Ideal Spatial Adaptation by Wavelet Shrinkage, a review

Mathematical Statistics (DS-GA-3001) project

Nikolaos Tsilivis, nt2231@nyu.edu

1 Introduction

We will review some interesting themes from the classical work of David Donoho and Iain Johnstone from 1992, titled "Ideal Spatial Adaptation by Wavelet Shrinkage". This paper is known for the introduction of wavelets into statistical estimation tasks.

2 Problem formulation

The problem of interest is that of function recovery from noisy, evenly sampled, observations. In other words, we are given one-dimensional data

$$y_i = f(t_i) + \epsilon_i, \qquad i = 1, \dots, n,\ t_i = i/n, \tag{1}$$

where ϵ_i are independently sampled from $\mathcal{N}(0, \sigma^2)$, and our goal is to estimate the noiseless part $f := (f(t_1), \dots, f(t_n))^\top$ on these n points. We will measure the performance of an estimate $\hat{f} := (\hat{f}(t_1), \dots, \hat{f}(t_n))^\top$ of f by the risk

$$R(\hat{f}, f) = \frac{1}{n}\mathbb{E}\|\hat{f} - f\|^2. \tag{2}$$

This is what we called in class as *Gaussian sequence model*, enhanced with the extra assumpion that our samples are equally spaced inside the $[0, 1]$ interval.

3 Introduced frameworks

Before proposing their approach, the two authors introduce two useful and novel (to my understanding) frameworks to think about the problem; one that unifies *spatially adaptive* methods, and a second that serves as a notion of *ideal* performance obtained by such methods.

3.1 Spatial Adaptivity

In the context of statistics, spatial adaptivity refers to the ability of an estimator to adapt its "behavior" (in means of adjusting its "hyperparameters") given the actual observations that it encounters. For example, a simple estimator for our problem (1) clusters the points $t_i, i = 1, \dots, n$, into L groups of same size, and outputs the average of each of these groups as the estimation of the function values that lie in the corresponding group (Fig. 1). As we saw in class, if we are willing to assume that our true function is smooth in terms of Lipschitz or Hölder smoothness, then this yields an estimation rate better than the one obtained from the maximum likelihood estimator. Notice, however, that the number of groups L is a "hyperparameter" of the estimator and, in the proof contained in the notes, knowledge about the exact degree of smoothness (Lipschitz constant) was assumed to optimally pick L. This may be too much to ask. Instead, one may try to "select" L from the actual observations, and adapt their estimator. This is the idea of spatial adaptivity.

We can summarise the above as follows. Spatially adaptive estimators $\hat{f}$ are defined as

$$\hat{f}(\cdot) = T(y, d(y))(\cdot), \tag{3}$$

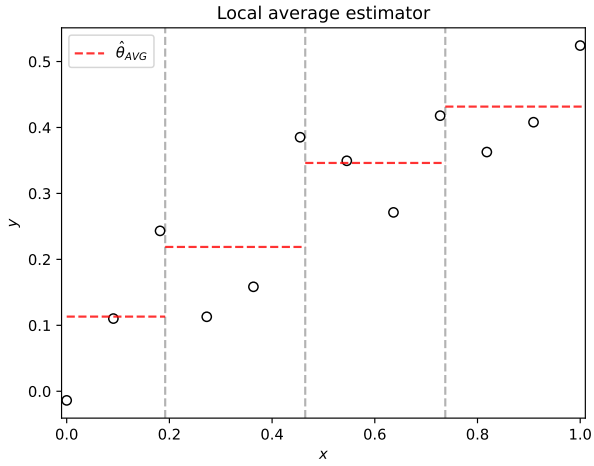

Figure 1: Local average estimator. We partition the observations into $L = 4$ groups and, then, compute the average on each interval.

where $T(y, \delta)$ is the reconstruction expression of the estimator with spatial parameter δ and $d(y)$ is a *data-adaptive* choice of δ. In the previous example, $T(y, \delta)$ would be the average of the "correct" group (the interested reader can use indicator functions to analytically express this) and δ would be the list of numbers that define the partition. There are several different approaches for designing the selector $d(y)$ and they can be viewed as solving a model selection problem. However, it may be difficult to argue about the theoretical performance of such procedures.

