A New Generation of Wavelet Shrinkage: Adaptive Strategies Based on Composition of Lorentz-type Thresholding and Besov-type Non-threshold Shrinkage

Lubomir Dechevsky^a and Niklas Grip^b and Joakim Gundersen^c

^aNarvik University College, P.O.B. 385, N-8505 Narvik, Norway;

^bDepartment of Mathematics, Luleå University of Technology, SE-971 87 Luleå, Sweden;

^cNarvik University College, P.O.B. 385, N-8505 Narvik, Norway

ABSTRACT

This article is a systematic overview of compression, smoothing and denoising techniques based on shrinkage of wavelet coefficients, and proposes (in Sections 5 and 6) an advanced technique for generating enhanced composite wavelet shrinkage strategies.

Keywords: stochastic signal analysis, image compression

1. INTRODUCTION

By now the theory of wavelet shrinkage via thresholding of the wavelet coefficients has reached a fairly advanced level of development. This theory yields good results both in parametric and nonparametric stochastic signal analysis and in deterministic data compression, when the target signal or image gathers most of its value from relatively few, relatively large, wavelet coefficients or, in other words, when the vector of its wavelet coefficients looks sparse. By the wavelet analogue of the well known Riemann-Lebesgue lemma in the study of trigonometric series, the typical case of quasi-sparseness of the wavelet-coefficient vector is when the signal is sufficiently smooth. In this case, it is usually sufficient to apply nonadaptive threshold shrinkage strategies such as, for example, hard and soft thresholding (in its various global, levelwise or block versions).

Another extreme case is provided by fractal signals and images which are continuous but nonsmooth everywhere (a simple classical example being the Weierstrass function). In this case, again according to the Riemann-Lebesgue lemma, the vector of wavelet coefficients looks locally full everywhere. This general case was specifically addressed in ¹, where a family of wavelet-shrinkage procedures of nonthreshold type were considered for very general classes of signals that belong to the general scale of Besov spaces and may have a full, non-sparse, vector of wavelet coefficients (see, in particular ¹, Appendix B, item B9). In stochastic signal analysis of selfsimilar or nonselfsimilar fractals which are everywhere nonsmooth it suffices to apply nonadaptive nonthreshold wavelet shrinkage of the above type (which, however, does not achieve any compression).

Of the most universal interest in signal and image processing applications is, however, the intermediate case of spatially inhomogeneous signals which exhibit both smooth regions and regions with (isolated, or continual fractal) singularities. For this most interesting and important for the applications case, nonadaptive thresholding shrinkage tends to oversmooth the signal in a neighbourhood of every singularity, while nonadaptive nonthresholding shrinkage tends to undersmooth the signal in the regions where the signal has regular behaviour and locally quasi-sparse wavelet-coefficient vector. It is, therefore, of interest, as noted in ², to develop a next, second, generation of composite wavelet shrinkage strategies having the new and very remarkable property to adapt to the local sparseness or nonsparseness of the vector of wavelet coefficients, with simultaneously improved performance near singularities as well as in smooth regions.

Further author information: (Send correspondence to L.D.)

L.D.: E-mail: ltd@hin.no, Telephone: (+47)76966369

N.G.: E-mail: grip@sm.luth.se, Telephone: (+46)702917793

J.G.: E-mail: joag@hin.no, Telephone: (+47)76966190

In the present study we incorporate all of the afore-mentioned wavelet shrinkage strategies within the very general setting of the so-called K-functional Tikhonov regularization of incorrectly posed inverse deterministic and stochastic problems, thereby obtaining a lucid uniform comparative characterization of the above-said approaches and their interrelations, and for the first time apply a new type of thresholding – the so-called Lorentz-type thresholding (based on decreasing rearrangement of the wavelet coefficient vector, as briefly outlined for the first time in ¹, Appendix B, item B10(b)). We develop new composite adaptive shrinkage procedures based on data-dependent composition of the new Lorentz-type thresholding with the nonthreshold shrinkage procedures of ¹. It can be shown that these new highly adaptive strategies achieve the best possible rate of compression over all signals with prescribed Besov regularity (smooth signals as well as fractals). Both univariate and multivariate signals are considered.

We illustrate the results of our study by a model example of data processing using the in-house wavelet library GM_Waves developed at Narvik University College in template-based C++, OpenGL and Qt. This example results from the simulator BedSim of the Swedish mining and processing consortium LKAB and the combustion simulator Kameleon FireEx of the Norwegian firm ComputIT AS – two of the industrial partners with whom our R&D group at Narvik University College has long-term research cooperation.

2. TYPES OF WAVELET SHRINKAGE

We discuss two types of wavelet shrinkage: a thresholding and non-thresholding type of shrinkage rules.

To fix terminology, a thresholding shrinkage rule sets to zero all coefficients with absolute values below a certain threshold level, $\lambda \geq 0$, whilst a non-thresholding rule shrinks the non-zero wavelet coefficients towards zero, without actually setting to zero any of them.

In general, shrinkage methods of threshold type are appropriate for estimating of relatively regular functions, while those of non-threshold type fit best for estimating of spatially inhomogeneous functions.

