the analytic form of LPM rule (4) is more complex and the exact risk analysis was carried out numerically. Computations performed in the software package MATHEMATICA produced Figure 3. We briefly describe the properties inferred from Figure 3.

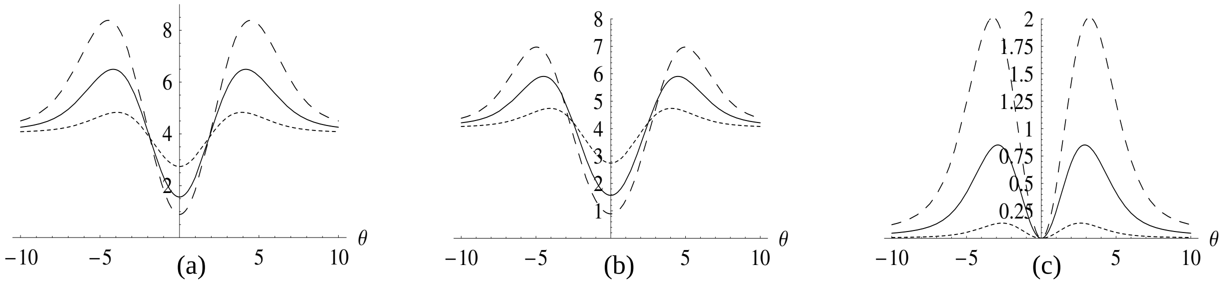


Figure 3: Exact risk plots for LPM rule, for $k = 0.6$ (short dash), $k = 0.75$ (solid), and $k = 0.9$ (lomg dash). For all three cases $\sigma^2 = 2^2$. (a) Risk; (b) Variance, and (c) Bias squared.

In Figure 3(a) the risks of rule (4) for $k = 0.6, 0.75,$ and $0.9,$ are presented. These risks are partitioned to variances and biases-squared given in panels Figure 3(b) and Figure 3(c). The shapes of risks are typical for hard thresholding rules. The risk is minimal at $\theta = 0$ and it stabilizes about the variance for $|\theta|$ large. For values of θ that are comparable to the threshold λ the risk is maximized. This signifies that largest contribution to the MSE is for the values of θ close to the threshold. This is to be expected since for θ's close to threshold, given that the noise averages to 0, the largest errors are made by the "keep-or-kill" policy. The variance plots Figure 3(b) generally resemble the plots for the risk. As is typical for hard thresholding rules, the squared bias Figure 3(c) is small in magnitude compared to variance and risk. This is a desirable property when the users are concerned about the bias of the rule and ultimately, the estimator $\hat{\mathbf{f}}$.

We note that the role of k in the shapes of risk, variance, and bias-squared is linked to the role of sample size and increased variance in standard shrinkage situations. This link will be discussed further in Section 4.

3 GENERALIZATIONS

In the previous we assumed that the variance of the noise, σ^2 was known. In applications, this variance is estimated from the data (usually using the finest level of detail in the wavelet decomposition) and plugged in the shrinkage rule. In this section we generalize the methodology by assuming the prior distribution on unknown variance.

We consider two generalizations of the model in (3). In the first, the variance is assigned an exponential prior, leading to a double exponential marginal likelihood, while in the second, the variance is assigned an inverse gamma prior, leading to a t marginal likelihood.

3.1 MODEL 1: EXPONENTIAL PRIOR ON UNKNOWN VARIANCE.

Assume that for a typical wavelet coefficient d the following model holds.

$$\begin{aligned}
d|\theta,\sigma^2 &\sim \mathcal{N}(\theta,\sigma^2),\\
\sigma^2 &\sim \mathcal{E}\left(\frac{1}{\mu}\right) \text{ with density } p(\sigma^2|\mu)=\mu e^{-\mu\sigma^2}, \mu>0,\\
\theta|\tau^2 &\sim \mathcal{N}(0,\tau^2),\\
\tau^2 &\sim (\tau^2)^{-k}, k>0.
\end{aligned}$$

It is well known that an exponential scale mixture of normals results in a double exponential distribution. Thus this model is equivalent to

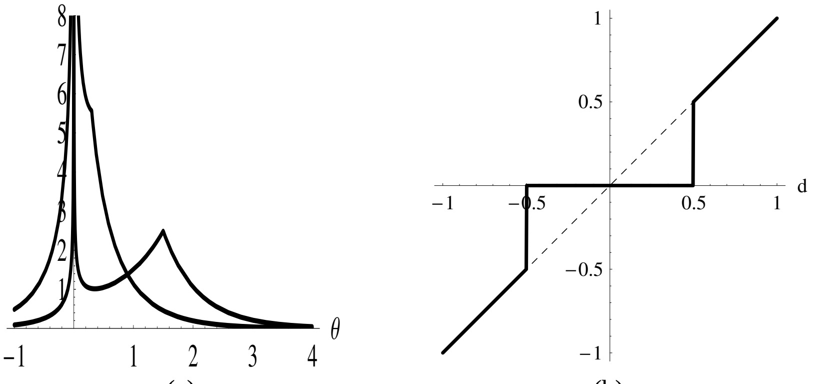
$$\begin{aligned}
d|\theta,\mu &\sim \mathcal{DE}\left(\theta,\frac{1}{\sqrt{2\mu}}\right), \quad \text{with density } f(d|\theta)=\frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|},\\
\theta|\tau^2 &\sim \mathcal{N}(0,\tau^2),\\
\tau^2 &\sim (\tau^2)^{-k}, k>0.
\end{aligned}$$

The resulting LPM rule turns out to be hard-thresholding,

$$\hat{\theta}=d\,\mathbf{1}(|d|\geq\lambda) \tag{5}$$

with $\lambda=\frac{2k-1}{\sqrt{2\mu}}$. Derivation of this fact is deferred to the Appendix.

Figure 4(a) shows the posterior distribution for Model 1 for values of d leading to unimodal (infinite at mode 0) and bimodal cases. The values are $d=0.3$ and $d=1.5$, $k=0.75$ and $\mu=1$. The LPM rule (5) is shown in Figure 4(b) for $k=0.75$ and $\mu=1/2$.



(a) (b)

Figure 4: (a) Influence on the posterior in Model 1 by two different values of d; (b) LPM rule for Model 1.

The double exponential marginal likelihood is a realistic model for wavelet coefficients. In fact, if a histogram of wavelet coefficients for many standard signals is plotted, it resembles a double exponential distribution. This observation first explicitly stated by Mallat (1989), is used in many Bayesian models in the wavelet domain, examples are BAMS wavelet shrinkage (Vidakovic and Ruggeri, 2001) or the wavelet image processing methodology of Simoncelli and coauthors (e.g., Simoncelli and Adelson, 1996).

