GDAŃSK UNIVERSITY OF TECHNOLOGY

The author of the doctoral dissertation: Artur Gańcza
Scientific discipline: Automation, electronics, electrical engineering and space technologies
[AEEiTK]

DOCTORAL DISSERTATION

Title of doctoral dissertation: Local basis function method for identification of nonstationary systems

Title of doctoral dissertation (in Polish): Lokalna metoda funkcji bazowych w identyfikacji obiektów niestacjonarnych

Supervisor	~~Second supervisor~~
signature	~~*signature*~~
Prof. dr hab. Inż. Maciej Niedźwiecki	
~~Auxiliary supervisor~~	~~Cosupervisor~~
~~*signature*~~	~~*signature*~~

Gdańsk, year 2023

POLITECHNIKA GDAŃSKA

STATEMENT

The author of the doctoral dissertation: Artur Gańcza

I, the undersigned, declare that I am aware that in accordance with the provisions of Art. 27 (1) and (2) of the Act of 4th February 1994 on Copyright and Related Rights (Journal of Laws of 2021, item 1062), the university may use my doctoral dissertation entitled:
<title>
for scientific or didactic purposes.[1]

Gdańsk, 31.08.2023

....
signature of the PhD student

Aware of criminal liability for violations of the Act of 4th February 1994 on Copyright and Related Rights and disciplinary actions set out in the Law on Higher Education and Science(Journal of Laws 2021, item 478), as well as civil liability, I declare, that the submitted doctoral dissertation is my own work.

I declare, that the submitted doctoral dissertation is my own work performed under and in cooperation with the supervision of <name of the supervisor>, the second supervision of <name of the second supervisor>, the auxiliary supervision of <name of the auxiliary supervisor>, the cosupervision of <name of the cosupervisor>*.

This submitted doctoral dissertation has never before been the basis of an official procedure associated with the awarding of a PhD degree.

All the information contained in the above thesis which is derived from written and electronic sources is documented in a list of relevant literature in accordance with Art. 34 of the Copyright and Related Rights Act.

I confirm that this doctoral dissertation is identical to the attached electronic version.

Gdańsk, 31.08.2023

....
signature of the PhD student

I, the undersigned, agree/do not agree* to include an electronic version of the above doctoral dissertation in the open, institutional, digital repository of Gdańsk University of Technology.

Gdańsk, 31.08.2023

....
signature of the PhD student

delete where appropriate

---

[1] Art 27. 1. Educational institutions and entities referred to in art. 7 sec. 1 points 1, 2 and 4–8 of the Act of 20 July 2018 – Law on Higher Education and Science, may use the disseminated works in the original and in translation for the purposes of illustrating the content provided for didactic purposes or in order to conduct research activities, and to reproduce for this purpose disseminated minor works or fragments of larger works.

2. If the works are made available to the public in such a way that everyone can have access to them at the place and time selected by them, as referred to in para. 1, is allowed only for a limited group of people learning, teaching or conducting research, identified by the entities listed in paragraph 1.

I am extremely grateful to my supervisor prof. Maciej Niedźwiecki for his guidance, helpful advice, and patience that cannot be overestimated.

Artur Gańcza

Abstract

This thesis is focused on the basis function method for the identification of nonstationary processes. The first chapter describes a group of models that can be identified using the basis function method. The next chapter describes the basic version of the basis function method, including its algebraic and statistical properties. The following section introduces the local basis function (LBF) method: its properties are described and similarities and differences between LBF and the basic basis function method are highlighted. The main difference lies in the approach to estimation. The primary version of the basis function method provides estimates for the entire analysis interval. The analysis window is then shifted so that estimates can be found for the next set of observations. In the case of the LBF method, the data from the analysis window are used to find parameter estimates for only one time instant within the analysis interval. The window is then moved to the subsequent observation and the estimation process is repeated. As a result, one obtains more accurate estimates at the expense of the increased computational burden.

This chapter also describes the methods of adaptive choice of hyperparameters crucial for the accuracy of the estimates, i.e. the number of basis functions and the length of the analysis window. The thesis concentrates on the parallel processing methods: one runs several algorithms simultaneously and each one of them is equipped with different settings. At each time instant, the estimator that minimizes a value of the local quality measure is chosen.

The following chapter describes the fast local basis function method (fLBF). It is a two-step procedure, which transforms the identification problem into a filtration problem. In the first step, one finds the so-called preestimates, which are approximately unbiased estimates of parameter trajectories. Therefore, they can be seen as true parameter trajectories contaminated with a zero-mean noise of finite variance. The next step is filtering, which yields final parameter estimates. In this chapter, it is also shown that under certain conditions, the estimates obtained using the fLBF method can be seen as a close approximation of estimates obtained using the LBF method. Additionally, the properties of the fLBF method are described, emphasizing similarities and differences between fLBF and LBF methods. Next, different preestimation methods are described, and differences between preestimation errors and computational burden are pointed out. Finally, the methods of adaptive choice of a number of basis functions and the length of the analysis window were described. It is also noted that the preestimates associated with different parameter trajectories can be processed separately.

The next chapter describes the use of regularization in the LBF and fLBF methods. The four critical questions to consider when applying regularization are: what is being constrained, what penalty is used, what prior knowledge is available, and what optimization technique is used. The chapter provides answers to these questions. First, the form of regularized estimators is derived. Next, it is explained how to design the regularization matrix. Finally, it is described how to choose the parameters defining the regularization matrix. The proposed methods are based on parallel processing.

The second last chapter contains description of simulations and experiments along with their results. This part of the thesis compares LBF estimators using different types of basis functions and different weighting sequences. It also includes a comparison of results yielded by methods of adaptive choice of hyperparameters. Lastly, the LBF method is compared with two other identification algorithms. The first algorithm was, until recently, regarded as the state-of-the-art in channel estimation for underwater acoustic communication. The second one is a variation of the basic basis function method, which employs wavelets as basis functions and incorporates an effective method for selecting the most important (from the modeling perspective) basis functions and input signals. The next part of this chapter presents the results of applying the fLBF algorithm to identification. It starts with the comparison of results obtained for different types of preestimates.

3

Next, the results obtained with the optimized impulse response of the filter used at the second stage, are presented. Next a comparison of methods with regularization is provided. The regularization matrix is optimized under the assumption of full statistical knowledge about parameter changes. The last part of this chapter describes results obtained for simulations performed using the simulator of underwater acoustic communication, as well as for the results obtained for data gathered in the lake experiment.

The thesis is concluded with a chapter containing a summary and general conclusions stemming from the simulations and experiments.

Streszczenie

Niniejsza praca poświęcona jest metodzie funkcji bazowych w identyfikacji procesów niestacjonarnych. W pierwszym rozdziale przedstawiono grupę modeli, których identyfikacja jest możliwa przy pomocy metod funkcji bazowych. W kolejnym rozdziale omówiona została podstawowa wersja metody funkcji bazowych - opisano jej właściwości algebraiczne i statystyczne. Następnie wprowadzono lokalną metodę funkcji bazowych (LBF), przedstawiono jej właściwości i wskazano na różnice, i podobieństwa pomiędzy nią i podstawową wersją metody funkcji bazowych. Najważniejszą różnicę stanowi podejście do estymacji. W podstawowej wersji metody funkcji bazowych uzyskuje się estymaty parametrów dla całego przedziału analizy. Następnie przesuwa się okno analizy w taki sposób żeby móc znaleźć oszacownaia wartości parametrów odpowiadające kolejnej grupie obserwacji. Natomiast w metodzie LBF dane pochodzące z lokalnego okna analizy służą do uzyskania estymat jedynie dla jednej chwili czasu wewnątrz przedziału analizy. Następnie okno jest przesuwane i procedura estymacji zostaje powtórzona dla kolejnej chwili czasu. Skutkuje to uzyskaniem dużo dokładniejszych estymat parametrów, kosztem zwiększonych nakładów obliczeniowych.

Rozdział ten opisuje także metody adaptacyjnego wyboru parametrów kluczowych dla jakości estymat, tj. liczby funkcji bazowych i szerokości okna analizy. W pracy skupiono się na metodach opartych o przetwarzanie równoległe. Metody te zakładają, że estymaty są wyznaczane jednocześnie przez kilka algorytmów pracujących z różnymi ustawieniami. W każdej chwili czasu wybiera się jako ostateczne te oszacowania parametrów, dla których pewien wskaźnik błędu osiąga najmniejszą wartość.

Kolejny rozdział skupia się na szybkiej lokalnej metodzie funkcji bazowych (fLBF). Jest to dwustopniowa procedura, która sprowadza problem identyfikacji do problemu filtracji. W pierwszym kroku znajdowane są tzw. preestymaty, które są w przybliżeniu nieobciążonymi oszacowaniami zmian prametrów modelu. Mogą więc być postrzegane jako prawdziwe trajektorie parametrów, zanieczyszczone szumem o zerowej wartości oczekiwanej i dużej wariancji. Kolejny krok to filtracja, która pozwala uzyskać końcowe estymaty. W rozdziale tym pokazano, że w określonych warunkach estymaty uzyskane za pomocą metody fLBF stanowią dobre przybliżenie estymat otrzymywanych przy pomocy metody LBF. Poza tym omówiono właściwości algorytmu fLBF, zwracając uwagę na podobieństwa i różnice miedzy metodami LBF i fLBF. Następnie opisano różne rodzaje preestymacji zwracając uwagę na różnice w błędach preestymacji oraz wymaganych nakładach obliczeniowych. Na koniec opisano metody adaptacyjnego wyboru liczby funkcji bazowych oraz szerokości okna analizy, zwracając uwagę na możliwość odrębnego przetwarzania preestymat odpowiadających poszczególnym parametrom modelu.

W kolejnym rozdziale omówiono zastosowanie regularyzacji w połączeniu z dwoma wcześniej wymienionymi metodami. Rozdział otwiera lista kluczowych pytań, na które należy odpowiedzieć stosując regularyzację, tj. jak skonstruować funkcję kary, czy wiedza wstępna na temat parametrów modelu jest dostęna i jeśli tak, to jak ją wykorzystać, i ostatecznie jak zaprojektować macierz regularyzacyjną i jak wybrać współczynnik regularyzacyjny. Kolejne części tego rozdziału stopniowo odpowiadają na przedstawione pytania. W pierwszej kolejności podano postać estymatorów z regularyzacją, następnie został opisany sposób wyboru macierzy regularyzacyjnej oraz jej znaczenie dla różnych właściwości chwilowej odpowiedzi impulsowej identyfikowanego systemu. W końcowej części rozdziału opisano jak w praktyce można wybrać wartości parametrów definiujących macierz regularyzacyjną. Przedstawione metody adaptacyjnego wyboru ponownie opierają się na przetwarzaniu równoległym.

Przedostatni rozdział zawiera opis symulacji i eksperymentów oraz ich wyniki. Najpierw zostały porównane wyniki uzyskane dla estymatorów LBF, korzystających z różnych rodzajów funkcji bazowych i różnych ciągów ważących. Zamieszczono także porównanie wyników otrzymanych przy zastosowaniu metod adaptacyjnego wyboru liczby funkcji bazowych i szerokości okna analizy.

6

Porównano rezultaty otrzymywane przy użyciu metody LBF, z wynikami, które można uzyskać stosując dwie inne metody identyfikacji. Pierwsza z nich była do niedawna uważana za wiodącą w zadaniach identyfikacji, w zastosowaniach związanych z akustyczną komunikacją podwodną. Druga z metod jest wariantem podstawowej metody funkcji bazowych, która wykorzystuje funkcje falkowe o różnych rozdzielczościach oraz efektywny mechanizm wyboru takich funkcji, które są najważniejsze z perspektywy modelowania. W kolejnej części rozdziału zaprezentowano wyniki uzyskane dla algorytmu fLBF. Porównane zostały wyniki uzyskane przy zastosowaniu różnych metod preestymacji, następnie zaprezentowano wyniki dla zoptymalizowanych odpowiedzi impulsowych filtru stosowanego w drugim etapie przetwarzania. Przedostatnia część przedstawia wyniki dla metod z regularyzacją i zoptymalizowaną macierzą regularyzacyjną (przy założeniu pełnej wiedzy statystycznej o zmianach parametrów). Ostatnia część tego rozdziału opisuje wyniki uzyskane dla symulacji wykonanych przy użyciu symulatora komunikacji podwodnej, a także rezultaty symulacji przeprowadzonych na danych zebranych w eksperymencie na jeziorze.

Pracę zamknięto rozdziałem podsumowującym i prezentującym ogólne wnioski płynące z symulacji, eksperymentów i wcześniejszych rozważań.

Abbreviations

AIC	Akaike information criterion
AR	autoregressive
ARX	autoregressive with exogenous input
BF	basis function
CV	cross-validation
DCD	dichotomous coordinate descent
DFT	discrete Fourier transform
DPSS	discrete prolate spheroidal sequences
EB	empirical Bayes
EWLS	exponentially weighted least squares
FD	full-duplex
FIR	finite impulse response
FPE	final prediction error
fLBF	fast local basis function
fRLBF	fast regularized local basis function
GCV	generalized cross-validation
i.i.d.	independent and identically distributed
IIR	infinite impulse response
KL	Karhunen-Loéve
LASSO	least absolute shrinkage and selection operator
LBF	local basis faunction
LMS	least mean squares
LOOCV	leave-one-out cross-validation
LPV	linear parameter-varying
LTI	linear time-invariant
LTV	linear time-varying
MA	moving-average
MAC	multiply and accumulate
MIMO	multiple-input-multiple-output

MSE	mean squared estimation error
MW	multi-wavelet
OLS	orthogonal least squares
RLBF	regularized local basis function
SI	self-interference
SISO	single-input-single-output
STLS	sequential thresholded least squares
SNR	singal-to-noise ratio
UAC	underwater acoustic communication
UWA	underwater acoustic
WLS	weighted least squares

List of Symbols

Operators

$\otimes$	Kronecker product
$\mathbf{x}^{\mathrm{T}}$	transpose of $\mathbf{x}$
$\mathbf{x}^{\mathrm{H}}$	complex-conjugate transpose (Hermitian adjoint) of $\mathbf{x}$
$\mathbf{1}_{n,i}$	$n \times 1$ indicator vector - vector containing one on a position i and zeros elsewhere
$\mathbf{x}_{(i)}$	i-th element of a vector x
$\mathbf{X}_{(ij)}$	element from i-th row and j-th column of a matrix X
$\mathbf{X}_{(\cdot i)}$	i-th column of a matrix X
$\mathbf{X}_{(i\cdot)}$	i-th row of a matrix X
x^*	complex-conjugate of x
$\mathbf{X}^{-1}$	inverse of $\mathbf{X}$
$||\mathbf{x}||$	ℓ^2 norm of $\mathbf{x}$
$||\mathbf{x}||_1$	ℓ^1 norm of $\mathbf{x}$
$\det(\mathbf{X})$	determinant of a square matrix $\mathbf{X}$
$\det_+(\mathbf{X})$	pseudodeterminant of a square matrix $\mathbf{X}$, i.e. the product of all nonzero eigenvalues of $\mathbf{X}$
$\mathrm{Tr}(\mathbf{X})$	trace of $\mathbf{X}$
$\mathrm{diag}\{x_1,\ldots,x_n\}$	diagonal matrix with $x_1,\ldots,x_n$ on the main diagonal
$\lambda_{\max}(\mathbf{X})$	the greatest eigenvalue of a square matrix $\mathbf{X}$
$\lambda_{\min}(\mathbf{X})$	the smallest eigenvalue of a square matrix $\mathbf{X}$
$\lambda_i(\mathbf{X})$	the i-th eigenvalue of a square matrix $\mathbf{X}$
$\arg(x)$	phase of a complex number x
$\mathrm{Re}(\mathbf{x})$	real part of $\mathbf{x}$
$\mathrm{Im}(\mathbf{x})$	imaginary part of $\mathbf{x}$
$\mathrm{E}(\mathbf{x})$	expectation of $\mathbf{x}$
$\mathrm{cov}(\mathbf{x})$	covariance matrix of $\mathbf{x}$
$\mathrm{var}(x)$	variance of the random variable x
$\mathrm{int}(x)$	an integer closest to x
$\lceil x \rceil$	an integer closest to x, but not smaller than x
$x_n \xrightarrow{\mathrm{dist.}} x$	a sequence of random variables $\{x_n\}$ converges in distribution to a random variable x

9

q^{-1}	backshift operator
δ_{ij}	Kronecker delta
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean equal to μ and variance equal to σ^2
$\mathcal{U}(a, b)$	uniform distribution defined on a support $[a, b]$
$\mathcal{O}(*)$	the big O notation (Landau's symbol)
$\text{card}(\mathcal{X})$	cardinality of set $\mathcal{X}$
$\arg\min_{\mathbf{x}}\{f(\mathbf{x})\}$	argument $\mathbf{x}$ corresponding to the minimum value of a function $f(\mathbf{x})$
$\arg\max_{\mathbf{x}}\{f(\mathbf{x})\}$	argument $\mathbf{x}$ corresponding to the maximum value of a function $f(\mathbf{x})$
$\max_{\mathbf{x}\in X^n} f(\mathbf{x})$	the maximum value of a function $\{f(\mathbf{x})\}$ on a set X^n
$P(X)$	Probability of an event X

Important variables and constants

i	imaginary unit obeying $\text{i}^2 = -1$
n	the order of a nonstationary system (the number of system parameters)
m	the number of chosen basis functions
k	the half-width of the analysis window
K	the length of the analysis window $K = 2k + 1$
t	discrete normalized time, $t \in \mathbb{Z}$
T	the number of data samples
τ	continous time, $\tau \in \mathbb{R}$
$y(t)$	output signal of a nonstationary system
$u(t)$	input signal of a nonstationary system
$e(t)$	measurement noise
σ_u^2	the variance of the input signal
σ_e^2	the variance of the measurement noise
$\theta_j(t)$	j-th time-varying parameter of a nonstationary system
λ	the forgetting constant $\lambda \in (0, 1)$
$w_k(i)$	the weighting function
$f_{l|k}(i)$	the l-th basis function
$h_{m|k}^{\text{LBF}}(i)$	the impulse response of a filter associated with the LBF method
$h_{m|k}^{\text{fLBF}}(i)$	the impulse response of a filter associated with the fLBF method
$l_{m|k}^{\text{LBF}}$	the equivalent number of observations for the LBF method
$\alpha_{j,m|k}^l(t)$	the l-th basis function coefficient for the j-th system parameter
$\widetilde{\theta}_j(t)$	the preestimate of j-th system parameter
$\boldsymbol{\varphi}(t)$	the $n \times 1$ vector of regression variables

$\boldsymbol{\theta}(t)$	the $n \times 1$ vector of time-varying system parameters
$\mathbf{z}(t)$	the $n \times 1$ vector of preestimation noise
$\mathbf{f}_{m\mid k}(i)$	the $m \times 1$ vector of basis functions
$\boldsymbol{\psi}_{m\mid k}(t,i)$	the $mn \times 1$ generalized regression vector $\boldsymbol{\psi}_{m\mid k}(t,i) = \boldsymbol{\varphi}(t) \otimes \mathbf{f}_{m\mid k}(i)$
$\boldsymbol{\alpha}_{j,m\mid k}(t)$	the $m \times 1$ vector of basis function coefficients for the j-th system parameter
$\boldsymbol{\alpha}_{m\mid k}(t)$	the $mn \times 1$ vector of basis function coefficients for all system parameters
$\widehat{\boldsymbol{\alpha}}_{m\mid k}(t)$	the $mn \times 1$ vector of estimates of basis function coefficients for all system parameters
$\widehat{\boldsymbol{\theta}}(t)$	the $n \times 1$ vector of estimates of time-varying system parameters
$\widetilde{\boldsymbol{\theta}}(t)$	the $n \times 1$ vector of preestimates
$\boldsymbol{\Phi}$	the $n \times n$ covariance matrix of the regression vector
$\mathbf{P}_{m\mid k}(t)$	the $mn \times mn$ generalized regression matrix
$\mathbf{I}_n$	the $n \times n$ identity matrix
I_k	the domain of basis functions $I_k = [-k,k]$

Other

$\mathbb{N}_+$	set of all natural numbers without zero

Chapter 1

Introduction and thesis overview

1.1 Nonstationary processes and their models

To understand the nature of nonstationary processes, one should start from the definition of a stationary random process. There are two types of stationarity: strict-sense stationarity and wide-sense stationarity. The definition of the first type is more restrictive. It states that a process is stationary in a strict sense if the joint probability density function of any set of process samples is time-shift invariant (does not depend on the time index). The second definition requires only that the mean and the autocorrelation function of the process are time-invariant. A mathematical formulation of these definitions can be found, e.g. in [57]. Obviously, strict-sense stationarity implies wide-sense stationarity while the reverse does not necessarily hold true. This work considers processes not obeying the wide-sense stationarity condition (and therefore also the strict-sense stationarity condition). Such processes will be further called nonstationary and they will be further divided into two categories, i.e. nonstationary signals, and nonstationary systems. In the first case, the behavior of the analyzed sequence can be explained using just the previous values of this sequence, and a purely stochastic component, whereas in nonstationary systems, the output is also affected by some additional input signal. Nonstationarity of random processes is often caused by intrinsic changes in the physical system that generates the signal, or by the change in environmental conditions, like temperature, humidity, etc. An example of a nonstationary system is a guided missile because its behavior changes with changing speed and altitude, due to the different air pressure and airflow around its components. Also, its mass decreases with time, as the missile burns its fuel.

Another "source" of nonstationarity is the linearization of nonlinear models. Nonlinear objects are ubiquitous in practice, and to work with them one usually linearizes the model around some operating point. When the operating point changes over time, linearization of the system must be repeated which makes its linear approximating model time-varying.

In this thesis, we focus mainly on linear nonstationary systems, because only in this case the analytical evaluation of the presented methods can be carried out. However, as remarked in [69], local basis function methods can be successfully applied to the identification of nonstationary systems and signals as well.

This thesis describes only noncausal methods of identification, which cannot be applied to real-time processing. Nevertheless, when the application allows for some fixed decision delay (almost real-time processing), the techniques presented here can be used. Later, an example of such an application will be given with some experimental results.

1.1.1 Models of linear nonstationary systems

Methods presented in this thesis were developed for nonstationary systems which can be modelled using linear regression model of a form

$$y(t) = \boldsymbol{\theta}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t) + e(t), \tag{1.1}$$

where $t \in \mathbb{Z}$ is a discrete time, $\{e(t)\}$ denotes the measurement noise, $\{y(t)\}$ is the complex-valued output sequence, $\boldsymbol{\theta}(t) = [\theta_1(t), \ldots, \theta_n(t)]^{\mathrm{T}}$ denotes the vector of time-varying, typically unknown parameters, and $\boldsymbol{\varphi}(t) = [\varphi_1(t), \ldots, \varphi_n(t)]^{\mathrm{T}}$ denotes the regression vector, composed of potentially

different input signals. The linear regression model can be used as a description of many different systems, including some of the nonlinear ones, as long as they are linear in parameters. An example of such could be the Hammerstein model [98]. For the discussion of what is, and what is not a subject of this thesis, see the paragraph 1.2.

Probably the simplest, but still powerful model belonging to the class of linear regression models is the finite impulse response model (FIR), defined by the equation

$$y(t) = \sum_{j=1}^{n} \theta_j^*(t) u(t - j + 1) + e(t) = \boldsymbol{\theta}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t) + e(t), \tag{1.2}$$

where in this example the regression vector $\boldsymbol{\varphi}(t) = [u(t), \dots, u(t - n + 1)]^{\mathrm{T}}$ consists of the past samples of a complex-valued input signal $\{u(t)\}$.

The aforementioned description is widely used as a parsimonious model of linear time-varying (LTV) communication channels, both terrestrial and underwater (see [99], [103] or [56] and references therein). The potential time-variability of a wireless communication channel is due to the movement of the transmitter/receiver or scatterers (objects reflecting or refracting the propagation waves), and the delay profile is a consequence of the multipath propagation of the electromagnetic or acoustic wave.

Most of the theoretical results described in this thesis were tested on data, either simulated or real, that arise from the full-duplex (FD) underwater acoustic communication (UAC). In this application, the transmitter and receiver are close to each other and use the same bandwidth. As a result, the signal recorded by the receiver is a mixture of a signal from the far-end transmitter, contaminated with a site-specific, or ambient noise, and the self-interference (SI) signal, which can be modeled as a convolution of a signal from the near-end transmitter with the time-varying impulse response of an underwater channel [92], [94]. The FIR structure of the model reflects the multipath propagation of the signal coming from the nearby transmitter. In this application, $\{y(t)\}$ corresponds to the recorded mixture of signals, $\{u(t)\}$ is a known signal, sent by the near-end antenna, $\{e(t)\}$ is a mixture of a far-end signal and the noise and, $\{\boldsymbol{\theta}(t)\}$ is a vector of unknown, time-varying coefficients of the channel impulse response. To secure the reliable performance of communication devices, $e(t)$ needs to be extracted from $y(t)$, which can be done by subtracting from $y(t)$ the estimated self-interference signal $\widehat{\boldsymbol{\theta}}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t)$, where $\widehat{\boldsymbol{\theta}}(t)$ is the estimated vector of channel parameters.

Interestingly, in this application a fixed decision delay is often acceptable, allowing one to use the noncausal identification methods [92], [94]. Since the techniques investigated in this thesis are based on the idea of fixed-lag smoothing, self-interference cancellation in UAC seems to be the perfect application for the proposed algorithms.

It is worth pointing out that the time-varying FIR model can be also used to approximate the infinite impulse response (IIR) system with a time-varying impulse response, provided that the system of interest is exponentially stable. Such a system can be described by the following ARX (autoregressive with exogenous input) model

$$\begin{aligned} y(t) &= \sum_{i=1}^{r} a_i^*(t) y(t - i) + \sum_{j=1}^{p} b_j^*(t) u(t - j + 1) + e(t), \\ y(t) &= \frac{B(t, q^{-1})}{A(t, q^{-1})} u(t) + \frac{1}{A(t, q^{-1})} e(t), \end{aligned} \tag{1.3}$$

where $B(t, q^{-1}) = b_1^*(t) + \dots + b_p^*(t) q^{-p+1}$, $A(t, q^{-1}) = 1 - a_1^*(t) q^{-1} - \dots - a_r^*(t) q^{-r}$ and q^{-1} is a time lag operator $q^{-1} u(t) = u(t - 1)$.

The stability of a system with an autoregressive (AR) part is guaranteed when at each time instant t, the roots of a "frozen" polynomial $A(t, q^{-1})$ stay strictly away from the unit disc and trajectories of parameters $a_i(t)$, $i = 1, \dots, r$ obey some additional smoothness constraints [58].

It is worth mentioning that some classes of nonstationary signals can be modeled using the AR model, which is a special case of (1.3) when there is no observable input signal. For more information about modeling of nonstationary systems and signals, check [29] and [66].

To briefly explain how different identification methods work, data from a simple two-tap FIR system described by

$$y(t) = \theta_1(t) u(t) + \theta_2(t) u(t - 1) + e(t), \tag{1.4}$$

was generated for parameter changes depicted in figure 1.1 and signal-to-noise ratio (SNR) equal to 25 dB. The input signal is autoregressive $u(t) = 0.8u(t-1) + v(t)$, where $v(t) \sim \mathcal{N}(0,1)$. The SNR is defined as follows

$$\text{SNR} = \frac{\text{E}[|\boldsymbol{\theta}^{\text{H}}(t)\boldsymbol{\varphi}(t)|^2]}{\sigma_e^2}. \tag{1.5}$$
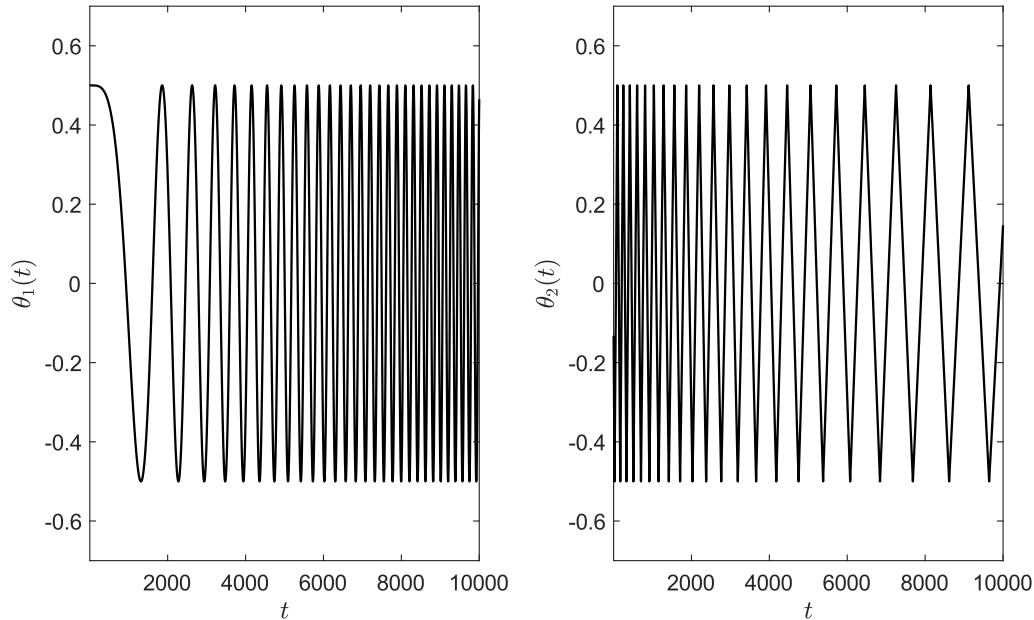

Figure 1.1: Parameter changes for the illustrative scenario.

This data set will be further used for illustrative purposes.

1.2 What this research is not about

The problem of identification of linear time-varying (LTV) systems, described in the previous section, might look similar to the identification of linear parameter varying (LPV) systems (see e.g. [7], [102]). In such systems, parameter changes are invoked by changes in some scheduling variables. As remarked in [34] both LTV and LPV systems belong to the wider class of varying-coefficient models, which can be identified using some general methods. However, the best performance can be achieved with problem-specific solutions. Therefore, this research focuses only on the identification of LTV systems.

All of the models introduced above describe single-input-single-output (SISO) systems. These descriptions can be extended to multiple-input-multiple-output (MIMO) systems. In such extended models, scalar variables have to be replaced by vector variables and vectors of time-varying parameters have to be substituted by matrices. This formulation poses multiple challenges and is not covered in this dissertation.

It is worth emphasizing that the analysis is carried out only for nonstationary FIR systems because some of the crucial assumptions, needed to obtain analytical results, are not met for nonstationary ARX systems and nonstationary AR signals. This does not mean that the proposed methods cannot be used for the identification of time-varying models in both cases mentioned above. The results of such experiments are presented in the next chapters. Moreover, all systems analyzed in this thesis are not reducible to stationary systems. For reducible processes, a global, time-invariant "hypermodel" can be found, while systems analyzed in this dissertation can be only locally approximated by a time-invariant description.

Another, wide class of systems is described by various types of nonlinear models. The problem of identification of nonlinear systems is very general, and there exist various identification techniques developed for specific nonlinear models. Some of these models can be expressed or at least

approximated by linear, with respect to parameters, and time-varying models, but this thread was not covered in this thesis.

The subject of this research is the identification of time-varying parametric models of nonstationary systems. The obtained models can be further used e.g. for self-interference cancellation in communication channels or detection of abrupt changes in an analysed system. Finally, this thesis does not cover methods of spectral analysis of nonstationary signals.

1.3   Overview of known identification methods developed for nonstationary systems/signals

Identification of LTV systems can be accomplished using one of two different approaches. If parameters vary sufficiently slowly, one might assume that they are locally constant and for the purpose of their estimation use the localized version of classical identification techniques such as weighted least squares (WLS) or gradient-based least mean squares (LMS) [5], [66], [98]. However, when parameter changes are faster, better results can be obtained by using an explicit model of parameter changes, usually called a hypermodel. The hypermodel can be either stochastic or deterministic.

1.3.1   Stochastic model of parameter changes

In the first case, one can use the integrated random walk (IRW) description of parameter changes. Then the task of estimation of model parameters can be transformed to the problem of tracking of a vector of properly defined state variables. In such a setup the Kalman filter is a commonly used tool to solve this problem, as described in papers [41], [67], [80], and recently in the article [17]. The bridge between the stochastic and deterministic approaches can be established when it is assumed that parameter trajectories are realizations of some random process with a known correlation matrix. In such a case, one can use functions resulting from the Karhunen-Loéve theorem [39] (such functions will be further reffered to as KL functions) to locally model parameter changes.

1.3.2   Deterministic model of parameter changes

The alternative is to assume a deterministic model of parameter changes. In this approach, one uses assumption that inside the analysis window of a finite length, parameter changes can be described by linear combinations of some known functions of time, the so-called basis functions. The origins of the basis function (BF) method can be traced down to the paper [87], where Taylor approximation was used to model parameter changes of an AR process. Then the basis properties were utilized to derive expressions for the bias and variance of this estimator, needed to evaluate the mean squared parameter estimation error (MSE). The article proposed the method of weighted least squares for parameter estimation and compared two different weighting sequences.

Later, the BF approach was further extended and used for identification of parameters of models of many different processes like speech, EEG signals or seismic activity records. An example of an article describing application of the BF method to identification of an autoregressive model of speech signal is [49]. In this article the high computational burden associated with large dimensions of the regression matrix was pointed out and as a solution, the algorithm based on the Gram-Schmidt procedure to orthogonalize regressors and obtain estimates in an order-recursive manner was proposed. This procedure also involves an order selection rule, based on the stability assessment of the created AR model. The method of orthogonalization in the BF approach was later rediscovered as the orthogonal least squares (OLS) method, and described in [14] for the purpose of identification of nonlinear stationary processes. Later it was also successfully applied to identification of LTV systems [47], [106], [107].

Soon, the other choices of basis functions (like cosinusoidal, prolate spheroidal wave functions [97], or Daubechies wavelets) were investigated and the identification schemes for ARMA (autoregressive-moving-average) models were proposed [29], [33], [104].

The article [64] provides the frequency and tracking performance analysis of BF estimator for nonstationary FIR systems. In this article, it was shown that the expected trajectories of parameter estimates can be seen as the effect of filtering of the true parameter trajectories with a linear filter associated with the BF method. Of course, the impulse response of such a filter depends on the chosen basis and the type of the weighting sequence (if weighted least squares algorithm is used).

In most applications, the basis in the BF method contains the constant function. This is typical because when no prior knowledge about parameter changes is available then the first guess should be that parameters are time-invariant. It can be shown [66] that when either the constant function is one of the chosen functions, or it belongs to the subspace spanned by the basis functions, then the filter associated with the BF method is a lowpass filter.

It is worth noting that BF methods are unbiased only if the design assumption is fulfilled, namely if parameter trajectories belong to the subspace spanned by the chosen functions. Articles [64] and [37] confirm that this subspace is in fact a crucial factor in estimation and two different bases, spanning the same subspace will produce exactly the same estimator. This information is the main clue when choosing the type of the basis functions. For example, if one expects some periodicity in parameter changes, the Fourier basis will be more suitable than the Taylor basis. Additionally, [37] discusses the conditional biases and variances of the BF method, provides the assumptions needed for consistency, and provides the information about the convergence rate of the algorithm.

An interesting approach was developed in a series of papers [47], [106] and [107]. These articles describe parameter changes with wavelets, such an approach was first proposed in [104]. In the papers cardinal B-splines were used. Using wavelets is the first step towards localized analysis since the support of wavelets is limited and the basis typically contains wavelets of different resolutions. These facts make them well-suited for local modelling of parameter changes. The price for sparseness and different resolutions of wavelet basis is a very high dimensionality. In [104] the approach based on consecutive addition of regressors was proposed. To determine which wavelets carry important information, Student's t-test was used. In the aforementioned series of papers, the OLS algorithm with AIC criterion [14] was used. This approach was successfully tested on EEG signals.

All papers discussed so far present the framewise approach, in which the estimated basis function coefficients are used to obtain system parameter estimates for the entire analysis interval. The next step in the development of BF methods was performed in [69], where the samples surrounding the observation at the current time instant are used to find basis function coefficients. These coefficients are then used to obtain estimates of model parametrs for the current time instant only. Next, the analysis window is shifted by one sample and the procedure is repeated for the new position of the analysis window. The paper [69] also contains the analytical comparison between the BF and the new LBF (local BF) approach including comparison of their tracking performance and statistical properties. The BF method is computationally demanding because the number of estimated coefficients is a product of the number of system parameters and the basis order. In the new approach, the least squares (LS) procedure is carried out for every position of the analysis window which additionally increases computational costs. This is why a lot of emphasis was put on the use of recursively computable basis functions and weighting sequences, which allows one to recursively update the regression matrix. Recently the dichotomous coordinate descent (DCD) algorithm, firstly described in [111], was combined with the basis function method [93], which allows one to substantially reduce the computational burden associated with the inversion of the regression matrix.

An overview and detailed analysis of methods of identification of nonstationary processes can be found in [66].

1.4   Research justification

The state-of-the-art LBF and multi-wavelet (MW) approaches (see [47], [106] and [107]), described in the previous section, provide high-precision estimates of system trajectories. However, the computational burden associated with these methods is quite high. This prevents their application in many fields requiring almost real-time processing. This means that, in such applications, a fast version of the LBF algorithm (fLBF), providing estimates of similar or higher accuracy, would be very helpful. Therefore, the development of the new fLBF method is the main topic of the thesis. The second topic is focused on the improvement of the accuracy of LBF and fLBF estimates. In many applications, the accuracy of estimates is the most significant feature and even its slight improvement can be of great practical importance, especially when the associated computational cost remains low. As remarked in the next section, regularization can be a useful tool for increasing estimation accuracy by achieving a better bias-variance trade-off. This dissertation also describes the methods for adaptive choice of design parameters. As stated in the previous section, since
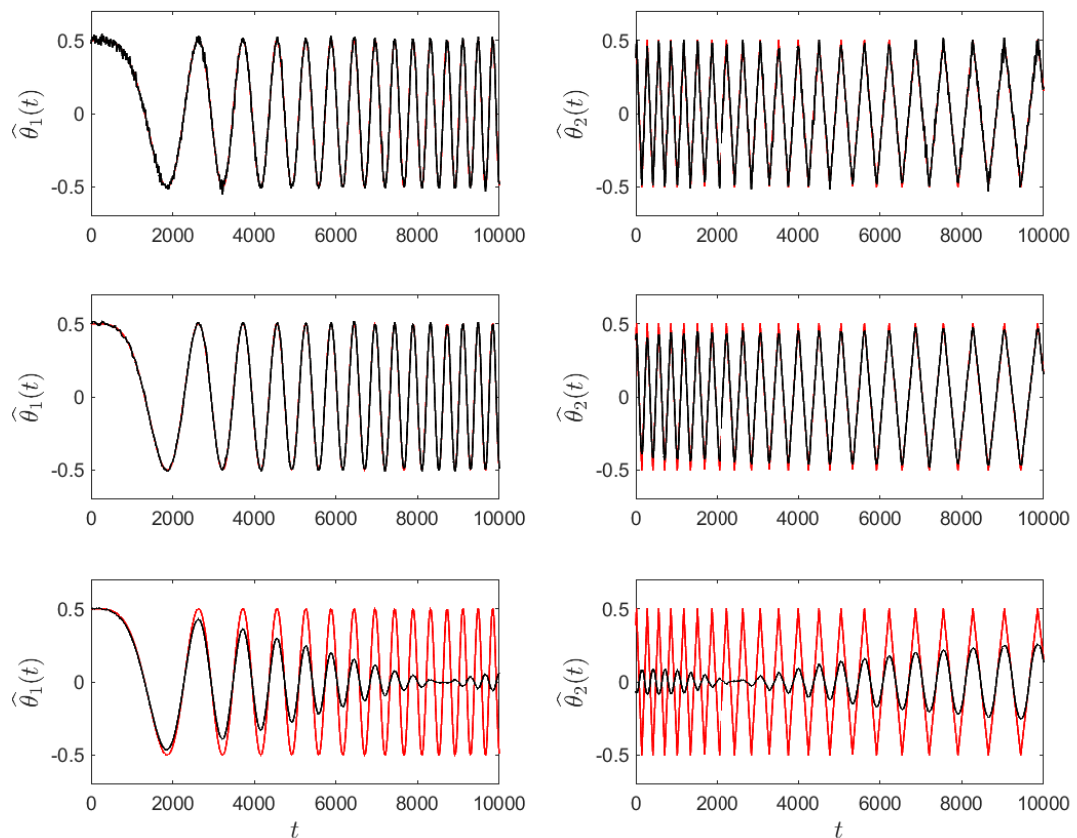
Figure 1.2: LBF estimates (black lines) of the exemplary two-tap FIR system parameter trajectories for $m = 5$, polynomial functions and window length equal to $K = 101$ (left and right upper plots), $m = 3$ and $K = 201$ (left and right middle plots) and $m = 1$, and $K = 400$ (left and right lower plots). All of the plots were obtained for polynomial basis functions (see Chapter 2.)

typically the ratio of system nonstationarity is unknown and can vary with time, the adaptive methods for hyperparameters tuning are an important part of identification schemes.

1.5 Overview of new results presented in the thesis

Apart from providing some new analytical results for LBF estimators, the subject of this dissertation is developement of a new fast version of the LBF method (fast LBF - fLBF). This novel approach consists of two steps - preestimation and postfiltering. The first part of this new approach produces raw estimates of system trajectories that are unbiased regardless of the type and speed of parameter changes. However, the cost that one needs to pay for unbiasedness is a very high variance of the preestimation noise (exemplary preestimates are shown in figure 1.3). This is why preestimates cannot be used as final estimates and additional filtering is needed. It is easy to show that when the impulse response of the filter used in the second step is the same as the impulse response of the filter associated with the LBF method, then the fLBF estimates closely approximate the LBF estimates.

1.5.1 Preestimation

Preestimates were introduced in [76] for the purpose of causal estimation and later used for noncausal estimation in [70], [71] and [73]. In [71] the direct approximation of the expression for ideal preestimates was presented and its connection to the LBF method was established.
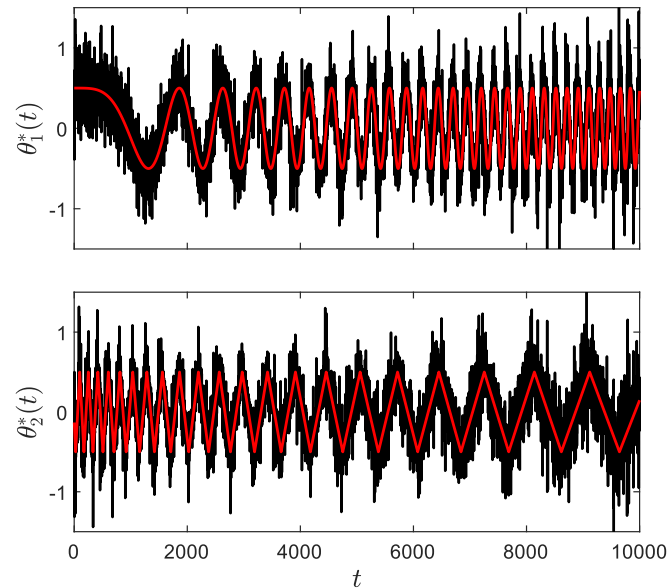
Figure 1.3: Indirect preestimates (black lines) and the true parameter trajectories (red lines).