Recently, several thresholding approaches have been proposed to generate wavelet estimators with improved performance on moderate samples. But these estimators still have their problems, especially when the sample sizes are small to moderate, the noise-to-signal ratio is relatively large, the white noise is not Gaussian, and the estimated function is not sufficiently smooth. The only opportunity to get good fits for small samples seems to be to introduce in the estimation model all available information about essential features of the signal. However, optimization with respect to threshold values leads to non-linear, non-smooth, non-convex problems with high computational complexity. An additional problem when estimating functions with low regularity and fractals is that thresholding methods tend to oversmooth the signal since they are well adapted for functions which gather their value on relatively few, relatively large wavelet coefficients (sparse vector of wavelet coefficients). Continuous fractals and functions with jumps do not usually fall into this case, but rather gather their value uniformly from many wavelet coefficients on infinitely many levels and the vector of their wavelet coefficients does not necessarily look sparse. Due to these facts, for such types of functions non-threshold shrinkage is appropriate.

The shrinkage estimator is obtained by replacing the empirical $\hat{\beta}_{jk}$ in with their shrunken analogues $\tilde{\beta}_{jk}$ (coefficients of the wavelet expansion representing the empirical estimator (see (13, 14) below for the details).

3. GENERAL THRESHOLDING RULES

The hard and soft thresholding rules proposed by Donoho and Johnstone ^{3, 4, 5} for smooth functions, are given respectively by:

$$\delta(x; \lambda) = \begin{cases} |x|, & \text{if } |x| > \lambda \\ 0, & \text{if } |x| \leq \lambda \end{cases}$$

(1)

and

$$\delta(x; \lambda) = \begin{cases} \text{sgn}(x)(|x| - \lambda), & \text{if } |x| > \lambda \\ 0, & \text{if } |x| \leq \lambda \end{cases}$$

where $\lambda \in [0, \infty)$ is the threshold.

Asymptotically both hard and soft shrinkage estimates achieve within a $\log n$ factor of the ideal performance (with the aid of an oracle). Soft thresholding (a continuous function) is a 'shrink' or 'kill' rule, while hard

Figure 1. Thresholding rules: (1) hard-, (2) firm-, (3) hyperbole-, (4) garrote- and (5) soft-types.

thresholding (a discontinuous function) is a 'keep' or 'kill' rule. Due to the discontinuity of the hard threshold rule, the hard shrinkage estimates tend to have bigger variance and can be sensitive to small changes in the data. The soft thresholding estimates tend to have bigger bias, due to the shrinkage of large coefficients.

Non-negative garrote shrinkage remedies the drawbacks of the hard and the soft shrinkage. The function was first introduced by Breiman,⁶ and is defined as follows:

$$\delta(x; \lambda) = \begin{cases} x - \frac{\lambda^2}{x}, & \text{if } |x| > \lambda \\ 0, & \text{if } |x| \leq \lambda \end{cases} \quad (2)$$

Since the non-negative garrote shrinkage function is continuous and approaches the identity line as $|x|$ gets large, it is less sensitive than the hard shrinkage to small fluctuations in the data, and less biased than the soft shrinkage. In simulation studies, the non-negative garrote shows to have consistently lower prediction error than subset selection when the true model has many small nonzero coefficients.

Gao and Bruce⁷ introduced the firm shrinkage rule:

$$\delta(x, \lambda_1, \lambda_2) = \begin{cases} \text{sgn} \frac{\lambda_2 \cdot (|x| - \lambda_1)}{\lambda_2 - \lambda_1}, & \text{if } |x| \in (\lambda_1, \lambda_2] \\ x, & \text{if } |x| > \lambda_2 \\ 0, & \text{if } |x| \leq \lambda_1 \end{cases} \quad (3)$$

By choosing appropriate thresholds (λ_1, λ_2) , firm thresholding outperforms both hard and soft thresholding; it has all the benefits of the best of the hard and soft without the drawbacks of either. The only disadvantage of the firm shrinkage is that it requires two thresholds. This makes threshold selection problems much harder and computationally more expensive for adaptive threshold selection procedures such as cross-validation.

Note, that

$$\lim_{\lambda_2 \rightarrow \infty} \delta^{firm}(x, \lambda_1, \lambda_2) = \delta^{soft}(x, \lambda_1) \quad (4)$$

and

$$\lim_{\lambda_2 \rightarrow \lambda_1} \delta^{firm}(x, \lambda_1, \lambda_2) = \delta^{hard}(x, \lambda_1). \quad (5)$$

A so-called hyperbole thresholding⁸ function is given by:

$$\delta(x; \lambda) = \begin{cases} \text{sgn}(x) \sqrt{x^2 - \lambda^2}, & \text{if } |x| \geq \lambda \\ 0, & \text{if } |x| < \lambda \end{cases} \quad (6)$$

It also, like the non-negative garrote function, combines properties of both hard- and soft-thresholding rules.