3.2 MODEL 2: INVERSE GAMMA PRIOR ON UNKNOWN VARIANCE.

The inverse gamma prior on the unknown variance of a normal likelihood is the most common and well understood prior. The resulting marginal likelihood on the wavelet coefficients is t, which models heavy tails of empirical distributions of wavelet coefficients well. Model 2 with an inverse gamma prior will not realistically model the behavior of wavelet coefficients in the neighborhood of 0, but will account for heavy tails encountered in empirical distributions of wavelet coefficients. Model 2 is given by

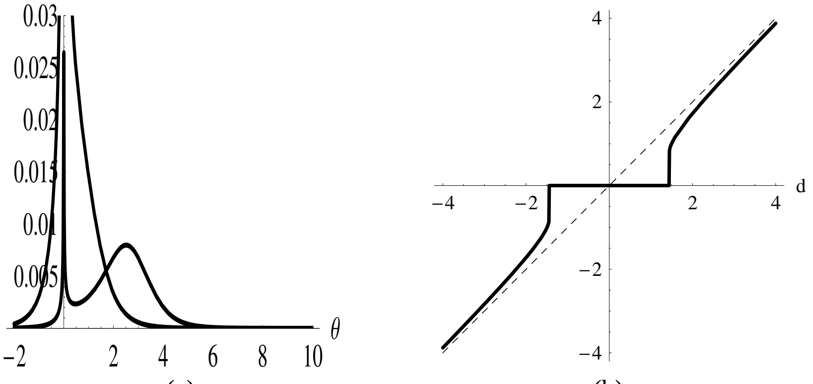
$$\begin{aligned} d|\theta,\sigma^2 &\sim \mathcal{N}(\theta,\sigma^2), \\ \sigma^2 &\sim \mathcal{IG}(\alpha,\beta) \text{ with density } p(\sigma^2|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-1-\alpha} e^{\frac{-\beta}{\sigma^2}}, \alpha > 0, \beta > 0, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, k > 0. \end{aligned}$$

The resulting LPM rule is

$$\hat{\theta} = \frac{(2\alpha+4k-1)d + \text{ sign}(d)\sqrt{(2\alpha+1)^2 d^2 + 16(1-2k)(k+\alpha)\beta}}{4(k+\alpha)}\,\mathbf{1}(|d| \geq \lambda), \tag{6}$$

where

$$\lambda = \frac{2}{2\alpha-1}\,\sqrt{(2k-1)(k+\alpha)\beta}\,.$$



(a)

(b)

Figure 5: (a) Influence on the posterior in Model 2 by two different values of d; (b) LPM rule for Model 2.

Figure 5(a) shows the posterior distribution for the Model 2 for values of d leading to unimodal and bimodal cases. The values are: $d = 0.7$ and $d = 2.7$, $k = 0.85$, $\alpha = 2.5$ and $\beta = 2$. The LPM rule (6) is shown in Figure 5(b) for $k = 0.85$, $\alpha = 2.5$ and $\beta = 2$.

4 SIMULATIONS AND APPLICATIONS

In this section we apply the proposed thresholding rules. First, we discuss the selection of the hyperparameters for each model. This is important for an automatic application of the methodology. Next, we compare performance of the proposed rules to eight other commonly used methods (both global and adaptive). Finally, we apply the shrinkage methodology to a real-life example involving Atomic Force Microscopy.

4.1 SELECTION OF HYPERPARAMETERS

In any Bayesian modeling task the selection of the hyperparameters is instrumental for good performance of the model. It is also desirable to have an automatic way to select the hyperparameters, thus making the shrinkage procedure automatic, i.e., free of subjective user intervention. More about specification of hyperparameters in Bayesian models in the wavelet denoising context can be found in Chipman, Kolaczyk, and and McCulloch (1997), Vidakovic and Ruggeri (2001), and Angelini and Sapatinas (2004), among others.

The hyperparameters should be selected so that the resulting methodology is robust with respect to a wide range of input signals (sample sizes, signal regularity, size of noise, etc). In contemporary wavelet practice the values of the hyperparameters are usually assessed by empirical Bayes arguments due to enormous variability of potential input data (Clyde and George, 1999; 2000). Straightforward Empirical Bayes techniques such as predictive moment matching, or MLII method, are most commonly used efficient methods for hyperparameter specification. In this paper we determine hyperparameters by moment-matching.

In this paper we consider three Bayesian models on wavelet coefficients, (i) the basic model with σ^2 assumed known, and two generalizations in which the variance σ^2 is modeled by (ii) exponential and (iii) inverse-gamma priors. The elicitation of corresponding hyperparameters is discussed for each case.

The Basic Model. In the basic model the only hyperparameter is the *power parameter* k. Even though the proper posterior is obtained for $k < 1$, the existence of the second, non-zero mode does not depend on the "properness" of the posterior. Thus we will consider all $k > 1/2$. Note that the condition $k > 1/2$ is needed to ensure that Gamma function $\Gamma(2k-1)$ is finite and non-negative.

The sample size of the input signal should influence our selection of k. Figure 6 shows the Bumps signal at SNR =5 and sample sizes $n = 512, 1024, 2048,$ and 4096, smoothed by LPM thresholding rule (4) for various values of k. The minimum AMSE is achieved at $k = 1.0, 1.2, 1.4,$ and 1.6 respectively. Thus the increasing of the sample size increases the optimal k.

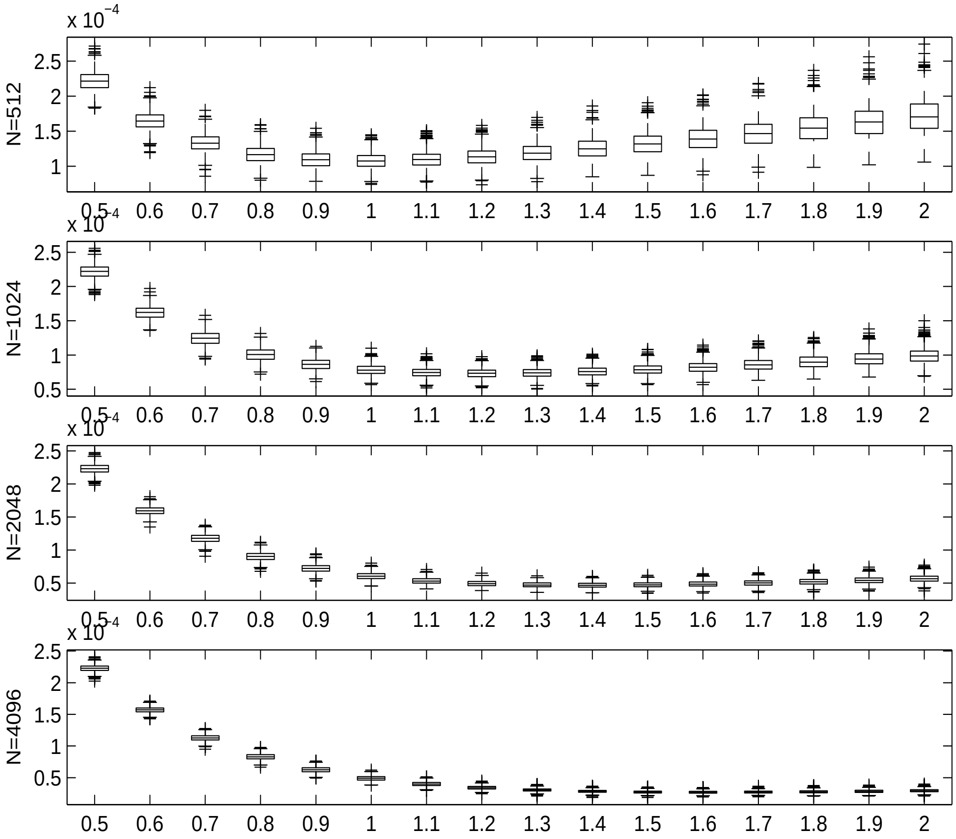


Figure 6: The AMSE for the Bumps function for four different sample sizes $n = 512$ (top), $n = 1024$, $n = 2048$, $n = 4096$ (bottom), evaluated at different values of power parameter k. The level of noise is such that SNR=5. The thresholding rule used was (4).

