Postprint of: Niedźwiecki M., Easy recipes for cooperative smoothing, Automatica, Vol. 46, Iss. 4 (2010), pp. 716-720,
DOI: 10.1016/j.automatica.2010.01.030

Easy Recipes for Cooperative Smoothing

Maciej Niedźwiecki [a]

[a] *Faculty of Electronics, Telecommunications and Computer Science, Department of Automatic Control*
Gdańsk University of Technology, ul. Narutowicza 11/12, Gdańsk , Poland
Tel: +48 58 3472519, Fax: +48 58 3415821, e-mail: maciekn@eti.pg.gda.pl

Abstract

In this paper we suggest how several competing signal smoothers, differing in design parameters, or even in design principles, can be combined together to yield a better and more reliable smoothing algorithm. The proposed heuristic, but statistically well motivated, fusion mechanism allows one to combine practically all kinds of smoothers, from simple local averaging or order statistic filters, to parametric smoothers designed for different hypothetical signal and/or noise models. It also allows one to account for the distribution of measurement noise, and in particular – to cope with heavy-tailed disturbances, such as Laplacian noise, or light-tailed disturbances, such as uniform noise.

Key words: adaptive signal processing, signal smoothing

1 Introduction

Consider the problem of noncausal estimation of the signal $s(i)$ based on its noisy measurements $y(i)$:

$$y(i) = s(i) + v(i), \quad i = \ldots, -1, 0, 1, \ldots \tag{1}$$

where i denotes normalized time and $\{v(i)\}$ is the sequence of independent, identically distributed (i.i.d.) random variables, representing additive measurement noise. To simplify our further considerations, we will assume that an infinite observation history is available $\mathcal{Y} = \{y(i), i \in (-\infty, \infty)\}$. Note that for a given time instant t, $\mathcal{Y}$ can be decomposed into the set of "past" measurements $\mathcal{Y}_-(t) = \{y(i), i < t\}$, "present" measurement $y(t)$, and the set of "future" measurements $\mathcal{Y}_+(t) = \{y(i), i > t\}$:

$$\mathcal{Y} = \{\mathcal{Y}_-(t), y(t), \mathcal{Y}_+(t)\}.$$

Any estimate $\widehat{s}(t) = f[\mathcal{Y}]$, that relies on both "past", "present" and "future" measurements, is called the smoothed estimate. Since estimation accuracy of noncausal estimation schemes that incorporate smoothing exceeds accuracy of their causal counterparts, smoothing is used in many off-line signal processing applications, where the analyzed signals are prerecorded, rather than acquired sequentially in a sample-by-sample manner.
Our objective will be to find the estimate $\widehat{s}(t)$ that minimizes the mean-squared error

$$\mathrm{E}\left\{[s(t) - \widehat{s}(t)]^2\right\} \to \min . \tag{2}$$

Rather than proposing a new smoothing paradigm, in this paper we will suggest how several competing smoothers, differing in design parameters, or even in design principles, can be combined together yielding a better and more reliable smoothing algorithm.

2 Bayesian Pattern Matching

To work out a rational fusion mechanism, that could be used to combine different smoothers, we will start from solving a simpler problem, further referred to as pattern matching.
Denote by $T(t) = [t - m, t + m]$ the local evaluation frame, centered at t and covering $M = 2m + 1$ time instants. We will assume that, for all $i \in T(t)$, the true signal coincides with one of K signal patterns, further denoted by $s_k(i), k = 1, \ldots, K$. The hypotheses

$$H_k : s(i) = s_k(i), \quad i \in T(t) \tag{3}$$

will be regarded as equiprobable

$$\pi(H_k) = \frac{1}{K} , \quad k = 1, \ldots, K . \tag{4}$$

We will assume that measurement noise is distributed according to the generalized Gaussian law (Saralees, 2005)

$$v \sim \mathcal{GN}(\mu, \alpha, \beta) :$$
$$p(v; \mu, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{|v - \mu|}{\alpha}\right)^{\beta}\right\} \tag{5}$$

