THRESHOLDING OF WAVELET COEFFICIENTS AS MULTIPLE HYPOTHESES TESTING PROCEDURE

Felix Abramovich
School of Mathematics, University of Bristol
Bristol BS8 1TW, UK

Yoav Benjamini
Department of Statistics & Operations Research
Raymond & Beverly Sackler Faculty of Exact Sciences
Tel Aviv University, Ramat Aviv 69978, Israel

Abstract

Given noisy signal, its finite discrete wavelet transform is an estimator of signal's wavelet expansion coefficients. An appropriate thresholding of coefficients for further reconstruction of de-noised signal plays a key-role in the wavelet decomposition/reconstruction procedure. [DJ1] proposed a global threshold $\lambda = \sigma\sqrt{2\log n}$ and showed that such a threshold *asymptotically* reduces the expected risk of the corresponding wavelet estimator close to the possible minimum. To apply their threshold for *finite* samples they suggested to always keep coefficients of the first coarse j_0 levels.

We demonstrate that the choice of j_0 may strongly affect the corresponding estimators. Then, we consider the thresholding of wavelet coefficients as a multiple hypotheses testing problem and use the False Discovery Rate (FDR) approach to multiple testing of [BH1]. The suggested procedure controls the expected proportion of incorrectly kept coefficients among those chosen for the wavelet reconstruction. The resulting procedure is inherently adaptive, and responds to the complexity of the estimated function. Finally, comparing the proposed FDR-threshold with that fixed global of Donoho and Johnstone by evaluating the relative Mean-Square-Error across the various test-functions and noise levels, we find the FDR-estimator to enjoy robustness of MSE-efficiency.

1 Introduction

Suppose we are given data

$$y_i = g(t_i) + \varepsilon_i, \qquad i = 1, ..., n, \tag{1}$$

where $t_i = i/n$, $n = 2^{J+1}$ for some J, ε_i's are i.i.d. normal variables with zero mean and variance σ^2 . We wish to estimate the unknown response function $g(\cdot)$ from the data without assuming any particular parametric form of g by expanding g into wavelet series generated by translations and dilations of a single function called *mother wavelet* (examples of mother wavelets are given in [Dub1]). Unlike classical Fourier sine and cosine functions, wavelets are localized both in time and frequency domains. This characteristic property allows parsimonious wavelet expansion for a wide

set of function spaces. This set includes such commonly used classes as Sobolev and Hölder scales of smooth functions, but in addition less traditional spaces, like the space of functions of bounded variations that contains spatially inhomogeneous functions, as well (see [Mey1], [DJ2] for precise details).

In the absence of random noise in the data we may find $m = n - 1 = 2^{J+1} - 1$ wavelet coefficients of g's expansion, $d_{jk}, j = 0, ..., J; k = 0, ..., 2^j - 1$, by performing Mallat's fast discrete wavelet transform (DWT) of the vector of the noiseless data $\mathbf{g} = (g(t_1), ..., g(t_n))'$. "Noisy" data only allow us to find a vector of the coefficients of the MLE estimates $\hat{\mathbf{d}}$ which is the DWT of the vector $\mathbf{y}$ of the observed data. White noise contaminates all wavelet coefficients $\hat{d}_{jk}$ equally (the DWT of the noise vector ε is also a white noise). However, it is reasonable to assume that only a few $\hat{d}_{jk}$ contain information about the real signal while others represent a random noise. The goal is to extract these significant coefficients and to ignore others. Such an extraction can be naturally performed by thresholding the $\hat{d}_{jk}$'s :

$$\hat{d}^*_{jk} = \begin{cases} \hat{d}_{jk} & , \ |\hat{d}_{jk}| \geq \lambda \\ 0 & , \ |\hat{d}_{jk}| < \lambda \end{cases}$$

where λ is the threshold value.

The well-known [DJ1] *global* threshold $\lambda = \sigma\sqrt{2 \log n}$ can be shown to imply a wavelet estimator which risk is *asymptotically* "close" to the minimal risk corresponding to the optimal (but unknown) thresholding rule. However, it should be noted that such a threshold depends on the data only through the estimated σ and for fixed n is otherwise the same, for all samples and for all kinds of functions. For *finite* samples Donoho and Johnstone further suggested to always keep the coefficients of the first "coarse" j_0 levels, even if these coefficients do not pass the thresholding level. In their paper they used $j_0 = 5$. Obviously, any fixed choice of j_0 does not change the *asymptotic* properties. Intuitively, the proper choice of j_0 should depend on the smoothness of the estimated function. It might be argued that j_0 should be greater for oscillating functions but smaller for smooth ones. The examples considered in Section 4 illustrate the fact that the choice of j_0 in practice may strongly affect the corresponding estimators. Some other thresholding rules are proposed in [DJ2], [Nas1].