Another feature of the signal is also important for specifying k - signal regularity. The power parameter k is small if the signal (to be estimated) is irregular. Figure 7 illustrates this relationship. Four standard test signals, Bumps, Blocks, HeaviSine and Doppler of size $n = 1024$ are considered at SNR=5. Bumps is an irregular signal. The optimal k was 1.2. HeaviSine is the most regular signal with optimal value 1.8. Blocks and Doppler exhibit irregularities of different nature (Blocks is a piecewise constant, but discontinuous, while Doppler is smooth but with time varying frequency). For both the optimal value of k was 1.6.

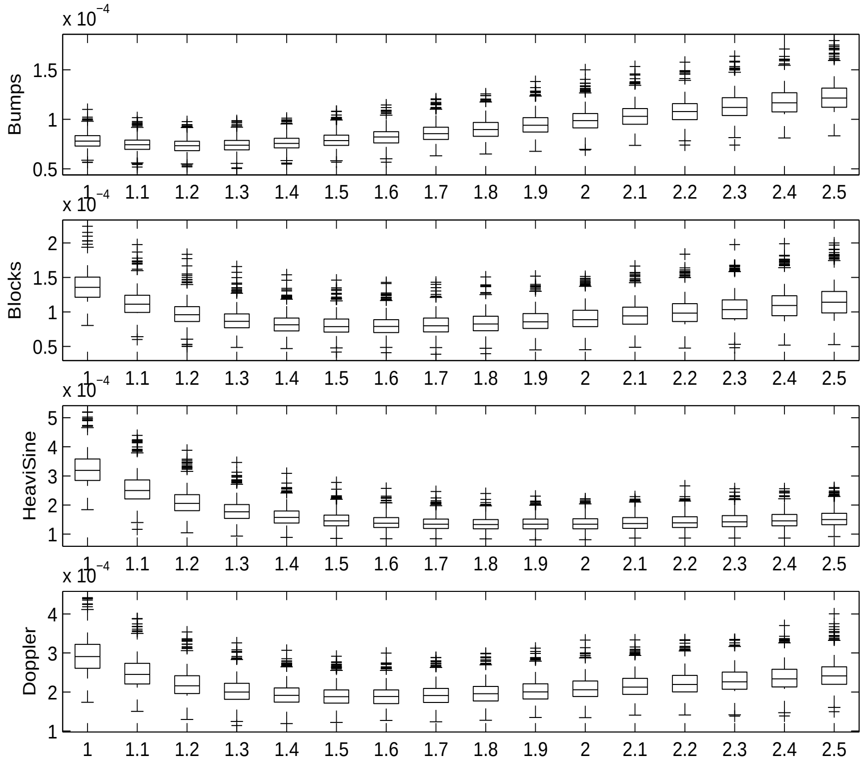


Figure 7: Boxplots of the AMSE for the various values of the power parameter k for four test signals Bumps, Blocks, HeaviSine, and Doppler. Sample size was $n = 1024$ and SNR = 5.

Taking into account the above analysis, a single universal value of k for an automatic use of the rule (4) should be chosen from the interval $(1, 2)$.

Figure 8 shows the true signals and the noisy signals based on $n = 1024$ design points at SNR=5 along with the reconstruction obtained after thresholding the coefficients by LPM method with optimal k. we can see that LPM method does a very good job at removing the noise. From Figure 9 we can see the change in smoothness of recovered signals with the change of k.

Model 1. In the model with an exponential prior on σ^2 in addition to the power parameter k we also have the hyperparameter μ which is the reciprocal of the scale parameter. Given an estimator $\hat{\xi}$ of the noise variance, a moment-matching choice for μ would be $\hat{\mu} = \frac{1}{\hat{\xi}}$. Donoho and Johnstone (1994) suggested to estimate the noise level σ by the median absolute deviation (MAD) of the wavelet coefficients at the finest level adjusted by 1/0.6745, our choice is to consider $\hat{\xi} = \mathrm{MAD}^2$. In this model we will consider $k > 1/2$ with no upper bounds because, even if the posterior distribution is not proper, the choice of $k > 1$ does not affect the existence of the non-zero mode.

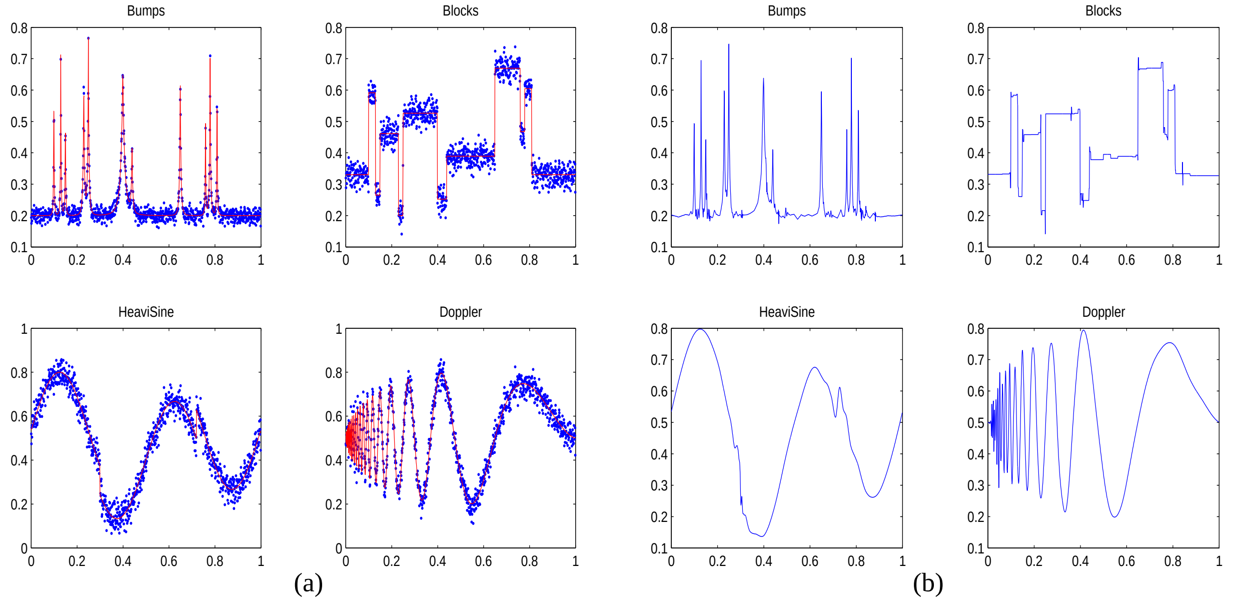


Figure 8: (a) Test signals with superimposed noisy versions at SNR=5. The sample size is $n = 1024$. (b) Estimates obtained by using the LPM method with optimal k.