Later, in [70] the indirect preestimation method, introduced earlier in [76], was investigated. It was also shown that the preestimation errors for the indirect method are smaller than for the direct one. This article also provided the theoretical lower bound on the variance of preestimation errors.

[73] provides the iterative procedure for improving the accuracy of preestimates by reducing one of the two major components of preestimation errors.

In [74] the preestimation technique is extended to complex-valued systems and the new, computationally cheaper variant of the indirect preestimates is presented.

The aforementioned articles describe causal preestimation techniques (except [71]) which are later used in a noncausal estimation scheme. As noted in [77], preestimation can be also noncausal (especially when the input-output sequences are prerecorded). It was demonstrated that the proper design of bidirectional preestimates can be beneficial when some parameter trajectories are subject to abrupt changes.

1.5.2   Postfiltering

The very first filtering technique adopted for the fLBf method in [71] and [70] was the Savitzky-Golay filtering [90], [91], the reason being that when the basis functions in Savitzky-Golay filter coincide with the basis functions in the LBF method, then the fLBF estimates can be seen as tight approximations of LBF estimates (in such a case the impulse response of the filter used in the postfiltering step is the same as the impulse response of the filter associated with the LBF method). Obviously, any well-established filtering technique can be applied at this stage, which was shown in [17] where the Kalman filter was used for this purpose, and in [77] where the windowing technique was used to design a general purpose lowpass filter.

1.5.3   Hyperparameter tunning

The crucial design parameters for the BF methods are the length of the local analysis window K and the number of basis functions m. Typically, when m increases, so does the model capacity which results in a greater variance but the smaller bias of parameter estimates. The opposite effect can be obtained by increasing K, because including more information in the identification process clearly reduces the variance at the cost of increased bias. An example of such an effect is depicted in figure 1.2. Therefore, when the speed and type of parameter changes are not known, one should use some procedure for adaptive choice of m and K. It is worth noting that since

the aforementioned parameters have the opposite effect on the bias-variance trade-off, one could as well fix one of them and tune the other one. One of the known solutions to this problem is model validation based on statistical tests [14], [49], [104]. Another approach would be to use the modified Akaike's final prediction error (FPE) criterion, [69], [68] or leave-one-out cross-validation (LOOCV) statistic [6], [37], [69]. It is also worth noting that some regularization techniques can be interpreted as mechanisms for rejecting the spurious parameters [13], [62], [100] and therefore can be seen as model order selection methods.

In [69] the algorithm for adaptive choice of hyperparameters m and k was proposed. In this scheme, several estimators, equipped with different settings, compete against each other. Finally, at each time instant, the estimator minimising the local quality measure is chosen. [69] compares two quality measures for the LBF algorithm - the sum of squared leave-one-out cross-validation errors and generalized Akaike's Final Prediction Error criterion.

The LOOCV criterion was also proposed for the fLBF method in the paper [70]. The same strategy was proven useful also for choosing regularization gain [22] and regularization matrix [74], [79].

1.5.4 Regularization

Suppose that $\{\boldsymbol{\theta}(t)\}$ is a deterministic function. The mean squared parameter estimation error can be decomposed into a bias and variance components

$$\mathrm{E}\{[\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}(t)]^2\} = ||\overline{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)||^2 + \mathrm{Tr}\Big\{\mathrm{cov}\Big[\widehat{\boldsymbol{\theta}}(t)\Big]\Big\} \tag{1.6}$$

where

$$\overline{\boldsymbol{\theta}}(t) = \mathrm{E}[\widehat{\boldsymbol{\theta}}(t)]$$

and the expectation is taken over different realizations of measurement noise and input signal. The improvement of the MSE score can be achieved by using biased estimators characterized by smaller variance [18], [36]. This can be achieved by means of regularization. Originally regularization was introduced as a way of imposing smoothness constraints on the solution of the least squares problem [108] and later it was rediscovered as a way of solving ill-posed inverse problems [81], [101].

The idea of regularization was recently heavily exploited in the identification of time-invariant systems. The new trend started from continuous-time systems [15], [82] and was later extended to discrete-time systems [16], [51], [52], [60], [83]. The above-mentioned articles provide justification for the regularization and demonstrate possible advantages of this technique when the FIR description of a system is used. They also compare different choices of regularization kernels [15], and analyse their influence on the MSE score [60]. Also, several algorithms for choosing regularization constants [16], [51] were compared and the Bayesian interpretation of regularization was provided.

As far as the identification of time-varying systems is concerned, regularization was present only in methods assuming a stochastic model of parameter changes [41]. In this article, the empirical Bayes procedure (earlier referred to as Type II Bayesian procedure [3]) was used to tune the regularization method.

Inspired by these results, first attempts to introduce the regularization technique in the BF method were made [22], [23], [24], [74], [79]. All of these were constructed by addressing four fundamental questions about regularization

1. What is being constrained? - system parameters $\boldsymbol{\theta}(t)$ or hypermodel coefficients $\boldsymbol{\alpha}(t)$ (defined later).

2. What penalty is being used? - ℓ^2, ℓ^1 or reweighted ℓ^2.

3. What prior knowledge is available? - no prior knowledge, partial prior knowledge (e.g. power profile of the time-varying impulse response or degree of smoothness of the impulse response), or full (statistical) prior knowledge.

4. What optimization technique is used? - cross-validation or empirical Bayes.

In [24] the formula for the optimal diagonal regularization matrix (in terms of the MSE) was developed when no prior knowledge about parameter changes is available. In this early work on

regularization in LTV systems identification, the excessive values of coefficients $\boldsymbol{\alpha}(t)$ were penalized. It is worth mentioning that in this paper the FPE criterion was also proposed for choosing m and K, and the results were compared with the approach based on LOOCV.

In [74] the extension of the aforementioned article was provided. Namely, the formula for the optimal full regularization matrix was proposed when the penalty is applied to parameters $\boldsymbol{\theta}(t)$. Moreover, in this article, the adaptive technique for choosing the regularization matrix was developed. It is worth mentioning that the optimal regularization matrix from this article requires prior knowledge about the covariance matrix of parameters $\boldsymbol{\theta}(t)$.

[79] presents the approach utilising partial prior knowledge to design the regularization matrix. This article proposes two different ways of designing the regularization matrix. The first one relies on the knowledge about the degree of smoothness of the time-varying impulse response of the system. The second one uses the assumption about exponential stability of the impulse response of a "frozen" system at each time instant t. To optimize the value of regularization gain, the empirical Bayes approach was used.

[23] and [22] again use the assumption that no prior knowledge about parameter changes is available. In the first paper, the regularization gain is chosen using generalized cross-validation (GCV) [27] that already has been proven useful in regularized identification of LTI systems [59]. The second paper compares traditional ridge regression with reweighted ℓ^2 regularization which can be interpreted as the approximation of ℓ^1 regression [100].

1.5.5 The author's main contribution

The author considers the following achievements as his most important contributions to the research

- Development of the procedure for improving unidirectional preestimates, called enhanced preestimation. The aforementioned technique is an iterative procedure of improving preestimates based on the understanding of the source of preestimation errors. This algorithm was included in the conference paper [C3], and enabled an analytical explanation of differences between various types of preestimates.

- Derivation of the optimal in the mean squared sense, diagonal regularization matrix for identification of nonstationary FIR systems and development of suboptimal but feasible formula for regularization matrix. This achievement was the main subject of the paper [J2] published in IEEE Transactions on Signal Processing.

- Derivation of the closed-form solution of the generalized cross-validation optimization problem in identification of nonstationary FIR systems and statistical interpretation of the obtained formula. This result was presented in the conference paper [C5]

- Development of reweighted ℓ^2 algorithm in the identification of time-varying systems with its interpretation in terms of approximation of ℓ^1 regression. The proposal of leave-one-out cross-validation based scheme for adaptive choice of the regularization gain. These results became a part of the conference paper [C6]

- Development of an algorithm deciding adaptively whether the identified parameter is locally constant or time-varying. This scheme is one of the main contributions of the article published in Signal Processing [J7] and work presented during ICASSP 2022 [C7].

- Proposition of application of basis functions stemming from Karhunen-Loéve decomposition to tracking of rapidly fading communication channels, and showing their connection to the optimal basis functions. This results were gathered in a paper published in Signal Processing [J8].

- Development of robustified version of the fast local basis function method, based on iterative reweighting technique. This method was described in a conference paper [C10].

Co-authored publications

<u>Journal articles</u>

[J1] M. Niedźwiecki, M. Ciołek and A. Gańcza, "A new look at the statistical identification of nonstationary systems". In: *Automatica*, vol. 118 (2020), paper no. 109037.

[J2] A. Gańcza, M. Niedźwiecki and M. Ciołek, "Regularized local basis function approach to identification of nonstationary processes". In: *IEEE Transactions on Signal Processing*, vol. 69 (2021), paper no. 9363574.

[J3] M. Ciołek, M. Niedźwiecki and A. Gańcza, "Decoupled Kalman filter based identification of time-varying FIR systems". In: *IEEE Access*, vol. 9 (2021), paper no. 9435188.

[J4] M. Niedźwiecki, M. Ciołek, A. Gańcza and P. Kaczmarek, "Application of regularized Savitzky–Golay filters to identification of time-varying systems". In: *Automatica*, vol. 133 (2021), paper no. 109865.

[J5] M. Niedźwiecki, A. Gańcza and P. Kaczmarek, "Identification of fast time-varying communication channels using the preestimation technique". In: *IFAC-PapersOnLine*, vol. 54 (2021), pp. 351-356.

[J6] L. Shen, Y. Zakharov, M. Niedźwiecki and A. Gańcza, "Finite-window RLS algorithms". In: *Signal Processing*, vol. 198 (2022), paper no. 108599.

[J7] M. Niedźwiecki, A. Gańcza, L. Shen and Y. Zakharov, "Adaptive identification of sparse underwater acoustic channels with a mix of static and time-varying parameters". In: *Signal Processing*, vol. 200 (2022), paper no. 108664.

[J8] M. Niedźwiecki and A. Gańcza, "Karhunen-Loéve-based approach to tracking of rapidly fading wireless communication channels". In: *Signal Processing* (2023), vol. 209, paper no. 109043.

Conference articles

[C1] M. Niedźwiecki, M. Ciołek and A. Gańcza, "Fast basis function estimators for identification of nonstationary stochastic processes". In: *27th European Signal Processing Conference EUSIPCO 2019* (A Coruna, Spain). 2019.

[C2] M. Niedźwiecki, M. Ciołek and A. Gańcza, "Local basis function estimators for identification of nonstationary systems". In: *58th IEEE Conference on Decision and Control CDC 2019* (Nice, France) pp. 777-783. 2019.

[C3] M. Niedźwiecki, A. Gańcza and M. Ciołek, "On the preestimation technique and its application to identification of nonstationary systems". In: *59th IEEE Conference on Decision and Control CDC 2020* (Jeju Island, Republic of Korea), pp. 286-293. 2020.

[C4] M. Niedźwiecki, M. Ciołek, A. Gańcza and P. Kaczmarek, "A new method of noncausal identification of time-varying systems". In: *IEEE Conference on Signal Processing: Algorithms, Architectures, Arrangements and Applications, SPA 2020.* (Poznań, Poland), pp. 48-52. 2020.

[C5] A. Gańcza and M. Niedźwiecki, "Regularized identification of time-varying FIR systems based on Generalized Cross-Validation". In: *29th European Signal Processing Conference EUSIPCO 2021* (Dublin, Ireland). 2021.

[C6] A. Gańcza and M. Niedźwiecki, "Regularized identification of fast time-varying systems - comparison of two regularization strategies". In: *60th IEEE Conference on Decision and Control CDC 2021* (Austin, USA). 2021.

[C7] M. Niedźwiecki, A. Gańcza, L. Shen and Y. Zakharov, "Adaptive identification of underwater acoustic channel with a mix of static and time-varying parameters". In: *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Singapore). 2022.

[C8] M. Niedźwiecki and A. Gańcza, "Optimally regularized local basis function approach to identification of time-varying systems". In: *61st IEEE Conference on Decision and Control CDC 2022* (Cancún, Mexico). 2022.

[C9] M. Niedźwiecki, A. Gańcza, L. Shen and Y. Zakharov "On bidirectional preestimates and their application to identification of fast time-varying systems". In: *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Rhodes, Greece). 2023.

[C10] A. Gańcza, M Chełkowska and N. Kleinschmidt "Towards robust identification of nonstationary systems". In: *31st European Signal Processing Conference EUSIPCO 2023* (Helsinki, Finland). 2023.

[C11] M. Niedźwiecki and A. Gańcza, "Fast algorithms for identification of time-varying systems with both smooth and discontinuous parameter changes". In: *62nd IEEE Conference on Decision and Control CDC 2023* (Singapore, Singapore). 2023.

1.6   Objectives of the thesis

The aim of this thesis is to describe the new methods for the identification of nonstationary systems - the LBF and fLBF algorithms. Within the thesis one will find the answers to the following questions, which did not have clear answers before this research:

1. What is the performance of the fLBF method compared to the LBF method?

2. How to adaptively choose the number of hyperparameters in investigated methods?

3. How to eliminate the delay that occurs in preestimates based on exponentially weighted least squares estimates of system parameters?

4. What is the performance of the fLBF algorithm with different preestimation techniques?

5. What is the role of regularization in LBF and fLBF methods?

6. How to design the regularization matrix and how to optimize its hyperparameters?

Note that the answers to these questions are not gathered in one place, they unfold gradually throughout the thesis.

Chapter 2

Basis function method

2.1 Introduction

In the first part of this chapter, the traditional BF estimator is characterized and its statistical properties are derived. Later, it is pointed out that the LBF method can be obtained by a simple change in estimation strategy and that the statistical properties of this estimator follow directly from the properties of a traditional BF method.

The starting point of a BF method and its essential part is an assumption

> (A2.1) Inside the local analysis window $T_k(t) = [t-k, t+k]$ of length $K = 2k+1$, centered at t, parameter trajectories can be expressed as a linear combination of m linearly independent functions of time
>
> $$\theta_j(t+i) = \sum_{l=1}^{m} f_{l|k}^*(i)\alpha_{j,l}(t), \quad j = 1, \ldots, n; \; i \in [-k, k] \cap \mathbb{Z} = I_k, \qquad (2.1)$$
>
> where $\alpha_{j,l}(t)$ is a basis function coefficient and $\{f_l(i)\}$ represents a known function of time.

Note, that it is assumed, that coefficients $\alpha_{j,l}(t)$ are constant inside the analysis window. However, since their values might change with the position of the analysis window, they are written down as functions of time.

It is worth noting that this assumption, although crucial for derivation, is not needed to be fulfilled in practice. In most real-life applications, no prior knowledge about parameter changes is available, and BF methods can be still applied and they produce satisfactory results.

This work presents all methods for odd values of K. This is not a requirement for the BF method and all of the derivations remain true also for even lengths of the analysis interval. This notation was adopted only because the LBF and fLBF methods are defined in this particular way, hence the author wanted to avoid any possible confusion stemming from changes in notation.

Remark

Since, for now, we do not specify the way the next position of the analysis window depends on the previous position, the more precise notation would be $\alpha_{j,l}[T_k(t)]$, which for the sake of brevity is replaced from now on by simplified notation $\alpha_{j,l}(t)$.

2.1.1 Basis functions

Basis functions are bounded deterministic sequences defined on I_k. For the majority of the time, the only requirement for them will be that the vectors of basis functions have to constitute a basis in the m-dimensional space. Only for the purpose of derivation of statistical properties of basis function methods, we will need to impose some restrictions on the way the basis functions are generated, however, the aforementioned restrictions are not very limiting and can be easily fulfilled in practice.

The ideal basis should of course span the subspace that contains the true parameter trajectories. Moreover, it is often required that the basis is sparse, guaranteeing computational advantages. If

23

there is available a certain prior knowledge about the type of parameter changes, the basis can be chosen so as to reflect dominant patterns present in parameter trajectories, e.g. for piecewise constant parameter trajectories, the Haar wavelets could be applied successfully. When parameter trajectories are known to have limited bandwidth, then a certain number of prolate spheroidal wave functions are guaranteed to provide the best local approximation in the mean squared sense [44]. One can also assume that parameter trajectories are realizations of a random process with a known correlation matrix. In this case, the functions originating from the Karhunen-Loéve theorem [39] (further referred to as KL functions) are proven to provide an optimal reconstruction of the unknown trajectory. Note that the last approach can be seen as a bridge between stochastic and deterministic modelling of parameter variations.

When no prior knowledge is available (which is typical), basis functions of a high generalization capability should be used. Examples of such general-purpose basis functions, that are frequently used, are

- polynomial basis, which allows one for the local Taylor approximation of parameter trajectories

$$f_{l|k}(i) = \left(\frac{i}{k}\right)^{l-1}, \quad i \in I_k, \; l = 1, \dots, m$$

  The shape of first five Legendre polynomials (orthonormal polynomial basis functions) was presented in figure 2.1.

- sine-cosine basis, which allows one for the the local Fourier approximation of parameter trajectories

$$f_{l|k}(i) = \begin{cases} \cos \frac{\pi i (l-1)}{2k}, & l \equiv 0 \;(\mathrm{mod}2) \\ \sin \frac{\pi i (l-1)}{2k}, & l \equiv 1 \;(\mathrm{mod}2) \end{cases}, \quad i \in I_k, \; l = 1, \dots, m$$

- complex exponential basis

$$f_{l|k}(i) = \frac{1}{\sqrt{K}} e^{\frac{\omega_l \mathrm{i}}{K}}, \quad i \in I_k, \; l = 1, \dots, m,$$

  where $\omega_l = (-1)^{l-1} 2\pi \lceil \frac{l}{2} \rceil$ and $\lceil x \rceil$ denotes the smallest integer larger than x (the so-called ceil function).

Note that, both polynomial and sine-cosine functions are already presented in a normalized (but not orthogonal) form and complex exponential basis functions are already orthonormal.

2.1.2 BF estimator

If the vector notation is incorporated, assumption (A2.1) can be expressed as follows

$$\theta_j(t+i) = \mathbf{f}_{m|k}^{\mathrm{H}}(i)\boldsymbol{\alpha}_{j,m|k}(t), \quad j = 1, \dots, n; \; i \in I_k, \tag{2.2}$$

where

$$\mathbf{f}_{m|k}(i) = [f_{1|k}(i), \dots, f_{m|k}(i)]^{\mathrm{T}}$$
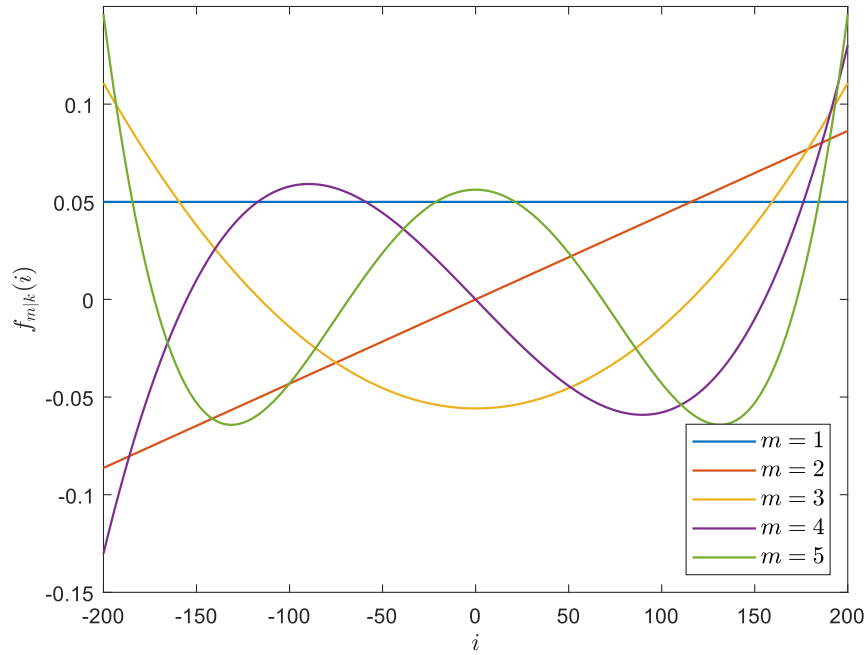$$\boldsymbol{\alpha}_{j,m|k}(t) = [\alpha_{j,m|k}^1(t), \dots, \alpha_{j,m|k}^m(t)]^{\mathrm{T}},$$

or even in a more compact form

$$\boldsymbol{\theta}(t+i) = \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}_{m|k}(t), \quad i \in I_k, \tag{2.3}$$

where

$$\boldsymbol{\alpha}_{m|k}(t) = [\boldsymbol{\alpha}_{1,m|k}^{\mathrm{T}}(t), \dots, \boldsymbol{\alpha}_{n,m|k}^{\mathrm{T}}(t)]^{\mathrm{T}}$$
$$\mathbf{F}_{m|k}(i) = \mathbf{I}_n \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(i),$$

and $\otimes$ denotes the Kronecker product of corresponding vectors/matrices. The subscript $m|k$ was added to indicate that the shape and properties of basis functions, as well as the properties of the BF estimator, depend on the choice of these two hyperparameters.

Figure 2.1: Five first Legendre polynomials for $k = 200$.

Under (A2.1) the system (1.2) can be locally described using a higher-order, but time-invariant model

$$y(t + i) = \boldsymbol{\alpha}_{m|k}^{\mathrm{H}}(t)\boldsymbol{\psi}_{m|k}(t, i) + e(t + i), \quad i \in I_k, \tag{2.4}$$

where $\boldsymbol{\psi}_{m|k}(t, i) = \boldsymbol{\varphi}(t) \otimes \mathbf{f}_{m|k}(i)$ denotes the generalized regression vector. The estimates of coefficients $\boldsymbol{\alpha}_{m|k}(t)$ can be found using the WLS method

$$\widehat{\boldsymbol{\alpha}}_{m|k}(t) = \arg\min_{\boldsymbol{\alpha}} \sum_{i=-k}^{k} w_k(i)|y(t + i) - \boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\psi}_{m|k}(t, i)|^2, \tag{2.5}$$

where $\{w_k(i) \in \mathbb{R}_+, \ i \in I_k\}$ denotes a symmetrical, nonnegative weighting sequence, which is nonincreasing for $i \geq 0$ and $w_k(0) = 1$. It is worth mentioning that the most reasonable choice of a weighting function for the BF approach is a uniform (nonpreferential) weighting $w_k(i) \equiv 1$, $i \in I_k$. However, other (especially bell-shaped) weighting sequences are frequently used in the LBF method since they can substantially increase the estimation accuracy by applying more emphasis on data closer to the centre of the analysis window. Two important examples of such weighting sequences are

- cosinusoidal window

$$w_k(i) = \cos\frac{\pi i}{2k}, \quad i \in I_k.$$

- Hann window

$$w_k(i) = 0.5\left(1 + \cos\frac{\pi i}{k}\right).$$

Note that the cost function optimized in (2.5) is a scalar, real-valued function of a multidimensional complex-valued variable, which makes the gradient-based optimization less straightforward. This can be done using the so-called Wirtinger derivative (also known as the $\mathbb{CR}$-calculus) [21], [109]. For the explicit solution to this problem and additional information on this topic see appendix A of this dissertation. Although the optimization procedure is more challenging than for real numbers, the result remains very similar

$$\widehat{\boldsymbol{\alpha}}_{m|k}(t) = \mathbf{P}_{m|k}^{-1}(t)\mathbf{p}_{m|k}(t), \tag{2.6}$$

$$\mathbf{P}_{m|k}(t) = \sum_{i=-k}^{k} w_k(i) \boldsymbol{\psi}_{m|k}(t,i) \boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,i) > 0$$

$$\mathbf{p}_{m|k}(t) = \sum_{i=-k}^{k} w_k(i) y^*(t+i) \boldsymbol{\psi}_{m|k}(t,i),$$

(2.7)

provided that the generalized regression matrix $\mathbf{P}_{m|k}(t)$ is nonsingular.

The parameter estimates can be obtained using

$$\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+i|t) = \mathbf{F}_{m|k}(i) \widehat{\boldsymbol{\alpha}}_{m|k}(t), \quad i \in I_k.$$

(2.8)

The notation $\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+i|t)$ is used to emphasize, that the estimates are obtained for the analysis window, centered around t.

Remark - numerical well-conditioning

It is clearly visible that the solution to (2.6) depends on the well-conditioning of the matrix $\mathbf{P}_{m|k}(t)$. For $m = 1$, in which case the basis function method reduces down to the WLS method, the stochastic conditions for invertibility of the generalized regression matrix $\mathbf{P}_{m|k}(t)$ were given in [75]. However, in general when $m \geq 1$ the invertibility conditions are unknown. Nevertheless, in practice, the generalized regression matrix could only be ill-conditioned when the ratio K/mn becomes close to 1.

Remark - segmentation

It would be rather naive to assume that coefficients $\boldsymbol{\alpha}(t)$ and the set of chosen basis functions can characterize parameter trajectories in the entire time domain. This would mean that the time-varying model of a nonstationary process is in fact reducible to the time-invatiant model, and that the methods of identification of stationary systems could be used for such process. In practice, some segmentation of the available data is performed firstly (see [66] or [45] for a discussion on this topic), and then the BF method is applied to each segment. This approach, however, might invoke discontinuities in parameters' estimates on the frame borders. Additionally, the accuracy of estimates usually deteriorates at the borders of the analysis window, because the estimates at the end of a frame are functions of the past or "future" data only. This problem can be alleviated by applying a simple "clipping" technique [65]. At the cost of a slight increase of computational complexity (since it requires using overlapping frames), it allows one to obtain higher accuracy estimates.

2.2   Algebraic properties of the basis function method

For brevity and clarity of analytical results, the following assumption is adopted hereafter

> (A2.2) The chosen basis functions are w_k-orthonormal, namely
>
> $$\sum_{i=-k}^{k} w_k(i) \mathbf{f}_{m|k}(i) \mathbf{f}_{m|k}^{\mathrm{H}}(i) = \mathbf{I}_m$$

From an analytical standpoint (A2.2) is by no means critical. Moreover, any set of linearly independent basis functions can be w_k-orthonormalized, e.g. using the modified Gram-Schmidt method [28].

2.2.1   Change of coordinates

Suppose that the originally chosen functions are transformed using a nonsingular matrix $\mathbf{A}$ which does not change the subspace spanned by the basis functions but only transforms the coordinates

$\widetilde{\mathbf{f}}_{m|k}(i) = \mathbf{A}\mathbf{f}_{m|k}(i)$. The estimates obtained with the new basis functions have the form

$$\widetilde{\boldsymbol{\alpha}}_{m|k}(t) = \arg\min_{\boldsymbol{\alpha}} \sum_{i=-k}^{k} w_k(i)|y(t+i) - \boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\varphi}(t+i) \otimes \mathbf{A}\mathbf{f}_{m|k}(i)|^2 = \widetilde{\mathbf{P}}_{m|k}^{-1}(t)\widetilde{\mathbf{p}}_{m|k}(t), \tag{2.9}$$

where

$$\begin{aligned}
\widetilde{\mathbf{P}}_{m|k}(t) &= \sum_{i=-k}^{k} w_k(i)\widetilde{\boldsymbol{\psi}}_{m|k}(t,i)\widetilde{\boldsymbol{\psi}}_{m|k}^{\mathrm{H}}(t,i) = \sum_{i=-k}^{k} w_k(i) \left[\boldsymbol{\varphi}(t+i) \otimes \mathbf{A}\mathbf{f}_{m|k}(i)\right] \left[\boldsymbol{\varphi}^{\mathrm{H}}(t+i) \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(i)\mathbf{A}^{\mathrm{H}}\right] \\
&= (\mathbf{I}_n \otimes \mathbf{A})\mathbf{P}_{m|k}(t)(\mathbf{I}_n \otimes \mathbf{A}^{\mathrm{H}}),
\end{aligned}$$
$$\tag{2.10}$$

$$\widetilde{\mathbf{p}}_{m|k}(t) = \sum_{i=-k}^{k} w_k(i)\widetilde{\boldsymbol{\psi}}_{m|k}(t,i)y^*(t+i) = \sum_{i=-k}^{k} w_k(i) \left[\boldsymbol{\varphi}(t+i) \otimes \mathbf{A}\mathbf{f}_{m|k}(i)\right] y^*(t+i) = (\mathbf{I}_n \otimes \mathbf{A})\mathbf{p}_{m|k}(t).$$
$$\tag{2.11}$$

Note that since

$$\widetilde{\boldsymbol{\alpha}}_{m|k}(t) = (\mathbf{I}_n \otimes \mathbf{A}^{\mathrm{H}})^{-1}\widehat{\boldsymbol{\alpha}}_{m|k}(t), \tag{2.12}$$

and

$$\widetilde{\boldsymbol{\theta}}_{m|k}(t+i|t) = \widetilde{\mathbf{F}}_{m|k}(i)\widetilde{\boldsymbol{\alpha}}_{m|k}(t), \tag{2.13}$$

where $\widetilde{\mathbf{F}}_{m|k}(i) = \mathbf{I}_n \otimes \widetilde{\mathbf{f}}_{m|k}^{\mathrm{H}}(i)$, it holds that

$$\widetilde{\boldsymbol{\theta}}_{m|k}(t+i|t) = [\mathbf{I}_n \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(i)\mathbf{A}^{\mathrm{H}}](\mathbf{I}_n \otimes \mathbf{A}^{\mathrm{H}})^{-1}\widehat{\boldsymbol{\alpha}}_{m|k}(t) = \widehat{\boldsymbol{\theta}}_{m|k}(t+i|t), \tag{2.14}$$

where

The transitions follow from the well-known property of the Kronecker product $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{A}\mathbf{C} \otimes \mathbf{B}\mathbf{D})$ which holds provided that dimensions of corresponding matrices match. According to (2.14), as long as the functions span the same subspace, the result of estimation remains unchanged. Since w_k-orthonormalization is a linear transformation, orthonormal and not orthonormal functions spanning the same subspace produce exactly the same estimation results.

2.3 Statistical properties of the BF method

2.3.1 Introduction

Before we derive the statistical properties of the BF estimator, some results regarding the convergence of the generalized regression matrix to its expected value will be needed. All derivations require the assumption that the input signal $\{u(t)\}$ is a zero-mean wide-sense stationary random process with autocorrelation function $r_u(s) = \mathrm{E}[u(t)u^*(t-s)]$, $s \in \mathbb{Z}$.

To prove the convergence, one needs to ensure that the basis functions are well defined for k growing to infinity and one also needs to establish some bounds for them. Thus far we have not been specifying how the sequences $\{f_{l|k}(i), \ i \in I_k, \ l = 1, \ldots, m, \ m, k \in \mathbb{N}_+\}$ can be generated. For the sake of this proof we will assume that they are obtained by w_k-orthonormalization of sequences $\{g_{l|k}(i), \ i \in I_k, \ l = 1, \ldots, m, \ m, k \in \mathbb{N}_+\}$ obtained by sampling some continuous-time integrable functions $g_l : [-1, 1] \to \mathbb{C}$

$$g_{l|k}(i) = g_l\left(\frac{i}{k}\right), \ i \in I_k.$$

We will also assume that functions $g_l(\tau)$, $l = 1, \ldots, m$ are bounded, namely

$$\exists_{M_l > 0}\forall_{\tau \in [-1,1]} \quad |g_l(\tau)| < M_l, \quad l = 1, \ldots, m,$$

and that $w_k(i)$, $i \in I_k$ was obtained by sampling the positive-valued continuous-time function $w : [-1, 1] \to (0, 1]$:

$$w_k(i) = w\left(\frac{i}{k}\right), \quad i \in I_k.$$

Under these assumptions the following lemma can be proved:

Lemma 2.1
$$\exists_{c_1>0}\exists_{k_0>0}\forall_{k>k_0}\forall_{i\in I_k} \quad |f_{l|k}(i)| < \frac{c_1}{\sqrt{k}}, \quad l=1,\ldots,m. \tag{2.15}$$

Proof
The w_k-orthonormalization can be written down as a linear transform

$$\mathbf{f}_{m|k}(i) = \mathbf{G}_{m|k}^{-1}\mathbf{g}_{m|k}(i), \quad i \in I_k, \tag{2.16}$$

where $\mathbf{g}_{m|k}(i) = [g_{1|k}(i),\ldots,g_{m|k}(i)]^{\mathrm{T}}$,

$$\mathbf{Q}_{m|k} = \sum_{i=-k}^{k} w_k(i)\mathbf{g}_{m|k}(i)\mathbf{g}_{m|k}^{\mathrm{H}}(i),$$

and $\mathbf{G}_{m|k}$ is a square, nonsingular matrix, such that $\mathbf{Q}_{m|k} = \mathbf{G}_{m|k}\mathbf{G}_{m|k}^{\mathrm{H}}$. Note that this is not a matrix square root in general [28], however, for a positive semidefinite Hermitian matrix, there exists a unique Hermitian positive semidefinite matrix $\mathbf{G}_{m|k}$ which is also a matrix square root of $\mathbf{Q}_{m|k}$, namely $\mathbf{Q}_{m|k} = \mathbf{G}_{m|k}\mathbf{G}_{m|k}^{\mathrm{H}} = \mathbf{G}_{m|k}\mathbf{G}_{m|k}$. Nonetheless, the decomposition $\mathbf{Q}_{m|k} = \mathbf{G}_{m|k}\mathbf{G}_{m|k}^{\mathrm{H}}$ is also nonunique, since, having $\mathbf{G}_{m|k}$, a new matrix $\widetilde{\mathbf{G}}_{m|k}$ which also obeys the decomposition $\mathbf{Q}_{m|k} = \widetilde{\mathbf{G}}_{m|k}\widetilde{\mathbf{G}}_{m|k}^{\mathrm{H}}$ can be easily found as $\widetilde{\mathbf{G}}_{m|k} = \mathbf{G}_{m|k}\mathbf{U}$, where $\mathbf{U}$ is an $m \times m$ unitary matrix.

Note that

$$\lim_{k\to\infty}\frac{1}{k}\mathbf{Q}_{m|k} = \lim_{k\to\infty}\frac{1}{k}\sum_{i=-k}^{k} w_k(i)\mathbf{g}_{m|k}(i)\mathbf{g}_{m|k}^{\mathrm{H}}(i) = \lim_{k\to\infty}\frac{1}{k}\sum_{i=-k}^{k} w\left(\frac{i}{k}\right)\mathbf{g}\left(\frac{i}{k}\right)\mathbf{g}^{\mathrm{H}}\left(\frac{i}{k}\right)$$
$$= \int_{-1}^{1} w(\tau)\mathbf{g}(\tau)\mathbf{g}^{\mathrm{H}}(\tau)d\tau = \mathbf{Q} > 0, \tag{2.17}$$

where $\mathbf{g}(\tau) = [g_1(\tau),\ldots,g_m(\tau)]^{\mathrm{T}}$. This means that any element $q_{m|k}^{ij} = [\mathbf{Q}_{m|k}]_{(ij)}$, $i,j = 1,\ldots,m$ of the matrix $\mathbf{Q}_{m|k}$ obeys

$$\lim_{k\to\infty}\frac{1}{k}q_{m|k}^{ij} = q^{ij}, \quad i,j = 1,\ldots,m, \tag{2.18}$$

where $q^{ij} = \mathbf{Q}_{(ij)}$ is an element of the matrix $\mathbf{Q}$. Consequently, any element of a matrix $\mathbf{G}_{m|k}^{-1}$: $\widetilde{g}_{m|k}^{ij} = [\mathbf{G}_{m|k}^{-1}]_{(ij)}$ obeys

$$\widetilde{g}_{m|k}^{ij} = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right). \tag{2.19}$$

Combining this result with (2.16), one obtains that

$$\exists_{c_1>0}\exists_{k_0>0}\forall_{k>k_0}\forall_{i\in I_k} \quad |f_{l|k}(i)| < \frac{c_1}{\sqrt{k}}, \quad l=1,\ldots,m. \quad \blacksquare \tag{2.20}$$

Having this result, one can start proving that

$$\lim_{k\to\infty}\mathbf{P}_{m|k}(t) = \mathrm{E}[\mathbf{P}_{m|k}(t)] = \mathbf{P}_0 = \mathbf{\Phi} \otimes \mathbf{I}_m, \tag{2.21}$$

in a mean squared sense, where

$$\mathbf{\Phi} = \mathrm{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)] = \begin{bmatrix} r_u(0) & \ldots & r_u(n-1) \\ \vdots & \ddots & \vdots \\ r_u^*(n-1) & \ldots & r_u(0) \end{bmatrix}. \tag{2.22}$$

Since the matrices $\mathbf{P}_{m|k}(t)$ and $\mathbf{P}_0$ are by definition Hermitian block matrices with Hermitian blocks, one can choose any element from the upper triangle of any block of matrix $\mathbf{P}_{m|k}(t)$ and

prove its convergence to the corresponding element of $\mathbf{P}_0$. Define

$$d_{n_1 n_2 m_1 m_2}(t, k) = \sum_{i=-k}^{k} w_k(i) f_{m_1|k}(i) f^*_{m_2|k}(i) [u(t - n_1 + i) u^*(t - n_2 + i) - r_u(n_2 - n_1)], \tag{2.23}$$

$$n_1, n_2 = 1, \ldots, n, \quad n_1 \le n_2$$
$$m_1, m_2 = 1, \ldots, m, \quad m_1 \le m_2$$

the element from the $[m(n_1 - 1) + m_1]$-th row and $[m(n_2 - 1) + m_2]$-th column of a Hermitian matrix

$$\mathbf{D}_{m|k}(t) = \mathbf{P}_{m|k}(t) - \mathbf{P}_0. \tag{2.24}$$

The essence of the proof is to show that

$$\lim_{k \to \infty} \mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t, k)|^2] = 0. \tag{2.25}$$

One should start from

$$\begin{aligned}
\mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t, k)|^2] = \mathrm{E}\Bigg\{ & \sum_{i=-k}^{k} \sum_{j=-k}^{k} w_k(i) w_k(j) f_{m_1|k}(i) f^*_{m_1|k}(j) f^*_{m_2|k}(i) f_{m_2|k}(j) \times \\
& \times [u(t - n_1 + i) u^*(t - n_2 + i) - r_u(n_2 - n_1)] \times \\
& \times [u^*(t - n_1 + j) u(t - n_2 + j) - r^*_u(n_2 - n_1)] \Bigg\} \\
= & \sum_{i=-k}^{k} \sum_{j=-k}^{k} \zeta_k(i, j) \mathrm{E}[u(t - n_1 + i) u^*(t - n_2 + i) u^*(t - n_1 + j) u(t - n_2 + j) \\
& - r_u(n_2 - n_1) u^*(t - n_1 + j) u(t - n_2 + j) \\
& - r^*_u(n_2 - n_1) u(t - n_1 + i) u^*(t - n_2 + i) + |r_u(n_2 - n_1)|^2],
\end{aligned} \tag{2.26}$$

where $\zeta_k(i, j) = w_k(i) w_k(j) f_{m_1|k}(i) f^*_{m_1|k}(j) f^*_{m_2|k}(i) f_{m_2|k}(j)$, $i, j \in I_k$. Note that

$$\mathrm{E}[r_u(n_2 - n_1) u^*(t - n_1 + j) u(t - n_2 + j)] = r_u(n_2 - n_1) r^*_u(n_2 - n_1) = |r_u(n_2 - n_1)|^2,$$

and analogically

$$\mathrm{E}[r^*_u(n_2 - n_1) u(t - n_1 + i) u^*(t - n_2 + i)] = r^*_u(n_2 - n_1) r_u(n_2 - n_1) = |r_u(n_2 - n_1)|^2.$$

Hence

$$\begin{aligned}
\mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t, k)|^2] = \sum_{i=-k}^{k} \sum_{j=-k}^{k} \zeta_k(i, j) \Bigg\{ & \mathrm{E}[u(t - n_1 + i) u^*(t - n_2 + i) \times \\
& \times u^*(t - n_1 + j) u(t - n_2 + j)] - |r_u(n_2 - n_1)|^2 \Bigg\}.
\end{aligned} \tag{2.27}$$

Because $\mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2]$ is real and nonnegative, one may write down the following expression

$$
\mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2] = |\mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2]| = \left| \sum_{i=-k}^{k} \sum_{j=-k}^{k} \zeta_k(i,j) \Big\{ \mathrm{E}[u(t-n_1+i)u^*(t-n_2+i) \times \right.
$$

$$
\left. \times\, u^*(t-n_1+j)u(t-n_2+j)] - |r_u(n_2-n_1)|^2 \Big\} \right|
$$

$$
\leq \sum_{i=-k}^{k} \sum_{j=-k}^{k} \left| \zeta_k(i,j) \Big\{ \mathrm{E}[u(t-n_1+i)u^*(t-n_2+i) \times \right.
$$

$$
\left. \times\, u^*(t-n_1+j)u(t-n_2+j)] - |r_u(n_2-n_1)|^2 \Big\} \right|
$$

$$
= \sum_{i=-k}^{k} \sum_{j=-k}^{k} |\zeta_k(i,j)| \left| \Big\{ \mathrm{E}[u(t-n_1+i)u^*(t-n_2+i) \times \right.
$$

$$
\left. \times\, u^*(t-n_1+j)u(t-n_2+j)] - |r_u(n_2-n_1)|^2 \Big\} \right|.
$$

(2.28)

The first element of the double sum can be upper-bounded using lemma 2.1. and the fact that $\forall_{i \in I_k}\, w_k(i) \leq 1$

$$
\exists_{c_1 > 0} \exists_{k_0 > 0} \forall_{k > k_0} \forall_{i,j \in I_k} \quad |\zeta_k(i,j)| \leq \frac{c_1^4}{k^2}.
$$

(2.29)

To proceed with the proof, one needs to restrict the second modulus inside the double sum. This is why, one needs an additional assumption about the mixing (asymptotic independence) conditions for $\{u(t)\}$. We will start with the one that is usually met in telecommunication applications.

> (A2.3) The input signal $\{u(t)\}$ is a zero-mean N-dependent sequence of circular complex random variables with finite statistical moments up to the fourth order moment $\mathrm{E}[|u(t)^4|] = c_u^4 < \infty$.

Formally, a circularity of a complex random variable x means that for any real a, x and $xe^{a\mathrm{i}}$ have the same distribution [21]. For the purpose of this proof only the following properties of a complex circular random variable x are important $\mathrm{E}(x) = 0$, $\mathrm{E}[\mathrm{Re}(x)^2] = \mathrm{E}[\mathrm{Im}(x)^2] = \frac{1}{2}\mathrm{var}(x)$ and $\mathrm{E}[\mathrm{Re}(x)\mathrm{Im}(x)] = 0$. Complex random variables obeying the conditions above are sometimes called "proper" [1], [61].

Under this assumption, for some combination of indexes i and j in a double sum, the expectation of a product of random variables can be replaced by the product of expectation. More precisely, when $|i-j| > N$, $|i-n_1+j-n_2| > N$ and $|i-n_2+j-n_1| > N$, each one of the random variables $u(t-n_1+i)$ and $u(t-n_2+i)$ is independent from each one of the pair $u(t-n_1+j)$ and $u(t-n_2+j)$. In such a case

$$
\mathrm{E}[u(t-n_1+i)u^*(t-n_2+i)u^*(t-n_1+j)u(t-n_2+j)] =
$$
$$
\mathrm{E}[u(t-n_1+i)u^*(t-n_2+i)]\mathrm{E}[u^*(t-n_1+j)u(t-n_2+j)] = r_u(n_2-n_1)r_u^*(n_2-n_1) = \quad (2.30)
$$
$$
|r_u(n_2-n_1)|^2,
$$

and as a consequence nonzero elements will occur in the second modulus of (2.28) only when the aforementioned conditions for i and j are not met. It is worth noting, that these three conditions can be replaced by one, which is more restrictive: $|i-j| > N + |n_2-n_1|$, which can be further majorized by $|i-j| > N + n$.