where μ is the location parameter, $\alpha > 0$ is the scale parameter, $\beta > 0$ is the shape parameter, and $\Gamma(x) = \int_0^\infty e^{-z} z^{x-1} dz$, for $x > 0$, denotes the Euler's gamma function (extension of the factorial function).
Generalized Gaussian is a parametric family of symmetric distributions that includes normal distribution when $\beta = 2$ (with mean μ and variance $\alpha^2/2$), and Laplace distribution when $\beta = 1$ (with mean μ and variance $2\alpha^2$). When $\beta \to \infty$, the density (5) converges pointwise to a uniform density on $(\mu - \alpha, \mu + \alpha)$.
We will assume that $\mu = 0$ (zero-mean measurement noise), and that $\beta \geq 1$ is a predetermined (user-defined) shape parameter. We will *not* assume that the scale parameter $\alpha > 0$ is known – in our Bayesian analysis α will be treated as a nuisance parameter with assigned noninformative (improper) prior distribution

$$\pi(\alpha|H_k) = \pi(\alpha) \propto \frac{1}{\alpha} \tag{6}$$

where $\propto$ denotes proportionality.
Under the assumptions made, the optimal, in the mean-squared sense, approximation of $s(t)$ can be obtained in the form (Lewis, 1986)

$$\widehat{s}(i) = \sum_{k=1}^{K} \mu_k(t) s_k(i) \ , \quad i \in T(t) \tag{7}$$

where

$$\mu_k(t) = P(H_k|\mathcal{Y}_T(t)) = \frac{\int_0^\infty p(\mathcal{Y}_T(t), \alpha, H_k) d\alpha}{p(\mathcal{Y}_T(t))} \tag{8}$$

denote posterior probabilities of signal patterns $s_k(\cdot)$, given the data $\mathcal{Y}_T(t) = \{y(i), i \in T(t)\}$. Straightforward calculations lead to

$$\mu_k(t) = \frac{\varphi_k(t)}{\sum_{k=1}^K \varphi_k(t)} \tag{9}$$

$$\varphi_k(t) \propto \int_0^\infty p(\mathcal{Y}_T(t)|\alpha, H_k) \pi(\alpha|H_k) \pi(H_k) d\alpha \ . \tag{10}$$

2.1 Posterior Probabilities
Note that

$$\begin{aligned} & p(\mathcal{Y}_T(t)|\alpha, H_k) \\ & = \prod_{i \in T(t)} \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{ - \left(\frac{|\varepsilon_k(i)|}{\alpha} \right)^\beta \right\} \\ & = \left[\frac{\beta}{2\alpha\Gamma(1/\beta)} \right]^M \exp\left\{ - \frac{\sum_{i \in T(t)} |\varepsilon_k(i)|^\beta}{\alpha^\beta} \right\} \end{aligned} \tag{11}$$

where $\varepsilon_k(i) = y(i) - s_k(i)$.

Combining (11) with (4)–(6), and carrying out integration in (10), one arrives at (see Appendix)

$$\varphi_k(t) = \left[\sum_{i \in T(t)} |\varepsilon_k(i)|^\beta \right]^{-M/\beta} \ . \tag{12}$$

Note that when $\beta \to \infty$ (the uniform noise case), it holds that $\varphi_k(t) \to \left[\ \max_{i \in T(t)} |\varepsilon_k(i)| \ \right]^{-M}$.
2.2 Sliding Analysis Window
In the method described above, posterior probabilities $\mu_k(t)$ are used to approximate the signal in the entire evaluation frame $T(t)$. To gain more flexibility, one can repeat the pattern matching procedure for consecutive values of t, generating a sequence of point estimates

$$\widehat{s}(t) = \sum_{k=1}^{K} \mu_k(t) s_k(t) \ , \quad \forall t \tag{13}$$

instead of a single interval estimate (7). Sliding window approach is computationally more involved, but yields better approximations than the interval approach.