As could be seen from Figure 4(b) the rule resulting from this model coincides with a hard-thresholding rule and clearly differs from the basic model rule displayed in Figure 2(b). Therefore, we anticipate different behavior of the optimal k as the sample size increases. Our simulations reveal that the optimal power parameter is dependent on the sample size and on the regularity of the signal in this model, as well, but the minimum AMSE is achieved at larger values of k compared to the basic model under the same test conditions. For instance, for the Bumps signal at SNR=5 and $n = 512, 1024, 2048$ and 4096 the optimal values of k are $2.1, 2.4, 2.6$ and 2.8, respectively, while for the four standard test signals Bumps, Blocks, HeaviSine and Doppler of size $n = 1024$ at SNR=5 the optimal values of k are $2.4, 2.7, 3.0$ and 2.7. Therefore, for an automatic use of the thresholding rule in the exponential model, a single universal value of k should be selected from the interval $(2, 3)$.

Model 2. In the model with an inverse gamma prior on σ^2 in addition to the power parameter k we also have two new hyperparameters α and β which specify the prior. As in Model 1 we will match the prior moments with the observed moments in order to specify the hyperparameters. The n-th moment of an inverse gamma random variable $X \sim \mathcal{IG}(\alpha, \beta)$ is

$$EX^n = \frac{\beta^n}{(\alpha-1)\dots(\alpha-n)}.$$

Thus, the first two moments matched with the corresponding empirical moments of wavelet coefficients from the finest level of detail will “estimate” α and β. This consideration and Gaussianity of the noise yields $\alpha = 2.5$ and $\beta = 1.5\ \hat{\xi}$, where $\hat{\xi}$ is some estimator of the variance of the noise. As in the previous models we use the robust $(MAD)^2$ estimator. An argument for the specification of α and β are given in the Appendix.

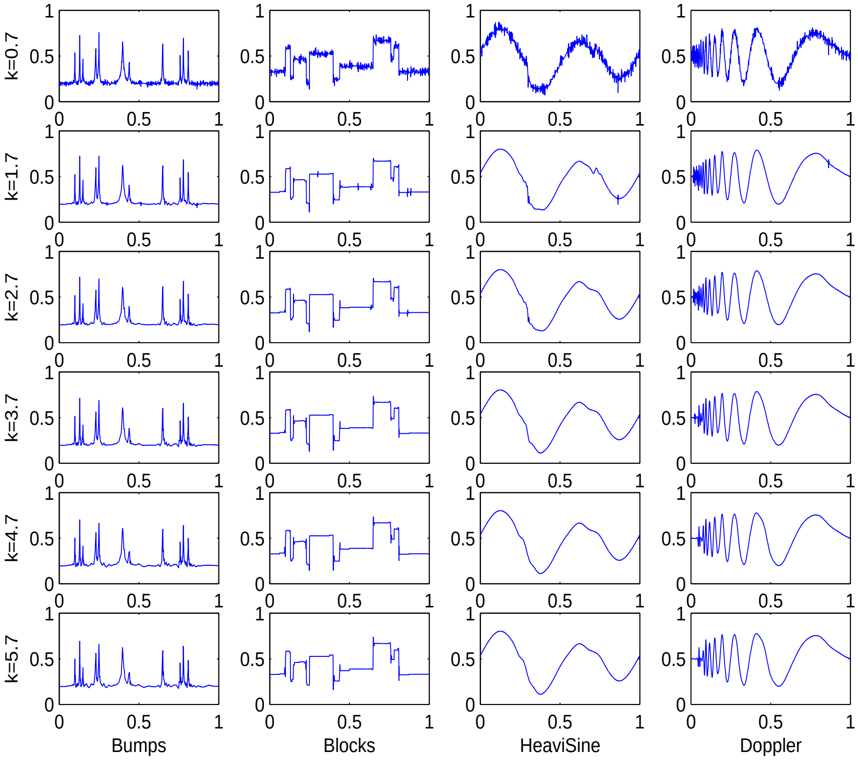


Figure 9: Estimates obtained using LPM method for roughly selected k, based on n=1024 points at SNR=5.

As in the previous two cases, we anticipate different behavior of the optimal k with respect to the sample size and regularity of test functions. For instance, for $\alpha = 2.5$ and β determined using α and an estimator of the variance of the noise for Bumps signal with SNR = 5 the AMSE minimizing values of k are 1.3, 1.6, 1.8, and 2.0 for sample sizes 512, 1024, 2048, and 4096, respectively. For the four standard test signals Bumps, Blocks, HeaviSine and Doppler of size $n = 1024$ at SNR=5 the optimal values of k are $1.6, 2.0, 2.3$ and 2.0. Therefore, for an automatic use of the thresholding rule in the inverse gamma model, a single universal value of k should be selected from the interval $(1, 3)$.

4.2 SIMULATIONS AND COMPARISONS

We present a simulation study of the performance of LPM method for the three models. The simulation is done with the "known truth", that is with test functions specified, and controlled signal-to-noise ratio. We also compare the average mean square error (AMSE) performance with several popular methods.

For our simulation study, four standard test functions (Bumps, Blocks, HeaviSine and Doppler) were added rescaled normal noise to produce a preassigned signal-to-noise ratio (SNR). For each method, test functions were simulated at $n = 512, 1024,$ and 2048 points equally spaced on the unit interval. Three commonly used SNR's were selected: SNR=3 (weak signal), 5 (moderate signal), and 7 (strong signal). The wavelet bases are also standard for the above test functions: Symmlet 8 for `HeaviSine` and `Doppler`, Daubechies 6 for `Bumps` and Haar for `Blocks`.

Closeness of the reconstruction to the theoretical signal of each method was measured by an average mean-square error (AMSE), calculated over 1000 simulation runs. In each case, the optimal power parameter k (minimizing AMSE) was used. All computations are carried out using MATLAB, with the WaveLab toolbox (see Buckheit, Chen, Donoho, Johnstone, and Scargle, 1995) and the GaussWaveDen toolbox (see Antoniadis, Bigot, and Sapatinas, 2001).

The results are summarized in two tables. Table 1 gives minimum AMSE for the three introduced models at three SNR levels and for four standard test functions, while Table 2 presents the corresponding optimal value of the power parameter k.

