Received: 4 July 2023 | Revised: 21 November 2023 | Accepted: 2 January 2024

DOI: 10.1049/cth2.12628

IET Control Theory & Applications

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Iterative-recursive estimation of parameters of regression models with resistance to outliers on practical examples

Janusz Kozłowski | **Zdzisław Kowalczuk**

Department of Robotics and Decision Systems, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland

Correspondence
Zdzisław Kowalczuk, Department of Robotics and Decision Systems, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Narutowicza 11/12, 80–233 Gdańsk, Poland.
Email: kova@pg.edu.pl

Abstract

Here, identification of processes and systems in the sense of the least sum of absolute values is taken into consideration. The respective absolute value estimators are recognised as exceptionally insensitive to large measurement faults or other defects in the processed data, whereas the classical least squares procedure appears to be completely impractical for processing the data contaminated with such parasitic distortions. Since the absolute value quality index cannot be minimised analytically, an iterative solution is used to find optimal estimates of the parameters of the underlying regression model. In addition, an approximate recursive estimator is proposed and implemented for on-line evaluation of system parameters. The convergence (basic property) of the iterative estimator is show to be proven and some aspects related to the absolute value criterion are explained. This allows for the formulation of practical conclusions and indication of directions for further research. In addition, the effectiveness of the described iterative-recursive estimation procedures is practically verified by appropriate numerical experiments.

1 | INTRODUCTION

A substantial progress in technical science, being observed continuously within the recent decades and, in particular, the associated development of digital technologies, resulted in a manifold of spheres of human activity supported by computer-aided systems. The evolution of these technologies has strongly influenced modernisation in many common areas, such as telecommunication, radio and television, Internet, banking and, last but not least, automation and robotics.

Considering the field of automation, significant progress can be observed in such detailed areas as process supervision, prediction and implementation of advanced control algorithms [1, 2], digital data processing, prediction and filtration [3, 4], and system identification discussed here [5, 6].

Parametric identification, understood as matching the parameters of the adopted mathematical description to the dynamics of the supervised process expressed in pairs of input-output trajectories, seems to be an important issue in many practical applications, such as description of physical phenomena, supervision of hazardous chemical processes, prediction of trends or determination of various econometric indicators.

Since processes are continuous in the real world, continuous-time models are most suitable for reliably describing the behaviour of such objects, among which we distinguish differential equations, state-space descriptions or Laplace transfer functions, and ('non-parametric') convolutional representations or spectral characteristics in frequency domain.

In most practical applications, estimation algorithms are mainly based on objective functions involving sums of (weighted) squared prediction errors. Such a square index may express energy losses, provided that the error signal has a suitable physical meaning. Such criteria are convenient for analytical minimisation and lead to known least squares (LS) estimation procedures [7].

However, there are situations in which the above energy interpretation is inadequate. In the case of trading and economy, the classic least squares estimator can of course be used to track the evolution of certain stock indicators or market trends. But then using the quadratic criterion leads to the interpretation of market gains or losses in terms of 'square dollars'. This example clearly shows that in practical applications, it may be more useful to stick to simple units and use absolute-valued index minimisation.

It is obvious that the analytical ease of minimising quadratic functions predestines the least squares method to solve many tasks, and the resulting estimation algorithm can be presented in a transparent recursive form that is convenient for numerical implementation.

All methods derived from square measures, however, are highly sensitive to large measurement errors called outliers. The appearance of such harmful phenomena can deteriorate the quality of measurement data processing by the LS method. The observed disadvantage can be attributed to the fact that the outlier is squared here, that is, it has a greater impact on the quality indicator than normal data.

In view of the above, it is worth considering a more balanced approach, which can be based, for example, on minimising the aforementioned absolute value criterion [8, 9]. The resulting estimator, which penalises prediction error less severely, is less sensitive to outlier data. However, this desirable property comes with the increased cost of numerical overhead associated with minimising the non-differentiable index [10, 11].

Basically, we can distinguish two approaches to outlier-insensitive identification:

- the use of specific detection algorithms to isolate and eliminate outliers in the processed data, which then allows for reliable identification using the classic least squares procedure,
- the use of identification methods (such as the non-quadratic quality criteria minimisation algorithms discussed) that are inherently insensitive to outliers.

In our opinion, the first methodology (a) is less convenient because the isolation of outliers in the dataset requires additional processing in the form of implementing a hypothesis verification method (e.g. a solution based on the Grubbs test for prior elimination of outliers). As a result, this approach is more suitable for off-line implementation. Furthermore, outlier detection tests are great for isolating single errors, but in the case of sequences of outliers, these tests may fail (there is no guarantee of success). On the other hand, applying a least squares algorithm to unreliable data (with retained errors) may lead to failure.

Accepting such arguments, in this study we consider the latter approach (b) based on the implementation of inherently data error-tolerant methods such as LA procedures. Developing the concept of robust identification [12, 13], in this paper we present new results related to outlier-insensitive identification of regression models. We discuss and derive iterative-recursive least absolute value (LA) estimators and verify them numerically in non-trivial simulation experiments (i.e. with outliers or other destructive errors that distort the measurements).

The proposed proprietary procedures along with improvements to avoid problems with small divisors are our contribution to the field of system identification. An equally important theoretical contribution is the proof of convergence of the iterative weighted absolute value algorithm. Importantly, there are also grounds for stating that our solutions are better than other numerical optimisation methods (e.g. linear programming/simplex or gradient descent based on the Huber loss function).

The paper is organised as follows. In Section 2, we discuss linear regression in terms of the smallest sum of absolute values using a simple one-parameter model. In addition, on the attached numerical example, we verify the declared insensitivity to occasional measurement errors of large magnitudes (outliers).

In Section 3, we recall the well-known method of least squares. An outline of the rearrangements leading to the recursive LS procedure is helpful here to illustrate the reasoning presented in the following section.

The basic iterative-recursive smallest-absolute-error estimators are derived and discussed in Section 4.

Section 5 shows the results of practical applications of these 'non-square' estimators, which confirm the basic properties of LA procedures and their insensitivity to harmful outliers.

Finally, in Section 6, we highlight the original contribution and indicate directions for further research.

In addition, Appendix A proves the convergence theorem of the iterative weighted absolute value algorithm, and Appendix B explicitly comments on the problems of the non-differentiable LA criterion.

2 | LINEAR REGRESSION

As a gentle introduction to further considerations, let us recall the classic concept of linear regression based on the following static model

$$y(l) = \boldsymbol{\phi}(l)\,\theta + e(l) \tag{1}$$

where θ is the unknown proportionality factor, $\boldsymbol{\phi}(l)$ means the scalar input (excitation), $y(l)$ stands for the process output (response), and $e(l)$ represents the equation or prediction error (also called the residual). The optimal value of θ can be estimated based on available data $\{\boldsymbol{\phi}(1), \dots, \boldsymbol{\phi}(k)\}$ and $\{y(1), \dots, y(k)\}$.

Note that this seemingly simple model (1) has many physical applications (e.g. in the case of Hooke's law, θ stands for elasticity modulus, while the essence of piezoelectricity is the linear relationship between mechanical stress and the electric charge accumulating in crystalline materials).

The LS estimate of θ follows directly from the minimisation of the penalty function (quadratic index)

$$I(\theta) = \frac{1}{2}\sum_{l=1}^{k} e^2(l) = \frac{1}{2}\sum_{l=1}^{k} [y(l) - \boldsymbol{\phi}(l)\,\theta]^2 \tag{2}$$

By zeroing the derivative of (2)

$$\frac{dI}{d\theta} = \frac{dI}{de}\,\frac{de}{d\theta} = -\sum_{l=1}^{k} \boldsymbol{\phi}(l)\,e(l)$$

$$= -\sum_{l=1}^{k} \boldsymbol{\phi}(l)\,[y(l) - \boldsymbol{\phi}(l)\,\theta] = 0 \tag{3}$$

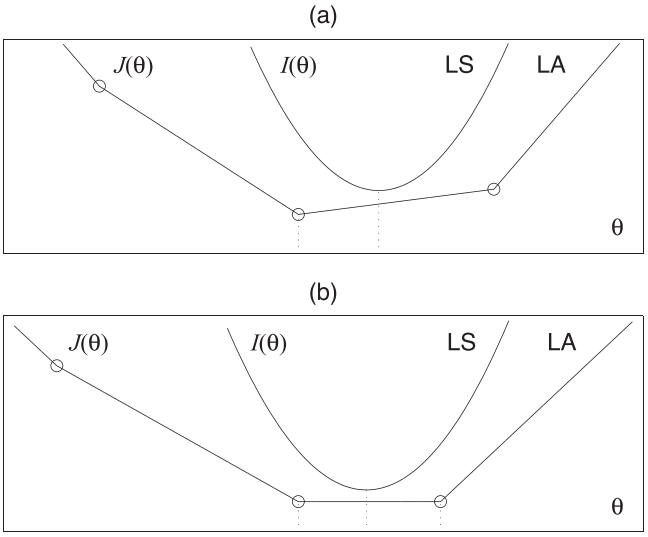

FIGURE 1 The course of the penalty function for the LS and LA rules, where the LA index may, depending on the measurement data, take the form with (a) a unique minimum or with (b) a flat minimum zone.

we get the LS estimate given by

$$\hat{\theta} = \left[\sum_{l=1}^{k} \phi^2(l)\right]^{-1} \left[\sum_{l=1}^{k} \phi(l)\, y(l)\right] \tag{4}$$

Since the strictly convex LS criterion $I(\theta)$ is unimodal (Figure 1) and the second derivative of (2) is positive

$$\frac{d^2 I}{d\theta^2} = \sum_{l=1}^{k} \phi^2(l) > 0 \tag{5}$$

solution (4) yields the global minimum of (2).

As mentioned, the estimators resulting from the minimisation of the quadratic criteria can be ineffective in reliably assessing the system parameters in the presence of large measurement errors, called outliers.

As a remedy to this problem, it is worth thinking about a method that penalises less severely for large prediction errors. Thus, let us consider linear regression in the sense of the least sum of absolute values.

The LA index corresponding to the regression model (1) takes the form

$$J(\theta) = \sum_{l=1}^{k} |\, e(l)\,| = \sum_{l=1}^{k} |\, y(l) - \phi(l)\,\theta\,| \tag{6}$$

An explanation of the origin of the shape of the LA functional used as a minimisation criterion, including the location of its kink points, is provided in Appendix B, where the quality index $J(\theta)$ need not be a unimodal function (Figure 1), especially in numerical representation. Other effects of minimising LA are discussed below.

Due to the piecewise-linear course of function (6) with kinks (points of discontinuity of its derivative as shown in Figures B1a and B2a in Appendix B), it is impossible to carry out its analytical differentiation (in order to minimise).