3 Cooperative Smoothing

Bayesian pattern matching is a good starting point for derivation of a more realistic smoothing procedure, further referred to as cooperative smoothing. The key elements of this new approach are: data-dependent patterns and cross-validatory pattern assessment.
3.1 Data-Dependent Patterns
The Bayesian pattern matching approach, described in the previous section, is certainly too rigid to be practically useful – unless some prior knowledge about $s(t)$ is available, the number of fixed patterns that should be used to obtain satisfactory approximation results becomes impractically large. This problem can be alleviated by considering patterns that are locally adapted to the signal. Such data-dependent patterns, generated by local smoothing procedures, will be further denoted by $\widehat{s}_k(t)$.
As an example of a local pattern generation technique, consider a nonparametric estimation method known as kernel regression (Härdle, 1990). The kernel smoother is defined as

$$\widehat{s}(t) = \frac{\sum_i y(i) K_\eta(t - i)}{\sum_i K_\eta(t - i)} \tag{14}$$

where $K_\eta(z) = 1/\eta K(z/\eta)$ denotes the kernel function, η denotes the kernel bandwidth and summation extends to all available samples. The kernel is a nonnegative, continuous, bounded and symmetric function, such that $\int K(z) dz = 1$.
When the kernel nonnegativity constraint is removed, many *universal smoothers*, i.e., those requiring no, or very little, prior knowledge about the approximated signal [such as local polynomial smoothers, known also as Savitzky-Golay smoothing filters (Orfanidis, 1996)], can be expressed in the form (14).

3.2 Selection of Smoothing Bandwidth

All smoothing methods are equipped with design parameters that allow one to to control degree of smoothing, usually referred to as smoothing bandwidth. In kernel regression, degree of smoothing depends on the kernel bandwidth η.

Selection of smoothing bandwidth is certainly one of the key problems in statistical approximation theory. From the statistical viewpoint, selection of bandwidth parameters is a tradeoff between the estimation variance, which decreases with growing bandwidth, and estimation bias, which increases with growing bandwidth. The problem is usually solved by considering estimates yielded by several smoothers $\widehat{s}_k(t), k = 1, \ldots, K$, of the same kind, but with different bandwidth settings, and selecting the best fitting candidate, or combining all candidates appropriately. Automatic bandwidth selectors, which constitute the core of such schemes, are usually based on some statistical principles, such as Akaike's information criterion (Cleveland & Loader, 1996), (Hurvich & Simonoff, 1998), wavelet shrinkage (Donoho & Johnstone, 1995), intersection of confidence intervals rule (Katkovnik, 1999), or adaptive regression by mixing approach (Yang, 2000). However, all existing methods of automatic bandwidth tuning are subject to at least one of the following limitations: 1. They result in competitive, rather than cooperative, smoothing schemes (winner-take-all strategy). 2. Their application is limited to a specific family of smoothers they were derived for. 3. They cannot account for non-Gaussian noise distribution. 4. They are computationally very intense. In contrast with this, the fusion mechanism proposed below allows one to combine practically all kinds of smoothers, from simple local averaging or order statistic filters, to parametric smoothers designed for different hypothetical signal and/or noise models. It is computationally simple and it allows one to account for the distribution of measurement noise.

3.3 Preliminary Considerations

One may argue that for large smoothing bandwidths, signal-adapted patterns are weakly correlated with measurement noise, allowing one to use the same fusion mechanism, which was developed in Section 2, namely

$$\widehat{s}(t) = \sum_{k=1}^{K} \mu_k(t) \widehat{s}_k(t) \ , \quad \forall t \tag{15}$$

where the weights $\mu_k(t)$, further called *credibility coefficients*, are evaluated according to (12), and residual errors are given by $\varepsilon_k(t) = y(t) - \widehat{s}_k(t) \ , \forall t$. However, for small bandwidths, smoothers cannot be evaluated based on the values of the corresponding residual errors. Note that after setting $\eta \to 0$ in (14), one obtains $\widehat{s}(t) \to y(t)$, i.e., $\varepsilon(t) \to 0, \forall t$. This means that the smallest-bandwidth smoother, which completely ignores the presence of measurement noise, will always obtain the highest score, no matter what metric is used to quantify residual errors. To circumvent this problem, we will define credibility coefficients in a slightly different way. Denote by $\widehat{s}_k^{\circ}(t)$ the *holey smoother* associated with $\widehat{s}_k(t)$, i.e., smoother that excludes $y(t)$ from the set of measurements used for estimation of $s(t)$

$$\widehat{s}_k^{\circ}(t) = f\left[\mathcal{Y}^{\circ}(t)\right], \quad \mathcal{Y}^{\circ}(t) = \{\mathcal{Y}_-(t-1), \mathcal{Y}_+(t+1)\}.$$