From the statistical viewpoint, thresholding, as was also pointed out by [DJ1], is closely related to another data-analytic approach to model building involving *multiple* hypotheses testing: for each coefficient, test whether it is zero, and keep only the significant ones. Classical approaches to hypotheses testing in this case face serious problems because of the large number of hypotheses being tested simultaneously: if the error is controlled at an *individual* level, the chance of keeping erroneously a coefficient is extremely high; if the *simultaneous* error is controlled, the chance of keeping a coefficient is very low. Recently, [BH1] have suggested the False Discovery error Rate (FDR) criterion as an alternative method in multiple hypotheses testing problems. This paper proposes a statistical procedure for thresholding wavelet coefficients which is based on the FDR-approach. In a way it controls the expected proportion of incorrectly kept coefficients among those chosen for the model. The resulting FDR-thresholding procedure is *inherently adaptive* due to the

adaptiveness of the criterion being controlled.

In Section 2 we describe the FDR criterion of [BH1] and construct the FDR-procedure for thresholding wavelet coefficients in Section 3. Several test-cases considered in Section 4 demonstrate the limitation of non-adaptive global thresholding and compare it with alternative FDR-thresholding procedure. Evaluating the relative Mean-Square-Errors across the various test-functions and noise levels, we find the FDR-estimator to enjoy robustness of MSE-efficiency.

2 Thresholding as multiple hypotheses testing problem

We consider here the problem of testing the $m = 2^{J+1} - 1$ hypotheses $H_{jk} : d_{jk} = 0$, where d_{jk}'s are wavelet coefficients of a true (but unknown) function g. Of these hypotheses, m_1 are false, or equivalently the corresponding coefficients should be kept in the wavelet expansion. The other $m_0 = m - m_1$ coefficients are in fact zeroes and ideally should all be dropped.

Separating the coefficients into those which are zero and those which are not zero may seem an idealization of the real situation: in practice very few coefficients of a true function are identically zero, while many of them will be merely very small. Nevertheless, if we consider a coefficient to be incorrectly kept in the model either if it is truely zero and kept, or if it is truly of one sign but is kept in the model with the wrong sign (directional error), then the case where such coefficients are considered to be exactly zero is the extreme case that needs to be controlled (see [Tuk1] for a discussion of this point of view).

As we view the problem of thresholding wavelet coefficients in the framework of hypotheses testing we have to face the problem caused by the multiplicity of the errors that have to be controlled simultaneously. One approach is the "don't worry" that ignores the problem altogether: conduct *each* test at the usual significance level, say 0.05, as if it were the only one tested. Allas, with 1023 hypotheses to be tested (for 1024 observations) about 50 would be found (1023×0.05 on the average) significant, even when the representation of the true function needs none. Hence, a stronger control of error is needed, and the most commonly used alternative is the "panic" approach: control the probability that no truly zero coefficient enters the model (Bonferroni's approach). The control of such a stringent criterion is well-known to reduce power, implying that too few coefficients will enter the model. It is therefore hardly used in practice in other similar problems such as variable selection in regression, or choosing autoregressive terms in time series analysis.

Adapting the general idea of [BH1] we analyse the performance of a thresholding procedure as follows. Let R be the number of coefficients that are not dropped by the thresholding procedure for a given sample, and are, thus, kept in the representation. Of these R coefficients, S are correctly kept in the model and V are erroneously kept, $R = V + S$. The error in such a procedure is expressed in terms of the random variable $Q = V/R$ - the proportion of the coefficients kept in the representation that should have been dropped. Naturally we define $Q = 0$ when $R = 0$ since no error of this type can be made when no coefficient is kept.

The False Discovery Rate of Coefficients (FDR) can be now defined as the expectation of Q,

and thus reflects the expected proportion of erroneously kept coefficients among the ones kept in the representation. Following [BH1] we suggest maximizing the number of kept coefficients subject to controlling of the FDR to some level q.