The desired derivative of the absolute value function $|e|$, however, can be described as

$$\frac{d|e|}{de} = \operatorname{sign}(e) = \frac{e}{|e|} \tag{7}$$

where the residual error $e = e(l)$ shown in (1) is actually a function of both the discrete-time moment (l) and the proportionality factor (θ): $e(l) = e(l, \theta) = y(l) - \phi(l)\theta$.

Of course, this derivative does not exist for $e = 0$ (at the minimum of this function). On the other hand, an approximate minimisation of the functional (6) may be used here, provided that an estimate $\hat{e}(l)$ of the prediction error $e(l)$ is available (for example, from another estimation procedure).

Taking into account the approximation (7) and the obvious relation $d\, e(l, \theta)/d\theta = -\phi(l)$, the derivative of the functional (6) can be represented as

$$\begin{aligned} \frac{dJ}{d\theta} &= \frac{dJ}{de}\,\frac{de}{d\theta} = -\sum_{l=1}^{k} \phi(l)\,\operatorname{sign}(e) \\ &= -\sum_{l=1}^{k} \phi(l)\,\frac{e(l)}{|e(l)|} \approx -\sum_{l=1}^{k} \frac{\phi(l)\, e(l)}{|\hat{e}(l)|} \\ &= -\sum_{l=1}^{k} \frac{\phi(l)\,[y(l) - \phi(l)\,\theta]}{|\hat{e}(l)|} = 0 \end{aligned} \tag{8}$$

where in the denominator of (8) we propose to use the aforementioned auxiliary estimate $\hat{e}(l)$ of the prediction error $e(l) = y(l) - \phi(l)\theta$.

Thus, the estimation of parameter θ in the approximate sense of LA can be given as

$$\hat{\theta} = \left[\sum_{l=1}^{k} \frac{\phi^2(l)}{|\hat{e}(l)|}\right]^{-1} \left[\sum_{l=1}^{k} \frac{\phi(l)\, y(l)}{|\hat{e}(l)|}\right] \tag{9}$$

An additional way to improve the accuracy of the estimate (9) is the mechanism of iteratively reaching the optimal value of the criterion $J(\theta)$ during a finite number of steps: $\hat{\theta}^{[r]}$ $(r = 0, 1, \ldots)$.

To start the iteration, we can, for example, use the easy-to-calculate LS estimate (4) as the initial value $\hat{\theta}^{[0]}$ (for $r = 0$). Next, with such a current estimate of θ, we can recalculate all prediction error samples (for the entire measurement record: $l = 1 \ldots k$) and substitute them, i.e. the updated values of $\hat{e}(l)$, into (9).

In this way we come directly to the following procedure of successive approximations $\hat{\theta}^{[r]}$ of the estimated parameter θ

$$\hat{e}^{[r]}(l) = y(l) - \phi(l)\,\hat{\theta}^{[r]} \tag{10}$$

$$\hat{\theta}^{[r+1]} = \left[\sum_{l=1}^{k} \frac{\phi^2(l)}{|\hat{e}^{[r]}(l)|}\right]^{-1} \left[\sum_{l=1}^{k} \frac{\phi(l)\, y(l)}{|\hat{e}^{[r]}(l)|}\right] \tag{11}$$

TABLE 1 Iterative LA estimates of the scalar parameter θ of the model (1).

r	$\hat{\theta}^{[r]}$	$J(\hat{\theta}^{[r]})$	$\hat{e}(l)\mid_{l=9}$
1	6.5369	176.4065	12.1666
2	7.1427	166.1072	6.7140
3	7.5010	160.1058	3.4891
4	7.6527	157.8305	2.1239
5	7.7772	155.9632	1.0035
6	7.8473	154.9119	0.3727
7	7.8798	154.4241	0.0801
8	7.8898	154.2940	0.0101

Note that by solving (10) for $y(l) = \phi(l)\,\hat{\theta}^{[r]} + \hat{e}^{[r]}(l)$ and then substituting this it into (11), we acquire its equivalent innovation form

$$\hat{\theta}^{[r+1]} = \hat{\theta}^{[r]} + R^{-1}(k)\,\psi(k) \tag{12}$$

$$R(k) = \sum_{l=1}^{k} \frac{\phi^2(l)}{|\hat{e}^{[r]}(l)|} \tag{13}$$

$$\psi(k) = \sum_{l=1}^{k} \frac{\phi(l)\,\hat{e}^{[r]}(l)}{|\hat{e}^{[r]}(l)|} \tag{14}$$

where $R(k)$ is the Hessian and '$-\psi(k)$' is the gradient of the function $J(\theta)$ (both factors are scalar in this case).

Processing in both cases (10)–(11) or (12)–(14) should be stopped when the observed decrease in the minimised criterion (6) in the next iteration turns out to be negligibly small, that is, it falls below a certain positive numerical threshold $\Delta_{\min}$

$$\left| J(\hat{\theta}^{[r]}) - J(\hat{\theta}^{[r+1]}) \right| < \Delta_{\min} \tag{15}$$

The idea of iterative processing described above, which is also applied in [14, 15], was originally presented in [8], where it was called "re-weighted least squares". It is of fundamental importance here that the sequence of values of the quality indicator (6) calculated in successive iterations of the estimate $\hat{\theta}$ is generally decreasing: $J(\hat{\theta}^{[r+1]}) < J(\hat{\theta}^{[r]})$. This property, which is relevant to the fitness and effectiveness of the given end condition (15), is analysed and proven in Appendix A.

The properties of the LS and LA methods considered above were exercised on a simple numerical example of linear regression. In this experiment, the model (1) was fitted to a set of ten pairs of numerical data ($k = 10$). The corresponding sequence of output values $y(l)$, for $l = 1 \ldots k$, was calculated for $\theta = 8$ and in the presence of zero-mean uniformly distributed white noise $e(l)$ with variance $\sigma_e^2 = 0.64$. Additionally, occasional zero outliers, $y(l) = 0$, were simulated for randomly selected calculation moments $l = 2$, 3, 6 and 8. The results obtained for linear regression using the LS and LA schemes are presented in Table 1. In Figure 2, the bullets indicate the measurements $y(l)$,

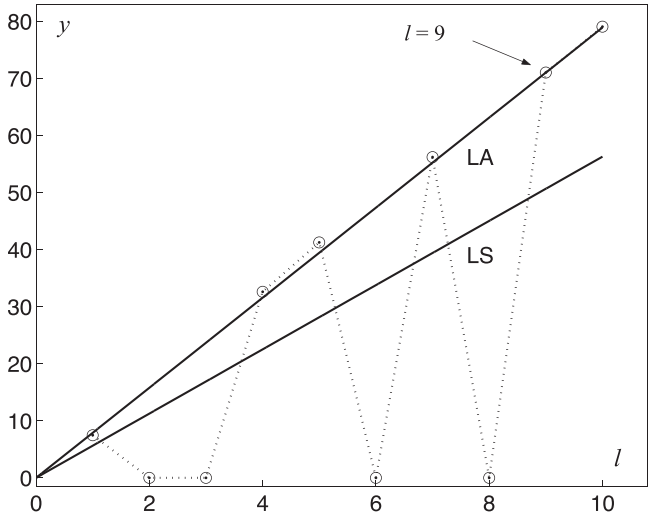

FIGURE 2 Linear regression for numerical data (bullets on the dotted line), obtained by the LS and LA methods: the theoretical tangent is 8 (the line is invisible covered by the solid LA line; the dashed line is used to better show the faults).

and the resulting tangents θ obtained in the LS and LA sense, respectively, are shown with solid lines.

As shown in Appendix B, the minimum of the LA index (6), which is a non-differentiable $J(\theta)$, is usually found at one of the kink points given by $y(l)/\phi(l)$. Figure 2 shows how the optimal LA regression line passes through the point $[\phi(9), y(9)]$, additionally indicated by an arrow. Thus, the optimal estimate of θ is equal to $\hat{\theta} = y(9)/\phi(9) = 7.8898$. Also, the smallest prediction error $\hat{e}(l) = y(l) - \phi(l)\hat{\theta}$ is assigned to the 9th measurement point ($l = 9$). The iterative convergence of its assessment to zero (for increasing index r) is presented in Table 1 together with the corresponding sequence of the estimates of θ, and the related decreasing index (6).

In this study, the iterative estimate of θ was obtained in 8 iterations with the threshold $\Delta_{\min} = 0.15$ used in the end condition (15). It is obvious that with a smaller threshold, the number of iterations (12)–(14) will be greater (e.g. for $\Delta_{\min} = 10^{-4}$, the number of runs is 15).

Figure 2 shows that in the presence of outliers, the LS estimate deteriorated significantly, giving the parameter $\hat{\theta} = 5.6282$. In contrast to LS, the iterative result of the LA evaluation is 7.8898 (with threshold $\Delta_{\min} = 0.15$ and 8 iterations), which confirms the advantage of absolute value estimators in terms of insensitivity to outliers.

The insensitivity to big errors shown above is a fundamental advantage of the discussed LA approach. Yet, this advantage comes at the cost of some numerical peculiarities associated with this iterative processing.

First, since the minimum of the piecewise-linear LA criterion (6) is usually located in its non-differentiable kink point (Figures 1a and B1a), you can rely on that the corresponding value of the prediction error in recursive calculations tends to zero (Table 1). The problem of inconvenient divisors close to zero in (11) (or in (13), (14)) can be solved here by regularisation

techniques. In the simplest approach, we replace very small values $|\hat{e}^{[r]}(l)|$ with a fixed positive numerical threshold $e_{\min}$ ($|\hat{e}^{[r]}(l)| \leftarrow e_{\min}$ for $|\hat{e}^{[r]}(l)| < e_{\min}$).

Second, because the basic absolute value criterion can reveal a flat minimum range of optimal solutions for θ (as in Figure 1b and Figure B2a). Meanwhile, the sequence of iteratively determined values of the LA criterion (for $r = 0, 1, \ldots$) is decreasing $J(\hat{\theta}^{[r+1]}) < J(\hat{\theta}^{[r]})$, as long as the gradient present in (12) was non-zero ($\psi(k) \neq 0$). This is a requirement for most regression methods (see also Appendix A).

However, for the 'flat' case with $\psi(k) = 0$, the innovative Equation (12) 'gets stuck' in the state $\hat{\theta}^{[r+1]} = \hat{\theta}^{[r]}$. Thus, the criterion $J(\theta)$ will not be further minimised because, due to the equality $J(\hat{\theta}^{[r+1]}) = J(\hat{\theta}^{[r]})$, condition (15) immediately breaks the iteration loop with $\hat{\theta} = \hat{\theta}^{[r]}$.

Later, in Section 4, the LA method is generalised for arbitrary regression models.