3.2 Ideal Adaptivity with Oracles

If we would like to theoretically design a spatially adaptive estimator of the form (3), then we may wonder what is the optimal performance we can hope for in this context? Fixing the reconstruction formula, restricts us to varying only the spatial parameter δ. But, what is the best choice of δ for an underlying (unknown) function?

An answer to this question requires *oracular* access to the function, which, in principle, can not be expected to be obtained from a finite number of observations. Such knowledge can only be provided by an *oracle* that consults f. Note, however, that the oracle will not tell us f itself, but will tell us, for our method $T(y, \delta)$, what is the best choice (measured by the risk) of δ given the true f.

The authors motivate this framework with the following example. Consider the case of estimating a function f that is piecewise polynomial of degree D, with L pieces, $I_1, \ldots, I_L$. It seems reasonable to define an estimator that is itself a piecewise polynomial of degree D. An oracle, then, can supply us with the *true partition* of f. Given this knowledge, we can design L linear models $Y_i = X_i\beta_i + E, \beta_i \in \mathbb{R}^{D+1}, Y_i \in \mathbb{R}^n$ (in order to do so, we need to observe that a polynomial model is nothing but a linear model with "transformed" covariates), and find the least squares estimator for each of the β_i (by X_i we mean the data that lie in the i-th set of the partition). We used this linear regression model in class when we talked about regularization and the lasso estimator in high-dimensional statistics. The "in-sample error" of the least squares estimator $\hat{\beta}_i$ for each of the $\beta_i's$ is then

$$\mathbb{E}\|X_i\beta_i - X_i\hat{\beta}_i\|_2^2 = \underbrace{(D+1)}_{\#\text{ of parameters}} \sigma^2. \tag{4}$$

The estimator across the whole domain can be written as

$$\hat{f}_{PP}(t) = \sum_{i=1}^{L} \beta_i \mathbb{1}_{t \in I_i} \tag{5}$$

and, finally, for the risk (2) it holds

$$R_{n,\sigma^2}(\hat{f}_{PP}, f) = \frac{1}{n} L(D+1)\sigma^2. \tag{6}$$

This is a risk that corresponds to an ideal situation, and provides (at least asymptotically) an upper bound of linear nature on the estimation rate we are hoping to achieve. The rates of non-adaptive methods are typically worse and grow as $O(\sqrt{n})$, making themselves unsatisfactory for this estimation task. The success of the wavelet methods that we will introduce shortly is that they are able to "close" this gap. Restricting ourselves into the same piecewise polynomial true function scenario, not only these wavelets perform "almost" as good as the ideal piecewise polynomial estimator when furnished with an oracle, but they can also yield an estimator $\hat{f}^*$ that depends on the data **alone** and is "almost" as good. Even more remarkably, the situation is similar for any underlying function f. That is, wavelets provide an, in a way, universal spatially adaptive method.

4 Spatial adaptivity with wavelets

Wavelets are families of functions that were introduced, in their current form, in the 70's and 80's, and served initially as tools for signal analysis and processing in many disciplines of physics and engineering. Intuitively, they can be thought as generalizations of the Fourier transform/series, allowing information extraction from a signal in localised time intervals with varying "scales". A so-called *mother* wavelet is being translated and dilated (vertically scaled) in order to generate a whole basis of wavelets, in the same way that a complex sinusoid gives birth to others with varying frequency in the context of Fourier analysis. See Fig. 2 for two continuous wavelets. For each mother wavelet, there are 3 important parameters

- M: number of vanishing *moments*,
- S: support *width*,
- j_0: cutoff frequency.