Two properties of the FDR are important to note:

a) If the data are pure noise, i.e., all true coefficients are zero, then controlling the FDR implies the control of the probability of including erroneously even one coefficient (Bonferroni's approach). Because of this property the traditional levels for significance testing were used, e.g., $q = .01$ or $q = .05$.

b) The FDR increases with an increase in the number of incorrectly kept coefficients, and decreases as more coefficients are chosen to be kept. If a number of large true coefficients are present, R will tend to be larger and, therefore, the FDR will tend to be smaller. Thus, the error rate will respond to the complexity of the estimated function.

Note that Donoho-Johnstone's thresholding can be also viewed as a multiple hypotheses testing procedure. Their thresholding rule is equivalent to rejecting each null-hypothesis $H_{jk} : d_{jk} = 0$ at a critical value $\sigma\sqrt{2\log n}$. Using the well-known asymptotics $\Phi(-a) \sim \phi(a)/a$ for large a, where Φ and ϕ are standard normal c.d.f. and p.d.f. respectively, one can verify that the corresponding significance level (the same for all tests) would be approximately $(n\sqrt{\pi \log n})^{-1}$. Thus, we see that not only Donoho and Johnstone's procedure is equivalent to the "panic" procedure of controlling the probability of even one erroneous inclusion of a coefficient at the level $(\sqrt{\pi \log n})^{-1}$, but the level at which this error is controlled approaches zero as n increases. No wonder that the loss of power is such that it requires the ad-hoc remedy of supressing the thresholding procedure for the first coarse levels.

Finally, we note that while this paper only deals with the estimation of functions on the real line, it is straightforward to extend the suggested thresholding algorithm to R^d and to recovering images on R^2 in particular. The details are obvious, and we do not give them here.

3 FDR-procedure

Applying the procedure of [BH1] for wavelet thresholding yields the following procedure:

1) For each $\hat{d}_{jk}$ calculate the corresponding two-sided p-value, p_{jk}, testing $H_{jk} : d_{jk} = 0$,

$$p_{jk} = 2(1 - \Phi(|\hat{d}_{jk}|/\sigma))$$

2) Order the p_{jk}'s according to their size, $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$, where each of the $p_{(i)}$'s corresponds to some coefficient d_{jk}.

3) Starting with $i = 1$, let k be the largest i for which $p_{(i)} \leq (i/m)q$. For this k calculate $\lambda_k = \sigma\Phi^{-1}(1 - p_{(k)}/2)$.

4) Threshold all coefficients at level λ_k.

[BH1] proved that for independent Gaussian noise in the model (1) the above procedure controls the FDR at the (unknown) level $(m_0/m)q \leq q$. The procedure also controls the FDR if the marginal distribution of the noise in model (1) is other than Gaussian, say F, with the only change in the above procedure being to replace Φ by F .

Computational note. Step 4) can be replaced by taking the k coefficients corresponding to the k smallest p-values. Furthermore, since a coefficient can be kept in the model only if the corresponding $p_{(i)} \leq q$, it has to be at least larger (in absolute value) than $\lambda_{min} = \sigma\Phi^{-1}(1 - q/2)$. Therefore, the above steps could be performed only for $|\hat{d}_{jk}| \geq \lambda_{min}$, making large computational savings in sorting, etc.

Note that in a specific sample thresholding is done effectively at some (adaptive) level between $\lambda_{max} = \sigma\Phi^{-1}(1 - q/2n)$ and $\lambda_{min} = \sigma\Phi^{-1}(1 - q/2)$. For practically used sample sizes $n = 2^{J+1}$, $J + 1 = 7, 8, ..., 14$ and the traditional $q = 0.05$, the Donoho-Johnstone global threshold λ satisfies $\lambda_{min} \leq \lambda \leq \lambda_{max}$. In fact, over this range λ_{max} is larger than λ by $5\% - 15\%$. Fig. 1 displays some FDR-thresholds for $n = 1024$, assuming $\sigma = 1$. While Donoho-Jonstone's $\lambda = 3.723$, if only one (the largest) coefficient enters the representation it should pass the threshold of 4.061. If exactly four coefficients are significant, the corresponding FDR-threshold is equal to the global DJ-threshold. As more coefficients enter the representation, the effective FDR-threshold is set at lower values.