3 | LS ESTIMATORS

Let us now consider a general regression model of a multi-parameter SISO system subjected to identification

$$y(l) = \boldsymbol{\varphi}^{\mathrm{T}}(l)\,\boldsymbol{\theta} + e(l) \tag{16}$$

$$\boldsymbol{\varphi}(l) = [\ \phi_1(l) \quad \ldots \quad \phi_m(l)\ ]^{\mathrm{T}} \tag{17}$$

$$\boldsymbol{\theta} = [\ \theta_1 \quad \ldots \quad \theta_m\ ]^{\mathrm{T}} \tag{18}$$

The regression vector $\boldsymbol{\varphi}(l)$ contains the scalar components $\phi_i(l)$ (signal measurements or deterministic terms, see Section 5), while $\boldsymbol{\theta}$ stands for the vector of estimated scalar parameters. The stochastic term $e(l)$, referred to as the prediction or residual error, includes disturbances and other under-modelling.

There are many systems identification procedures for estimating such an unknown parameter vector $\boldsymbol{\theta}$. One of the simplest methods of such evaluation is the least squares routine. To get a concrete foundation for our analysis, we will recall here the necessary information and formulae related to the batch and recursive versions of the LS method.

The classic weighted least squares estimation algorithm comes from minimising the following, strictly convex, quadratic criterion

$$I(\boldsymbol{\theta}) = \frac{1}{2}\sum_{l=1}^{k}\gamma(l)\,e^2(l) = \frac{1}{2}\sum_{l=1}^{k}\lambda^{k-l}[y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\,\boldsymbol{\theta}]^2 \tag{19}$$

where the weighting function $\gamma(l) > 0$ can be practically represented by a parameterised exponential window $\gamma(l) = \lambda^{k-l}$. The weighting factor λ from the practical range [0.9, 1] introduced here determines the rate of exponential forgetting, which is useful when tracking the parameters of a non-stationary process.

The concept of the effective number of observations, also called the length of the estimator's memory, is useful here, expressed by the formula $M = 1/(1-\lambda)$.

The LS index can be directly minimised by zeroing the gradient of (19)

$$\nabla_{\boldsymbol{\theta}} I = -\sum_{l=1}^{k}\lambda^{k-l}\,\boldsymbol{\varphi}(l)\,[y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\,\boldsymbol{\theta}] = \mathbf{0} \tag{20}$$

Hence, the LS estimation of the parameter $\boldsymbol{\theta}$ takes its algebraic form, which we will also call the batch form

$$\hat{\boldsymbol{\theta}}(k) = \left[\sum_{l=1}^{k}\lambda^{k-l}\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)\right]^{-1}\left[\sum_{l=1}^{k}\lambda^{k-l}\boldsymbol{\varphi}(l)y(l)\right] \tag{21}$$

It is easy to check that the Hessian of the criterion function (19) is positive definite

$$\nabla^2_{\boldsymbol{\theta}} I = \left[\sum_{l=1}^{k}\lambda^{k-l}\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)\right] > 0 \tag{22}$$

Summing up, the obtained solution (21) gives the global minimum of the convex/unimodal LS index (19).

The batch estimator (21) can be converted into its convenient recursive form by separating the current data (k) in the expressions of the estimator (21) as follows

$$\sum_{l=1}^{k}\lambda^{k-l}\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l) = \lambda\sum_{l=1}^{k-1}\lambda^{k-1-l}\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l) + \boldsymbol{\varphi}(k)\boldsymbol{\varphi}^{\mathrm{T}}(k) \tag{23}$$

$$\sum_{l=1}^{k}\lambda^{k-l}\boldsymbol{\varphi}(l)y(l) = \lambda\sum_{l=1}^{k-1}\lambda^{k-1-l}\boldsymbol{\varphi}(l)y(l) + \boldsymbol{\varphi}(k)y(k) \tag{24}$$

To avoid the numerically problematic inversion present in expression (21), the matrix inversion lemma is commonly used [7]. As a consequence, we can easily compute the inverse of the Hessian present in (21)–(23), hereinafter referred to as the covariance matrix $\boldsymbol{P}(k)$, and show the estimate (21) in the form of innovation.

The obtained recursive form of the weighted LS algorithm, containing the 'a priori' evaluation of the prediction error $\varepsilon(k)$, the update of the covariance matrix $\boldsymbol{P}(k)$ and the resulting innovation of the estimation vector, can be presented as [7]

$$\varepsilon(k) = y(k) - \boldsymbol{\varphi}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1) \tag{25}$$

$$\boldsymbol{P}(k) = \frac{1}{\lambda}\left[\boldsymbol{P}(k-1) - \frac{\boldsymbol{P}(k-1)\boldsymbol{\varphi}(k)\boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)}{\lambda + \boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{\varphi}(k)}\right] \tag{26}$$

$$\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{P}(k)\,\boldsymbol{\varphi}(k)\,\varepsilon(k) \tag{27}$$

where the starting value of the initially-diagonal covariance matrix $\boldsymbol{P}(0)$ is chosen quite arbitrarily, as long as the elements

on the main diagonal are reasonably large. For example, it can be $\boldsymbol{P}(0) = \text{diag}\,[10^5 \ \ldots \ 10^5]$, which means a high uncertainty of the initial assessment of $\boldsymbol{\theta}$.

It is easy to show that the simplest LS procedure (no weighting, $\lambda = 1$) generates unbiased estimates of $\boldsymbol{\theta}$, provided that the regression data $\boldsymbol{\varphi}(l)$ and the prediction error $e(l)$ are mutually uncorrelated: $\mathrm{E}\{\boldsymbol{\varphi}(l)e(l)\} = \boldsymbol{0}$.

Very often, unfortunately, the above condition is violated, because the prediction error does not always take the form of white noise (i.e. a sequence of zero-mean independent random variables). When this happens, that is, $e(l)$ takes the form of a correlated process, it is better to use the instrumental variable technique, which can significantly improve the accuracy of the estimation [16, 17].

As shown, algorithms derived from quadratic criteria are usually ineffective in processing measurement data contaminated with large errors. Therefore, in the following we will focus on estimation schemes that are insensitive to large outliers, in particular procedures resulting from the minimisation of a non-quadratic quality measure.

4 | LA ESTIMATORS

Given model (16)–(18), the estimator of weighted least absolute values results from minimising the following non-square index

$$J(\boldsymbol{\theta}) = \sum_{l=1}^{k} \gamma(l)\,\big|e(l)\big| = \sum_{l=1}^{k} \lambda^{k-l}\,\big|y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{\theta}\big| \tag{28}$$

where $\gamma(l)$ represents the weighting mechanism introduced earlier in Section 3.

According to the reasoning presented in Section 2, the LA index can be approximately minimised as long as an estimate of the prediction error $e(l)$ is available (from the LS procedure, for instance). Again, based on relation (7), the gradient of $J(\boldsymbol{\theta})$ can be shown as

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J &= -\sum_{l=1}^{k} \lambda^{k-l}\boldsymbol{\varphi}(l)\,\text{sign}(e)\\ &= -\sum_{l=1}^{k} \lambda^{k-l}\boldsymbol{\varphi}(l)\,\frac{e(l)}{|e(l)|} \approx -\sum_{l=1}^{k} \lambda^{k-l}\frac{\boldsymbol{\varphi}(l)\,e(l)}{|\hat{e}(l)|}\\ &= -\sum_{l=1}^{k} \lambda^{k-l}\frac{\boldsymbol{\varphi}(l)\,[y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\,\boldsymbol{\theta}]}{|\hat{e}(l)|} = \boldsymbol{0}\end{aligned} \tag{29}$$

where, as in (8), some estimate of the prediction error $e(l)$ can be used in the denominator. Then, the algorithm to evaluate the vector $\boldsymbol{\theta}$ takes the following batch form

$$\hat{\boldsymbol{\theta}} = \left[\sum_{l=1}^{k} \lambda^{k-l}\frac{\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)}{|\hat{e}(l)|}\right]^{-1}\left[\sum_{l=1}^{k} \lambda^{k-l}\frac{\boldsymbol{\varphi}(l)y(l)}{|\hat{e}(l)|}\right] \tag{30}$$

As before, the accuracy of the estimator (30) can be improved by the iterative approach ($r = 0, 1, \ldots$) to $\hat{\boldsymbol{\theta}}^{[r]}$, initialising this process with $\hat{\boldsymbol{\theta}}^{[0]}$ resulting from the LS estimate (21). All this allows for the following formulation of the procedure of successive approximations of $\boldsymbol{\theta}$, which is a vector generalisation of the system of scalar equations (10), (11)

$$\hat{e}^{[r]}(l) = y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\,\hat{\boldsymbol{\theta}}^{[r]} \tag{31}$$

$$\hat{\boldsymbol{\theta}}^{[r+1]} = \left[\sum_{l=1}^{k} \lambda^{k-l}\frac{\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)}{|\hat{e}^{[r]}(l)|}\right]^{-1}\left[\sum_{l=1}^{k} \lambda^{k-l}\frac{\boldsymbol{\varphi}(l)y(l)}{|\hat{e}^{[r]}(l)|}\right] \tag{32}$$

Note that in the above iterative batch estimator, at each step, the most recent $\boldsymbol{\theta}$ estimate is used to update all errors (31), for the entire data segment ($l = 1 \ \ldots \ k$).

Defining the output as $y(l) = \boldsymbol{\varphi}^{\mathrm{T}}(l)\,\hat{\boldsymbol{\theta}}^{[r]} + \hat{e}^{[r]}(l)$ according to (31) and substituting it into (32), we can describe the iterative estimator in the innovation form

$$\hat{\boldsymbol{\theta}}^{[r+1]} = \hat{\boldsymbol{\theta}}^{[r]} + \boldsymbol{R}^{-1}(k)\,\boldsymbol{\psi}(k) \tag{33}$$

$$\boldsymbol{R}(k) = \sum_{l=1}^{k} \lambda^{k-l}\frac{\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)}{|\hat{e}^{[r]}(l)|} \tag{34}$$

$$\boldsymbol{\psi}(k) = \sum_{l=1}^{k} \lambda^{k-l}\frac{\boldsymbol{\varphi}(l)\,\hat{e}^{[r]}(l)}{|\hat{e}^{[r]}(l)|} \tag{35}$$

where $\boldsymbol{R}(k)$ is the Hessian and '$-\boldsymbol{\psi}(k)$' is the gradient of the cost functional $J(\boldsymbol{\theta})$.

As in Section 2, the computational processes of both estimators (31)–(32) and (33)–(35) terminate on the basis of the observed stagnation in the minimisation criterion (28), exactly when $|J(\hat{\boldsymbol{\theta}}^{[r]}) - J(\hat{\boldsymbol{\theta}}^{[r+1]})|$ drops below the assumed threshold $\Delta_{\min}$, as in (15). And we solve the problem of divisors $|\hat{e}^{[r]}(l)|$ which are close to zero by replacing them with a certain value equal to a small positive constant called the threshold $e_{\min}$.

