FACTA UNIVERSITATIS (NIŠ)
SER.: ELEC. ENERG. vol. 16, Dec. 2003, 377-387

Performance Analysis of the Adaptive Algorithm for Bias-to-Variance Trade-off

Ljubiša Stanković

Abstract: Analysis of the algorithm for the mean square error minimization, through the bias-to-variance ratio optimization, is performed. This algorithm is based on the confidence intervals intersection. It does not require explicit knowledge of the estimation bias. The algorithm performance and reliability for various kinds of noise in the estimation are studied.

Keywords: Signal processing, adaptive algorithm, time-frequency distribution, reliability, noise analysis.

1 Introduction

In numerous signal processing methods and applications of noisy signals the result is a biased random variable. This is the case in filtering, smoothing, Fourier transform calculation, instantaneous frequency estimation, time-frequency distributions calculations, LMS adaptive algorithms, direction of arrival estimation, image and multidimensional signal processing, and many other applications not only in signal processing. The variance and bias in most of these cases are functions of one parameter (smoothing interval, number of samples, lag window, step value, number of sensors,...). Behavior of bias and variance is usually opposite with respect to this parameter. When the parameter increases then the variance increases (decreases) and the bias decreases (increases). The optimal parameter value can be determined by minimizing the estimation mean squared error (MSE), provided that some

Manuscript received June 22, 2003. An earlier version of this paper was presented at the 10th Telecommunications Forum TELFOR 2002, November 26-28, 2002, Belgrade, Serbia.

The author is with Faculty of Electrical Engineering, University of Montenegro, 81000 Podgorica, Serbia and Montenegro (e-mail: `l.stankovic@ieee.org`).

signal and noise parameters are explicitly known. However, these parameters are not available in advance. This is especially true for the signal changes, which determine the estimation bias. The adaptive algorithm for determination of the parameter value, in such a way that it is as close to the optimal one as possible, is proposed and used in [1] -[11]. The algorithm is based on the confidence intervals intersection [12]. This algorithm does not require knowledge of the estimation bias value, but only its theoretical behavior formula.

In the first part of this paper we have analyzed and proposed a method for a fine adjustment of the algorithm parameters. The second part of the paper introduces a reliability analysis of the algorithm, along with the statistical study and confirmation of the presented results.

2 Model and Optimal Parameter Value

Consider a noisy signal:

$$x(k) = f(k) + \epsilon(k) \tag{1}$$

with $f(k)$ being a signal and $\epsilon(k)$ being a stationary noise. Suppose that we want to estimate a quantity $Q(k)$ from this noisy signal. In general, this quantity is time-dependent. Also assume that the estimate of this quantity $\hat{Q}(k)$ depends on a parameter h (smoothing interval, number of samples, lag window width, adaptive step value, number of sensors,... [1]-[13]). Let the estimation bias be

$$b(k,h) = \sqrt{B(k)h^n} \tag{2}$$

and the variance

$$\sigma^2(h) = V/h^m. \tag{3}$$

Here, the parameter $B(k)$ depends somehow on the unknown signal $f(k)$. This parameter is unknown. The variance of the estimate is assumed to be time-invariant (as it is true in all of the above mentioned applications). It depends on the parameter h. Variance and squared bias dependence on h is of the $m-th$ and $n-th$ power, respectively. Note that in some cases form with $h \to 1/h$ in the bias and variance expressions is obtained. It does not change the form of algorithm.

The MSE for (2) and (3) is of the form

$$E\left\{(Q(k) - \hat{Q}(k))^2\right\} = \frac{V}{h^m} + B(k)h^n. \tag{4}$$

The MSE in (4) has a minimum with respect to h. This minimum occurs for the optimal value of $h = h_{opt}(k)$. It follows from

$$\frac{\partial E\left\{(Q(k)-\hat{Q}(k))^2\right\}}{\partial h} = \frac{-mV}{h^{m+1}} + nB(k)h^{n-1} = 0_{|h=h_{opt}} \tag{5}$$

in the form

$$h_{opt}(k) = [mV/(nB(k))]^{1/(m+n)}. \tag{6}$$

Note that this relation is not useful in practice, because its right hand-side contains $B(k)$ which depends on the unknown signal $f(k)$.