4. CURVE AND SURFACE FITTING BY WAVELET SHRINKAGE

We consider the following two statistical models: 1. Non-parametric regression; 2. Density estimation. For the non-parametric regression, the model is:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (7)$$

where X_i are independent, uniformly distributed on $[0, 1]^n$ random variables. For the independent identically distributed (i.i.d.) errors ε_i we assume $E\varepsilon_1 = 0$, $E\varepsilon_1^2 = \delta^2$. In the case of identically distributed density estimation, the sample consists of n i.i.d. observations Z_i , $i = 1, 2, \dots, n$, with unknown density $f(x)$, $x \in \mathbb{R}^n$. We shall address the problem of estimating f in the above two problems by using orthonormal wavelets.

Let $\dot{B}_{pq}^s(\mathbb{R}^n)$ and $B_{pq}^s(\mathbb{R}^n)$ be respectively the homogeneous and inhomogeneous Besov spaces with metric indices p, q and smoothness index s . For $f \in B_{pq}^s$, $0 < p \leq \infty$, $0 < q \leq \infty$, $n(\frac{1}{p} - 1)_+ < s < r$,

$$f(x) = \sum_{k \in \mathbb{Z}^d} \alpha_{0k} \phi_{0k}^{[0]}(x) + \sum_{j=0}^{\infty} \sum_{k \in \mathbb{Z}^d} \sum_{l=1}^{2^d-1} \beta_{jk}^{[l]} \psi_{jk}^{[l]}(x), \quad a.e. \quad x \in \mathbb{R}^d \quad (8)$$

holds, where $\alpha_{0k} = \langle \phi_{0k}^{[0]}, f \rangle = \int_{\mathbb{R}^d} \phi_{0k}^{[0]}(x) \overline{f(x)} dx$, $\beta_{jk}^{[l]} = \langle \psi_{jk}^{[l]}, f \rangle$ and $\bar{z}$ is the conjugate of $z \in \mathbb{C}$. Here the tensor-product basis $\{\varphi_{jk}^{[0]}, \psi_{jk}^{[l]}\}$ is defined in ¹. Convergence in (8) is in the quasi-norm topology of the inhomogeneous Besov space $B_{pq}^s(\mathbb{R}^d)$.

For the homogenous space $\dot{B}_{pq}^s(\mathbb{R}^d)$ it can be shown that

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k \in \mathbb{Z}^n} \sum_{i=1}^{2^n-1} \beta_{jk}^{[i]} \psi_{jk}^{[i]}(x), \quad a.e. \quad x \in \mathbb{R}^d \quad (9)$$

holds modulo polynomials of total degree less than r . An equivalent (quasi)-norm of f in $B_{pq}^s(\mathbb{R}^d)$ has the notation:

$$\|f\|_{B_{pq}^s(\mathbb{R}^d)} = \left\{ \left(\sum_{k \in \mathbb{Z}^d} |\alpha_{0k}|^p \right)^{q/p} + \sum_{j=0}^{\infty} \left[2^{j[s+n(\frac{1}{2}-\frac{1}{p})]} \left(\sum_{k \in \mathbb{Z}^d} \sum_{l=1}^{2^d-1} |\beta_{jk}^{[l]}|^p \right)^{\frac{1}{p}} \right]^q \right\}^{\frac{1}{q}}. \quad (10)$$

For the concept of quasi-norm, see the references in ⁹. Analogously, for the homogenous space $\dot{B}_{pq}^s(\mathbb{R}^d)$ we have:

$$\|f\|_{\dot{B}_{pq}^s(\mathbb{R}^d)} = \left\{ \sum_{j=-\infty}^{\infty} \left[2^{j[s+d(\frac{1}{2}-\frac{1}{p})]} \left(\sum_{k \in \mathbb{Z}^d} \sum_{l=1}^{2^d-1} |\beta_{jk}^{[l]}|^p \right)^{\frac{1}{p}} \right]^q \right\}^{\frac{1}{q}}. \quad (11)$$

The empirical wavelet estimator $\hat{f}(x)$ is defined via:

$$\hat{f}(x) = \sum_{k \in \mathbb{Z}^d} \hat{\alpha}_{0k} \phi_{0k}^{[0]}(x) + \sum_{j=0}^{\infty} \sum_{k \in \mathbb{Z}^d} \sum_{l=1}^{2^d-1} \hat{\beta}_{jk}^{[l]} \phi_{jk}^{[l]}(x), \quad a.e. \quad x \in \mathbb{R}^d, \quad (12)$$

where in the case of non-parametric regression (7):

$$\hat{\alpha}_{j_0 k} = \frac{1}{n} \sum_{i=1}^n \varphi_{j_0 k}^{[0]}(X_i) Y_i, \quad \hat{\beta}_{jk}^{[l]} = \frac{1}{n} \sum_{i=1}^n \psi_{jk}^{[l]}(X_i) Y_i, \quad (13)$$

and, in the case of density estimation:

$$\hat{\alpha}_{j_0 k} = \frac{1}{n} \sum_{i=1}^n \varphi_{j_0 k}^{[0]}(X_i), \quad \hat{\beta}_{jk}^{[l]} = \frac{1}{n} \sum_{i=1}^n \psi_{jk}^{[l]}(X_i), \quad (14)$$

are the empirical (in statistical sense) wavelet coefficients. In both the case of non-parametric regression and density estimation the estimator $\hat{f}$ can be obtained simply by replacing the coefficients in (8) and (9) by their empirical versions, but this procedure is not as fast as the discrete wavelet transform.