Appendix A states that the sequence of calculated values of the quality functional $J(\boldsymbol{\theta})$ as a convex function (with a lower bound and infimum) is decreasing during iterations: $J(\hat{\boldsymbol{\theta}}^{[r+1]}) \leq J(\hat{\boldsymbol{\theta}}^{[r]})$ (for $r = 0, 1, \ldots$). Such a sequence, bounded from below by a number not less than zero, must converge. Hence, we conclude that the iterative method (33)–(35) allows minimisation of the considered quality indicator in all cases, for sharp or flat minima. Note also that in the considered multi-parameter case with noisy measurement, the probability of some hyper-flat minimum is negligibly small.

From an implementation point of view, this can happen when the modulus of the gradient (35) accidentally approaches zero ($||\boldsymbol{\psi}(k)|| \cong 0$). Then we also have $||\hat{\boldsymbol{\theta}}^{[r+1]}|| \cong ||\hat{\boldsymbol{\theta}}^{[r]}||$. In this case, we can conclude that the criterion has reached its

minimum, because $J(\hat{\boldsymbol{\theta}}^{[r+1]}) \cong J(\hat{\boldsymbol{\theta}}^{[r]})$ means that the final condition (15), as in the case of one-parameter regression, based on the modulus of the difference between the latest functionals $|J(\hat{\boldsymbol{\theta}}^{[r]}) - J(\hat{\boldsymbol{\theta}}^{[r+1]})| < \Delta_{\min}$, can stop the iteration.

An approximate recursive implementation of the iterative LA estimator can be obtained in a similar way as before, starting from the recursive notation of the summation formulae in (32)

$$\sum_{l=1}^{k} \lambda^{k-l} \frac{\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)}{|\hat{e}(l)|} = \lambda \sum_{l=1}^{k-1} \lambda^{k-1-l} \frac{\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)}{|\hat{e}(l)|} + \frac{\boldsymbol{\varphi}(k)\boldsymbol{\varphi}^{\mathrm{T}}(k)}{|\hat{e}(k)|} \tag{36}$$

$$\sum_{l=1}^{k} \lambda^{k-l} \frac{\boldsymbol{\varphi}(l)y(l)}{|\hat{e}(l)|} = \lambda \sum_{l=1}^{k-1} \lambda^{k-1-l} \frac{\boldsymbol{\varphi}(l)y(l)}{|\hat{e}(l)|} + \frac{\boldsymbol{\varphi}(k)y(k)}{|\hat{e}(k)|} \tag{37}$$

Then, applying the matrix inversion lemma to (33), we obtain a recursive form of the LA algorithm, similar to (25)–(27). This means that our LA procedure takes the typical form where the regression (column) vector $\boldsymbol{\varphi}(k)$ is replaced by $\boldsymbol{\varphi}(k)/|\hat{e}(k)|$.

In this calculation we have the 'a priori' prediction error $\varepsilon(k)$, which is the best measure of the current on-line $\hat{e}(k)$ and can approximately replace it in the LA algorithm (leaving the problem of small divisors). Finally, the LA scheme including the update of the covariance matrix $\boldsymbol{P}(k)$ and the correction of the estimation vector can be presented as

$$\varepsilon(k) = y(k) - \boldsymbol{\varphi}^{\mathrm{T}}(k)\hat{\boldsymbol{\theta}}(k-1) \tag{38}$$

$$\begin{aligned} \boldsymbol{P}(k) &= \frac{1}{\lambda}\left[\boldsymbol{P}(k-1) - \frac{\boldsymbol{P}(k-1)\frac{\boldsymbol{\varphi}(k)}{|\varepsilon(k)|}\boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)}{\lambda + \boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\frac{\boldsymbol{\varphi}(k)}{|\varepsilon(k)|}}\right] \\ &= \frac{1}{\lambda}\left[\boldsymbol{P}(k-1) - \frac{\boldsymbol{P}(k-1)\boldsymbol{\varphi}(k)\boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)}{\lambda|\varepsilon(k)| + \boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{P}(k-1)\boldsymbol{\varphi}(k)}\right] \end{aligned} \tag{39}$$

$$\begin{aligned} \hat{\boldsymbol{\theta}}(k) &= \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{P}(k)\frac{\boldsymbol{\varphi}(k)}{|\varepsilon(k)|}\varepsilon(k) \\ &= \hat{\boldsymbol{\theta}}(k-1) + \boldsymbol{P}(k)\boldsymbol{\varphi}(k)\,\mathrm{sign}(\varepsilon(k)) \end{aligned} \tag{40}$$

As before, to initialise the estimator (38)–(40) we need a high-value matrix $\boldsymbol{P}(0) = \mathrm{diag}\,[10^5 \ \ldots \ 10^5]$, which means a high uncertainty of the initial evaluation of the parameter $\boldsymbol{\theta}$. This solution is not harmful, because thanks to the weighting mechanism ($\lambda < 1$) implemented in the LA algorithm, such starting data will be gradually eliminated from the estimator's memory.

To sum up, we have indicated above ways of solving problems with differentiability of the criterion and small divisors occurring in the iterative procedure LA.

5 | RESISTANCE TO OUTLIERS WITH EXAMPLES

In the numerical simulations presented below, we will demonstrate the announced phenomenal property of the LA method consisting in insensitivity to large outliers in the measurement data. Of practical importance is the fact that iterative and recursive LA procedures are used to solve non-trivial technical problems, such as dynamic vehicle weighing or voltage quality diagnostics in the power grid. In the summary of this section, we consider the linear programming (simplex) and gradient descent methods, which can be considered as alternative procedures for minimising the non-quadratic criterion. We will point out that taking into account the theoretical background and numerical complexity of these (competitive) methods, the advantage of the LA approach must be recognised.

5.1 | Dynamic vehicle weighing

Due to the principles used in automation, the recorded signals often resemble the step response of a damped second-order linear system. Such modelling has many applications. An example is the automatic weigh-in-motion systems installed at border crossings. The concept of dynamic weighing is designed to quickly verify whether a passing truck meets the provisions of the road traffic law in terms of the permissible axle load of the vehicle. Since the weighing process is dynamic, we must evaluate the weight based on the recorded step response of the mechanical sensor system.

The decaying oscillations of the weighing platform due to the step impact (when the car axle activates the sensor system) can be described as follows

$$y(t) = \Omega\left\{1 - \mathrm{e}^{-\zeta\omega t}\left[\cos(\beta\omega t) + \frac{\zeta}{\beta}\sin(\beta\omega t)\right]\right\} \tag{41}$$

where Ω represents the input stroke amplitude (weight carried by the axle), ω stands for the angular frequency of free oscillation (without damping), ζ is the damping constant ($0 < \zeta < 1$) and $\beta = \sqrt{1-\zeta^2}$.

The process (41) can be modelled using the classical discrete-time equation [18]

$$y(l) = \boldsymbol{\varphi}^{\mathrm{T}}(l)\,\boldsymbol{\theta} + e(l) \tag{42}$$

$$\boldsymbol{\varphi}(l) = [-y(l-1) \quad -y(l-2) \quad 1]^{\mathrm{T}} \tag{43}$$

$$\boldsymbol{\theta} = [a_1 \quad a_2 \quad b_1]^{\mathrm{T}} \tag{44}$$

where $\boldsymbol{\varphi}(l)$ and $\boldsymbol{\theta}$ are the regression and parameter vectors, respectively. The prediction error $e(l)$ is treated as an additive noise process that contaminates the measurements $y(l)$ (obtained for $l = 1 \ \ldots \ k$). Such a regression model is easily identified by the LS or LA algorithms. Based on the $\boldsymbol{\theta}$ estimate, a settled output can be found that corresponds to the estimated mass

$$\hat{\Omega} = \hat{y}(\infty) = \frac{\hat{b}_1}{1 + \hat{a}_1 + \hat{a}_2} \tag{45}$$

In the performed tests, the oscillating platform was simulated as (41) with the following parameters: $\Omega = 800$ kg,

TABLE 2 Estimates of parameters (44) and weight (45) obtained using the batch LS and iterative LA methods

	LS	LA
a_1	−1.9609	−1.9795
a_2	0.9683	0.9856
b_1	5.6729	4.9089
Ω	764.58	803.25

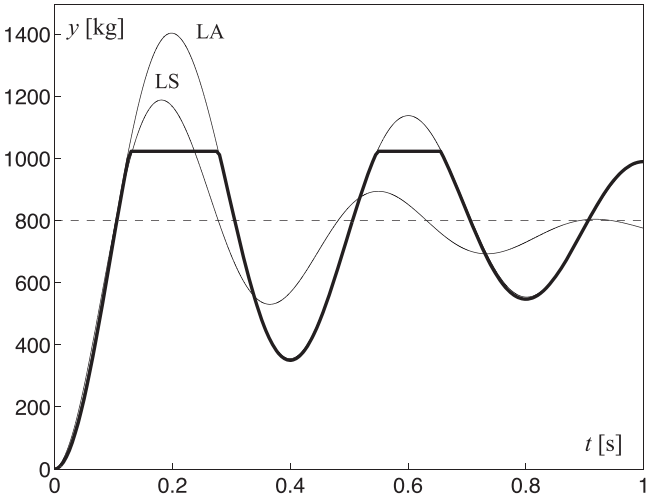

FIGURE 3 Signal (41) with destructive saturation (thick line) and its two reconstructions: LS and iterative LA (solid lines), with the actual weight setpoint equal to 800 (dashed line).

$\omega = 15.7737$ rad/s, and $\zeta = 9.1191 \times 10^{-2}$. The measurement data $y(l) = y(t)|_{t=lT}$, recorded for $l = 1 \ldots k$ ($k = 200$), were obtained with the sampling time T set to 5×10^{-3} s.

The effect of quantisation occurring in the AD converter was expressed in the form of measurements falsified with additive white noise with a uniform distribution and variance $\sigma^2 = 10^{-4}$. In addition, the effect of saturation (caused, for example, by incorrect calibration of the sensor) was simulated in order to express large measurement errors. Incorrectly selected linearity range of the sensor causes serious distortion of the recorded data. In this case, the sampled values $y(l)$ were limited to 1023 (i.e. $2^{10} - 1$).

The off-line estimate of $\boldsymbol{\theta}$ was obtained using the LS scheme (21) and the iterative LA procedure (33)–(35) using the threshold $\Delta_{\min} = 10^{-4}$. For the considered stationary dynamics, the weighting mechanism was turned off ($\lambda = 1$). The obtained estimates of parameters a_1, a_2 and b_1 together with the assessment of the weight Ω are presented in Table 2.

In addition, the analysed signal $y(t)$ was reconstructed on the basis of the vector $\boldsymbol{\theta}$ estimated by the LS and LA methods, as shown in Figure 3. The LA estimate was obtained after 24 iterative loops.

This simple example very convincingly shows the impact of large measurement errors on the quality of identification. As announced, outlier errors have a large impact on LS, while the LA procedure shows its main advantage of being highly insensitive to such parasitic phenomena. Also, the LA method seems to be more reliable in the case of identification of processes burdened with non-linear distortion of measurement data (due to typical sensor saturation, for instance).