In this work, however, we are concerned with discrete wavelets, i.e. functions defined on the integers.

Suppose we have $n = 2^{J+1}$ observations $y_1, \ldots, y_n$ from (1), where $J \in \mathbb{N}$, and let $y = (y_1, \ldots, y_n)^\top$. A choice of a wavelet function induces an orthogonal matrix $W \in \mathbb{R}^{n \times n}$ (property that stems from the definition of the wavelet), where each of its rows is a different "version" of the mother wavelet, which corresponds, conventionally, to the first row. This allows the transformation of y into a new vector

$$w = Wy. \tag{7}$$

The assumption of n being a power of 2 allows us to number the rows of W with two indices $j \in \{0, \ldots, J\}$ and $k \in \{0, \ldots, 2^j - 1\}$; j and k represent the dilation and translation, respectively, with respect to the mother wavelet. Since W is orthogonal ($W^{-1} = W^\top$), reconstruction is possible in the form of

$$y = W^\top w. \tag{8}$$

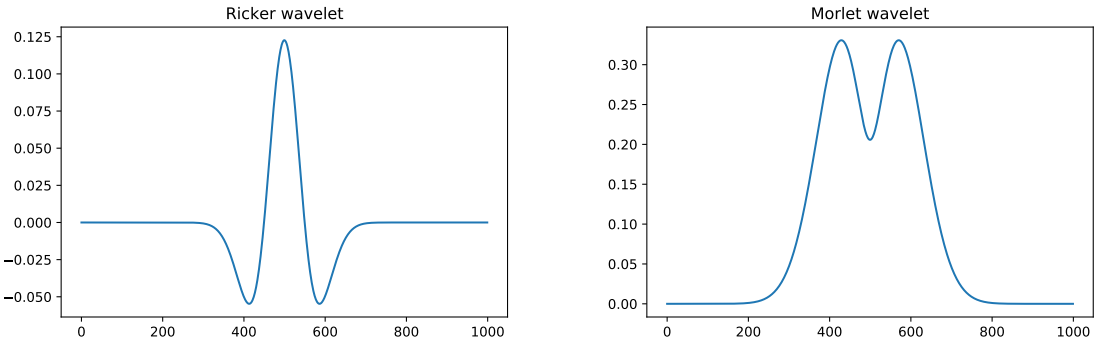

Figure 2: Two continuous wavelets. Left: Ricker wavelet with width=50. Right: Modulus of a Morlet wavelet with width=0.8 and "frequency"=1.

We say that w contains the *wavelet coefficients*, $w_{j,k}, j = 0, \dots, J, k = 0, \dots, 2^j - 1$, of y. For instance, if $n = 8 = 2^{2+1}$, then w contains in order $\underbrace{w_{0,0}}_{j=0}, \underbrace{w_{1,0}, w_{1,1}}_{j=1}, \underbrace{w_{2,0}, w_{2,1}, w_{2,2}, w_{2,3}}_{j=2}, w_{-1,0}$, where by convention we denote the last coefficient by $w_{-1,0}$. For each data sample y_i, it holds separately

$$y_i = \sum_{j,k} w_{j,k} W_{jk}(i), \tag{9}$$

where $W_{jk}(i)$ denotes the i-th element of the jk-th wavelet. It follows from the theory of continuous wavelet transforms that as long as $j \in [j_0, J - j_1]$ and $k \in (S, 2^j - S)$ for some j_1 (we are away from the boundary cases), then W_{jk} as a function of i enjoys the following two properties

1. M vanishing moments:
$$\sum_{i=1}^{n} i^l W_{jk}(i) = 0, \quad \forall l = 0, \dots, M. \tag{10}$$
2. It is non zero only in $[2^{J-j}(k - S), 2^{J-j}(k + S)]$.

Because of these localised properties, it is reasonable to expect that the wavelet coefficients $w_{j,k}$ of an unknown signal f are non-zero for only a few values of j and k. Indeed, in the special case of polynomials of degree $D \leq M$, property 1. above (10) immediately implies that $w_{j,k} \neq 0$ only when $j < j_0$.