In case of kernel smoothers, one should simply redefine the kernel function: $K_\eta^{\circ}(0) = 0, K_\eta^{\circ}(z) = K_\eta(z), \forall z \neq 0$. A very important property of every holey smoother is its pointwise independence of measurement noise

$$p(\widehat{s}_k^{\circ}(t)|v(t)) = p(\widehat{s}_k^{\circ}(t)) \ , \quad \forall t \ . \tag{16}$$

Owing to this property, the modified Bayesian-like combination rule, obtained when credibility coefficients are evaluated for *matching* errors

$$\varepsilon_k^{\circ}(t) = y(t) - \widehat{s}_k^{\circ}(t) \ , \quad \forall t$$

will not favor smoothers that "underestimate" the influence of measurement noise on the observed data.

Note that for the kernel smoother (14) it holds that $\varepsilon^{\circ}(t) = \delta\varepsilon(t)$, where $\varepsilon(t) = y(t) - \widehat{s}(t)$ denotes residual error and

$$\delta = \frac{\sum_i K_\eta(i)}{\sum_i K_\eta^{\circ}(i)} = \left[1 - \frac{K_\eta(0)}{\sum_i K_\eta(i)}\right]^{-1} > 1$$

is the penalty factor which grows with decreasing kernel bandwidth.

3.4 Proposed Smoothing Formula

The proposed adaptive cooperative smoothing formula, allowing one to combine results yielded by K competing smoothers $\widehat{s}_k(t)$, $k = 1, \ldots, K$, has the form

$$\widehat{s}(t) = \sum_{k=1}^{K} \mu_k^{\circ}(t) \widehat{s}_k(t) \ , \quad \forall t \tag{17}$$

where

$$\mu_k^{\circ}(t) = \frac{\varphi_k^{\circ}(t)}{\sum_{k=1}^{K} \varphi_k^{\circ}(t)} \ , \quad k = 1, \ldots, K \tag{18}$$

and the quantities $\varphi_k^{\circ}(t)$ should be evaluated according to

$$\varphi_k^{\circ}(t) = \left[\sum_{i \in T(t)} |\varepsilon_k^{\circ}(i)|^{\beta}\right]^{-M/\beta} . \tag{19}$$

For large values of M, the weighted estimation formula (17) *de facto* reduces itself to

$$\widehat{s}(i) = \widehat{s}_{k^*(t)}(t) \ , \quad i \in T(t) \tag{20}$$

where

$$k^*(t) = \arg \max_{1 \le k \le K} \mu_k^\circ(t) \ .$$

This is because for large evaluation frames even small differences in the matching error statistics produce large differences in the values of the corresponding credibility coefficients. Consequently, the major contribution to $\widehat{s}(t)$ in (7) is due to the "locally the best" smoother $\widehat{s}_{k^*(t)}(t)$. When $\beta = 2$, maximization of $\mu_k^\circ(t)$ is equivalent to minimization of $\sum_{i \in T(t)} [\varepsilon_k^\circ(i)]^2$. This can be regarded as a time-localized variant of the leave-one-out cross-validation (CV) approach, introduced by (Stone, 1974) and further developed by many authors – for more details see e.g. (Friedl & Stampfer, 2002). Our Bayesian framework is a natural way of bringing the notion of model credibility into cross-validatory analysis.

The asymptotic properties of CV-based selectors are well understood (Droge, 2006). When $M \to \infty$, cross-validation does not guarantee statistically consistent model/smoother selection (when used with local polynomial or kernel estimators it tends to undersmooth), but it is asymptotically optimal in the sense of Shibata (Shibata, 1981). Of course, none of these asymptotic statements is justifiable when M is small, which, unfortunately, is the only case that has practical relevance (to guarantee "alertness" of combination smoothers to changing estimation conditions, we recommend to use $20 \le M \le 50$).

