Bernoulli **6**(4), 2000, 699–708

Can adaptive estimators for Fourier series be of interest to wavelets?

SAM EFROMOVICH

Department of Mathematics and Statistics, University of New Mexico, Albuquerque NM 87131, USA. E-mail: efrom@math.unm.edu

There is a firm belief in the literature on statistical applications of wavelets that adaptive procedures developed for Fourier series, labelled by that literature as 'linear', are inadmissible because they are created for estimation of smooth functions and cannot attain optimal rates of mean integrated squared error convergence whenever an underlying function is spatially inhomogeneous, for instance, when it contains spikes/jumps and smooth parts. I use the recent remarkable results by Hall, Kerkyacharian and Picard on block-thresholded wavelet estimation to present a counterexample to that belief.

Keywords: Efromovich–Pinsker estimator; filtering; small sample sizes; spatial adaptation

1. Adaptive estimation for Fourier and wavelet series

Consider a problem of data-driven wavelet series estimation of spatially inhomogeneous curves (which may have both smooth parts and jumps, spikes, etc.) under a minimax mean integrated squared error (MISE). Wavelets are relative newcomers to the world of orthogonal series, and, in contrast to classical Fourier series, they have no problem approximating jumps, spikes, etc.

The creation of wavelets sparked a heated discussion in the mathematical literature on wavelet transforms and Fourier transforms; see the illuminating paper by Strang (1993). Unsurprisingly, a similar discussion was sparked in the statistical wavelet literature with the emphasis on the possibility of using adaptive procedures developed for Fourier series. The judgement of the wavelet literature, which refers to these estimators as 'linear' adaptive, was unanimously negative; see, for instance, the discussion in Donoho and Johnstone (1994; 1995) and Donoho *et al.* (1995). The main theoretical argument of the wavelet literature is that adaptive Fourier estimators, with the Efromovich–Pinsker (EP) block shrinkage estimator being the main example, mimic linear pseudo-estimates (which are based both on data and an underlying class of function) that do not attain an optimal rate over classes of spatially inhomogeneous functions. 'Thus, adaptive linear methods cannot attain the optimal rate either', conclude Donoho and Johnstone (1995, p. 1200). Moreover, their numerical experiments show that using the EP estimator with a wavelet basis in place of the Fourier basis and with the block size specially chosen gives a disappointing performance.

Such a theoretical conclusion and the supporting numerical results look very reasonable (after all, the nature of wavelets and Fourier systems is absolutely different), and I know of

no article or book where they have been contested. On the other hand, why is the wavelet literature so sure that data-driven estimators developed for Fourier series mimic these inadmissible linear pseudo-estimates? Is this due to that label 'linear'? Furthermore, what happens if in the EP estimator one simply changes a Fourier basis into a wavelet basis (without any other modifications)? To answer these questions and shed light on the issue, let us recall the basics of the minimax theory of data-driven Fourier series estimation. (Note that in what follows it is essential to distinguish between the improvement in estimation of spatially inhomogeneous functions obtained by using a wavelet basis in place of a Fourier one, which is not at all the point of the paper, and the comparison explored between different adaptive procedures for wavelet series estimators.)

2. Minimax paradigm of data-driven Fourier series estimation

First of all, let us recall three classical steps in exploring a minimax data-driven Fourier estimation.

1. Finding a lower bound for the minimax (over a given function class $\mathscr{H}$) risk.
2. Finding an upper bound for the minimax risk which should asymptotically coincide with the lower bound. Typically a pseudo-estimate, based both on data and $\mathscr{H}$, is used to establish the upper bound.
3. Finding a data-driven estimate – based only on data – which attains the lower bound. The best-known data-driven estimators mimic oracles which are based on data and estimated curves. (Note that the use of the underlying curve instead of the function class $\mathscr{H}$ differentiates the oracles from the pseudo-estimates used in step 2.)

It was a matter of mathematical luck that for smooth functions the pseudo-estimates used in step 2 were linear, that is, for a filtering model – linear filters with coefficients depending on $\mathscr{H}$, for nonparametric regression – linear combinations of responses with coefficients depending on $\mathscr{H}$, etc. On the other hand, it was discovered in the 1980s that linear pseudo-estimates could not be rate-optimal whenever an underlying curve was not smooth; see the original article by Nemirovskii *et al*. (1985) and a comprehensive review in Härdle *et al.* (1998). This fact simply tells us that in step 2 other pseudo-estimates should be used to support the lower bound; nevertheless, it has become the main argument in the wavelet literature in favour of rejection of data-driven methods suggested for Fourier bases. To assess the validity of such an argument, let us recall the EP estimator, which is a sharp minimax estimator suggested for Fourier series during the 1980s.