Final Results × 10^{-3}										
Function	n	SNR=3			SNR=5			SNR=7		
Bumps	512	0.2825	0.3116	0.2875	0.1079	0.1180	0.1095	0.0570	0.0621	0.0577
	1024	0.1953	0.2150	0.1993	0.0733	0.0802	0.0745	0.0373	0.0401	0.0379
	2048	0.1254	0.1371	0.1282	0.0469	0.0509	0.0477	0.0240	0.0257	0.0244
Blocks	512	0.3820	0.4111	0.3876	0.1202	0.1265	0.1213	0.0553	0.0568	0.0554
	1024	0.2752	0.3004	0.2790	0.0802	0.0827	0.0800	0.0359	0.0364	0.0357
	2048	0.1584	0.1692	0.1601	0.0480	0.0502	0.0483	0.0201	0.0204	0.0200
HeaviSine	512	0.4066	0.4305	0.4155	0.2243	0.2441	0.2300	0.1432	0.1575	0.1472
	1024	0.2769	0.2966	0.2835	0.1353	0.1443	0.1379	0.0890	0.0964	0.0914
	2048	0.1711	0.1786	0.1734	0.0950	0.1007	0.0969	0.0604	0.0666	0.0622
Doppler	512	0.7046	0.7706	0.7187	0.2725	0.2959	0.2767	0.1449	0.1557	0.1470
	1024	0.4491	0.4879	0.4590	0.1896	0.2062	0.1931	0.1032	0.1125	0.1054
	2048	0.2514	0.2649	0.2540	0.1064	0.1135	0.1081	0.0596	0.0639	0.0607

Table 1: AMSE for the Basic Model (left), Model 1 (center), and Model 2 (right) at different SNR levels and for the four standard test functions.

Final Results Optimal k										
Function	n	SNR=3			SNR=5			SNR=7		
Bumps	512	1.0	2.1	1.4	1.0	2.1	1.3	1.0	2.0	1.3
	1024	1.2	2.4	1.6	1.2	2.4	1.6	1.2	2.4	1.6
	2048	1.4	2.6	1.8	1.4	2.6	1.8	1.4	2.6	1.8
Blocks	512	1.4	2.5	1.8	1.4	2.6	1.8	1.5	2.7	1.9
	1024	1.5	2.6	1.9	1.6	2.7	2.0	1.6	2.8	2.1
	2048	1.6	2.8	2.1	1.7	2.9	2.2	1.8	2.9	2.2
HeaviSine	512	1.9	3.4	2.4	1.7	2.8	2.1	1.5	2.8	2.0
	1024	2.0	3.2	2.4	1.8	3.0	2.3	1.7	3.0	2.2
	2048	2.1	3.2	2.6	2.0	3.2	2.4	1.8	2.9	2.2
Doppler	512	1.4	2.6	1.8	1.4	2.6	1.8	1.4	2.5	1.8
	1024	1.6	2.8	2.1	1.6	2.7	2.0	1.5	2.7	1.9
	2048	1.8	3.0	2.3	1.8	3.0	2.2	1.7	2.9	2.2

Table 2: Values of optimal k for the Basic model (left), Model 1 (center), and Model 2 (right) at different SNR levels and for the four standard test functions.

We also compare LPM method with several established wavelet-based estimators for reconstructing noisy signals. In particular we consider the term-by-term Bayesian estimator *BAMS* of Vidakovic and Ruggeri (2001), the classical term-by-term estimators *VisuShrink* of Donoho and Johnstone (1994) and *Hybrid-SureShrink* of Donoho and Johnstone (1995), the scale invariant term-by-term Bayesian *ABE* method of Figueiredo and Nowak (2001), the "leave-out-half" version of the *Cross-Validation* method of Nason (1996), the term-by-term False Discovery Rate (*FDR*) method of Abramovich and Benjamini (1995), and finally *NeighCoeff* of Cai and Silverman (2001) and *BlockJS* of Cai (1999) which represent classical estimators that incorporate the blocking procedure to achieve a better performance. Note that, for excellent numerical performance, we consider the *VisuShrink* and the "leave-out-half" version of the *CrossValidation* methods with the hard threshold and the *BlockJS* with the option 'Augment' (see Antoniadis, Bigot, and Sapatinas, 2001).

The LPM is a global method, i.e., the model parameters/hyperparameters are common across the scales in wavelet decompositions. Models for which the parameters/hyperparameters are level-dependent are called adaptive. To avoid confusion, we note that term adaptive is also used in large sample theory for parameters/methods that do not affect the convergence rates. Four of the methods contrasted to LPM are global (*VisuShrink*, *ABE*, *CrossValidation* and *FDR*), while the four remaining methods (*BAMS*, *Hybrid-SureShrink*, *NeighCoeff* and *BlockJS*) are adaptive .

Figure 10 presents the boxplots of the AMSE computed for the above 9 methods based on $n = 1024$ design points at SNR=5. It is clear that LPM method outperforms well-known methods such as VisuShrink, Cross-Validation, FDR and BlockJS methods, and often performs comparably to (sometimes even better than) BAMS, Hybrid-SureShrink, ABE and NeighCoeff methods.

4.3 AN EXAMPLE IN ATOMIC FORCE MICROSCOPY

To illustrate the performance of the LPM thresholding method proposed here, we estimate an underlying smooth function in the noisy measurements from an atomic force microscopy (AFM) experiment.

AFM is a type of scanned proximity probe microscopy (SPM) that can measure the adhesion strength between two materials at the nanonewton scale (Binnig, Quate and Gerber, 1986). In AFM, a cantilever beam is adjusted until it bonds with the surface of a sample, and then the force required to separate the beam and sample is measured from the beam deflection. Beam vibration can be caused by factors such as thermal energy of the surrounding air or the footsteps of someone outside the laboratory. The vibration of a beam acts as noise on the deflection signal; in order for the data to be useful this noise must be removed.

The AFM data from the adhesion measurements between carbohydrate and the cell adhesion molecule (CAM) E-Selectin was collected by Bryan Marshal from the BME Department at Georgia Institute of Technology. The detailed technical description is provided in Marshall, McEver, and Zhu (2001).

In Figure 11 the top panel shows the original noisy data. The middle panel shows the LPM estimate with the default parameter $k = 1$, while the bottom panel shows LPM estimate with the parameter $k = 1.4$. The sample size was $n = 2^{11}$ and Symmlet 8-tap filter was used to obtain the estimate. We observe that the latter estimate exhibits slightly smoother behavior, especially in the long-middle part without oversmoothing the "ramp-like" structure which is the feature of interest here.