Furthermore, applying a wavelet transform on our problem (1) yields

$$Wy = Wf + We. \tag{11}$$

Since $e \sim \mathcal{N}(0, \sigma^2 I_n)$ and W is an orthogonal matrix, it is true that $We \sim \mathcal{N}(0, \sigma^2 W^\top I_n W) = \mathcal{N}(0, \sigma^2 I_n)$, so the wavelet coefficients of the noise are also Gaussian with the same mean and variance. Coupled with the previous "sparsity" property, this provides the key idea of statistical estimation with wavelets (which we quote from the paper)

Every empirical wavelet coefficient therefore contributes noise of variance σ^2, but only a very few wavelet coefficients contribute signal.

Thus, it is natural to use an estimator that selects only some of the wavelet coefficients of the data. In the language of Sec. 3.1, the definition of a *selective wavelet* estimator is

$$T_{SW}(y, \delta) = \sum_{(j,k) \in \delta} w_{j,k} W_{jk}, \tag{12}$$

where δ here is a list of pairs.

Further extending the arguments used above about the "sparsity" of the coeffiecients, one can show that, when f is a piecewise polynomial of degree D, there are only $O(J) = O(\log n)$ non-zero of them. If an oracle then provides these optimal wavelet coefficients $\delta^\star$, and we also observe that the reconstruction formula essentially specifies the least-squares estimate for this problem, we can conclude that the obtained ideal risk satisfies

$$R_{n,\sigma^2}(T_{SW}(y, \delta^\star), f) = O\left(\frac{\sigma^2 \log n}{n}\right). \tag{13}$$

The assertion is that, in the ideal regime, we didn't lose too much by using wavelets instead of piecewise polynomials (compare with (6)). We are only a factor of $\log n$ shy of it. Coming up next with an estimator that uses an actual data-dependent selector for the list of coeffiecients δ and is only $\log n$ itself away from the ideal $T_{SW}(y, \delta^\star)$, we will be able to show what was advertised in the end of Section 3.2; there is a spatial wavelet estimator that has a rate of $O(\frac{n}{\log^2 n})$, which is almost as good as the linear rate obtained from *ideal* piecewise polynomial estimation, and a lot better than the typical $O(\sqrt{n})$.

4.1 Adaptive wavelet shrinkage and oracle inequality

Take a look at equation (12). The following operations are taking place in sequence: (i) first, we perform the wavelet transform on the observations y to obtain $w_{j,k}$, then (ii) a few of them live to see another day inside δ, while the rest are set to 0, and (iii) an inverse wavelet transform is being applied to obtain the final estimator. Part (ii) is an estimator that decides on each wavelet coefficient, whether to keep it or not, and can be written in a spatial adaptive language as: $(T_H(w, \delta))_i = \delta_i w_i$, for $i = 1, \ldots, n$[1] with $\delta_i \in \{0, 1\}$. The subscript H hints *"hard thresholding"* and will be justified shortly. The risk of the selective wavelet estimator can thus be written as

$$\begin{aligned} R(T_{SW}(y, \delta), f) &= \frac{1}{n}\mathbb{E}\|T_{SW}(y, \delta) - f\|_2^2 \\ &= \frac{1}{n}\mathbb{E}\|W^\top (T_H(Wy, \delta)) - W^\top \theta\|_2^2 \\ &= \frac{1}{n}\mathbb{E}\|T_H(Wy, \delta) - \theta\|_2^2, \end{aligned} \tag{14}$$

where θ denotes the *true* wavelets coefficients of f and we used the fact that W is an orthogonal matrix. Therefore, in order to evaluate the performance of a wavelet estimator, we simply need to look at different estimators for the coefficients. As mentioned before, the coefficients also obey the Gaussian sequence model

$$w_i = \theta_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{15}$$

The following lemma establishes the ideal performance we aspire to achieve with the T_H estimator.