3. Efromovich–Pinsker estimator

This section is based on Efromovich and Pinsker (1984). A square-integrable signal $f(t)$, $0 \leqslant t < 1$, is observed in a white Gaussian noise and an observation $Y_n(t)$ satisfies $\mathrm{d}Y_n(t) = f(t)\mathrm{d}t + \sigma n^{-1/2}\,\mathrm{d}w(t)$ (recall that results for the filtering model are typically valid

for other nonparametric models including equidistant nonparametric regression and density estimation where n is the sample size).

Let $\{\varphi_j, j = 1, 2 \ldots\}$ be a Fourier basis in $L_2[0,1]$. Then $f(t) = \sum_{j=1}^{\infty} \theta_j \varphi_j(t)$, where θ_j are Fourier coefficients. Thus the filtering problem is equivalent to estimation of the Fourier coefficients based on observations $Y_j = \theta_j + n^{-1/2}\sigma\xi_j$, where ξ_j are independent and identically distributed standard normal. Note that at this stage the underlying basis is irrelevant, so it may be a wavelet basis as well.

The EP estimator attempts to mimic the linear oracle $f^{**}(t) = \sum_{j=1}^{\infty} w_j^* Y_j \varphi_j(t)$, where $w_j^* = \theta_j^2/(\theta_j^2 + \sigma^2 n^{-1})$. Note that $\arg\min_c \mathrm{E}\{(cY_j - \theta_j)^2\} = w_j^*$, thus the linear oracle dominates all other oracles of the form $\tilde{f}(t) = \sum_{j=1}^{\infty} \lambda_j(f) Y_j \varphi_j(t)$ including all known threshold ones. This simple fact is interesting in itself because, at least in the world of oracles, thresholding is not optimal.

Unfortunately, it is impossible to mimic the linear oracle using a statistic because it is too intelligent (it knows all the θ_j^2), so a slightly less intelligent oracle should be used. Thus, let the blocks $\{T_{1n}, T_{2n}, \ldots\}$ be (possibly) a sequence of partitions of the indices $\{1, 2, \ldots\}$, and then consider the block-linear oracle

$$\hat{f}^*(t) = \sum_{k=1}^{S_n} [\Theta_k/(\Theta_k + \sigma^2 n^{-1})] \sum_{j \in T_{kn}} Y_j \varphi_j(t). \tag{1}$$

Here $\Theta_k = L_{kn}^{-1} \sum_{j \in T_{kn}} \theta_j^2$ and L_{kn} is the cardinality of T_{kn}. Apparently, this oracle dominates any other oracle of the form $\hat{f}(t) = \sum_{k=1}^{S_n} \lambda_k(f) \sum_{j \in T_{kn}} Y_j \varphi_j(t)$.

The block-linear oracle has several nice features. For smooth functions, under very mild assumptions on the choice of blocks and S_n, it mimics the linear oracle. In its turn, the block-linear oracle may be mimicked by a statistic (data-driven estimate) which is called the EP estimate,

$$\hat{f}(t) = \sum_{k=1}^{S_n} [\hat{\Theta}_k/(\hat{\Theta}_k + \sigma^2 n^{-1})] I(\hat{\Theta}_k > d_{kn}\sigma^2 n^{-1}) \sum_{j \in T_{kn}} Y_j \varphi_j(t). \tag{2}$$

Here $\hat{\Theta}_k = \max(L_{kn}^{-1} \sum_{j \in T_{kn}} (Y_j^2 - \sigma^2 n^{-1}), 0)$ is the positive part of the unbiased estimate of Θ_k (thus, it is an admissible estimate), and $I(\cdot)$ is the indicator. Note that the EP estimator shrinks the observed Y_j towards the origin by the product of two well-known factors: the ratio between powers of the input and output signals on the block frequencies, and the hard-threshold factor.