In the case of non-parametric regression this shrinkage can be obtained for sample sizes n which are an exact power of 2, by the fast discrete wavelet transform in the following way:

- (a) Transform data y into the wavelet domain.
- (b) Shrink the empirical wavelet coefficients towards zero.
- (c) Transform the shrunk coefficients back to the data domain.

The methodology to estimate f is based on the principle of shrinking wavelet coefficients towards zero to remove noise, which means reducing the absolute value of the empirical coefficients.

Wavelet coefficients having small absolute value contain mostly noise. The important information at every resolution level is encoded in the coefficients on that level which have large absolute value.

One of the most important applications of wavelets - denoising - began after observing that shrinking wavelet coefficients towards zero and then reconstructing the signal has the effect of denoising and smoothing.

All the main results for the basic regression-function and density estimation models are applicable for the multivariate case. (We consider here the tensor-product bases to define multivariate orthonormal wavelets).

4.1 Composite estimators

By the well known Riemann-Lebesgue lemma in the study of trigonometric series, applied in the case of wavelets, a very smooth function without singularity points, gathers its value in (8, 9) and its (semi)-norm in (10, 11) from relatively large coefficients, or in other words, when the vector of wavelet coefficients of f in (8-11) looks sparse. In this case, non-adaptive threshold shrinkage suffices.

Another extreme case, again according to the Riemann-Lebesgue lemma, is provided by fractal functions which are continuous but non-smooth everywhere (e.g., the Weierstrass function in Fig. 3). In this case, the vector of wavelet coefficients in (8-11) looks locally full everywhere, and non-adaptive non-threshold shrinkage suffices.

Fig. 2 provides three typical examples of function curves containing both regular and singular points, and it is clearly seen that non-adaptive threshold and non-threshold estimators perform well only in a neighborhood of a regular, respectively singular, point. Creating data-driven adaptive composite estimators which tend to have thresholding effect near regular points, and non-threshold shrinkage effect near singular points is, in our opinion, one main strategic task in the design of the next generation of wavelet approximants for deterministic and indeterminate curve and surface fitting.

One first example of such composite shrinking strategies is firm thresholding. It is, indeed, still a rather primitive example, but even so, it consistently outperforms soft and hard thresholding.

4.2 Threshold assessment

A central question in many threshold procedures is how to choose the threshold. A threshold is a trade-off between closeness of fit and smoothness. A small threshold yields a result close to the input, but this result may still be noisy. A large threshold on the other hand, produces a signal with a lot of zero wavelet coefficients. This sparsity is a sort of smoothness: the output has a simple, smooth representation in the chosen basis. Paying too much attention to smoothness however destroys some of the signal singularities; in image processing for example, it may cause blur and artifacts.

Literature contains many papers devoted to this problem of threshold selection. A well known universal threshold⁵ pays attention to smoothness rather than to minimizing the mean square error. This threshold comes close to the minimax threshold, i.e. the threshold that minimizes the worst case mean square error in a typical function space⁴.

5. OTHER COEFFICIENT SELECTION PRINCIPLES

Coefficient shrinkage aims at a noise reduction, without losing too much signal information. In most algorithms, this manipulation depends on a classification of a coefficient, which is often binary: a coefficient is either noisy or relatively noise-free and important. To distinguish between these classes, we need a criterion, a sort of threshold on a measure of regularity. There are essentially two types of models on which this measure of regularity is based: Bayesian and non-Bayesian models. The former type thinks of the uncorrupted coefficients V as an instance from a density function $f_V(\nu)$, so we get a fully random model:

$$W = V + N, \quad (15)$$

where N is the noise. In principle, these methods compute the posterior density for the noise-free coefficients from Bayes' rule:

$$f_{V|W}(\nu|\omega) = \frac{f_V(\nu)f_{W|V}(\omega|\nu)}{f_W(\omega)}, \quad (16)$$

which allows, for instance, to estimate the underlying 'true' coefficients by the posterior mean $\tilde{\nu} = E(V|W)$. Depending on the chosen density functions $f_V(\nu)$ and $f_{W|V}(\omega|\nu)$, this mostly leads to shrinking rules discussed above.

The other type of classification considers the noise-free signal as a deterministic member of a smooth function space. These methods try to understand how signal regularity can be observed from wavelet coefficients.