4 Simulation results

Due to space limitations, we will present results of only one, albeit carefully designed, simulation experiment. The four test signals used in this experiment, called *Blocks*, *Bumps*, *HeaviSine* and *Doppler*, respectively (see Fig. 1), were proposed by Donoho and Johnstone in their seminal paper on wavelet-based denoising (Donoho & Johnstone, 1995). Since then they are commonly used for benchmarking different smoothing techniques. The popularity of this particular set of test signals is due to the fact that it was designed to represent various spatially inhomogenous phenomena, which make smoothing difficult, and which are encountered in many real-world signals in such areas as telecommunications, geophysics and biomedicine. Test signals, each containing 2048 samples, were extended by zeros at both ends (to avoid boundary problems) and corrupted with either Gaussian ($\beta = 2$) or Laplacian ($\beta = 1$) white noise with intensity $\sigma_v^2 = 1$, see Figs. 2 and 4, respectively. The average signal-to-noise ratio was in all cases the same and equal to 16.9 dB.

The bank of competing smoothers consisted of 5 local averaging filters $\widehat{s}(t) = \sum_{i=-n}^{n} y(t+i)/(2n+1)$ and 5 median filters [1] $\widehat{s}(t) = \text{med}\{y(t-n), \ldots, y(t+n)\}$, with

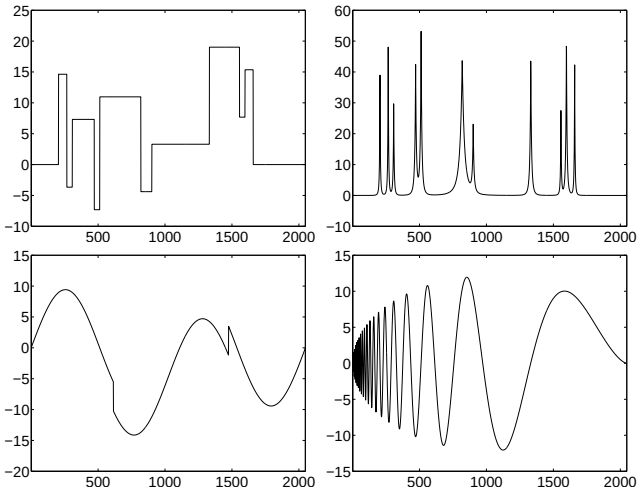

Fig. 1. Test signals: *Blocks* (top left), *Bumps* (top right), *HeaviSine* (bottom left), and *Doppler* (bottom right).

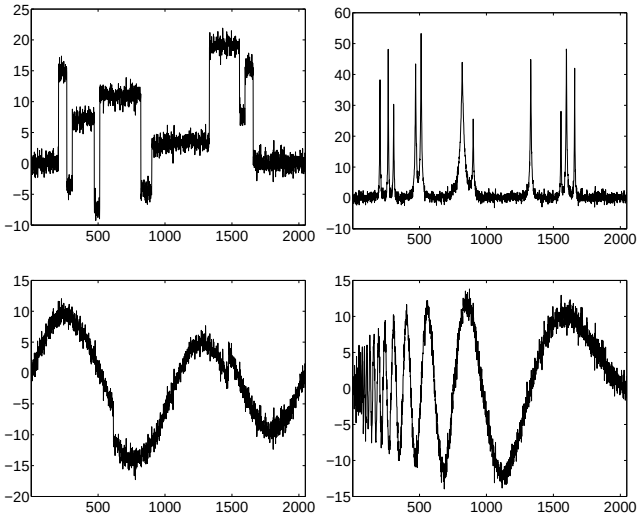

Fig. 2. Test signals with white Gaussian noise, $\sigma_v^2 = 1$, SNR=16.9 dB (SD(s)/σ_v=7).

fitting frames of lenghts $N_k = 2n_k + 1, k = 1, \ldots, 5$, forming (approximately) a geometric progression: $N_1 = 5$, $N_2 = 11$, $N_3 = 23$, $N_4 = 47$, $N_4 = 95$.

Three combination smoothers were considered: the naive smoother (15), further denoted by C_1, the competitive smoother (20), denoted by C_2, and the cooperative smoother (17), denoted by C_3. The width M of the evaluation frame was set equal 31 ($m = 15$).

The SNR scores, obtained for the Gaussian noise and Laplacian noise, are shown in Tabs. 1 and 2, respectively. All numbers were obtained by ensemble averaging over 100 realizations of $\{v(t)\}$. Typical results of smoothing are shown in Figs. 3 and 5. Note the very good performance of the cooperative (recommended) smoother C_3, on all occasions better than performance of the component smoothers. The competitive smoother,

[1] med$\{\cdot\}$ denotes the central value of the sequence obtained by ordering the original sequence: med$\{z_1, \ldots, z_j\}$ is defined as $\widetilde{z}_{(j+1)/2}$ for odd values of j, and $(\widetilde{z}_{j/2} + \widetilde{z}_{j/2+1})/2$ for even values of j, where $\widetilde{z}_i$ is the ith smallest sample among $\{z_1, \ldots, z_j\}$.