Lemma 1. *The ideal estimator $T_H(\cdot, \delta^*)$ for the above model is given by the hard thresholding estimator $T_H(w, \delta) = (w_i \mathbb{1}_{|\theta_i| > \sigma})_{i=1}^n$, and it attains risk equal to $\frac{1}{n}\sum_{i=1}^n \min(\theta_i^2, \sigma^2)$.*

Proof.

$$\begin{aligned} R(T_H(w, \delta), \theta) &= \frac{1}{n}\mathbb{E}\|T_H(w, \delta) - \theta\|_2^2 \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}(\delta_i w_i - \theta_i)^2. \end{aligned} \tag{16}$$

[1] for the sake of simplicity we abandon for the rest of this section the dyadic indexing.

For each i, there are two cases

$$\begin{aligned}\mathbb{E}(\delta_i w_i - \theta_i)^2 &= \begin{cases}\mathbb{E}(w_i-\theta_i)^2,\ \delta=1\\ \mathbb{E}(0-\theta_i)^2,\ \delta_i=0\end{cases}\\ &= \begin{cases}\sigma^2,\ \delta_i=1\\ \theta_i^2,\ \delta_i=0.\end{cases}\end{aligned} \tag{17}$$

An oracle can supply us, now, with the information of when we should keep the coefficient ($|\theta_i| > \sigma$) and when we should throw it away ($|\theta_i| \leq \sigma$). The ideal risk therefore is the sum of the individual contributions, $\frac{1}{n}\sum_{i=1}^n \min(\theta_i^2, \sigma^2)$. □

Remarkably, this oracular performance can be approached by a similar *data-adaptive* estimator. Define the "soft threshold" non-linearity as $\eta_S(w,\lambda) = \operatorname{sgn}(w)\max(|w|-\lambda, 0)$.

Theorem 2. *The estimator*

$$\hat{\theta}_i = \eta_S(w_i, \sigma\sqrt{2\log n}), \quad i = 1,\dots,n \tag{18}$$

satisfies

$$\mathbb{E}\|\hat{\theta}-\theta\|_2^2 \leq (2\log n + 1)(\sigma^2 + \sum_{i=1}^n \min(\theta_i^2, \sigma^2)). \tag{19}$$

Notice in the right hand side the ideal risk of the hard thresholding estimator. That is, asymptotically, $\hat{\theta}$ is at most a logarithmic factor away from the ideal performance.

Proof. As in the previous proof, we can consider each term separately. To simplify notation, we set $\sigma = 1$, and let $\lambda = \sqrt{2\log n}$. First, notice that $w_i \sim \mathcal{N}(\theta_i, 1)$. We have

$$\begin{aligned}\mathbb{E}(\hat{\theta}_i - \theta_i)^2 &= \mathbb{E}(\eta_S(w_i,\lambda) - \theta_i)^2\\ &= \mathbb{E}(\eta_S(w_i,\lambda) - w_i + w_i - \theta_i)^2\\ &= \mathbb{E}(\eta_S(w_i,\lambda) - w_i)^2 + \mathbb{E}(w_i-\theta_i)^2 + 2\mathbb{E}(\eta_S(w_i,\lambda)-w_i)(w_i-\theta_i).\end{aligned} \tag{20}$$

We will consider each term separately:

- "Variance" term

$$\mathbb{E}(w_i-\theta_i)^2 = \operatorname{Var}[w_i] = 1. \tag{21}$$

- "Bias" term

$$\begin{aligned}\mathbb{E}\left(\eta_S(w_i,\lambda)-w_i\right)^2 &= \mathbb{E}\left(\operatorname{sgn}(w_i)\max(|w_i|-\lambda,0) - w_i\right)^2\\ &= \mathbb{E}(\max(|w_i|-\lambda,0) - |w_i|)^2\\ &= \mathbb{E}(\max(-\lambda, -|w_i|))^2\\ &= \mathbb{E}\min(|w_i|,\lambda)^2\\ &= \mathbb{E}\min(w_i^2,\lambda^2).\end{aligned} \tag{22}$$