5.1 Basis selection methods

Wavelet coefficient manipulation methods proceed within one wavelet basis, or one pair of dual bases. More adaptive methods build up the basis in which the data signal is processed. The objective is to make the signal fit as well as possible into this self-constructed basis. The method uses an overcomplete set of functions, in the sense that a given signal can be expressed as a linear combination of more than one subset of this 'library' of functions. Well structured libraries lead to fast algorithms for best basis selection, i.e. finding the basis in which the coefficients α in the decomposition

$$f = \sum_i \alpha_i \phi_i \quad (17)$$

has minimal entropy. The concept of entropy has several possible definitions. The ℓ_1 -norm is one of these, and another well known example is:

$$E(\alpha) = \sum_i \frac{|\alpha_i|}{\|\alpha\|_2} \log \frac{|\alpha_i|}{\|\alpha\|_2}, \quad (18)$$

(compare to $\ell^2(\{\alpha_i\})$ in section 8). One could also consider the sum of squares of coefficients above a given threshold as an entropy measure.³ The wavelet packet transform belongs to these methods of best basis selection¹⁰. Other methods use different types of functions, like local trigonometric functions¹¹. If the input data are noisy, the noise can be eliminated by decomposition

$$y = \sum_{i=1}^m \alpha_i \phi_i + R^{(m)}, \quad (19)$$

so that:

$$\frac{1}{2} \|R^{(m)}\|_2^2 + \lambda E(\alpha) \quad (20)$$

is as small as possible. The result is a trade-off between a close approximation of the input and a sparse representation. In this cost function, λ plays the role of a smoothing parameter. The idea is that noise cannot be sparsely represented in any basis from the library. Some variants include:

1. Basis pursuit¹² uses the ℓ_1 -norm as entropy. The objective function (1.23) then reduces to the form of a linear programming problem. Moreover, this expression has the same form as the one leading to soft-thresholding in the fixed basis setting.
2. Matching pursuit¹³ is a greedy algorithm. In each step, it looks for the function from the library with the highest correlation with the residual after the previous step.
3. Best Orthogonal Basis¹⁴ is limited to orthogonal bases. Extensions are given in.¹¹

After this short overview on basis dependent techniques, in the rest of this exposition we concentrate on the shrinkage techniques for a fixed wavelet basis. These techniques are still not sufficiently well explored because their study requires rather advanced knowledge of functional analysis and operator theory.

5.2 Non-thresholding Shrinkage

Increasing interest to the study of fractals and singularity points of functions (discontinuities of the function or its derivatives, cusps, chirps, etc.) raises the necessity of non-threshold wavelet shrinkage.

At this point, while threshold rules can be considered as well studied, non-threshold rules, on the contrary, are fairly new and the corresponding theory is so far in an initial stage. This can be explained by the fact that traditionally only very smooth functions were being estimated.

We consider here a family of wavelet-shrinkage estimators of non-threshold type which are particularly well adapted for functions belonging to Besov spaces and have a full, non-sparse, vector of wavelet coefficients. The approach, proposed by Dechevsky, Ramsay and Penev in¹, parallels Wahba's spline smoothing technique for the case of fitting less regular curves.

The regularity of the curve is discussed in terms of the size of its semi-form in the homogenous Besov spaces, with a relatively small value of the smoothness index $s > 0$. The optimal solution of the penalization problem is in the form of a wavelet expansion whose coefficients are obtained by appropriate level- and space- dependent shrinking of the empirical wavelet coefficients. Thanks to the use of wavelets, both density and regression estimation can be treated in a somehow unified way.

The penalization model is defined via:

$$L(v, \hat{f}; B_{p_1 p_1}^{s_1}, \dot{B}_{pp}^s) := K_1(v, \hat{f}; (B_{p_1 p_1}^{s_1})^{p_1}, (\dot{B}_{pp}^s)^p), \quad (21)$$

where p, p_1, s, s_1 are such that $B_{p_1 p_1}^{s_1} \supset B_{pp}^s$. The risk of the estimator is the expected value of the quasi-norm of the difference between the function and the estimator in a certain Besov space, $\|f - \tilde{f}\|_{p, q, s}$.

The general function for shrinking in Besov spaces is given by $\tilde{\beta}_{jk} = \mu_{jk}\hat{\beta}_{jk}$, $\mu_{jk} = \mu_{jk}(v) \in [0, 1]$, and can be numerically computed from the following equation:

$$p_1^{\frac{1}{\eta}} \cdot (1 - \mu)^{p_1 - \frac{1}{\eta}} \cdot |\hat{\beta}_{jk}|^{(p_1 - p)} = v_j \cdot 2^{\frac{j \cdot \varepsilon}{2}} \cdot p^{\frac{1}{\eta}} \cdot \mu^{p - \frac{1}{\eta}} \quad (22)$$

Here v_j is a smoothing factor¹ and for the arguments in (22) we have:

$$0 < p \leq p_1 < \infty, \max \left\{ \frac{1}{p}, \frac{1}{p_1} \right\} < \eta < \infty, \varepsilon = 2ps - 2p_1s_1 + n(p - p_1), \quad (23)$$

here ε is the so-called critical-regularity index.