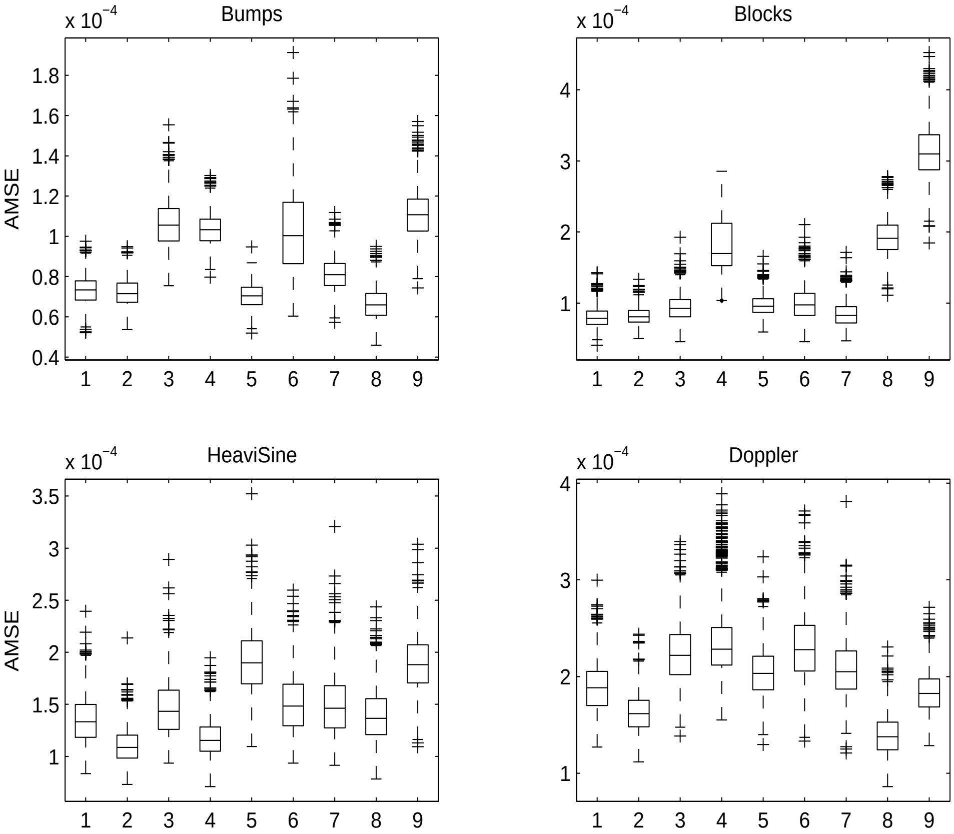


Figure 10: Boxplots of the AMSE for the various methods (1) LPM, (2) BAMS, (3) VisuShrink, (4) Hybrid, (5) ABE, (6) CV, (7) FDR, (8) NC, (9) BJS, based on $n = 1024$ points at SNR=5.

5 CONCLUSIONS

In this paper we developed a method for wavelet-filtering of noisy signals based on larger (in absolute value) posterior mode when the posterior is bimodal. Three variants of the model are considered. The resulting shrinkage rules are thresholding and their performance is comparable to some most popular shrinkage techniques. The method is fast and easily implementable.

We envision several avenues for future research. The LPM thresholding could possibly be improved by level-based specification of model hyperparameters. Such level adaptive formulations are more appropriate for signals and noises that exhibit scale-dependent heterogeneity.

In generalizing the basic model to account for unknown variance we considered only exponential and inverse gamma scale mixtures of normals. Scale mixtures of normals comprise a rich family of models and it would be possible to find an optimal mixing distribution. Specifically, an exponential power distribution (EPD) that contains as special cases normal and double exponential distributions can be obtained as a scale mixture of normals with positive stable distribution as a mixing distribution.

We adhere to the concept of reproducible research (Buckheit and Donoho, 1995). The m-files used for calculations and figures in this paper can be downloaded from Jacket's Wavelets page

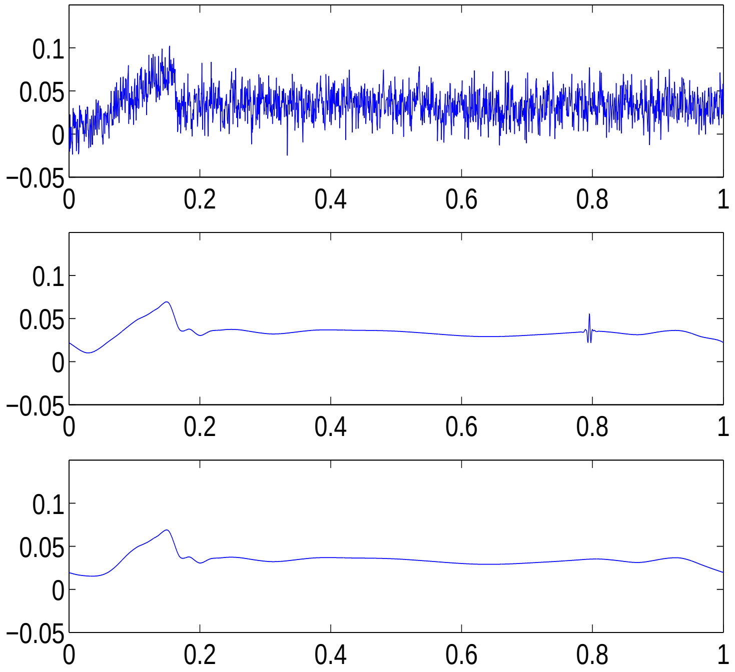


Figure 11: Original AFM measurements (top), LPM estimator with the default parameter $k = 1$ (middle), LMP estimator with the parameter $k = 1.4$ (bottom).

`http://www.isye.gatech.edu/~brani/wavelet.html`.

Acknowledgement. This work was done during visits of Luisa Cutillo and Yoon Young Jung to Georgia Institute of Technology and Brani Vidakovic to CNR-IMATI. Luisa Cutillo acknowledges the support from University "Federico II" under the program MOBILITA DI BREVE DURATA. Work of Yoon Young Jung was supported by Korean Science and Engineering Foundation under the program INTERNATIONAL RESEARCH INTERNSHIP. Partial support for Brani Vidakovic was provided by NSA Grant 240660R and NSF grant 0505490 at Georgia Institute of Technology. The authors thank Claudia Angelini for her help with simulations and modeling at the early stage of this project and Andrew Smith for constructive comments.

REFERENCES

Abramovich, F. and Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In *Wavelets and Statistics* (Antoniadis A. & Oppenheim G. Eds.) Lect. Notes Statist. **103**, 5-14, Springer-Verlag, New York.

Abramovich, F. and Sapatinas, T. (1999). Bayesian approach to wavelet decomposition and shrinkage. In *Bayesian Inference in Wavelet Based Models*, (Müller, P. & Vidakovic, B. Eds.) Lect. Notes Statist. **141**, 33–50, Springer-Verlag, New York.

Angelini, C. and Sapatinas, T. (2004) Empirical Bayes approach to wavelet regression using e-contaminated priors. *Journal of Statistical Computation and Simulation*, **74**, 741 – 764.

Angelini, C. and Vidakovic, B. (2004). Γ-Minimax Wavelet Shrinkage: A Robust Incorporation of Information about Energy of a Signal in Denoising Applications, *Statistica Sinica*, **14**, 103–125.

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study. J. Statist. Soft. **6**, 1–83.