For $i = j$ there are exactly K nonzero elements, for $i = j \pm 1$, one has $K-1$ nonzero quantities, and so on. Summarizing, the double sum contains less than $K + 2[K-1 + K-2 + \ldots + K - (N + |n_2-n_1|)]$ nonzero elements, and this is smaller than $(2N + 2n + 1)(2k + 1)$. Regarding the

nonzero quantities, they can be upper-bounded

$$
\begin{aligned}
\mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2] &\leq \sum_{i=-k}^{k} \sum_{\substack{j=-k \\ |i-j| \leq N+n}}^{k} |\zeta_k(i,j)| \left| \left\{ \mathrm{E}[u(t-n_1+i)u^*(t-n_2+i) \times \right. \right. \\
&\quad \times u^*(t-n_1+j)u(t-n_2+j)] - |r_u(n_2-n_1)|^2 \Big\} \Big| \\
&\leq \sum_{i=-k}^{k} \sum_{\substack{j=-k \\ |i-j| \leq N+n}}^{k} |\zeta_k(i,j)| \Big\{ \big| \mathrm{E}[u(t-n_1+i)u^*(t-n_2+i) \times \\
&\quad \times u^*(t-n_1+j)u(t-n_2+j)]\big| + |r_u(n_2-n_1)|^2 \Big\}.
\end{aligned}
\tag{2.31}
$$

Using the version of Cauchy-Schwarz inequality

$$
|\mathrm{E}(xy)|^2 \leq \mathrm{E}(|x|^2)\mathrm{E}(|y|^2),
$$

one obtains

$$
\begin{aligned}
|\mathrm{E}[u(t-&n_1+i)u^*(t-n_2+i)u^*(t-n_1+j)u(t-n_2+j)]| \\
&\leq \sqrt{|\mathrm{E}[|u(t-n_1+i)|^2|u(t-n_2+i)|^2]\mathrm{E}[|u(t-n_1+j)|^2|u(t-n_2+j)|^2]|} \\
&\leq \sqrt[4]{E[|u(t-n_1+i)|^4]\mathrm{E}[|u(t-n_2+i)|^4]\mathrm{E}[|u(t-n_1+j)|^4]\mathrm{E}[|u(t-n_2+j)|^4]} = c_u^4.
\end{aligned}
\tag{2.32}
$$

Note also that $|r_u(n_2-n_1)|^2 \leq \sigma_u^4$, where σ_u^2 is a variance of the input signal.

Combining (2.29) with (2.31) and (2.32), one finally obtains

$$
\exists_{c_1>0} \exists_{k_0>0} \forall_{k>k_0} \forall_{i \in I_k} \quad \mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2] \leq \frac{(2N+2n+1)(2k+1)c_1^4}{k^2}[c_u^4+\sigma_u^4] = \mathcal{O}\left(\frac{1}{k}\right),
\tag{2.33}
$$

which results in

$$
\lim_{k \to \infty} \mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2] = 0, \quad \text{q.e.d.}
\tag{2.34}
$$

If needed, the mixing condition from (A2.3) can be relaxed if one assumes that $\{u(t)\}$ is a sequence of Gaussian random variables.

> (A2.4) $\{u(t)\}$ is a zero-mean wide-sense stationary sequence of complex Gaussian random variables with exponentially decaying autocorrelation function $r_u(s) = \mathrm{E}[u(t)u^*(t-s)]$, $s \in \mathbb{Z}$:
> $$
> \exists_{c_2>0,\,\beta \in (0,1)} \forall_{s \in \mathbb{Z}} \quad |r_u(s)| \leq c_2 \beta^{|s|}.
> $$

Under (A2.4) and using the well-known extension of the Isserlis' theorem for the zero-mean complex Gaussian processes [88], which takes the form

$$
\mathrm{E}[x_1 x_2 x_3^* x_4^*] = \mathrm{E}[x_1 x_3^*]\mathrm{E}[x_2 x_4^*] + \mathrm{E}[x_1 x_4^*]\mathrm{E}[x_2 x_3^*],
\tag{2.35}
$$

one obtains

$$
\mathrm{E}[u(t-n_1+i)u^*(t-n_2+i)u^*(t-n_1+j)u(t-n_2+j)] = |r_u(j-i)|^2 + |r_u(n_2-n_1)|^2.
\tag{2.36}
$$

Combining this with (2.28), one arrives at

$$
\mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2] \leq \sum_{i=-k}^{k} \sum_{j=-k}^{k} |\zeta_k(i,j)| |[|r_u(j-i)|^2 + |r_u(n_2-n_1)|^2 - |r_u(n_2-n_1)|^2]|
\tag{2.37}
$$

$$
= \sum_{i=-k}^{k} \sum_{j=-k}^{k} |\zeta_k(i,j)| |r_u(j-i)|^2.
\tag{2.38}
$$

Hence, for sufficiently large k, this expression can be bounded by

$$\mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2] \leq \frac{c_1^4}{k^2} \sum_{i=-k}^{k} \sum_{j=-k}^{k} |r_u(j-i)|^2 \leq \frac{c_1^4 c_2^2}{k^2} \sum_{i=-k}^{k} \sum_{j=-k}^{k} \beta^{2|j-i|}. \tag{2.39}$$

The closed-form formula for $\sum_{i=-k}^{k} \sum_{j=-k}^{k} \beta^{2|j-i|}$ can be easily found, however, for this proof the following upper bound is sufficient

$$\forall_{k>0} \forall_{\beta \in (0,1)} \quad \sum_{i=-k}^{k} \sum_{j=-k}^{k} \beta^{2|j-i|} \leq \frac{4k+1}{1-\beta^2}, \tag{2.40}$$

which means that

$$\exists_{c_3 > 0} \exists_{k_0 > 0} \forall_{k > k_0} \quad \mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2] \leq \frac{c_3}{k}, \tag{2.41}$$

and consequently

$$\lim_{k \to \infty} \mathrm{E}[|d_{n_1 n_2 m_1 m_2}(t,k)|^2] = 0. \tag{2.42}$$

The rate of convergence is of order $\mathcal{O}\left(\frac{1}{k}\right)$.

Remark

Even though the assumptions (A2.3) and (A2.4) may look restrictive, numerical evidence suggests that convergence takes place under much weaker mixing conditions. For example, figure 2.2 shows the behavior (averaged by 10 independent realizations of the input signal) of the largest (the upper plot) and smallest (the lower plot) eigenvalue of a matrix $\mathbf{P}_{m|k}(t)$ for $m = 5$ and $n = 6$ in a situation when $u(t) = 0.8u(t-1) + e(t)$ and $\mathrm{Re}[e(t)] \sim \mathcal{U}(-0.5, 0.5)$, $\mathrm{Im}[e(t)] \sim \mathcal{U}(-0.5, 0.5)$. Note that this situation does not fit either assumption (A2.3) or (A2.4) because the input signal is neither N-dependent nor Gaussian.

2.3.2   Bias

For the derivation of the covariance matrix and expected trajectory of BF estimates, we will need the following assumption

> (A2.5) $\{e(t)\}$, independent of $\{\boldsymbol{\varphi}(t)\}$, is a sequence of zero-mean, i.i.d. complex circular random variables with finite variance σ_e^2.

The result from the previous section justifies the following approximation valid for sufficiently large values of k

$$\mathbf{P}_{m|k}^{-1}(t) \cong \mathbf{P}_o^{-1}, \tag{2.43}$$

which leads to the simplified version of the BF estimator

$$\widehat{\boldsymbol{\alpha}}_{m|k}(t) \cong \widetilde{\boldsymbol{\alpha}}_{m|k}(t) = \mathbf{P}_o^{-1} \mathbf{p}_{m|k}(t), \tag{2.44}$$

and

$$\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t) \cong \widetilde{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t) = \mathbf{F}_{m|k}(i) \widetilde{\boldsymbol{\alpha}}_{m|k}(t), \quad s \in I_k \tag{2.45}$$

Hence

$$\overline{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t) = \mathrm{E}\left[\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t)\right] \cong \mathrm{E}\left[\widetilde{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t)\right] = \mathrm{E}\left[\mathbf{F}_{m|k}(s) \mathbf{P}_0^{-1} \sum_{i=-k}^{k} w_k(i) \boldsymbol{\psi}_{m|k}(t,i) y^*(t+i)\right]$$

$$= \mathbf{F}_{m|k}(s) \mathbf{P}_0^{-1} \mathrm{E}\left\{ \sum_{i=-k}^{k} w_k(i)[\boldsymbol{\varphi}(t+i) \otimes \mathbf{f}_{m|k}(i)][\boldsymbol{\varphi}^{\mathrm{H}}(t+i)\boldsymbol{\theta}(t+i) + e^*(t+i)]\right\}, \tag{2.46}$$

and since

$$\mathrm{E}\left\{[\boldsymbol{\varphi}(t+i) \otimes \mathbf{f}_{m|k}(i)]\boldsymbol{\varphi}^{\mathrm{H}}(t+i)\right\} = \boldsymbol{\Phi} \otimes \mathbf{f}_{m|k}(i), \tag{2.47}$$

the final result is

$$\overline{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t) \cong \mathbf{F}_{m|k}(s) \sum_{i=-k}^{k} w_k(i) \mathbf{P}_0^{-1}[\boldsymbol{\Phi} \otimes \mathbf{f}_{m|k}(i)]\boldsymbol{\theta}(t+i) = \mathbf{f}_{m|k}^{\mathrm{H}}(s) \sum_{i=-k}^{k} w_k(i) \mathbf{f}_{m|k}(i)\boldsymbol{\theta}(t+i). \tag{2.48}$$
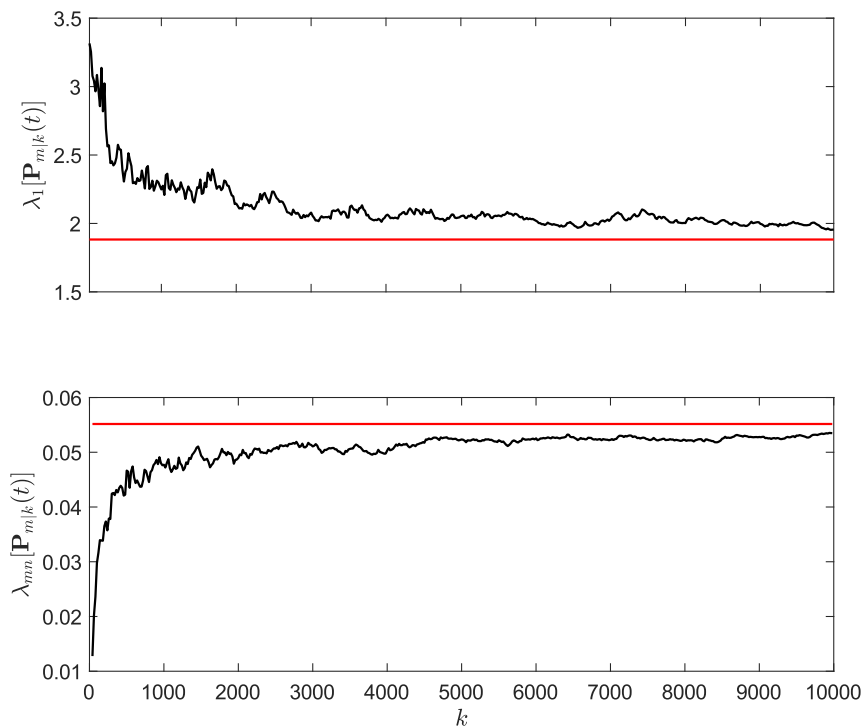
Figure 2.2: The largest (the upper plot) and smallest (the lower plot) eigenvalue of matrix $\mathbf{P}_{m|k}(t)$ for $m = 5$ and $n = 6$ in a situation when $u(t) = 0.8u(t-1) + e(t)$ and $\text{Re}[e(t)] \sim \mathcal{U}(-0.5, 0.5)$, $\text{Im}[e(t)] \sim \mathcal{U}(-0.5, 0.5)$, depicted using black lines. The red lines show the true eigenvalue of a matrix $\mathbf{P}_0$. Results were avereged over 10 independent realizations of the input signal.

Remark 1 - projection

When the flat (unity) weighting is applied, i.e., $w_h(i) \equiv 1$, $i \in I_k$, then (2.48) can be interpreted as the orthogonal projection of true parameter trajectories onto the subspace spanned by the basis functions, which confirms that the choice of the coordinates cannot influence the estimation result, only the change of the basis can. For uniform weighting $w_k(i) \equiv 1$, $i \in I_k$ (2.48) is an orthogonal projection onto the subspace spanned by the basis functions.

Remark 2 - condition for approximate unbiasedness

When the design assumption (A2.1) is fulfilled, which means that parameter trajectories can be described exactly as a linear combination of chosen basis functions, then (2.48) becomes

$$\overline{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t) \cong \mathbf{f}_{m|k}^{\mathrm{H}}(s) \sum_{i=-k}^{k} w_k(i) \mathbf{f}_{m|k}(i) [\mathbf{I}_n \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(i)] \boldsymbol{\alpha}_{m|k}(t) = \mathbf{F}_{m|k}(s) \boldsymbol{\alpha}_{m|k}(t) = \boldsymbol{\theta}(t+s), \quad s \in I_k,$$

(2.49)

and the BF estimator is approximately unbiased which stays in agreement with the remark above.

Remark 3 - filtering

Note that in the context of (2.48), the trajectory of expected estimates can be interpreted as the effect of filtering the true parameter trajectory with a noncausal, time-varying filter associated with the basis function method

$$\overline{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t) \cong \sum_{i=-k}^{k} h_{m|k}^{\mathrm{BF}}(s,i) \boldsymbol{\theta}(t+i),$$

(2.50)

where

$$h_{m|k}^{\mathrm{BF}}(s,i) = w_k(i) \mathbf{f}_{m|k}^{\mathrm{H}}(s) \mathbf{f}_{m|k}(i), \quad s, i \in I_k$$

is an impulse response associated with the BF method.

2.3.3 Covariance of the estimation errors

In order to compute the covariance matrix, the same approximation as in the previous subsection can be used. Suppose that assumption (A2.1) holds true. Then

$$\mathrm{cov}\left[\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t)\right] \cong \mathrm{cov}\left[\widetilde{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t)\right] = \mathbf{F}_{m|k}(s) \mathrm{cov}\left[\widetilde{\boldsymbol{\alpha}}_{m|k}(t)\right] \mathbf{F}_{m|k}^{\mathrm{H}}(s).$$

(2.51)

Note also, that by combining (2.44) with (1.2), one arrives at

$$\widetilde{\boldsymbol{\alpha}}_{m|k}(t) - \boldsymbol{\alpha}_{m|k}(t) = \left[\mathbf{P}_0^{-1} \sum_{i=-k}^{k} w_k(i) \boldsymbol{\psi}_{m|k}(t,i) \boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,i) - \mathbf{I}_{mn}\right] \boldsymbol{\alpha}_{m|k}(t)$$

$$+ \mathbf{P}_0^{-1} \sum_{i=-k}^{k} w_k(i) \boldsymbol{\psi}_{m|k}(t,i) e^*(t+i)$$

$$= \mathbf{v}_1(t) + \mathbf{v}_2(t).$$

(2.52)

Consequently

$$\mathrm{cov}[\widetilde{\boldsymbol{\alpha}}_{m|k}(t)] = \mathrm{E}\{[\mathbf{v}_1(t) + \mathbf{v}_2(t)][\mathbf{v}_1(t) + \mathbf{v}_2(t)]^{\mathrm{H}}\} = \mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_1^{\mathrm{H}}(t)] + \mathrm{E}[\mathbf{v}_2(t)\mathbf{v}_2^{\mathrm{H}}(t)],$$

(2.53)

because under (A2.4), and (A2.5): $\mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_2^{\mathrm{H}}(t)] = \mathrm{E}[\mathbf{v}_2(t)\mathbf{v}_1^{\mathrm{H}}(t)] = 0$. It can be also shown [64] that under (A2.5) $\mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_1^{\mathrm{H}}(t)] = O\left(\frac{1}{K}\right)$, hence,

$$\lim_{k \to \infty} \mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_1^{\mathrm{H}}(t)] = 0,$$

in a mean squared sense (hence also in probability). Therefore, for sufficiently long analysis windows

$$
\begin{aligned}
\operatorname{cov}[\widetilde{\boldsymbol{\alpha}}_{m|k}(t)] &\cong \mathrm{E}[\mathbf{v}_2(t)\mathbf{v}_2^{\mathrm{H}}(t)] \\
&= \mathbf{P}_0^{-1}\mathrm{E}\left[\sum_{i=-k}^{k}\sum_{j=-k}^{k}w_k(i)w_k(j)\boldsymbol{\psi}_{m|k}(t,i)\boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,j)e^*(t+i)e(t+j)\right]\mathbf{P}_0^{-1}.
\end{aligned}
\tag{2.54}
$$

One can use the well-known extension of the Isserlis' theorem [88] to show that

$$
\mathrm{E}[\mathbf{x}\mathbf{x}^{\mathrm{H}}yy^*] = \mathrm{E}[\mathbf{x}\mathbf{x}^{\mathrm{H}}]\mathrm{E}[yy^*] + \mathrm{E}[\mathbf{x}y^*]\mathrm{E}[y\mathbf{x}^{\mathrm{H}}].
\tag{2.55}
$$

Under the assumption (A2.5), and assuming that the noise is complex Gaussian, one can use the above result to obtain

$$
\operatorname{cov}[\widetilde{\boldsymbol{\alpha}}_{m|k}(t)] \cong \sigma_e^2(\boldsymbol{\Phi}^{-1}\otimes\mathbf{H}_{m|k}),
\tag{2.56}
$$

where

$$
\mathbf{H}_{m|k} = \sum_{i=-k}^{k}w_k^2(i)\mathbf{f}_{m|k}(i)\mathbf{f}_{m|k}^{\mathrm{H}}(i).
$$

Hence

$$
\begin{aligned}
\operatorname{cov}\left[\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t)\right] &\cong \sigma_e^2[\mathbf{I}_n\otimes\mathbf{f}_{m|k}^{\mathrm{H}}(s)][\boldsymbol{\Phi}^{-1}\otimes\mathbf{H}_{m|k}][\mathbf{I}_n\otimes\mathbf{f}_{m|k}(s)] \\
&= \sigma_e^2[\boldsymbol{\Phi}^{-1}\otimes\mathbf{f}_{m|k}^{\mathrm{H}}(s)\mathbf{H}_{m|k}\mathbf{f}_{m|k}(s)] \\
&= \sigma_e^2[\boldsymbol{\Phi}^{-1}\otimes\sum_{i=-k}^{k}w_k(i)\mathbf{f}_{m|k}^{\mathrm{H}}(s)\mathbf{f}_{m|k}(i)w_k(i)\mathbf{f}_{m|k}^{\mathrm{H}}(i)\mathbf{f}_{m|k}(s)] \\
&= \sigma_e^2\boldsymbol{\Phi}^{-1}\sum_{i=-k}^{k}\{h_{m|k}^{\mathrm{BF}}(s,i)[h_{m|k}^{\mathrm{BF}}(s,i)]^*\} = \sigma_e^2\boldsymbol{\Phi}^{-1}\sum_{i=-k}^{k}|h_{m|k}^{\mathrm{BF}}(s,i)|^2 = \frac{\sigma_e^2\boldsymbol{\Phi}^{-1}}{l_{m|k}^{\mathrm{BF}}(s)},
\end{aligned}
\tag{2.57}
$$

where

$$
l_{m|k}^{\mathrm{BF}}(s) = \left\{\sum_{i=-k}^{k}|h_{m|k}^{\mathrm{BF}}(s,i)|^2\right\}^{-1},\ s\in I_k
\tag{2.58}
$$

is the local equivalent number of observations. The name of this quantity originates from the fact that for constant parameters, using the BF method is locally equivalent to using the least squares technique with analysis window of length equal to $l_{m|k}^{\mathrm{BF}}(s)$. The local equivalent number of observations is typically not constant inside the analysis window because $h_{m|k}^{\mathrm{BF}}(s,i)$ is made of inner products between vectors of basis functions at different times. This means that the accuracy of estimates at different points inside the analysis window will differ. An example of $l_{m|k}^{\mathrm{BF}}(s)$ for $m=5$ Legendre polynomials and rectangular weighting seqience is depicted in figure 2.3.3 When the estimates of basis function coefficients are utilized to obtain parameter estimates in the entire analysis window, as in [66], the notions of an average value of the error covariance matrix and the average equivalent memory of the BF estimator can be very useful. In this work, the main focus is on local methods, hence these two quantities are not presented. However, their definitions can be found in [66].

2.4 Frequency characteristics of the BF method

In this section derivation of frequency characteristics of the BF method is provided. It is worth noting that in this subchapter all expectations are over different trajectories of parameter trajectories. Thus far, parameter trajectories were treated as deterministic, but now the following assumption will be adopted

(A2.6) $\{\boldsymbol{\theta}(t)\}$, independent of $\{u(t)\}$ and $\{e(t)\}$, is a zero-mean, wide-sense stationary process with a spectral density matrix $\mathbf{S}_{\theta}(\omega)$.
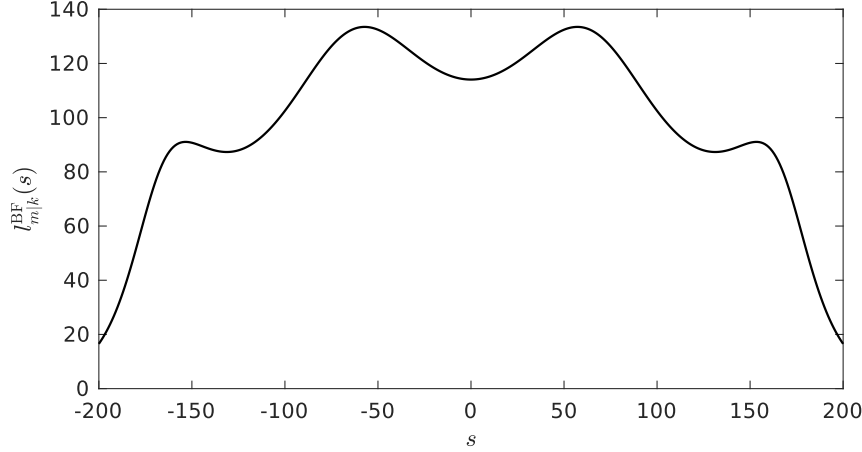
Figure 2.3: Shape of the local equivalent number of observations for $k = 200$ and $m = 3$ Legendre polynomials, and rectangular weighting sequence.

Let us start by defining the cross-covariance matrices of system parameters as

$$\mathbf{Q}(i) = \mathrm{E}[\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\mathrm{H}}(t+i)], \quad i \in \mathbb{Z}, \tag{2.59}$$

then the spectral density matrix is defined as follows

$$\mathbf{S}_\theta(\omega) = \sum_{s=-\infty}^{\infty} \mathrm{E}[\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\mathrm{H}}(t+s)]e^{-\mathrm{i}\omega s}. \tag{2.60}$$

Using the relationship (2.48) derived in the previous section, one can easily find the instantenous frequency response associated with the BF method by computing the discrete Fourier transform (DFT) of the impulse response of filter associated with the BF estimator

$$H_{m|k}^{\mathrm{BF}}(s,\omega) = \sum_{i=-k}^{k} h_{m|k}^{\mathrm{BF}}(t,i)e^{-\mathrm{i}\omega i} = A_{m|k}^{\mathrm{BF}}(s,\omega)e^{\mathrm{i}\phi_{m|k}^{\mathrm{BF}}(s,\omega)}, s \in I_k, \tag{2.61}$$

where $A_{m|k}^{\mathrm{BF}}(s,\omega) = |H_{m|k}^{\mathrm{BF}}(s,\omega)|$ and $\phi_{m|k}^{\mathrm{BF}}(s,\omega) = \arg[H_{m|k}(s,\omega)]$ are the amplitude and phase responses respectively.

In this subsection, we are interested in examining the frequency distribution of the bias component of the mean squared estimation error

$$\overline{B}_{m|k}^{\mathrm{BF}}(s) = \mathrm{E}\left[\left|\left|\boldsymbol{\theta}(t+s) - \overline{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t)\right|\right|^2\right]. \tag{2.62}$$

First, note that (2.48) can be expressed in the form

$$\overline{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s|t) \cong \boldsymbol{\Theta}(t)\mathbf{h}_{m|k}(s), \tag{2.63}$$

where $\boldsymbol{\Theta}(t) = [\boldsymbol{\theta}(t-k),\dots,\boldsymbol{\theta}(t+k)]$ is an $n \times K$ matrix and $\mathbf{h}_{m|k}(s) = [h_{m|k}^{\mathrm{BF}}(s,-k),\dots,h_{m|k}^{\mathrm{BF}}(s,k)]^{\mathrm{T}}$, and similarly

$$\boldsymbol{\theta}(t+s) = \boldsymbol{\Theta}(t)\mathbf{1}_{K,s+k+1}, \tag{2.64}$$

where $\mathbf{1}_{K,s+k+1}$ is an $K \times 1$ indicator vector, namely a column vector which contains one on the position $s+k+1$ and zeros elsewhere. Therefore

$$\boldsymbol{\theta}(t+s) - \overline{\boldsymbol{\theta}}_{m|k}^{\mathrm{BF}}(t+s) = [\mathbf{1}_{K,s+k+1} - \mathbf{h}_{m|k}(s)]\boldsymbol{\Theta}(t), \tag{2.65}$$

and consequently

$$\overline{B}_{m|k}^{\mathrm{BF}}(s) \cong [\mathbf{1}_{K,s+k+1} - \mathbf{h}_{m|k}(s)]^{\mathrm{H}}\mathrm{E}[\boldsymbol{\Theta}^{\mathrm{H}}(t)\boldsymbol{\Theta}(t)][\mathbf{1}_{K,s+k+1} - \mathbf{h}_{m|k}(s)]. \tag{2.66}$$

Since elements of the resulting matrix $\mathrm{E}[\mathbf{\Theta}^{\mathrm{H}}(t)\mathbf{\Theta}(t)]$ are related to cross-covariance matrices of system parameters, the element from the i-th row and j-th column of $\mathrm{E}[\mathbf{\Theta}^{\mathrm{H}}(t)\mathbf{\Theta}(t)]$ is given by

$$\mathrm{E}[\mathbf{\Theta}^{\mathrm{H}}(t)\mathbf{\Theta}(t)]_{(ij)} = \mathrm{Tr}[\mathbf{Q}(i-j)] = \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{Tr}[\mathbf{S}_{\theta}(\omega)]e^{\mathrm{i}\omega(i-j)}d\omega. \tag{2.67}$$

Therefore, one finally obtains

$$\overline{B}_{m|k}^{\mathrm{BF}}(s) \cong \frac{1}{\pi}\int_{0}^{\pi} E_{m|k}^{\mathrm{BF}}(s,\omega)\mathrm{Tr}[\mathbf{S}_{\theta}(\omega)]d\omega, \tag{2.68}$$

where

$$E_{m|k}^{\mathrm{BF}}(s,\omega) = |1 - H_{m|k}^{\mathrm{BF}}(s,\omega)|^2 \tag{2.69}$$

is the instantaneous parameter matching characteristic of a BF estimator. This result is by no means surprising and fully coincides with the previous remarks about the projection and approximate unbiasedness of the BF method. It also supports the claim, that when the spectral density function of each parameter is known, basis functions can be chosen so as to capture all of the dominant dynamical patterns of parameter changes. This also gives one an insight into the way the basis function method will affect (on average) certain frequency components present in parameter trajectories.

Remark

One can easily conduct a similar analysis for each parameter separately, obtaining

$$\overline{B}_{j,m|k}^{\mathrm{BF}}(s) = \mathrm{E}\left[\left|\theta_j(t+s) - \overline{\theta}_{j,m|k}^{\mathrm{BF}}(t+s|t)\right|^2\right] \cong \frac{1}{\pi}\int_{0}^{\pi} E_{m|k}^{\mathrm{BF}}(s,\omega)S_{\theta_j}(\omega)d\omega, \tag{2.70}$$

where $S_{\theta_j}(\omega)$ denotes the spectral density function of a j-th parameter.

2.5 Selection of regressors

The models of time-varying systems are often sparse, which means that one needs to decide which of the estimated parameters are of the smallest importance and discard them. Keeping spurious parameters in the model would result in overfitting, thus decreasing the performance of the model. The same problem applies to the basis function expansion of each parameter. The speed and type of variations might be different for each parameter trajectory. As a consequence, one might need different basis functions to successfully model changes of different parameters. Therefore, one needs to be able to adaptively select the parameters and basis functions for each one of them. This problem can be solved by adaptive selection of regressors - basis functions from the expansion of each parameter. If the parameter is not critical for the model, all basis function coefficients will be set to zero in its expansion resulting in discarding the parameter from the final model. The adaptive choice of the most important regressors is most often implemented in one of two ways: ascending or descending. In the first case, one starts from the first-order model and sequentially adds new terms to the model until the reduction of the variance of residual errors is insignificant. The descending approach works in the opposite way. One starts from the full-order model and then drops the least significant regressors in consecutive iterations. The ideas described in this section are tightly connected to the techniques used in the identification of nonlinear systems, where overparametrization is a serious problem. In fact, some of the methods mentioned in the thesis originate from that field.

2.5.1 The ascending approach

Let us introduce the family of first order models $\{\mathcal{M}_1^1(t), \dots, \mathcal{M}_1^{mn}(t)\}$ defined as follows

$$\mathcal{M}_1^p(t) : y(t+i) = [\alpha_{k,j}^l(t)]^* f_{l|k}(i)u(t+i-j+1) + e_1^p(t+i),$$
$$p = (j-1)m + l, \ i \in I_k, \ l = 1, \dots, m, \ j = 1, \dots, n. \tag{2.71}$$

Estimates, obtained for the p-th first order model are given by

$$\widehat{\alpha}_k^{(1,p)}(t) = \arg\min_{\alpha} \sum_{i=-k}^{k} w_k(i)|y(t+i) - \alpha^* f_{k,l}(i)u(t+i-j+1)|^2, \quad p = 1,\ldots,nm. \tag{2.72}$$

To each of the models, the local sum of squared residuals can be assigned

$$J_k^{(1,p)}(t) = \sum_{i=-k}^{k} w_k(i)|y(t+i) - [\widehat{\alpha}_k^{(1,p)}(t)]^* f_{k,l}(i)u(t+i-j+1)|^2 \quad p = 1,\ldots,nm. \tag{2.73}$$

As the final first order model $\mathcal{M}_1(t) = \mathcal{M}_1^{p_{\mathrm{opt}}^1}(t)$, we choose the one with the smallest value of cost function (2.73)

$$p_{\mathrm{opt}}^1 = \arg\min_{p=1,\ldots,nm} J_k^{(1,p)}(t). \tag{2.74}$$

Based on the regressor $f_{l_1|k}(i)u(t+i-j_1+1)$ corresponding to the p_{opt}^1 we construct family of second order models $\{\mathcal{M}_2^1(t),\ldots,\mathcal{M}_2^{mn-1}(t)\}$ defined similarly to (2.71)

$$\mathcal{M}_2^p(t) : y(t+i) = [\alpha_{k,j_1}^2 f_{l_1|k}(i)]^* u(t+i-j_1+1) + [\alpha_{k,j}^2(t)]^* f_{l|k}(i)u(t+i-j+1) + e_2^p(t+i),$$
$$p = (j-1)m + l, p \neq p_{\mathrm{opt}} \ i \in I_k, \ l = 1,\ldots,m, \ j = 1,\ldots,n, \tag{2.75}$$

such that $\forall_p \ \mathcal{M}_1(t) \subset \mathcal{M}_2^p(t)$. Next, the estimates corresponding to each model are found and the sum of squared residuals for each model is evaluated analogically to the way described for the first-order models. After we find the second most important regressor $f_{k,l_2}(i)u(t+i-j_2+1)$, the family of third-order models is constructed and so on. The point of this procedure is to build a sparse model, so the stopping criterion should be also proposed. The sum of squared residuals cannot be used for this purpose because as the model order grows, the sum of squared residuals uniformly decreases. Simultaneously, after passing a certain size, modelling capability decreases and the model starts to fit the noise. This is why the criterion should include a penalty term on a model order, preventing the model from overparametrization. The second problem is how to perform the aforementioned procedure in a computationally efficient manner. These two issues were discussed in [14]. In this article, the orthogonalization of regressors was proposed to solve the least squares problem using QR decomposition. Thanks to that, estimates from the first iteration can be used in the second iteration and so on. For the stopping criterion, the minimization of the Akaike Information Criterion (AIC) [2] was suggested as one of the possible options. The methods discussed in [14], called orthogonal least squares (OLS), were derived for time-invariant nonlinear systems. However, it was shown in [47] and [107] that they can be easily adapted for the identification of nonstationary processes with the basis function method.

2.5.2   The descending approach

The descending approach reverses the procedure from the ascending approach: we start from the full-order model and then we iteratively discard some regressors until a stopping criterion is met. An interesting approach to this problem was described in [11]. The authors propose to build a basis of candidate functions to describe the nonlinear system dynamics and then to use available data for sparse regression. As noted by the authors, LASSO [100] is a well-established method for deducing sparse models from data. However, this method in the general case is computationally very demanding. This is why the authors propose sequential thresholded least squares (STLS). The method starts by performing ordinary least squares. In the second step, all regressors corresponding to estimated coefficients smaller than some threshold are discarded from the model. The next step computes the least squares estimates for the truncated model and the procedure is repeated. First of all, it is worth noting that this method can be easily adapted to the basis function method, just like in the case of the OLS technique. Secondly, the authors of [11] propose to set the threshold based on the cross-validation. For time-varying systems, traditional cross-validation is often impossible since typically, one only has one realization of the input/output data. Therefore, one could modify this approach and use the AIC criterion to choose how many regressors should be discarded at each step, as well as to determine how many steps should be performed. Similar approach, utilizing wavelets was presented in [104], where authors used AIC and P test to determine the final model order.

2.6 Local Basis Function method

As mentioned before, the LBF method introduced and thoroughly described in [69] is a variant of the BF estimator. Unlike the clasical BF method, the LBF estimator yields the estimates only for the current time instant, not for the entire analysis interval. Consequently, the analysis in the LBF method is carried out in a sliding-window mode, resulting in a series of point estimates obtained for consecutive positions of the window. In a mathematical description of the LBF estimator the relationships (2.4) - (2.6) remain unchanged, and differences start from (2.8), which now has the form

$$\widehat{\boldsymbol{\theta}}_{m|k}^{\text{LBF}}(t) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{m|k}(t). \tag{2.76}$$

The properties of the LBF method can be easily derived, knowing the general properties of BF estimators. The algebraic properties are exactly the same, since the estimation formula is the same, and is based on the least squares method. The statistical properties of the LBF method can be obtained by substituting into formulas derived in the previous section $s = 0$. The following subsections compare the properties of LBF and BF methods.

2.6.1 Bias

The formula for the expected path of LBF estimates can be expressed as

$$\overline{\boldsymbol{\theta}}_{m|k}^{\text{LBF}}(t) = \overline{\boldsymbol{\theta}}_{m|k}^{\text{BF}}(t + s|t)|_{s=0} \cong \sum_{i=-k}^{k} h_{m|k}^{\text{BF}}(0, i)\boldsymbol{\theta}(t + i) = \sum_{i=-k}^{k} h_{m|k}^{\text{LBF}}(i)\boldsymbol{\theta}(t + i), \tag{2.77}$$

where

$$h_{m|k}^{\text{LBF}}(i) = h_{m|k}^{\text{BF}}(0, i) = w_k(i)\mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(i), \quad i \in I_k, \tag{2.78}$$

denotes an impulse response of a time-invariant linear filter associated with the LBF method. Thus, all remarks about projection and unbiasedness hold true also for the LBF estimator. Additionally, when the weighting function has the property $w_k(0) = 1$, then the LBF estimator projects parmaeter trajectories onto the subspace spanned by the functions $\widetilde{\mathbf{f}}_{m|k}(i) = w_k(i)\mathbf{f}_{m|k}(i)$, $i \in I_k$. Note that to obtain final estimates at time instant t, the same amount of information is extracted from the past data and from the "future" (with respect to the time instant t) data, which guarantees more accurate estimates of parameter trajectories. The exemplary impulse responses generated using Legendre polynomials are presented in figure 2.4.

2.6.2 Covariance matrix of the estimation error

The covariance matrix associated with this method can be obtained in the same way

$$\text{cov}\left[\widehat{\boldsymbol{\theta}}_{m|k}^{\text{LBF}}(t)\right] = \text{cov}\left[\widehat{\boldsymbol{\theta}}_{m|k}^{\text{BF}}(t + s|t)\right]\bigg|_{s=0} \cong \frac{\sigma_e^2\boldsymbol{\Phi}^{-1}}{l_{m|k}^{\text{BF}}(0)} = \frac{\sigma_e^2\boldsymbol{\Phi}^{-1}}{l_{m|k}^{\text{LBF}}}. \tag{2.79}$$

The equivalent number of observations of the local basis function estimator can be expressed as

$$l_{m|k}^{\text{LBF}} = \left\{ \sum_{i=-k}^{k} |h_{m|k}^{\text{LBF}}(i)|^2 \right\}^{-1}. \tag{2.80}$$

To get insights into the relationship between covariance matrices of BF and LBF estimators within the range of the analysis window, one can compare the equivalent numbers of observations. Again, the key observation is that for each time instant, LBF uses the same equivalent number of observations, in contrary to the BF estimator. Since typically $l_{m|k}^{\text{BF}}(0)$ is the local maximum of $\{l_{m|k}^{\text{BF}}(i), i \in I_k\}$, one can expect the covariance matrix of BF estimates to be greater (in a matrix sense) than the covariance matrix of the LBF estimates for some time instants within the range of the analysis window.
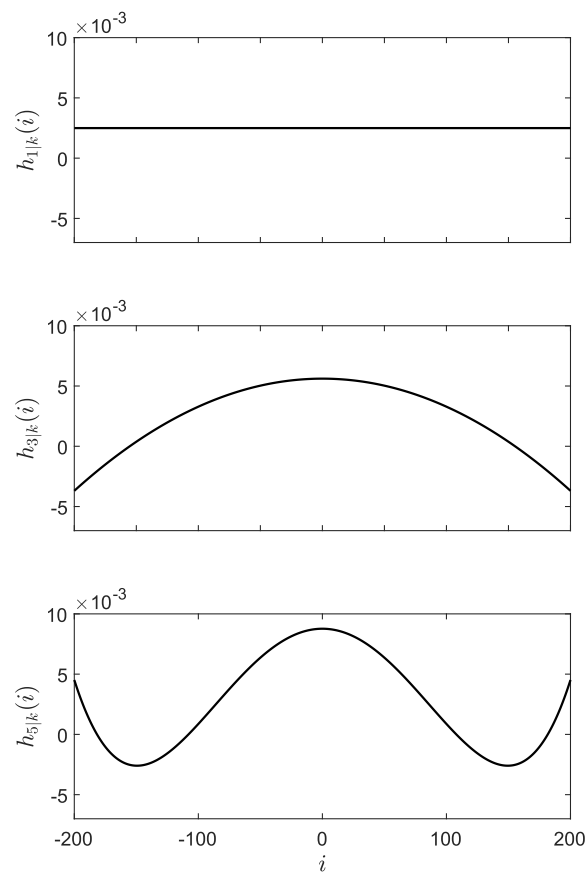
Figure 2.4: Impulse responses associated with the LBF method for 1 (the uppmost plot), 3 (plot in the middle) and 5 Legendre polynomials (the lowest plot).

2.6.3 Frequency characteristics

Using the same technique as above, one can derive frequency characteristics of the LBF method from corresponding characteristics of a BF estimator. Consequently, we have the discrete Fourier transform of an impulse response associated with the LBF method

$$H_{m|k}^{\text{LBF}}(\omega) = \sum_{i=-k}^{k} h_{m|k}^{\text{LBF}}(i)e^{-\mathrm{i}\omega i} = A_{m|k}^{\text{LBF}}(\omega)e^{\mathrm{i}\phi_{m|k}^{\text{LBF}}(\omega)}, \tag{2.81}$$

where $A_{m|k}^{\text{LBF}}(\omega) = |H_{m|k}^{\text{LBF}}(\omega)|$ and $\phi_{m|k}^{\text{LBF}}(\omega) = \arg[H_{m|k}^{\text{LBF}}(\omega)]$ are the amplitude and phase responses respectively.

The MSE of LBF estimator, for any particular time instant

$$\overline{B}_{m|k}^{\text{LBF}} \cong \frac{1}{\pi} \int_0^\pi E_{m|k}^{\text{LBF}}(\omega)\text{Tr}[\mathbf{S}_\theta(\omega)]d\omega, \tag{2.82}$$

where

$$E_{m|k}^{\text{LBF}}(\omega) = |1 - H_{m|k}^{\text{LBF}}(\omega)|^2, \tag{2.83}$$

is a parameter matching characteristic of the LBF method.

2.6.4 MSE analysis

In this section, we give the expression for the MSE of LBF estimates and show the connection between the chosen basis functions and the MSE score. To do so, we need to assume some statistical knowledge about the type of parameter changes.

