Accepted manuscript of: M. Niedźwiecki and A. Gańcza, "Optimally regularized local basis function approach to identification of time-varying systems," 2022 IEEE 61st Conference on Decision and Control (CDC), Cancun, Mexico, 2022, pp. 227-234, doi: 10.1109/CDC51059.2022.9992328.

Optimally regularized local basis function approach to identification of time-varying systems

Maciej Niedźwiecki and Artur Gańcza

***Abstract*— Accurate identification of stochastic systems with fast-varying parameters is a challenging task which cannot be accomplished using model-free estimation methods, such as weighted least squares, which assume only that system coefficients can be regarded as locally constant. The current state of the art solutions are based on the assumption that system parameters can be locally approximated by a linear combination of appropriately chosen basis functions. The paper shows that when the internal correlation structure of estimated parameters is known, the tracking performance of the local basis function estimation algorithms can be further improved by means of regularization. The optimal form of the regularization matrix is derived analytically and it is shown that the best settings of the regularized algorithm can be determined in the computationally efficient way using cross-validation.**

I. Introduction

Consider the problem of identification of a time-varying finite impulse response (FIR) system governed by

$$y(t) = \sum_{j=1}^{n} \theta_j^*(t) u(t-j+1) + e(t) = \boldsymbol{\theta}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t) + e(t) \quad (1)$$

where $t = \ldots, -1, 0, 1, \ldots$ denotes discrete (normalized) time, $y(t)$ denotes the complex-valued output signal, $\boldsymbol{\varphi}(t) = [u(t), \ldots, u(t-n+1)]^{\mathrm{T}}$ denotes regression vector made up of past values of the complex-valued input signal, $\boldsymbol{\theta}(t) = [\theta_1(t), \ldots, \theta_n(t)]^{\mathrm{T}}$ denotes the vector of time-varying system impulse response coefficients, and $\{e(t)\}$ denotes white noise independent of $\{u(t)\}$ and $\{\boldsymbol{\theta}(t)\}$. The symbol $*$ denotes complex conjugate and H – complex conjugate transpose (Hermitian transpose).

The FIR model (1) is used, among others, to approximate nonstationary communication channels, both terrestial and underwater [1], [2]. Channel identification, i.e., estimation of its impulse response, is needed, for example, for equalization purposes – recovery of the transmitted signal $\{u(t)\}$ from the received sequence $\{y(t)\}$ [1].

When system coefficients vary slowly with time, their estimation can be carried out using the time-localized versions of the least squares or maximum likelihood approach [3] - [5]. The corresponding estimation algorithms, such as weighted least squares or weighted maximum likelihood, are not based on any explicit model of parameter variation – it is only assumed that system parameters can be regarded as "locally constant", i.e., that the system is locally stationary. In the case of fast varying systems such a simple estimation strategy fails because the achievable estimation accuracy is not sufficient to guarantee satisfactory operation of the underlying model-based decision system [6].

Fast parameter changes can be tracked successfully if the system model (1) is extended with an explicit model (hypermodel) of parameter variation. When such a hypermodel is stochastic, the problem of parameter estimation can be reformulated as a problem of filtering/smoothing in the appropriately defined state space. It can then be solved using the Kalman filtering methodology [7] - [9].

An alternative solution, pursued in this paper, is based on adopting a deterministic hypermodel of parameter changes. In this approach each parameter trajectory is locally approximated by a linear combination of known functions of time, called basis functions (BF) [10] - [15].

In the majority of studies devoted to the BF approach, basis functions are used to generate interval estimates of parameter trajectories. Recently a new class of identification algorithms was described, which combines the BF approach with the local estimation technique [16], [17]. The proposed local basis function (LBF) estimators provide a sequence of point estimates of system parameters corresponding to different locations of a sliding analysis window of a fixed width. As shown in [16], such a point approach yields more accurate estimates than the interval one. In the follow-up paper [18] it was shown that the tracking performance of the local basis function estimation algorithms can be further improved by means of regularization. The two-stage regularization scheme proposed in [18] does not assume any specific knowledge of the internal correlation structure of estimated parameters. In the current submission it will be shown that identification results can be further improved if such a statistical insight is available, which is the case in some (e.g. telecommunication) applications. The optimal form of the regularization matrix is derived analytically and it is shown that the best settings of the regularized LBF algorithm (the width of the local analysis interval, the number of basis functions) can be determined in a computationally efficient way using the leave-one-out cross-validation approach.

II. Local basis function estimators

In the LBF approach the time-varying system parameters are approximated, in the sliding analysis window $T(t) = [t-k, t+k]$ of width $K = 2k+1$, by a linear combination of known linearly independent functions of time

*This work was partially supported by the National Science Center under the agreement UMO-2018/29/B/ST7/00325. Both authors are with the Gdańsk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Department of Automatic Control, ul. Narutowicza 11/12, Gdańsk, Poland: artgancz@student.pg.edu.pl, maciekn@eti.pg.edu.pl

$f_1(i),\ldots,f_m(i), i \in I_k = [-k,k]$, called basis functions. Typical choices of basis functions are powers of time (local Taylor expansion) or harmonic functions (local Fourier expansion). In the case considered we will assume that each coefficient of the estimated impulse response can be expressed in the form

$$\begin{aligned} \theta_j(t+i) &= \mathbf{f}^{\mathrm{T}}(i)\boldsymbol{\alpha}_j(t), \quad i \in I_k \\ j &= 1,\ldots,n \end{aligned} \tag{2}$$

where $\mathbf{f}(i) = [f_1(i),\ldots,f_m(i)]^{\mathrm{T}}$ and $\boldsymbol{\alpha}_j(t)$ denotes the vector of complex-valued basis expansion coefficients. The hypermodel (2) can be expressed in a more compact form

$$\boldsymbol{\theta}(t+i) = \mathbf{F}(i)\boldsymbol{\alpha}(t), \quad i \in I_k \tag{3}$$

where

$$\mathbf{F}(i) = \mathbf{I}_n \otimes \mathbf{f}^{\mathrm{T}}(i), \quad \boldsymbol{\alpha}(t) = [\boldsymbol{\alpha}_1^{\mathrm{T}}(t),\ldots,\boldsymbol{\alpha}_n^{\mathrm{T}}(t)]^{\mathrm{T}} \tag{4}$$

and $\otimes$ denotes the Kronecker product of the respective vectors and/or matrices.

Some caution is needed when interpreting the hypermodel (3). According to (3), the vector of hyperparameters $\boldsymbol{\alpha}$ is assumed to be *constant* in the entire analysis window $T(t) = [t-k, t+k]$. However, since the value of $\boldsymbol{\alpha}$ may change along with the position of the window $T(t)$, it is written down as a function of time (note that the two statements made above are *not* contradictory). The estimation of $\boldsymbol{\alpha}$ will be carried out independently for each position of the sliding analysis window, i.e., it will be repeated for consecutive values of t, which is typical of local estimation frameworks.

Denote by $w(i), i \in I_k, w(0) = 1$, a symmetric, nonnegative, bell-shaped window which will be used to put more emphasis on data gathered at instants close to t than on instants far from t. For convenience, but without any loss of generality, we will assume that the adopted basis functions are w-orthonormal, namely

$$\sum_{i=-k}^{k} w(i)\mathbf{f}(i)\mathbf{f}^{\mathrm{T}}(i) = \mathbf{I}_m \ . \tag{5}$$

where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. Orthonormalization of any set of basis functions can be carried out sequentially using the well-known Gram-Schmidt procedure [5].