5.2 | Evaluation of harmonics in periodic signals

Among the many important issues in the diagnosis of power grids, it is extremely important to assess the quality of the generated AC voltage. A suitable periodic voltage must maintain the required angular frequency $\omega = 2\pi \times 50$ rad/s and an amplitude of 325 V, which gives an effective voltage of 230 V. Unfortunately, with non-linear loads connected to the mains, AC voltage degradation due to other harmonics becomes real. To face this, we can implement estimation procedures to effectively assess the 'purity' of the voltage [19].

The multi-harmonic mains voltage can be shown as

$$y(t) = \sum_{i=1}^{n} \Omega_i \sin(\omega_i t + \eta_i)$$

$$= \sum_{i=1}^{n} [a_i \sin(\omega_i t) + b_i \cos(\omega_i t)] \tag{46}$$

where we have successive frequencies ω_i, $i = 1 \ldots n$: $\omega_1 = \omega$, $\omega_2 = 2\omega$, …, $\omega_n = n\omega$, while the amplitudes Ω_i of individual harmonics and the arguments η_i of their phase shift are defined as

$$\Omega_i = \sqrt{a_i^2 + b_i^2} \tag{47}$$

$$\eta_i = \operatorname{atan2}(b_i, a_i) \tag{48}$$

Assuming that the frequency ω remains constant, the process (46)–(48) can be described by a trigonometric series in discrete-time

$$y(l) = \boldsymbol{\varphi}^{\mathrm{T}}(l)\,\boldsymbol{\theta} + e(l) \tag{49}$$

$$\boldsymbol{\varphi}(l) = [\sin(\varpi l) \ \ldots \ \sin(n\varpi l) \ \ \cos(\varpi l) \ \ldots \ \cos(n\varpi l)]^{\mathrm{T}} \tag{50}$$

$$\boldsymbol{\theta} = [a_1 \ \ldots \ a_n \ \ b_1 \ \ldots \ b_n]^{\mathrm{T}} \tag{51}$$

where $\boldsymbol{\varphi}(l)$ and $\boldsymbol{\theta}$ stand for the regression and parameter vectors, respectively, $\varpi = \omega T$ is the normalised frequency, and T is the sampling time. Again, the prediction error $e(l)$ comes from additive noise that interferes with the measurements $y(l)$, for $l = 1 \ldots k$.

Note that using the LS algebraic procedure (21) to identify the parameters of the model (49)–(51) the classical Fourier formulae can be obtained, provided that the period ($2\pi/\omega$) of the considered signal (46) is a multiple of the sampling time (T) used.

TABLE 3 On-line estimates of harmonics (47) using the recursive LS and LA methods.

	LS	LA
Ω_1	64.1088 ± 1.0837	268.6253 ± 0.0258
Ω_2	89.9357 ± 0.6867	39.3352 ± 0.0106
Ω_3	111.4615 ± 0.8019	15.7391 ± 0.0151

TABLE 4 Off-line estimates of harmonics (47) using the batch LS and iterative LA methods.

	LS	LA
Ω_1	71.3807	268.7928
Ω_2	87.3943	39.3025
Ω_3	107.4741	15.6674

There are many specific situations in which large measurement errors can unexpectedly distort the correctness of the processed data. For example, in optical rotary measurements based on simple binary encoders, the reading of two consecutive values (i.e. $2^{\mathrm{NoB}} - 1$ and 2^{NoB}) can be affected by a large outlier resulting from the transient and simultaneous switching of 100/NoB bits. Similarly, random errors corrupting the transmitted data may arise due to the interference of unsecured transmission. Such distortions may also result from software errors—an example may be data conversion not thought through by the programmer (e.g. falsifying the sign bit).

These problems have been known for a long time. Various measures are used to overcome them. For example, Gray code encoders contribute to reliable optical readings when multiple bits are switched simultaneously. A good solution may be a parity bit or a checksum, which reduce the risk of undetected bit errors in the transmission. Programming errors can be partially eliminated by using more advanced compilers that can accurately identify all suspicious data conversions.

In the conducted numerical studies, the signal in the form of (46)–(48) with the parameters $n = 3$, $\Omega_1 = 270$ V, $\Omega_2 = 40$ V, $\Omega_3 = 15$ V, was simulated. Samples $y(l) = y(t)\,|_{t=lT}$, recorded for $l = 1 \ldots k$ ($k = 10^5$), were obtained with a sampling time of $T = 10^{-4}$ s and a simulation time of 10 s. The falsifying effect of quantisation induced in the AD converter was simulated as an additive uniformly distributed white noise with variance $\sigma^2 = 2.5 \times 10^{-1}$. This time, the measurement errors resulted from the incorrectly selected resolution of the ADC. Namely, for the assumed 9-bit conversion, which represents integers in the two's complement format in the range [−256, 255], an incautious attempt to read a number outside this range may wrongly interpret the overflow as a sign bit change, and thus strongly falsify the measurement. Consequently, for example, measurements 256, 257 and 258 will be machine interpreted as −256, −254 and −253, respectively.

The on-line estimate of $\boldsymbol{\theta}$ was obtained using the recursive LS algorithm (25)–(27) and the (approximate) recursive LA procedure (38)–(40). Based on the estimated parameters (51), the sought-after amplitudes in (46) were determined by means of (47). The weighting parameter was set at $\lambda = 0.9995$. This increased the efficiency of the algorithms, thanks to the gradual removal of errors from the memory of the estimators. The estimates and their standard deviations (averaged for 50 realisations) are shown in Table 3.

Moreover, based on a single period $2\pi/\omega = 2 \times 10^{-2}$ s ($k = 200$) of the sampled process (46), off-line $\boldsymbol{\theta}$ was estimated using the LS algorithm (21) and the iterative LA procedure (33)–(35),

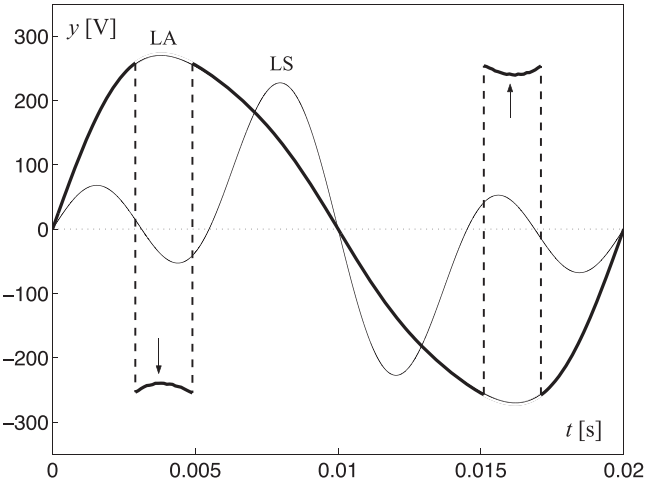

FIGURE 4 Signal (46) with numerical jumps (thick line) and its off-line reconstructions: LS and iterative LA (solid lines).

assuming $\lambda = 1$ and $\Delta_{\min} = 10^{-4}$. The LA result was obtained after 62 loops of the iterative processing (Table 4).

Reconstructions of the periodic signal (46), based on estimates obtained by the off-line LS and LA methods, are compared in Figure 4.

The conducted numerical research confirms that the LS method in its algebraic and recursive implementations is highly sensitive to obviously falsified measurement data. Meanwhile, the iterative-recursive implementations of the LA method show very high insensitivity to simulated giant outliers. It is also worth noting that the proposed recursive (on-line) LA estimation is completely reliable even though its algorithm is only approximate.

5.3 | Comparison with other numerical methods

Among other non-quadratic criteria optimisation methods, the most famous are linear programming and gradient descent algorithms. Therefore, in order to compare the developed LA procedure with such approaches, a simplex scheme and Gauss–Newton algorithm were implemented to solve the identification problems discussed above.

To adapt the simplex method to handle minimisation of the LA index (28), it is necessary to use auxiliary variables $v(l)$ such that $|y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{\theta}| \le v(l)$. Then, the following linear function

should be minimised

$$\sum_{l=1}^{k} \gamma(l)v(l) = \sum_{l=1}^{k} \lambda^{k-l} v(l) \tag{52}$$

under the following algebraic constraints ($l = 1 \ldots k$)

$$-v(l) + \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{\theta} \le y(l) \tag{53}$$

$$-v(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{\theta} \le -y(l) \tag{54}$$

It should be noted that it is rather impractical to use the simplex method here. This is due to the large number of measurements (e.g. $k = 200$) implying the creation of the same (k) number of new variables and twice more ($2k$) limitations (53), (54) when we consider the evaluation of only a few parameters (3 or 6, as is the case in the examples). Moreover, with the exponential complexity of the simplex procedure, its execution time increases enormously (and is approximately 30 times longer than in the case of the LA method).

On the other hand, since the Gauss-Newton procedure cannot be directly applied to the non-differentiable LA indicator, the idea of "smoothing" breakpoints (28) on a user-defined interval $[-\delta, \delta]$ can be applied by implementing the Huber loss function

$$f(e) = \begin{cases} 0.5e^2 & \text{for} \quad |e| \le \delta \\ \delta\ (\ |e| - 0.5\delta\ ) & \text{for} \quad |e| > \delta \end{cases} \tag{55}$$

Based on (55), the LA quality indicator was modified to

$$\mathfrak{I}(\boldsymbol{\theta}) = \sum_{l=1}^{k} \gamma(l) f(e(l)) = \sum_{l=1}^{k} \lambda^{k-l} f(y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{\theta}) \tag{56}$$

for minimisation using the Gauss–Newton scheme

$$\hat{\boldsymbol{\theta}}^{[r+1]} = \hat{\boldsymbol{\theta}}^{[r]} - \left\{ [\nabla^2_{\boldsymbol{\theta}} \mathfrak{I}(\boldsymbol{\theta})]^{-1} [\nabla_{\boldsymbol{\theta}} \mathfrak{I}(\boldsymbol{\theta})] \right\} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{[r]}} \tag{57}$$

where $\nabla_{\boldsymbol{\theta}} \mathfrak{I}(\boldsymbol{\theta})$ and $\nabla^2_{\boldsymbol{\theta}} \mathfrak{I}(\boldsymbol{\theta})$ denote the gradient and Hessian for (56), respectively.

The appearance of the Huber loss function in (56) is convenient because this modified criterion becomes locally quadratic (in the vicinity of the kink points in the LA index). In fact, the complexity $O(n^3)$ of the Gauss-Newton method is the same as that of the LA procedure, but in terms of execution time the iterative LA algorithm is slightly better (since we do not evaluate the function $f(e)$ for all errors $e(1) \ldots e(k)$ in each iteration).

The results obtained using the discussed alternative methods were compared with those obtained with the LA algorithm. For example, in the case of dynamic vehicle weight assessment (Section 5.1), the estimates of Ω ($\Omega = 800$) were as follows:

LA: 803.25, Simplex: 811.11, Gradient: 810.08.