3 Adaptive Algorithm

Here, we will review the adaptive algorithm [1]-[6] which can produce $h_{opt}(k)$ or, due to the discrete nature of h a value close to $h_{opt}(k)$, without having to know $B(k)$. For the optimal value of h relation (5) holds. Multiplying (5) by h, we get the relationship between the bias and standard deviation (3) for $h = h_{opt}$,

$$b(k, h_{opt}) = \sqrt{\frac{m}{n}}\sigma(h_{opt}). \tag{7}$$

Thus, the bias-to-standard deviation ratio is signal independent at $h = h_{opt}$, $b(k, h_{opt})/\sigma(h_{opt}) = \sqrt{m/n}$.

In the analysis that follows it will be assumed, without loss of generality, that the bias is positive. The estimate $\hat{Q}$ is a random variable distributed around Q with the bias $b(k, h)$ and the standard deviation $\sigma(h)$. Thus, we may write the relation:

$$\left|Q(k) - \left(\hat{Q}(k) - b(k,h)\right)\right| \leq \kappa\sigma(h), \tag{8}$$

where the inequality holds with probability $P(\kappa)$ depending on parameter κ. We will assume that κ is such that $P(\kappa) \to 1$.

Let us introduce a set of discrete parameter h values, $h \in H$,

$$H = \{h_s \,|\; h_s = 2h_{s-1},\; s = 1, 2, ..., J\}. \tag{9}$$

Define upper and lower bounds of the confidence intervals $D_s = [L_s, U_s]$ of the IF estimates as

$$L_s = \hat{Q}_s(k) - (\kappa + \Delta\kappa)\,\sigma(h_s);\; U_s = \hat{Q}_s(k) + (\kappa + \Delta\kappa)\,\sigma(h_s), \tag{10}$$

where $\hat{Q}_s(k)$ is an estimate of Q, with parameter $h = h_s$, and $\sigma(h_s)$ is its standard deviation. Assume that the parameter denoted by $h_{s^+} \in H$ is of h_{opt} order, $h_{s^+} \sim h_{opt}$. We can write $h_{s^+} = 2^p h_{opt}$, where p is a constant close to 0. According to (9) all other parameter values can be written as a function of h_{s^+} as

$$h_s = h_{s^+} 2^{(s-s^+)} = h_{opt} 2^{s-s^+ + p}, \;\; (s - s^+) = ..., -1, 0, 1, ...$$

With this notation, having in mind (7), the standard deviation and the bias from (3) can be expressed by

$$\begin{aligned} \sigma(h_s) &= \sqrt{\frac{V}{h_s^m}} = \sigma(h_{opt}) 2^{-(s-s^+ + p)m/2}, \qquad (11) \\ b(k, h_s) &= \sqrt{B(k) h_s^n} = \sqrt{\frac{m}{n}} \sigma(h_{opt}) 2^{(s-s^+ + p)n/2} \end{aligned}$$

For small values of h_s, when $s \ll s^+$, the bias is negligible, thus $Q(k) \in D_s$ (with probability $P(\kappa + \Delta\kappa) \to 1$). Then, obviously, $D_{s-1} \cap D_s \neq \emptyset$, since at least the true value $Q(k)$, belongs to both confidence intervals. For $s \gg s^+$ the variance is small, but the bias is large. It is clear that there always exists such a large s such that $D_s \cap D_{s+1} = \emptyset$ for a finite $\kappa + \Delta\kappa$.