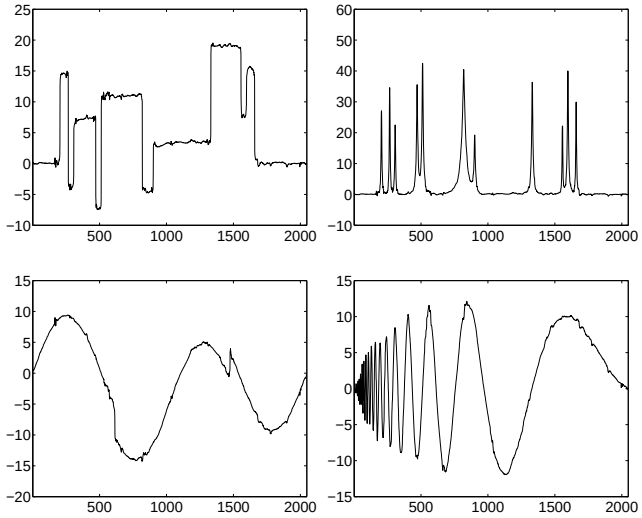

Fig. 3. Denoised test signals (Gaussian noise, C_3).

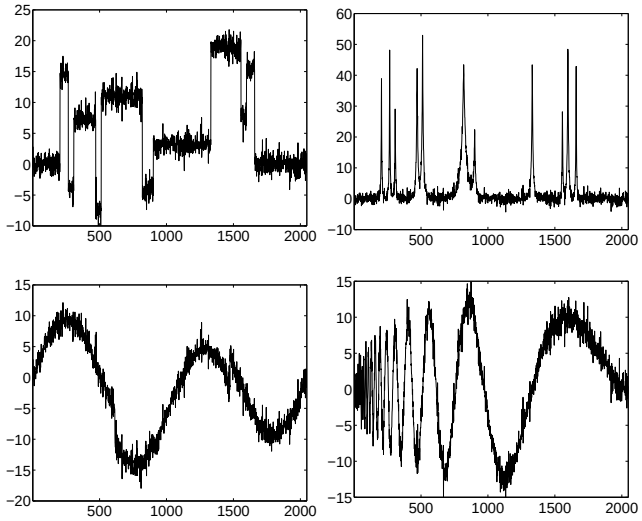

Fig. 4. Test signals with white Laplacian noise, $\sigma_v^2 = 1$, SNR=16.9 dB (SD(s)/σ_v=7).

which can be regarded as a simplified version of the cooperative smoother, yields consistently worse results than C_3, which shows clearly the benefits of weighting. Finally, the naive smoother C_1, incorporating unmodified credibility coefficients $\mu_k(t)$, performs considerably worse than C_2 and C_3. This demonstrates the importance of the pointwise independence condition enforced in the modified scheme.

Combination of linear (averaging) and nonlinear (median) smoothers allows one to obtain estimation procedure that can be used to smooth discontinuous signals, i.e., procedure that attenuates measurement noise but, at the same time, does not distort step-like signal features. As a matter of fact, when applied to *Blocks*, *HeaviSine* and *Doppler*, our simple *ad hoc* smoother (which was not optimized in any way) is doing a remarkably good job as it outperforms the state-of-the-art wavelet

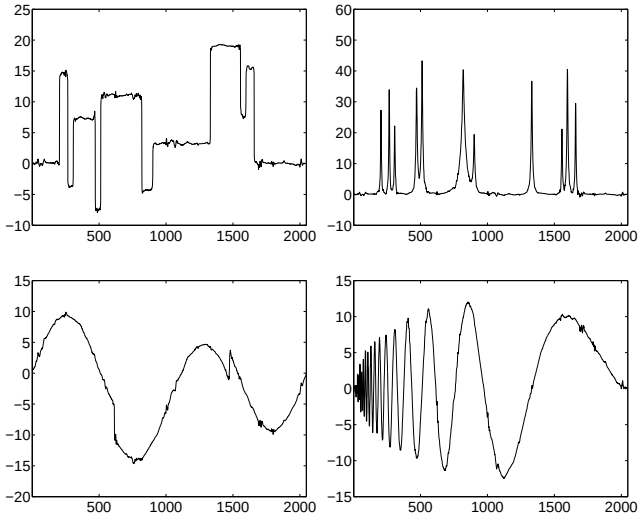

Fig. 5. Denoised test signals (Laplacian noise, C_3).