Combining (1) with (3), one arrives at

$$y(t+i) = \boldsymbol{\alpha}^{\mathrm{H}}(t)\boldsymbol{\psi}(t,i) + e(t+i), \quad i \in I_k \tag{6}$$

where $\boldsymbol{\psi}(t,i) = \boldsymbol{\varphi}(t+i) \otimes \mathbf{f}(i)$ denotes the generalized regression vector. The LBF estimates of $\boldsymbol{\theta}(t)$ were defined in [16] in the form

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}^{\mathrm{LBF}}(t) &= \arg\min_{\boldsymbol{\alpha}} \sum_{i=-k}^{k} w(i)|y(t+i) - \boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\psi}(t,i)|^2 \\ &= \mathbf{P}^{-1}(t)\mathbf{p}(t) \\ \widehat{\boldsymbol{\theta}}^{\mathrm{LBF}}(t) &= \mathbf{F}_0\widehat{\boldsymbol{\alpha}}^{\mathrm{LBF}}(t) \end{aligned} \tag{7}$$

where

$$\begin{aligned} \mathbf{P}(t) &= \sum_{i=-k}^{k} w(i)\boldsymbol{\psi}(t,i)\boldsymbol{\psi}^{\mathrm{H}}(t,i) \\ \mathbf{p}(t) &= \sum_{i=-k}^{k} w(i)y^*(t+i)\boldsymbol{\psi}(t,i) \end{aligned} \tag{8}$$

and $\mathbf{F}_0 = \mathbf{F}(0) = \mathbf{I}_n \otimes \mathbf{f}_0^{\mathrm{T}}$, $\mathbf{f}_0 = \mathbf{f}(0)$.

III. Regularized LBF estimators

Regularization is a technique which was originally introduced to solve ill-conditioned inverse problems [19]. As shown later, regularization also allows one to improve the bias-variance trade-off of the applied estimation schemes, and hence – to increase their accuracy [20]. In the approach discussed in this paper both aspects of regularization will be taken advantage of. The idea is to add to the minimized cost function a term, often referred to as a regularizer, which reduces the norm of the solution. In agreement with this principle, we will introduce the L_2 regularizer of the form

$$||\boldsymbol{\theta}(t)||^2_{\mathbf{R}} = \boldsymbol{\theta}^{\mathrm{H}}(t)\mathbf{R}\boldsymbol{\theta}(t) = \boldsymbol{\alpha}^{\mathrm{H}}(t)\mathbf{F}_0^{\mathrm{T}}\mathbf{R}\mathbf{F}_0\boldsymbol{\alpha}(t) \tag{9}$$

where $\mathbf{R} = \mathbf{D}^{\mathrm{H}}\mathbf{D} > 0$ denotes the $n \times n$ positive definite regularization matrix. Note that such regularization, different from that applied in [18], penalizes the norm of $\boldsymbol{\theta}(t)$, the estimation of which is a real purpose of system identification, and only indirectly penalizes the norm of the vector of hyperparameters $\boldsymbol{\alpha}(t)$, which is not of our primary interest. The regularized LBF estimators (RLBF) will be defined in the form

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R}) = \arg\min_{\boldsymbol{\alpha}} \Bigg\{ &\sum_{i=-k}^{k} w(i)|y(t+i) - \boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\psi}(t,i)|^2 \\ &+ ||\boldsymbol{\alpha}||^2_{\mathbf{F}_0^{\mathrm{T}}\mathbf{R}\mathbf{F}_0} \Bigg\} = \mathbf{S}^{-1}(t)\mathbf{p}(t) \\ \widehat{\boldsymbol{\theta}}^{\mathrm{RLBF}}(t|\mathbf{R}) &= \mathbf{F}_0\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R}) \end{aligned} \tag{10}$$

where

$$\begin{aligned} \mathbf{S}(t) &= \mathbf{P}(t) + \mathbf{F}_0^{\mathrm{T}}\mathbf{R}\mathbf{F}_0 = \mathbf{P}(t) + \mathbf{B}^{\mathrm{H}}\mathbf{B} \\ \mathbf{B} &= \mathbf{D}\mathbf{F}_0 = \mathbf{D} \otimes \mathbf{f}_0^{\mathrm{T}}. \end{aligned} \tag{11}$$

IV. Computational complexity of LBF and RLBF algorithms

Since LBF and RLBF estimates are evaluated in the sliding window mode, i.e., computations are repeated for every new location t of the analysis window $T(t)$, the computational burden is high. It can be lowered if the applied window $w(i)$ and the vector of basis functions $\mathbf{f}(i)$ are recursively computable. Taking advantage of this property, one can easily derive recursive algorithms for computation of the $mn \times 1$ vector $\mathbf{p}(t)$ and the $mn \times mn$ generalized regression matrix $\mathbf{P}(t)$, needed to evaluate LBF and RLBF estimates – for more details see [16]. In this way the computational cost of evaluation of $\mathbf{p}(t)$ and $\mathbf{P}(t)/\mathbf{S}(t)$ can be lowered from $O(mnK)$ and $O(m^2n^2K)$ flops (multiply-add operations) per time update, to $O(mn)$ and $O(m^2n^2)$ flops, respectively,

i.e., it becomes independent of the window width $K = 2k+1$.

Once the quantities $\mathbf{p}(t)$ and $\mathbf{P}(t)/\mathbf{S}(t)$ are updated, the LBF or RLBF estimates can be evaluated by solving the corresponding systems of linear equations $\mathbf{P}(t)\widehat{\boldsymbol{\alpha}}^{\mathrm{LBF}}(t) = \mathbf{p}(t)$ or $\mathbf{S}(t)\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t) = \mathbf{p}(t)$, respectively. In both cases the associated computational burden is roughly equal to $\mathrm{O}(m^3n^3)$ flops. We note, however, that this cost can be significantly reduced to as little as $3mn$ flops per time update if the iterative dichotomous coordinate descent (DCD) technique, described in [21], is applied. The DCD approach, which is a low-complexity computational scheme suitable for finite precision implementations, was originally proposed to speed up evaluation of finite-memory recursive least squares estimates. Later it was successfully applied to computation of LBF estimates – see [6].

As demonstrated in [17], when the number of estimated hyperparameters mn becomes comparable with (is not much smaller than) the width of the analysis window K, the matrix $\mathbf{P}(t)$ is often poorely conditioned, which may result in some sort of bursting behavior of the LBF algorithm. Note that regularization improves numerical conditioning of the identification problem since the condition number of the matrix $\mathbf{S}(t)$ (the ratio of its maximum and minimum singular values) is usually smaller than that of $\mathbf{R}(t)$.

V. Optimization of the regularization matrix

We will optimize the regularization matrix $\mathbf{R}$ when some prior knowledge about statistical properties of $\boldsymbol{\theta}(\cdot)$ is available. In the sequel we will assume that

(A1) $\{u(t)\}$ is a zero-mean circular white noise with covariance matrix $\boldsymbol{\Phi} = \mathrm{cov}[\boldsymbol{\varphi}(t)] = \sigma_u^2\mathbf{I}_n$.

(A2) $\{e(t)\}$, independent of $\{u(t)\}$, is a zero-mean circular white noise with variance σ_e^2.

(A3) $\{\boldsymbol{\theta}(t)\}$ is a sequence, independent of $\{u(t)\}$ and $\{e(t)\}$, with known correlation matrix $\mathrm{E}[\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\mathrm{H}}(t)] = \mathbf{Q} > 0$.

Circular white noise is a sequence of independent and identically distributed (i.i.d.) random variables with independent real and imaginary parts. Note that assumptions (A1)-(A2) are typical of wireless communication systems.