The **idea behind the algorithm** is that $\Delta\kappa$ in D_s can be found in such a way that the largest s, for which the sequence of the pairs of the confidence intervals D_{s-1} and D_s has at least a point in common, is $s = s^+$. Such a value of $\Delta\kappa$ exists because the bias and the variance are monotonically increasing and decreasing functions of h, respectively, (11). As soon as this value of $\Delta\kappa$ is found (or assumed), an intersection of the confidence intervals D_{s-1} and D_s,

$$\left| \hat{Q}_{s-1}(k) - \hat{Q}_s(k) \right| \leq (\kappa + \Delta\kappa)[\sigma(h_{s-1}) + \sigma(h_s)], \qquad (12)$$

works as an indicator of the event $s = s^+$, i.e., the event $h_s \sim h_{opt}$.

4 Parameters in the Adaptive Algorithm

Two ways of choosing algorithm parameters κ, $\Delta\kappa$, and p will be described next. Their performance does not differ significantly.

Table 1. Parameters in the adaptive algorithm for some m, n, κ, corresponding to the values appearing in practical applications presented in the cited references.

m	1	1	3	3	3	3	3	3
n	4	4	4	4	4	8	8	8
κ	2	3	2	3	5	2	3	5
$\Delta\kappa$	0.86	1.29	0.39	0.58	0.97	0.09	0.14	0.23
p	1.15	1.38	0.34	0.51	0.72	-0.13	-0.03	0.11
p_1	1.34	1.58	0.59	0.76	0.97	0.19	0.30	0.43

4.1 Heuristic Approach

When our knowledge about the variance and bias behavior, given by (4), is not quite reliable, an approximative approach for κ, $\Delta\kappa$, and p determination can be used. Then, we can assume, for example, a value of $\kappa \cong 2.5$, such that $P(\kappa) \cong 0.99$ for Gaussian distribution of estimation error. The value of $\Delta\kappa$ should take into account the bias for the expected optimal parameter value (7). It is common to assume that, for the optimal value of h, the bias and variance are of the same order, resulting in $\Delta\kappa \cong 1$. Then we can expect that the obtained value is close to h_{opt}, thus $p \cong 0$, and the algorithm is completely defined, since all parameters for **the key algorithm equation (12)** are defined. This simple heuristic form has been successfully used in many of the cited references, and *it is highly recommended for most of practical applications.*

4.2 Analytical Approach

When the knowledge about the variance and bias behavior is reliable, i.e., when (4) accurately describes estimation error, then we can calculate all algorithm parameters. According to the algorithm basic idea, only three confidence intervals, D_{s^+-1}, D_{s^+}, and D_{s^++1}, should be considered. The confidence intervals D_{s^+-1} and D_{s^+} **should have,** while D_{s^+} and D_{s^++1} **should not have,** at least one point in common. Assuming that relation (8) holds, and that the bias is positive, this condition means that the minimal possible value of upper D_{s^+-1} bound, (10), denoted by $\min\{U_{s^+-1}\}$, is always greater than or equal to the maximal possible value of the lower D_{s^+} bound, denoted $\max\{L_{s^+}\}$, i.e., $\min\{U_{s^+-1}\} \geq \max\{L_{s^+}\}$. The condition that D_{s^+} and D_{s^++1} do not intersect is given by $\max\{U_{s^+}\} < \min\{L_{s^++1}\}$. The maximal and minimal values of $\hat{Q}(k)$ follows from (8), as $Q(k)+b(k,h)-\kappa\sigma(h) \leq$

$\hat{Q}(k) \leq Q(k) + b(k,h) + \kappa\sigma(h)$. By replacing these values into (10) the above two inequalities result in

$$b(h_{s^+-1}) + \Delta\kappa\sigma(h_{s^+-1}) \geq b(h_{s^+}) - \Delta\kappa\sigma(h_{s^+}),$$
$$b(h_{s^+}) + (2\kappa + \Delta\kappa)\sigma(h_{s^+}) < b(h_{s^++1}) - (2\kappa + \Delta\kappa)\sigma(h_{s^++1}). \tag{13}$$

Since the inequalities are written for the worst case, we can calculate the algorithm parameters by using the corresponding equalities. With (11) we get

$$\begin{aligned} \Delta\kappa &= 2\kappa/[2^{(m+n)/2} - 1], \quad \text{and} \\ 2^p &= \left[\Delta\kappa\sqrt{n/m}\left(2^{m/2}+1\right)/\left(1-2^{-n/2}\right)\right]^{2/(m+n)}. \end{aligned} \tag{14}$$

Values of the parameters $\Delta\kappa$ and p for various distributions, i.e., various values of m and n, are given in Table I.