- "Cross" term

$$\begin{aligned}\mathbb{E}(\eta_S(w_i,\lambda)-w_i)(w_i-\theta_i) &= \mathbb{E}\eta_S(w_i,\lambda)(w_i-\theta_i) - \mathbb{E}w_i(w_i-\theta_i)\\ &= \mathbb{E}\eta_S(w_i,\lambda)(w_i-\theta_i) - \mathbb{E}w_i^2 + \theta_i^2\\ &= \mathbb{E}\eta_S(w_i,\lambda)(w_i-\theta_i) - 1\\ &= \mathbb{E}\operatorname{sgn}(w_i)\max(|w_i|-\lambda,0)(w_i-\theta_i) - 1.\end{aligned} \tag{23}$$

By the law of total expectation, we write

$$\begin{aligned}\mathbb{E}\text{sgn}(w_i)\max(|w_i|-\lambda,0)(w_i-\theta_i) &= \mathbb{P}\{w_i\geq\lambda\}\mathbb{E}(w_i-\lambda)(w_i-\theta_i)\\ &\quad+\mathbb{P}\{w_i\leq-\lambda\}\mathbb{E}(w_i+\lambda)(w_i-\theta_i)\\ &= (\mathbb{E}w_i^2-(\lambda+\theta_i)\mathbb{E}w_i+\lambda\theta_i)\mathbb{P}\{w_i\geq\lambda\}\\ &\quad+(\mathbb{E}w_i^2+(\lambda-\theta_i)\mathbb{E}w_i-\lambda\theta_i)\mathbb{P}\{w_i\leq-\lambda\}\\ &= \mathbb{P}\{w_i\geq\lambda\}+\mathbb{P}\{w_i\leq-\lambda\}\\ &= \mathbb{P}\{|w_i|\geq\lambda\}.\end{aligned} \tag{24}$$

Combining everything together yields

$$\begin{aligned}\mathbb{E}(\eta_S(w_i,\lambda)-\theta_i)^2 &= 1+\mathbb{E}\min(w_i^2,\lambda^2)-2\mathbb{P}\{|w_i|<\lambda\}\\ &\leq \begin{cases}1+\mathbb{E}\lambda^2-2\mathbb{P}\{|w_i|<\lambda\}\\ 1+\mathbb{E}w_i^2-2\mathbb{P}\{|w_i|<\lambda\}\end{cases}\\ &= \begin{cases}1+\lambda^2-2\mathbb{P}\{|w_i|<\lambda\}\\ 1+(1+\theta_i^2)-2\mathbb{P}\{|w_i|<\lambda\},\end{cases}\\ &\leq \begin{cases}1+\lambda^2\\ \theta_i^2+2\mathbb{P}\{|w_i|\geq\lambda\},\end{cases}\end{aligned} \tag{25}$$

where the two cases come from the two inequalities derived from $\min(a,b)\leq\{a,b\}$. Now the first term is $1+2\log n\leq(1+2\log n)(1+\frac{1}{n}), n\geq 1$, and the second can be shown to satisfy (through calculations with the Gaussian distribution) $\theta_i^2+2\mathbb{P}\{|w_i|\geq\lambda\}\leq(2\log n+1)(\frac{1}{n}+\theta_i^2), n\geq 2$. Therefore, we have

$$\mathbb{E}(\hat{\theta}_i-\theta_i)^2\leq(2\log n+1)(\frac{1}{n}+\min(\theta_i^2,1)) \tag{26}$$

and summing all the coefficients together concludes the proof. □

Two immediate corollaries follow from the previous theorem:

- $\hat{\theta}$ when combined with the matrices $W^\top, W$ ($W^\top\circ\hat{\theta}\circ W$) yield an estimator, called ***VisuShrink***, for any function f that "almost" achieves the ideal wavelet performance (see (14)).
- The previous point together with equation (13) imply that $W^\top\circ\hat{\theta}\circ W$ is at most $\log^2 n$ times worse than the *oracular* piecewise polynomial estimator on the set up of (13). As we promised, a wavelet estimator manages to close the performance gap.