We will derive the formula for the mean square parameter estimation error matrix in the case where the parameter trajectory obeys the model (3). After straightforward calculations, one obtains

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R}) - \boldsymbol{\alpha}(t) = &-[\mathbf{P}(t) + \mathbf{B}^{\mathrm{H}}\mathbf{B}]^{-1}\mathbf{B}^{\mathrm{H}}\mathbf{B}\boldsymbol{\alpha}(t) \\ &+ [\mathbf{P}(t) + \mathbf{B}^{\mathrm{H}}\mathbf{B}]^{-1}\boldsymbol{\xi}(t)\end{aligned} \tag{12}$$

where

$$\boldsymbol{\xi}(t) = \sum_{i=-k}^{k} w(i)e^*(t+i)\boldsymbol{\psi}(t,i). \tag{13}$$

Under the assumptions made above it can be shown that for growing k the regression matrix $\mathbf{P}(t)$ converges in the mean squared sense to the constant matrix $\bar{\mathbf{P}} = \mathrm{E}[\mathbf{P}(t)] = \sigma_u^2\mathbf{I}_{mn}$ – see [16]. This justifies the following approximation valid for sufficiently large values of k

$$\begin{aligned}\mathbf{S}^{-1}(t) = [\mathbf{P}(t) + \mathbf{B}^{\mathrm{H}}\mathbf{B}]^{-1} &\cong [\sigma_u^2\mathbf{I}_{mn} + \mathbf{B}^{\mathrm{H}}\mathbf{B}]^{-1} \\ &= \frac{1}{\sigma_u^2}[\mathbf{I}_{mn} + \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1}\end{aligned} \tag{14}$$

where $\widetilde{\mathbf{B}} = \widetilde{\mathbf{D}}\mathbf{F}_0$ and $\widetilde{\mathbf{D}} = \mathbf{D}/\sigma_u$. Using this approximation, one obtains

$$\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R}) - \boldsymbol{\alpha}(t) \cong \boldsymbol{\rho}_1(t) + \boldsymbol{\rho}_2(t) \tag{15}$$

where

$$\begin{aligned}\boldsymbol{\rho}_1(t) &= -[\mathbf{I}_{mn} + \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1}\widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}\boldsymbol{\alpha}(t) \\ \boldsymbol{\rho}_2(t) &= \frac{1}{\sigma_u^2}[\mathbf{I}_{mn} + \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1}\boldsymbol{\xi}(t)\ .\end{aligned} \tag{16}$$

In the sequel we will use the following result

Lemma 1

It holds that

$$\begin{aligned}[\mathbf{I}_{mn} + \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1} = \mathbf{I}_{mn} &- \mathbf{I}_n \otimes \left[\frac{\mathbf{f}_0\mathbf{f}_0^{\mathrm{T}}}{\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0}\right] \\ &+ [\mathbf{I}_n + \widetilde{\mathbf{R}}]^{-1} \otimes \left[\frac{\mathbf{f}_0\mathbf{f}_0^{\mathrm{T}}}{\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0}\right]\end{aligned} \tag{17}$$

where

$$\widetilde{\mathbf{R}} = \frac{\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0}{\sigma_u^2}\,\mathbf{R}\ . \tag{18}$$

Proof - see Appendix 1.

Using (17) and the identity

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD}) \tag{19}$$

which holds true for Kronecker products (provided that all dimensions match), one arrives at

$$\mathbf{F}_0[\mathbf{I}_{mn} + \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1} = [\mathbf{I}_n + \widetilde{\mathbf{R}}]^{-1} \otimes \mathbf{f}_0^{\mathrm{T}} \tag{20}$$

Note that

$$\begin{aligned}\widehat{\boldsymbol{\theta}}^{\mathrm{RLBF}}(t|\mathbf{R}) - \boldsymbol{\theta}(t) &= \mathbf{F}_0[\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R}) - \boldsymbol{\alpha}(t)] \\ &\cong \boldsymbol{\delta}_1(t) + \boldsymbol{\delta}_2(t)\end{aligned} \tag{21}$$

where $\boldsymbol{\delta}_1(t) = \mathbf{F}_0\boldsymbol{\rho}_1(t)$ and $\boldsymbol{\delta}_2(t) = \mathbf{F}_0\boldsymbol{\rho}_2(t)$.

Since $\mathbf{B}\boldsymbol{\alpha}(t) = \mathbf{D}\boldsymbol{\theta}(t) = (\mathbf{D}\otimes 1)\boldsymbol{\theta}(t)$, one obtains

$$\widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}\boldsymbol{\alpha}(t) = \frac{1}{\sigma_u^2}(\mathbf{D}^{\mathrm{H}}\otimes\mathbf{f}_0)(\mathbf{D}\otimes 1)\boldsymbol{\theta}(t). \tag{22}$$

Hence, combining (16), (20) and (22), and using the identity (19), one arrives at

$$\begin{aligned}\boldsymbol{\delta}_1(t) &= -[\mathbf{I}_n + \widetilde{\mathbf{R}}]^{-1}\mathbf{D}^{\mathrm{H}}\mathbf{D}\frac{\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0}{\sigma_u^2}\boldsymbol{\theta}(t) \\ &= -[\mathbf{I}_n + \widetilde{\mathbf{R}}]^{-1}\widetilde{\mathbf{R}}\boldsymbol{\theta}(t)\end{aligned} \tag{23}$$

and consequently (since the matrix $\widetilde{\mathbf{R}}$ is Hermitian)

$$\mathrm{E}[\boldsymbol{\delta}_1(t)\boldsymbol{\delta}_1^{\mathrm{H}}(t)] = [\mathbf{I}_n + \widetilde{\mathbf{R}}]^{-1}\widetilde{\mathbf{R}}\mathbf{Q}\widetilde{\mathbf{R}}[\mathbf{I}_n + \widetilde{\mathbf{R}}]^{-1} \tag{24}$$

where the expectation is carried out over different realizations of $\{\boldsymbol{\theta}(t)\}$.

In order to evaluate $\mathrm{E}[\boldsymbol{\delta}_2(t)\boldsymbol{\delta}_2^{\mathrm{H}}(t)]$, observe that

$$\begin{aligned}\mathrm{E}[\boldsymbol{\xi}(t)\boldsymbol{\xi}^{\mathrm{H}}(t)] &= \mathrm{E}\Bigg[\sum_{i=-k}^{k}\sum_{j=-k}^{k} w(i)w(j)e^{*}(t+i)e(t+j) \\ &\quad\times[\boldsymbol{\varphi}(t+i)\otimes\mathbf{f}(i)][\boldsymbol{\varphi}^{\mathrm{H}}(t+j)\otimes\mathbf{f}^{\mathrm{T}}(j)]\Bigg] \\ &= \sigma_e^2\,\mathrm{E}\Bigg[\sum_{i=-k}^{k} w^2(i)[\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+i)] \\ &\quad\otimes[\mathbf{f}(i)\mathbf{f}^{\mathrm{T}}(i)]\Bigg] = \sigma_e^2\sigma_u^2[\mathbf{I}_n\otimes\mathbf{W}] \end{aligned} \tag{25}$$

where

$$\mathbf{W} = \sum_{i=-k}^{k} w^2(i)\mathbf{f}(i)\mathbf{f}^{\mathrm{T}}(i) \tag{26}$$

and the expectation is over $\boldsymbol{\varphi}(t)$ and $e(t)$. Combining (16), (20) and (25), and using (19), one obtains

$$\begin{aligned}\mathrm{E}[\boldsymbol{\delta}_2(t)\boldsymbol{\delta}_2^{\mathrm{H}}(t)] &= \frac{\sigma_e^2}{\sigma_u^2}\Big[[\mathbf{I}_n+\widetilde{\mathbf{R}}]^{-1}\otimes\mathbf{f}_0^{\mathrm{T}}\Big][\mathbf{I}_n\otimes\mathbf{W}] \\ &\times\Big[[\mathbf{I}_n+\widetilde{\mathbf{R}}]^{-1}\otimes\mathbf{f}_0\Big] = \frac{\sigma_e^2}{\sigma_u^2 N_k}\,[\mathbf{I}_n+\widetilde{\mathbf{R}}]^{-1}[\mathbf{I}_n+\widetilde{\mathbf{R}}]^{-1}\end{aligned} \tag{27}$$

where

$$N_k = [\mathbf{f}_0^{\mathrm{T}}\mathbf{W}\mathbf{f}_0]^{-1} = \left\{\sum_{i=-k}^{k}[w(i)\mathbf{f}_0^{\mathrm{T}}\mathbf{f}(i)]^2\right\}^{-1} \tag{28}$$

denotes the so-called equivalent width of the analysis window, different from its effective width $L_k = \sum_{i=-k}^{k} w(i)$ – see [5].