Statistical properties of the EP estimator which are important for our discussion are formulated in the following proposition. To shed light on assumptions, note that the integrated squared error (ISE) of the block-linear oracle (1) is

$$\int_0^1 (\hat{f}^*(t) - f(t))^2 \, \mathrm{d}t = n^{-1} \sum_{k=1}^{S_n} L_{kn} \Lambda_{kn} + \sum_{k > S_n} L_{kn} \Theta_{kn}, \tag{3}$$

where $\Lambda_{kn} = \Theta_{kn}/(\Theta_{kn} + \sigma^2 n^{-1})$. Write $a \wedge b$ for $\min(a, b)$.

Theorem 1. *Consider a non-constant square-integrable signal f. Let the threshold levels d_{kn} of the EP estimator* (2) *be bounded, $L_{kn}d_{kn}^3 > c_0 > 0$, and, for some $a < 1$,*

$$\sum_{k=1}^{S_n} \exp\{-L_{kn} \max(ad_{kn}(d_{kn} \wedge 1)/8, (ad_{kn} - 1)/2)\} < C^* < \infty. \tag{4}$$

Then for some finite constant C the MISE of the EP estimator is no larger than a constant times the MISE of the block-linear oracle (1)*, that is,*

$$\mathrm{E}\left\{\int_0^1 (\hat{f}(t) - f(t))^2 \,\mathrm{d}t\right\} \leqslant C\,\mathrm{E}\left\{\int_0^1 (\hat{f}^*(t) - f(t))^2 \,\mathrm{d}t\right\}. \tag{5}$$

Moreover, if the threshold levels decay – more precisely, if

$$\sum_{k=1}^{S_n} d_{kn}^{1/2} L_{kn}\Lambda_{kn} = o(1) \sum_{k=1}^{S_n} L_{kn}\Lambda_{kn} \tag{6}$$

and $\sum_{k=1}^{S_n} L_{kn}\Lambda_{kn} \to \infty$ as $n \to \infty$ – then the EP estimator sharply mimicks the block-linear oracle, that is,

$$E\left\{\int_0^1 (\hat{f}(t) - f(t))^2 \,\mathrm{d}t\right\} \leqslant (1 + o(1))\mathrm{E}\left\{\int_0^1 (\hat{f}^*(t) - f(t))^2 \,\mathrm{d}t\right\}. \tag{7}$$

This assertion was established in Efromovich and Pinsker (1984) for a more general model $Y_j = \theta_j + \sigma n^{-1/2}\varepsilon_j$ where the noise ε_j had only eight finite moments. In this case assumption (4) should be replaced by $\sum_{k=1}^{S_n} L_{kn}^{-1} d_{kn}^{-3} < C^*$. To establish (5) and (7) for the Gaussian model, use the exponential inequality

$$P(\hat{\Theta}_k > dn^{-1}, \Theta_k < cdn^{-1}) \leqslant$$

$$\min(3\exp\{-L_{kn}(1-c)d((1-c)d \wedge 1)/8\}, ((1-c)d_{kn})^{-1}\exp\{-L_{kn}((1-c)d_{kn} - 1)/2\}),$$

which holds for sufficiently small $c > 0$, in Lemma 3 of that paper in place of the Chebyshev inequality. Also note that the proposition holds for any orthonormal system of functions.

4. The Hall–Kerkyacharian–Picard block-threshold wavelet oracle

Let ϕ and ψ denote father and mother wavelet functions and let f have the wavelet expansion $f(t) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k}\phi_{j_0 k}(t) + \sum_{j \geqslant j_0} \sum_{k=0}^{2^j-1} \theta_{jk}\psi_{jk}(t)$, where $\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k)$, $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$ and α_{jk}, θ_{jk} are the corresponding wavelet coefficients whose natural estimates are $\hat{\alpha}_{jk} = \int_0^1 \phi_{jk}(t)\mathrm{d}Y_n(t)$ and $\hat{\theta}_{jk} = \int_0^1 \psi_{jk}(t)\mathrm{d}Y_n(t)$. Hall *et al*. (1995; 1998) made the remarkable discovery that a block-thresholded oracle is asymptotically minimax over a wide class of both smooth and spatially inhomogeneous functions including all practical and

test examples studied in the wavelet literature such as 'blocks', 'Doppler', 'jumpsine', piecewise smooth functions with increasing number of pieces, etc. The oracle is

$$\tilde{f}^*(t) = \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{J^*}\sum_{s=1}^{2^j/L_{jn}} I\left(L_{jn}^{-1}\sum_{k\in T_{jsn}}\theta_{jk}^2 > c^*\sigma^2 n^{-1}\right)\sum_{k\in T_{jsn}}\hat{\theta}_{jk}\psi_{jk}(t). \quad (8)$$