In the assessment of harmonics in the power grid (Section 5.2), all results were also very similar.

There are obvious reasons for obtaining comparable results by all the considered procedures: (1) The simplex method is inherently suitable for minimising such piecewise-linear criteria (albeit at the cost of huge computational overhead associated with the generation of many auxiliary variables). (2) The gradient minimisation of the artfully smoothed LA index (56) is also effective in obtaining acceptable results.

In summary, we believe that there are solid reasons to consider our LA methodology superior to the other optimisation programs discussed here.

First, we found that the iterative LA procedure converges, while there is no such guarantee for gradient search. Second, the numerical difficulty of the Gauss–Newton method with the Huber-smoothed criterion (56) is still incommensurate with the numerical simplicity of our approach to the existence of breakpoints of the LA criterion using regularisation (in our LA procedure, divisors $e(l)$ close to zero are replaced by a small positive threshold $e_{\min}$). Third, we propose this approximate recursive version of the LA method with complexity $O(n^2)$, while neither the simplex method nor the Gauss-Newton scheme can take such a practical form on-line (without matrix inversion).

This section has presented the results of numerical tests indicating the unique properties of the LA method. We believe that our examples are practical, illustrative and intriguing. Other interesting aspects regarding robust identification can be found in the literature [20].

6 | CONCLUSIONS

The article develops and implements the idea of identifying parameters of processes and systems in the sense of the least sum of absolute values. The main motivation to address non-square estimation methods was the demonstrated insensitivity of LA procedures to significant disruptive phenomena, such as large measurement errors (outliers).

Unlike the classic LS algorithm, which by definition is very sensitive to this type of errors, the results of LA identification turn out to be reliable regardless of occasional outliers or other distortions in the processed data. Also when identifying physical continuous-time models, the LA index retains the proper physical meaning, while with the squared quantities that make up the LS index, the induced physical interpretation can sometimes be confusing (or incomprehensible, e.g. losses expressed in 'square dollars').

The numerical verification of the reported properties of the LA strategy was based on imaginative examples, such as weighing of vehicles in motion and diagnosis of mains voltage. The comparison of the LS and LA estimation results undoubtedly showed that the LA procedure is exceptionally well suited to processing measurements heavily contaminated with large errors.

It is important that the iterative LA is apparently more effective compared to the simplex scheme, which is characterised by burdensome exponential complexity. The numerical complexity of the gradient method is essentially equal to that of

the LA procedure, but LA is convergent, while there is no such guarantee for the Gauss-Newton algorithm. However, neither the simplex nor the gradient method can be implemented on-line.

(1) Main contribution

The convergence of the weighted iterative LA method (33)–(35), presented in Appendix A, is an important theoretical contribution to the field of process identification analysed in this paper. Other problems related to the non-differentiable LA criterion are also discussed in Appendix B. In addition, an approximate recursive implementation of the weighted LA method (38)–(40) has been presented in the form of a practical estimation scheme that can be conveniently implemented on-line.

The accuracy of the promoted LA approach was supported by numerical studies, which showed its insensitivity to harmful outliers in measurement data (including non-linear distortions of recorded signals). It is worth noting that the useful weighting mechanism, which allows tracking the time-varying parameters of the observed system, can be applied to all considered iterative-recursive forms of the LA estimator.

It may be important for industrial engineers that iterative (precise) and recursive (approximate) LA procedures can actually be combined. Namely, this can be achieved by iteratively processing LA (33)–(35), respectively, between sampling times.

However, we must be aware that the cumbersome matrix inversion underlying identification makes iterative calculations time consuming. Thus, to ensure that the proposed LA procedure 'recursive with embedded iterations' satisfies the real-time constraints (i.e. it ended iterations before the upcoming sampling time), we can relax the requirements imposed on the accuracy of the estimation by increasing the threshold $\Delta_{\min}$ accordingly.

(2) Further study

Taking into account the reported results, the further research in this area is to focus on the following issues:

(a) Consistent identification in the LA sense: Both the LS and LA methods have an asymptotic bias in the estimate θ, unless the prediction error $e(l)$ represented in the regression model (16)–(18) is zero-mean white noise. In order to make the estimation process immune to correlated noise, and thus eliminate the systematic error of parameter estimation, the idea of instrumental variables can be put into practice. Useful hints in this direction can be found in the literature [16, 21].
(b) Effective tracking of variable parameters of non-stationary systems: Proper selection of the weighting factor λ is of key importance for the correct identification of processes with variable parameters. Unfortunately, the choice of λ is most often intuitive or based on rough predictions of the type of non-stationarity of the system (e.g. assumptions about fast or gradual evolution of its parameters). To overcome this problem, the well-known idea of parallel estimation can be used [22].
(c) It can then be presumed that at least one of the estimators working in a typical 'battery' of 3 competing filters with differently tuned factors ($\lambda_1 < \lambda_2 < \lambda_3$) will be (sub-)optimally matched. The outlier-resistant LA schemes used in such a configuration will make the system even more resistant, allowing for effective change detection in the parameters of the object.
(d) Error-proof identification of non-trivial industrial objects: Due to the physical nature of industrial systems, continuous-time differential equation models [23, 24] or the more involved state-space representations [25, 26] seem to be more adequate to describe the basic process dynamics. Non-trivial means, for example, solving the problem of identifying an unknown input lag, identifying models with non-linear expressions, or handling infinite-dimensional models (e.g. represented by partial differential equations). In the literature there are solutions suitable for modelling and identification of delay systems [27, 28], distributed parameter systems [29], and specific non-linear stationary [30] and non-stationary objects [31, 32].

Finally, it is worth noting that there are modern, non-classical methods (based on neural networks or genetic algorithms) that can also be successfully used to reliably identify processes [33, 34].

AUTHOR CONTRIBUTIONS

Janusz Kozłowski: Conceptualization; methodology; validation; investigation; visualization; software; writing—original draft preparation; data curation; writing—reviewing and editing. **Zdzisław Kowalczuk**: Conceptualization; methodology; validation; investigation; writing—reviewing and editing; supervision; text correction.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

FUNDING INFORMATION

The authors did not receive any specific funding for this work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Zdzisław Kowalczuk https://orcid.org/0000-0001-9174-546X

REFERENCES

1. Byrski, W., Drapała, M., Byrski, J.: An adaptive identification method based on the modulating functions technique and exact state observers for modeling and simulation of a nonlinear MISO glass melting process. Int. J. Appl. Math. Comput. Sci. 29(4), 739–757 (2019). https://doi.org/10.2478/amcs-2019-0055
2. Suchomski, P., Kowalczuk, Z.: Analytical design of stable delta-domain generalized predictive control. Optim. Control Appl. Methods 23(5), 239–273 (2002). https://doi.org/10.1002/oca.712

3. Middleton, R.H., Goodwin, G.C.: Digital Control and Estimation. A Unified Approach. Prentice-Hall, Upper Saddle River, NJ (1990)
4. Schoukens, J.: Modeling of continuous time systems using a discrete time representation. Automatica 26(3), 579–583 (1990). https://doi.org/10.1016/0005-1098(90)90029-H
5. Kowalczuk, Z., Kozłowski, J.: Continuous-time approaches to identification of continuous-time systems. Automatica 36(8), 1229–1236 (2000). https://doi.org/10.1016/S0005-1098(00)00033-9
6. Unbehauen, H., Rao, G.P.: Identification of Continuous Systems. North Holland, Amsterdam (1987)
7. Ljung, L.: System Identification: Theory for the User. Prentice-Hall, Upper Saddle River, NJ (1987)
8. Schlossmacher, E.: An iterative technique for absolute deviations curve fitting. J. Am. Stat. Assoc. 68(344), 857–859 (1973). https://doi.org/10.1080/01621459.1973.10481436
9. Straszak, D., Vishnoi, N.K.: Iteratively reweighted least squares and slime mold dynamics: connection and convergence. Math. Program. 194 (1-2), 685–717 (2021). https://doi.org/10.1007/s10107-021-01644-z
10. Gentle, J.E.: Least absolute values estimation: an introduction. Commun. Stat. - Simul. Comput. 6(4), 313–328 (1977). https://doi.org/10.1080/03610917708812047
11. Janiszowski, K.B.: Towards estimation in the sense of the least sum of absolute errors. IFAC Proc. 31(20), 605–610 (1998). https://doi.org/10.1016/S1474-6670(17)41862-3
12. Eakambaram, S., Rex Irudhaya Raj, A.: Robust regression using least absolute deviations method. Int. J. Mech. Eng. 7(5), 53–57 (2022)
13. Thanoon, F.H.: Robust regression by least absolute deviations method. Int. J. Stat. Appl. 5(3), 109–112 (2015). https://doi.org/10.5923/j.statistics.20150503.02
14. Kowalczuk, Z., Kozłowski, J.: Non-quadratic quality criteria in parameter estimation of continuous-time models. IET Control Theory Appl. 5(13), 1494–1508 (2011). https://doi.org/10.1049/iet-cta.2010.0310
15. Zhang, G., Shi, Y., Sheng, Y.: Least absolute deviation estimation for uncertain vector autoregressive model with imprecise data. Int. J. Uncertainty Fuzziness Knowledge Based Syst. 31(3), 353–370 (2023). https://doi.org/10.1142/S0218488523500186
16. Kozłowski, J., Kowalczuk, Z.: Resistant to correlated noise and outliers discrete identification of continuous non-linear non-stationary dynamics objects. In: Kowalczuk, Z. (Ed.). Intelligent and Safe Computer Systems in Control and Diagnostics. pp. 317–327. Springer Nature AG, Cham (2023). https://doi.org/10.1007/978-3-031-16159-9_26
17. Sagara, S., Yang, Z.J., Wada, K.: Identification of continuous systems using digital low-pass filters. Int. J. Syst. Sci. 22(7), 1159–1176 (1991). https://doi.org/10.1080/00207729108910693
18. Kozłowski, J., Kowalczuk, Z.: Intelligent monitoring the vertical dynamics of wheeled inspection vehicles. IFAC-PapersOnLine. 52(8), 251–256 (2019). https://doi.org/10.1016/j.ifacol.2019.08.079
19. Kozłowski, J., Kowalczuk, Z.: Robust to measurement faults parameter estimation algorithms in issues on systems diagnosis. In: Kowalczuk, Z., Wiszniewski, B. (Eds.). Automation and Informatics: Information technologies – Diagnostics. pp. 221–240. PWNT, Gdańsk, Poland (2007)
20. Kozłowski, J., Kowalczuk, Z.: Identification of continuous systems – practical issues of insensitivity to perturbations. In: Kościelny, J.M., Syfert, M., Sztyber, A. (Ed.). Advanced Solutions in Diagnostics and Fault Tolerant Control. pp. 180–191. Springer IP AG, Cham (2018). https://doi.org/10.1007/978-3-319-64474-5_15
21. Soderstrom, T., Stoica, P.: Comparison of some instrumental variable methods – consistency and accuracy aspects. Automatica 17(1), 101–115 (1981). https://doi.org/10.1016/0005-1098(81)90087-X
22. Kowalczuk, Z.: Competitive identification for self-tuning control: robust estimation design and simulation experiments. Automatica 28(1), 193–201 (1992). https://doi.org/10.1016/0005-1098(92)90021-7
23. Johansson, R.: Identification of continuous-time models. IEEE Trans. Signal Process. 42(4), 887–897 (1994). https://doi.org/10.1109/78.285652
24. Unbehauen, H., Rao, G.P.: Continuous-time approaches to system identification – a survey. Automatica 26(1), 23–35 (1990). https://doi.org/10.1016/0005-1098(90)90155-B
25. Kowalczuk, Z.: On discretization of continuous-time state-space models: A stable normal approach. IEEE Trans. Circuits Syst. 38(1), 1460–1477 (1991). https://doi.org/10.1109/31.108500
26. Young, P.: Parameter estimation for continuous-time models – a survey. Automatica 17(1), 23–39 (1981). https://doi.org/10.1016/0005-1098(81)90082-0
27. Kozłowski, J., Kowalczuk, Z.: On-line parameter and delay estimation of continuous-time dynamic systems. Int. J. Appl. Math. Comput. Sci. 25(2), 223–232 (2015). https://doi.org/10.1515/amcs-215-0017
28. Zhao, Z.Y., Sagara, S.: Consistent estimation of time delay in continuous-time systems. Trans. Soc. Instrum. Control Eng. 27(1), 64–69 (1991). https://doi.org/10.9746/sicetr1965.27.64
29. Sagara, S., Zhao, Z.Y.: Identification of system parameters in distributed parameter systems. In: Proceedings of the 11th IFAC World Congress. Tallinn, Estonia, pp. 471–476 (1990). https://doi.org/10.1016/S1474-6670(17)51960-6
30. Inoue, K., Kumamaru, K., Nakahashi, Y., Nakamura, H., Uchida, M.: A quick identification method of continuous-time nonlinear systems and its application to power plant control. In: Proceedings of the 10th IFAC Symposium on System Identification. Copenhagen, Denmark, pp. 283–288 (1994). https://doi.org/10.1016/S1474-6670(17)47729-9
31. Kozłowski, J., Kowalczuk, Z.: Discrete identification of continuous nonlinear and non-stationary dynamical systems that is insensitive to noise correlation and measurement outliers. Arch. Control Sci. 33(2), 391–411 (2023). https://doi.org/10.24425/acs.2023.146281
32. Schoukens, J., Ljung, L.: Nonlinear system identification: a user-oriented road map. IEEE Control Syst. Mag. 39(6), 28–99 (2019). https://doi.org/10.1109/MCS.2019.2938121
33. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, MA (1989)
34. Uciński, D., Patan, M.: Sensor network design for the estimation of spatially distributed processes. Int. J. Appl. Math. Comput. Sci. 20(3), 459–481 (2010). https://doi.org/10.2478/v10006-010-0034-2