Note that $\mathrm{E}[\boldsymbol{\delta}_1(t)\boldsymbol{\delta}_2^{\mathrm{H}}(t)] = \mathrm{E}[\boldsymbol{\delta}_2(t)\boldsymbol{\delta}_1^{\mathrm{H}}(t)] = 0$, which leads to the following expression for the mean square parameter tracking error matrix

$$\begin{aligned}&\mathrm{E}\Big\{[\widehat{\boldsymbol{\theta}}^{\mathrm{RLBF}}(t|\mathbf{R})-\boldsymbol{\theta}(t)][\widehat{\boldsymbol{\theta}}^{\mathrm{RLBF}}(t|\mathbf{R})-\boldsymbol{\theta}(t)]^{\mathrm{H}}\Big\} \\ &\quad\cong \mathrm{E}[\boldsymbol{\delta}_1(t)\boldsymbol{\delta}_1^{\mathrm{H}}(t)] + \mathrm{E}[\boldsymbol{\delta}_2(t)\boldsymbol{\delta}_2^{\mathrm{H}}(t)] \\ &\quad= \eta[\mathbf{I}_n+\widetilde{\mathbf{R}}]^{-1}[\widetilde{\mathbf{R}}\widetilde{\mathbf{Q}}\widetilde{\mathbf{R}}+\mathbf{I}_n][\mathbf{I}_n+\widetilde{\mathbf{R}}]^{-1} \\ &\quad= \mathrm{MSE}(\widetilde{\mathbf{R}})\end{aligned} \tag{29}$$

where

$$\eta = \frac{\sigma_e^2}{\sigma_u^2 N_k}, \quad \widetilde{\mathbf{Q}} = \frac{\mathbf{Q}}{\eta}\,. \tag{30}$$

Optimization of (30) will be based on the results of the following Theorem 1

Theorem 1

For any nonnegative definite matrix $\widetilde{\mathbf{R}}$ it holds that

$$\mathrm{MSE}(\widetilde{\mathbf{R}}) \geq \mathrm{MSE}(\widetilde{\mathbf{Q}}^{-1}) \tag{31}$$

Proof - see Appendix 2.

According to Theorem 1, the optimal choice of $\widetilde{\mathbf{R}}$ is given by $\widetilde{\mathbf{R}}_{\mathrm{opt}} = \widetilde{\mathbf{Q}}^{-1}$, i.e.,

$$\mathbf{R}_{\mathrm{opt}} = \frac{\sigma_e^2}{N_k\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0}\mathbf{Q}^{-1}\,. \tag{32}$$

Remark 1

So far we have been assuming that the variance σ_e^2 is constant and known. When the noise intensity varies with time, σ_e^2 can be replaced in (32) with the following LBF estimate

$$\begin{aligned}\widehat{\sigma}_e^2(t) &= \frac{1}{L_k}\sum_{i=-k}^{k} w(i)\Big|y(t+i)-[\widehat{\boldsymbol{\alpha}}^{\mathrm{LBF}}(t)]^{\mathrm{H}}\boldsymbol{\psi}(t,i)]\Big|^2 \\ &= \frac{1}{L_k}\Big[c(t)-[\widehat{\boldsymbol{\alpha}}^{\mathrm{LBF}}(t)]^{\mathrm{H}}\mathbf{p}(t)\Big]\end{aligned} \tag{33}$$

where

$$c(t) = \sum_{i=-k}^{k} w(i)|y(t+i)|^2.$$

VI. Adaptive regularization

In order to use the optimal regularization formula (23), one needs to know the correlation profile of the process $\{\boldsymbol{\theta}(t)\}$. As an example, consider the underwater acoustic (UWA) communication system. When the UWA system is fixed in the position, such a statistic can be determined experimentally by averaging identification results obtained in many trials. However, even in this simple case, the correlation matrix $\mathbf{Q} = \mathrm{E}[\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\mathrm{H}}(t)]$ is likely to depend on environmental factors such as the water temperature and weather conditions. Transmitter/receiver motion makes the picture even more complicated [2]. Therefore, to make the system more robust, at each time instant t the cancellation unit may be allowed to choose the best fitting variant amongst a certain number of the available correlation profiles. As a selection rule, one can use the leave-one-out cross validation approach. In this framework, the degree of fit of the model is defined as the local sum of squared unbiased interpolation errors (deleted residuals)

$$\varepsilon_0(t|\mathbf{R}) = y(t) - [\widehat{\boldsymbol{\theta}}_0^{\mathrm{RLBF}}(t|\mathbf{R})]^{\mathrm{H}}\boldsymbol{\varphi}(t) \tag{34}$$

where $\widehat{\boldsymbol{\theta}}_0^{\mathrm{RLBF}}(t|\mathbf{R})$ denotes the holey estimate of $\boldsymbol{\theta}(t)$, obtained by excluding from the estimation process, governed by (10), the "central" measurement $y(t)$

$$\begin{aligned}&\widehat{\boldsymbol{\alpha}}_0^{\mathrm{RLBF}}(t|\mathbf{R}) = \\ &= \arg\min_{\boldsymbol{\alpha}}\Bigg\{\sum_{\substack{i=-k\\ i\neq 0}}^{k} w(i)|y(t+i)-\boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\psi}(t,i)|^2 \\ &\qquad + ||\boldsymbol{\alpha}||^2_{\mathbf{F}_0^{\mathrm{T}}\mathbf{R}\mathbf{F}_0}\Bigg\} = \mathbf{S}_0^{-1}(t)\mathbf{p}_0(t) \\ &\widehat{\boldsymbol{\theta}}_0^{\mathrm{RLBF}}(t|\mathbf{R}) = \mathbf{F}_0\widehat{\boldsymbol{\alpha}}_0^{\mathrm{RLBF}}(t|\mathbf{R})\end{aligned} \tag{35}$$

where (note that $w(0)=1$)

$$\begin{aligned}\mathbf{S}_0(t) &= \mathbf{S}(t) - \boldsymbol{\psi}(t,0)\boldsymbol{\psi}^{\mathrm{H}}(t,0) \\ \mathbf{p}_0(t) &= \mathbf{p}(t) - y^{*}(t)\boldsymbol{\psi}(t,0).\end{aligned} \tag{36}$$

Lemma 2

It holds that

$$\varepsilon_0(t|\mathbf{R}) = \frac{\varepsilon(t|\mathbf{R})}{1-\beta(t)} \tag{37}$$

where $\beta(t) = \boldsymbol{\psi}^{\mathrm{H}}(t,0)\mathbf{S}^{-1}(t)\boldsymbol{\psi}(t,0)$ and

$$\varepsilon(t|\mathbf{R}) = y(t) - [\widehat{\boldsymbol{\theta}}^{\mathrm{RLBF}}(t|\mathbf{R})]^{\mathrm{H}}\boldsymbol{\varphi}(t) \tag{38}$$

denotes the residual error.