For further, and very fine tuning of the algorithm parameters, one may want that the adaptive parameter h is unbiased in logarithmic, instead of in linear scale (due to definition (9)). This results in $\Delta p \cong \left[\log_2\left((1+2^{(m+n)/2})/2\right)\right]\frac{2}{m+n} - \frac{1}{2}$, with total logarithmic shift $p_1 = p + \Delta p$, presented in Table 1. Therefore the adaptive parameter (as an estimate of the optimal parameter value) should be $\hat{h}_{opt} = h_{s^+}/2^{p_1}$

Note that the set H of parameter h values is a priori assumed and fixed. Therefore, as long as we can calculate the total logarithmic shift p_1, we can use it in the following ways:

a) To calculate distribution with the new value $h_a = h_{s^+}/2^{p_1}$ as the best estimate of h_{opt},

b) To remain within the assumed set of $h_s \in H$, and to decide only whether to correct the obtained h_{s^+} or not. If $|p_1| \leq 1/2$ the correction is smaller than the parameter h discretization step. Thus, if we remain within the assumed set H we can use $h_a = h_{s^+}$. For $1/2 < p_1 \leq 3/2$ it is better to use $h_a = h_{s^+}/2 = h_{s^+-1}$, as the adaptive parameter h value. Fortunately, the loss of accuracy for the adaptive parameters h_a, as far as they are of h_{opt} order, is not significant since the MSE varies slowly around its stationary point. Thus, in numerical implementations we can use only the parameters from the given set H.

5 Algorithm Reliability Analysis

Here we will analyze the probability that the algorithm produces "a false result" when the algorithm parameters are chosen according to Section IV.B.

As "a false result" will be considered a value h_a obtained from the algorithm, such that it is not one of two the closest values (the first greater or the first smaller) from the discrete set H to the optimal parameter h_{opt}. The false result may follow from the fact that probability that (8) is satisfied is not $P(\kappa) = 1$, except for the finite distributed errors. Since we start analysis with the lowest bias and a large variance a false result will appear if two confidence intervals, when the bias is small, do not intersect. Now, we will find that probability. Assume that the error $\hat{Q} - Q$ takes a value $x > 0$ for a parameter $h = h_{s-1}$. Probability of this event is $p_{s-1}(x)dx$, where $p_{s-1}(x)$ is a pdf of the error $\hat{Q} - Q$ calculated with parameter h_{s-1}. The false result will be produced if the error $\hat{Q} - Q$ with $h = h_s$ is such that two confidence intervals, for h_{s-1} and h_s, do not intersect although the bias is very small (assume zero), (12). This outcome happens when the estimated value with h_s is greater than $x + (\kappa + \Delta\kappa)[\sigma(h_{s-1}) + \sigma(h_s)$ or lower than $-(\kappa + \Delta\kappa)[\sigma(h_{s-1}) + \sigma(h_s) + x$. Thus, the overall false result probability is:

$$P_F = 2\int_0^{\infty}\int_{x+(\kappa+\Delta\kappa)[\sigma(h_{s-1})+\sigma(h_s)]}^{\infty} p_{s-1}(x)p_s(y)dydx$$

$$+2\int_0^{\infty}\int_{-\infty}^{-(\kappa+\Delta\kappa)[\sigma(h_{s-1})+\sigma(h_s)]+x} p_{s-1}(x)p_s(y)dydx \tag{15}$$

Special cases:

1. **Limited distribution of error**, $p_s(x) = 0$ for $|x| > (\kappa + \Delta\kappa)\sigma(h_s)$: Then we have $P_F = 0$, i.e., there is no probability that we will get a false result. For example, for a uniformly distributed error $(\kappa + \Delta\kappa) > \sqrt{3}$ guaranties $P_F = 0$.