shrinkage procedure *SureShrink* proposed in (Donoho & Johnstone, 1995). For Gaussian noise the corresponding SNR scores achieved by *SureShrink* are equal to 24.6 dB (*Blocks*), 26.0 dB (*HeaviSine*) and 21.2 dB (*Doppler*) for the family of Haar wavelets, and 22.8 dB, 28.6 dB and 23.8 dB, respectively, for Daubechies D4 wavelets – see Table 2 in (Donoho & Johnstone, 1995). The low score achieved by the proposed smoother for *Bumps* is caused by insufficient spatial resolution of component filters – when the smoothing range of the highest-resolution filters is lowered from $N_1 = 5$ to $N_1 = 3$, the peak clipping effect is much reduced and the SNR score rises from 18.6 dB to 21.9 dB, the result comparable with those yielded by *SureShrink* (20.4 dB for Haar wavelets and 22.9 dB for D4 wavelets).

Of course, a more systematic study, carried out for different time scales and different signal-to-noise ratios, is needed to check whether the simple smoothing algorithm described above is a serious competitor to wavelet-based smoothers. Our purpose here was only to demonstrate the potential of cooperative approach to smoothing. The material presented in this section is therefore just an *example* of using this technique.

5 Conclusion

We have shown how several competing smoothers, differing in design parameters, or even in design principles, can be combined together yielding a better and more reliable smoothing algorithm. The proposed fusion mechanism was inspired by solution to the problem of Bayesian pattern matching, where signal approximation is obtained as a weighted combination of a certain number of fixed signal "patterns". The new scheme is computationally simple and can be used to combine practically all kinds of smoothers. It also allows one to account for the distribution of measurement noise.

References

Cleveland, W.S. & C. Loader (1996). Smoothing by local regression: Principles and methods. in *Statistical Theory*

Table 1
Average SNR scores (in dB) evaluated for 10 competing smoothers (5 average-based smoothers A(N) and 5 median smoothers M(N), where N denotes the smoothing range) and 3 combination smoothers (C_1, C_2, C_3), for 4 test signals contaminated with additive white Gaussian noise (SNR=16.9 dB).

$\widehat{s}(\cdot)$	Blocks	Bumps	HeaviSine	Doppler
A(5)	**18.9±0.1**	**17.9±0.1**	23.6±0.2	**23.2±0.2**
A(11)	16.6±0.1	11.9±0.1	26.4±0.3	21.0±0.1
A(23)	13.7±0.1	7.0±0.1	**27.6±0.3**	16.3±0.1
A(47)	10.7±0.1	3.7±0.1	26.6±0.2	12.2±0.1
A(95)	6.9±0.1	2.3±0.1	24.0±0.2	8.6±0.1
M(5)	22.1±0.2	**17.7±0.2**	22.2±0.2	**21.6±0.2**
M(11)	24.6±0.3	11.2±0.2	25.3±0.3	20.6±0.2
M(23)	**25.4±0.4**	6.2±0.1	27.7±0.4	16.5±0.2
M(47)	24.2±0.4	2.8±0.1	**28.3±0.5**	12.4±0.2
M(95)	8.2±0.1	1.7±0.1	25.9±0.3	8.8±0.1
C_1	24.2±0.4	18.0±0.2	24.9±0.4	23.4±0.3
C_2	25.6±0.6	18.4±0.2	27.7±0.8	25.1±0.4
C_3	**26.6±0.5**	**18.6±0.2**	**29.2±0.8**	**25.9±0.4**

Table 2
Average SNR scores (in dB) evaluated for 10 competing smoothers (5 average-based smoothers A(N) and 5 median smoothers M(N), where N denotes the smoothing range) and 3 combination smoothers (C_1, C_2, C_3), for 4 test signals contaminated with additive white Laplacian noise (SNR=16.9 dB).