Proof - see Appendix 3.

According to (37), the deleted residuals can be easily computed in terms of residual errors, which means that implementation of the holey estimation scheme is not necessary to evaluate (34).

Consider now the case where several RLBF algorithms, equipped with different regularization matrices $\mathbf{R} \in \mathcal{R} = \{\mathbf{R}_1, \ldots, \mathbf{R}_M\}$, are run simultaneously yielding interpolation errors $\varepsilon_0(t|\mathbf{R}_i), i = 1, \ldots, M$. Selection of the best-fitting value of $\mathbf{R}$ can be made using the following cross-validation decision rule

$$\mathbf{R}_{\mathrm{opt}}(t) = \arg\min_{\mathbf{R}\in\mathcal{R}} \sum_{i=-L}^{L} |\varepsilon_0(t+i|\mathbf{R})|^2 \tag{39}$$

where L determines the size of the local decision window.

The range of applicability of the adaptive decision rule (39) is not restricted to selection of the regularization matrix $\mathbf{R}$. Tracking capabilities of LBF estimators may strongly depend on the number of basis functions m and on the choice of the width of the analysis interval $K = 2k+1$. It is known that small values of m and large values of k increase the bias component of the mean square parameter estimation error (MSE) and decrease its variance component. For large values of m and small values of k, one observes the opposite effect. Hence, to minimize MSE, which is the sum of its bias and variance components, the values of m and k should be chosen so as to guarantee a good bias-variance trade-off [16]. Similar as in the case of selection of $\mathbf{R}$, the best compromise can be seeked using the parallel estimation approach: not one, but several RLBF algorithms yielding the estimates $\widehat{\boldsymbol{\theta}}^{\mathrm{RLBF}}_{m|k}(t), k \in \mathcal{K}, m \in \mathcal{M}_k$, corresponding to different values of m and k, can be run in parallel and compared using the accumulated interpolation error statistic specified above. Moreover, the same approach can be used to select the most suitable basis set although, as demonstrated in [16], when standard general purpose functional bases are used (Taylor, Fourier, Slepian), this choice is usually of less importance.

Remark 2

Evaluation of $\beta(t)$, which is a part of (37), requires computation of $\mathbf{S}^{-1}(t)$ which may be, but must not be, a byproduct of evaluation of $\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t)$. The associated computational burden is O(m^3n^3) per time update. This computational burden can be reduced to O(n^2) if the following approximation, based on (14) and (17), is used

$$\begin{aligned}\widehat{\beta}(t) &= \frac{1}{\sigma_u^2}\,[\boldsymbol{\varphi}^{\mathrm{H}}(t)\otimes\mathbf{f}_0^{\mathrm{T}}][\mathbf{I}_{mn} + \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1}[\boldsymbol{\varphi}(t)\otimes\mathbf{f}_0] \\ &= \frac{\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0}{\sigma_u^2}\,\boldsymbol{\varphi}^{\mathrm{H}}(t)[\mathbf{I}_n + \widetilde{\mathbf{R}}]^{-1}\boldsymbol{\varphi}(t).\end{aligned} \tag{40}$$

Suppose that $\widetilde{\mathbf{R}}$ is chosen in the optimal form, i.e., $\widetilde{\mathbf{R}} = \eta\mathbf{Q}^{-1}$. The matrix $\mathbf{Q}$ can be written down in the form $\mathbf{Q} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{H}}$ where $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ is the diagonal matrix made up of the eigenvalues of $\mathbf{Q}$, and $\mathbf{V}$, $\mathbf{V}\mathbf{V}^{\mathrm{H}} = \mathbf{V}^{\mathrm{H}}\mathbf{V} = \mathbf{I}_n$, denotes the unitary matrix made up of its normalized eigenvectors. Both $\boldsymbol{\Lambda}$ and $\mathbf{V}$ can be computed prior to identification. Using this decomposition, one easily finds out that

$$[\mathbf{I}_n + \widetilde{\mathbf{R}}]^{-1} = [\mathbf{I}_n + \eta\mathbf{Q}^{-1}]^{-1} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{V}^{\mathrm{H}} \tag{41}$$

where

$$\boldsymbol{\Gamma} = \mathrm{diag}\left\{\frac{\lambda_1}{\eta+\lambda_1}, \ldots, \frac{\lambda_n}{\eta+\lambda_n}\right\}. \tag{42}$$

When the true variance σ_e^2 is replaced with its estimate (33), the only modification needed is replacement of η in (41)-(42) with its (time-varying) estimate

$$\widehat{\eta}(t) = \frac{\widehat{\sigma}_e^2(t)}{\sigma_u^2 N_k}. \tag{43}$$

VII. Computer simulations

The application, studied recently, which particularly well fits the technique developed in this paper, is adaptive self-interference cancellation in full-duplex (FD) underwater acoustic communication systems [6]. FD UWA systems, designed to maximize the limited capacity of acoustic links, simultaneously transmit and receive data in the same frequency band. Due to the close spacing of the transmit and receive antennas, the far-end signal is strongly contaminated by the so-called self-interference (SI) introduced by the near-end transmitter. Self-interference is a multipath propagation effect caused, among others, by multiple reflections of the emitted signal from the water surface and/or the bottom. The model of the received signal is given by (1), where $\{u(t)\}$ denotes the near-end (known) signal and $\{e(t)\}$ is a mixture of the far-end signal and the channel noise (ambient and/or site-specific). Note that in this case our goal is extraction of the signal $\{e(t)\}$ from $\{y(t)\}$, which can be easily done provided that channel parameters are known. Adaptive (on-line) identification of the channel is needed due to its time variability – the effect caused by the transmitter/receiver motion and/or by inherent changes in the propagation medium. An interesting feature of this application is that it allows one to work with a decision delay, which means that estimation of channel parameters can be based not only on past signal samples but also on a certain number of "future" (with respect to the moment of interest) ones. Hence, channel identification can be carried out using noncausal estimation algorithms with improved tracking capabilities, such as the ones described in this paper.

Simulation was carried out for the model of the self-interference channel of the full-duplex UWA system, decribed in [6]. Following [6], it was assumed that all complex-valued analog signals are sampled at the rate of 1 kHz, and that the bandwidth of channel coefficient variation is 1 Hz, which can be regarded as fast changes in the UWA case. The channel was modeled as a 50-tap FIR filter with complex-valued coefficients that vary independently of each other. The time varying impulse response coefficients were generated by lowpass filtering of discrete time circular (with independent real and imaginary components) white Gaussian noise with the variance chosen according to

$$\mathrm{var}[\theta_j(t)] = \zeta^{j-1}, \quad j = 1, \ldots, 50$$

which reflects the decaying power delay profile caused by the spreading and absorption loss. The value of ζ was set to 0.69 so that the ratio between the variance of the first arrivals ($j = 1$) and that of the latest arrivals ($j = 50$) was equal to 80 dB [6]. Typical trajectories of system parameters are shown in Fig. 1.