In general, the equation (22) can not be solved explicitly. However, for any (j, k) the solution can be found by very quickly convergent iterations of the dyadic or Fibonacci bisection method. Moreover, in many important partial cases (22) can be solved explicitly. This is due to the fact that for the constants

$$A = \left(p_1^{\frac{1}{\eta}} |\hat{\beta}_{jk}|^{p_1 - p} \right)^{\frac{1}{p_1 - 1/\eta}} \quad B = \left(v_j \cdot 2^{\frac{j \cdot \varepsilon}{2}} \cdot p^{\frac{1}{\eta}} \right)^{\frac{1}{p_1 - 1/\eta}}$$

(22) gets the form:

$$A(1 - \mu) = B\mu^{\frac{p - 1/\eta}{p_1 - 1/\eta}}, \quad (24)$$

and for $p_1 = p < \infty$, $s > s_1$, $v = t^p$, $p, t \in [0, \infty)$, becomes linear with the unique solution:

$$\mu_{jk} = \mu_j = \frac{1}{1 + (t \cdot 2^{j(s - s_1)})^{\frac{p}{p - 1/\eta}}} \quad (25)$$

It is seen from (22-25) that, the closer p is to $1/\eta$, the more sensitive the model is to variations of s , s_1 and choice of v_j .

Analogously, for $\frac{p_1 - 1/\eta}{p - 1/\eta} = 2$, $\frac{p_1 - 1/\eta}{p - 1/\eta} = 3$, and $\frac{p_1 - 1/\eta}{p - 1/\eta} = 4$ the equation (22) can be solved explicitly.

For particular cases of the metric and smoothness parameters, the Besov rules include both the non-threshold shrinkage rules and soft and hard thresholding rules. The garrote and firm thresholding are included in certain classes of Besov composite estimators. The particular combination of Besov parameters, for which the classical shrinkage estimators occur, provide additional insight into some fine properties of functions of one and several variables. For example, soft thresholding occurs in cases where p_1 is much larger than p , while ε has small to moderate values; if ε is large, then the perspective Besov rule (with p_1 much larger than p) approximates the hard threshold rule.

On Fig. 2 is compared the distribution of approximation/ estimation error for a typical threshold shrinkage method (soft thresholding - left column) and non-threshold (Besov for $p = p_1$ - right column) for the same set of noisy observations (the non-parametric regression case) for the four functional curves considered as test examples in .¹ It is seen from these two figures that the non-adaptive threshold method oversmooths in singular points, while the non-adaptive non-threshold method overfits in regular points; an adaptive composite estimator is needed (see next section). In terms of absolute error (plotted centered around 0), the non-threshold estimator fares better. In the case of the Weierstass function which is non-smooth everywhere, the non-threshold estimator performs much better everywhere (see Fig. 3).

5.3 Lorentz Curve Thresholding

Usually, time series contains an energy disbalance in the wavelet domain. Vidakovic⁸, proposed a thresholding method based on the Lorentz curve for the energy in the wavelet decomposition.

The Lorentz curve corresponding to a nonnegative random variable X , $EX = \mu < \infty$, is defined as

Figure 2. Thresholding (left column) and non-thresholding (right column) rules.

$$L(p) = \frac{1}{\mu} \int_0^{\xi_p} x dF(x), \quad (26)$$

where ξ_p is the unique population p th quantile. Distributions with finite mean uniquely define their Lorentz curve, and vice versa. The scale-free character of $L(p)$ allows development of goodness-of-fit tests for families of

Figure 3. Thresholding and non-thresholding rules.

distributions depending on an unknown scale parameter.

Replace the $p_0 \times 100\%$ of the coefficients with the smallest energy with zero.

$$p_0 = \frac{1}{n} \sum_i \mathbf{1}(d_i \leq \bar{d}_2) \quad (27)$$

where $\bar{d}_2$ is the mean of the energies $(d_1^2, d_2^2, \dots, d_n^2)$. The value p_0 represents the proportion at which the gain by thresholding an additional element will be smaller than the loss in the energy. Both losses are measured on a scale 0-1, and are equally weighted.

5.4 General Lorentz thresholding

In Item B9(b) of Appendix B of ¹ it was shown that in the case of the embedding $B_{pp}^\sigma \hookrightarrow B_{pp}^s$, $s > \sigma$, there is an explicit shrinking formula for the shrinking factor μ_{jk} (formula (B6)). Another major embedding within the Besov-space scale is the *Sobolev-type* embedding: $B_{\pi\pi}^\sigma \hookrightarrow B_{pp}^s$ if $\sigma - \frac{1}{\pi} = s - \frac{1}{p}$ and $0 < p \leq \pi \leq \infty$. From (22) it can be seen that there is no explicit formula of the type of (25) for this setting of σ , s , π and p . This makes even more interesting the fact that the K -functional approach indicates that there is a remarkable new thresholding method narrowly specialized for: σ , s , π , p : $\sigma - \frac{1}{\pi} = s - \frac{1}{p} =: \tau \in \mathbb{R}$, $0 < p \leq \pi \leq \infty$. The essential additional fact here is that τ is the same for both spaces in the K -functional. Because of this invariance of τ we can invoke the following formula

$$K_1(v, f; L_\pi(U, d\mu), L_p(U, d\mu)) \asymp \left(\int_{v^{-\lambda}}^\infty f^*(\xi)^\pi d\xi \right)^{1/\pi} + v \left(\int_0^{v^{-\lambda}} f^*(\xi)^p d\xi \right)^{1/p},$$

where $(U, d\mu)$ is an arbitrary measure space ($d\mu$ can be also an atomic measure), $f^*(\xi)$ is the *decreasing rearrangement* of f with respect to the measure $d\mu$ (see ¹⁵, Section 1.3), and