Binnig, G., Quate, C.F., and Gerber, Ch. (1986). Atomic force microscope. *Phys. Rev. Lett.* **56**, 930–933.

Bruce, A.G., and Gao, H-Y., (1996). Understanding WaveShrink: Variance and bias estimation. Biometrika **83**, 727–745.

Buckheit, J.B., Chen, S., Donoho, D.L., Johnstone, I.M. and Scargle, J. (1995). About WaveLab. Technical Report, Department of Statistics, Stanford University, USA.

Buckheit, J. and Donoho, D. L. (1995). Wavelab and reproducible research, in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim eds., LNS 103, Springer-Verlag, New York.

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

Cai, T.T. and Silverman, B.W. (2001). Incorporating information on neighbouring coefficients into wavelet estimation. Sankhy$\bar{a}$ B, **63**, 127–148.

Chipman, H.A., Kolaczyk, E.D. and McCulloch, R.E., (1997) Adaptive Bayesian wavelet shrinkage. J. Amer. Stat. Assoc. **92**, 1413-1421.

Clyde, M. and George, E.I. (1999). Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian Inference in Wavelet Based Models*, (Müller, P. & Vidakovic, B. Eds.), Lect. Notes Statist., **141**, pp. 309–322, Springer-Verlag, New York.

Clyde, M. and George, E.I. (2000). Flexible empirical Bayes estimation for wavelets. J. R. Statist. Soc. B, **62**, 681–698.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. Biometrika **85**, 391–401.

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

Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. J. Am. Statist. Assoc. **90**, 1200–1224.

Donoho, D.L., Johnstone I., Kerkyacharian G. and Picard D. (1995). Wavelet shrinkage: asymptopia J.Roy. Statist. Soc. B **57**, 301-370.

Figueidero, M.A.T. and Nowak, R.D. (2001). Wavelet-based image estimation: an empirical Bayes approach using Jeffrey's noninformative prior. IEEE Trans. Image Process. **10**, 1322–1331.

Gao, H-Y., and Bruce, A.G., (1997) WaveShrink with firm shrinkage. Statistica Sinica **7**, 855-874.

Mallat, S. G. (1989). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. Pattern Analysis Machine Intel.*, **11**, 674–693.

Marshall, B., McEver, R. and Zhu, C. (2001). Kinetic rates and their force dependence on the P-Selectin/PSGL-1 interaction measured by atomic force microscopy. Proceedings of ASME 2001, Bioengineering Conference, BED - Vol. 50.

Müller, P. and Vidakovic, B. (1999). MCMC methods in wavelet shrinkage. In: *Bayesian Inference in Wavelet-Based Models*. Editors Müller, P. and Vidakovic, B., Springer-Verlag, Lecture Notes in Statistics **141**, 187–202.

Nason, G.P. (1996). Wavelet shrinkage using cross-validation. J. R. Statist. Soc. B, **58**, 463–479.

Robert, C. (2001). *Bayesian Choice*, Second Edition, Springer-Verlag, New York.

Ruggeri, F. and Vidakovic, B. (2005), Bayesian modeling in the wavelet domain, to appear in *Handbook of Statistics - vol. 25 - Bayesian Thinking, Modeling and Computation*, D. Dey and C.R. Rao, Eds., North Holland.

Simoncelli, E. and Adelson, E. (1996). Noise removal via Bayesian Coring, *Proceedings of 3rd IEEE International Conference on Image Processing*, Vol. I, pp. 379–382. Lausanne, Switzerland. 16-19 September 1996.

Vidakovic, B., (1998). Non linear wavelet shrinkage with Bayes rules and Bayes factors. J. Amer. Stat. Soc. **45** B, 173-179.

Vidakovic, B. (1999). Statistical Modeling by Wavelets. Wiley, New York.

Vidakovic, B. and Ruggeri, F. (1999). Expansion estimation by Bayes rules, *J. Stat. Plann. Infer.,* **79,** 223–235.

Vidakovic, B. and Ruggeri, F. (2001). BAMS Method: Theory and Simulations. Sankhya B **63**, 234–249.

APPENDIX

DERIVATION OF RULE (5)

Assume that for a typical wavelet coefficient d the following model holds.

$$\begin{aligned} d|\theta,\sigma^2 &\sim \mathcal{N}(\theta,\sigma^2), \\ \sigma^2 &\sim \mathcal{E}\left(\frac{1}{\mu}\right) \text{ with density } p(\sigma^2|\mu)=\mu e^{-\mu\sigma^2},\ \mu>0, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, k>0. \end{aligned}$$

It well known that the marginal likelihood, as a scale mixture of normals, is

$$d|\theta \sim \mathcal{DE}\left(\theta,\frac{1}{\sqrt{2\mu}}\right), \quad \text{with density } f(d|\theta)=\frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|}.$$

Therefore the model can be rewritten as

$$\begin{aligned} d|\theta &\sim \frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|}, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, k>0. \end{aligned}$$

The joint distribution of d and θ is proportional to

$$\begin{aligned} p(d,\theta) &\propto \int_0^\infty p(d|\theta)p(\theta|\tau^2)p(\tau^2)d\tau^2 \\ &= \frac{1}{2}\sqrt{\frac{\mu}{\pi}}e^{-\sqrt{2\mu}|d-\theta|}\int_0^\infty e^{-\theta^2/(2\tau^2)}\frac{1}{(\tau^2)^k}d\tau^2 \\ &= \frac{1}{2}\sqrt{\frac{\mu}{\pi}}e^{-\sqrt{2\mu}|d-\theta|}\int_0^\infty y^{(k-1/2)-1}e^{-\theta^2 y/2}dy \\ &= \frac{1}{2}\sqrt{\frac{\mu}{\pi}}e^{-\sqrt{2\mu}|d-\theta|}\Gamma\left(k-\frac{1}{2}\right)\left(\frac{\theta^2}{2}\right)^{1/2-k}, k>1/2. \end{aligned}$$

Furthermore we have

$$p(\theta|d) \propto p(d,\theta) \propto e^{-\sqrt{2\mu}|d-\theta|}(\theta^2)^{1/2-k}.$$

The likelihood of θ

$$l(\theta) = e^{-\sqrt{2\mu}|d-\theta|}(\theta^2)^{1/2-k} \tag{7}$$

is integrable if and only if $k < 1$.

The eventual modes of the posterior $p(\theta|d)$ exist if and only if they maximize the function (7), that is if and only if they maximize $L(\theta) = \log[l(\theta)]$. More explicitly

$$L(\theta) = \log[l(\theta)] = -\sqrt{2\mu}|d-\theta| + 1 - 2k\log\theta. \tag{8}$$

Consider its derivative

$$L' = \sqrt{2\mu}\,\text{sign}(d-\theta) + \frac{1-2k}{|\theta|}\,\text{sign}(\theta) = \sqrt{2\mu}\,\text{sign}(d-\theta) + \frac{1-2k}{\theta}, \tag{9}$$

and WLOG, suppose $d > 0$. Observe that the critical points of (9) are $\hat{\theta}_1 = 0$ and $\hat{\theta}_2 = \lambda = \frac{2k-1}{\sqrt{2\mu}}$.When $d < \lambda$ there exists only one mode in zero. When $d > \lambda$ there exists two modes, the smaller is zero and the larger is d; in fact the function (8) is decreasing between zero and lambda, increasing between lambda and d and decreasing after d.