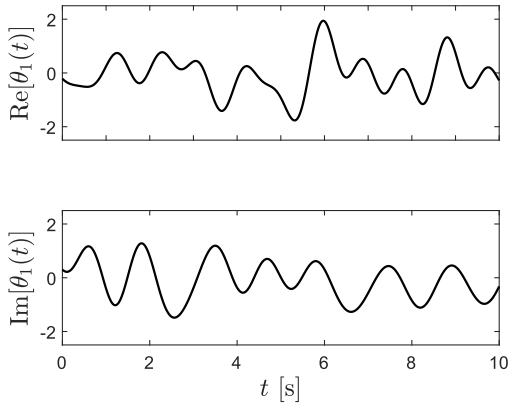

Fig. 1. Typical trajectories of system parameters.

The generated input signal was circular white binary $u(t) = \pm 1 \pm j$ and the measurement noise was circular white Gaussian with variance σ_e^2 equal to 0.0065, 0.065 and 0.65, which corresponds to the input signal-to-noise ratio

$$\mathrm{SNR} = \frac{\mathrm{E}[|\boldsymbol{\theta}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t)|^2]}{\sigma_e^2} = \frac{\sigma_u^2}{\sigma_e^2} \sum_{j=1}^{50} \mathrm{var}[\theta_j(t)]$$

equal to 30 dB, 20 dB and 10 dB, respectively.

The estimation design parameters were set to $k = 100$, $w(i) = \cos \frac{\pi i}{2k}, i \in I_k$, (recursively computable cosinusoidal window) and $m = 3$ (Legendre basis). Based on the available prior knowledge of the estimated impulse response (exponentially decaying and spatially uncorrelated), the regularization matrix was adopted in the form

$$\mathbf{R}(t) = \frac{\widehat{\sigma}_e^2(t)}{N_k \mathbf{f}_0^{\mathrm{T}} \mathbf{f}_0} \, \mathrm{diag}\{1, \gamma^{-1}, \gamma^{-2}, \ldots, \gamma^{-49}\}.$$

Three hypothetical values of γ were considered: 0.5, 0.7 and 0.9, none of which was equal to ζ. Optimization was carried out numerically using (39) by searching, at each time instant t, for the best value of $\gamma \in \{0.5, 0.7, 0.9\}$.

TABLE I

FIT[%] SCORES OBTAINED FOR 3 SIGNAL-TO-NOISE RATIOS FOR THE ALGORITHMS DESCRIBED IN THE TEXT.

Alg. \SNR	30 dB	20 dB	10 dB
LBF	95.7	86.6	57.5
RLBF_1	95.0	89.8	79.5
RLBF_2	97.3	92.6	80.8
RLBF_3	95.8	86.8	62.6
A_1	96.8	91.3	78.0
A_2	97.0	91.4	70.2
tsRLBF	96.5	89.1	66.1

Performance was evaluated in terms of the following normalized root mean squared error measure of fit used in [20]

$$\mathrm{FIT}(t) = 100 \left(1 - \left[\frac{\sum_{j=1}^{50} |\theta_j(t) - \widehat{\theta}_j(t)|^2}{\sum_{j=1}^{50} |\theta_j(t) - \bar{\theta}(t)|^2} \right]^{1/2} \right) \quad (44)$$

where $\bar{\theta}(t) = \frac{1}{50} \sum_{j=1}^{50} \theta_j(t)$. The maximum value of $\mathrm{FIT}(t)$, equal to 100, corresponds to the perfect match between the true and estimated impulse response. The final scores, further referred to as FIT (%), were obtained by combined time averaging (10000 time steps) and ensemble averaging (20 realizations of scaling coefficients) of the instantaneous/realization-constrained measures. Data generation was started 1000 time instants prior to $t = 1$ and was continued for 1000 time instants after $t = T_s$, where $T_s = 10000$ denotes simulation time.

Table 1 compares results obtained for the LBF algorithm, three RLBF algorithms with fixed values of γ: RLBF_1 ($\gamma = 0.5$), RLBF_2 ($\gamma = 0.7$) and RLBF_3 ($\gamma = 0.9$), and 2 algorithms with adaptive scheduling of γ, using $\beta(t)$ and $\widehat{\beta}(t)$ (A_1 and A_2, respectively). In all cases L was set to 30. The last row of Table 1 shows results obtained for the two-stage regularized LBF algorithm, denoted by tsRLBF, proposed in [18].

According to the results summarized in Table 1, regularization improves channel identification results (in spite of the discrepancy between the true value of γ and the assumed one). Furthermore, adaptive scheduling of γ yields performance comparable with that given by the best algorithms incorporated in the parallel estimation scheme. As expected, regularization provides the largest performance improvements for small values of SNR. Interestingly (and somewhat surprisingly), for SNR$\geq$ 20 dB the simplified decision rule works slightly better than the original rule which incorporates the exact value of $\beta(t)$. Finally, note that the new adaptive regularization scheme performs better than the t-sRLBF algorithm, which does not incorporate prior knowledge of the correlation structure of $\boldsymbol{\theta}(t)$.

Figure 2 shows FIT scores for all 20 process realizations, i.e., for all excitation patterns corresponding to different random choices of scaling coefficients. Note that adaptive al-

gorithms with regularization yield *consistently* better results than the not regularized LBF algorithm (i.e., results that are better not only in the mean sense but also for *every* process realization).

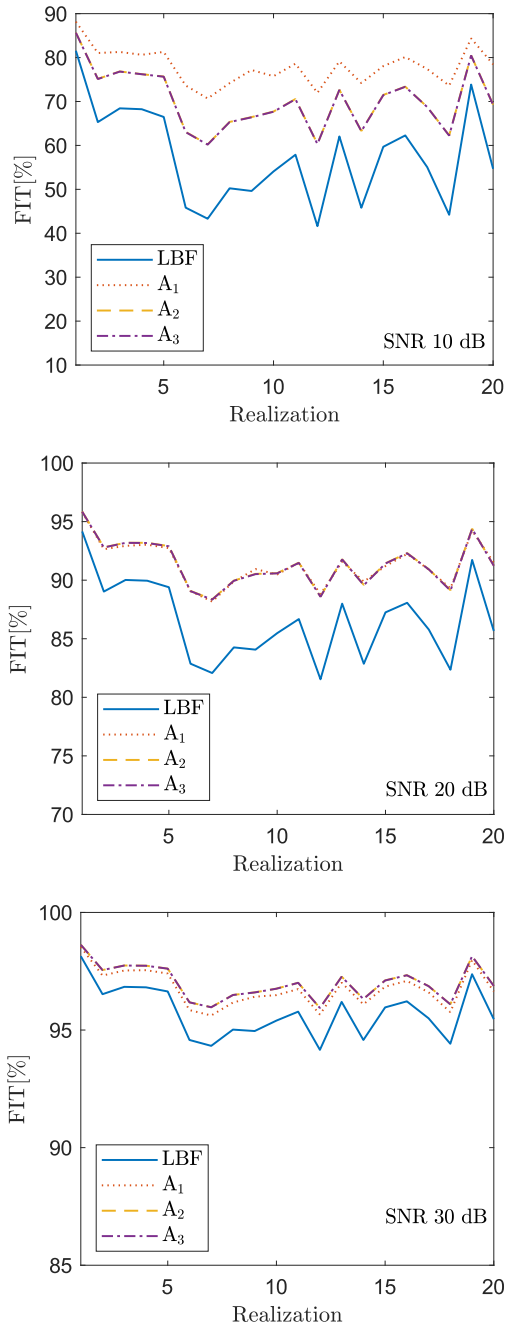

Fig. 2. FIT scores obtained for all 20 process realizations. For adaptive algorithms A_2 and A_3 the lines practically coincide.