$\lambda: \frac{1}{\lambda} = \frac{1}{p} - \frac{1}{\pi}$. Formula (5.4) cannot be found explicitly in our reference sources but it can be obtained from the Peetre-Kre formula (¹⁵, Theorem 3.6.1, together with 3.14.5,6 and Theorem 5.2.1 (2) for $q = p$ in their notation).

Suppose, as usual, that f and ψ are compactly supported. For simplicity of presentation, assume that $j_0 = 0$ for any sample size n . Formula (5.4) implies the following strategy for thresholding the β -wavelet coefficients:

Step 1. Consider all $(j, k) : \text{supp} \psi_{jk} \cap \text{supp} f \neq \emptyset$. Denote the set of all such (j, k) by $I(f, \psi)$. From the proof of Theorem 1 in Appendix A of ¹ that the number M of elements of $I(f, \psi)$ does not exceed $c \cdot 2^{j_1}$, where $c = c(f, \psi)$.

Step 2. Consider the decreasing rearrangement $\{b_\nu, \nu = 1, \dots, M\}$ of the finite set $\{2^{j(\tau+1/2)} |\hat{\beta}_{jk}| : (j, k) \in I(f, \psi)\}$. By definition, $b_\nu \geq b_{\nu+1}$, $\nu = 1, \dots, M-1$; for any $\nu = 1, \dots, M-1$, there exists a unique $(j_\nu, k_\nu) \in I(f, \psi)$, such that $2^{j(\tau+1/2)} |\hat{\beta}_{jk}| = b_\nu$; $(j_{\nu_1}, k_{\nu_1}) \neq (j_{\nu_2}, k_{\nu_2})$ if and only if $\nu_1 \neq \nu_2$. The new, hard-threshold wavelet estimator f_v^* is defined by

$$f_v^*(x) = \sum_k \hat{\alpha}_{0k} \varphi_{0k}(x) + \sum_{\nu=1}^{\lfloor v^{-\frac{\pi p}{\pi-p}} \rfloor} \hat{\beta}_{j_\nu k_\nu} \psi_{j_\nu k_\nu}(x), \quad x \in \mathbb{R}.$$

Step 3. If $b_{\lfloor v^{-\frac{\pi p}{\pi-p}} \rfloor - 1} = b_{\lfloor v^{-\frac{\pi p}{\pi-p}} \rfloor}$ up to some resolution threshold (tending to zero as $n \rightarrow \infty$ and much smaller than the threshold levels in standard thresholding methods), then the terms corresponding to all the last ν 's included in the RHS of (5.4) which fall into this resolution threshold (that is, are formally equal to $b_{\lfloor v^{-\frac{\pi p}{\pi-p}} \rfloor}$) are being *removed*. Step 3 ensures the uniqueness of f_v^* . If the sequence $\{b_\nu\}$ is strictly decreasing (relative to the resolution threshold chosen), step 3 is not needed.

Let us discuss (5.4) in brief. The significance of every $|\hat{\beta}_{jk}|$ is being assessed with respect to its level j . As $v \rightarrow 0$ with $n \rightarrow \infty$, the most significant features appear first, and less significant details emerge only for sufficiently small v . The criterion for significance of the coefficients is the regularity assumption expressed in the value of τ . Notice the qualitative differences between the strategies for "more regular functions" ($\tau > -1/2$) and for "less regular functions" ($\tau < -1/2$). The boundary value $\tau = -1/2$ exactly corresponds to the critical regularity $\varepsilon = -2(\pi - p)(\tau + 1/2) = 0$ (cf. ¹⁶, Theorem 1, $d = 1$). The new thresholding approach suggested by (5.4) is quite useful in applications involving composite estimators. It is also very appropriate for the case when n is a power of two and the orthogonal discrete wavelet transform (ODWT) is being applied.

6. CRITERION FOR COMPOSING LORENTZ AND BESOV SHRINKAGE

One criterion, which is announced here for the first time, is to use the real interpolation spaces between $B_{\pi\pi}^\sigma$ and B_{pp}^s . It is known that the Besov scale is closed under real interpolation, i.e., $(B_{\pi\pi}^\sigma, B_{pp}^s)_{\theta, p(\theta)} = B_{p(\theta)p(\theta)}^{s(\theta)}$, where $0 \leq \theta \leq 1$, and $p(\theta)$ is defined by $\frac{1}{p(\theta)} = \frac{1-\theta}{\pi} + \frac{\theta}{p}$.

The parameter θ which is a coefficient in a convex combination is determining the degree to which the composite estimator is of Lorentz-type and the degree to which it is a Besov shrinkage-type estimator.