DERIVATION OF RULE (6)

The model considered was

$$\begin{aligned} d|\theta,\sigma^2 &\sim \mathcal{N}(\theta,\sigma^2), \\ \sigma^2 &\sim \mathcal{IG}(\alpha,\beta) \text{ with density } p(\sigma^2|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-1-\alpha}e^{\frac{-\beta}{\sigma^2}}, \alpha > 0, \beta > 0, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, k > 0. \end{aligned}$$

It is well known that t distribution is a scale mixture of normals, with mixing distribution being an inverse gamma.

$$d|\theta \sim \frac{1}{\sqrt{2\beta}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}, \quad \text{where } \mathcal{B}\left(\frac{1}{2},\alpha\right) = \frac{\Gamma(\frac{1}{2})\Gamma(\alpha)}{\Gamma(\frac{1}{2}+\alpha)}.$$

Therefore the model can be rewritten as

$$\begin{aligned} d|\theta &\sim \frac{1}{\sqrt{2\beta}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}, \alpha > 0, \beta > 0, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, k > 0. \end{aligned}$$

The joint distribution of d and θ is proportional to

$$\begin{aligned} p(d,\theta) &\propto \int_0^\infty p(d|\theta)p(\theta|\tau^2)p(\tau^2)d\tau^2 \\ &= \int_0^\infty \frac{1}{\sqrt{2\beta}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}} \frac{1}{\sqrt{2\pi\tau^2}}e^{-\theta^2/(2\tau^2)}\frac{1}{(\tau^2)^k}d\tau^2 \\ &= \frac{1}{2\sqrt{\beta\pi}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}\int_0^\infty (\tau^2)^{-(k+1/2)}e^{-\theta^2/(2\tau^2)}d\tau^2 \\ &= \frac{1}{2\sqrt{\beta\pi}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}\int_0^\infty y^{(k-1/2)-1}e^{-\theta^2 y/2}dy \\ &= \frac{1}{2\sqrt{\beta\pi}\mathcal{B}(\frac{1}{2},\alpha)}\Gamma\left(k-\frac{1}{2}\right)\left(\frac{\theta^2}{2}\right)^{1/2-k}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}, k>1/2 \end{aligned}$$

Furthermore, we have

$$p(\theta|d) \propto p(d,\theta) \propto |\theta|^{1-2k}[(d-\theta)^2+2\beta]^{-\alpha-1/2}.$$

The likelihood of θ

$$l(\theta) = |\theta|^{1-2k}[(d-\theta)^2+2\beta]^{-\alpha-1/2}, \quad (10)$$

is integrable for any $k > \frac{1}{2}$.
The eventual modes of the posterior $p(\theta|d)$ exist if and only if they maximize the function (10). Since

$$\begin{aligned} l' &= (1-2k)|\theta|^{-2k}\,\text{sign}(\theta)[(d-\theta)^2+2\beta]^{-\alpha-1/2} + |\theta|^{1-2k}(2\alpha+1)(d-\theta)[(d-\theta)^2+2\beta]^{-\alpha-3/2} \\ &= |\theta|^{-2k}\,\text{sign}(\theta)[(d-\theta)^2+2\beta]^{-\alpha-3/2}\{(1-2k)[(d-\theta)^2+2\beta]+(2\alpha+1)(d-\theta)\theta\}, \end{aligned}$$

it follows that

$$\begin{aligned} |\theta|^{-2k} &> 0, \forall\theta \in \mathcal{R}-\{0\}, \\ \text{sign}(\theta) &> 0, \forall\theta > 0, \\ [(d-\theta)^2+2\beta]^{-\alpha-3/2} &> 0, \forall\theta \in \mathcal{R}, \end{aligned}$$

and

$$l' = 0 \Leftrightarrow (1-2k)[(d-\theta)^2+2\beta]+(2\alpha+1)(d-\theta)\theta = 0,$$

with solutions

$$\theta_{1,2} = \frac{(2\alpha+4k-1)d \pm \sqrt{(2\alpha+1)^2d^2+16(1-2k)(k+\alpha)\beta}}{4(k+\alpha)}.$$

The roots are real if and and only if $(2\alpha+1)^2d^2+16(1-2k)(k+\alpha)\beta > 0$, i.e.,

$$|d| \geq \lambda = \frac{2}{2\alpha-1}\sqrt{(2k-1)(k+\alpha)\beta}\,. \quad (11)$$

If the condition (11) is not satisfied then the MAP is given by $\hat{\theta} = 0$. Now assume that (11) holds and $d > 0$. In this case both solutions $\theta_{1,2}$ are real and positive and the posterior is decreasing from zero to the smaller root, increasing between the two roots and decreasing again after the larger root. We have two posterior modes, the smaller is zero and the larger is

$$\hat{\theta} = \frac{(2\alpha + 4k - 1)d + \sqrt{(2\alpha+1)^2 d^2 + 16(1-2k)(k+\alpha)\beta}}{4(k+\alpha)}.$$

It is easy to see that $\hat{\theta}$ is always smaller then d, resulting in a shrinkage rule.

SELECTION OF HYPERPARAMETERS α AND β IN MODEL 2.

Note that for wavelet coefficients $(d_1, \dots, d_m)$ from the finest level of detail the mean is close to 0, $\bar{d} \approx 0$. That means that $s_d^2 = \frac{1}{m-1}\sum(d_i - \bar{d})^2$ and $\frac{1}{m}\sum d_i^2 = \overline{d^2}$ are both comparable estimators of the variance. Also, even central empirical moments are approximately equal to the noncentral moments. The following two equations are approximately moment matching:

$$\overline{d^2} = \frac{\beta}{\alpha - 1}, \qquad \overline{d^4} = \frac{\beta^2}{(\alpha-1)(\alpha-2)},$$

where $\overline{d^4} = \frac{1}{m}\sum d_i^4$. From these equations we derive

$$\alpha = \frac{2\overline{d^4} - (\overline{d^2})^2}{\overline{d^4} - (\overline{d^2})^2},$$

which is free of the scale of wavelet coefficients. Since in the finest level of detail the contribution of signal is minimal and the wavelet coefficients are close to zero-mean normal random variables the Law of Large Numbers argument gives $\overline{d^2} \approx \sigma^2$ and $\overline{d^4} \approx 3\sigma^4$, which specifies the “shape” hyperparameter

$$\alpha = 2.5.$$

Hyperparameter β is determined from $\overline{d^2} = \frac{\beta}{\alpha-1}$, but instead of $\overline{d^2}$ we can use any estimator of variance of d. In simulations, we used the robust $(MAD)^2$.