Even more remarkably, through a bias and variance decomposition proof, the authors show that ideal wavelet performance is close to ideal piecewise polynomial for *any* underlying function, and the first point above leads to the second, but now for any f. As promised, modulo a $\log^2 n$ factor, wavelets are universal spatially adaptive methods.

Finally, the paper considers further refinements of the above procedure to assess how tight the bounds are. In particular, there is a *min-max* optimal threshold, λ^*, for the thresholding that guarantees that no estimator can achieve better performance than what (19) essentially achieves. The proof of min-max optimality takes a probabilistic apparoach, by specifying a suitable prior distribution over the coefficients θ and concludes the optimality through its Bayes risk (similarly to what we saw in one of the recitations; although a bit more invloved of a proof). Hard thresholding is also studied, and is shown that asymptotically the performance is identical as the one presented here.

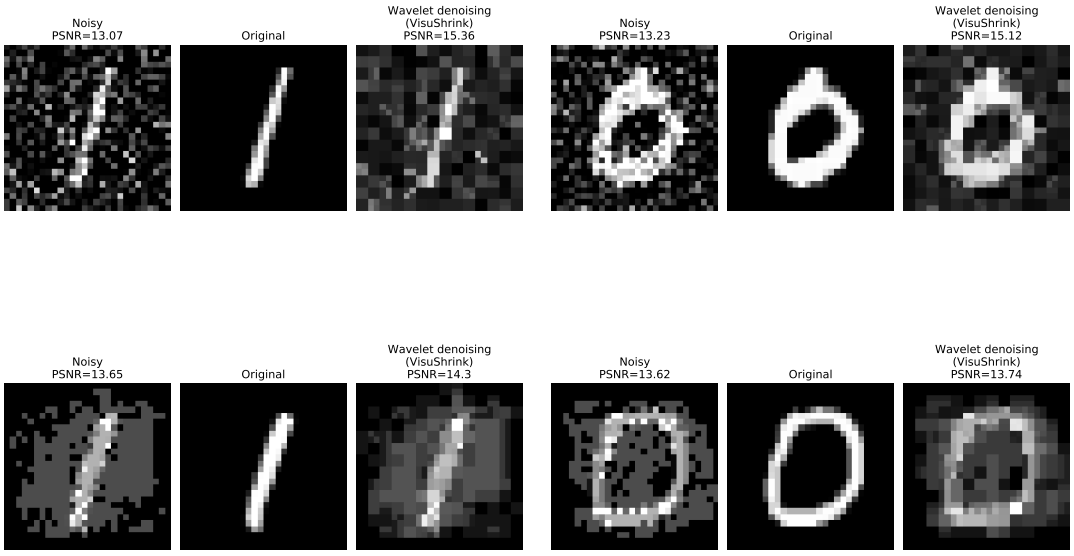

Figure 3: Handwritten digits denoising.

5 Bonus: Application in Image Denoising

While this paper (and the review) was focused on one-dimensional function estimation, wavelets have had tremendous success in applications in computer vision. In computer vision, we are usually concerned with two-dimensional images.

In this very small subsection, we evaluate the performance of VisuShrink in the task of denoising handwritten digits from the popular MNIST dataset. We checked the performance of the estimator both on Gaussian corrupted digits (first row of Fig. 3), and on adversarially (with respect to the function of an infinitely wide neural network) corrupted images with the same order of noise (second row). One interesting thing is that all noisy data have similar initial signal to ratio values, but wavelet denoising is more successful both in terms of visual appearance and snr for the Gaussian case.