The parameter θ can be computed via cross-validation, considerations for asymptotic minimax rate, Bayesian techniques, or any other procedure for statistical parametric estimation. The resulting composite estimator is in this case highly adaptive to the local smoothness of the estimated function. We omit a more detailed exposition of this topic here, since it merits a much more detailed study.

A visual example is given in Figure 4 showing how data compression affects the shape of temperature isosurfaces (isotherm surfaces) in the iron-ore pelletizing process, as simulated via the simulator BedSim of LKAB.

7. APPENDIX: NOISE MODEL AND DEFINITIONS

In the regression case, most noise reduction algorithms start from the following additive model of a discrete signal f of N data points corrupted with noise η :

$$y = f + \eta \quad (28)$$

The vector y represents the input signal. The noise η is a vector of random variables, while the untouched values f are a purely deterministic signal. The length of these vectors is N .

We suppose that the noise has zero mean, i.e. $E\eta = 0$ and define

$$Q = E[(\eta - E\eta)(\eta - E\eta)^T] = E\eta\eta^T \quad (29)$$

the covariance matrix of η . On the diagonal we find the variances. If this matrix is diagonal, i.e. if $E\eta_i\eta_j = 0$ for $i \neq j$, the noise is defined as white or uncorrelated. If all the data points come from the same probability density, we say that the points are identically distributed. This implies of course

$$\sigma_i^2 = \sigma^2, \quad \forall i = 1, \dots, N \quad (30)$$

Noise with constant variance is called homoscedastic. Non-homoscedastic noise is heteroscedastic. Homoscedastic, white noise has a simple covariance matrix:

$$Q = \sigma^2 I. \quad (31)$$

An important density model is the joint Gaussian:

$$\phi_\eta(\eta) = \frac{1}{(2\pi)^{N/2} \sqrt{\det Q}} e^{-\frac{1}{2} \eta^T Q^{-1} \eta}. \quad (32)$$

If Gaussian noise variables are uncorrelated, they are also independent. The reverse implication holds for all densities. A classical assumption in regression theory is that of independent, identically distributed noise (i.i.d.). For Gaussian variables this is equivalent with supposing stationary and white noise.

0.0% Compression

0.0% Compression

70.9% Compression

88.7% Compression

84.9% Compression

93.7% Compression

Figure 4. Different levels of compression of the pellets-bed temperature in the processes of making iron-ore pellets. In the 1st column the temperature corresponds to a relatively non-smooth isotherm surface closer to the beginning of the process; in the 2nd column the temperature corresponds to a relatively smooth isotherm surface closer to the end of the process.

REFERENCES

1. L. T. Dechevsky, J. O. Ramsay, and S. I. Penev, "Penalized wavelet estimation with Besov regularity constraints," *Mathematica Balkanica (N. S.)* **13**(3-4), pp. 257–356, 1999.
2. T. Moguchaya, N. Grip, L. T. Dechevsky, B. Bang, A. Lakså, and B. Tong, "Curve and surface fitting by wavelet shrinkage using GM.Waves," pp. 263–274, 2005.
3. D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika* **81**(3), pp. 425–455, 1994.
4. D. L. Donoho and I. M. Johnstone, "Minimax estimation via wavelet shrinkage," *Annals of Statistics* **26**(3), pp. 879–921, 1998.
5. D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptopia?," *Journal of the Royal Statistical Society Series B* **57**(2), pp. 301–369, 1995.
6. L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics* **37**(4), pp. 373–384, 1995.
7. H.-Y. Gao and A. G. Bruce, "Waveshrink with firm shrinkage," *Statist. Sinica* **7**(4), pp. 855–874, 1997.
8. B. Vidakovic, *Statistical Modeling by Wavelets*, Wiley, New York, 1999.
9. L. T. Dechevsky, "Atomic decomposition of function spaces and fractional integral and differential operators," in *Transform Methods and Special Functions, Part A*, P. Rusev, I. Dimovski, and V. Kiryakova, eds., **2**, pp. 367–381, Fractional Calculus & Applied Analysis, 1999.
10. R. R. Coifman, Y. Meyer, and V. Wickerhauser, "Size properties of wavelet packets," in *Wavelets and their Applications*, pp. 453–470, Jones and Bartlett, 1992.
11. I. Cohen, I. Raz, and D. Malah, "Orthonormal shift-invariant adaptive local trigonometric decomposition.," *Signal Processing* **57**(1), pp. 43–64, 1997.
12. S. Chen and D. L. Donoho, "Atomic decomposition by basis pursuit," Preprint Technical Report 479, Department of Statistics, Stanford University, 1995.
13. S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing* **41**(12), pp. 3397–3415, 1993.
14. R. R. Coifman and M. V. Wickerhauser, "Entropybased algorithms for best basis selection," *IEEE Transactions on Information Theory* **38**(2), pp. 713–718, 1992.
15. J. Bergh and J. L.
16. B. Delyon and A. Juditsky, "On minimax wavelet estimators," *Applied and Computational Harmonic Analysis* **3**, pp. 215–228, 1996.