The procedure can be motivated as a samplewise implementation of the "maximization subject to control" aproach. If $p_{(i)}$ corresponds to a potential threshold, exactly i coefficients will pass the threshhold and be kept in the representation. The expected number of incorrectly kept coefficients is $m_0 p_{(i)} \leq m p_{(i)}$, as for these coefficients the estimated p-values are uniformly ditributed. Thus, the expected proportion of incorrectly kept coefficients among those kept in the model can be given by $m_0 p_{(i)}/i \leq m p_{(i)}/i$, which we wish to control below q. Selecting as many as possible coefficients means choosing the largest possible i, leading to step 3) of the procedure.

4 Examples

We consider the performance of two FDR-estimators with $q = 0.01$ (FDR01) and $q = 0.05$ (FDR05). We compare them with three versions of Donoho and Johnstone estimators corresponding to three different thresholding starting levels of j_0: $j_0 = 1$ (DJ1), $j_0 = 3$ (DJ3), the default value in Nason and Silverman (1994), and $j_0 = 5$ (DJ5) used by Donoho and Johnstone. All the thresholding procedures were tried on the following test cases (see Fig. 2) :

1) $g(t) = (t - 0.4)^2$

2) $g(t) = \min(2t, -2(t - 1))$ (triangular function)

3) $g(t) = (t - 0.3)_+ - (t - 0.7)_+$ (block function)

Fig. 1. Threshold for k-th largest coefficient (n=1024, q=0.05)

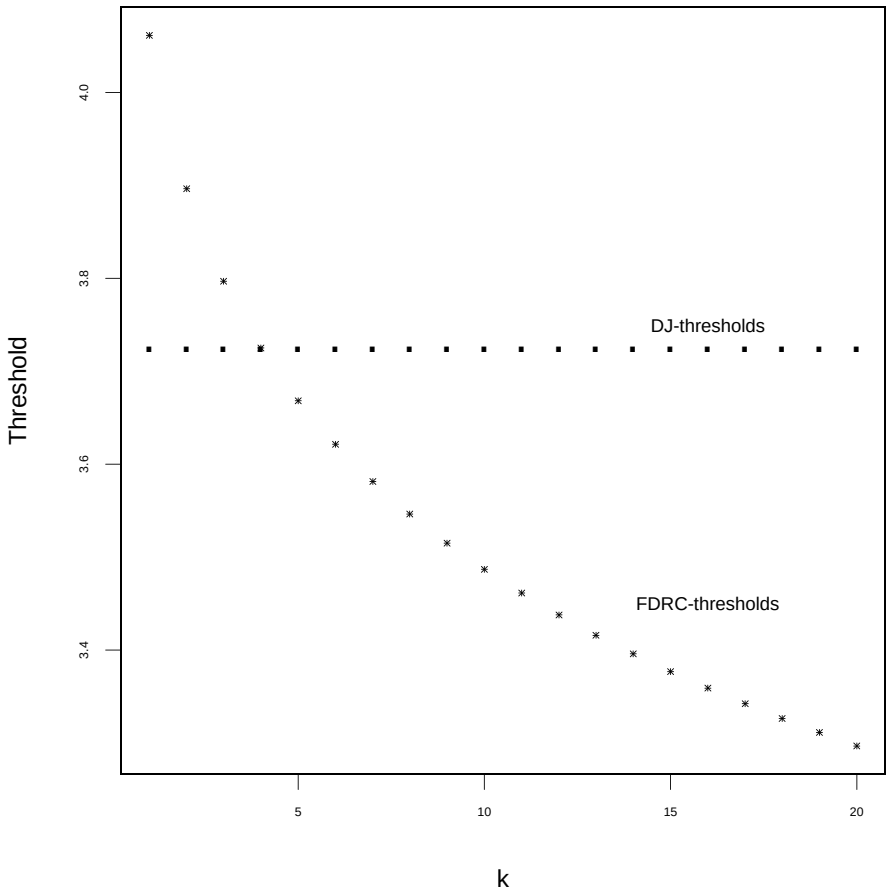

Figure 1: Threshold for the k-th largest coefficient ($n = 1024$, $q = 0.05$).