Here $2^{j_0-1} \leqslant n^{1/(2N+1)} \leqslant 2^{j_0}$ and N is the wavelet regularity; $T_{jsn} = \{k : L_{jn}(s-1) \leqslant k < L_{jn}s\}$, in which L_{jn} is the length of blocks on the jth resolution scale; $J^* = \lfloor \log_2 n \rfloor$; $\lfloor x \rfloor$ is integer part of x; and c^* is a positive constant. There is a wide variety of blocks which imply rate optimality, for instance, logarithmic blocks; see also Cay and Brown (1998). Furthermore, Cay (1999) established rate optimality of block thresholding over Besov spaces.

Theorem 1 implies the following corollary.

Corollary 1. *Consider a non-constant square-integrable function f. Let the threshold levels* d_{jn} *be bounded,* $L_{jn}d_{jn}^3 > c_0 > 0$*, and, for some* $a < 1$*,*

$$\sum_{j=j_0}^{J^*}(2^j/L_{jn})\exp\{-L_{jn}\max(ad_{jn}(d_{jn}\wedge 1)/8, (ad_{jn}-1)/2)\} < C^* < \infty. \quad (9)$$

Then the MISE of the wavelet EP estimator,

$$\tilde{f}(t) = \sum_{k=0}^{2^{j_0}-1}\hat{\alpha}_{j_0 k}\phi_{j_0 k}(t) + \sum_{j=j_0}^{J^*}\sum_{s=1}^{2^j/L_{jn}}\frac{\hat{\Theta}_{js}}{\hat{\Theta}_{js}+\sigma^2 n^{-1}} I\left(\hat{\Theta}_{js} > \frac{d_{jn}\sigma^2}{n}\right)\sum_{k\in T_{jsn}}\hat{\theta}_{jk}\psi_{jk}(t), \quad (10)$$

is no larger than a constant times the MISE of the oracle (8)*, that is,*

$$\mathrm{E}\left\{\int_0^1(\tilde{f}(t)-f(y))^2 dt\right\} \leqslant C\,\mathrm{E}\left\{\int_0^1(\tilde{f}^*(t)-f(t))^2\right\}. \quad (11)$$

For instance, (9) holds if $d_{jn} = d > 3$ is a constant and $L_{jn} = \lfloor \ln(n) \rfloor$. Thus, if a block-threshold oracle is spatially adaptive (rate minimax) then the wavelet EP estimator is spatially adaptive as well. If σ^2 is unknown then a traditional robust median-scale estimate, based on $\{\hat{\theta}_{J^*k}\}$, is used. Furthermore, as in Cay (1999), the estimator is pointwise rate-optimal. Also, as in (6)–(7), sharp mimicking is also possible.

5. Numerical study

Donoho and Johnstone (1995, p. 1216) correctly conclude that using the 'linear' adaptive estimators LPJS and WaveJS is disappointing. Bear in mind that the EP estimator (2) is a prototype of these estimators and that they are similar to (8) with some specific parameters. In short, numerical examples show that these estimators perform essentially worse than SureShrink, which is the benchmark of the term-by-term threshold-adaptive estimates.

I will explain below why these two 'linear' estimators perform so badly, but first let us look at two particular cases of the EP estimator which illustrate its robustness. Let us refer to the first one as the EP estimator with 'increasing L and $d = 0$' because it has block sizes $L_{jn} = b_n 2^{\lfloor (j-j_0)/3 \rfloor}$ and zero threshold levels $d_{jn} = d = 0$; here b_n is the largest power of 2 (2^k, where $k = 1, 2, \ldots$) which is at most $\log_2(n)$. Note that this EP estimator has no thresholding factor. Let us refer to the second estimator as the EP estimator with 'constant L and $d = 5$' because it has $L_{jn} = b_n$ and $d_{jn} = 5$. There is nothing special in these particular parameters, I simply want to show a broad spectrum of the possibilities.

These EP estimators are compared with the default SureShrink estimator supported by the toolkit S+WAVELETS^{TM}. This toolkit allows one to consider equidistant nonparametric regression problems, thus I consider such a model with n observations, j_0 being the default one (six finest scales are considered by the default), and the default Symmlet 8 wavelet is used. Two particular underlying regression functions are the classical 'Doppler' and 'jumpsine' of that toolkit. Figures 1 and 2 show particular cases of estimation of these two functions for $n = 1024$ and signal-to-noise ratio (snr) equal to 3. The EP estimates appear to perform exceptionally well in comparison with SureShrink.