2. **Gaussian distributed error** $\hat{Q} - Q$, with $p_s(x) = \exp\left(-x^2/(2\sigma_s^2)\right) / (\sigma_s\sqrt{2\pi})$: The false result probability is:

$$P_F = \operatorname{erfc}\left(\frac{(\kappa + \Delta\kappa)}{\sqrt{2}} \frac{\alpha + 1}{\sqrt{\alpha^2 + 1}}\right) \tag{16}$$

where $\alpha = \sigma(h_{s-1})/\sigma(h_s) = 2^{m/2}$ and $\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty}\exp(-t^2)dt$. The last expression is derived by considering integration domain of (15) in the two-dimensional space (x, y).

For $m = 1$ and other parameters from Table 1, we get that for $\kappa = 2$ false result probability is $P_F < 0.0001$. For $m = 3$ we get $P_F < 0.002$ with $\kappa = 2$. For $\kappa = 3$ these probabilities are $P_F < 10^{-8}$ and $P_F < 10^{-5}$ for $m = 1$ and $m = 3$, respectively.

3. **Heavy-tailed Laplacian error** $\hat{Q}-Q$, with $p_s(x)=\exp(-|x/\sigma_s|)/(2\sigma_s)$: In this case:

$$P_F=\frac{\exp(-\xi(1+1/\alpha))\alpha^2-\exp(-\xi(1+\alpha))}{\alpha^2-1} \tag{17}$$

where $\xi=\sqrt{2}(\kappa+\Delta\kappa)$. This is a heavy-tailed case which requires larger values of κ in order to satisfy the probability that $P(\kappa)$ is close to 1. For example, for $m=3$ and $\kappa=2$ we get $P_F<0.02$, while for $\kappa=3$ we get $P_F<0.002$. For $\kappa=5$ we get $P_F<0.00002$. Obviously for heavy-tailed distributed errors, like for the Laplacian one, larger values of κ are required in order to produce highly reliable results.

6 Illustrations with Statistical Study

Example 1 **(Gaussian error):** We have simulated a biased random variable as

$$\triangle Q=\hat{Q}-Q=\mathbf{a}\sqrt{V/h^m}+\sqrt{B(k)h^n}, \tag{18}$$

whose MSE is of the form (4). Here, $\mathbf{a}=\mathcal{N}(0,1)$ is a Gaussian (zero-mean, unity-variance) random variable, $m=3$, $n=4$, and $V=1$. The bias parameter $B(k)$ in $\triangle Q$ is logarithmically changed within $\frac{1}{7}\log_2(mV/nB(k))\in[-4,4]$, with the step 0.008. Thus, in total 1000 experiments are done for one case.

-For each value of 1000 parameter $B(k)$ values we have calculated optimal parameter according to (6), and plotted $\log_2 h_{opt}$ as a thick gray line in Fig.1.