4) $g(t) = \sum_j h_j K(t - t_j)$, where $K(t) = (1 + \text{sign}(t))/2$,
$(t_j) = (.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81)$,
$(h_j) = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 5.1, -4, 2)$
([DJ1] Blocks example)

5) $g(t) = \sum_j h_j K((t - t_j)/w_j)$, where $K(t) = \max((1 - |t|)^4, 0)$,
(t_j) are the same as in the previous example,
$(h_j) = (4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4, 2)$,
$(w_j) = (.005, .005, .006, .01, .01, .03, .01, .01, .005, .008, .005)$
(similar to [DJ1] Bumps example)

For each case we ran simulations with four different values of σ to satisfy the signal-to-noise ratio $\delta = SD(\mathbf{g})/\sigma = 7, 4, 2$ and 1 respectively. 1024 design points were taken equally spaced on $[0, 1]$ and the data were generated for 500 replications of every combination of cases and σ's by adding to $g(t(i/n))$ independent random noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

To find the vector of wavelet estimates $\hat{\mathbf{g}}$ Mallat's ([Mal1]) well-known algorithm of decomposition-reconstruction was used. On the decomposition step we found the wavelet coefficients $\hat{d}_{jk}$ by the fast DWT of Mallat using the compactly supported mother wavelet D_4 from Daubechies's family (see [Dub1]). Thresholding $\hat{d}_{jk}$ and performing the fast inverse DWT of the thresholded coefficients on the reconstruction step we derived the corresponding function estimates. The noise level σ was estimated by the standard deviation of the wavelet coefficients $\hat{d}_{Jk}$ at the finest level J and performed quite satisfactorily. All the programming was done in the statistical package S-Plus using the S wavelet software developed by Nason and Silverman (the description may be found in Nason and Silverman, 1994), and the built in normal random numbers generator. The goodness of fit of each estimator was measured by its mean squared error $\text{MSE} = n^{-1}||\mathbf{g} - \hat{\mathbf{g}}||_{l^2}$ averaged over all 500 replications. The standard error of MSE was about $0.5 - 2.0\%$ of its estimated mean value.

For each test-case, we found the best estimator among the five ones, i.e., the one achieving the minimum MSE. Then the relative MSE for each estimator was evaluated as $(\min \text{MSE}_j)/\text{MSE}_i, i = 1, ..., 5$. The results of simulation studies are summarized in Table 1.

Considering first the three DJ-estimators one sees that the relative MSE varies strongly for different j_0 and depends on the smoothness of the function. For example, for relatively smooth functions (Cases 1, 2) DJ3 is highly preferable over DJ5, while in oscillating examples (Cases 4, 5) DJ5 performs much better.

Comparing the performance of FDR and DJ-estimators one concludes (see Table 1) that for smooth functions (Cases 1, 2) FDR01 performs slightly better than FRD05 (for such functions we would like to be more conservative in including additional coefficients in the representation) but both of them give in to DJ3 which is undoubtedly the best estimator for these Cases for all δ's. However, for functions with rapid local changes (Cases 4, 5) FDR05 is highly preferable over FDR01 and DJ3, and even somewhat better than DJ5, the best (for these cases) among DJ-estimators.

Then we found the minimal relative MSE (MRMSE) of each estimator over all the cases (see bolded numbers in Table 1 and Fig. 3). The MRMSE reflects the loss of effectiveness at the most