But how stable are these properties if one repeats these simulations over and over? To answer this question, let us define an experiment as a combination of an underlying function, sample size n and snr. Consider our two functions, n being 512, 1024 and 2048, and snr being 3, 5 and 7. Suppose that for every experiment: (i) we carry out 100 independent Monte Carlo simulations; (ii) we calculate the sample mean (over these 100 simulations) of the ratio of the ISE of each of our EP estimators in turn to the ISE of SureShrink; (iii) we calculate the sample mean (over these 100 simulations) of the ratio of the number of non-zero wavelet coefficients used by each of our EP estimators in turn to the number of non-zero wavelet coefficients used by SureShrink.

Step (ii) allows us to compare the estimates in terms of ISE, whereas step (iii) allows us to compare the data-compression properties of the estimates. Note that if the ratio is smaller than 1 then the EP estimator is better than SureShrink, and vice versa if greater than 1. The results are presented in Table 1.

These results confirm the preliminary conclusion, made by inspection of particular graphs, that the EP estimator may be of interest to wavelets. Also note that the first EP estimator (the one with no thresholding) performs exceptionally well in terms of ISE but slightly worse than both the second EP estimator and SureShrink in terms of data compression (the latter is not a surprise). On the other hand, the second EP estimator is comparable with SureShrink in terms of ISE and yields a much better data compression.

Recall that the only goal of this numerical study is to show that data-driven estimators developed for Fourier series may be of interest to wavelets. The results presented clearly indicate that this is the case.

Finally, I promised to explain why the 'linear' adaptive estimators suggested by Donoho and Johnstone perform so badly. The main reason is that they use $L_{jn} = 2^j$ (a resolution scale is considered as a block) together with $d_{jn} = 0$ (no thresholding factor). As a result, all wavelet coefficients from a resolution scale are shrunk by the same factor, and thus the disappointing outcome is straightforward. In short, while there is a wide choice of reasonable parameters for the EP estimator, extremes should be avoided.