> (A2.7) $\{\theta_j(t)\}$, $j = 1, \dots, n$ are sequences of mutually uncorrelated complex random variables with the same, up to the scaling factors $\zeta_j > 0$ $j = 1, \dots, n$, autocorrelation function $\mathrm{E}[\theta_j(t)\theta_j^*(t - s)] = \zeta_j r_\theta(s)$, $s \geq 0$

The simplest, yet useful example of an autocorrelation function would be the autocorrelation function of a bandlimited signal with uniform (nonpreferential) power spectral density $S(\omega)$ defined as

$$S(\omega) = \begin{cases} S_0 & \text{for } |\omega| \leq \omega_0 \\ 0 & \text{for } |\omega| > \omega_0 \end{cases}, \tag{2.84}$$

where S_0 is a positive constant and $\omega_0 \in [-\pi, \pi]$ is a normalized cut-off frequency. For such a signal, the autocorrelation function is given by $r(s) = \sigma^2 \text{sinc}(\omega_0 s)$. Another important example is a function, that is frequently used when working with terrestrial wireless communication systems under the assumption of wide-sense stationary uncorrelated scattering [48]. In such a case, one often adopts the Jakes Doppler power profile, resulting in the following autocorrelation function $r(s) = \sigma^2 J_0(\omega_d s)$, where $J_0(\cdot)$ is the zero-order Bessel function of the first kind and ω_d denotes the normalized maximum Doppler frequency, which is equal to the speed of the mobile device and scatterers, divided by the carrier wavelength [56]. Comparison of these two autocorrelation functions is shown in figure 2.5 for $\sigma^2 = 1$ and $\omega_0 = \omega_d = \pi/20$. The upper plot shows the first of the above-mentioned functions and the lower plot shows the function stemming from the Jakes' model. As mentioned in the previous chapter, the MSE can be written down as a sum of bias and variance components (1.6). Combining (1.6) with (2.77), the bias component can be expressed as

$$B_{m|k}^{\text{LBF}} \cong ||\boldsymbol{\theta}(t) - \boldsymbol{\Theta}(t)\mathbf{h}_{m|k}||^2, \tag{2.85}$$

where $\boldsymbol{\Theta}(t) = [\boldsymbol{\theta}(t-k), \dots, \boldsymbol{\theta}(t+k)]$, $\mathbf{h}_{m|k} = [h_{m|k}^{\text{LBF}}(-k), \dots, h_{m|k}^{\text{LBF}}(k)]^{\text{T}}$. Under the assumption (A2.7), the expected value of the bias component (in this subsection, the expectation is carried over different realizations of parameter trajectories) can be expressed as

$$\overline{B}_{m|k}^{\text{LBF}} = \mathrm{E}[B_{m|k}^{\text{LBF}}] \cong \eta[r_\theta(0) - \mathbf{r}_\theta^{\text{H}}\mathbf{h}_{m|k}^* - \mathbf{r}_\theta^{\text{T}}\mathbf{h}_{m|k} + \mathbf{h}_{m|k}^{\text{H}}\mathbf{R}_\theta\mathbf{h}_{m|k}], \tag{2.86}$$

where $\eta = \sum_{j=1}^{n} \zeta_j$, $\mathbf{r}_\theta = [r_\theta(-k), \dots, r_\theta(k)]^{\text{T}}$, and

$$\mathbf{R}_\theta = \begin{bmatrix} r_\theta(0) & \dots & r_\theta(K-1) \\ \vdots & \ddots & \vdots \\ r_\theta^*(K-1) & \dots & r_\theta(0) \end{bmatrix} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\text{H}},$$
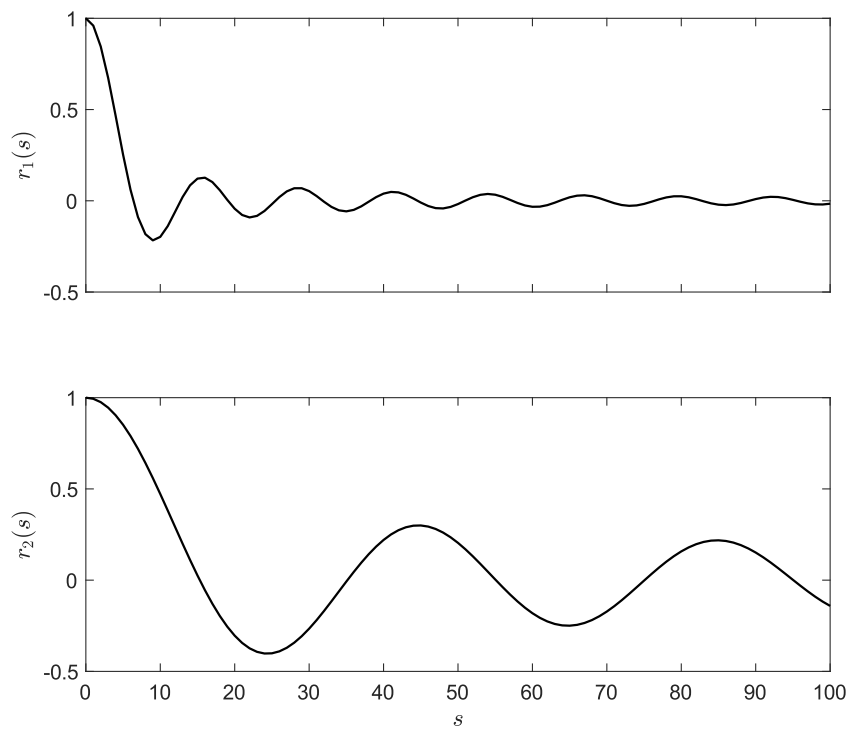
Figure 2.5: Comparison of autocorrelation function of a bandlimited signal (the upper plot) and the autocorrelation fucntion stemming from the Jakes' model (the lower plot), for $\sigma^2 = 1$ and $\omega_0 = \omega_d = \pi/6$.

$\mathbf{\Lambda} = \text{diag}[\lambda_1, \dots, \lambda_K]$, is a matrix made up of the eigenvalues organized in non-ascending order $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K \geq 0$ and $\mathbf{Q} = [\mathbf{q}_1 | \dots | \mathbf{q}_K]$ is a matrix of corresponding orthonormal eigenvectors $\mathbf{q}_l$ of $\mathbf{R}_\theta$.

When parameter trajectories belong to the subspace spanned by the basis functions, the variance component from (1.6) can be expressed as

$$V_{m|k}^{\text{LBF}} = \text{Tr}\left\{ \text{cov}\left[\widehat{\boldsymbol{\theta}}_{m|k}^{\text{LBF}}(t) \right] \right\} \cong \kappa \sigma_e^2 \mathbf{h}_{m|k}^{\text{H}} \mathbf{h}_{m|k}, \tag{2.87}$$

where $\kappa > 0$ is a sum of eigenvalues of $\mathbf{\Phi}^{-1}$. Even if parameter trajectories do not belong to the subspace spanned by the basis functions, increasing m increases the accuracy of this approximation.

Consequently, the expected mean squared estimation error can be obtained in the form

$$M_{m|k}^{\text{LBF}} = \text{E}[B_{m|k}^{\text{LBF}}] + \text{E}[V_{m|k}^{\text{LBF}}] \cong \eta \left[r_\theta(0) - \mathbf{r}_\theta^{\text{H}} \mathbf{h}_{m|k}^* - \mathbf{r}_\theta^{\text{T}} \mathbf{h}_{m|k} + \mathbf{h}_{m|k}^{\text{H}} \left(\mathbf{R}_\theta + \frac{\kappa \sigma_e^2}{\eta} \mathbf{I}_K \right) \mathbf{h}_{m|k} \right]. \tag{2.88}$$

The expression (2.88) shows indirectly how the chosen basis functions affect the MSE of the LBF estimator. Unfortunately, it ties the impulse response associated with a chosen basis with the MSE score, which says little about basis functions themselves. One can minimize (2.88) and find the functions providing the optimal impulse response, but it would require using K basis functions, which is impossible in practice (unless $n = 1$) because of the problems with inverting generalized regularization matrix. To avoid such troubles, the length of the analysis interval should be suitably large, namely $K \gg mn$.

However, one can use functions, resulting from the Karhunen-Loéve theorem [39] (KL functions) and it turns out that for such a choice of the basis, the optimal number of functions can be easily found.

KL basis functions

The original Karhunen-Loéve theorem states that any second-order random process can be expressed using the infinite expansion of orthogonal functions [39], similar in some sense to the Fourier expansion for deterministic signals. This result was developed independently by several researchers [38], [40], [42], [54], and has since became useful in many fields of engineering [39]. From the identification perspective, the most important consequence of the Karhunen-Loéve theorem is that the second-order discrete-time random sequence of length K can be exactly described as a linear combination of K orthogonal sequences, that we call here KL functions.

They are defined as follows

$$[f_{l|k}(-k), \dots f_{l|k}(k)]^{\text{T}} = \mathbf{q}_l, \; l = 1, \dots, K. \tag{2.89}$$

Note that because $\mathbf{R}_\theta$ is Toeplitz, these functions are even. Namely, let $\mathbf{J}_K$ be an antidiagonal matrix of size $K \times K$, with antidiagonal entries equal to one and all other elemnts equal to zero. Left-side multiplication by such a matrix flips the matrix "upside down" and right-side multiplication flips the matrix left to right. Then, for each eigenvector we have that

$$\mathbf{R}_\theta \mathbf{q}_j = \lambda_j \mathbf{q}_j \iff \mathbf{J}_K \mathbf{R}_\theta \mathbf{q}_j = \lambda_j \mathbf{J}_K \mathbf{q}_j \iff \mathbf{J}_K \mathbf{R}_\theta^* \mathbf{q}_j^* = \lambda_j \mathbf{J}_K \mathbf{q}_j^* \iff$$
$$\mathbf{J}_K \mathbf{R}_\theta^* \mathbf{J}_K \mathbf{J}_K \mathbf{q}_j^* = \lambda_j \mathbf{J}_K \mathbf{q}_j^* \iff \mathbf{R}_\theta \mathbf{q}_j^* = \lambda_j \mathbf{q}_j^*, \; j = 1, \dots, K.$$

The last transitions follow from the facts that $\mathbf{J}_K \mathbf{R}_\theta^* \mathbf{J}_K = \mathbf{R}_\theta$ and $\mathbf{J}_K \mathbf{J}_K = \mathbf{I}_K$. This means that the i-th element of $\mathbf{q}_j$, $j = 1, \dots, K$ obeys $q_j(i) = q_j^*(K - i + 1)$, $i = 1, \dots, K$, and $q_j(k + 1) \in \mathbb{R}$.

An example of KL functions for bandlimited signal is shown in figure 2.6.

In practice, one cannot use K basis functions (unless $n = 1$). However, using (2.88), one can find the optimal number of KL basis functions. Assume that instead of using all K eigenvectors, one uses only the first m of them. In such a case, the impulse response is given by

$$\mathbf{h}_{m|k}^{\text{KL}} = \mathbf{Q}_m \mathbf{Q}_m^{\text{H}} \mathbf{1}_{K,k+1} = \mathbf{Q}_m \mathbf{f}_{m|k}^{\text{KL}}(0), \tag{2.90}$$

where $\mathbf{Q}_m$ is a $K \times m$ matrix containing first m columns of $\mathbf{Q}$. Because

$$\mathbf{Q}^{\text{H}} \mathbf{Q}_m = \left[\frac{\mathbf{I}_m}{\mathbf{0}} \right]_{K \times m},$$
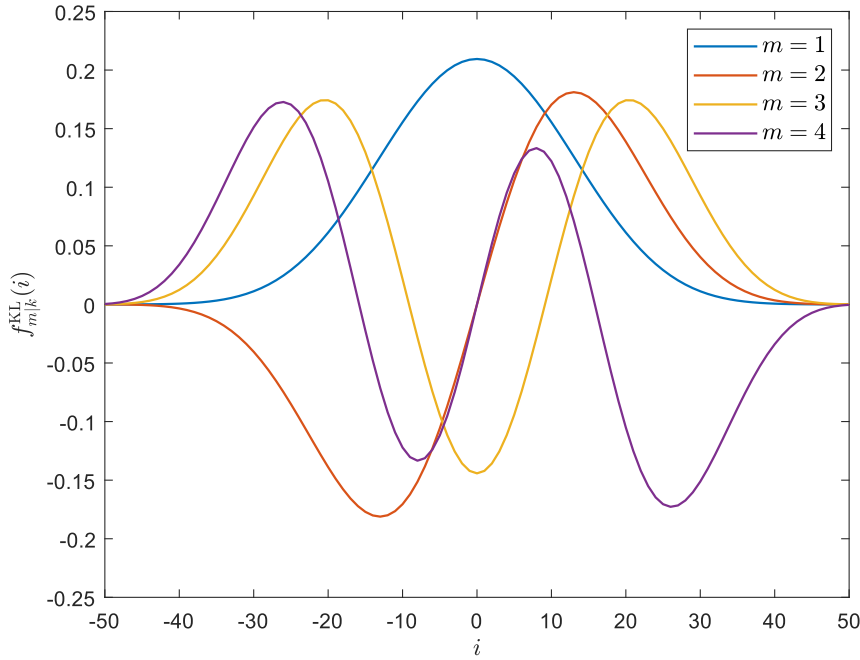
Figure 2.6: Four first KL functions for a bandlimited parameter trajectory with $\omega_0 = 0.1$ and $\sigma_\theta^2 = 1$.

it is straightforward to show that

$$\mathbf{r}_\theta^{\mathrm{H}}(\mathbf{h}_{m|k}^{\mathrm{KL}})^* = \mathbf{r}_\theta^{\mathrm{T}}\mathbf{h}_{m|k}^{\mathrm{KL}} = (\mathbf{h}_{m|k}^{\mathrm{KL}})^{\mathrm{H}}\mathbf{R}_\theta\mathbf{h}_{m|k}^{\mathrm{KL}} = \sum_{l=1}^{m}\lambda_l[f_l^{\mathrm{KL}}(0)]^2. \tag{2.91}$$

Therefore, the bias can be expressed as

$$\mathrm{E}[B_{m|k}^{\mathrm{KL}}] = \eta\left\{r_\theta(0) - \sum_{l=1}^{m}\lambda_l[f_l^{\mathrm{KL}}(0)]^2\right\}, \tag{2.92}$$

and the variance is equal

$$\mathrm{E}[V_{m|k}^{\mathrm{KL}}] \cong \kappa\sigma_e^2\sum_{l=1}^{m}[f_l^{\mathrm{KL}}(0)]^2. \tag{2.93}$$

Consequently, the MSE for the KL basis has the form

$$M_{m|k}^{\mathrm{KL}} \cong \eta r_\theta(0) + \sum_{l=1}^{m}\left\{[\kappa\sigma_e^2 - \eta\lambda_l][f_l^{\mathrm{KL}}(0)]^2\right\}. \tag{2.94}$$

By looking for m minimizing (2.94), one can find the optimal number of KL basis functions. However, as emphasized several times before, the number of basis functions cannot be arbitrarily large due to possible problems with the invertibility of the generalized regression matrix. To guarantee invertibility of the generalized regression matrix $\mathbf{P}_{m|k}(t)$ the number of basis functions m must obey the condition $mn \leq K$. This requirement guarantees that the amount of information is greater than the number of estimated coefficients, which in practice secures well-conditioning of the generalized regression matrix. To avoid numerical problems we will search m in the interval $[1, m_0]$, where $m_0 = \lfloor (K - 10)/n \rfloor$:

$$m_{\mathrm{opt}} = \underset{1 \leq m \leq m_0}{\arg\min} M_{m|k}^{\mathrm{KL}} = \underset{1 \leq m \leq m_0}{\arg\max}\left\{m : \lambda_m > \frac{\kappa\sigma_e^2}{\eta}\right\}. \tag{2.95}$$

Numerical experiments suggest that this safeguard is sufficient to avoid numerical problems.

Remark - DPSS

For bandlimited discrete signals, functions stemming from the Karhunen-Loéve theorem are equivalent to discrete prolate spheroidal sequences (DPSS) [96]. It is also worth noting that in such a case, the autocorrelation matrix might be poorly conditioned for $K \geq 32$, causing numerical problems during computation of its eigenvectors [105]. However, in the article [30] the fast algorithm for computing DPSS were provided. It is based on the eigendecomposition of a suitably chosen tridiagonal symmetric matrix $\mathbf{D}_\theta$ that commutes with $\mathbf{R}_\theta$ [96]

$$[\mathbf{D}_\theta]_{(ij)} = \begin{cases} \frac{1}{2}i(K-i) & \text{if } j = i+1 \\ \left(\frac{K-1}{2} - i + 1\right)^2 \cos\omega_0 & \text{if } j = i \\ \frac{1}{2}(i-1)(K-i+1) & \text{if } j = i-1 \\ 0 & \text{if } |j-i| > 1 \end{cases} \quad i,j = 1,\ldots,K. \tag{2.96}$$

After computing the eigenvector $\mathbf{q}_l$, $l = 1,\ldots,K$ of matrix $\mathbf{D}_\theta$, the corresponding eigenvalue of $\mathbf{R}_\theta$ can be found using

$$\lambda_l = ||\mathbf{R}_\theta \mathbf{q}_l||, \quad l = 1,\ldots,K. \tag{2.97}$$

Remark - bias-variance trade-off

Expressions (2.92) and (2.93) well illustrate the need for the bias-variance trade-off. When the basis functions from defined by the expression (2.89) are adopted, $r_\theta(0) = \sum_{l=1}^{K} \lambda_l [f_l^{\mathrm{KL}}(0)]^2$, so the bias component decreases when we increase the number of basis functions. It is interesting to note, that if K KL basis functions could be used, the LBF estimator would be unbiased. The opposite effect takes place for variance (2.93). Because $\mathbf{Q}$ is a unitary matrix, the variance rises up to $\kappa\sigma_e^2$ with the increase of m.

2.6.5 Computational aspects

The LBF method, when implemented directly, can be computationally very expensive. The cost of computing generalized regression matrix $\mathbf{P}_{m|k}(t)$ and vector $\mathbf{p}_{m|k}(t)$ from (2.7) is equal to $\mathcal{O}(Km^2n^2)$ and $\mathcal{O}(Kmn)$ MACs (multiply and accumulate operations), respectively. To solve (2.6), additional $\mathcal{O}(m^3n^3)$ MACs are needed. The first part of the computational burden, associated with the update of $\mathbf{P}_{m|k}(t)$ and $\mathbf{p}_{m|k}(t)$ can be decreased to $\mathcal{O}(m^2n^2)$ and $\mathcal{O}(mn)$, respectively, when the basis functions and the weighting sequence are recursively computable [69]. Namely, assume that there exists a matrix $\mathbf{\Gamma}_{m|k}$ such that

$$\mathbf{f}_{m|k}(i-1) = \mathbf{\Gamma}_{m|k}\mathbf{f}_{m|k}(i), \tag{2.98}$$

and a number β_k, such that

$$w_k(i-1) = \beta_k w_k(i). \tag{2.99}$$

An obvious example of a recursively computable weighting sequence is uniform weighting $w_k(i) = 1, \quad i \in I_k$.

Examples of recursively computable functions

- Polynomial functions

  Let

  $$f_l(i) = \left(\frac{i}{k}\right)^{l-1}, \, i \in I_k,$$

  then it is straightforward to check that

  $$\mathbf{\Gamma}_{m|k} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ -\frac{1}{k} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\binom{m-1}{m-1}}{(-k)^{m-1}} & \frac{\binom{m-1}{m-2}}{(-k)^{m-2}} & \ldots & 1 \end{bmatrix}.$$

- Sine-cosine functions

  For

  $$\mathbf{f}_{m|k}(i) = \left[1, \sin\frac{\pi i}{2k}, \cos\frac{\pi i}{2k}, \ldots, \sin\frac{\pi i m_0}{2k}, \cos\frac{\pi i m_0}{2k}\right]^{\mathrm{T}},$$

  $m = 2m_0 + 1$, matrix $\mathbf{\Gamma}_{m|k}$ is defined as follows

  $$\mathbf{\Gamma}_{m|k} = \mathrm{bl\,diag}\{1, \mathbf{G}_{1|k}, \ldots, \mathbf{G}_{m_0|k}\},$$

  where

  $$\mathbf{G}_{l|k} = \begin{bmatrix} \cos\frac{\pi l}{2k} & -\sin\frac{\pi l}{2k} \\ \sin\frac{\pi l}{2k} & \cos\frac{\pi l}{2k} \end{bmatrix}, \ l = 1, \ldots, m_0.$$

- Complex exponential functions

  Define

  $$\mathbf{f}_{m|k}(i) = \frac{1}{\sqrt{K}}\left[1, e^{\frac{-i2\pi i}{K}}, e^{\frac{i2\pi i}{K}}, \ldots, e^{\frac{-i2\pi m_0 i}{K}}, e^{\frac{i2\pi m_0 i}{K}}\right], \ m = 2m_0 + 1.$$

  From the computational point of view, this basis is the most suitable one because in this case $\mathbf{\Gamma}_{m|k}$ is a diagonal matrix

  $$\mathbf{\Gamma}_{m|k} = \mathrm{diag}\left\{1, e^{\frac{2\pi i}{K}}, e^{\frac{-2\pi i}{K}}, \ldots, e^{\frac{2\pi m_0 i}{K}}, e^{\frac{-2\pi m_0 i}{K}}\right\}.$$

  Moreover, this basis is already orthogonal.

Suppose that one starts from one of the function types listed above and then performs a linear transformation of the chosen basis, e.g. the w_k-orthonormalization. It turns out that the new basis is still recursively computable. Denote by $\mathbf{A}$ the matrix transforming the original basis defined by $\mathbf{f}_{m|k}(i)$ to the new basis $\widetilde{\mathbf{f}}_{m|k}(i)$, namely

$$\widetilde{\mathbf{f}}_{m|k}(i) = \mathbf{A}\mathbf{f}_{m|k}(i).$$

Then the new transition matrix is defined as

$$\widetilde{\mathbf{\Gamma}}_{m|k} = \mathbf{A}\mathbf{\Gamma}_{m|k}\mathbf{A}^{-1}. \tag{2.100}$$

Note also, that if the original matrix $\mathbf{\Gamma}_{m|k}$ is diagonalizable, then the linear transformation matrix $\mathbf{A}$ can be easily found that transforms the basis to the basis with lower computational complexity. If $\mathbf{\Gamma}_{m|k}$ is not diagonalizable, there is always a transformation that allows one to bring the transition matrix to Jordan cannonical form using, for example, the Golub-Kahan-Lanczos algorithm [28], which may also reduce the computational complexity.

Examples

For the sine-cosine basis the transition matrix can be diagonalized, since its block components are just the rotation matrices, so it holds that

$$\mathbf{G}_{l|k} = \begin{bmatrix} \cos\frac{\pi l}{k} & -\sin\frac{\pi l}{k} \\ \sin\frac{\pi l}{k} & \cos\frac{\pi l}{k} \end{bmatrix} = \begin{bmatrix} -\frac{\sqrt{2}}{2}i & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2}i \end{bmatrix}\begin{bmatrix} \cos\frac{\pi l}{k} - i\sin\frac{\pi l}{k} & 0 \\ 0 & \cos\frac{\pi l}{k} + i\sin\frac{\pi l}{k} \end{bmatrix}\begin{bmatrix} \frac{\sqrt{2}}{2}i & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2}i & \frac{\sqrt{2}}{2} \end{bmatrix},$$

so matrix A is defined as

$$\mathbf{A} = \mathrm{bl\,diag}\{1, \mathbf{A}_1, \ldots, \mathbf{A}_{m_0}\}$$

$$\mathbf{A}_l = \begin{bmatrix} \frac{\sqrt{2}}{2}i & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2}i & \frac{\sqrt{2}}{2} \end{bmatrix}, \ l = 2, \ldots, m_0.$$

As a consequence the transformed basis becomes very similar to the complex exponential basis, defined above. ∎

It is straightforward to check that the transition matrix associated with the polynomial basis is not diagonalizable since all of its eigenvalues are equal to 1. What is more important, this matrix

always has only one independent eigenvector pointing in a direction $\mathbf{x} = [0, \ldots, 0, 1]^{\mathrm{T}}$. This means that only the Jordan canonical form (JCF) can be reached for this matrix. Since JCF is an upper bidiagonal matrix, it is typically sparser than the initial matrix $\mathbf{\Gamma}_{m|k}$.

For the basis functions and weighting sequence, obeying (2.98) and (2.99), respectively, it holds that

$$
\begin{aligned}
\mathbf{P}_{m|k}(t) = \beta_k [\mathbf{I}_n \otimes \mathbf{\Gamma}_{m|k}] \left[\mathbf{P}_{m|k}(t-1) - w_k(-k)\mathbf{A}(t-k-1) \otimes \mathbf{B}_{m|k}(-k) \right] [\mathbf{I}_n \otimes \mathbf{\Gamma}_{m|k}^{\mathrm{H}}] \\
+ w_k(k)\mathbf{A}(t+k) \otimes \mathbf{B}_{m|k}(k)
\end{aligned}
\tag{2.101}
$$

$$
\begin{aligned}
\mathbf{p}_{m|k}(t) = \beta_k [\mathbf{I}_n \otimes \mathbf{\Gamma}_{m|k}] \left[\mathbf{p}_{m|k}(t-1) - w_k(-k)\mathbf{c}(t-k-1) \otimes \mathbf{f}_{m|k}(-k) \right] [\mathbf{I}_n \otimes \mathbf{\Gamma}_{m|k}^{\mathrm{H}}] \\
+ w_k(k)\mathbf{c}(t+k) \otimes \mathbf{f}_{m|k}(k),
\end{aligned}
\tag{2.102}
$$

where $\mathbf{A}(t) = \boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)$, $\mathbf{B}_{m|k}(i) = \mathbf{f}_{m|k}(i)\mathbf{f}_{m|k}^{\mathrm{H}}(i)$, and $\mathbf{c}(t) = y^*(t)\boldsymbol{\varphi}(t)$. This algorithm excludes the last observation from the analysis window, then "shifts" the basis functions and weighting sequence, and finally attaches the new observation to the analysis window. Note that for the basis functions and weighting sequences obeying (2.98) and (2.99), respectively, the inverse of a matrix $\mathbf{P}_{m|k}(t)$ can be also computed recursively using the Sherman-Morrison formula [98]. However, due to the presence of the matrix $[\mathbf{I}_n \otimes \mathbf{\Gamma}_{m|k}]$ and its Hermitian transpose, the number of real-valued multiply and accumulate operations (later called MACs for short), required to update $\mathbf{P}_{m|k}^{-1}(t)$, is of order $\mathcal{O}(n^2 m^3)$ (because of the sparse nature of this matrix). The number of MACs required for the direct implementation is of order $\mathcal{O}(n^3 m^3)$, hence, there is no real benefit in recursively updating the inverse of $\mathbf{P}_{m|k}(t)$.

Recursively computable complex weighting sequences

Some popular weighting functions, like cosinusoidal or Hann window, are not recursively computable. However, they can be evaluated using the real part of another, complex-valued function $v_k(i)$, $i \in I_k$ which is recursively computable, namely $v_k(i-1) = \beta_k v_k(i)$.

- Cosinusoidal window
  The recursively computable complex function is defined as follows

  $$
  v_k(i) = e^{\mathrm{i}\frac{\pi i}{2k}}, \quad i \in I_k,
  $$

  with $\beta_k = e^{-\mathrm{i}\frac{\pi}{2k}}$. The cosinusoidal window can be obtained as

  $$
  w_k(i) = \cos\frac{\pi i}{2k} = \operatorname{Re}[v_k(i)], \quad i \in I_k.
  $$

- Hann window
  The recursively computable complex function is similar to the function for cosinusoidal function

  $$
  v_k(i) = e^{\mathrm{i}\frac{\pi i}{k}}, \quad i \in I_k,
  $$

  with $\beta_k = e^{-\mathrm{i}\frac{\pi}{k}}$. The Hann window can be computed as follows

  $$
  w_k(i) = 0.5\{1 + \operatorname{Re}[v_k(i)]\}, \quad i \in I_k.
  $$

Let $v_k^r(i) = \operatorname{Re}[v_k(i)]$ and $v_k^i(i) = \operatorname{Im}[v_k(i)]$. Then, to perform recursive computations, one should compute and update two matrices/vectors $\mathbf{P}_{m|k}^r(t|t)/\mathbf{p}_{m|k}^r(t|t)$ and $\mathbf{P}_{m|k}^i(t|t)/\mathbf{p}_{m|k}^i(t|t)$ that appears in the counterpart problem with complex weighting function

$$
\begin{aligned}
\mathbf{R}_{m|k}(t) &= \sum_{i=-k}^{k} v_k^r(i)\boldsymbol{\psi}(t,i)\boldsymbol{\psi}^{\mathrm{H}}(t,i) + \mathrm{i}\sum_{i=-k}^{k} v_k^i(i)\boldsymbol{\psi}(t,i)\boldsymbol{\psi}^{\mathrm{H}}(t,i) = \mathbf{P}_{m|k}^r(t|t) + \mathrm{i}\mathbf{P}_{m|k}^i(t|t) \\
\mathbf{r}_{m|k}(t) &= \sum_{i=-k}^{k} v_k^r(i)\boldsymbol{\psi}(t,i)y^*(t+i) + \mathrm{i}\sum_{i=-k}^{k} v_k^i(i)\boldsymbol{\psi}(t,i)y^*(t+i) = \mathbf{p}_{m|k}^r(t|t) + \mathrm{i}\mathbf{p}_{m|k}^i(t|t).
\end{aligned}
\tag{2.103}
$$

Note that the above formula is not a decomposition into a real and imaginary part. Since the regression vectors and basis functions are complex-valued, both $\mathbf{P}^r_{m|k}(t|t)$ and $\mathbf{P}^i_{m|k}(t|t)$ are complex-valued matrices (the same holds for the corresponding vectors). Only when all of the signals in the system and the basis functions are real-valued, can (2.103) be interpreted as the decomposition into the real and imaginary part of the matrix $\mathbf{R}_{m|k}(t)$ and the vector $\mathbf{r}_{m|k}(t)$.

The first step is to subtract the old observation from these matrices/vectors

$$
\begin{aligned}
\mathbf{P}^r_{m|k}(t|t-1) &= \mathbf{P}^r_{m|k}(t-1|t-1) - v^r_k(-k)\mathbf{A}(t-k-1) \otimes \mathbf{B}_{m|k}(-k) \\
\mathbf{P}^i_{m|k}(t|t-1) &= \mathbf{P}^i_{m|k}(t-1|t-1) - v^i_k(-k)\mathbf{A}(t-k-1) \otimes \mathbf{B}_{m|k}(-k) \\
\mathbf{p}^r_{m|k}(t|t-1) &= \mathbf{p}^r_{m|k}(t-1|t-1) - v^r_k(-k)\mathbf{c}(t-k-1) \otimes \mathbf{f}_{m|k}(-k) \\
\mathbf{p}^i_{m|k}(t|t-1) &= \mathbf{p}^i_{m|k}(t-1|t-1) - v^i_k(-k)\mathbf{c}(t-k-1) \otimes \mathbf{f}_{m|k}(-k)
\end{aligned}
\tag{2.104}
$$

Now assume that there exist $\beta_k \in \mathbb{C}$ such that $v_k(i-1) = \beta_k v_k(i)$, and $\beta_k = a_k + ib_k$. Then

$$
\begin{aligned}
\mathbf{P}^r_{m|k}(t|t) &= [\mathbf{I}_n \otimes \mathbf{\Gamma}_{m|k}][a_k\mathbf{P}^r_{m|k}(t|t-1) - b_k\mathbf{P}^i_{m|k}(t|t-1)][\mathbf{I}_n \otimes \mathbf{\Gamma}^{\mathrm{H}}_{m|k}] + v^r_k(k)\mathbf{A}(t+k) \otimes \mathbf{B}_{m|k}(k) \\
\mathbf{P}^i_{m|k}(t|t) &= [\mathbf{I}_n \otimes \mathbf{\Gamma}_{m|k}][b_k\mathbf{P}^r_{m|k}(t|t-1) + a_k\mathbf{P}^i_{m|k}(t|t-1)][\mathbf{I}_n \otimes \mathbf{\Gamma}^{\mathrm{H}}_{m|k}] + v^i_k(k)\mathbf{A}(t+k) \otimes \mathbf{B}_{m|k}(k) \\
\mathbf{p}^r_{m|k}(t|t) &= [\mathbf{I}_n \otimes \mathbf{\Gamma}_{m|k}][a_k\mathbf{p}^r_{m|k}(t|t-1) - b_k\mathbf{p}^i_{m|k}(t|t-1)][\mathbf{I}_n \otimes \mathbf{\Gamma}^{\mathrm{H}}_{m|k}] + v^r_k(k)\mathbf{c}(t+k) \otimes \mathbf{f}_{m|k}(k) \\
\mathbf{p}^i_{m|k}(t|t) &= [\mathbf{I}_n \otimes \mathbf{\Gamma}_{m|k}][b_k\mathbf{p}^r_{m|k}(t|t-1) + a_k\mathbf{p}^i_{m|k}(t|t-1)][\mathbf{I}_n \otimes \mathbf{\Gamma}^{\mathrm{H}}_{m|k}] + v^i_k(k)\mathbf{c}(t+k) \otimes \mathbf{f}_{m|k}(k).
\end{aligned}
\tag{2.105}
$$

For the cosinusoidal window $\mathbf{P}_{m|k}(t) = \mathbf{P}^r_{m|k}(t|t)$ and $\mathbf{p}_{m|k}(t) = \mathbf{p}^r_{m|k}(t|t)$, while for the Hann window $\mathbf{P}_{m|k}(t) = 0.5[1 + \mathbf{P}^r_{m|k}(t|t)]$ and $\mathbf{p}_{m|k}(t) = 0.5[1 + \mathbf{p}^r_{m|k}(t|t)]$.

It is a well-known fact, that the sliding-window-type algoritmhs are not exponentially, but only mariginally stable, which means that numerical errors will increase at a linear rate. It is caused by the fact that the spectral radius of matrix $\mathbf{\Gamma}_{m|k}$ is equal to one. This means that the largest eigenvalue (and possibly some other eigenvalues as well) lies on the unit circle in a complex plane. Therefore, computation of $\mathbf{P}_{m|k}(t)$ and $\mathbf{p}_{m|k}(t)$ should be resetted regularly using the explicit formulae (2.7). It is easy to check that for basis functions described above all eigenvalues of matrix $\mathbf{\Gamma}_{m|k}$ lie on the unit circle in a complex plane. Numerical experiments have shown that the explicit formulae have to be used every 5 to 20 samples, depending on the value of mn. The higher the value of mn, the more frequent resetting is needed.

The second component of the LBF algorithm is solving of the set of linear equations

$$
\mathbf{P}_{m|k}(t)\widehat{\boldsymbol{\alpha}}_{m|k}(t) = \mathbf{p}_{m|k}(t).
\tag{2.106}
$$

It can be done numerically, reducing the complexity. Recently, the dichotomous coordinate descent (DCD) algorithm was adapted for this task [93], allowing to reduce computaional complexity of this step to as low as $\mathcal{O}(mn)$. The summary of this algortihm with the comparison of exact number of multiply-add operations nedded for different steps is given in the next section.

2.6.6   The dichotomous coordinate descent approach for LBF and computational costs

The dichotomous coordinate descent (DCD) algorithm, described in detail in [111], can be used to recursively solve a system of equations, using just a few iterations. It was used in [93] alongside with the basis functions method. The technique presented in this article was called SRLS-L, but it is worth noting that it is just a different implementation of the LBF method. Note that after adopting assumption (A2.1), one might rearrange (2.4)

$$
\begin{aligned}
y(t+i) &= \boldsymbol{\alpha}^{\mathrm{H}}_{m|k}(t)\boldsymbol{\psi}_{m|k}(t,i) + e(t+i) = \sum_{j=1}^{n}\sum_{l=1}^{m} \alpha^*_{j,l}(t)u(t-j+1)f_{l|k}(i) + e(t+i) \\
&= \sum_{l=1}^{m} f_{l|k}(i)\sum_{j=1}^{n} \alpha^*_{j,l}(t)u(t-j+1) + e(t+i) = \boldsymbol{\beta}^{\mathrm{H}}_{m|k}(t)\boldsymbol{\phi}_{m|k}(t,i) + e(t+i), \quad i \in I_k,
\end{aligned}
\tag{2.107}
$$

where $\boldsymbol{\phi}_{m|k}(t,i) = \mathbf{f}_{m|k}(i) \otimes \boldsymbol{\varphi}(t+i)$ is just a permutated version of $\boldsymbol{\psi}(t,i)$. Since the LBF method is based on solving a system of linear equations, any permutation in data will result in the same permutation of estimated coefficients, and hence the ultimate result will remain the same[1] [66]. Anyway, to apply the DCD algorithm, one needs to write the solution of the system of linear equations in a recursive form, which for LBF is

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{LBF}}(t) = \widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{LBF}}(t-1) + \Delta\boldsymbol{\alpha}_{m|k}^{\mathrm{LBF}}(t), \tag{2.108}$$

where the equation that will be solved instead of

$$\mathbf{P}_{m|k}(t)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{LBF}}(t) = \mathbf{p}_{m|k}(t)$$

takes the form

$$\mathbf{P}_{m|k}(t)\Delta\boldsymbol{\alpha}_{m|k}^{\mathrm{LBF}}(t) = \mathbf{r}_{m|k}(t), \tag{2.109}$$

and $\mathbf{r}_{m|k}(t)$ is a residual vector computed as

$$\mathbf{r}_{m|k}(t) = \mathbf{p}_{m|k}(t) - \mathbf{P}_{m|k}(t)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{LBF}}(t-1). \tag{2.110}$$

This notation allows one to use the DCD algorithm described in [111]. There is one more difference between the description used here and in [93]. Namely, the authors of [93] use additional regularization stabilizing the solution, so instead of linear regression they perform the so-called ridge regression [36], [101].

Computational costs

Here we provide the exact computational costs of all algorithms designed to find the LBF estimates. In all calculations, it was taken into account that single complex-valued multiplication requires 4 MACs.

The direct approach

We will start with the direct implementation of the LBF method. To compute all generalized regression vectors inside the analysis window, one needs $4nmK$ MACs. Computation of the generalized regression matrix takes approximately $2(nm)^2K$ MACs because this matrix is by definition Hermitian. Computation of the vector $\mathbf{p}_{m|k}(t)$ requires $4nmK$ MACs. Finally, solving the system of linear equations directly, i.e. the number of MAC opperations is of order $\mathcal{O}(n^3m^3)$. The total cost of the direct implementation is also of order $\mathcal{O}(n^3m^3)$.[2]

Recursive updates

The costs in this sections will characterize operations stemming from equations (2.104) and (2.105). Computing the first line from (2.104) requires $\mathcal{O}(n^2m^2)$ operations ($\mathbf{B}_{m|k}(-k)$ can be precomputed). For the vector $\mathbf{p}_{m|k}^r(t|t-1)$ the cost is $\mathcal{O}(nm)$ MACs, which results in a number of operations of order $\mathcal{O}(n^2m^2)$.

Computing $[a_k\mathbf{P}_{m|k}^r(t|t-1) - b_k\mathbf{P}_{m|k}^i(t|t-1)]$ requires $\mathcal{O}(n^2m^2)$ MACs, because the matrices are Hermitian. Multiplications $[\mathbf{I}_n \otimes \boldsymbol{\Gamma}_{m|k}][a_k\mathbf{P}_{m|k}^r(t|t-1) - b_k\mathbf{P}_{m|k}^i(t|t-1)][\mathbf{I}_n \otimes \boldsymbol{\Gamma}_{m|k}^{\mathrm{H}}]$ require $\mathcal{O}(n^2m^3)$ MACs (because of the block diagonality), because the matrix $\boldsymbol{\Gamma}_{m|k}$ can have a special structure reducing the number of operations. The rest of calculations needed for the update are of order $\mathcal{O}(n^2m^2)$, so the total cost is of order $\mathcal{O}(n^2m^3)$.

The DCD algorithm

The DCD algorithm updates the solution using Euclidean coordinate systems and the step size is chosen using the binary representation of the solution, by proceeding from the most significant bits to the least significant ones. According to [111], in the worst-case scenario, a single pass of the algorithm requires only $4N_inm+N_b$ additions, where N_i denotes the number of algorithm iterations (typically not greater than 16), and N_b is a number of significant bits (typically not greater than 24). Computing the residual vector requires (assuming that $\mathbf{p}_{m|k}(t)$ was already computed) $4(nm)^2$ MACs. Summarizing, the DCD approach requires $4(nm)^2 + 4N_inm + N_b$ MACs. For example, taking $n = 20$, $m = 5$, $N_i = 16$ and $N_b = 24$ results in approximately 1 000 000 MACs for the direct method, and only 46 424 MACs (in the worst case scenario) when using DCD iterations.

---

[1]In fact, it can be proven that any orthogonal transformation of data will result in exactly the same transformation of estimated parameters.

[2]All calculations are for the single position of the analysis window and do not contain constant costs associated with computations that can be done beforehand.

2.6.7 Hyperparameter optimization

The length of analysis window and the number of basis functions are two very important hyperparameters strongly affecting identification results. One method of tuning them is to run several algorithms equipped with different settings and at each time instant t choose the one that provides the best results, evaluated in terms of some local quality measure. Thus far, two local performance measures were proposed: the modified version of the final prediction error (FPE) criterion [68], originally derived by Akaike [4], and the local sum of squared leave-one-out interpolation errors. The first of them is based on the idea of evaluating the estimates on a different data set. Denote by $\Omega'_k(t) = \{\boldsymbol{\varphi}'(t+i), e'(t+i),\ i \in I_k\}$ the realization of input-output data, independent of the realization $\Omega_k(t) = \{\boldsymbol{\varphi}(t+i), e(t+i),\ i \in I_k\}$ used for identification purposes. The FPE measure is defined as

$$\varpi_{m|k}(t) = \mathrm{E}_{\Omega_k(t),\Omega'_k(t)}\left\{ \left| y'(t) - \left[\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{LBF}}(t)\right]^{\mathrm{H}} \boldsymbol{\varphi}'(t) \right|^2 \right\}. \tag{2.111}$$

This can be expressed as

$$\varpi_{m|k}(t) = \mathrm{E}_{\Omega_k(t),\Omega'_k(t)}\left\{ \left| e'(t) - \left[\Delta\boldsymbol{\theta}_{m|k}^{\mathrm{LBF}}(t)\right]^{\mathrm{H}} \boldsymbol{\varphi}'(t) \right|^2 \right\}, \tag{2.112}$$

where $\Delta\boldsymbol{\theta}_{m|k}^{\mathrm{LBF}}(t) = \boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{LBF}}(t)$. Under the assumptions (A2.4), (A2.6) and using (2.57), (2.112) becomes

$$\varpi_{m|k}(t) = \sigma_e^2\left(1 + \frac{n}{l_{m|k}^{\mathrm{LBF}}}\right). \tag{2.113}$$

The final formula can be obtained using

$$\begin{aligned}
\widehat{\sigma}_{m|k}^2(t) &= \frac{1}{L_w}\sum_{i=-k}^{k} w_k(i)|y(t+i) - \widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{H}}(t)\boldsymbol{\psi}_{m|k}(t,i)|^2 \\
&= \frac{1}{L_w}\sum_{i=-k}^{k} w_k(i)|e(t+i) - \Delta\boldsymbol{\alpha}_{m|k}^{\mathrm{H}}(t)\boldsymbol{\psi}_{m|k}(t,i)|^2 \\
&= \frac{1}{L_w}\left[\sum_{i=-k}^{k} w_k(i)|e(t+i)|^2 - \boldsymbol{\zeta}^{\mathrm{H}}(t)\mathbf{P}_{m|k}^{-1}(t)\boldsymbol{\zeta}(t)\right],
\end{aligned} \tag{2.114}$$

where $\boldsymbol{\zeta}(t) = \sum_{i=-k}^{k} w_k(i)\boldsymbol{\psi}(t,i)e^*(t+i)$ and $L_w = \sum_{i=-k}^{k} w_k(i)$ is an effective window width. The last transition follows from the fact that

$$\begin{aligned}
\Delta\boldsymbol{\alpha}_{m|k}(t) = \widehat{\boldsymbol{\alpha}}_{m|k}(t) - \boldsymbol{\alpha}(t) &= \mathbf{P}_{m|k}^{-1}(t)\sum_{i=-k}^{k} w_k(i)\boldsymbol{\psi}(t,i)y^*(t+i) \\
&= \mathbf{P}_{m|k}^{-1}(t)\sum_{i=-k}^{k} w_k(i)\boldsymbol{\psi}(t,i)e^*(t+i) = \mathbf{P}_{m|k}^{-1}(t)\boldsymbol{\zeta}(t),
\end{aligned} \tag{2.115}$$

which follows directly from (2.4) and (2.7).