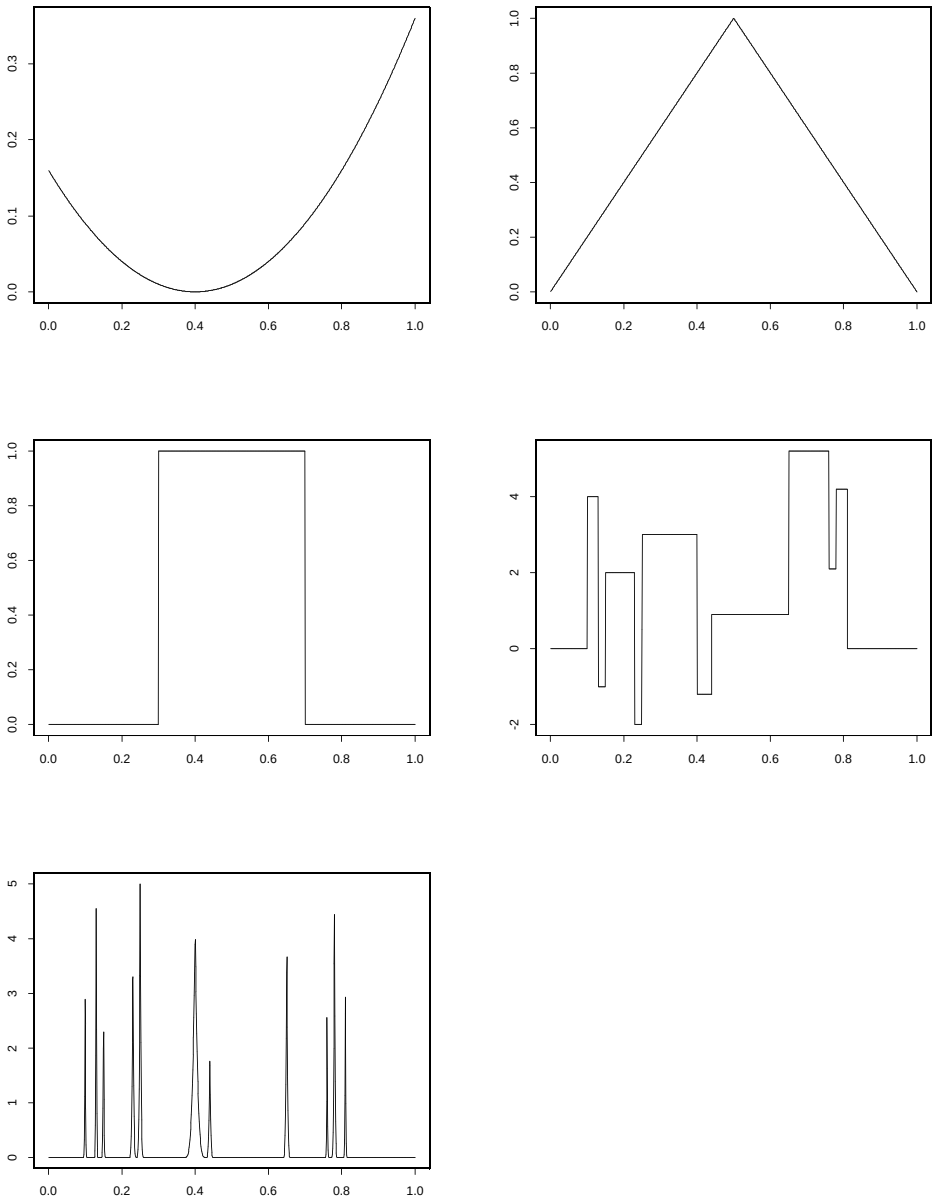

Figure 2: Test-cases.

challenging test-case for each estimator and is a natural measure of its robustness to the test-cases. This measure is further studied as a function of signal-to-noise ratio, which characterizes the robustness of the procedures in face of different noise levels. Fig. 3 clearly shows that the proper choice of j_0 in the Donoho-Johnstone algorithm should depend on the noise level. For large noise (δ small) the optimal j_0 is small since the wavelet coefficients are strongly influenced by noise and we should threshold them starting from the very coarse levels in order to decrease the noise in the reconstruction. For moderate noise "significant" coefficients (especially those at low levels) reflect the real signal and should not be dropped. In contrast to the behavior of the DJ-estimators, FDR-estimators are much less sensitive to the noise level due to their adaptiveness, and perform quite satisfactorily for all noise levels and test cases studied.

In conclusion, it might be interesting to note that the FDR-approach for choosing coefficients in the wavelet representation is philoshophically very different from traditional methods. Usually, we seek the most compact presentation possible and enrich the model only if some condition is met. Here we try to keep as many coefficients as possible but subject them to a certain control rule. While this approach might be carried over to other problems of model selection, its usefulness should be demonstrated at each case separately, as it was done here.

Acknowledgements. The authors are grateful to Mark Low for a very fruitful discussion of possible applications of the FDR-approach in a wide range of statistical problems including wavelets.

Fig.3. Minimal Relative MSE (MRMSE)