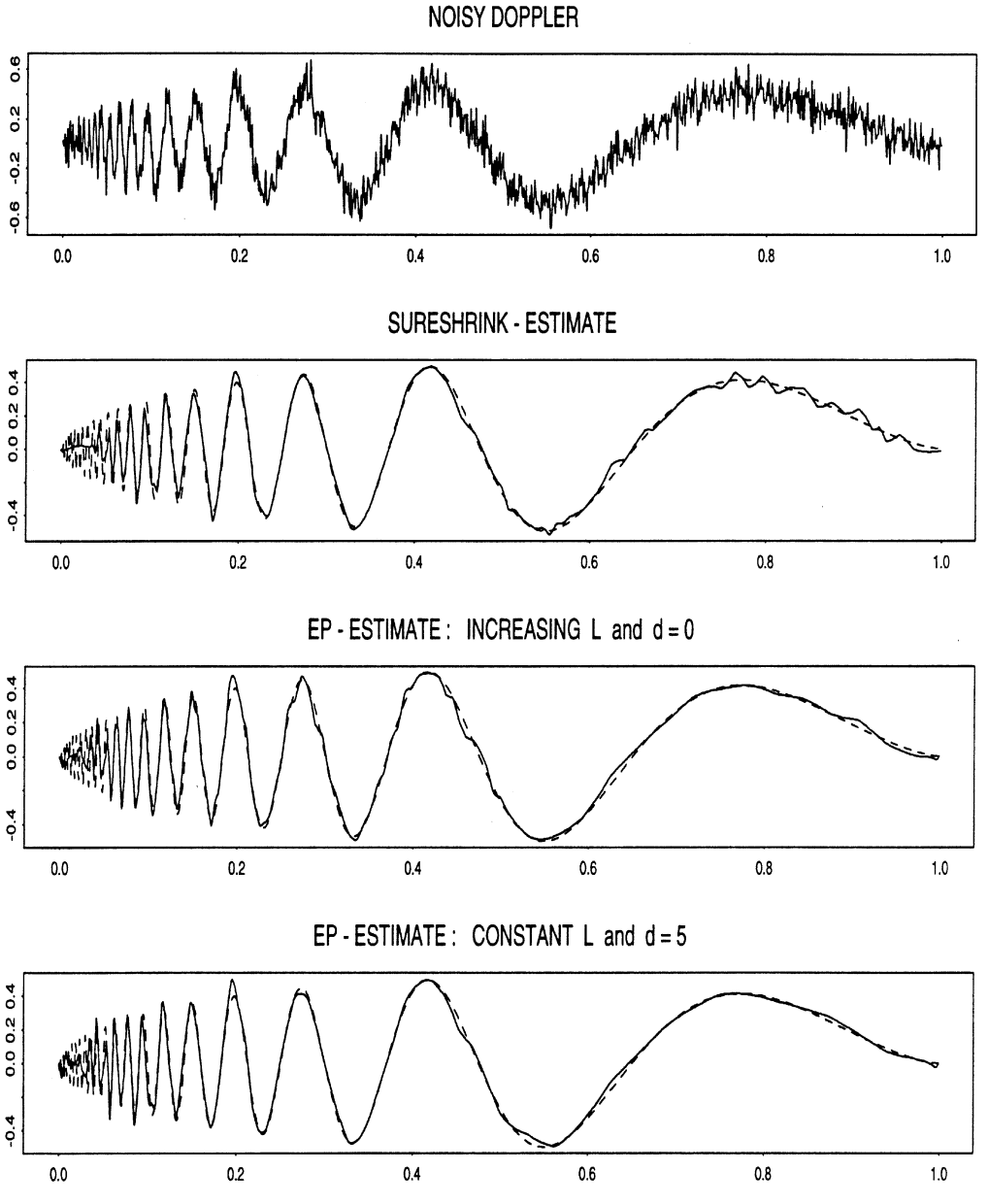

Figure 1. Estimation of noisy Doppler signal ($n = 1024$, signal-to-noise ratio 3) by three estimates shown by solid lines. The dashed line shows the underlying signal.