Under (A2.5), (A2.6), and the assumption that the noise is Gaussian, one obtains

$$\begin{aligned}
\mathrm{E}[\widehat{\sigma}_{m|k}^2(t)] &= \sigma_e^2 - \frac{1}{L_w}\mathrm{E}\left[\sum_{i=-k}^{k}\sum_{j=-k}^{k} w_k(i)w_k(j)\boldsymbol{\psi}^{\mathrm{H}}(t,i)\mathbf{P}_{m|k}^{-1}(t)\boldsymbol{\psi}(t,j)e(t+i)e^*(t+j)\right] \\
&\cong \sigma_e^2 - \frac{1}{L_w}\sum_{i=-k}^{k}\sum_{j=-k}^{k} w_k(i)w_k(j)\mathrm{E}[\boldsymbol{\varphi}^{\mathrm{H}}(t+i)\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t+j)e(t+i)e^*(t+j)]\mathbf{f}_{m|k}^{\mathrm{H}}(i)\mathbf{f}_{m|k}(j) \\
&= \sigma_e^2\left[1 - \frac{n}{L_w}\sum_{i=-k}^{k} w_k^2(i)\mathbf{f}_{m|k}^{\mathrm{H}}(i)\mathbf{f}_{m|k}(i)\right].
\end{aligned} \tag{2.116}$$

Finally, by combining (2.113) and (2.116), one arrives at

$$\text{FPE}(t) \cong \varpi_{m|k}(t) = \frac{1 + \frac{n}{l_{m|k}^{\text{LBF}}}}{1 - \frac{n}{N_{m|k}}} \widehat{\sigma}_{m|k}^2(t), \tag{2.117}$$

where

$$N_{m|k} = \frac{L_w}{\sum_{i=-k}^{k} w_k^2(i) \mathbf{f}_{m|k}^{\text{H}}(i) \mathbf{f}_{m|k}(i)},$$

and

$$\widehat{\sigma}_{m|k}^2(t) = \frac{1}{L_w} \sum_{i=-k}^{k} w_k(i) \left| y(t+i) - \widehat{\boldsymbol{\alpha}}_{m|k}^{\text{H}}(t) \boldsymbol{\psi}(t,i) \right|^2 = \frac{1}{L_w} \left[d_k(t) - \widehat{\boldsymbol{\alpha}}_{m|k}^{\text{H}}(t) \mathbf{p}_{m|k}(t) \right], \tag{2.118}$$

where $d_k(t) = \sum_{i=-k}^{k} w_k(i) |y(t+i)|^2$.

The second approach is based on the idea of leave-one-out cross-validation. The cost function is defined as follows

$$J_{o,m|k}^{\text{LBF}}(t) = \sum_{i=-L}^{L} \left| \varepsilon_{o,m|k}^{\text{LBF}}(t+i) \right|^2, \tag{2.119}$$

where

$$\varepsilon_{o,m|k}^{\text{LBF}}(t) = y(t) - [\widehat{\boldsymbol{\theta}}_{o,m|k}^{\text{LBF}}(t)]^{\text{H}} \boldsymbol{\varphi}(t) = y(t) - \widehat{\boldsymbol{\alpha}}_{o,m|k}^{\text{H}}(t) \boldsymbol{\psi}(t,0) \tag{2.120}$$

denotes leave-one-out interpolation error computed using the holey estimates of system trajectories, i.e. the estimates obtained by excluding the central observation from the analysis window

$$\widehat{\boldsymbol{\theta}}_{o,m|k}^{\text{LBF}}(t) = \mathbf{F}_{m|k}(0) \widehat{\boldsymbol{\alpha}}_{o,m|k}(t) \tag{2.121}$$

and

$$\widehat{\boldsymbol{\alpha}}_{o,m|k}(t) = \arg\min_{\boldsymbol{\alpha}} \sum_{\substack{i=-k \\ i \neq 0}}^{k} w_k(i) |y(t+i) - \boldsymbol{\alpha}^{\text{H}} \boldsymbol{\psi}_{m|k}(t,i)|^2 = \mathbf{P}_{o,m|k}^{-1}(t) \mathbf{p}_{o,m|k}(t), \tag{2.122}$$

$$\mathbf{P}_{o,m|k}(t) = \sum_{\substack{i=-k \\ i \neq 0}}^{k} w_k(i) \boldsymbol{\psi}_{m|k}(t,i) \boldsymbol{\psi}_{m|k}^{\text{H}}(t,i) = \mathbf{P}_{m|k}(t) - w_k(0) \boldsymbol{\psi}(t,0) \boldsymbol{\psi}^{\text{H}}(t,0)$$

$$\tag{2.123}$$

$$\mathbf{p}_{o,m|k}(t) = \sum_{\substack{i=-k \\ i \neq 0}}^{k} w_k(i) y^*(t+i) \boldsymbol{\psi}_{m|k}(t,i) = \mathbf{p}_{m|k}(t) - w_k(0) \boldsymbol{\psi}(t,0) y^*(t).$$

Typically, weighting functions used for localization purposes in LBF method are bell-shaped with $w_k(0) = 1$, which simplifies the derivation.

Using (2.123) and the well-known Sherman-Morrison formula [28], and assuming that $w_k(0) = 1$, one gets

$$\widehat{\boldsymbol{\alpha}}_{o,m|k}(t) = \left[\mathbf{P}_{m|k}^{-1}(t) + \frac{\mathbf{P}_{m|k}^{-1}(t) \boldsymbol{\psi}(t,0) \boldsymbol{\psi}^{\text{H}}(t,0) \mathbf{P}_{m|k}^{-1}(t)}{1 - q_{m|k}(t)} \right] [\mathbf{p}_{m|k}(t) - \boldsymbol{\psi}(t,0) y^*(t)]$$

$$= \left[\mathbf{I}_{mn} + \frac{\mathbf{P}_{m|k}^{-1}(t) \boldsymbol{\psi}(t,0) \boldsymbol{\psi}^{\text{H}}(t,0)}{1 - q_{m|k}(t)} \right] \widehat{\boldsymbol{\alpha}}_{m|k}(t) - \frac{\mathbf{P}_{m|k}^{-1}(t) \boldsymbol{\psi}(t,0) y^*(t)}{1 - q_{m|k}(t)}, \tag{2.124}$$

where

$$q_{m|k}(t) = \boldsymbol{\psi}_{m|k}^{\text{H}}(t,0) \mathbf{P}_{m|k}^{-1}(t) \boldsymbol{\psi}_{m|k}(t,0) \in \mathbb{R}_+, \tag{2.125}$$

is by construction a real, nonnegative number.

Observe that LOOCV errors (2.120) can be now written as

$$\varepsilon_{o,m|k}^{\text{LBF}}(t) = y(t) - \widehat{\boldsymbol{\alpha}}_{m|k}^{\text{H}}(t) \left[\mathbf{I}_{mn} + \frac{\boldsymbol{\psi}(t,0)\boldsymbol{\psi}^{\text{H}}(t,0)\mathbf{P}_{m|k}^{-1}(t)}{1 - q_{m|k}(t)} \right] \boldsymbol{\psi}(t,0) - \frac{\boldsymbol{\psi}^{\text{H}}(t,0)\mathbf{P}_{m|k}^{-1}(t)y(t)}{1 - q_{m|k}(t)}\boldsymbol{\psi}(t,0)$$

$$= \frac{y(t) - \widehat{\boldsymbol{\alpha}}_{m|k}^{\text{H}}(t)\boldsymbol{\psi}(t,0)}{1 - q_{m|k}(t)} = \frac{\varepsilon_{m|k}^{\text{LBF}}(t)}{1 - q_{m|k}(t)},$$

(2.126)

which allows one to compute (2.119) without evaluating the holey estimates of system parameters. This algorithm can still be very time-consuming, unless $\mathbf{P}_{m|k}^{-1}(t)$ is a by-product of computing (2.6). To reduce computational burden, for sufficiently large k, $q_{m|k}(t)$ can be replaced by its expected value

$$\overline{q}_{m|k} = \text{E}[q_{m|k}(t)] \cong \text{E}\Big\{ [\boldsymbol{\varphi}^{\text{H}}(t) \otimes \mathbf{f}_{m|k}^{\text{H}}(0)][\boldsymbol{\Phi}^{-1} \otimes \mathbf{I}_m][\boldsymbol{\varphi}(t) \otimes \mathbf{f}_{m|k}(0)] \Big\} = n\mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0). \quad (2.127)$$

These two criteria can be also applied to choose the set of basis functions used for estimation.

2.6.8 Remark on the real-valued case

Note that most of the formulae and derivations presented in this chapter remain valid when the signals $\{u(t)\}$, $\{y(t)\}$ or/and model parameters $\boldsymbol{\theta}(t)$ are real-valued. The only difference might appear when all the signals are real and the chosen basis functions are complex-valued. Then the final estimates will be the real parts of the estimates derived in this chapter. On the contrary, when the system is complex and basis functions $\{f_l(i), \ i \in I_k, \ l = 1, \ldots, m\}$ are real-valued, all equations remain the same. The only difference is that when all of the listed quantities are real-valued, the recursive expressions for cosinusoidal and Hann window (2.103) - (2.105) will be much simpler.

Chapter 3

Fast local basis function method

The original LBF method, although proven to provide high-quality estimates of nonstationary system parameters, has two drawbacks. Firstly, when implemented directly, it is very computationally expensive. Secondly, all basis function identification schemes developed thus far, attempt to estimates all parameters simultaneously, treating them in the same way, whereas it might be beneficial to apply different estimation settings to different parameters. Both of these issues can be solved using the new, two-step algorithm - fast local basis function (fLBF) method, presented in this chapter. The first step of this new method is called preestimation. It allows one to obtain approximately unbiased estimates of parameter trajectories, regardless of the type or speed of their variation. This property is not typical of identification of nonstationary processes and allows one to obtain unique insight into the dynamics of the system. However, the price for unbiasedness is a strong preestimation noise. This is the reason why preestimates cannot be used as final estimates and need to be filtered first. Hence, the second step is called postfiltering. An important advantage of this approach is that it gives an opportunity to adjust the filtering technique to each parameter separately based either on some *a priori* knowledge or on the insights from the preestimation step.

3.1 Concept presentation

3.1.1 Preestimation

In this section we propose an estimator that is unbiased, regardless of the type and speed of parameter variation, and we describe its statistical properties. Such a method yields estimates that can be characterized by the following equation

$$\widetilde{\boldsymbol{\theta}}(t) = \boldsymbol{\theta}(t) + \mathbf{z}(t), \tag{3.1}$$

where $\mathbf{z}(t)$ is a zero-mean noise. Such an estimator is a "maximum bandwidth" estimator [9], in the sense that it can track (in a mean sense) arbitrarily fast changes of system parameters. Estimators that share this property will be called in this thesis preestimators. The name stems from the fact that the unbiasedness of these estimators comes at a price of high variance of the estimation errors. Because of this, the estimates they yield cannot be used as final estimates and require further processing - posfiltering.

Suppose that the system obeys assumptions (A2.3) and (A2.5), and that the covariance matrix $\boldsymbol{\Phi}$ of the regression vector is known. An estimator that fulfils this condition and takes the form

$$\widetilde{\boldsymbol{\theta}}(t) = \boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t)y^*(t). \tag{3.2}$$

was proposed in [71]. This particular form of preestimator will be further called prototype estimator or protoestimator for short.

Under (A2.5) the expected value of protoestimates is equal to

$$\mathrm{E}[\widetilde{\boldsymbol{\theta}}(t)] = \boldsymbol{\Phi}^{-1}\mathrm{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)]\boldsymbol{\theta}(t) + \boldsymbol{\Phi}^{-1}\mathrm{E}[\boldsymbol{\varphi}(t)e^*(t)] = \boldsymbol{\theta}(t). \tag{3.3}$$

The preestimation errors can be expressed as

$$\Delta\widetilde{\boldsymbol{\theta}}(t) = \widetilde{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t) = [\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t) - \mathbf{I}_n]\boldsymbol{\theta}(t) + \boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t)e^*(t), \tag{3.4}$$

53

and the covariance matrix of preestimation errors is given by

$$
\begin{aligned}
\text{cov}[\widetilde{\boldsymbol{\theta}}(t)] = \text{E}\{\Delta\widetilde{\boldsymbol{\theta}}(t)[\Delta\widetilde{\boldsymbol{\theta}}(t)]^{\text{H}}\} &= \text{E}\{[\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t) - \mathbf{I}_n]\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\text{H}}(t)[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)\boldsymbol{\Phi}^{-1} - \mathbf{I}_n]\} \\
&\quad + \boldsymbol{\Phi}^{-1}\text{E}[e(t)e^*(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)]\boldsymbol{\Phi}^{-1} \\
&= \boldsymbol{\Phi}^{-1}\text{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\text{H}}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)]\boldsymbol{\Phi}^{-1} - \boldsymbol{\theta}(t)\boldsymbol{\theta}^{\text{H}}(t) + \boldsymbol{\Phi}^{-1}\text{E}[e(t)e^*(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)]\boldsymbol{\Phi}^{-1}.
\end{aligned}
$$
$$(3.5)$$

To obtain more insightful analytical results we will additionally assume that the input signal and noise are complex Gaussian. Using the identity that can be derived from the moment theorem for zero-mean complex Gaussian processes (an extension of an Isserlis' theorem) [88]

$$
\text{E}[\mathbf{xy}^{\text{H}}\mathbf{Spq}^{\text{H}}] = \text{E}[\mathbf{xy}^{\text{H}}]\mathbf{S}\text{E}[\mathbf{pq}^{\text{H}}] + \text{E}[\mathbf{y}^{\text{H}}\mathbf{Sp}]\text{E}[\mathbf{xq}^{\text{H}}], \tag{3.6}
$$

where $\mathbf{S}$ is a symmetric matrix and $\mathbf{x}, \mathbf{y}, \mathbf{p}, \mathbf{q}$ are zero-mean, jointly normally distributed complex random variables, one obtains that

$$
\begin{aligned}
\text{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\text{H}}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)] &= \text{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)]\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\text{H}}(t)\text{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)] \\
&\quad + \text{E}[\boldsymbol{\varphi}^{\text{H}}(t)\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\text{H}}(t)\boldsymbol{\varphi}(t)]\text{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)] \\
&= \boldsymbol{\Phi}\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\text{H}}(t)\boldsymbol{\Phi} + \boldsymbol{\theta}^{\text{H}}(t)\boldsymbol{\Phi}\boldsymbol{\theta}(t)\boldsymbol{\Phi},
\end{aligned} \tag{3.7}
$$

and

$$
\text{E}[e(t)e^*(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)] = \sigma_e^2\boldsymbol{\Phi}. \tag{3.8}
$$

Hence

$$
\text{cov}[\widetilde{\boldsymbol{\theta}}(t)] = \boldsymbol{\theta}^{\text{H}}(t)\boldsymbol{\Phi}\boldsymbol{\theta}(t)\boldsymbol{\Phi}^{-1} + \sigma_e^2\boldsymbol{\Phi}^{-1} \geq \sigma_e^2\boldsymbol{\Phi}^{-1}. \tag{3.9}
$$

Note that the covariance matrix of the estimation errors depends on the way parameters vary with time. It can be upper-bounded

$$
\text{cov}[\widetilde{\boldsymbol{\theta}}(t)] \leq \frac{\lambda_{\max}(\boldsymbol{\Phi})||\boldsymbol{\theta}(t)||^2}{\lambda_{\min}(\boldsymbol{\Phi})}\mathbf{I}_n + \frac{\sigma_e^2}{\lambda_{\min}(\boldsymbol{\Phi})}\mathbf{I}_n = \frac{\lambda_{\max}(\boldsymbol{\Phi})||\boldsymbol{\theta}(t)||^2 + \sigma_e^2}{\lambda_{\min}(\boldsymbol{\Phi})}\mathbf{I}_n, \tag{3.10}
$$

which stems from the facts that $\boldsymbol{\theta}^{\text{H}}(t)\boldsymbol{\Phi}\boldsymbol{\theta}(t) \leq \lambda_{\max}(\boldsymbol{\Phi})||\boldsymbol{\theta}(t)||^2$ and $\boldsymbol{\Phi}^{-1} \leq \frac{1}{\lambda_{\min}(\boldsymbol{\Phi})}\mathbf{I}_n$. Because the covariance matrix of preestimation errors depends strongly on the values of true parameters, the variance of preestimation errors may be very high. This justifies the need for the second step - postfiltering.

In the special case, where $\{u(t)\}$ is a sequence of independent circular random variables (which is often the case in telecommunication applications), the covariance matrix of the regression vector is equal to $\boldsymbol{\Phi} = \sigma_u^2\mathbf{I}_n$. In this situation

$$
\text{cov}[\widetilde{\boldsymbol{\theta}}(t)] = ||\boldsymbol{\theta}(t)||^2\mathbf{I}_n + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I}_n = \left[||\boldsymbol{\theta}(t)||^2 + \frac{\sigma_e^2}{\sigma_u^2}\right]\mathbf{I}_n \cong \sigma_z^2\mathbf{I}_n, \tag{3.11}
$$

which means that preestimation errors are mutually uncorrelated.[1]

Figure 3.1 shows the protoestimates obtained for one of the illustrative systems described in the first chapter. The figure illustrates the capability of preestimates to "x-ray" the system and simultaneously shows the need for postfiltering.

Remark

For real-valued systems and signals, the covariance matrix of protoestimates will be different (see page 2 in [73] for the derivation), because according to the Isserlis' theorem

$$
\text{E}[xypq] = \text{E}[xy]\text{E}[pq] + \text{E}[xp]\text{E}[yq] + \text{E}[xq]\text{E}[yp],
$$

which holds true for zero-mean random Gaussian variables, we have that

$$
\text{E}[\mathbf{xy}^{\text{T}}\mathbf{Spq}^{\text{T}}] = \text{E}[\mathbf{xy}^{\text{T}}]\mathbf{S}\text{E}[\mathbf{pq}^{\text{T}}] + \text{E}[\mathbf{xp}^{\text{T}}]\mathbf{S}\text{E}[\mathbf{yq}^{\text{T}}] + \text{E}[\mathbf{xq}^{\text{T}}]\text{E}[\mathbf{y}^{\text{T}}\mathbf{Sp}],
$$

---

[1]This result holds also when the input signal is a sequence of mutually uncorrelated circular random variables.
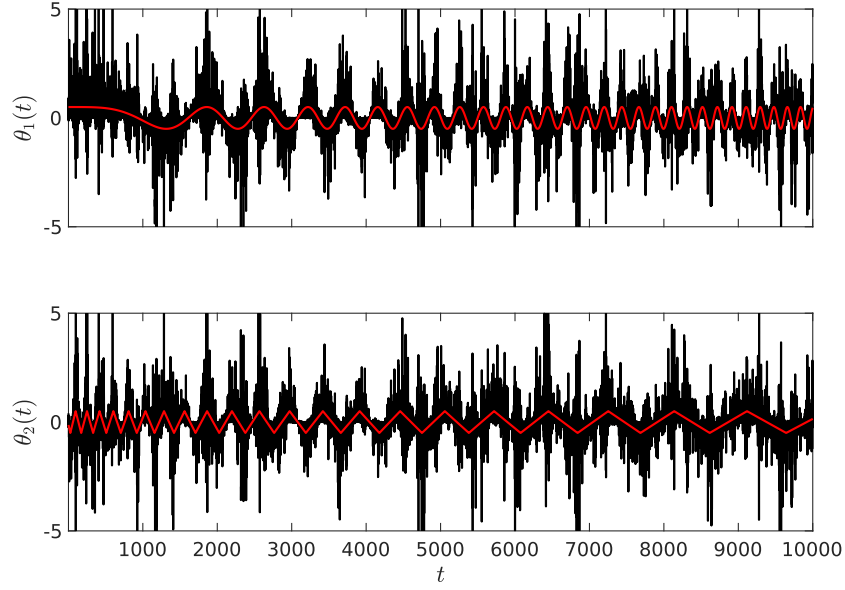
Figure 3.1: Protoestimates (black lines) superimposed on true parameter trajectories (red lines) for a two-tap FIR system.

where $\mathbf{x}, \mathbf{y}, \mathbf{p}, \mathbf{q}$ are zero-mean jointly normally distributed and $\mathbf{S}$ is a symmetric matrix. Therefore, for real-valued case, one obtains

$$\text{cov}[\widetilde{\boldsymbol{\theta}}(t)] = \boldsymbol{\theta}(t)\boldsymbol{\theta}^{\mathrm{T}}(t) + \boldsymbol{\Phi}^{-1}\boldsymbol{\theta}^{\mathrm{T}}(t)\boldsymbol{\Phi}\boldsymbol{\theta}(t) + \sigma_e^2\boldsymbol{\Phi}^{-1}. \tag{3.12}$$

Similarly, the upper and lower bounds can be found as

$$\sigma_e^2\boldsymbol{\Phi}^{-1} \leq \text{cov}[\widetilde{\boldsymbol{\theta}}(t)] \leq \left[1 + \frac{\lambda_{\max}(\boldsymbol{\Phi})}{\lambda_{\min}(\boldsymbol{\Phi})}\right] ||\boldsymbol{\theta}(t)||^2\mathbf{I}_n + \sigma_e^2\boldsymbol{\Phi}^{-1}. \tag{3.13}$$

Note that when the input signal is a sequence of uncorrelated random variables

$$\text{cov}[\widetilde{\boldsymbol{\theta}}(t)] = \boldsymbol{\theta}(t)\boldsymbol{\theta}^{\mathrm{T}}(t) + \left[||\boldsymbol{\theta}(t)||^2 + \frac{\sigma_e^2}{\sigma_u^2}\right]\mathbf{I}_n, \tag{3.14}$$

and the preestimation errors are usually correlated.

3.1.2 Postfiltering

In the second, postfiltering stage, one can use any well-established filtration technique to obtain the final fLBF estimates. In the article [17] Kalman smoothing was used as postfiltering method, while in the paper [77], the application of noncausal lowpass FIR filters was described. In this thesis, we will use the Savitzky-Golay filtration [91]. This filtering method boils down to adopting the assumption (A2.1) and solving the following problem

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t) = \arg\min_{\boldsymbol{\alpha}} \sum_{i=-k}^{k} w_k(i)||\widetilde{\boldsymbol{\theta}}(t) - \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}||^2, \tag{3.15}$$

which under (A2.2) becomes

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t) = \sum_{i=-k}^{k} \widetilde{\boldsymbol{\theta}}(t+i) \otimes w_k(i)\mathbf{f}_{m|k}(i). \tag{3.16}$$

The final estimates take the form

$$\widehat{\boldsymbol{\theta}}_{m|k}^{\text{fLBF}}(t) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t) = \sum_{i=-k}^{k} w_k(i)\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(i)\widetilde{\boldsymbol{\theta}}(t+i) = \sum_{i=-k}^{k} h_{m|k}^{\text{fLBF}}(i)\widetilde{\boldsymbol{\theta}}(t+i), \tag{3.17}$$

where $h_{m|k}^{\mathrm{fLBF}}(i) = w_k(i)\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(i)$ is the impulse response associated with the selected basis functions. The first transition follows from the properties of the Kronecker product. When the assumption (A2.1) is fulfilled, the fLBF estimates (3.17) are unbiased, namely

$$
\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)] &= \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)\mathrm{E}[\widetilde{\boldsymbol{\theta}}(t+i)] = \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)\boldsymbol{\theta}(t+i) \\
&= \sum_{i=-k}^{k} w_k(i)\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(i)\mathbf{F}_{m|k}(i)\boldsymbol{\alpha}(t) = \mathbf{F}_{m|k}(0)\boldsymbol{\alpha}(t) = \boldsymbol{\theta}(t).
\end{aligned}
\tag{3.18}
$$

The estimation errors can be written down as

$$
\begin{aligned}
\Delta\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t) &= \widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t) - \boldsymbol{\theta}(t) = \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)\widetilde{\boldsymbol{\theta}}(t+i) - \boldsymbol{\theta}(t) \\
&= \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+i)\boldsymbol{\theta}(t+i) - \boldsymbol{\theta}(t) + \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t+i)e^*(t+i) \\
&= \mathbf{v}_1(t) + \mathbf{v}_2(t),
\end{aligned}
\tag{3.19}
$$

where under (A2.1)

$$
\mathbf{v}_1(t) = \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)[\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+i) - \mathbf{I}_n]\boldsymbol{\theta}(t+i)
$$

$$
\mathbf{v}_2(t) = \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t+i)e^*(t+i).
$$

The covariance matrix of estimation errors is defined as

$$
\begin{aligned}
\mathrm{cov}[\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)] &= \mathrm{E}\{\Delta\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)[\Delta\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)]^{\mathrm{H}}\} = \mathrm{E}\{[\mathbf{v}_1(t) + \mathbf{v}_2(t)][\mathbf{v}_1(t) + \mathbf{v}_2(t)]^{\mathrm{H}}\} \\
&= \mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_1^{\mathrm{H}}(t)] + \mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_2^{\mathrm{H}}(t)] + \mathrm{E}[\mathbf{v}_2(t)\mathbf{v}_1^{\mathrm{H}}(t)] + \mathrm{E}[\mathbf{v}_2(t)\mathbf{v}_2^{\mathrm{H}}(t)],
\end{aligned}
\tag{3.20}
$$

It is easy to check that $\mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_2^{\mathrm{H}}(t)] = \mathrm{E}[\mathbf{v}_2(t)\mathbf{v}_1^{\mathrm{H}}(t)] = 0$ under (A2.5), because the noise is independent of the input signal. It also holds that

$$
\mathrm{E}[\mathbf{v}_2(t)\mathbf{v}_2^{\mathrm{H}}(t)] = \sum_{i=-k}^{k}\sum_{j=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)h_{m|k}^{\mathrm{fLBF}}(j)\boldsymbol{\Phi}^{-1}\mathrm{E}[\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+j)]\mathrm{E}[e^*(t+i)e(t+j)]\boldsymbol{\Phi}^{-1},
\tag{3.21}
$$

where the expectation of the product was separated into the product of expectations because of the independence of the input signal and the measurement noise. Because $\{e(t)\}$ is a sequence of mutually independent random variables, this expression can be further simplified to

$$
\mathrm{E}[\mathbf{v}_2(t)\mathbf{v}_2^{\mathrm{H}}(t)] = \sum_{i=-k}^{k} [h_{m|k}^{\mathrm{fLBF}}(i)]^2\sigma_e^2\boldsymbol{\Phi}^{-1} = \frac{\sigma_e^2\boldsymbol{\Phi}^{-1}}{l_{m|k}^{\mathrm{fLBF}}},
\tag{3.22}
$$

where $l_{m|k}^{\mathrm{fLBF}} = \left\{ \sum_{i=-k}^{k}[h_{m|k}^{\mathrm{fLBF}}(i)]^2 \right\}^{-1}$ is an equivalent number of observations used by the fLBF method. Note that

$$
\begin{aligned}
\mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_1^{\mathrm{H}}(t)] &= \sum_{i=-k}^{k}\sum_{j=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)h_{m|k}^{\mathrm{fLBF}}(j)\mathrm{E}\{[\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+i) - \mathbf{I}_n]\boldsymbol{\theta}(t+i)\boldsymbol{\theta}(t+j)\times \\
&\qquad\times [\boldsymbol{\varphi}(t+j)\boldsymbol{\varphi}^{\mathrm{H}}(t+j)\boldsymbol{\Phi}^{-1} - \mathbf{I}_n]\} \\
&= \sum_{i=-k}^{k}\sum_{j=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)h_{m|k}^{\mathrm{fLBF}}(j)\{\boldsymbol{\Phi}^{-1}\mathrm{E}[\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+i)\boldsymbol{\theta}(t+i)\boldsymbol{\theta}(t+j)\boldsymbol{\varphi}(t+j)\times \\
&\qquad\times \boldsymbol{\varphi}^{\mathrm{H}}(t+j)]\boldsymbol{\Phi}^{-1} - \boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j)\},
\end{aligned}
\tag{3.23}
$$

where the last transition follows from the fact that

$$\mathbf{\Phi}^{-1}\mathrm{E}[\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+i)]\boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j) = \mathrm{E}[\boldsymbol{\varphi}(t+j)\boldsymbol{\varphi}^{\mathrm{H}}(t+j)]\mathbf{\Phi}^{-1}\boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j) = \boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j).$$

To obtain analytical results that are easier to interpret, we will adopt the assumption (A2.3). This allows one to again use the results (3.6), leading to

$$
\begin{aligned}
\mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_1^{\mathrm{H}}(t)] &= \sum_{i=-k}^{k}\sum_{j=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)h_{m|k}^{\mathrm{fLBF}}(j)\{\mathbf{\Phi}^{-1}\mathbf{\Phi}\boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j)\mathbf{\Phi}\mathbf{\Phi}^{-1} \\
&\quad + \mathbf{\Phi}^{-1}\mathrm{Tr}[\mathbf{\Phi}_{ij}^{\mathrm{H}}\boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j)]\mathbf{\Phi}_{ij}\mathbf{\Phi}^{-1} - \boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j)\} \qquad (3.24) \\
&= \sum_{i=-k}^{k}\sum_{j=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)h_{m|k}^{\mathrm{fLBF}}(j)\mathrm{Tr}[\mathbf{\Phi}_{ij}^{\mathrm{H}}\boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j)]\mathbf{\Phi}^{-1}\mathbf{\Phi}_{ij}\mathbf{\Phi}^{-1},
\end{aligned}
$$

where

$$\mathbf{\Phi}_{ij} = \mathrm{E}[\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+j)]. \qquad (3.25)$$

Now we will show that the modulus of any element of a matrix $E[\mathbf{v}_1(t)\mathbf{v}_1^{\mathrm{H}}(t)]$:
$v_{n_1 n_2}(t) = \mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_1^{\mathrm{H}}(t)]_{(n_1,n_2)}$, $n \geq n_2 \geq n_1 \geq 1$ tends to zero as k increases. First, note that

$$\mathrm{E}[\mathbf{v}_1(t)\mathbf{v}_1^{\mathrm{H}}(t)] \leq \frac{1}{\lambda_{\min}^2(\mathbf{\Phi})}\sum_{i=-k}^{k}\sum_{j=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)h_{m|k}^{\mathrm{fLBF}}(j)\mathrm{Tr}[\mathbf{\Phi}_{ij}^{\mathrm{H}}\boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j)]\mathbf{\Phi}_{ij}. \qquad (3.26)$$

As a consequence, one obtains

$$|v_{n_1 n_2}(t)| \leq \frac{1}{\lambda_{\min}^2(\mathbf{\Phi})}\sum_{i=-k}^{k}\sum_{j=-k}^{k}|h_{m|k}^{\mathrm{fLBF}}(i)h_{m|k}^{\mathrm{fLBF}}(j)||\mathrm{Tr}[\mathbf{\Phi}_{ij}^{\mathrm{H}}\boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j)]||[\mathbf{\Phi}_{ij}]_{(n_1 n_2)}|. \qquad (3.27)$$

We will assume that each parameter trajectory is bounded, which means that $\exists_{M_j>0}\forall_{t\in\mathbb{Z}}\ |\theta_j(t)| \leq M_j$. As a result $\exists_{M>0}\forall_{i,j\in\mathbb{Z}}\ ||\boldsymbol{\theta}^{\mathrm{H}}(j)\boldsymbol{\theta}(i)||^2 < M$. Consequently $\boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j) \leq M\mathbf{I}_n$ and

$$
\begin{aligned}
|\mathrm{Tr}[\mathbf{\Phi}_{ij}^{\mathrm{H}}\boldsymbol{\theta}(t+i)\boldsymbol{\theta}^{\mathrm{H}}(t+j)]| &\leq M|\mathrm{Tr}[\mathbf{\Phi}_{ij}^{\mathrm{H}}]| \leq M\sum_{s=1}^{n}|\mathrm{E}[u(t+i+s-1)u^*(t+j+s-1)]| \\
&= M\sum_{s=1}^{n}|r_u(j-i)| \leq Mnc_2\beta^{|j-i|}.
\end{aligned}
\qquad (3.28)
$$

Note also that

$$|[\mathbf{\Phi}_{ij}]_{(n_1 n_2)}| = |r_u(j - n_2 - i + n_1)| \leq c_2\beta^{|j-n_2-i+n_1|}. \qquad (3.29)$$

Finally, using lemma 2.1 and the same arguments as in the previous chapter[2], one obtains that

$$\exists_{c_1>0}\exists_{k_0>0}\forall_{k>k_0}\forall_{i,j\in I_k}\quad |h_{m|k}^{\mathrm{fLBF}}(i)h_{m|k}^{\mathrm{fLBF}}(j)| \leq \frac{c_1^4}{k^2}. \qquad (3.30)$$

Combining (3.27) - (3.30), one arrives at the following formula

$$
\begin{aligned}
\exists_{D>0}\exists_{k_0>0}\forall_{k>k_0}\quad |v_{n_1 n_2}(t)| &\leq \frac{D}{k^2}\sum_{i=-k}^{k}\sum_{j=-k}^{k}\beta^{|j-i|}\beta^{|j-n_2-i+n_1|} \\
&\leq \frac{D}{k^2}\sum_{i=-k}^{k}\sum_{j=-k}^{k}\beta^{2|j-i|}\beta^{-|n_2-n_1|} = \mathcal{O}\left(\frac{1}{k}\right).
\end{aligned}
\qquad (3.31)
$$

The last transition follows from the fact that $|j-n_2-i+n_1| \geq |j-i|-|n_2-n_1|$, hence $\beta^{|j-n_2-i+n_1|} \leq \beta^{|j-i|}\beta^{-|n_2-n_1|}$.

---

[2]see pages 20-22.

Remark

Note that such a formulation of identification strategy allows one to use different basis functions and different lengths of analysis window for each system parameter. The decoupled problem can be formulated as follows

$$\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fLBF}}(t) = \arg\min_{\boldsymbol{\alpha}_j} \sum_{i=-k}^{k} w_{j,k}(i)|\widetilde{\theta}_j(t) - \mathbf{f}_{j,mk}^{\text{H}}(i)\boldsymbol{\alpha}_j|^2 = \sum_{i=-k}^{k} \widetilde{\theta}_j(t)w_{j,k}(i)\mathbf{f}_{j,m|k}(i), \quad j = 1,\dots,n.$$

$$(3.32)$$

The final estimates can be written down as

$$\widehat{\theta}_{j,m|k}^{\text{fLBF}}(t) = \mathbf{f}_{j,mk}^{\text{H}}(0)\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fLBF}}(t) = \sum_{i=-k}^{k} w_{j,k}(i)\mathbf{f}_{j,m|k}^{\text{H}}(0)\mathbf{f}_{j,m|k}(i)\widetilde{\theta}_j(t)$$

$$= \sum_{i=-k}^{k} h_{j,m|k}^{\text{fLBF}}(i)\widetilde{\theta}_j(t), \quad j = 1,\dots,n.$$

$$(3.33)$$

When basis functions, lengths of analysis window and weigthing sequences are the same for every parameter, then the formulation above boils down to (3.17).

3.1.3 Connection with the LBF method

Note that (3.17) can be written down as

$$\widehat{\boldsymbol{\theta}}_{m|k}^{\text{fLBF}}(t) = \sum_{i=-k}^{k} w_k(i)\mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(i)\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t+i)y^*(t+i).$$

$$(3.34)$$

Combining (2.77) and (2.46) yields

$$\widehat{\boldsymbol{\theta}}_{m|k}^{\text{LBF}}(t) \cong \widetilde{\boldsymbol{\theta}}_{m|k}^{\text{BF}}(t+s|t)|_{s=0} = \mathbf{F}_{m|k}(0)\mathbf{P}_0^{-1} \sum_{i=-k}^{k} w_k(i)\boldsymbol{\psi}_{m|k}(t,i)y^*(t+i)$$

$$= [\mathbf{I}_n \otimes \mathbf{f}_{m|k}^{\text{H}}(0)][\boldsymbol{\Phi}^{-1} \otimes \mathbf{I}_m] \sum_{i=-k}^{k} w_k(i)[\boldsymbol{\varphi}(t+i) \otimes \mathbf{f}_{m|k}(i)]y^*(t+i)$$

$$= \sum_{i=-k}^{k} w_k(i)\mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(i)\boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t+i)y^*(t+i) = \widehat{\boldsymbol{\theta}}_{m|k}^{\text{fLBF}}(t),$$

$$(3.35)$$

which is justified for sufficiently large values of k. In other words, when one chooses the same basis functions and the same weighting sequence in the LBF and fLBF methods, then the fLBF can be regarded as a close approximation of the LBF technique.

Note that the statistical properties, i.e. the formulae for bias and covariance matrix, of the LBF method were derived in the previous chapter for a simplified estimator for which $\mathbf{P}_{m|k}(t)$ was replaced by $\mathbf{P}_0$, just as in the equation (3.35). This means that for sufficiently large values of k, the statistical properties of the fLBF and LBF methods should be approximately the same.

3.2 Direct preestimates

If the covariance matrix of input signal is not known, one can use the following formula for the direct preestimate

$$\widetilde{\boldsymbol{\theta}}^d(t) = \widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)\boldsymbol{\varphi}(t)y^*(t),$$

$$(3.36)$$

where

$$\widehat{\boldsymbol{\Phi}}_\lambda(t) = \frac{1}{L(t)} \sum_{i=1}^{t} \lambda^{t-i}\boldsymbol{\varphi}(i)\boldsymbol{\varphi}^{\text{H}}(i),$$

$$(3.37)$$

is the exponentially weighted estimate of $\boldsymbol{\Phi}$, where $\lambda \in (0,1)$ is the so-called forgetting constant and $L(t) = \sum_{i=1}^{t} \lambda^{t-i} = \lambda L(t-1) + 1 = \frac{1-\lambda^t}{1-\lambda}$, $L(0) = 0$ denotes the effective length of the wieghting sequence. One can show that under (A2.3) [66]

$$\lim_{\lambda \to 1} \widehat{\boldsymbol{\Phi}}_\lambda(t) = \mathrm{E}[\widehat{\boldsymbol{\Phi}}_\lambda(t)] = \boldsymbol{\Phi}, \tag{3.38}$$

in a mean squared sense.

For values of λ sufficiently close to 1, $\widehat{\boldsymbol{\Phi}}_\lambda(t) \cong \boldsymbol{\Phi}$ and direct preestimates can be seen as a close approximation of protoestimates, and as a consequence

$$\mathrm{E}[\widetilde{\boldsymbol{\theta}}^d(t)] \cong \boldsymbol{\theta}(t). \tag{3.39}$$

The preestimation errors take the form

$$\Delta\widetilde{\boldsymbol{\theta}}^d(t) = \widetilde{\boldsymbol{\theta}}^d(t) - \boldsymbol{\theta}(t) = [\widehat{\boldsymbol{\Phi}}_\lambda(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t) - \mathbf{I}_n]\boldsymbol{\theta}(t) + \widehat{\boldsymbol{\Phi}}_\lambda(t)\boldsymbol{\varphi}(t)e^*(t). \tag{3.40}$$

One can again use the approximation $\widehat{\boldsymbol{\Phi}}_\lambda(t) \cong \boldsymbol{\Phi}$ valid for λ sufficiently close to 1, to obtain that

$$\mathrm{cov}[\widetilde{\boldsymbol{\theta}}^d(t)] = \mathrm{E}\left\{\Delta\widetilde{\boldsymbol{\theta}}^d(t)[\Delta\widetilde{\boldsymbol{\theta}}^d(t)]^{\mathrm{H}}\right\} \cong \mathrm{cov}[\widetilde{\boldsymbol{\theta}}(t)]. \tag{3.41}$$

Remark

Note that the direct preestimates can be also obtained using different estimate of the input signal covariance matrix, like

$$\widehat{\boldsymbol{\Phi}}_{k_0}(t) = \frac{1}{2k_0 + 1} \sum_{i=-k_0}^{k_0} \boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+i), \tag{3.42}$$

which can approximate $\boldsymbol{\Phi}$ more accurately. This formulation was proposed in the paper [71] where direct preestimates were introduced.

Remark 2

The inverse of matrix $\widehat{\boldsymbol{\Phi}}_\lambda(t)$ can be easily updated using the Sherman-Morrison formula, which reduces the number of required computations.

3.3 Indirect preestimates

The indirect preestimates were proposed in [76] for the purpose of causal estimation and then applied in [70] to noncausal estimation. They are based on the exponentially weighted least squares estimates of system parameters

$$\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{t} \lambda^{t-i}|y(i) - \boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\varphi}(i)|^2, \tag{3.43}$$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) = \mathbf{R}^{-1}(t)\mathbf{r}(t), \tag{3.44}$$

where

$$\begin{aligned}\mathbf{R}(t) &= \sum_{i=1}^{t} \lambda^{t-i}\boldsymbol{\varphi}(i)\boldsymbol{\varphi}(i)^{\mathrm{H}} = \lambda\mathbf{R}(t-1) + \boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t) \\ \mathbf{r}(t) &= \sum_{i=1}^{t} \lambda^{t-i}\boldsymbol{\varphi}(i)y^*(i) = \lambda\mathbf{r}(t-1) + \boldsymbol{\varphi}(t)y^*(t),\end{aligned} \tag{3.45}$$

and the initial conditions are $\mathbf{R}(0) = 0$, $\mathbf{r}(0) = 0$.

The indirect preestimates are defined as follows

$$\widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) = L(t)\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) - \lambda L(t-1)\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1). \tag{3.46}$$

Note that $\mathbf{R}(t) = L(t)\widehat{\boldsymbol{\Phi}}_\lambda(t)$ and for λ sufficiently close to 1, the following approximation is valid

$$\mathbf{R}(t) \cong \mathrm{E}[\mathbf{R}(t)] = L(t)\boldsymbol{\Phi}, \tag{3.47}$$

which leads to

$$\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) \cong \frac{1}{L(t)}\boldsymbol{\Phi}^{-1}\mathbf{r}(t). \tag{3.48}$$

Applying this to (3.46) yields

$$
\begin{aligned}
\widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) &= L(t)\mathbf{R}^{-1}\mathbf{r}(t) - \lambda L(t-1)\mathbf{R}^{-1}(t-1)\mathbf{r}(t-1) \\
&\cong \boldsymbol{\Phi}^{-1}[\mathbf{r}(t) - \lambda \mathbf{r}(t-1)] = \boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t)y^*(t) = \widetilde{\boldsymbol{\theta}}(t).
\end{aligned}
\tag{3.49}
$$

This is the reason why the estimator (3.46) was given the name indirect preestimator in the article [70]. The formula (3.46) can be regarded as an indirect approximation of the protoestimator.

3.3.1   Properties of the indirect preestimates

Using the formula (3.48), one can conclude that the indirect preestimates are approximately unbiased

$$\mathrm{E}[\widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t)] \cong \mathrm{E}[\widetilde{\boldsymbol{\theta}}(t)] = \boldsymbol{\theta}(t). \tag{3.50}$$

Thus far, one could expect that the preestimation errors for the indirect preestimator will be similar to preestimation errors of the protoestimator and direct preestimator. Quite surprisingly, the truth turns out to be different. To show this, we will start with the recursive formula for the EWLS estimates [98]

$$\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) = \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1) + \mathbf{R}^{-1}(t)\boldsymbol{\varphi}(t)[\varepsilon^{\mathrm{EWLS}}(t)]^*, \tag{3.51}$$

where

$$\varepsilon^{\mathrm{EWLS}}(t) = y(t) - [\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1)]^{\mathrm{H}}\boldsymbol{\varphi}(t) \tag{3.52}$$

denotes one-step-ahead prediction error. Using (3.51) one can rewrite (3.46) as

$$
\begin{aligned}
\widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) &= L(t)\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) - \lambda L(t-1)\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1) \\
&= L(t)\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1) + L(t)\mathbf{R}^{-1}(t)\boldsymbol{\varphi}(t)[\varepsilon^{\mathrm{EWLS}}(t)]^* - \lambda L(t-1)\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1) \\
&= \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1) + \widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\boldsymbol{\theta}(t) + \widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)\boldsymbol{\varphi}(t)e^*(t) \\
&\quad - \widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1),
\end{aligned}
\tag{3.53}
$$

where the last transition follows from (1.2), (3.48) and (3.51). Using (3.53), one gets

$$\Delta\widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) = \widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) - \boldsymbol{\theta}(t) = [\widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t) - \mathbf{I}_n][\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1)] + \widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)\boldsymbol{\varphi}(t)e^*(t). \tag{3.54}$$

One can now easily compare preestimation errors of direct and indirect preestimates described by (3.40) and (3.54), respectively. The difference boils down to the first term that is equal to $[\widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t) - \mathbf{I}_n]\boldsymbol{\theta}(t)$ in direct preestimation and $[\widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t) - \mathbf{I}_n][\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1)]$ in indirect preestimation. Since typically $||\boldsymbol{\theta}(t)|| \gg ||\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1)||$, preestimation errors in indirect preestimation scheme have substantially smaller variability than preestimation errors in direct preeestimates. It is also worth noting that when parameters are slowly changing and $\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t-1) \cong \boldsymbol{\theta}(t)$, indirect preestimation errors are approximately white.