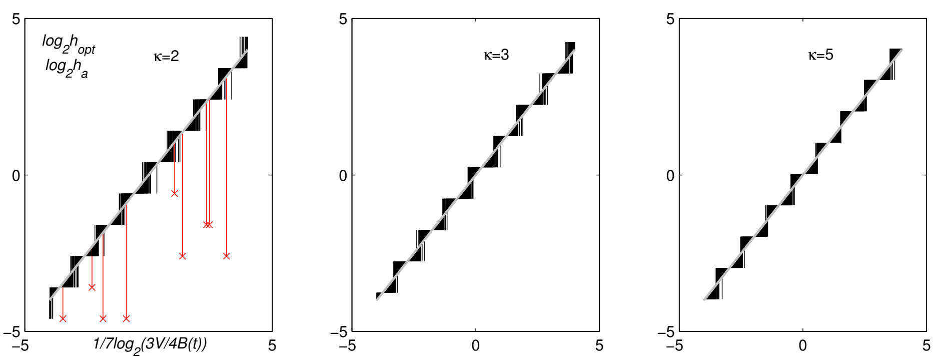

Fig. 1. Gaussian distribution of error: Optimal parameter value h_{opt} (straight gray line), and adaptive values h_a (end of the vertical lines, starting from the optimal parameter line) for $m=3$, $n=4$, $V=1$. The variance to bias ratio $V/B(t)$ is logarithmically changed, in 1000 points. The adaptive value $h_a=h_{s+}/2^{p_1}$ is obtained by correcting h_{s+} by 2^{-p_1}, Table 1. False results are indicated by "x".

-Now, we have assumed that the bias parameter was not known, as it is the case. For a given unknown $B(k)$, the value of $\triangle\hat{Q}(k)$ was simulated for each $h_s \in H$ according to (18). The assumed set of possible parameter values was $H = \{1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8, 16, 32, 64\}$, and $\kappa = 2$. The key algorithm relation (12) was tested each time, with the known standard deviation $\sigma(h_s) = \sqrt{V/h_s^m}$. The largest value of h_s when the key equation (12) was still satisfied was denoted by h_{s+}. Value $\Delta\kappa = 0.39$, corresponding to $m = 3$, $n = 4$, $\kappa = 2$, was used (Table I). The adaptive values $h_a = h_{s+}/2^{p_1}$, $p_1 = 0.59$ (Table I), produced in this way are connected with the optimal parameter line, by thin vertical lines in Fig.1.

The same simulation is repeated with $\kappa = 3$ and $\kappa = 5$.

We can conclude that the presented algorithm almost always chooses the parameters h_s from H which is one of two the nearest values to the optimal one. However, for relatively small $\kappa = 2$ there are few complete misses of the optimal parameter value. The number of these misses ("false results") is in full accordance with the algorithm reliability analysis from the previous section, eq.(16).

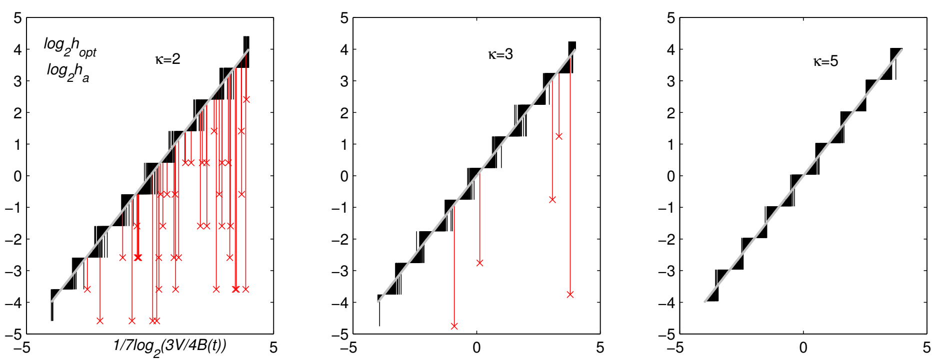

Fig. 2. Laplacian (heavy-tailed) distributed error: Optimal parameter value h_{opt} (straight gray line), and adaptive values h_a (end of the vertical lines, starting from the optimal parameter line) for $m = 3$, $n = 4$, $V = 1$. The variance-to-bias ratio $V/B(t)$ is logarithmically changed, in 1000 points. The adaptive value $h_a = h_{s+}/2^{p_1}$ is obtined by correcting h_{s+} by 2^{-p_1}, Table I.