How to cite this article: Kozłowski, J., Kowalczuk, Z.: Iterative-recursive estimation of parameters of regression models with resistance to outliers on practical examples. IET Control Theory Appl. 18, 1099–1113 (2024). https://doi.org/10.1049/cth2.12628

APPENDIX A: Convergence of the LA method

Thesis: The sequence $J^{[r]} = J(\hat{\boldsymbol{\theta}}^{[r]})$ of the index values (28) determined in consecutive iterations ($r = 0, 1, \ldots$) in accordance with the standard iterative estimation scheme (33)–(35) is decreasing: $J^{[r+1]} - J^{[r]} < 0$.

Proof. The LA quality index (28) is given by

$$J(\boldsymbol{\theta}) = \sum_{l=1}^{k} \gamma(l) \; \big| e(l) \big| = \sum_{l=1}^{k} \gamma(l) \; \big| y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{\theta} \big| \qquad \text{(A1)}$$

where the useful weighting factor $\gamma(l) > 0$ can be represented by a classical exponential window $\gamma(l) = \lambda^{k-l}$. The iterative ($r = 0, 1, \ldots$) LA estimate minimising the index (A1) follows from

$$\hat{\boldsymbol{\theta}}^{[r+1]} = \hat{\boldsymbol{\theta}}^{[r]} + \boldsymbol{R}^{-1}(k)\boldsymbol{\psi}(k) \qquad \text{(A2)}$$

where the error $\hat{e}^{[r]}(l)$, Hessian $\boldsymbol{R}(k)$ (called 'information matrix'), and gradient '$-\boldsymbol{\psi}(k)$' are given by

$$\hat{e}^{[r]}(l) = y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\hat{\boldsymbol{\theta}}^{[r]} \tag{A3}$$

$$\boldsymbol{R}(k) = \sum_{l=1}^{k} \gamma(l) \frac{\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)}{|\hat{e}^{[r]}(l)|} \tag{A4}$$

$$\boldsymbol{\psi}(k) = \sum_{l=1}^{k} \gamma(l) \frac{\boldsymbol{\varphi}(l)\hat{e}^{[r]}(l)}{|\hat{e}^{[r]}(l)|} \tag{A5}$$

The matrix $\boldsymbol{R}(k)$ is positive definite, because for a consistent vector ($||\boldsymbol{v}|| \neq 0$) the corresponding quadratic form is positive, provided $\hat{e}^{[r]}(l) \neq 0$ (for $l = 1, \ldots k$)

$$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}(k)\boldsymbol{v} = \sum_{l=1}^{k} \gamma(l) \frac{[\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\varphi}(l)]^2}{|\hat{e}^{[r]}(l)|} > 0 \tag{A6}$$

By introducing $\boldsymbol{P}(k) = \boldsymbol{R}^{-1}(k)$, the iterative Equation (A2) can be rearranged as

$$\boldsymbol{\varphi}^{\mathrm{T}}(l)\hat{\boldsymbol{\theta}}^{[r+1]} = \boldsymbol{\varphi}^{\mathrm{T}}(l)[\hat{\boldsymbol{\theta}}^{[r]} + \boldsymbol{P}(k)\boldsymbol{\psi}(k)] \tag{A7}$$

$$y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\hat{\boldsymbol{\theta}}^{[r+1]} = y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)[\hat{\boldsymbol{\theta}}^{[r]} + \boldsymbol{P}(k)\boldsymbol{\psi}(k)] \tag{A8}$$

$$\hat{e}^{[r+1]}(l) = \hat{e}^{[r]}(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \tag{A9}$$

$$\gamma(l)\,|\hat{e}^{[r+1]}(l)| = \gamma(l)\,|\hat{e}^{[r]}(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k)| \tag{A10}$$

By summing over l on both sides of (A10), we get

$$\sum_{l=1}^{k} \gamma(l)\,|\hat{e}^{[r+1]}(l)| = \sum_{l=1}^{k} \gamma(l)\,|\hat{e}^{[r]}(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k)| \tag{A11}$$

Let $J^{[r]} = J(\hat{\boldsymbol{\theta}}^{[r]})$ be the value of the index (A1) computed in the r-th iteration. Assuming that the values of $\hat{e}^{[r]}(l)$ are always non-zero (for $l = 1 \ldots k$), we can rewrite Equation (A11) as follows

$$\begin{aligned} J^{[r+1]} &= \sum_{l=1}^{k} \gamma(l)\,|\hat{e}^{[r+1]}(l)| \\ &= \sum_{l=1}^{k} \sqrt{\frac{|\hat{e}^{[r]}(l)|}{\gamma^{-1}(l)}} \left| \frac{\hat{e}^{[r]}(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k)}{\sqrt{\gamma^{-1}(l)|\hat{e}^{[r]}(l)|}} \right| \end{aligned} \tag{A12}$$

Let us now recall the Schwarz inequality, which for any functions $u(l)$ and $w(l)$ gives

$$\left[\sum_{l=1}^{k} u(l)w(l)\right]^2 \leq \left[\sum_{l=1}^{k} u^2(l)\right] \left[\sum_{l=1}^{k} w^2(l)\right] \tag{A13}$$

Using (A12) and (A13), the square of $J^{[r+1]}$ takes the form

$$\begin{aligned} (J^{[r+1]})^2 &= \left[\sum_{l=1}^{k} \sqrt{\frac{|\hat{e}^{[r]}(l)|}{\gamma^{-1}(l)}} \left| \frac{\hat{e}^{[r]}(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k)}{\sqrt{\gamma^{-1}(l)|\hat{e}^{[r]}(l)|}} \right| \right]^2 \\ &\leq \left\{\sum_{l=1}^{k} \left[\sqrt{\frac{|\hat{e}^{[r]}(l)|}{\gamma^{-1}(l)}}\right]^2\right\} \\ &\quad \times \left\{\sum_{l=1}^{k} \left[\frac{\hat{e}^{[r]}(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k)}{\sqrt{\gamma^{-1}(l)|\hat{e}^{[r]}(l)|}}\right]^2\right\} \\ &= \left\{\sum_{l=1}^{k} \gamma(l)\,|\hat{e}^{[r]}(l)|\right\} \\ &\quad \times \left\{\sum_{l=1}^{k} \gamma(l) \frac{[\hat{e}^{[r]}(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k)]^2}{|\hat{e}^{[r]}(l)|}\right\} \\ &= J^{[r]} \sum_{l=1}^{k} \gamma(l) \frac{[\hat{e}^{[r]}(l)]^2}{|\hat{e}^{[r]}(l)|} \\ &\quad - 2J^{[r]} \left[\sum_{l=1}^{k} \gamma(l) \frac{\hat{e}^{[r]}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)}{|\hat{e}^{[r]}(l)|}\right] \boldsymbol{P}(k)\boldsymbol{\psi}(k) \\ &\quad + J^{[r]} \sum_{l=1}^{k} \gamma(l) \frac{[\boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k)]^2}{|\hat{e}^{[r]}(l)|} \\ &= J^{[r]} \sum_{l=1}^{k} \gamma(l)\,|\hat{e}^{[r]}(l)| \\ &\quad - 2J^{[r]} \left[\sum_{l=1}^{k} \gamma(l) \frac{\hat{e}^{[r]}(l)\boldsymbol{\varphi}(l)}{|\hat{e}^{[r]}(l)|}\right]^{\mathrm{T}} \boldsymbol{P}(k)\boldsymbol{\psi}(k) \\ &\quad + J^{[r]} \sum_{l=1}^{k} \gamma(l) \frac{[\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}^{\mathrm{T}}(k)\boldsymbol{\varphi}(l)][\boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k)]}{|\hat{e}^{[r]}(l)|} \\ &= (J^{[r]})^2 - 2J^{[r]}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \\ &\quad + J^{[r]}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}^{\mathrm{T}}(k) \left[\sum_{l=1}^{k} \gamma(l) \frac{\boldsymbol{\varphi}(l)\boldsymbol{\varphi}^{\mathrm{T}}(l)}{|\hat{e}^{[r]}(l)|}\right] \boldsymbol{P}(k)\boldsymbol{\psi}(k) \\ &= (J^{[r]})^2 - 2J^{[r]}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \\ &\quad + J^{[r]}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}^{\mathrm{T}}(k)\boldsymbol{P}^{-1}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \\ &= (J^{[r]})^2 - 2J^{[r]}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \\ &\quad + J^{[r]}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \\ &= (J^{[r]})^2 - J^{[r]}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \end{aligned} \tag{A14}$$