Using the approximation stemming from (3.38), one arrives at

$$\mathrm{cov}[\widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t)] \ge \sigma_e^2 \mathrm{E}[\widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\widehat{\boldsymbol{\Phi}}_\lambda^{-1}(t)] \cong \sigma_e^2 \boldsymbol{\Phi}^{-1}. \tag{3.55}$$

3.3.2 Steady-state preestimation formula

For a sufficiently large value of t, the steady-state version of formulae described above can be used. They can be obtained by substituting $L(t)$ with $L(\infty) = \frac{1}{1-\lambda}$ in (3.46), which results in

$$\widetilde{\boldsymbol{\theta}}^{\text{EWLS}}(t) = \frac{1}{1-\lambda}[\widehat{\boldsymbol{\theta}}^{\text{EWLS}}(t) - \lambda\widehat{\boldsymbol{\theta}}^{\text{EWLS}}(t-1)] \tag{3.56}$$

The steady-state formula (3.56) has an additional, interesting interpretation. According to [66] the expected path of the EWLS estimates in a steady state can be approximately seen as the output of a linear time-invariant lowpass filter

$$\overline{\boldsymbol{\theta}}^{\text{EWLS}}(t) = \text{E}[\widehat{\boldsymbol{\theta}}^{\text{EWLS}}(t)] \cong H^{\text{EWLS}}(q^{-1})\boldsymbol{\theta}(t), \tag{3.57}$$

where

$$H^{\text{EWLS}}(q^{-1}) = \frac{1-\lambda}{1-\lambda q^{-1}}. \tag{3.58}$$

Note that (3.56) can be rewritten as highpass filtration of the EWLS estimates

$$\widetilde{\boldsymbol{\theta}}^{\text{EWLS}}(t) = \frac{1}{H^{\text{EWLS}}(q^{-1})}\widehat{\boldsymbol{\theta}}^{\text{EWLS}}(t). \tag{3.59}$$

The EWLS estimates can be expressed as

$$\widehat{\boldsymbol{\theta}}^{\text{EWLS}}(t) = \overline{\boldsymbol{\theta}}^{\text{EWLS}}(t) + \boldsymbol{\eta}(t), \tag{3.60}$$

where $\boldsymbol{\eta}(t)$ is a zero-mean noise. Therefore,

$$\widetilde{\boldsymbol{\theta}}^{\text{EWLS}}(t) \cong \boldsymbol{\theta}(t) + \frac{1}{H^{\text{EWLS}}(q^{-1})}\boldsymbol{\eta}(t). \tag{3.61}$$

This means that they can be seen as the effect of "inverse filtering" of the EWLS estimates. From this point of view, it becomes clear that one should carefully choose the value of λ for the purpose of preestimation. On the one hand, small values of the forgetting constant are desirable because decreasing λ reduces the estimator memory which allows one to successfully track faster changes in parameter trajectories. On the other hand, estimating a large number of parameters using an estimator with a very short memory can cause practical problems. This is why one should search for a trade-off when choosing the forgetting constant. It turns out that the rule of thumb that provides satisfactory results can be formulated in the following form

$$\lambda = \max\left\{0.9, 1 - \frac{2}{n}\right\}. \tag{3.62}$$

The expression $1 - \frac{2}{n}$ follows from the assumption that the equivalent steady-state memory of the EWLS estimator [66]

$$l_\infty = \frac{1+\lambda}{1-\lambda} \cong \frac{2}{1-\lambda}$$

should not be smaller than the number of estimated parameters which yields

$$\frac{2}{1-\lambda} \geq n \iff \lambda \geq 1 - 2/n. \tag{3.63}$$

Although even for a very low value of l_∞, one usually does not encounter numerical problems when computing the inverse of a matrix $\mathbf{R}(t)$, from a statistical point of view it is questionable to use the estimator with an equivalent memory smaller than the number of estimated parameters. Such an estimator can simply lack sufficient information to track changes of all parameters in a reliable way.

3.4 LMS-based preestimates

Although the EWLS algorithm can be implemented using fast transversal filters [35], ensuring the computational complexity proportional to the number of estimated parameters, one can use even faster, gradient-based least mean squares (LMS) estimates to obtain preestimates

$$\widehat{\boldsymbol{\theta}}^{\text{LMS}}(t) = \widehat{\boldsymbol{\theta}}^{\text{LMS}}(t-1) + \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)\varepsilon^*(t)$$

$$\varepsilon(t) = y(t) - [\widehat{\boldsymbol{\theta}}^{\text{LMS}}(t-1)]^{\text{H}}\boldsymbol{\varphi}(t),$$

(3.64)

where $\mu > 0$ is a stepsize parameter, typically of a small value.

The following preestimation formula can be proposed

$$\widetilde{\boldsymbol{\theta}}^{\text{LMS}}(t) = \frac{1}{1-\lambda_L}[\widehat{\boldsymbol{\theta}}^{\text{LMS}}(t) - \lambda_L\widehat{\boldsymbol{\theta}}^{\text{LMS}}(t-1)],$$

(3.65)

where $\lambda_L = 1 - \mu$. Note that this formula closely resembles the steady-state formula (3.56) for EWLS-based indirect preestimates.

3.4.1 Mean squared stability of the preestimation scheme

The mean squared stability of the LMS-based preestimation algorithm depends on the mean squared stability of the LMS algorithm. In this section we derive the upper bound for the step-size parameter, which guarantees the mean squared stability of the LMS algorithm under some simplifying assumptions. We assume that the parameters are constant. We will also adopt the following assumption

(A3.1) $\{\boldsymbol{\varphi}(t)\}$, independent of $\{e(t)\}$, is a sequence of zero-mean mutually independent and identically distributed n-dimensional complex random variables with the covariance matrix $\text{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)] = \sigma_u^2\mathbf{I}_n$.

The estimation errors can be defined as

$$\boldsymbol{\rho}(t) = \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^{\text{LMS}}(t) = \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^{\text{LMS}}(t-1) - \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)y^*(t) + \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)\widehat{\boldsymbol{\theta}}^{\text{LMS}}(t-1)$$

$$= \boldsymbol{\rho}(t-1) - \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)\boldsymbol{\theta} + \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)\widehat{\boldsymbol{\theta}}^{\text{LMS}}(t-1) - \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)e^*(t) \quad (3.66)$$

$$= \left[\mathbf{I}_n - \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)\right]\boldsymbol{\rho}(t-1) - \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)e^*(t).$$

Note that $\boldsymbol{\rho}(t-1) = \boldsymbol{\theta}(t-1) - \widehat{\boldsymbol{\theta}}^{\text{LMS}}(t-1)$, which is independent of $\boldsymbol{\varphi}(t)$ under (A3.1). This leads to

$$\text{E}[\boldsymbol{\rho}(t)] = (1-\mu)\text{E}[\boldsymbol{\rho}(t-1)],$$

(3.67)

which means that the convergence in the mean sense is guaranteed for $\mu \in (0,1)$.

Now let us inspect the mean squared convergence. First note that

$$\boldsymbol{\rho}(t) = \mathbf{v}_1(t) - \mathbf{v}_2(t),$$

where

$$\mathbf{v}_1(t) = \left[\mathbf{I}_n - \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\text{H}}(t)\right]\boldsymbol{\rho}(t-1)$$

$$\mathbf{v}_2(t) = \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)e^*(t).$$

Hence

$$\text{E}[||\boldsymbol{\rho}(t)||^2] = \text{E}[\boldsymbol{\rho}^{\text{H}}(t)\boldsymbol{\rho}(t)] = \text{E}[\mathbf{v}_1^{\text{H}}(t)\mathbf{v}_1(t)] - \text{E}[\mathbf{v}_1^{\text{H}}(t)\mathbf{v}_2(t)] - \text{E}[\mathbf{v}_2^{\text{H}}(t)\mathbf{v}_1(t)] + \text{E}[\mathbf{v}_2^{\text{H}}(t)\mathbf{v}_2(t)].$$

Because $e(t)$ is independent of $\boldsymbol{\varphi}(t)$ and of $\boldsymbol{\rho}(t-1)$, it holds that $\mathrm{E}[\mathbf{v}_1^{\mathrm{H}}(t)\mathbf{v}_2(t)] = \mathrm{E}[\mathbf{v}_2^{\mathrm{H}}(t)\mathbf{v}_1(t)] = 0$. For the same reason

$$\mathrm{E}[\mathbf{v}_2^{\mathrm{H}}(t)\mathbf{v}_2(t)] = \frac{\mu^2}{(\sigma_u^2)^2}\mathrm{E}[|e(t)|^2]\mathrm{E}[\boldsymbol{\varphi}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t)] = \frac{n\mu^2\sigma_e^2}{\sigma_u^2}. \tag{3.68}$$

The last quantity takes the form

$$\begin{aligned}
\mathrm{E}[\mathbf{v}_1^{\mathrm{H}}(t)\mathbf{v}_1(t)] &= \mathrm{E}\left\{\boldsymbol{\rho}^{\mathrm{H}}(t-1)\left[\mathbf{I}_n - \frac{\mu}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\right]^2\boldsymbol{\rho}(t-1)\right\} \\
&= \mathrm{E}[||\boldsymbol{\rho}(t-1)||^2] - 2\frac{\mu}{\sigma_u^2}\mathrm{E}[\boldsymbol{\rho}^{\mathrm{H}}(t-1)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\boldsymbol{\rho}(t-1)] \\
&\quad + \frac{\mu^2}{(\sigma_u^2)^2}\mathrm{E}[\boldsymbol{\rho}^{\mathrm{H}}(t-1)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\boldsymbol{\rho}(t-1)].
\end{aligned} \tag{3.69}$$

The second term, due to independecne between $\boldsymbol{\rho}(t-1)$ and $\boldsymbol{\varphi}(t)$, is equal to

$$\mathrm{E}[\boldsymbol{\rho}^{\mathrm{H}}(t-1)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\boldsymbol{\rho}(t-1)] = \sigma_u^2\mathrm{E}[||\boldsymbol{\rho}(t-1)||^2].$$

To obtain results, that are easier to analyze, we will additionally assume that the regression vectors are jointly normally distributed. Under this assumption, one can use the results from the extension of the Isserlis' theorem [88], to obtain

$$\mathrm{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)] = \sigma_u^4(n+1)\mathbf{I}_n, \tag{3.70}$$

which means that

$$\mathrm{E}[||\boldsymbol{\rho}(t)||^2] = [1 - 2\mu + (n+1)\mu^2]\mathrm{E}[||\boldsymbol{\rho}(t-1)||^2] + \frac{n\mu^2\sigma_e^2}{\sigma_u^2}. \tag{3.71}$$

Hence, to guarantee convergence in the mean squared sense, the following condition has to be met

$$\mu < \frac{2}{n+1}. \tag{3.72}$$

Even though this bound was derived for constant parameters and restrictive assumptions about mixing conditions and distribution of regression vectors, numerical experiments have shown that when these assumptions are not met (i.e. parameters are time-varying and the regression vectors are dependent and not Gaussian), this stability criterion (3.72) still allows one to pretty accurately determine when the LMS algorithm becomes unstable.

Remark

For the real-valued case, (3.70) looks differently because of the discrepancy in the moment theorem for Gaussian variables for real and complex random variables, namely

$$\mathrm{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)] = \sigma_u^4(n+2)\mathbf{I}_n, \tag{3.73}$$

which leads to

$$\mu < \frac{2}{n+2}. \tag{3.74}$$

3.4.2 Bias and preestimation errors

Now, we show the expressions for the expected value and preestimation noise for LMS-based preestimates under the assumption (A3.1). First, note that

$$\begin{aligned}
\widetilde{\boldsymbol{\theta}}^{\mathrm{LMS}}(t) &= \frac{\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t) - (1-\mu)\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1)}{\mu} \\
&= \frac{1}{\mu}\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1) + \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)y^*(t) - \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1) - \frac{1}{\mu}\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1) + \widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1) \\
&= \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\boldsymbol{\theta}(t) + \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)e^*(t) - \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1) + \widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1) \\
&= \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\boldsymbol{\theta}(t) + \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)e^*(t) + [\mathbf{I}_n - \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)]\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1).
\end{aligned} \tag{3.75}$$

Hence,

$$\mathrm{E}[\widetilde{\boldsymbol{\theta}}^{\mathrm{LMS}}(t)] = \frac{1}{\sigma_u^2}\mathrm{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)]\boldsymbol{\theta}(t) + \frac{1}{\sigma_u^2}\mathrm{E}[\boldsymbol{\varphi}(t)e^*(t)] + \mathrm{E}\{[\mathbf{I}_n - \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)]\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1)\}. \quad (3.76)$$

The second term of the formula above is equal zero because the noise is independent of the regression vector. Based on the assumption (A3.1), $\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1)$ depends on $\boldsymbol{\varphi}(t-1)$ but not on $\boldsymbol{\varphi}(t)$, so the expectation can be replaced by the product of expectations, which is equal to zero. Therefore

$$\mathrm{E}[\widetilde{\boldsymbol{\theta}}^{\mathrm{LMS}}(t)] = \boldsymbol{\theta}(t). \quad (3.77)$$

The preestimation errors can be expressed in the form

$$\Delta\widetilde{\boldsymbol{\theta}}^{\mathrm{LMS}}(t) = \widetilde{\boldsymbol{\theta}}^{\mathrm{LMS}}(t) - \boldsymbol{\theta}(t) = \left[\frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t) - \mathbf{I}_n\right][\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1)] + \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)e^*(t), \quad (3.78)$$

which is very similar to the expression (3.54). Similarly as in the case of the EWLS-based preestimates

$$\mathrm{cov}[\widetilde{\boldsymbol{\theta}}^{\mathrm{LMS}}(t)] \geq \frac{1}{\sigma_u^4}\mathrm{E}[e(t)e^*(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)] = \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I}_n. \quad (3.79)$$

Remark

Note that the assumption (A3.1) can be relaxed resulting in approximate unbiasedness. Because $\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)$ is a fast changing quantity and $\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1)$ is slowly changing, based on the averaging theory (see [43] for more details), one can write down

$$\mathrm{E}\{[\mathbf{I}_n - \frac{1}{\sigma_u^2}\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)]\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1)\} \cong \mathrm{E}\Big\{\{\mathbf{I}_n - \frac{1}{\sigma_u^2}\mathrm{E}[\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)]\}\widehat{\boldsymbol{\theta}}^{\mathrm{LMS}}(t-1)\Big\} = 0.$$

The same argument can be used to derive the approximate criterion for the mean squared stability of the LMS algorithm (see [32]).

3.5   Enhanced preestimates

Enhancement of preestimates is an iterative procedure introduced in [73] aiming at reducing the variance of preestimation errors without affecting the approximate unbiasedness property. The enhanced preestimates are obtained using the fLBF estimates of system parameters

$$\boldsymbol{\theta}^{\dagger}(t) = \widetilde{\boldsymbol{\theta}}^{\mathrm{d}}(t) + [\mathbf{I}_n - \widehat{\boldsymbol{\Phi}}_{\lambda}^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)]\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t) = \widetilde{\boldsymbol{\theta}}^{\mathrm{d}}(t) + \widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t) - \frac{\widetilde{\boldsymbol{\theta}}^{\mathrm{d}}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)}{y^*(t)}\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)$$
$$= \widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t) + \left[1 - \frac{\boldsymbol{\varphi}^{\mathrm{H}}(t)\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)}{y^*(t)}\right]\widetilde{\boldsymbol{\theta}}^{\mathrm{d}}(t). \quad (3.80)$$

Note that $\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)$ and $\widehat{\boldsymbol{\Phi}}_{\lambda}^{-1}(t)$ are slowly changing, while $\boldsymbol{\varphi}(t)$ is a fast-changing quantity. Hence, one can use the results of the averaging theory to conclude that

$$\mathrm{E}[\boldsymbol{\theta}^{\dagger}(t)] \cong \boldsymbol{\theta}(t), \quad (3.81)$$

which holds under (A2.1).

It is easy to check that preestimation errors for the enhanced preestimates can be expressed as

$$\Delta\widetilde{\boldsymbol{\theta}}^{\dagger}(t) = \boldsymbol{\theta}^{\dagger}(t) - \boldsymbol{\theta}(t) = [\widehat{\boldsymbol{\Phi}}_{\lambda}^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t) - \mathbf{I}_n][\boldsymbol{\theta}(t) - \widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)] + \widehat{\boldsymbol{\Phi}}_{\lambda}^{-1}(t)\boldsymbol{\varphi}(t)e^*(t). \quad (3.82)$$

Since typically the accuracy of fLBF estimates is higher than the accuracy of the ELWS estimates, the variance of preestimation errors is lower for enhanced preestimates than for indirect preestimates.

Note that the oracle enhanced preestimate would use $\boldsymbol{\theta}(t)$ instead of $\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)$ resulting in

$$\widetilde{\boldsymbol{\theta}}^{\mathrm{o}}(t) = \boldsymbol{\theta}(t) + \widehat{\boldsymbol{\Phi}}_{\lambda}^{-1}(t)\boldsymbol{\varphi}(t)e^*(t).$$

Such an oracle preestimate would be characterized by the smallest possible variance of preestimation errors. It is worth noting that this sets the lower bound on the covariance matrix of the enhanced preestimates

$$\mathrm{cov}[\widetilde{\boldsymbol{\theta}}^{\dagger}(t)] \geq \sigma_e^2 \mathrm{E}[\widehat{\boldsymbol{\Phi}}_{\lambda}^{-1}(t)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t)\widehat{\boldsymbol{\Phi}}_{\lambda}^{-1}(t)] \cong \sigma_e^2 \boldsymbol{\Phi}^{-1}. \tag{3.83}$$

Exemplary preestimation errors are shown in the figure 3.2.

Summarizing, the algorithm to use enhanced preestimates would be

1. Compute direct and indirect preestimates.

2. Postfilter indirect preestimates.

3. Use direct preestimates and filtered indirect preestimates to obtain enhanced preestimates.

4. Filter enhanced preestimates to get the final estimates.
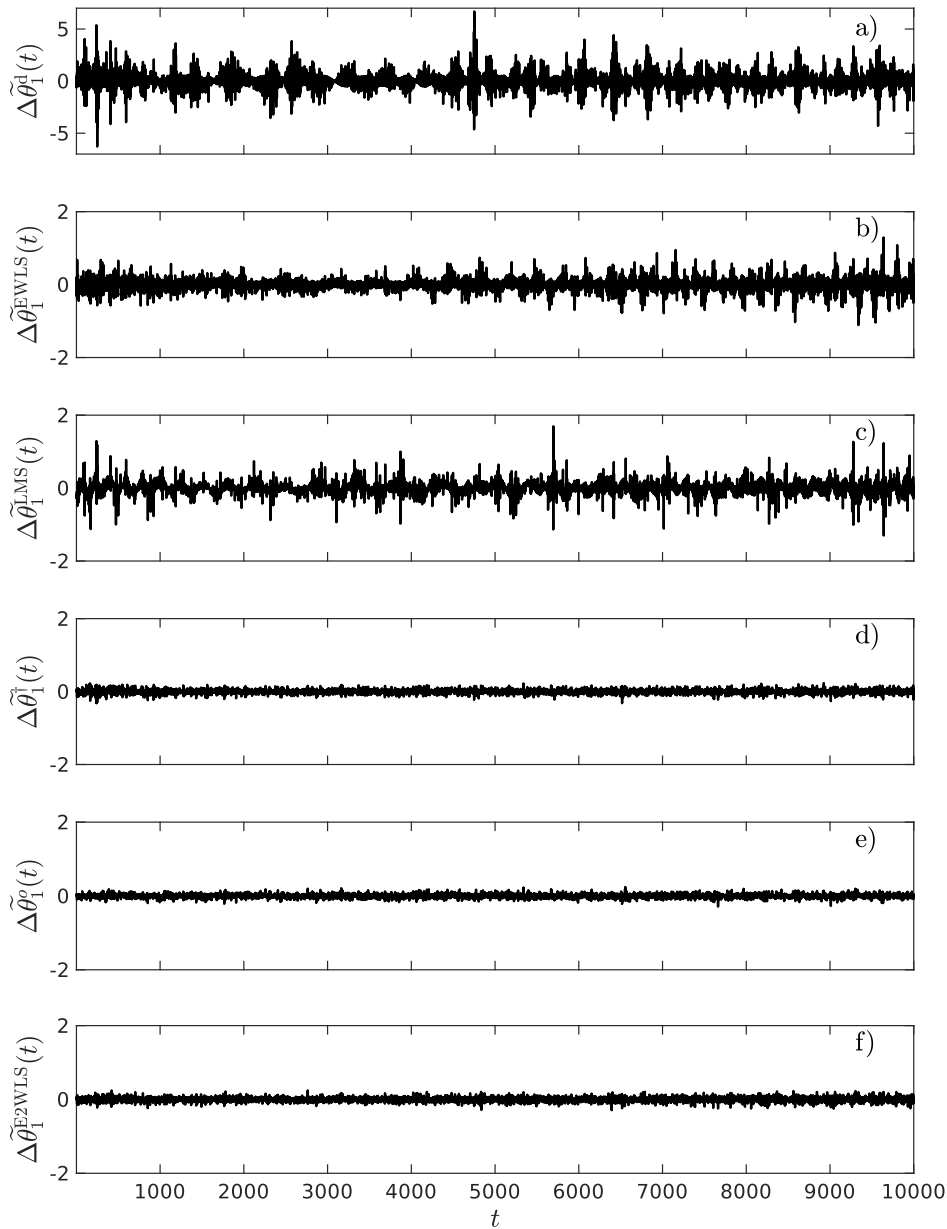
Figure 3.2: Preestimation errors for direct preestimates - plot a), indirect preestimates - plot b), LMS-based preestimates - plot c), enhanced preestimates - plot d), oracle preestimates - plot e) and bidirectional E^2WLS-based preestimates - plot f). Note that the scale of the first plot differs from the scales of other plots to avoid clipping.

3.6 Delay cancellation

Although the initial analysis suggests that the EWLS-based preestimates are unbiased, they result from a purely causal estimation scheme. The EWLS algorithm uses only data from the past, which means that EWLS estimates will lag behind the true parameter trajectories. Assume that the system parameters are lowpass signals. According to [66], for slowly varying parameters the mean path of the EWLS estimates can be, to some extent, regarded as a delayed version of the true parameter trajectory

$$\mathrm{E}[\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t)] \cong \boldsymbol{\theta}(t - \Delta_f), \tag{3.84}$$

where $\Delta_f = \mathrm{int}\left[\frac{\lambda}{1-\lambda}\right]$ is a nominal (low-frequency) delay of filter (3.58) associated with the EWLS estimator and $\mathrm{int}(x)$ is an integer closest to x. Detailed numerical analysis shows that the EWLS-based preestimates tend to inherit the delay associated with the EWLS-based estimates in a nontrivial way. However, simulations suggest that this delay becomes a significant factor when a number of parameters n grows and when the forgetting constant λ approaches 1. Therefore, in this section, two methods that allow one to cope with the delay, are presented. In the next section we present modifications in preestimation scheme, that are designed to avoid the delay.

3.6.1 Fixed-delay shift

The first and the simplest possibility is based on the formula (3.84) and boils down to shifting preestimates (or final fLBF estimates) using Δ_f. It is worth mentioning that [66] offers an alternative method of evaluating the "average" delay of the EWLS estimates. It is a time-domain approach in which one chooses the delay that minimizes the following criterion

$$\Delta_t = \arg\min_{\Delta} \mathrm{E}\{||\overline{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) - \boldsymbol{\theta}(t - \Delta)||^2\}, \tag{3.85}$$

where $\overline{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) = \mathrm{E}[\widehat{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t)]$.

It was shown in [66] that for sufficiently large values of t and assuming that $\boldsymbol{\theta}(t)$ is a random walk process, Δ_t is a solution to the following equation

$$\sum_{i=0}^{\Delta_t} (1 - \lambda)\lambda^i \cong 0.5. \tag{3.86}$$

After straightforward, yet tedious calculations, one can find the exact formula for Δ_t, however, the result is more instructive when one uses the following integral approximation

$$\sum_{i=0}^{\Delta_t} (1 - \lambda)\lambda^i \cong \int_0^{\Delta_t} \gamma e^{-\gamma\tau} d\tau, \quad \gamma = -\log\lambda \cong 1 - \lambda. \tag{3.87}$$

This results in

$$\int_0^{\Delta_t} \gamma e^{-\gamma\tau} d\tau = 0.5 \iff e^{-\gamma\tau}|_{\Delta_t}^0 = 0.5 \iff \Delta_t = -\frac{\log 2}{\log \lambda} \cong \frac{\log 2}{1 - \lambda}. \tag{3.88}$$

The practically usefull formula is

$$\Delta_t \cong \mathrm{int}\left[\frac{\log 2}{1 - \lambda}\right]. \tag{3.89}$$

According to the performed computer simulations, there is no clear benefit from using either of the two formulas for the delay. In some situations, using Δ_f provided better results, while in others, Δ_t turned out to be better suited than Δ_f. Regardless of the formula used to compute the delay, the fixed-delay shift turns out to provide worse results than the adaptive procedure presented in the next paragraph.

3.6.2   Adaptive shift

Instead of using the fixed delay to shift preestimates, one can use a more flexible approach. Final, approximately debiased estimates can be obtained using the following formula

$$\widehat{\boldsymbol{\theta}}^{\mathrm{dfLBF}}(t) = \widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t + d(t)), \tag{3.90}$$

where

$$d(t) = \arg\min_{d \in \mathcal{D}} \sum_{i=-L_0}^{L_0} |y(t+i) - [\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t+d+i)]^{\mathrm{H}}\boldsymbol{\varphi}(t+i)|^2, \tag{3.91}$$

and $\mathcal{D} = [\Delta_f - d_0, \Delta_f + d_0]$ is a set of all candidate delays, centered around the nominal delay Δ_f. In this approach, at each time instant, one chooses the delay that minimizes the local sum of interpolation errors. Such an adaptive method will obviously provide better results than a fixed-delay shift.

3.7   Bidirectional preestimates

One can avoid introducing delay by using noncausal algorithm at the preestimation stage. We will start by defining the steady-state bidirectional exponentially weighted least squares ($\mathrm{E}^2\mathrm{WLS}$) estimates as

$$\widehat{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}}(t) = \arg\min_{\boldsymbol{\theta}} \sum_{i=-\infty}^{\infty} \lambda^{|t-i|}|y(i) - \boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\varphi}(i)|^2, \tag{3.92}$$

where $\lambda \in (0,1)$ is the forgetting constant. It is easy to check that the explicit formula for the estimates is given by

$$\widehat{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}} = [\mathbf{R}_2(t)]^{-1}\mathbf{r}_2(t)(t), \tag{3.93}$$

where

$$\begin{aligned}
\mathbf{R}_2(t) &= \sum_{i=-\infty}^{\infty} \lambda^{|t-i|}\boldsymbol{\varphi}(i)\boldsymbol{\varphi}(i)^{\mathrm{H}} \\
\mathbf{r}_2(t) &= \sum_{i=-\infty}^{\infty} \lambda^{|t-i|}\boldsymbol{\varphi}(i)y^*(i).
\end{aligned} \tag{3.94}$$

The effective memory of this steady-state estimator takes the form

$$L_\infty = \frac{1+\lambda}{1-\lambda}. \tag{3.95}$$

We will show that the mean paths of estimated trajectories are approximately delay-free. For λ sufficiently close to 1, one can use the following approximation

$$\mathbf{R}_2(t) \cong \mathrm{E}[\mathbf{R}_2(t)] = L_\infty \boldsymbol{\Phi}, \tag{3.96}$$

which leads to

$$\begin{aligned}
\overline{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}}(t) = \mathrm{E}[\widehat{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}}(t)] = \mathrm{E}[\mathbf{R}_2^{-1}(t)\mathbf{r}_2(t)] &\cong \frac{1}{L_\infty}\boldsymbol{\Phi}^{-1} \sum_{i=-\infty}^{\infty} \lambda^{|t-i|}\mathrm{E}[\boldsymbol{\varphi}(i)y^*(i)] \\
&= \frac{1}{L_\infty}\boldsymbol{\Phi}^{-1} \sum_{i=-\infty}^{\infty} \lambda^{|t-i|}\{\mathrm{E}[\boldsymbol{\varphi}(i)\boldsymbol{\varphi}^{\mathrm{H}}(i)]\boldsymbol{\theta}(i) + \mathrm{E}[\boldsymbol{\varphi}(i)e^*(i)]\} = \frac{1}{L_\infty} \sum_{i=-\infty}^{\infty} \lambda^{|t-i|}\boldsymbol{\theta}(i) \\
&= \frac{1}{L_\infty} \sum_{i=-\infty}^{\infty} \lambda^{|i|}\boldsymbol{\theta}(t-i).
\end{aligned} \tag{3.97}$$

The impulse response of a filter associated with this estimator is given by

$$h^{\mathrm{E}^2\mathrm{WLS}}(i) = \frac{1}{L_\infty}\lambda^{|i|}, \quad i \in \mathbb{Z}. \tag{3.98}$$

The freuquency characteristic of this filter is defined as

$$H^{\mathrm{E^2WLS}}(\omega) = \frac{1}{L_\infty} \sum_{i=\infty}^{\infty} \lambda^{|i|} e^{-\mathrm{i}\omega i} = \frac{1}{L_\infty} \left[\sum_{i=-\infty}^{0} \lambda^{-i} e^{-\mathrm{i}\omega i} + \sum_{i=1}^{\infty} \lambda^i e^{-\mathrm{i}\omega i} \right]$$

$$= \frac{1}{L_\infty} \left[\frac{1}{1-\lambda e^{\mathrm{i}\omega}} + \frac{\lambda}{1-\lambda e^{-\mathrm{i}\omega}} \right] = \frac{1-\lambda}{1+\lambda} \frac{1-\lambda^2}{1-2\lambda\cos\omega+\lambda^2} = \frac{(1-\lambda)^2}{1-2\lambda\cos\omega+\lambda^2}.$$

$$(3.99)$$

Note that the frequency characteristic is real-valued, which means that this filter does not introduce any delay.

Preestimates, based on the E²WLS estimates can be defined as follows

$$\widetilde{\boldsymbol{\theta}}^{\mathrm{E^2WLS}}(t) = \frac{(1+\lambda^2)\widehat{\boldsymbol{\theta}}^{\mathrm{E^2WLS}}(t) - \lambda\widehat{\boldsymbol{\theta}}^{\mathrm{E^2WLS}}(t-1) - \lambda\widehat{\boldsymbol{\theta}}^{\mathrm{E^2WLS}}(t+1)}{(1-\lambda)^2}. \qquad (3.100)$$

Using the approximation (3.114), one obtains

$$\widetilde{\boldsymbol{\theta}}^{\mathrm{E^2WLS}}(t) \cong \frac{1}{1-\lambda^2} \boldsymbol{\Phi}^{-1}[\mathbf{r}_2(t) + \lambda^2\mathbf{r}_2(t) - \lambda\mathbf{r}_2(t-1) - \lambda\mathbf{r}_2(t+1)]. \qquad (3.101)$$

Note that

$$\mathbf{r}_2(t) = \sum_{i=-\infty}^{t} \lambda^{t-i}\boldsymbol{\varphi}(i)y^*(i) + \sum_{i=t+1}^{\infty} \lambda^{i-t}\boldsymbol{\varphi}(i)y^*(i)$$

$$\lambda^2\mathbf{r}_2(t) = \sum_{i=-\infty}^{t} \lambda^{t-i+2}\boldsymbol{\varphi}(i)y^*(i) + \sum_{i=t+1}^{\infty} \lambda^{i-t+2}\boldsymbol{\varphi}(i)y^*(i)$$

$$\lambda\mathbf{r}_2(t-1) = \sum_{i=-\infty}^{t} \lambda^{t-i}\boldsymbol{\varphi}(i)y^*(i) + \sum_{i=t+1}^{\infty} \lambda^{i-t+2}\boldsymbol{\varphi}(i)y^*(i) - \boldsymbol{\varphi}(t)y^*(t) + \lambda^2\boldsymbol{\varphi}(t)y^*(t)$$

$$\lambda\mathbf{r}_2(t+1) = \sum_{i=-\infty}^{t} \lambda^{t-i+2}\boldsymbol{\varphi}(i)y^*(i) + \sum_{i=t+1}^{\infty} \lambda^{i-t}\boldsymbol{\varphi}(i)y^*(i),$$

hence

$$\widetilde{\boldsymbol{\theta}}^{\mathrm{E^2WLS}}(t) \cong \boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t)y^*(t) = \widetilde{\boldsymbol{\theta}}(t), \qquad (3.102)$$

and the bidirectional preestimates are approximately unbiased.

Unfortunately, implementation of such estimator is impossible in practice, due to required infinite amount of data. One can only implement one of the finte-time versions of this algorithm, that are approximately delay-free.

3.7.1 Offline implementation

When one has access to the entire segment of prerecorded data[3], the fixed-interval implementation can be used (this method will be called for short offline E²WLS)

$$\widehat{\boldsymbol{\theta}}^{\mathrm{E^2WLS}}(t|T) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{T} \lambda^{|t-i|} |y(i) - \boldsymbol{\theta}^{\mathrm{H}}\boldsymbol{\varphi}(i)|^2. \qquad (3.103)$$

Similar as before

$$\widehat{\boldsymbol{\theta}}^{\mathrm{E^2WLS}}(t|T) = [\mathbf{R}_2(t|T)]^{-1}\mathbf{r}_2(t)(t|T), \qquad (3.104)$$

where

$$\mathbf{R}_2(t|T) = \sum_{i=1}^{T} \lambda^{|t-i|}\boldsymbol{\varphi}(i)\boldsymbol{\varphi}(i)^{\mathrm{H}}$$

$$\mathbf{r}_2(t|T) = \sum_{i=1}^{T} \lambda^{|t-i|}\boldsymbol{\varphi}(i)y^*(i). \qquad (3.105)$$

---

[3]Such a situation can occur e.g. when two UAC devices working in an FD mode exchange frames of data instead of communicating in a continuous manner.

Note that the offline E^2WLS estimates can be expressed using unidirectional EWLS algortihms working forward (3.43) and backward in time. The latter is defined as follows

$$\widehat{\boldsymbol{\theta}}_-^{\mathrm{EWLS}}(t) = \arg\min_{\boldsymbol{\theta}} \sum_{i=t}^{T} \lambda^{i-t} |y(i) - \boldsymbol{\theta}^{\mathrm{H}} \boldsymbol{\varphi}(i)|^2, \tag{3.106}$$

$$\widehat{\boldsymbol{\theta}}_-^{\mathrm{EWLS}}(t) = \mathbf{R}_-^{-1}(t)\mathbf{r}_-(t), \tag{3.107}$$

where

$$\begin{aligned}
\mathbf{R}_-(t) &= \sum_{i=t}^{T} \lambda^{i-t} \boldsymbol{\varphi}(i)\boldsymbol{\varphi}(i)^{\mathrm{H}} = \lambda \mathbf{R}_-(t+1) + \boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\mathrm{H}}(t) \\
\mathbf{r}_-(t) &= \sum_{i=1}^{T} \lambda^{i-t} \boldsymbol{\varphi}(i) y^*(i) = \lambda \mathbf{r}_-(t+1) + \boldsymbol{\varphi}(t) y^*(t),
\end{aligned} \tag{3.108}$$

which is initialized with $\mathbf{R}_-(T+1) = 0$ and $\mathbf{r}_-(T+1) = 0$.

The effective memory of this estimator is given by

$$L_-(t) = \sum_{i=t}^{T} \lambda^{i-t} = \lambda L_-(t+1) + 1 = \frac{1 - \lambda^{T-t+1}}{1 - \lambda}, \tag{3.109}$$

with the initial condition $L_-(T+1) = 0$.

Combining forward and backward-time EWLS estimators one obtains

$$\begin{aligned}
\mathbf{R}_2(t|T) &= \mathbf{R}(t) + \lambda \mathbf{R}_-(t+1) \\
\mathbf{r}_2(t|T) &= \mathbf{r}(t) + \lambda \mathbf{r}_-(t+1),
\end{aligned} \tag{3.110}$$

which leads to a simple recursive expression, that boils down to processing the causal estimator in a backward time

$$\begin{aligned}
\mathbf{R}_2(t|T) &= \lambda \mathbf{R}_2(t+1|T) + (1 - \lambda^2)\mathbf{R}(t) \\
\mathbf{r}_2(t|T) &= \lambda \mathbf{r}_2(t+1|T) + (1 - \lambda^2)\mathbf{r}(t),
\end{aligned} \tag{3.111}$$

with the initial conditions being $\mathbf{R}_2(T+1|T) = \lambda \mathbf{R}(T)$ and $\mathbf{r}_2(T+1|T) = \lambda \mathbf{r}(T)$. Obviously, the analogous expression can be coined for the forward-time regression using backward-time EWLS estimator.

Similarly, the effective memory of the offline bidirectional estimator can be written down as

$$L_2(t|T) = \sum_{i=1}^{T} \lambda^{|t-i|} L(t) + \lambda L_-(t+1) = \lambda L_2(t+1|T) + (1 - \lambda^2) L(t), \tag{3.112}$$

where $L_2(T+1|T) = \lambda L(T)$.

Bidirectional offline preestimates will be defined as follows

$$\begin{aligned}
\widetilde{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}}(t|T) = \frac{1}{1 - \lambda^2} &\left[(1 + \lambda^2) L_2(t|T)\widehat{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}}(t|T) - \lambda L_2(t-1|T)\widehat{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}}(t-1|T) - \right. \\
&\left. \lambda L_2(t+1|T)\widehat{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}}(t+1|T) \right].
\end{aligned} \tag{3.113}$$

It turns out that the preestimates implemented like this have approximately the same properties as preestimates (3.100), when $t \gg 1$ and $t \ll T$. In such situation, one can use the following approximation

$$\mathbf{R}_2(t|T) \cong \mathrm{E}[\mathbf{R}_2(t|T)] = L_2(t|T)\boldsymbol{\Phi}, \tag{3.114}$$

which leads to

$$L_2(t|T)\widehat{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}}(t|t) - \lambda L_2(t+1|T)\widehat{\boldsymbol{\theta}}^{\mathrm{E}^2\mathrm{WLS}}(t+1|T) \cong \boldsymbol{\Phi}^{-1}[\mathbf{r}_2(t|T) - \lambda \mathbf{r}_2(t+1|T)] = (1 - \lambda^2)\boldsymbol{\Phi}^{-1}\mathbf{r}(t). \tag{3.115}$$
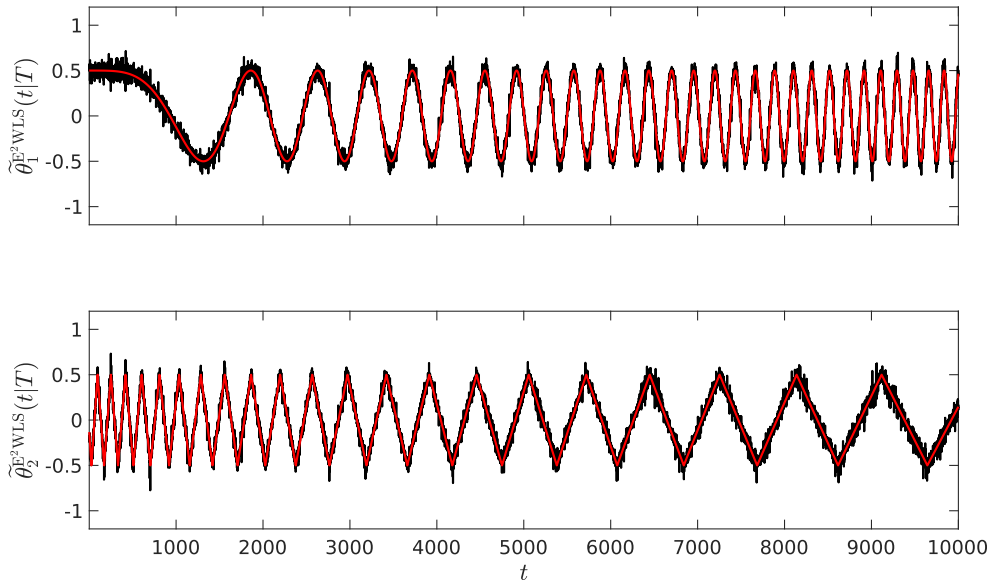
Figure 3.3: Offline E^2WLS preestimates (black lines) superimposed on true parameter trajectories (red lines).

Note that the formula (3.2) can be expressed as

$$\widetilde{\boldsymbol{\theta}}(t) = \boldsymbol{\Phi}^{-1}\boldsymbol{\varphi}(t)y^*(t) = \boldsymbol{\Phi}^{-1}[\mathbf{r}(t) - \lambda\mathbf{r}(t-1)]. \tag{3.116}$$

Combining (3.115) with (3.116) one obtains that

$$\widetilde{\boldsymbol{\theta}}^{E^2WLS}(t|T) \cong \boldsymbol{\Phi}^{-1}[\mathbf{r}(t) - \lambda\mathbf{r}(t-1)] = \widetilde{\boldsymbol{\theta}}(t), \tag{3.117}$$

which means that the offline bidirectional preestimates are approximately unbiased. An example of such preestimates is shown in figure 3.3.