*Example 2 (***Laplacian error***):* All parameters as in the previous two examples are taken here, but the Laplacian distributed error $Q - \hat{Q}$ is assumed. Laplacian random variable of unity variance $\mathbf{a} = \mathcal{L}(0,1)$ is formed by using $\mathbf{a} = (\mathbf{a}_1\mathbf{a}_2 + \mathbf{a}_3\mathbf{a}_4)/\sqrt{2}$ where $\mathbf{a}_i$ are Gaussian random variables $\mathbf{a}_i = \mathcal{N}(0,1)$. Values of $\kappa = 2$, $\kappa = 3$, and $\kappa = 5$, are considered. Since the noise is a kind of heavy-tailed noise the lowest $\kappa = 2$ here produces quite low $P(\kappa)$, with a small reliability of the algorithm, Fig.2. Number of false result points is in full agreement with (17). Thus, in order to improve the algorithm performance higher values of κ should be used.

7 Conclusion

The algorithm for parameter optimization, in a quite general formulation of the estimation problem, is considered. Reliability study for a general form of noise is done. It has been shown that even in the cases of some heavy tailed noises, like the Laplacian noise, the algorithm can produce accurate and reliable results. The algorithm has a quite general form, and it can be used in a variety of signal processing problems.

Acknowledgment

This work is supported by the Volkswagen Stiftung, Federal Republic of Germany.

References

[1] V. Katkovnik and Lj. Stanković, "Instantaneous frequency estimation using the wigner distribution with varying and data-driven window length," *IEEE Trans. on Signal Processing*, vol. 46, no. 9, pp. 2315–2325, Sept. 1998.

[2] Lj. Stanković and V. Katkovnik, "Algorithm for the instantaneous frequency estimation using time-frequency distributions with variable window width," *IEEE Sig. Proc. Let.*, vol. 5, no. 9, pp. 224–228, Sept.1998.

[3] Lj. Stanković and V. Katkovnik, "Instantaneous frequency estimation using higher order distributions with adaptive order and window length," *IEEE Trans. Information Theory*, vol. 46, no. 1, pp. 302–311, Jan. 2000.

[4] V. Katkovnik and Lj. Stanković, "Periodogram with varying and data-driven window length," *Signal Processing*, vol. 67, no. 3, pp. 345–358, 1998.

[5] Lj. Stanković, "Adaptive instantaneous frequency estimation using tfds," in *Time-Frequency Signal Analysis and Processing* (B. Boashash, ed.), Prentice Hall, 2003.

[6] Lj. Stanković and V. Katkovnik, "The wigner distribution of noisy signals with adaptive time-frequency varying window," *IEEE Trans. on Signal Processing*, vol. 47, no. 4, pp. 1099–1108, April 1999.

[7] B. Krstajić, Z. Uskoković, and Lj. Stanković, "An approach to variable step-size lms algorithm," *IEE Electronics Letters*, vol. 38, no. 16, pp. 927–928, Aug. 2002.

[8] A. Gershman, Lj. Stanković, and V. Katkovnik, "Sensor array signal tracking using a data-driven window approach," *Signal Processing*, pp. 2507–2515, Dec. 2000.

[9] Lj. Stanković, S. Stanković, and I. Djurović, "Space/spatial-frequency analysis based filtering," *IEEE Trans. on Signal Processing*, vol. 48, no. 8, pp. 2343–2352, Aug. 2000.

[10] Lj. Stanković, "On the time-frequency analysis based filtering," *Annales des Telecommunications*, vol. 55, no. 5-6, pp. 216–225, May 2000.

[11] I. Djurović and Lj. Stanković, "Adaptive windowed fourier transform," *Signal Processing*, vol. 83, 2003, in print.

[12] A. Goldenshluger and A. Nemirovski, "On spatial adaptive estimation of non-parametric regression," *Math. Meth. of Statistics*, vol. 6, no. 2, pp. 135–170, 1997.

[13] Z. Hussain and B. Boashash, "Adaptive instantaneous frequency estimation of multicomponent fm signals using quadratic time-frequency distributions," *IEEE Trans. on Signal Processing*, vol. 50, no. 8, pp. 1866–1676, Aug. 2002.