In the second simulation experiment the value of the rate of decay was fixed ($\gamma = 0.7$), while the values of the design parameters k and m were selected in an adaptive way using the simplified version of the cross-validation test. Two variants of the window width k were considered $\mathcal{K} = \{100, 200\}$. The corresponding choices of m were restricted to $\mathcal{M}_{100} = \{1, 3\}$ and $\mathcal{M}_{200} = \{1, 3, 5\}$. The combination $\{k = 100, m = 5\}$ was excluded as it would require estimation of $nm = 250$ hyperparameters from $K = 2k + 1 = 201$ data points. Note that, for all combinations of m and k, the regularized LBF algorithms yield better results than their not regularized versions.

TABLE II

FIT[%] SCORES OBTAINED FOR 3 SIGNAL-TO-NOISE RATIOS AND DIFFERENT COMBINATIONS OF m AND k FOR THE ALGORITHMS DESCRIBED IN THE TEXT.

	Method	k/m	1	3	5
10 dB	LBF	100	75.9	57.5	x
		200	78.9	77.1	68.2
	RLBF	100	86.9	80.8	x
		200	85.6	87.6	84.1
		A_3	80.9		
20 dB	LBF	100	88.1	86.6	x
		200	84.2	92.7	90.0
	RLBF	100	92.5	92.6	x
		200	87.8	95.3	94.0
		A_3	92.8		
30 dB	LBF	100	90.2	95.7	x
		200	84.9	97.5	96.8
	RLBF	100	93.5	97.3	x
		200	88.0	98.2	97.8
		A_3	97.4		

VIII. Conclusion

It was shown that regularization can improve parameter tracking capabilities of the local basis function algorithms used for identification of fast-varying FIR systems. First, the optimal regularization matrix was designed in the case where the correlation matrix of the estimated vector of system parameters is known. Then the adaptive regularization scheme was proposed, based on parallel estimation and cross-validation. It was shown that the obtained results are consistently better than those yielded by the local basis function algorithm without regularization. They are also better than results provided by the general purpose two-stage regularized algorithm proposed earlier.

References

[1] M. K. Tsatsanis and G. B. Giannakis, "Modeling and equalization of rapidly fading channels," *Int. J. Adaptive Contr. Signal Process.*, vol. 10, pp. 159-176, 1996.

[2] M. Stojanovic and J. Preisig, "Underwater acoustic communication channels: Propagation models and statistical characterization," *IEEE Communications Magazine*, vol. 47, pp. 84-89, 2009.

[3] T. Söderström and P. Stoica, *System Identification*, Prentice-Hall, 1988.

[4] S. Haykin, *Adaptive Filter Theory*. Prentice-Hall, 1996.

[5] M. Niedźwiecki, *Identification of Time-varying Processes*, Wiley, 2000.

[6] L. Shen, Y. Zakharov, B. Henson, N. Morozs and P. Mitchell, "Adaptive filtering for full-duplex UWA systems with time-varying self-interference channel," IEEE Access, vol. 8, pp. 187590-187604, 2020.

[7] J. P. Norton, "Optimal smoothing in the identification of linear time-varying systems," *Proc. IEEE*, vol. 122, pp. 663-668, 1975.

[8] G. Kitagawa and W. Gersch, *Smoothness Priors Analysis of Time Series*. Springer-Verlag, 1996.
[9] M. Niedźwiecki, "Locally adaptive cooperative Kalman smoothing and its application to identification of nonstationary stochastic systems," *IEEE Transactions on Signal Processing*, vol. 60, pp. 48-59, 2012.
[10] J. M. Liporace, "Linear estimation of nonstationary signals," *Journal of the Acoustical Society of America*, vol. 58, pp. 1288–1295, 1975.
[11] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Trans. Acoust. speech Signal Process.*, vol. 31, pp. 899-911, 1981.
[12] M. Niedźwiecki, "Functional series modeling approach to identification of nonstationary stochastic systems," *IEEE Transactions on Automatic Control*, vol. 33, pp. 955-961, 1988.
[13] M. K. Tsatsanis and G. B. Giannakis "Time-varying system identification and model reduction using wavelets," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3512-3523, 1994.
[14] G. B. Giannakis and C. Tepedelenlioğlu, "Basis expansion models and diversity techniques for blind identification and equalization of time-varying channels," *Proceedings of IEEE*, vol. 86, pp. 1969–1986, 1998.
[15] H. L. Wei, J. J. Liu and S. A. Billings, "Identification of time-varying systems using multi-resolution wavelet models," *International Journal of Systems Science*, vol. 33, pp. 1217-1228, 2002.
[16] M. Niedźwiecki and M. Ciołek, "Generalized Savitzky-Golay filters for identification of nonstationary systems," *Automatica*, vol. 108, no. 108522, 2019.
[17] M. Niedźwiecki, M. Ciołek and A. Gańcza, "A new look at the statistical identification of nonstationary systems,"*Automatica*, vol. 118, no. 109037, 2020.
[18] A. Gańcza, M. Niedźwiecki and M.Ciołek, "Regularized local basis function approach to identification of nonstationary processes," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1665-1680, 2021.
[19] A. Tikhonov and V. Arsenin, *Solutions of Ill-posed Problems*. Winston/Wiley, 1977.
[20] L. Ljung, T. Chen, "What can regularization offer for estimation of dynamical systems?," *Proc. of the 11th IFAC Workshop on Adaptation and Learning in Control and Signal Processing*, Caen, France, pp. 1-8, 2013.
[21] Y.V. Zakharov, G.P. White and J. Liu, "Low-complexity RLS algorithms using dichotomous coordinate descent iterations," *IEEE Transactions on Signal Processing*, vol. 56, pp. 3150-3161, 2008.
[22] T. Chen, H. Ohlsson and L. Ljung "On the estimation of transfer functons, regularizations and Gaussian processes - Revisited," *Automatica*, vol. 48, pp. 1525-1535, 2012.

APPENDIX 1 [proof of Lemma 1]

Using the Woodbury matrix identity [3] , one obtains

$$[\mathbf{I}_{mn}+\widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1}=\mathbf{I}_{mn}-\widetilde{\mathbf{B}}^{\mathrm{H}}[\mathbf{I}_n+\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^{\mathrm{H}}]^{-1}\widetilde{\mathbf{B}}$$

Note that

$$\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^{\mathrm{H}}=(\widetilde{\mathbf{D}}\otimes\mathbf{f}_0^{\mathrm{T}})(\widetilde{\mathbf{D}}^{\mathrm{H}}\otimes\mathbf{f}_0)=\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0\widetilde{\mathbf{D}}\widetilde{\mathbf{D}}^{\mathrm{H}}.$$

Hence

$$\begin{aligned}&[\mathbf{I}_{mn}+\widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1}\\&\quad=\mathbf{I}_{mn}-[\widetilde{\mathbf{D}}^{\mathrm{H}}\otimes\mathbf{f}_0][\mathbf{I}_n+\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0\widetilde{\mathbf{D}}\widetilde{\mathbf{D}}^{\mathrm{H}}]^{-1}[\widetilde{\mathbf{D}}\otimes\mathbf{f}_0^{\mathrm{T}}]\\&\quad=\mathbf{I}_{mn}-\left\{\widetilde{\mathbf{D}}^{\mathrm{H}}[\mathbf{I}_n+\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0\widetilde{\mathbf{D}}\widetilde{\mathbf{D}}^{\mathrm{H}}]^{-1}\widetilde{\mathbf{D}}\right\}\otimes[\mathbf{f}_0\mathbf{f}_0^{\mathrm{T}}].\end{aligned}$$

Observe that

$$[\mathbf{I}_n+\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0\widetilde{\mathbf{D}}^{\mathrm{H}}\widetilde{\mathbf{D}}]^{-1}=\mathbf{I}_n-\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0\widetilde{\mathbf{D}}^{\mathrm{H}}[\mathbf{I}_n+\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0\widetilde{\mathbf{D}}\widetilde{\mathbf{D}}^{\mathrm{H}}]^{-1}\widetilde{\mathbf{D}}$$

Combining the last two results, and noting that $\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0\widetilde{\mathbf{D}}^{\mathrm{H}}\widetilde{\mathbf{D}}=(\mathbf{f}_0^{\mathrm{T}}\mathbf{f}_0/\sigma_u^2)\mathbf{R}=\widetilde{\mathbf{R}}$, one arrives at (17).