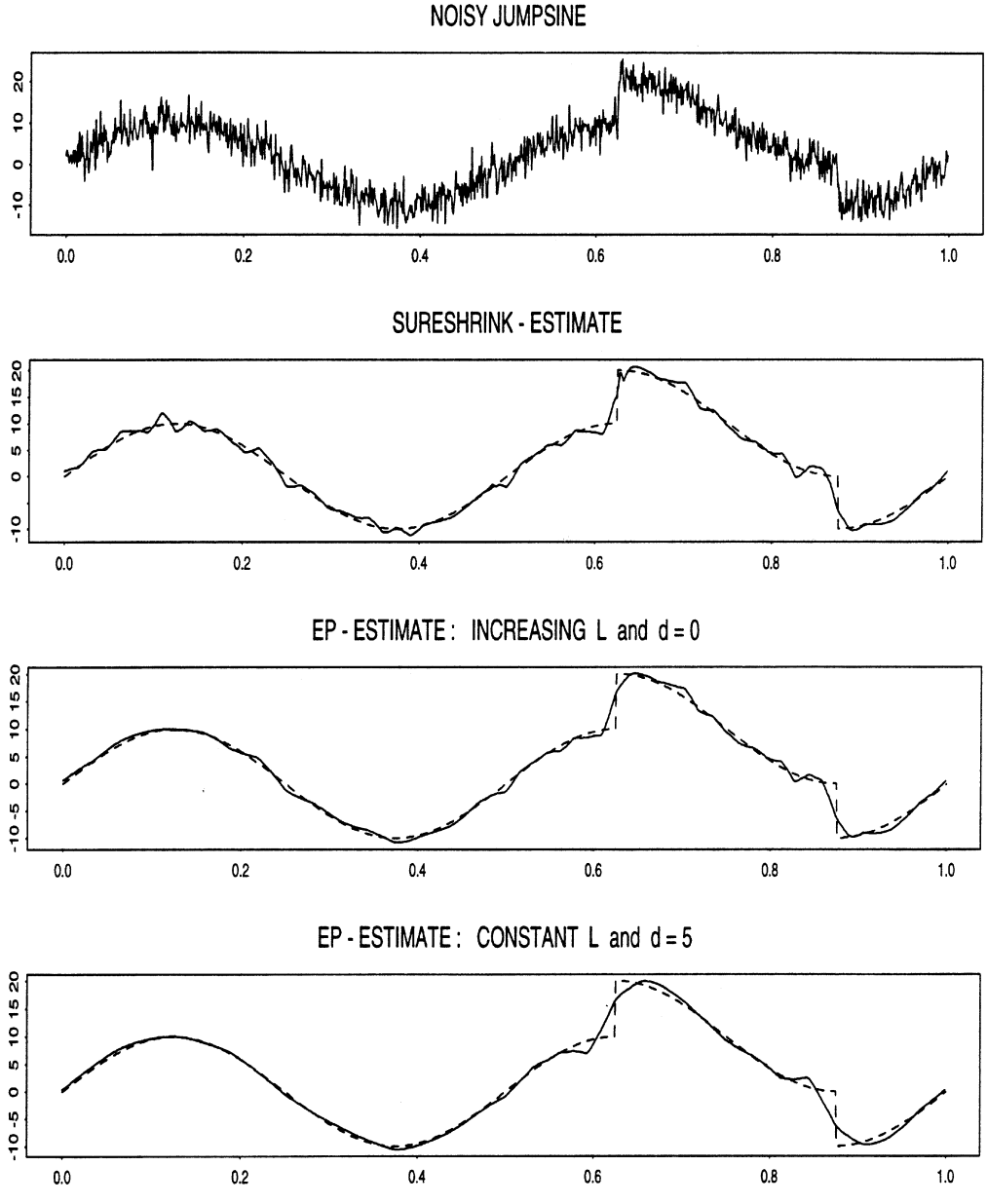

Figure 2. Estimation of noisy jumpsine signal ($n = 1024$, signal-to-noise ratio 3) by three estimates shown by solid lines. The dashed line shows the underlying signal.

Table 1. Sample means of ratios between an EP estimate and SureShrink for the case ISE and the case of data compression. Each element in the body of the table has two entries: the top entry is the sample mean of ratios for the EP estimator with increasing L and $d = 0$, and the bottom entry is the sample mean of ratios for the EP estimator with constant L and $d = 5$.

n	512			1024			2048		
snr	3	5	7	3	5	7	3	5	7
ISE									
'Doppler'	0.84	0.55	0.47	0.57	0.45	0.46	0.43	0.46	0.48
	1.03	0.56	0.47	0.77	0.47	0.51	0.53	0.50	0.53
'jumpsine'	0.90	0.96	1.11	0.82	0.87	1.02	0.76	0.87	0.95
	1.03	1.20	1.36	1.11	1.13	1.31	0.87	1.06	1.07
Data compression									
'Doppler'	1.43	1.75	1.35	1.14	1.04	0.90	0.75	0.80	0.70
	0.8	1.08	0.92	0.73	0.79	0.64	0.62	0.57	0.49
'jumpsine'	1.41	1.49	1.95	0.87	1.21	1.40	0.76	1.01	1.12
	0.61	0.77	0.88	0.42	0.62	0.75	0.38	0.51	0.74

6. Conclusion

I have shown that known adaptive procedures developed for Fourier series may be of interest to wavelet series as well. In particular, changing a Fourier basis into a wavelet basis in the EP estimator implies rate-optimal estimation of spatially inhomogeneous functions. Moreover, for the case of reasonable sample sizes these estimators may compete with classical wavelet estimators in terms of visual appeal, approximation and data compression.

References

Cay, T.T. and Brown, L. (1998) Wavelet shrinkage for nonequispaced samples. *Ann. Statist.*, **26**, 1783–1799.

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

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

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

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B*, **57**, 301–369.

Efromovich, S. and Pinsker, M. (1984) A learning algorithm for nonparametric filtering. *Automat. Remote Control*, **10**, 1434–1440.

Hall, P., Kerkyacharian, G. and Picard, D. (1995) On the minimax optimality of block thresholded wavelet estimators. Technical report, Australian National University, Canberra.

Hall, P., Kerkyacharian, G. and Picard, D. (1998) Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, **26**, 922–942.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998) *Wavelets, Approximation, and Statistical Applications*, Lecture Notes in Statist. 129. New York: Springer-Verlag.

Nemirovskii, A., Polyak, B. and Tsybakov, A. (1985) Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problems Inform. Transmission*, **21**, 258–272.

Strang, G. (1993) Wavelet transforms versus Fourier transforms. *Bull. Amer. Math. Soc. N.S.*, **28**, 288–305.

Received January 1997 and revised May 1999