3.7.2 Fixed-delay implementation

The second way of implementing the E^2WLS preestimates is by using the fixed-delay algorithms. These algorithms allow one to work in almost-real time. The definitions are very similar to these from the previous section, the ddifference is that T should be replaced by $t + t_0$ in formulae (3.103) - (3.109). The obvious drawback of the truncation is that it makes all the recursions more complicated. Note, however, that the recursive approach from (3.108) is still possible but in a slightly different manner, resembling the sliding-window-type recursions. In fact, even the inverse of the matrix from (3.108) can be still computed recursively despite its truncated form. As mentioned previously, the formulation of the truncated E^2WLS estimator is very similar. Nonetheless, it is worth explaining the difference between truncated and the normal backward-time EWLS estimator (3.106). The truncated anticausal algorithm is defined as follows

$$\widehat{\boldsymbol{\theta}}_-^{tEWLS}(t) = \arg\min_{\boldsymbol{\theta}} \sum_{i=t}^{t+t_0} \lambda^{i-t}|y(i) - \boldsymbol{\theta}^{H}\boldsymbol{\varphi}(i)|^2 = [\mathbf{R}_-^t(t)]^{-1}\mathbf{r}_-^t(t), \tag{3.118}$$

where

$$\begin{aligned} \mathbf{R}_-^t(t) &= \sum_{i=t}^{t+t_0} \lambda^{i-t}\boldsymbol{\varphi}(i)\boldsymbol{\varphi}(i)^{H} \\ \mathbf{r}_-^t(t) &= \sum_{i=t}^{t+t_0} \lambda^{i-t}\boldsymbol{\varphi}(i)y^*(i). \end{aligned} \tag{3.119}$$

This time, the recurssion from (3.108) does not hold, but instead we have that

$$
\mathbf{R}_-^t(t) = \frac{1}{\lambda}[\mathbf{R}_-^t(t-1) - \boldsymbol{\varphi}(t-1)\boldsymbol{\varphi}^{\mathrm{H}}(t-1)] + \lambda^{t_0}\boldsymbol{\varphi}(t+t_0)\boldsymbol{\varphi}^{\mathrm{H}}(t+t_0)
$$
$$
\mathbf{r}_-^t(t) = \frac{1}{\lambda}[\mathbf{r}_-^t(t-1) - \boldsymbol{\varphi}(t-1)y^*(t-1)] + \lambda^{t_0}\boldsymbol{\varphi}(t+t_0)y^*(t+t_0),
$$

(3.120)

Using the Sherman-Morrison formula [28] twice, one can compute recursively the inverse of matrix $\mathbf{R}_-^t(t)$. Denote by $\mathbf{P}_-^t(t) = [\mathbf{R}_-^t(t)]^{-1}$, the first step is

$$
\mathbf{P}_-^t(t|t-1) = \lambda[\mathbf{R}_-^t(t-1) - \boldsymbol{\varphi}(t-1)\boldsymbol{\varphi}^{\mathrm{H}}(t-1)]^{-1}
$$
$$
= \lambda\mathbf{P}_-^t(t-1) + \frac{\lambda\mathbf{P}_-^t(t-1)\boldsymbol{\varphi}(t-1)\boldsymbol{\varphi}^{\mathrm{H}}(t-1)\mathbf{P}_-^t(t-1)}{1 - \boldsymbol{\varphi}^{\mathrm{H}}(t-1)\mathbf{P}_-^t(t-1)\boldsymbol{\varphi}(t-1)}.
$$

(3.121)

The second step is

$$
\mathbf{P}_-^t(t) = \mathbf{P}_-^t(t|t-1) - \frac{\lambda^{t_0}\mathbf{P}_-^t(t|t-1)\boldsymbol{\varphi}(t+t_0)\boldsymbol{\varphi}^{\mathrm{H}}(t+t_0)\mathbf{P}_-^t(t-1)}{1 + \lambda^{t_0}\boldsymbol{\varphi}^{\mathrm{H}}(t+t_0)\mathbf{P}_-^t(t|t-1)\boldsymbol{\varphi}(t+t_0)}.
$$

(3.122)

These recursions can be helpful when using the fixed-delay implementation with the simplified preestimates described below.

When t is large and λ^{t_0} is sufficiently small, the fixed-delay implementation has approximately the same properties as the steady-state $\mathrm{E}^2\mathrm{WLS}$ algorithm. However, if the algorithm is meant to work in almost real-time, then one should carefully choose t_0 since it decides upon the amount of the processing delay introduced by the algorithm. The rule of thumb, that works well in practice turns out to be

$$
t_0 \geq \left\lceil \frac{8}{1-\lambda} \right\rceil,
$$

where $\lceil x \rceil$ is an integer closest to x but not smaller than x.

3.7.3   Simplified bidirectional preestimates

The bidirectional preestimates have an obvious drawback, which is their high computational complexity. Unfortunately, there are no fast algorithms for the computation of $\mathrm{E}^2\mathrm{WLS}$ estimates, resembling the fast transversal filters or computations based on the matrix inversion lemma [28]. A simple solution to this problem would be to use an approximation of the bidirectional preestimates. For λ sufficiently close to 1

$$
\mathbf{R}_-(t) \cong \mathrm{E}[\mathbf{R}_-(t)] = L_-(t)\boldsymbol{\Phi}.
$$

(3.123)

Combining this with (3.116), one can define the unidirectional anticausal preestimates as

$$
\widetilde{\boldsymbol{\theta}}_-^{\mathrm{EWLS}}(t) = \frac{L_-(t)\widehat{\boldsymbol{\theta}}_-^{\mathrm{EWLS}}(t) - \lambda L_-(t+1)\widehat{\boldsymbol{\theta}}_-^{\mathrm{EWLS}}(t+1)}{L_-(t)}.
$$

(3.124)

The simplified bidirectional preestimates become

$$
\widetilde{\boldsymbol{\theta}}^{\pm}(t) = \frac{L(t)\widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) + L_-(t)\widetilde{\boldsymbol{\theta}}_-^{\mathrm{EWLS}}(t)}{L(t) + L_-(t)}.
$$

(3.125)

Using the same argument as in the previous sections, one can conclude that both $\widetilde{\boldsymbol{\theta}}_-^{\mathrm{EWLS}}(t)$ and $\widetilde{\boldsymbol{\theta}}^{\pm}(t)$ are approximately unbiased. The steady-state version of such preestimates boils down to

$$
\widetilde{\boldsymbol{\theta}}^{\pm}(t) = \frac{\widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t) + \widetilde{\boldsymbol{\theta}}_-^{\mathrm{EWLS}}(t)}{2}.
$$

(3.126)

It is easy to show that the simplified bidirectional preestimates are approximately unbiased. Furthermore, if one assumes that

$$
\mathrm{E}[\widetilde{\boldsymbol{\theta}}^{\mathrm{EWLS}}(t)] \cong \boldsymbol{\theta}(t - \Delta_f)
$$
$$
\mathrm{E}[\widetilde{\boldsymbol{\theta}}_-^{\mathrm{EWLS}}(t)] \cong \boldsymbol{\theta}(t + \Delta_f),
$$

(3.127)

then it can be showed that these preestimates are delay-free. Assume that the discrete-time parameter trajecory $\{\boldsymbol{\theta}(t), t \in \mathbb{Z}\}$ is a result of sampling a smooth continous-time trajectory $\{\boldsymbol{\theta}(\tau), \tau \in \mathbb{R}\}$. Then, using the first-order Taylor approximation $\boldsymbol{\theta}(\tau + \tau_0) \cong \boldsymbol{\theta}(\tau) + \tau_0 \boldsymbol{\theta}'(\tau)$, one arrives at

$$\begin{aligned} \boldsymbol{\theta}(t - \Delta_f) &\cong \boldsymbol{\theta}(t) - \Delta_f \boldsymbol{\theta}'(t) \\ \boldsymbol{\theta}(t + \Delta_f) &\cong \boldsymbol{\theta}(t) + \Delta_f \boldsymbol{\theta}'(t), \end{aligned} \tag{3.128}$$

which means that

$$\mathrm{E}[\widetilde{\boldsymbol{\theta}}^{\pm}(t)] \cong \boldsymbol{\theta}(t), \tag{3.129}$$

the steady-state simplified preestimates are approximately delay-free.

Note that the simplified scheme can be used with both fixed-interval and fixed-delay implementations.

3.8 MSE analysis and optimal impulse response

In this section, we again adopt the assumption (A2.7), under which the MSE of the fLBF estimator can be found. First, define the estimation errors

$$\Delta \widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t) = \widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t) - \boldsymbol{\theta}(t) = \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i) \boldsymbol{\theta}(t + i) + \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i) \mathbf{z}(t + i) - \boldsymbol{\theta}(t). \tag{3.130}$$

The MSE is equal to

$$\begin{aligned} M_{m|k}^{\mathrm{fLBF}} = \mathrm{E}[||\Delta \widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t)||^2] &= \mathrm{E}[||\boldsymbol{\Theta}(t)\mathbf{h}_{m|k} - \boldsymbol{\theta}(t)||^2] + \mathrm{E}\{[\boldsymbol{\Theta}(t)\mathbf{h}_{m|k} - \boldsymbol{\theta}(t)]^{\mathrm{H}} \mathbf{Z}(t)\mathbf{h}_{m|k}\} \\ &+ \mathrm{E}\{\mathbf{h}_{m|k}^{\mathrm{H}} \mathbf{Z}^{\mathrm{H}}(t)[\boldsymbol{\Theta}(t)\mathbf{h}_{m|k} - \boldsymbol{\theta}(t)]\} + \mathrm{E}[||\mathbf{Z}(t)\mathbf{h}_{m|k}||^2], \end{aligned} \tag{3.131}$$

where the expectation is carried over different realizations of parameter trajectories, and $\mathbf{h}_{m|k} = [h_{m|k}^{\mathrm{fLBF}}(-k), \ldots, h_{m|k}^{\mathrm{fLBF}}(k)]^{\mathrm{T}}$, $\boldsymbol{\Theta}(t) = [\boldsymbol{\theta}(-k), \ldots, \boldsymbol{\theta}(t + k)]$, $\mathbf{Z}(t) = [\mathbf{z}(t - k), \ldots, \mathbf{z}(t + k)]$.

Under (A2.4), (A2.5) and (A2.6) the preestimation noise $\mathbf{z}(t)$ is approximately white and uncorrelated with $\boldsymbol{\theta}(t)$, hence $\mathrm{E}\{[\boldsymbol{\Theta}(t)\mathbf{h}_{m|k} - \boldsymbol{\theta}(t)]^{\mathrm{H}} \mathbf{Z}(t)\mathbf{h}_{m|k}\} = \mathrm{E}\{\mathbf{h}_{m|k}^{\mathrm{H}} \mathbf{Z}^{\mathrm{H}}(t)[\boldsymbol{\Theta}(t)\mathbf{h}_{m|k} - \boldsymbol{\theta}(t)]\} = 0$.

Furthermore, under (A2.7)

$$\mathrm{E}[||\boldsymbol{\Theta}(t)\mathbf{h}_{m|k} - \boldsymbol{\theta}(t)||^2] = \eta[r_\theta(0) - \mathbf{r}_\theta^{\mathrm{H}} \mathbf{h}_{m|k}^* + \mathbf{h}_{m|k}^{\mathrm{H}} \mathbf{R}_\theta \mathbf{h}_{m|k}], \tag{3.132}$$

where $\mathbf{r}_\theta$ and $\mathbf{R}_\theta$ are defined exactly the same as in Section 2.6.4. In fact, the expression above is the same as the formula (2.86).

We will also make an additional assumption, that is commonly fulfilled in telecommunication applications.

> (A3.2) $\{u(t)\}$ is a sequence of mutually uncorrelated circular random variables with variance σ_u^2.

Under (A3.2), it holds that

$$\mathrm{cov}[\mathbf{z}(t)] \cong \sigma_z^2 \mathbf{I}_n, \tag{3.133}$$

and

$$\mathrm{E}[||\mathbf{Z}(t)\mathbf{h}_{m|k}||^2] \cong \sigma_z^2 \mathbf{h}_{m|k}^{\mathrm{H}} \mathbf{h}_{m|k} \mathbf{I}_K. \tag{3.134}$$

Summarizing,

$$M_{m|k}^{\mathrm{fLBF}} \cong \eta \left[r_\theta(0) - \mathbf{r}_\theta^{\mathrm{H}} \mathbf{h}_{m|k}^* + \mathbf{h}_{m|k}^{\mathrm{H}} \left(\mathbf{R}_\theta + \frac{n\sigma_z^2}{\eta} \mathbf{I}_K \right) \mathbf{h}_{m|k} \right], \tag{3.135}$$

which is very similar to (2.88). The difference is that in this case, one can actually find the optimal impulse response $\mathbf{h}_{m|k}^{\mathrm{opt}}$ and apply it in the fLBF method. A similar approach for the WLS method was presented in [48]. One can take one of the two approaches here. The first would be to perform unconstrained optimization, leading to

$$\mathbf{h}_k^{\mathrm{opt}} = \left(\mathbf{R}_\theta + \frac{n\sigma_z^2}{\eta} \mathbf{I}_K \right)^{-1} \mathbf{r}_\theta^*. \tag{3.136}$$

The other way would be to adopt a constraint $\sum_{i=-k}^{k} h_{m|k}^{\text{fLBF}}(i) = 1 \iff \mathbf{c}^{\text{T}}\mathbf{h}_{m|k} = 1$, $\mathbf{c} = [1, \ldots, 1]^{\text{T}}$. The impulse response of a suboptimal lowpass filter can be then found using the method of Lagrange multipliers (see Appendix D in [35]). Its impulse response would be defined as

$$\mathbf{h}_k^{\text{sopt}} = \left(\mathbf{R}_\theta + \frac{n\sigma_z^2}{\eta}\mathbf{I}_K \right)^{-1} (\mathbf{r}_\theta^* + \gamma), \tag{3.137}$$

where

$$\gamma = \frac{1 - \mathbf{c}^{\text{T}} \left(\mathbf{R}_\theta + \frac{n\sigma_z^2}{\eta}\mathbf{I}_K \right)^{-1} \mathbf{r}_\theta^*}{\mathbf{c}^{\text{T}} \left(\mathbf{R}_\theta + \frac{n\sigma_z^2}{\eta}\mathbf{I}_K \right)^{-1} \mathbf{c}}. \tag{3.138}$$

Remark 1 - optimal basis functions

It turns out that the optimal impulse response can be easily connected with the basis functions incorporated in the Savitzky-Golay filter, when $w_k(i) \equiv 1$, $i \in I_k$. Note that the optimal impulse response can be expressed in the form

$$\mathbf{h}_k^{\text{opt}} = \left(\mathbf{R}_\theta + \frac{n\sigma_z^2}{\eta}\mathbf{I}_K \right)^{-1} \mathbf{r}_\theta^* = \mathbf{Q}\boldsymbol{\Upsilon}\mathbf{Q}^{\text{H}}\mathbf{1}_{K,k+1}, \tag{3.139}$$

where $\boldsymbol{\Upsilon} = \text{diag}\left\{ \frac{\eta\lambda_1}{\eta\lambda_1 + n\sigma_z^2}, \ldots, \frac{\eta\lambda_K}{\eta\lambda_K + n\sigma_z^2} \right\}$, and $\mathbf{Q}$ is a matrix of eigenvectors of $\mathbf{R}_\theta$, while λ_i, $i \in [1, K]$ are eigenvalues of the matrix $\mathbf{R}_\theta$.

Hence, the optimal basis should be defined as

$$[f_{l|K}(-k), \ldots, f_{l|K}(k)]^{\text{T}} = \sqrt{\frac{\lambda_l\eta}{\lambda_l\eta + n\sigma_z^2}}\mathbf{q}_l, \; l = 1, \ldots, K, \tag{3.140}$$

to produce the filter with the same impulse response as the optimal one.

Note also that in the noiseless scenario these functions reduce to KL functions. They are also almost indistinguishable from the KL functions when $\lambda_l \gg \sigma_z^2/\eta$. Another important thing is that in the fLBF method, one can use K orthonormal basis functions without any numerical problems.

Figure 3.4 shows the shapes of the optimal (the top plot) and suboptimal impulse response (in the middle plot), and the impulse response resulting from the KL functions (the bottom plot). These plots were obtained assuming uniform power spectral density function of parameter changes (2.84), with $\zeta_1 = \zeta_2 = \ldots = \zeta_{20} = 1$, normalized cut-off frequency $\omega_0 = \pi/50$ and $\sigma_z^2 = 2$.

The differences between the first two plots are very small. In fact, for this particular example $\mathbf{c}^{\text{T}}\mathbf{h}_k^{\text{opt}} \cong 0.993$. Note that

$$\mathbf{h}_k^{\text{sopt}} = \mathbf{h}_k^{\text{opt}} + \gamma \left(\mathbf{R}_\theta + \frac{n\sigma_z^2}{\eta}\mathbf{I}_K \right)^{-1} \mathbf{c}. \tag{3.141}$$

The suboptimal window is an effect of modifying the optimal one to make sure that the constraint $\mathbf{c}^{\text{T}}\mathbf{h}_{m|k} = 1$ is satisfied. In this example, the modification needs to be very small because the constraint is almost met. More differences are visible between the first and the last plot (especially at both ends of the analysis window), which is due to the fact that in this example $\sigma_z^2 > 0$ and hence, for some eigenvectors $\frac{\sigma_z^2}{n\lambda_l}$ is substantially greater than 0.

Remark 2 - decoupled optimization

Note that the MSE for the entire system is the sum of the MSEs for every parameter. Since, using the fLBF method, one can decouple the estimation problem, one can also relax the assumption (A2.7) and assume that every parameter is characterized by a different autocorrelation function. Then, one can find the impulse response of the optimal filter for each parameter, and filter each preestimated trajectory differently.

Remark 3 - regularization

Note that impulse response (3.136) can be interpreted as the impulse response associated with the KL basis after applying regularization. The higher the variance of the preestimation noise compared to the variance of the parameter trajectories, the stronger the regularization effect.
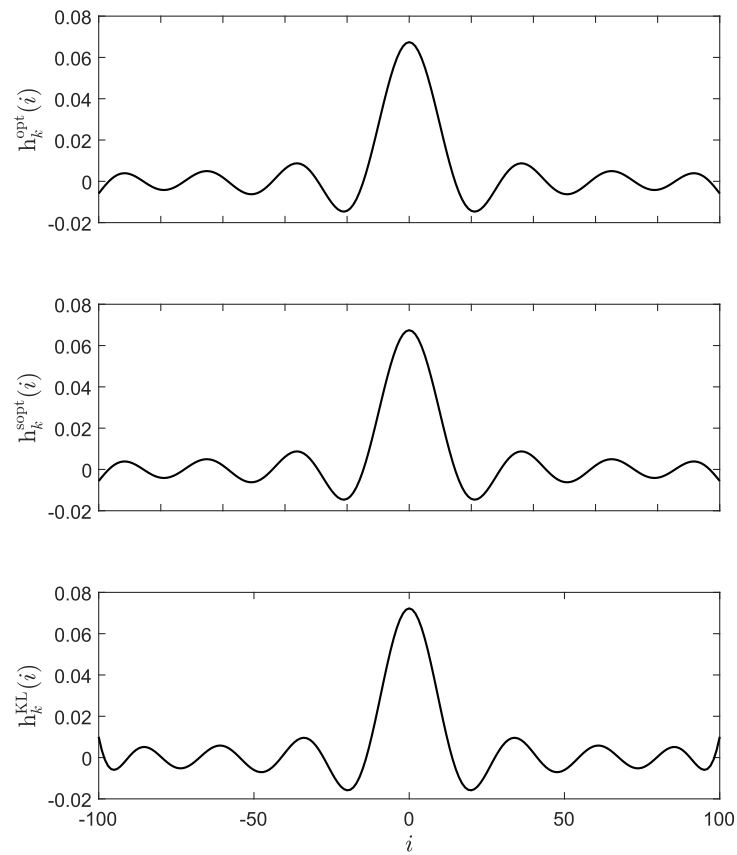
Figure 3.4: Shapes of the optimal (the top plot) and the suboptimal impulse response (the middle plot), and the impulse response resulting from the KL functions (the bottom plot). The plots were obtained assuming uniform power spectral density function of parameter changes (2.84), with $\zeta_1 = \zeta_2 = \ldots = \zeta_{20} = 1$, normalized cut-off frequency $\omega_0 = \pi/50$ and $\sigma_z^2 = 2$.

Remark 4 - delay

Note that the debiasing, either using the fixed delay or with the adaptive choice of the local delay of preestimates, can be incorporated into the design of an optimal (or suboptimal) filter. For instance, assuming that the average delay of preestimates is known and equal to Δ_f, it is easy to check that the formula for the optimal filter requires replacing $\mathbf{r}_\theta$ with a new, shifted one

$$\widetilde{\mathbf{r}}_\theta = [r_\theta(-k - \Delta_f), \dots, r_\theta(k - \Delta_f)]^\mathrm{T}. \tag{3.142}$$

3.9   Hyperparameter optimization in postfiltering

When using Savitzky-Golay filtration, one faces the same dilemma as in the LBF method, namely, how to choose the number of basis functions m, and the length of analysis window K? It turns out that the same approach, based on parallel estimation can be applied in the fLBF method. Just as previously, the FPE and LOOCV methods can be adapted to determine which settings are locally best suited for identification.

The fLBF method is a unique identification technique, which allows one to decouple the estimation of system parameters. This means, that the quality measure for choosing hyperparameters can be either based on modeling errors related to the system output, just like in the LBF approach, or one can use several quality measures, based on errors related to filtering preestimates. The first strategy is called centralized while the second one is called decentralized.

While tuning hyperparameters, one can make the decision for each parameter separately (resulting in potentially different m and K for every system parameter) or jointly. In the second approach, one chooses the same hyperparameters for every parameter trajectory. Both of these approaches can be implemented in centralized and decentralized strategies. However, choosing hyperparameters separately for every preestimate is more natural and implementational easier in a decentralized approach.

3.9.1   Centralized approach

Whether the hyperparameters are selected together or separately for each parameter, one has to run several estimators $\widehat{\boldsymbol{\theta}}_j^{\mathrm{fLBF}}(t)$, $j \in \mathcal{MK}$ in parallel, where $\mathcal{MK}$ is a set of all different settings. At each time instant, one chooses the one that minimizes the centralized quality measure

$$\widehat{\boldsymbol{\theta}}^{\mathrm{fLBF}}(t) = \widehat{\boldsymbol{\theta}}_{j(t)}^{\mathrm{fLBF}}(t), \tag{3.143}$$

where

$$j(t) = \arg\min_{j \in \mathcal{MK}} J_c(t). \tag{3.144}$$

The quality measure is based on the leave-one-out cross-validation approach, namely

$$J_c^{\mathrm{fLBF}}(t) = \sum_{i=-L}^{L} \left| \varepsilon_{o,m|k}^{\mathrm{fLBF}}(t+i) \right|^2, \tag{3.145}$$

where

$$\varepsilon_{o,m|k}^{\mathrm{fLBF}}(t) = y(t) - [\widehat{\boldsymbol{\theta}}_{o,m|k}^{\mathrm{fLBF}}(t)]^\mathrm{H} \boldsymbol{\varphi}(t), \tag{3.146}$$

denotes leave-one-out interpolation error, computed for the holey estimates of system trajectories

$$\widehat{\boldsymbol{\theta}}_{o,m|k}^{\mathrm{fLBF}}(t) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\mathrm{fLBF}}(t), \tag{3.147}$$

and

$$\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\mathrm{fLBF}}(t) = \arg\min_{\boldsymbol{\alpha}} \sum_{\substack{i=-k \\ i \neq 0}}^{k} w_k(i) ||\widetilde{\boldsymbol{\theta}}(t) - \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}||^2. \tag{3.148}$$

It is straightforward to check that

$$
\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\text{fLBF}}(t) = \left[\sum_{\substack{i=-k \\ i\neq 0}}^{k} w_k(i)\mathbf{F}_{m|k}^{\text{H}}(i)\mathbf{F}_{m|k}(i) \right]^{-1} \left[\sum_{\substack{i=-k \\ i\neq 0}}^{k} w_k(i)\widetilde{\boldsymbol{\theta}}(t+i) \otimes \mathbf{f}_{m|k}(i) \right]
$$

$$
= [\mathbf{I}_{nm} - w_k(0)\mathbf{F}_{m|k}^{\text{H}}(0)\mathbf{F}_{m|k}(0)]^{-1}[\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t) - w_k(0)\widetilde{\boldsymbol{\theta}}(t) \otimes \mathbf{f}_{mk}(0)].
$$

$$(3.149)$$

Since it was assumed that $w_k(0) = 1$, using the Sherman-Morrison formula [28] and the properties of Kronecker product, one obtains

$$
\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\text{fLBF}}(t) = \mathbf{I}_n \otimes [\mathbf{I}_m - \mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\text{H}}(0)]^{-1}[\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t) - \widetilde{\boldsymbol{\theta}}(t) \otimes \mathbf{f}_{mk}(0)]
$$

$$
= \mathbf{I}_n \otimes \left[\mathbf{I}_m + \frac{\mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\text{H}}(0)}{1 - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)} \right] [\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t) - \widetilde{\boldsymbol{\theta}}(t) \otimes \mathbf{f}_{mk}(0)]
$$

$$
= \mathbf{I}_n \otimes \left[\mathbf{I}_m + \frac{\mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\text{H}}(0)}{1 - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)} \right] \widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t) - \frac{1}{1 - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)}\widetilde{\boldsymbol{\theta}}(t) \otimes \mathbf{f}_{m|k}(0).
$$

$$(3.150)$$

Combining the result above with (3.147), one gets

$$
\widehat{\boldsymbol{\theta}}_{o,m|k}^{\text{fLBF}}(t) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\text{fLBF}}(t) = [\mathbf{I}_n \otimes \mathbf{f}_{m|k}^{\text{H}}(0)]\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\text{fLBF}}(t)
$$

$$
= \mathbf{I}_n \otimes \left[\mathbf{f}_{m|k}^{\text{H}}(0) + \frac{\mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\text{H}}(0)}{1 - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)} \right] \widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t)
$$

$$
- \frac{1}{1 - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)}\widetilde{\boldsymbol{\theta}}(t) \otimes \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)
$$

$$
= \frac{1}{1 - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)}[\mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t) - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)\widetilde{\boldsymbol{\theta}}(t)]
$$

$$
= \frac{1}{1 - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)}[\widehat{\boldsymbol{\theta}}_{m|k}^{\text{fLBF}}(t) - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)\widetilde{\boldsymbol{\theta}}(t)].
$$

$$(3.151)$$

As a consequence

$$
\varepsilon_{o,m|k}^{\text{fLBF}}(t) = y(t) - [\widehat{\boldsymbol{\theta}}_{o,m|k}^{\text{fLBF}}(t)]^{\text{H}}\boldsymbol{\varphi}(t) = \frac{1}{1 - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)}[\varepsilon_{m|k}^{\text{fLBF}}(t) - \mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)\widetilde{\varepsilon}(t)]
$$

$$
= \frac{1}{1 - h_{m|k}^{\text{fLBF}}(0)}[\varepsilon_{m|k}^{\text{fLBF}}(t) - h_{m|k}^{\text{fLBF}}(0)\widetilde{\varepsilon}(t)],
$$

$$(3.152)$$

where

$$
\varepsilon_{m|k}^{\text{fLBF}}(t) = y(t) - [\widehat{\boldsymbol{\theta}}_{m|k}^{\text{fLBF}}(t)]^{\text{H}}\boldsymbol{\varphi}(t),
$$

$$(3.153)$$

denotes the fLBF interpolation error, and

$$
\widetilde{\varepsilon}(t) = y(t) - \widetilde{\boldsymbol{\theta}}^{\text{H}}(t)\boldsymbol{\varphi}(t),
$$

$$(3.154)$$

is a preestimate interpolation error. This means, that leave-one-out interpolation errors can be found without computing new estimates of system parameters. Note also, that the expression analogous to (3.152) can be derived for any filter with the impulse response $\{h(i),\ i \in I_k\}$, even if it is not associated with any basis functions [77].

3.9.2 Decentralized approach

Unlike in the centralized approach, in this approach, every parameter has its own quality measure based not on the errors related to the system output, but on the filtering performance. The procedure is as follows, at each time instant, and for every parameter, one selects hyperparameters from sets of possible hyperparameter candidates $\mathcal{M}_j, \mathcal{K}_j,\ j = 1, \ldots, n$, for which the estimator minimizes the local quality measure

$$
\{m_j(t), k_j(t)\} = \underset{m_j \in \mathcal{M}_j, k_j \in \mathcal{K}_j}{\arg\min}\ J_j(t), \quad j = 1, \ldots, n,
$$

$$(3.155)$$

and

$$\widehat{\theta}_{j,m|k}^{\text{fLBF}}(t) = \widehat{\theta}_{j,m_j(t)|k_j(t)}^{\text{fLBF}}(t). \tag{3.156}$$

Such a quality measure can be derived either from the FPE criterion or again using the LOOCV approach.

FPE criterion

Let $\Omega'_{j,k}(t) = \{z'_j(t+i),\ i \in I_k\},\ j = 1, \ldots, n$, denote the realization of preestimation noise, independent of the realization $\Omega_{j,k}(t) = \{z_j(t+i),\ i \in I_k\},\ j = 1, \ldots, n,$, used for filtering. The FPE for each parameter trajcetroy is defined as follows

$$\upsilon_{j,m|k}(t) = \text{E}_{\Omega'_{j,k}(t)\Omega_{j,k}(t)} \left[|\widetilde{\theta}'_j(t) - \widehat{\theta}_{j,m|k}^{\text{fLBF}}(t)|^2 \right], \quad j = 1, \ldots, n. \tag{3.157}$$

This can be expressed as

$$\upsilon_{j,m|k}(t) = \text{E}_{\Omega'_{j,k}(t)\Omega_{j,k}(t)} \left[\left| \theta_j(t) + z'_j(t) - \sum_{i=-k}^{k} h_{m|k}^{\text{fLBF}}(i)\theta_j(t+i) - \sum_{i=-k}^{k} h_{m|k}^{\text{fLBF}}(i)z_j(t+i) \right|^2 \right]. \tag{3.158}$$

When the assumption (A2.1) is fulfilled, then $\sum_{i=-k}^{k} h_{m|k}^{\text{fLBF}}(i)\theta_j(t+i) = \theta_j(t)$, and (3.157) becomes

$$
\begin{aligned}
\upsilon_{j,m|k}(t) &= \text{E}_{\Omega'_{j,k}(t)\Omega_{j,k}(t)} \left[\left| z'_j(t) - \sum_{i=-k}^{k} h_{m|k}^{\text{fLBF}}(i)z_j(t+i) \right|^2 \right] \\
&= \text{E}_{\Omega'_{j,k}(t)\Omega_{j,k}(t)}[||z'_j(t)|^2] + \text{E}_{\Omega'_{j,k}(t)\Omega_{j,k}(t)} \left[\left| \sum_{i=-k}^{k} h_{m|k}^{\text{fLBF}}(i)z_j(t+i) \right|^2 \right] \\
&= \sigma_{z_j}^2 + \text{E}_{\Omega'_{j,k}(t)\Omega_{j,k}(t)} \left[\left| \sum_{i=-k}^{k} h_{m|k}^{\text{fLBF}}(i)z_j(t+i) \right|^2 \right],
\end{aligned}
\tag{3.159}
$$

where the last transition follows from the fact, that $z'_j(t)$ is independent of $z_j(t)$ and zero-mean. Under (A3.2) it holds that

$$\upsilon_{j,m|k}(t) \cong \sigma_z^2 \left(1 + \frac{1}{l_{m|k}^{\text{fLBF}}} \right). \tag{3.160}$$

The final expression can be obtained using

$$
\begin{aligned}
\widehat{\sigma}_z^2 &= \frac{1}{L_w} \sum_{i=-k}^{k} w_k(i) |\widetilde{\theta}_j(t+i) - \mathbf{f}_{m|k}^{\text{H}}(i)\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fLBF}}(t)|^2 \\
&= \frac{1}{L_w} \sum_{i=-k}^{k} w_k(i) |\theta_j(t+i) + z_j(t+i) - \mathbf{f}_{m|k}^{\text{H}}(i)\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fLBF}}(t)|^2 \\
&= \frac{1}{L_w} \sum_{i=-k}^{k} w_k(i) |z_j(t+i) - \mathbf{f}_{m|k}^{\text{H}}(i)\Delta\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fLBF}}(t)|^2.
\end{aligned}
\tag{3.161}
$$

Under (A2.1) it holds that

$$
\begin{aligned}
\Delta\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fLBF}}(t) &= \widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fLBF}}(t) - \boldsymbol{\alpha}_j(t) = \sum_{i=-k}^{k} w_k(i)\mathbf{f}_{m|k}(i)\widetilde{\theta}_j(t+i) - \boldsymbol{\alpha}_j(t) \\
&= \sum_{i=-k}^{k} w_k(i)\mathbf{f}_{m|k}(i)\mathbf{f}_{m|k}^{\text{H}}(i)\boldsymbol{\alpha}_j(t) - \boldsymbol{\alpha}_j(t) + \sum_{i=-k}^{k} w_k(i)\mathbf{f}_{m|k}(i)z_j(t+i) \\
&= \sum_{i=-k}^{k} w_k(i)\mathbf{f}_{m|k}(i)z_j(t+i),
\end{aligned}
\tag{3.162}
$$

hence

$$\widehat{\sigma}_z^2 = \frac{1}{L_w} \sum_{i=-k}^{k} w_k(i) \left| z_j(t+i) - \mathbf{f}_{m|k}^{\mathrm{H}}(i) \sum_{s=-k}^{k} w_k(s) \mathbf{f}_{m|k}(s) z_j(t+s) \right|^2$$

$$= \frac{1}{L_w} \sum_{i=-k}^{k} w_k(i) \left| z_j(t+i) - \mathbf{f}_{m|k}^{\mathrm{H}}(i) \boldsymbol{\xi}_{j,m|k}(t) \right|^2, \tag{3.163}$$

where $\boldsymbol{\xi}_{j,m|k}(t) = \sum_{s=-k}^{k} w_k(s) \mathbf{f}_{m|k}(s) z_j(t+s)$. Under (A3.2), the expected value of $\widehat{\sigma}_z^2$ has the form

$$\mathrm{E}[\widehat{\sigma}_z^2] = \frac{1}{L_w} \sum_{i=-k}^{k} w_k(i) \Big\{ \mathrm{E}[|z_j(t+i)|^2] - \mathrm{E}[z_j^*(t+i) \mathbf{f}_{m|k}^{\mathrm{H}}(i) \boldsymbol{\xi}_{j,m|k}(t)] - \mathrm{E}[z_j(t+i) \boldsymbol{\xi}_{j,m|k}^{\mathrm{H}}(t) \mathbf{f}_{m|k}(i)]$$

$$+ \mathrm{E}[\boldsymbol{\xi}_{j,m|k}^{\mathrm{H}}(t) \mathbf{f}_{m|k}(i) \mathbf{f}_{m|k}^{\mathrm{H}}(i) \boldsymbol{\xi}_{j,m|k}(t)] \Big\}$$

$$= \sigma_z^2 - \frac{2}{L_w} \sum_{i=-k}^{k} w_k^2(i) \mathbf{f}_{m|k}(i) \mathbf{f}_{m|k}^{\mathrm{H}}(i) \sigma_z^2 + \frac{1}{L_w} \sum_{i=-k}^{k} w_k^2(i) \mathbf{f}_{m|k}(i) \mathbf{f}_{m|k}^{\mathrm{H}}(i) \sigma_z^2$$

$$= \sigma_z^2 \left(1 - \frac{1}{N_{m|k}} \right). \tag{3.164}$$

Combining the result above with (3.160), one arrives at

$$\mathrm{FPE}_j(t) = v_{j,m|k}(t) \cong \frac{1 + \frac{1}{l_{m|k}^{\mathrm{fLBF}}}}{1 - \frac{1}{N_{m|k}}} \widehat{\sigma}_z^2(t), \tag{3.165}$$

where $\widehat{\sigma}_z^2$ can be evaluated as

$$\widehat{\sigma}_z^2(t) = \frac{1}{L_w} \sum_{i=-k}^{k} w_k(i) |\widetilde{\theta}_j(t+i) - \mathbf{f}_{m|k}^{\mathrm{H}}(i) \widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fLBF}}(t)|^2 = \frac{1}{L_w} \sum_{i=-k}^{k} w_k(i) |\widetilde{\theta}_j(t+i)|^2 - \frac{1}{L_w} ||\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fLBF}}(t)||^2. \tag{3.166}$$

Note that (3.165) has the same form as (2.117).

Leave-one-out cross-validation

The alternative to the FPE-based criterion is to use LOOCV, which involves the quality measure defined as

$$J_j(t) = \sum_{i=-L}^{L} |\varepsilon_{o,j}^{\mathrm{fLBF}}(t+i)|^2, \tag{3.167}$$

where

$$\varepsilon_{o,j}^{\mathrm{fLBF}}(t) = \widetilde{\theta}_j(t) - \widehat{\theta}_{o,j}^{\mathrm{fLBF}}(t), \tag{3.168}$$

is a leave-one-out interpolation error, and

$$\widehat{\theta}_{o,j}^{\mathrm{fLBF}}(t) = \frac{1}{\sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i) - h_{m|k}^{\mathrm{fLBF}}(0)} \sum_{\substack{i=-k \\ i \neq 0}}^{k} h_{m|k}^{\mathrm{fLBF}}(i) \widetilde{\theta}_j(t), \tag{3.169}$$

is a holey estimate of parameter trajectory. It is immediate to see that when the applied filter is a lowpass filter, then

$$\widehat{\theta}_{o,j}^{\mathrm{fLBF}}(t) = \frac{1}{1 - h_{m|k}^{\mathrm{fLBF}}(0)} [\widehat{\theta}_{j,m|k}^{\mathrm{fLBF}}(t) - h_{m|k}^{\mathrm{fLBF}}(0) \widetilde{\theta}_j(t)]. \tag{3.170}$$

As a consequence

$$\varepsilon_{o,j}^{\mathrm{fLBF}}(t) = \frac{1}{1 - h_{m|k}^{\mathrm{fLBF}}(0)} \varepsilon_{j,m|k}^{\mathrm{fLBF}}(t), \tag{3.171}$$

where

$$\varepsilon_{j,m|k}^{\text{fLBF}}(t) = \widetilde{\theta}_j(t) - \widehat{\theta}_{j,m|k}^{\text{fLBF}}(t). \tag{3.172}$$

Just like in the centralized case, the leave-one-out cross-validation approach can be used even if the filtration is performed with a finite impulse response filter that is not associated with any basis functions.

Remark

If one wants to choose hyperparameters jointly for all system parameters using a decentralized approach, then the following quality measure can be adopted

$$J(t) = \sum_{j=1}^n J_j(t). \tag{3.173}$$

However, it is worth noting that since typically the variance of preestimation noise is much larger than the variance of the measurement noise $\sigma_z^2 \gg \sigma_e^2$, the decentralized approach for choosing hyperparameters in a joint manner might yield inferior results to the results provided by the centralized strategy.

Sign test

One can also adopt the approach based on counting the sign changes of residuals, for choosing the number of basis functions m. If a model consisting of m basis functions is locally describing the true trajectory sufficiently accurately, then under (A2.5) and (A3.1), the residual errors

$$\epsilon_{j,m|k}^{\text{fLBF}}(t,i) = \widetilde{\theta}_j(t+i) - \mathbf{f}_{m|k}^{\text{H}}(i)\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fLBF}}(t), \quad i \in I_k, \tag{3.174}$$

can be approximately viewed as a realization of white noise. As a consequence, the signs of residuals should change on average every second sample. Define

$$\delta_{j,m|k}^{\text{R}}(t,i) = \begin{cases} 1 & \text{if } \text{Re}[\epsilon_{j,m|k}^{\text{fLBF}}(t,i)] > 0 \\ -1 & \text{if } \text{Re}[\epsilon_{j,m|k}^{\text{fLBF}}(t,i)] \leq 0 \end{cases}, \quad i \in I_k, \tag{3.175}$$

the signs of real part of residuals. Furthermore, define the set of indices for which the $\delta_{m|k}^{\text{R}}(t,i)$, $i \in I_k$ changes sign

$$\mathcal{D}_{j,m|k}^{\text{R}}(t) = \{i : \delta_{j,m|k}^{\text{R}}(t,i-1)\delta_{j,m|k}^{\text{R}}(t,i) < 0, \quad i \in [-k+1, k]\}. \tag{3.176}$$

Denote also by

$$d_{j,m|k}^{\text{R}}(t) = \text{card}[\mathcal{D}_{j,m|k}^{\text{R}}(t)] \in [0, K-1], \tag{3.177}$$

the number of sign changes of the real part of residuals inside the analysis window. If the sequence of residuals is a realization of a sequence of mutually independent, random variables with symmetrical probability distribution function, then the signs of residuals should obey the binomial distribution (see [25]). Hence, the test for the quality of a model will be based on checking whether $d_{j,m|k}^{\text{R}}(t)$ is above some statistically justified threshold.

Consider the following null hypothesis:

$\mathcal{H}_{0,j}^{\text{R}}(t)$: $\{\text{Re}[\epsilon_{j,m|k}^{\text{fLBF}}(t+i)], \quad i \in I_k\}$ is a sequence of independent random variables obeying the condition

$$P\left(\text{Re}[\epsilon_{j,m|k}^{\text{fLBF}}(t+i)] > 0\right) = P\left(\text{Re}[\epsilon_{j,m|k}^{\text{fLBF}}(t+i)] \leq 0\right), \quad \forall_{i \in I_k},$$

and $P(\cdot)$ denotes the probability.

If the null hypothesis is true, then one has

$$P\left(d_{j,m|k}^{\text{R}}(t) = d | \mathcal{H}_{0,j}^{\text{R}}(t)\right) = \frac{(K-1)!}{2^{K-1}d!(K-1-d)!}. \tag{3.178}$$

As a consequence, one obtains

$$\forall_{d_0\in[1,K-1]} \quad P\left(d_{j,m|k}^{\mathrm{R}}(t) \leq d_0|\mathcal{H}_{0,j}^{\mathrm{R}}(t)\right) = \sum_{d=0}^{d_0} \frac{(K-1)!}{2^{K-1}d!(K-1-d)!} = \eta_0^{\mathrm{R}}. \tag{3.179}$$

One can also define the analogous quantities $\delta_{j,m|k}^{\mathrm{I}}(t,i)$, $\mathcal{D}_{j,m|k}^{\mathrm{I}}(t)$, $d_{j,m|k}^{\mathrm{I}}(t)$ and η_0^{I} for the imaginary part of the residuals as well as the null hypothesis $\mathcal{H}_{0,j}^{\mathrm{I}}(t)$. The sign test can be used to choose the number of basis functions in the fLBF method. If the number of local signs changes in the real part $d_{j,m|k}^{\mathrm{R}}(t) > d_0$ or in the imaginary part $d_{j,m|k}^{\mathrm{I}}(t) > d_0$, then one should increase m by adding a new basis function to the set of currently used functions and repeat the identification process inside the current analysis window. More precisely, one should start with one, constant basis function, which is sufficient to model static parameters, run the sign test, and in case it indicates that one basis function does not suffice to model parameter dynamics, one should add another basis function, repeat the estimation process and the sign test. The procedure of adding new basis functions and repeating the estimation should be repeated until the sign test indicates that the currently used model explains parameter dynamics sufficiently.

The threshold d_0 can be calculated based on the chosen probability of type I errors

$$\eta_0 = P\left(d_{j,m|k}^{\mathrm{R}}(t) \leq d_0 \cup d_{j,m|k}^{\mathrm{I}}(t) \leq d_0|\mathcal{H}_{0,j}^{\mathrm{R}}(t) \cap \mathcal{H}_{0,j}^{\mathrm{I}}(t)\right) = P\left(d_{j,m|k}^{\mathrm{R}}(t) \leq d_0|\mathcal{H}_{0,j}^{\mathrm{R}}(t)\right)$$
$$+ P\left(d_{j,m|k}^{\mathrm{I}}(t) \leq d_0|\mathcal{H}_{0,j}^{\mathrm{I}}(t)\right) - P\left(d_{j,m|k}^{\mathrm{R}}(t) \leq d_0 \cap d_{j,m|k}^{\mathrm{I}}(t) \leq d_0|\mathcal{H}_{0,j}^{\mathrm{R}}(t) \cap \mathcal{H}_{0,j}^{\mathrm{I}}(t)\right). \tag{3.180}$$

Assuming that the real and imaginary part of residuals are independent, one obtains

$$\eta_0 = \eta_0^{\mathrm{R}} + \eta_0^{\mathrm{I}} - \eta_0^{\mathrm{R}}\eta_0^{\mathrm{I}} = 1 - (1-\eta_0^{\mathrm{R}})(1-\eta_0^{\mathrm{I}}) = 1 - (1-\eta_0^{\mathrm{R}})^2$$
$$= 1 - \left[1 - \sum_{d=0}^{d_0} \frac{(K-1)!}{2^{K-1}d!(K-1-d)!}\right]^2, \tag{3.181}$$

where the last transition follows from the fact that the threshold is the same for the real and imaginary parts, hence $\eta_0^{\mathrm{R}} = \eta_0^{\mathrm{I}}$. One should always choose a sufficiently small value of a probability of type I errors, typically $\eta_0 = 0.01$ or $\eta_0 = 0.05$.