APPENDIX 2 [proof of Theorem 1]

The relationship (31) is a variant of a similar result shown in [22] (see Theorem 1 there).

Let $\widetilde{\mathbf{P}}=\widetilde{\mathbf{R}}^{-1}$. Note that (29) can be rewritten in the form

$$\mathrm{MSE}(\widetilde{\mathbf{P}})=\eta[\mathbf{I}_n+\widetilde{\mathbf{P}}]^{-1}[\widetilde{\mathbf{P}}\widetilde{\mathbf{P}}+\widetilde{\mathbf{Q}}][\mathbf{I}_n+\widetilde{\mathbf{P}}]^{-1}.$$

Set $\mathbf{X}=-[\mathbf{I}_n+\widetilde{\mathbf{P}}]^{-1}$, $\mathbf{X}_0=[\mathbf{I}_n+\widetilde{\mathbf{Q}}]^{-1}$ and note that $\mathbf{X}\widetilde{\mathbf{P}}=-(\mathbf{I}_n+\mathbf{X})$, $\mathbf{X}_0\widetilde{\mathbf{Q}}=-(\mathbf{I}_n+\mathbf{X}_0)$. Using this notation, one obtains

$$\begin{aligned}\mathrm{MSE}(\widetilde{\mathbf{P}})&=\eta(\mathbf{I}_n+\mathbf{X})(\mathbf{I}_n+\mathbf{X})+\eta\mathbf{X}\widetilde{\mathbf{Q}}\mathbf{X}\\\mathrm{MSE}(\widetilde{\mathbf{Q}})&=\eta(\mathbf{I}_n+\mathbf{X}_0)(\mathbf{I}_n+\mathbf{X}_0)+\eta\mathbf{X}_0\widetilde{\mathbf{Q}}\mathbf{X}_0\end{aligned}$$

In the sequel we will use the following simple identity

$$\begin{aligned}\mathbf{ACA}-\mathbf{BCB}&=(\mathbf{A}-\mathbf{B})\mathbf{C}(\mathbf{A}-\mathbf{B})\\&\quad+\mathbf{ACB}+\mathbf{BCA}-2\mathbf{BCB}.\end{aligned}$$

Applying this identity and noting that the matrices $\mathbf{X}$, $\mathbf{X}_0$ and $\widetilde{\mathbf{Q}}$ are Hermitian, and $(\mathbf{I}_n+\mathbf{X}_0)=-\mathbf{X}_0\widetilde{\mathbf{Q}}=-\widetilde{\mathbf{Q}}\mathbf{X}_0$, one arrives at

$$\begin{aligned}&\boldsymbol{\Delta}_1=(\mathbf{I}_n+\mathbf{X})(\mathbf{I}_n+\mathbf{X})-(\mathbf{I}_n+\mathbf{X}_0)(\mathbf{I}_n+\mathbf{X}_0)\\&=(\mathbf{X}-\mathbf{X}_0)(\mathbf{X}-\mathbf{X}_0)-\mathbf{X}\widetilde{\mathbf{Q}}\mathbf{X}_0-\mathbf{X}_0\widetilde{\mathbf{Q}}\mathbf{X}+2\mathbf{X}_0\widetilde{\mathbf{Q}}\mathbf{X}_0\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Delta}_2&=\mathbf{X}\widetilde{\mathbf{Q}}\mathbf{X}-\mathbf{X}_0\widetilde{\mathbf{Q}}\mathbf{X}_0=(\mathbf{X}-\mathbf{X}_0)\widetilde{\mathbf{Q}}(\mathbf{X}-\mathbf{X}_0)\\&\quad+\mathbf{X}\widetilde{\mathbf{Q}}\mathbf{X}_0+\mathbf{X}_0\widetilde{\mathbf{Q}}\mathbf{X}-2\mathbf{X}_0\widetilde{\mathbf{Q}}\mathbf{X}_0.\end{aligned}$$

This leads to $\mathrm{MSE}(\widetilde{\mathbf{P}})-\mathrm{MSE}(\widetilde{\mathbf{Q}})=\eta(\boldsymbol{\Delta}_1+\boldsymbol{\Delta}_2)$ $=\eta(\mathbf{X}-\mathbf{X}_0)[\mathbf{I}_n+\widetilde{\mathbf{Q}}](\mathbf{X}-\mathbf{X}_0)\geq 0$, which proves (31).

APPENDIX 3 [proof of Lemma 2]

Using the matrix inversion lemma [3], one obtains

$$\mathbf{S}_0^{-1}(t)=\mathbf{S}^{-1}(t)+\frac{\mathbf{S}^{-1}(t)\boldsymbol{\psi}(t,0)\boldsymbol{\psi}^{\mathrm{H}}(t,0)\mathbf{S}^{-1}(t)}{1-\beta(t)}$$

which, after straightforward calculations, leads to

$$\begin{aligned}&\widehat{\boldsymbol{\alpha}}_0^{\mathrm{RLBF}}(t|\mathbf{R})=\mathbf{S}_0^{-1}(t)\mathbf{p}_0(t)=\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R})\\&+\frac{1}{1-\beta(t)}\,\mathbf{S}^{-1}(t)\boldsymbol{\psi}(t,0)\boldsymbol{\psi}^{\mathrm{H}}(t,0)\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R})\\&-\frac{1}{1-\beta(t)}\,\mathbf{S}^{-1}(t)\boldsymbol{\psi}(t,0)y^*(t)\\&=\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R})-\frac{1}{1-\beta(t)}\,\mathbf{S}^{-1}(t)\boldsymbol{\psi}(t,0)\varepsilon^*(t|\mathbf{R})\end{aligned}$$

where the last transition follows from the fact that

$$\begin{aligned}\boldsymbol{\psi}^{\mathrm{H}}(t,0)\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R})&=\left[[\widehat{\boldsymbol{\alpha}}^{\mathrm{RLBF}}(t|\mathbf{R})]^{\mathrm{H}}\boldsymbol{\psi}(t,0)\right]^*\\&=\left[[\widehat{\boldsymbol{\theta}}^{\mathrm{RLBF}}(t|\mathbf{R})]^{\mathrm{H}}\boldsymbol{\varphi}(t)\right]^*.\end{aligned}$$

Since

$$\begin{aligned}\varepsilon_0^*(t|\mathbf{R})&=\left[y(t)-[\widehat{\boldsymbol{\theta}}_0^{\mathrm{RLBF}}(t|\mathbf{R})]^{\mathrm{H}}\boldsymbol{\varphi}(t)\right]^*\\&=y^*(t)-\boldsymbol{\psi}^{\mathrm{H}}(t,0)\widehat{\boldsymbol{\alpha}}_0^{\mathrm{RLBF}}(t|\mathbf{R})\end{aligned}$$

one finally obtains

$$\varepsilon_0^*(t|\mathbf{R})=\varepsilon^*(t|\mathbf{R})+\frac{\beta(t)}{1-\beta(t)}\varepsilon^*(t|\mathbf{R})=\frac{\varepsilon^*(t|\mathbf{R})}{1-\beta(t)}$$

which is nothing but (26).