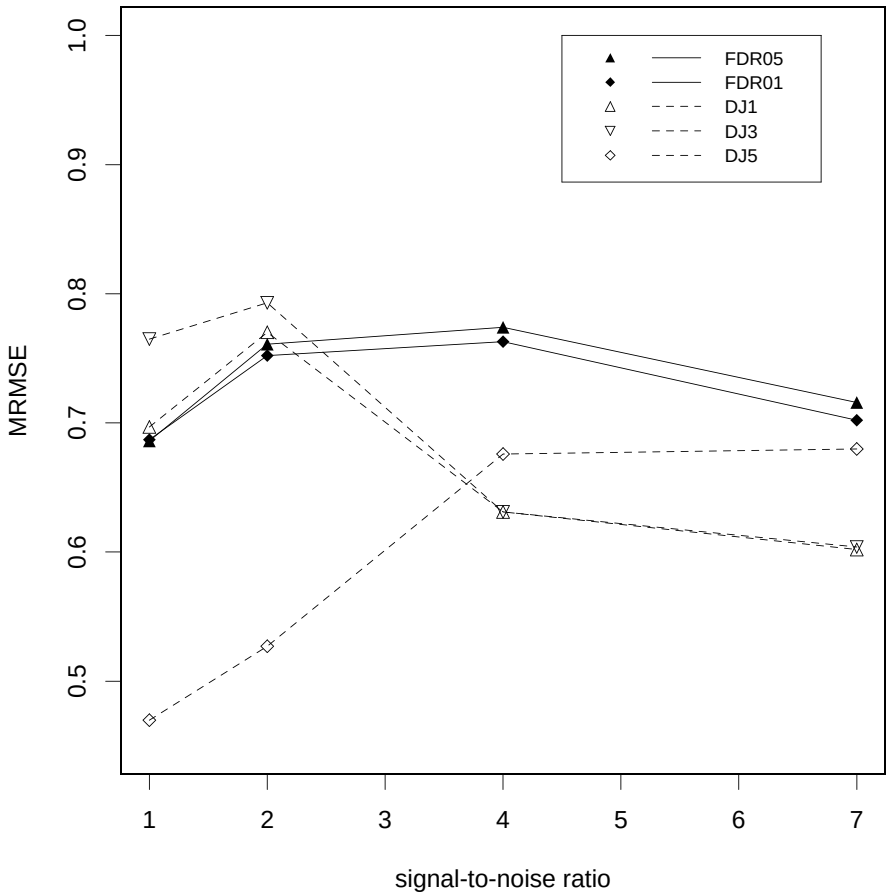

Figure 3: Minimal Relative MSE (MRMSE).

Table 1. Relative MSE (averaged over 500 replications).

		FDR05	FDR01	DJ1	DJ3	DJ5
$\delta = 7$						
	Case 1	0.859	0.962	0.917	1.000	0.690
	Case 2	**0.716**	**0.702**	0.714	0.818	1.000
	Case 3	1.000	0.951	0.923	0.923	0.941
	Case 4	1.000	0.784	0.658	0.658	0.713
	Case 5	1.000	0.747	**0.602**	**0.604**	**0.680**
$\delta = 4$						
	Case 1	**0.774**	0.806	0.798	1.000	**0.676**
	Case 2	0.799	0.849	0.844	1.000	0.885
	Case 3	1.000	0.875	0.857	0.873	0.956
	Case 4	1.000	0.848	0.774	0.781	0.849
	Case 5	1.000	**0.763**	**0.631**	**0.631**	0.719
$\delta = 2$						
	Case 1	**0.761**	0.786	0.796	1.000	**0.527**
	Case 2	0.767	0.791	0.793	1.000	0.729
	Case 3	0.816	**0.752**	**0.770**	0.867	1.000
	Case 4	0.992	0.934	0.913	0.922	1.000
	Case 5	1.000	0.857	0.787	**0.793**	0.880
$\delta = 1$						
	Case 1	0.760	0.805	0.767	1.000	**0.470**
	Case 2	**0.686**	**0.687**	**0.697**	1.000	0.553
	Case 3	0.722	0.694	0.718	0.793	1.000
	Case 4	0.801	0.733	0.737	**0.765**	1.000
	Case 5	1.000	0.857	0.845	0.878	0.959

References

[BH1] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc., Ser. **B 57** (1995) 289-300.

[Dub1] Daubechies, I.: *Ten Lectures on Wavelets.* SIAM (1992).

[DJ1] Donoho, D.L., Johnstone, I.M.: Ideal spatial adaption by wavelet shrinkage. Biometrika (to appear).

[DJ2] Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. J. Am. Stat. Ass (1994) (to appear).

[Mey1] Meyer, Y.: *Wavelets and Operators.* Cambridge University Press (1992).

[Mal1] Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Recogn. and Machine Intel. **11** (1989) 674-693.

[Nas1] Nason, G.P.: Wavelet regression by cross-validation. Tech. Report 447, Dep. of Stat., Stanford University (1994).

[NS1] Nason, G.P., Silverman, B.W.: The discrete wavelet transform in S. J. Comp. Graph. Statist. **3** (1994) 163-191.

[Tuk1] Tukey, J.W.: The philosophy of multiple comparison. Statist. Sci. **6** (1991) 100-116.