In the calculations above, the evident tautology and the symmetry of the covariance matrix were used, that is,

$$\boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{P}(k)\boldsymbol{\psi}(k) = \boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}^{\mathrm{T}}(k)\boldsymbol{\varphi}(l) \tag{A15}$$

$$\boldsymbol{P}^{\mathrm{T}}(k) = \boldsymbol{P}(k) = \boldsymbol{R}^{-1}(k) \tag{A16}$$

As a result, we show that

$$\left(J^{[r+1]}\right)^2 \leq \left(J^{[r]}\right)^2 - J^{[r]}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \tag{A17}$$

what is equivalent to

$$\left(J^{[r+1]}\right)^2 - \left(J^{[r]}\right)^2 \leq -J^{[r]}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \tag{A18}$$

Finally, using $(a^2 - b^2) = (a - b)(a + b)$, we get

$$J^{[r+1]} - J^{[r]} \leq \frac{-J^{[r]}}{J^{[r+1]} + J^{[r]}}\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) \tag{A19}$$

Conclusions: As shown in (A6), the information matrix $\boldsymbol{R}(k)$ is positive definite, and therefore the matrix $\boldsymbol{P}(k) = \boldsymbol{R}^{-1}(k)$ is also positive definite. Therefore, by definition, $\boldsymbol{\psi}^{\mathrm{T}}(k)\boldsymbol{P}(k)\boldsymbol{\psi}(k) > 0$, provided the gradient $||\boldsymbol{\psi}(k)|| \neq 0$. Of course, since the index (A1) is positive ($J^{[r]} > 0$), inequality (A19) is strict: $J^{[r+1]} - J^{[r]} < 0$.

However, if the modulus of $\boldsymbol{\psi}(k)$ is zero ($||\boldsymbol{\psi}(k)|| = 0$), the iterative Equation (A2) leads to $||\hat{\boldsymbol{\theta}}^{[r+1]}|| = ||\hat{\boldsymbol{\theta}}^{[r]}||$ and there is no further progress in iterative minimisation: $J^{[r+1]} = J^{[r]}$. This corresponds to the flat zone of (A1).

By definition, the convex LA criterion (A1) is lower bounded. Since each decreasing and lower bounded sequence $J^{[r]}$ ($r = 0, 1, \ldots$) is convergent, we conclude that the iterative method (A2)–(A5) minimises (A1).

Importantly, the iterative LA method converges also in the case of the weighing mechanism $\gamma(l)$ used in the indicator (A1), as long as the appropriate weighing sequence (not only in the popular version with the exponential profile λ^{k-l}) meets the condition $\gamma(l) > 0$.

Remark: It should be noted that in the presented reasoning all values of the prediction error $\hat{e}^{[r]}(l)$, for $l = 1 \ldots k$, were assumed as non-zero. This is necessary because the calculation of $\boldsymbol{R}(k)$ and $\boldsymbol{\psi}(k)$ involves dividing by the absolute value of this error.

Therefore, to avoid the numerical problem of small divisors, the values of $|\hat{e}^{[r]}(l)|$ close to zero should be replaced by a fixed positive value $e_{\min}$, which functions also as a threshold in the estimation algorithm.

APPENDIX B: LA quality criterion analysis

The quality index LA (6) can be a unimodal function (with a unique minimum) or a multimodal function with a flat zone. To illustrate these cases, we consider the data collected in Tables B1 and B2.

TABLE B1 1D linear regression measurement data leading to the LA functional with a unique minimum.

l	1	2	3
$\phi(l)$	1.0	1.5	2.0
$y(l)$	1.4	3.5	6.5

TABLE B2 1D linear regression measurement data leading to the LA functional with a flat zone.

l	1	2	3
$\phi(l)$	1.0	1.5	2.5
$y(l)$	1.2	3.5	7.5

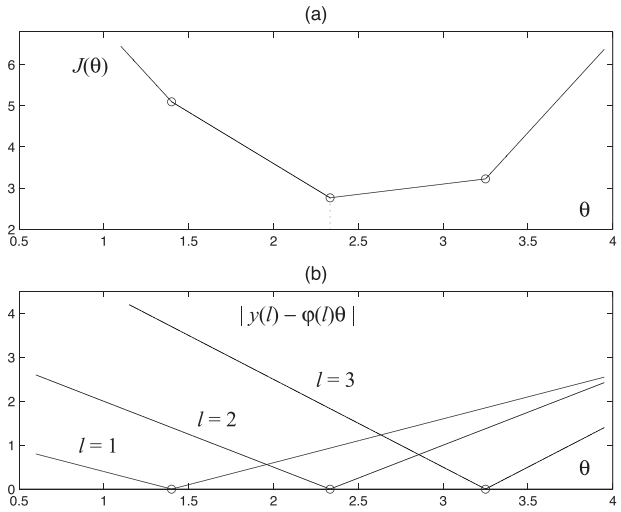

FIGURE B1 Construction of the LA functional in 1D with a unique minimum based on the data in Table B1.

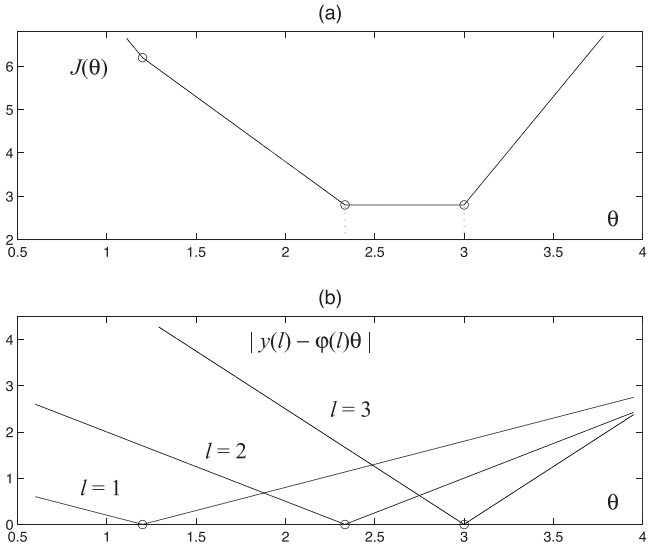

FIGURE B2 Construction of the waveform of the LA indicator with a 1D flat zone based on data from Table B2.

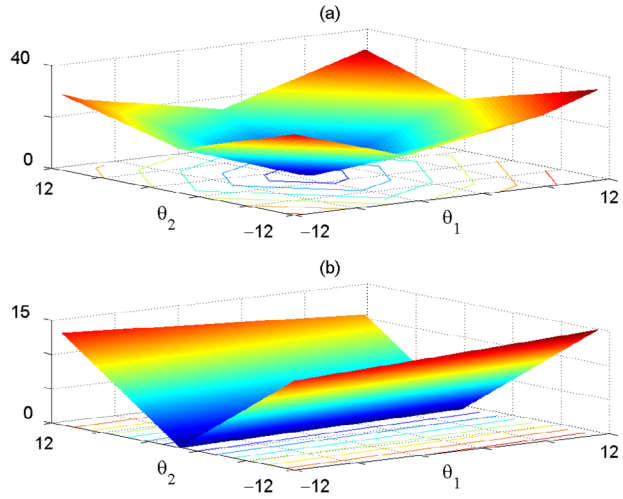

FIGURE B3 Creation of the LA indicator in 2D according to data from Table B3: (a) the resulting shape of the criterion J and (b) a selected single 'angle bar' representing a partial J_1.

The LA criterion (6) is given by the sum of the absolute value terms $|y(l) - \boldsymbol{\phi}(l)\boldsymbol{\theta}|$, where $l = 1, 2, 3$. The final piecewise-linear functional (Figure B1a) is formed as a composition of ingredients or members (Figure B1b), which are absolute values. This gives us a unique minimum for θ as $\hat{\theta} = y(2)/\boldsymbol{\phi}(2) \approx 2.33$.

In the second case, shown in Figure B2 and based on the measurement data from Table B2, a flat zone has been shaped in the range of the searched parameter $\hat{\theta} \in [y(2)/\boldsymbol{\phi}(2), y(3)/\boldsymbol{\phi}(3)]$, exactly $\hat{\theta} \in [2.33, 3]$.

Note that in the case of the one-variable criterion (6), the coordinates of the kink points are determined as $y(k)/\boldsymbol{\phi}(k)$, while the slopes of the single LA components $|y(l) - \boldsymbol{\phi}(l)\theta|$ are $\pm\boldsymbol{\phi}(k)$, as shown in Figures B1b and B2b.

TABLE B3 Measurement data used in the identification of a two-parameter (2D) model (16)–(18).

l	1	2	3
$\boldsymbol{\varphi}^{\mathrm{T}}(l)$	[–0.1 1.1]	[–1.2 0.2]	[–0.4 –0.3]
$y(l)$	1.0	1.5	0.5

Thus, the slopes of the segments of the piecewise-linear functions LA (Figures B1a and B2a) result from a simple accumulation of all components $\pm\boldsymbol{\phi}(k)$. Taking into account the data from Table B2, the inclinations of subsequent segments are determined as follows:

$-1-1.5-2.5 = -5$, $1-1.5-2.5 = -3$, $1+1.5-2.5 = 0$, $1+1.5+2.5 = 5$.

The third slope is equal to zero, which is still a realistic case of a flat minimum (Figure B2a) of criterion (6). Such an event can easily occur in finite precision calculations.

It is worth mentioning that the demonstrated compensation effect (zero slope) is typical for DACs using the first-order interpolation.

Samples of selected effects of the LA criterion (28) are shown in Figure B3. The regression vector (17) has two coordinates (2D) and the measurement data are given in Table B3. The individual components $|y(l) - \boldsymbol{\varphi}^{\mathrm{T}}(l)\boldsymbol{\theta}|$, $l = 1, 2, 3$, are represented by 'angle bars' as shown in Figure B3b, considering only the first column of data in Table B3. In this case (J_1), due to the dominance of the second coordinate of the regression vector, the partial criterion J_1 depends mainly on the second coordinate (with its optimum at $\theta_2 = 1/1.1 = 0.91$). Whereas the resulting LA criterion J of a 'piecewise-planar' form for the three observations ($l = 1, 2, 3$), is shown in Figure B3a.