$\widehat{s}(\cdot)$	Blocks	Bumps	HeaviSine	Doppler
A(5)	**18.9±0.1**	**17.9±0.1**	23.7±0.3	**23.2±0.2**
A(11)	16.6±0.1	11.9±0.1	26.4±0.3	21.0±0.1
A(23)	13.7±0.1	7.0±0.1	**27.6±0.3**	16.4±0.1
A(47)	10.7±0.1	3.7±0.1	26.6±0.2	12.2±0.1
A(95)	6.9±0.1	2.4±0.1	24.0±0.2	8.6±0.1
M(5)	24.1±0.3	**17.7±0.2**	24.3±0.3	**23.1±0.3**
M(11)	**26.7±0.5**	11.3±0.2	27.9±0.4	21.1±0.3
M(23)	25.0±0.6	6.2±0.1	**29.8±0.6**	16.5±0.2
M(47)	24.2±0.4	2.8±0.1	29.3±0.6	12.4±0.2
M(95)	8.1±0.1	1.7±0.1	25.9±0.3	8.8±0.1
C_1	25.1±0.4	18.0±0.2	25.7±0.4	23.9±0.3
C_2	27.5±0.7	18.4±0.2	29.2±0.8	25.8±0.4
C_3	**28.4±0.7**	**18.6±0.2**	**30.7±0.8**	**26.6±0.4**

and Computational Aspects of Smoothing, W. Hardel & M. Schimek, Eds., Heidelberg, Germany, Physica-Verlag, 10–49.

Donoho, D. & I. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *American Statistical Assoc.*, vol. 90, 1200–1224.

Droge, B. (2006). Asymptotic properties of model selection procedures in linear regression. *Statistics*, vol. 40, 1–38.

Friedl, H. & E. Stampfer (2002). Cross-validation. in *Encyclopedia of Environmetrics*, A.H. El-Shaarawi & W.W. Piegorsch, Eds., vol. 1, 452–460, New York: Wiley.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Hurvich, C.M. & J.S. Simonoff (1998). Smoothing parameter selection in nonparametric regression using an improved AIC criterion. *J. R. Stat. Soc.*, ser. B, vol. 60, 271–293.

Katkovnik, V. (1999). A new method for varying adaptive bandwidth selection. *IEEE Trans. Signal Process.*, vol. 47, 2567–2571.

Lewis, F (1986). *Optimal Estimation*. New York: Wiley.

Saralees, N. (2005). A generalized normal distribution. *Journal of Applied Statistics*, vol. 32, 685–694.

Shibata, R., (1981). An optimal selection of regression variables. *Biometrika*, vol. 68, 45–54.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Stone, M., (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc.*, vol. B36, 111–147.

Orfanidis S. (1996). *Introduction to Signal Processing*. Prentice Hall.

Yang, Y. (2000). Combining different procedures for adaptive regression. *J. Multiv. Anal*, vol. 74, 135–161.

APPENDIX

derivation of (12)

Let

$$\gamma_k = \sum_{i \in T(t)} |\varepsilon_k(i)|^\beta, \quad c = \frac{1}{K}\left[\frac{\beta}{2\Gamma(1/\beta)}\right]^M .$$

Putting $\pi(\alpha) = 1/\alpha$, one arrives at

$$\begin{aligned} J &= \int_0^\infty p(\mathcal{Y}_T(t)|\alpha, H_k)\pi(\alpha|H_k)\pi(H_k)d\alpha \\ &= c\int_0^\infty \alpha^{-(M+1)}\exp\left\{-\gamma_k\alpha^{-\beta}\right\}d\alpha \, . \end{aligned}$$

Using the substitution $x = \gamma_k\alpha^{-\beta}$, one obtains

$$J = \frac{c}{\beta}\,\gamma_k^{-M/\beta}\int_0^\infty x^{(M-\beta)/\beta}e^{-x}dx = c'\gamma_k^{-M/\beta}$$

where $c' = (c/\beta)\Gamma(M/\beta)$. Since c' is a constant independent of k, it can be omitted in definition of $\varphi_k(t)$, which leads to (12).