Remark 1

Note that when the basis functions are w_k-orthonormal, then the residuals from the equation (3.174) can be computed in an order-recursive fashion. Note that

$$\widehat{\boldsymbol{\alpha}}_{j,m+1|k}^{\mathrm{fLBF}}(t) = \begin{bmatrix} \widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fLBF}}(t) \\ \widehat{\alpha}_{j,m+1|k}^{m+1}(t) \end{bmatrix}, \tag{3.182}$$

where

$$\widehat{\alpha}_{j,m+1|k}^{m+1}(t) = \sum_{i=-k}^{k} w_k(i)\widetilde{\theta}_j(t+i)f_{m+1|k}(i). \tag{3.183}$$

Hence

$$\epsilon_{j,m+1|k}^{\mathrm{fLBF}}(t,i) = \epsilon_{j,m|k}^{\mathrm{fLBF}}(t,i) - f_{m+1|k}^*(i)\widehat{\alpha}_{j,m+1|k}^{m+1}(t), \quad i \in I_k. \tag{3.184}$$

Remark 2

For the real-valued systems, i.e. systems in which input and output signals, and parameters are real-valued, the probability of type I errors has a simpler form

$$\eta_0 = P(d_{j,m|k}^{\mathrm{R}}(t) \leq d_0|\mathcal{H}_{0,j}^{\mathrm{R}}(t)) = \sum_{d=0}^{d_0} \frac{(K-1)!}{2^{K-1}d!(K-1-d)!}. \tag{3.185}$$

Remark 3

If the resudiuals are i.i.d. and symmetrically distributed, the numbers of sign changes $d_{j,m|k}^{\mathrm{R}}(t)$ and $d_{j,m|k}^{\mathrm{I}}(t)$ can be seen as sums of independent and identically distributed random variables and, according to the central limit theorem, they have the asymptotical normal distribution

$$
\begin{aligned}
d_{j,m|k}^{\mathrm{R}}(t) &\xrightarrow{\text{dist.}} \mathcal{N}(\overline{d}_k, \sigma_d^2) \\
d_{j,m|k}^{\mathrm{I}}(t) &\xrightarrow{\text{dist.}} \mathcal{N}(\overline{d}_k, \sigma_d^2),
\end{aligned}
\tag{3.186}
$$

where (see [98])

$$
\begin{aligned}
\overline{d}_k &= \mathrm{E}[d_{j,m|k}^{\mathrm{R}}(t)] = \mathrm{E}[d_{j,m|k}^{\mathrm{I}}(t)] = \frac{K-1}{2} = k \\
\sigma_d^2 &= \mathrm{var}[d_{j,m|k}^{\mathrm{R}}(t)] = \mathrm{var}[d_{j,m|k}^{\mathrm{I}}(t)] = \mathrm{E}\{[d_{j,m|k}^{\mathrm{R}}(t)]^2\} - k^2 = \frac{K-1}{4} = \frac{k}{2},
\end{aligned}
\tag{3.187}
$$

which means that

$$
\begin{aligned}
\frac{d_{j,m|k}^{\mathrm{R}}(t) - k}{\sqrt{2k}/2} &\xrightarrow{\text{dist.}} \mathcal{N}(0,1) \\
\frac{d_{j,m|k}^{\mathrm{I}}(t) - k}{\sqrt{2k}/2} &\xrightarrow{\text{dist.}} \mathcal{N}(0,1),
\end{aligned}
\tag{3.188}
$$

and for K large enough, with approximately 95% confidence it holds

$$
\frac{|d_{j,m|k}^{\mathrm{R}}(t) - k|}{\sqrt{2k}/2} \leq 1.96,
\tag{3.189}
$$

and the same holds true for the imaginary part of residuals. Using the reasoning leading to (3.181), the probability that both numbers of sign changes will not obey the above condition, given that residuals are i.i.d. and symmetrically distributed, is equal to $\eta_0 = 0.0975 \approx 0.01$. Hence, for sufficiently large values of K, one can use this simplified version of a sign test. Note that this simplified test poses a lower and an upper bound on the number of sign changes. For example, for $\eta_0 = 0.01$ and $K = 1021$ the full sign test gives a threshold of 468, while the simplified test indicates that the number of sign changes should be in the interval $[478, 542]$, for $K = 401$, one obtains the exact threshold equal to 173, and the interval provided by the simplified test is $[180, 220]$.

3.10   Computational complexity

The computation of EWLS-based preestimates requires $\mathcal{O}(n)$ operations when implemented using fast transversal filters [5], [35], [53]. It is worth noting that one should take some precautions when implementing them with finite precision [35]. However, one can obtain similar computational complexity (also of order $\mathcal{O}(n)$) with DCD algorithm [111]. Computing preestimates also requires $\mathcal{O}(n)$ operations.

In the second step - postfiltering, one can take one of the two approaches. The first one is to find the estimates of basis function coefficients first, using equation (3.16) (when the basis is w_k-orthonormal), and then calculate the estimates of system parameters. The second approach - filtering, is to filter the preestimates using the impulse response associated with the chosen basis functions (equation (3.17)).

In the first case, computational complexity can be substantially reduced (compared with direct implementation of equation (3.16)) with recursive computations. The idea of the recursive computations was presented in Section 2.6.5. Note that the formulae (2.101) and (2.102) can be easily adapted to the fLBF case. In fact, one does not need to update the regression matrix, and when the basis is w_k-orthonormal, one does not need to inverse it either. In formulas for updating $\mathbf{p}_{m|k}(t)$, one only needs to replace $y(t)$ with 1 and $\boldsymbol{\varphi}(t)$ with $\widetilde{\boldsymbol{\theta}}(t)$.

In the second approach, one can use the FFT-based approach (even when working in an almost real-time fashion, the technique presented in [95] can be used).

Chapter 4

Regularization in basis function methods

4.1 Introduction

The ultimate goal of identification is to minimize mean squared estimation errors. These errors can be decomposed into bias and variance components. As described in previous chapters, reducing one of them usually increases the value of the second one. This is why one should seek a bias-variance trade-off. One of the powerful tools, that help to reach such a trade-off, is regularization. In this chapter we show how to incorporate regularization in the local basis function and the fast local basis function methods, resulting in regularized LBF (RLBF) and regularized fLBF (fRLBF) methods. The content of this chapter evolves around the four questions that one may ask about regularization:

1. What is being constrained?

2. What penalty is used?

3. What prior knowledge is available?

4. What optimization technique is used?

4.2 Regularized local basis function method

Before one can formulate expressions for the RLBF estimator, one needs to answer the first two of the aforementioned questions. One can put constraints on the vector of basis function coefficients, or on the vector of system parameters. Regarding the second question, one can choose e.g. the ℓ^2 or ℓ^1 norm. First, note that the functional form of the final formula will be the same when the constraint is put on the vector of basis function coefficients and on the vector of system parameters. However, the resulting regularization matrix will be different because under (A2.1) $\boldsymbol{\theta}(t) = \mathbf{F}_{m|k}(0)\boldsymbol{\alpha}(t)$. Therefore, we first show the more natural formulation, when one constrains the basis function coefficients. Later we explain, how one can use this formulation to impose constraints on the vector of system parameters $\boldsymbol{\theta}(t)$.

4.2.1 Formulation for the ℓ^2 penalty

Let $\mathbf{G}$ be an $mn \times mn$ Hermitian nonnegative-definite regularization matrix and $||\boldsymbol{\alpha}(t)||_{\mathbf{G}}^2 = \boldsymbol{\alpha}^{\mathrm{H}}(t)\mathbf{G}\boldsymbol{\alpha}(t)$. The RLBF estimates are defined as follows

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t) = \arg\min_{\boldsymbol{\alpha}} \left\{ \sum_{i=-k}^{k} w_k(i)|y(t+i) - \boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\psi}_{m|k}(t,i)|^2 + \mu||\boldsymbol{\alpha}||_{\mathbf{G}}^2 \right\}$$

$$\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{RLBF}}(t) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t),$$

(4.1)

where $\mu > 0$ is a regularization constant. This constant is very important since it controls the balance between the influence of user requirements for the system parameters or basis function coefficients and the information contained in the available data. Therefore, the optimization of this constant is a crucial factor and will be discussed in detail in the next sections. Assume now, that one wants to put constraints on the vector of system parameters using the Hermitian nonnegative regularization matrix $\mathbf{H} = \mathbf{D}^{\mathrm{H}}\mathbf{D}$, where $\mathbf{D}$ is an $l \times n$ matrix, $l \geq n$, this can be achieved by noting that

$$||\boldsymbol{\theta}(t)||_{\mathbf{H}}^2 = \boldsymbol{\theta}^{\mathrm{H}}(t)\mathbf{H}\boldsymbol{\theta}(t) = \boldsymbol{\alpha}^{\mathrm{H}}(t)\mathbf{F}_{m|k}^{\mathrm{H}}(0)\mathbf{H}\mathbf{F}_{m|k}(0)\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}^{\mathrm{H}}(t)\mathbf{G}\boldsymbol{\alpha}(t) = ||\boldsymbol{\alpha}(t)||_{\mathbf{G}}^2, \qquad (4.2)$$

provided that (A2.1) holds true. Hence, one can use the expression (4.1) when the constraints are imposed on the vector of system parameters $\boldsymbol{\theta}(t)$ as well as when they are formulated for the vector of basis function coefficients $\boldsymbol{\alpha}(t)$.

It can be shown that

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t) = \mathbf{S}_{m|k}^{-1}(t)\mathbf{p}_{m|k}(t), \qquad (4.3)$$

where

$$\mathbf{S}_{m|k}(t) = \mathbf{P}_{m|k}(t) + \mu\mathbf{G} = \sum_{i=-k}^{k} w_k(i)\boldsymbol{\psi}_{m|k}(t,i)\boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,i) + \mu\mathbf{G}. \qquad (4.4)$$

Note that, one can still use the recursive formulae introduced in one of the previous chapters to update the matrix $\mathbf{S}_{m|k}(t)$.

From now on, we will assume in all derivations for the RLBF method (unless stated otherwise) that

$$\mathbf{G} = \mathbf{F}_{m|k}^{\mathrm{H}}(0)\mathbf{H}\mathbf{F}_{m|k}(0) = \mathbf{F}_{m|k}^{\mathrm{H}}(0)\mathbf{D}^{\mathrm{H}}\mathbf{D}\mathbf{F}_{m|k}(0) = \mathbf{B}^{\mathrm{H}}\mathbf{B} = \mathbf{H} \otimes \mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\mathrm{H}}(0), \qquad (4.5)$$

where $\mathbf{B} = \mathbf{D}\mathbf{F}_{m|k}(0) = \mathbf{D} \otimes \mathbf{f}_{mk}^{\mathrm{H}}(0)$.

4.2.2 Formulation for the ℓ^1 penalty

The RLBF estimator with the ℓ^1 penalty is defined as follows

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t) = \arg\min_{\boldsymbol{\alpha}} \left\{ \sum_{i=-k}^{k} w_k(i)|y(t+i) - \boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\psi}_{m|k}(t,i)|^2 + \mu||\boldsymbol{\alpha}||_1 \right\}$$
$$\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{RLBF}}(t) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t). \qquad (4.6)$$

As it was explained in [100], such a formulation of the problem promotes the sparsity of basis function coefficients. However, the closed-form solution for such a problem is not known in general, and the minimization problem with the ℓ^1 norm is considered difficult. One can either solve (4.6) numerically as an optimization problem or apply an iterative reweighting technique, which is a useful tool in approximation theory [8], [12]. Note that the ℓ^1 norm can be written down as

$$|x| = \frac{xx^*}{|x|},$$

which means that the formula (4.6) can be approximated by a RLBF method with ℓ^2 penalty and suitably designed regularization matrix, namely

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t,p) = \arg\min_{\boldsymbol{\alpha}} \left\{ \sum_{i=-k}^{k} w_k(i)|y(t+i) - \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}|^2 + \mu||\boldsymbol{\alpha}||_{\mathbf{G}_p}^2 \right\}$$
$$\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{RLBF}}(t,p) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t,p), \quad p \in \mathbb{N}_+, \qquad (4.7)$$

where

$$\mathbf{G}_p = \mathrm{diag}\{|\widehat{\alpha}_{1,1}^{\mathrm{RLBF}}(t,p-1)|,\ldots,|\widehat{\alpha}_{m,1}^{\mathrm{RLBF}}(t,p-1)|,\ldots|\widehat{\alpha}_{1,n}^{\mathrm{RLBF}}(t,p-1)|,\ldots,|\widehat{\alpha}_{m,n}^{\mathrm{RLBF}}(t,p-1)|\}^{-1},$$
$$p \in \mathbb{N}_+,$$
$$(4.8)$$

and $\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t,0) = \widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{LBF}}(t)$. When the ℓ^1 penalty is applied to the vector of system parameters $\boldsymbol{\theta}(t)$, the matrix $\mathbf{G}_p$ takes the form

$$\mathbf{G}_p = \mathbf{F}_{m|k}^{\mathrm{H}}(0)\mathrm{diag}\{|\widehat{\theta}_1^{\mathrm{RLBF}}(t,p-1)|,\ldots,|\widehat{\theta}_n^{\mathrm{RLBF}}(t,p-1)|,\}^{-1}\mathbf{F}_{m|k}(0), \quad p \in \mathbb{N}, \qquad (4.9)$$

and $\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{RLBF}}(t,0) = \widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{LBF}}(t)$. Such estimators were investigated in [22].

4.3 Regularization matrix design

Designing the regularization matrix appropriately, one can impose certain restrictions on the shape and smoothness of the estimated instantaneous impulse response of a time-varying system. Its structure can also reflect the available prior knowledge about parameter changes and in this sense, it addresses the third of the questions listed in the introduction to this chapter. One can find more information about the design of the regularization matrices in the publication [15]. It discusses two perspectives on the design of regularization matrices. Even though the article describes the identification of time-invariant systems, it contains many valuable insights.

4.3.1 Design with no prior knowledge about parameter changes

If no prior knowledge about the parameter variations is available, one can use the regularization matrix either to implement the reweighted estimation scheme, as it was shown earlier, or to prevent the estimation procedure from ill-conditioning, e.g. by using the ridge regression.

4.3.2 Design based on partial prior knowledge about parameter variations

When the partial or full (statistical) knowledge about parameter changes is available, one can parametrize the regularization matrix in terms of some prior-related hyperparameter vector $\boldsymbol{\beta}$, resulting in $\mathbf{H}(\boldsymbol{\beta})$.

Time-domain smoothness priors

Historically, the first attempt to incorporate prior knowledge into the design of a regularization matrix was made by Whittaker [108]. Later on, his approach was rediscovered and/or extended by many authors - see e.g. [3], [41], [81]. This method, based on time-domain smoothness priors was later used for Kalman-filter-based identification of nonstationary processes [41].

Define the forward lag-shift operator δ, working in the domain of a frozen impulse response of a system, namely

$$\delta\theta_j(t) = \theta_{j+1}(t), \quad j = 1,\ldots,n-1.$$

One can impose smoothness constraints on $\{\theta_j(t)\}$ by penalizing the excessive values of the following p-th order finite differences

$$\nabla^p\theta_j(t) = (1-\delta)^p\theta_j(t) = \sum_{i=0}^{p}(-1)^i\binom{p}{i}\delta^{p-i}\theta_j(t) = \sum_{i=0}^{p}c_{p,i}\theta_{j+p-i}(t), \quad j = 1,\ldots,n,$$

where the boundary conditions are $\theta_j(t) = 0$, $j > n$. Such boundary conditions allow to achieve an additional goal. Besides imposing the smoothness constraints, the regularization based on the above formula helps to achieve the tapering (gradual decay to zero) of the last p impulse response coefficients, which is often beneficial in practical applications.

Using time-domain smoothness priors in identification results in penalizing the excessive values of a vector $\mathbf{D}(p)\boldsymbol{\theta}(t)$, where

$$
\mathbf{D}(p) = \begin{bmatrix}
c_{p,0} & c_{p,1} & \dots & c_{p,p} & 0 & \dots & 0 & 0 \\
0 & c_{p,0} & \dots & c_{p,p-1} & c_{p,p} & \dots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & c_{p,0} & c_{p,1} & \dots & c_{p,p} & 0 \\
0 & 0 & \dots & c_{p,0} & c_{p,1} & \dots & c_{p,p-1} & c_{p,p} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & \dots & c_{p,0} & 0 \\
0 & 0 & \dots & 0 & 0 & \dots & 0 & c_{p,0}
\end{bmatrix}, \tag{4.10}
$$

is an $n \times n$ matrix. As a consequence, the regularization matrix takes the form $\mathbf{H}(p) = \mathbf{D}^{\mathrm{T}}(p)\mathbf{D}(p)$ because using $||\boldsymbol{\theta}(t)||^2_{\mathbf{H}(p)}$ will boil down to

$$
||\boldsymbol{\theta}(t)||^2_{\mathbf{H}(p)} = \boldsymbol{\theta}^{\mathrm{H}}(t)\mathbf{H}(p)\boldsymbol{\theta}(t) = \boldsymbol{\theta}^{\mathrm{H}}(t)\mathbf{D}^{\mathrm{T}}(p)\mathbf{D}(p)\boldsymbol{\theta}(t) = ||\mathbf{D}(p)\boldsymbol{\theta}(t)||^2 = ||\mathbf{x}(t)||^2,
$$

where $\mathbf{x}(t) = \mathbf{D}(p)\boldsymbol{\theta}(t) = [\nabla^p\theta_1(t), \dots, \nabla^p\theta_n(t)]^{\mathrm{T}}$ is a vector of p-th order finite differences.

For example, when $p = 1$, one promotes local constancy of the impulse response, when $p = 2$, one promotes local linearity of the impulse response, when $p = 3$, one promotes local convexity of the impulse response, and so on. The case of $p = 0$ corresponds to the traditional ridge regression [101], in which one penalizes the excessive values of the impulse response coefficients.

Frequency-domain smoothness priors

This approach was originally proposed in the article [26] for time-invariant processes. However, it can be easily adapted to the time-varying, complex-valued case. Define the frozen frequency response of a time-varying system as

$$
\Theta(t, \omega) = \sum_{j=1}^{n} \theta_j(t)e^{-\mathrm{i}\omega j}. \tag{4.11}
$$

One can adopt the measure of smoothness of the frozen frequency response, defined as

$$
\int_{-\pi}^{\pi} \left| \frac{\partial^r \Theta(t, \omega)}{\partial \omega^r} \right|^2 d\omega = 2\pi \sum_{j=1}^{n} j^{2r}|\theta_j(t)|^2. \tag{4.12}
$$

The above formula follows directly from Parseval's theorem and the differentiation rule in the frequency domain (see [85]). Using this formula, one can impose some constraints on the smoothness of the frozen frequency response of the system by applying regularization on the vector $\mathbf{D}(r)\boldsymbol{\theta}(t)$, where

$$
\mathbf{D}(r) = \mathrm{diag}\{1^r, 2^r, \dots, n^r\}, \tag{4.13}
$$

which results in the regularization matrix $\mathbf{H}(r) = \mathbf{D}^2(r)$.

Such a choice of regularization matrix promotes also the stability of the instantaneous impulse response because the higher penalty is imposed on the last coefficients of the frozen impulse response, which promotes their small values.

Stability priors

Recently many researchers proposed different regularization matrices whose structure is based upon the assumption of exponential stability of the identified system. In such a case, one expects that the coefficients of the impulse response decay at an exponential rate. In the UAC case, the exponential decay of impulse response coefficients has a physical justification. The decay of a power profile of system parameters is caused by the spreading and absorption loss (see [99]).

A popular choice of a regularization matrix is based on the so-called kernels: diagonal-correlated (DC) kernel, and tuned correlation (TC) kernel. The DC kernel takes the form

$$
\mathbf{K}_{\mathrm{DC}}(\beta_1, \beta_2) = \begin{bmatrix}
\beta_1 & \beta_1^{\frac{3}{2}}\beta_2 & \beta_1^2\beta_2^2 & \dots \\
\beta_1^{\frac{3}{2}}\beta_2 & \beta_1^2 & \beta_1^{\frac{5}{2}}\beta_2 & \dots \\
\beta_1^2\beta_2^2 & \beta_1^{\frac{5}{2}}\beta_2 & \beta_1^3 & \dots \\
\vdots & \vdots & \vdots & \ddots
\end{bmatrix}, \quad \beta_1 \in (0, 1), |\beta_2| < 1. \tag{4.14}
$$

The TC kernel is a special case of a DC kernel, for $\beta_2 = \sqrt{\beta_1}$, namely

$$\mathbf{K}_{\mathrm{TC}}(\beta) = \begin{bmatrix} \beta & \beta^2 & \beta^3 & \dots \\ \beta^2 & \beta^2 & \beta^3 & \dots \\ \beta^3 & \beta^3 & \beta^3 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \beta \in (0,1). \tag{4.15}$$

The regularization matrix, based on kernels is defined as follows

$$\mathbf{H}(\boldsymbol{\beta}) = \mathbf{K}_*^{-1}(\boldsymbol{\beta}), \quad * \in \{\mathrm{DC}, \mathrm{TC}\}. \tag{4.16}$$

Recently, in [55], a different perspective on the design of regularization matrices was presented. The authors noted that regularization matrices based on both DC and TC kernels can be decomposed using the Cholesky decomposition $\mathbf{H}(\beta) = \mathbf{D}^{\mathrm{T}}(\boldsymbol{\beta})\mathbf{D}(\boldsymbol{\beta})$, and the operation $\mathbf{D}(\beta)\boldsymbol{\theta}(t)$ can be seen as prefiltering. They have shown that these kernels impose not only exponential stability constraints but also some smoothness constraints on the coefficients of the impulse response. The authors of this article presented also the explicit formulas for the entries of matrices $\mathbf{D}(\boldsymbol{\beta})$ and $\mathbf{H}(\boldsymbol{\beta})$ for DC and TC kernels.

4.3.3 Hyperperameter optimization based on LOOCV

When the regularization matrix $\mathbf{H}$ is parametrized, one can use the parallel estimation approach similar to the one described in the second chapter, to choose the value of the hyperparameters defining the regularization matrix. Here, the regularization constant will be also treated as a such hyperparameter, hence in this and the following sections regarding RLBF, the following modified problem formulation will be used

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = \arg\min_{\boldsymbol{\alpha}} \left\{ \sum_{i=-k}^{k} w_k(i)|y(t+i) - \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}|^2 + ||\boldsymbol{\alpha}||_{\mathbf{G}(\mathbf{H})}^2 \right\}$$
$$\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}), \tag{4.17}$$

where $\mathbf{G}(\mathbf{H}) = \mathbf{F}_{m|k}^{\mathrm{H}}(0)\mathbf{H}\mathbf{F}_{m|k}(0)$.

One can run several algorithms with different regularization matrices, and at each time instant decide which one should constitute the final estimate. The criterion for choosing the final estimate can be again based on the leave-one-out cross-validation.

Consider running several RLBF algorithms, each one with a different regularization matrix $\mathbf{H} \in \mathcal{H} = \{\mathbf{H}_1, \dots, \mathbf{H}_g\}$. At each time instant t, one can choose the estimates corresponding to the regularization matrix for which the local estimate of the variance of leave-one-out interpolation errors is the smallest

$$J_{0,m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = \sum_{i=-L}^{L} \left| \varepsilon_{0,m|k}^{\mathrm{RLBF}}(t+i|\mathbf{H}) \right|^2, \tag{4.18}$$

where $\varepsilon_{0,m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = y(t) - [\widehat{\boldsymbol{\theta}}_{0,m|k}^{\mathrm{RLBF}}(t|\mathbf{H})]^{\mathrm{H}}\boldsymbol{\varphi}(t)$ denotes the leave-one-out interpolation error, and

$$\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = \arg\min_{\boldsymbol{\alpha}} \left\{ \sum_{\substack{i=-k \\ i\neq 0}}^{k} w_k(i)|y(t+i) - \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}|^2 + ||\boldsymbol{\alpha}||_{\mathbf{G}}^2 \right\}$$
$$\widehat{\boldsymbol{\theta}}_{o,m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\mathrm{RLBF}}(t|\mathbf{H}), \tag{4.19}$$

are the holey regularized estimates. Note that

$$\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = \mathbf{S}_{o,m|k}^{-1}(t|\mathbf{H})\mathbf{p}_{o,m|k}(t), \tag{4.20}$$

where, since $w_k(0) = 1$,

$$\mathbf{S}_{o,m|k}(t|\mathbf{H}) = \mathbf{S}_{m|k}(t|\mathbf{H}) - \boldsymbol{\psi}_{m|k}(t,0)\boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,0),$$
$$\mathbf{p}_{o,m|k}(t) = \mathbf{p}_{m|k}(t) - y^*(t)\boldsymbol{\psi}_{m|k}(t,0). \tag{4.21}$$

One can apply the well-known Sherman-Morrison formula [28] to (4.20), which leads to

$$\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = \left[\mathbf{S}_{m|k}^{-1}(t|\mathbf{H}) + \frac{\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\boldsymbol{\psi}_{m|k}(t,0)\boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,0)\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})}{1 - \boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,0)\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\boldsymbol{\psi}_{m|k}(t,0)}\right] [\mathbf{p}_{m|k}(t) - y^*(t)\boldsymbol{\psi}_{m|k}(t,0)]$$

$$= \widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) + \frac{\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\boldsymbol{\psi}_{m|k}(t,0)\boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,0)}{1 - q_{m|k}(t|\mathbf{H})}\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})$$

$$- \frac{\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\boldsymbol{\psi}_{m|k}(t,0)}{1 - q_{m|k}(t|\mathbf{H})}y^*(t),$$

$$(4.22)$$

where $q_{m|k}(t|\mathbf{H}) = \boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,0)\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\boldsymbol{\psi}_{m|k}(t,0)$.

Note that $\left[\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right]^{\mathrm{H}}\boldsymbol{\psi}_{m|k}(t,0) = \left[\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right]^{\mathrm{H}}\boldsymbol{\varphi}(t)$, hence

$$\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = \widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) - \frac{\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\boldsymbol{\psi}_{m|k}(t,0)}{1 - q_{m|k}(t|\mathbf{H})}\left[\varepsilon_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right]^*, \quad (4.23)$$

where $\varepsilon_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = y(t) - \left[\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right]^{\mathrm{H}}\boldsymbol{\varphi}(t)$. Finally, one obtains

$$\varepsilon_{o,m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) = y(t) - \left[\widehat{\boldsymbol{\alpha}}_{o,m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right]^{\mathrm{H}}\boldsymbol{\psi}_{m|k}(t,0) = y(t) - \left[\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right]^{\mathrm{H}}\boldsymbol{\psi}_{m|k}(t,0)$$

$$+ \frac{q_{m|k}(t|\mathbf{H})}{1 - q_{m|k}(t|\mathbf{H})}\varepsilon_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) \quad (4.24)$$

$$= \frac{\varepsilon_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})}{1 - q_{m|k}(t|\mathbf{H})}.$$

The formula (4.24) allows one to evaluate (4.18) without computing holey estimates of system parameters. However, the formula (4.24) is still computationally demanding because evaluation of $q_{m|k}(t|\mathbf{H})$ requires inversion of matrix $\mathbf{S}_{m|k}(t|\mathbf{H})$. For the sufficiently large value of k, the following approximation can be used

$$\mathbf{S}_{m|k}^{-1}(t|\mathbf{H}) \cong [\boldsymbol{\Phi} \otimes \mathbf{I}_m + \mathbf{G}(\mathbf{H})]^{-1}. \quad (4.25)$$

One can use this approximation to compute the approximate formula

$$\widehat{q}_{m|k}(t|\mathbf{H}) = \boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,0)[\boldsymbol{\Phi} \otimes \mathbf{I}_m + \mathbf{G}(\mathbf{H})]^{-1}\boldsymbol{\psi}_{m|k}(t,0). \quad (4.26)$$

One can also replace $\widehat{q}_{m|k}(t|\mathbf{H})$ by its expected value

$$\overline{q}_{m|k}(\mathbf{H}) = \mathrm{E}[\widehat{q}_{m|k}(t|\mathbf{H})] = \mathrm{Tr}\{[\boldsymbol{\Phi} \otimes \mathbf{I}_m + \mathbf{G}]^{-1}\boldsymbol{\Phi} \otimes \mathbf{I}_m\}. \quad (4.27)$$

Under the assumption (A3.1), the above formulae can be simplified. Firstly,

$$\overline{q}_{m|k}(\mathbf{H}) = \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathrm{Tr}\{[\mathbf{I}_n + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{H}]^{-1}\}. \quad (4.28)$$

Denote by $\lambda_j(\mathbf{H})$, $i = 1, \ldots, n$ the eigenvalues of a matrix $\mathbf{H}$, then

$$\overline{q}_{m|k}(\mathbf{H}) = \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\sum_{j=1}^{n}\frac{1}{1 + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\lambda_j(\mathbf{H})}. \quad (4.29)$$

To simplify the formula (4.26), one needs additional lemma. First, note that the regularization matrix can be put down in the form

$$\mathbf{G}(\mathbf{H}) = \mathbf{F}_{m|k}^{\mathrm{H}}(0)\mathbf{H}\mathbf{F}_{m|k}(0) = \mathbf{F}_{m|k}^{\mathrm{H}}(0)\mathbf{D}^{\mathrm{H}}\mathbf{D}\mathbf{F}_{m|k}(0) = \mathbf{B}^{\mathrm{H}}\mathbf{B},$$

where $\mathbf{B} = \mathbf{D}\mathbf{F}_{m|k}(0)$ is an $l \times mn$ matrix and $\mathbf{D}$ is an $l \times n$ matrix, $l \geq n$. The second transition follows from the fact that every Hermitian, positive semidefinite matrix $\mathbf{H}$ can be expressed in the form $\mathbf{H} = \mathbf{D}^{\mathrm{H}}\mathbf{D}$, where $\mathbf{D}$ is an $l \times n$ matrix, $l \geq n$. The following lemma can be proven

Lemma 4.1

$$[\mathbf{I}_{mn}+\mu\mathbf{B}^{\mathrm{H}}\mathbf{B}]^{-1} = \mathbf{I}_{mn}-\mathbf{I}_n\otimes\left[\frac{\mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\mathrm{H}}(0)}{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)}\right]+[\mathbf{I}_n+\mu\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{H}]^{-1}\otimes\left[\frac{\mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\mathrm{H}}(0)}{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)}\right],$$

<u>Proof of Lemma 4.1</u>
Using the matrix inversion lemma [28] and the properties of the Kronecker product, one obtains

$$[\mathbf{I}_{mn} + \mu\mathbf{B}^{\mathrm{H}}\mathbf{B}]^{-1} = \mathbf{I}_{mn} - \mu[\mathbf{D}^{\mathrm{H}} \otimes \mathbf{f}_{m|k}(0)]\{\mathbf{I}_l + \mu[\mathbf{D} \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(0)][\mathbf{D}^{\mathrm{H}} \otimes \mathbf{f}_{m|k}(0)]\}^{-1}[\mathbf{D} \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(0)]$$

$$= \mathbf{I}_{mn} - \mu\mathbf{D}^{\mathrm{H}}[\mathbf{I}_l + \mu\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{D}\mathbf{D}^{\mathrm{H}}]^{-1}\mathbf{D} \otimes \mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\mathrm{H}}(0),$$
$$(4.30)$$

Note that one can use again the matrix inversion lemma, which leads to

$$[\mathbf{I}_n + \mu\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{D}^{\mathrm{H}}\mathbf{D}]^{-1} = \mathbf{I}_n - \mu\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{D}^{\mathrm{H}}[\mathbf{I}_l + \mu\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{D}\mathbf{D}^{\mathrm{H}}]^{-1}\mathbf{D}$$

$$\iff \mu\mathbf{D}^{\mathrm{H}}[\mathbf{I}_l + \mu\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{D}\mathbf{D}^{\mathrm{H}}]^{-1}\mathbf{D}$$
$$(4.31)$$
$$= \frac{1}{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)}\left\{\mathbf{I}_n - [\mathbf{I}_n + \mu\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{D}^{\mathrm{H}}\mathbf{D}]^{-1}\right\}.$$

Combining equations (4.30) and (4.31), and noting that $\mathbf{D}^{\mathrm{H}}\mathbf{D} = \mathbf{H}$, one obtains the result presented in the Lemma 4.1. ∎.

Using this result, one can replace (4.26) with

$$\widehat{q}_{m|k}(t|\mathbf{H}) = \boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,0)\boldsymbol{\psi}_{m|k}(t,0) - [\boldsymbol{\varphi}^{\mathrm{H}}(t) \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(0)]\left\{\mathbf{I}_n \otimes \left[\frac{\mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\mathrm{H}}(0)}{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)}\right]\right\}[\boldsymbol{\varphi}(t) \otimes \mathbf{f}_{m|k}(0)]$$

$$+ [\boldsymbol{\varphi}^{\mathrm{H}}(t) \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(0)]\left\{[\mathbf{I}_n + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{H}]^{-1} \otimes \left[\frac{\mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\mathrm{H}}(0)}{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)}\right]\right\}[\boldsymbol{\varphi}(t) \otimes \mathbf{f}_{m|k}(0)]$$

$$= \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\boldsymbol{\varphi}^{\mathrm{H}}(t)[\mathbf{I}_n + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{H}]^{-1}\boldsymbol{\varphi}(t).$$
$$(4.32)$$

4.3.4 Hyperparameter optimization based on the empirical Bayes approach

When the weighting sequence is rectangular $w_k(i) \equiv 1$, $i \in I_k$ and the signals are Gaussian and circular, the RLBF estimator has the Bayesian interpretation. Note that minimization of (4.17) is equivalent to maximization of the quantity

$$\exp\left\{-\frac{1}{\sigma_e^2}\sum_{i=-k}^{k}|y(t+i) - \boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\psi}_{m|k}(t,i)|^2\right\} \times \exp\left\{-\frac{1}{\sigma_e^2}\|\boldsymbol{\alpha}\|_{\mathbf{G(H)}}^2\right\}. \tag{4.33}$$

Hence, the regularization matrix $\mathbf{H}$ can be chosen using the type II maximum likelihood method [3] recently referred to as the empirical Bayes approach [16]. The first part of (4.33) can be associated with the conditional data distribution $p(\mathcal{Y}(t)|\boldsymbol{\alpha}, \mathbf{H})$, where $\mathcal{Y}(t) = \{y(t+i),\ i \in I_k\}$, and the second term can be attributed to the prior distribution of the basis function coefficients $\boldsymbol{\alpha}$: $\pi(\boldsymbol{\alpha}|\mathbf{H})$. The posterior distribution of $\boldsymbol{\alpha}$ can be seen as

$$p(\boldsymbol{\alpha}|\mathcal{Y}(t), \mathbf{H}) \propto p(\mathcal{Y}(t)|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}|\mathbf{H}), \tag{4.34}$$

Using properties of the probability density function of circular complex normal random variables [1], [110], one obtains

$$p(\mathcal{Y}(t)|\boldsymbol{\alpha}) = \frac{1}{(\pi\sigma_e^2)^K}\exp\left\{-\frac{1}{\sigma_e^2}\sum_{i=-k}^{k}|y(t+i) - \boldsymbol{\alpha}^{\mathrm{H}}\boldsymbol{\psi}_{m|k}(t,i)|^2\right\}, \tag{4.35}$$

and

$$\pi(\boldsymbol{\alpha}|\mathbf{H}) = \det_+\left[\frac{1}{\pi\sigma_e^2}\mathbf{G}(\mathbf{H})\right]\exp\left\{-\frac{1}{\sigma_e^2}\boldsymbol{\alpha}^{\mathrm{H}}\mathbf{G}(\mathbf{H})\boldsymbol{\alpha}\right\}, \tag{4.36}$$

where $\det_+(\mathbf{X})$ denotes pseodudeterminant of the matrix $\mathbf{X}$ - the product of all nonzero eigenvalues of $\mathbf{X}$. The above results from the following theorem stemming from the properties of degenerate normal distribution of multivariate real varaibles [86], [110].

Theorem 4.1 Let $\mathbf{z} = \mathbf{x} + \mathrm{i}\mathbf{y}$, $\mathbf{x} \sim \mathcal{N}(0, \mathbf{P})$, $\mathbf{y} \sim \mathcal{N}(0, \mathbf{P})$ be a zero-mean degenerate circular normal random variable, with a covariance matrix $\mathbf{Z} = \mathrm{E}[\mathbf{z}\mathbf{z}^{\mathrm{H}}]$ which is singular, namely $\mathrm{rank}(\mathbf{Z}) = r < \dim(\mathbf{Z}) = n$. The probability density function of such a random variable is given by

$$p_Z(\mathbf{z}) = \frac{1}{\pi^r\det_+(\mathbf{Z})}\exp\{-\mathbf{z}^{\mathrm{H}}\mathbf{Z}^+\mathbf{z}\},$$

where $\mathbf{Z}^+$ denotes the Moore-Penrose pseudoinverse.

Since $\mathbf{G}(\mathbf{H}) = [\mathbf{D}^{\mathrm{H}} \otimes \mathbf{f}_{m|k}(0)][\mathbf{D} \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(0)] = \mathbf{H} \otimes \mathbf{f}_{m|k}(0)\mathbf{f}_{m|k}^{\mathrm{H}}(0)$, the formula (4.36) becomes

$$\pi(\boldsymbol{\alpha}|\mathbf{H}) = \left[\frac{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)}{\pi\sigma_e^2}\right]^n\det(\mathbf{H})\exp\left\{-\frac{1}{\sigma_e^2}\boldsymbol{\alpha}^{\mathrm{H}}\mathbf{G}(\mathbf{H})\boldsymbol{\alpha}\right\}. \tag{4.37}$$

The likelihood function $L(\mathbf{H}, \sigma_e^2, t)$ can be computed as

$$\begin{aligned}
L(\mathbf{H}, \sigma_e^2, t) &= \int_{\mathbb{C}^{mn}} p(\mathcal{Y}(t)|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}|\mathbf{H})d\boldsymbol{\alpha} \\
&= \frac{[\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)]^n\det(\mathbf{H})}{(\pi\sigma_e^2)^{K+n}} \times \\
&\quad \times \int_{\mathbb{C}^{mn}}\exp\left\{-\frac{1}{\sigma_e^2}\left[\boldsymbol{\alpha}^{\mathrm{H}}\mathbf{S}_{m|k}(t|\mathbf{H})\boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathrm{H}}\mathbf{p}_{m|k}(t) - \mathbf{p}_{m|k}^{\mathrm{H}}(t)\boldsymbol{\alpha} + \sum_{i=-k}^{k}|y(t+i)|^2\right]\right\}d\boldsymbol{\alpha}.
\end{aligned} \tag{4.38}$$

One can compute this integral using the matrix version of the technique known as completing the square

$$\begin{aligned}
\mathbf{x}^{\mathrm{H}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{H}}\mathbf{x} + \mathbf{x}^{\mathrm{H}}\mathbf{b} + c &= \mathbf{x}^{\mathrm{H}}\mathbf{A}\mathbf{x} + \mathbf{b}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{H}}\mathbf{b} + c + \mathbf{b}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{b} - \mathbf{b}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{b} \\
&= (\mathbf{x}^{\mathrm{H}} + \mathbf{b}^{\mathrm{H}}\mathbf{A}^{-1})\mathbf{A}\mathbf{x} + (\mathbf{x}^{\mathrm{H}} + \mathbf{b}^{\mathrm{H}}\mathbf{A}^{-1})\mathbf{b} + c - \mathbf{b}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{b} \\
&= (\mathbf{x}^{\mathrm{H}} + \mathbf{b}^{\mathrm{H}}\mathbf{A}^{-1})(\mathbf{A}\mathbf{x} + \mathbf{b}) + c - \mathbf{b}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{b} \\
&= (\mathbf{x} + \mathbf{A}^{-1}\mathbf{b})^{\mathrm{H}}\mathbf{A}(\mathbf{x} + \mathbf{A}^{-1}\mathbf{b}) + c - \mathbf{b}^{\mathrm{H}}\mathbf{A}^{-1}\mathbf{b},
\end{aligned} \tag{4.39}$$

which can be applied for any complex-valued $\mathbf{b}$, c, and any hermitian positive definite matrix $\mathbf{A}$ (provided that the dimensions match). Using this technique, one obtains

$$\begin{aligned}
L(\mathbf{H}, \sigma_e^2, t) &= \frac{[\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)]^n\det(\mathbf{H})}{(\pi\sigma_e^2)^{K+n}} \times \\
&\quad \times \int_{\mathbb{C}^{mn}}\exp\left\{-\frac{1}{\sigma_e^2}[\boldsymbol{\alpha} - \mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\mathbf{p}_{m|k}(t)]^{\mathrm{H}}\mathbf{S}_{m|k}(t|\mathbf{H})[\boldsymbol{\alpha} - \mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\mathbf{p}_{m|k}(t)]\right. \\
&\quad\quad \left. -\frac{1}{\sigma_e^2}\sum_{i=-k}^{k}|y(t+i)|^2 + \frac{1}{\sigma_e^2}\mathbf{p}_{m|k}^{\mathrm{H}}(t)\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\mathbf{p}_{m|k}(t)\right\}d\boldsymbol{\alpha} \\
&= \frac{[\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)]^n\det(\mathbf{H})}{(\pi\sigma_e^2)^N\det[\mathbf{S}_{m|k}(t|\mathbf{H})]}\exp\left\{-\frac{1}{\sigma_e^2}\left[\sum_{i=-k}^{k}|y(t+i)|^2 - \mathbf{p}_{m|k}^{\mathrm{H}}(t)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right]\right\},
\end{aligned} \tag{4.40}$$

where $N = K + n - nm$. The last transition follows from (4.3).

In practice, it is more convenient to work with the minus log-likelihood function

$$\mathcal{L}(\mathbf{H}, \sigma_e^2, t) = -\log L(\mathbf{H}, \sigma_e^2, t) = N \log \sigma_e^2 - \log[\det(\mathbf{H})] + \log\{\det[\mathbf{S}_{m|k}(t|\mathbf{H})]\}$$
$$+ \frac{1}{\sigma_e^2}\left\{\sum_{i=-k}^{k} |y(t+i)|^2 - \mathbf{p}_{m|k}^{\mathrm{H}}(t)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right\} + C, \quad C \in \mathbb{R}. \tag{4.41}$$

One can now find the maximum likelihood estimate of σ_e^2 by solving

$$\frac{\partial \mathcal{L}}{\partial \sigma_e^2}(\mathbf{H}, \sigma_e^2, t) = \frac{N}{\sigma_e^2} - \frac{1}{\sigma_e^4}\left\{\sum_{i=-k}^{k} |y(t+i)|^2 - \mathbf{p}_{m|k}^{\mathrm{H}}(t)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right\} = 0, \tag{4.42}$$

from the equation above one obtains

$$\widehat{\sigma}_e^2(t|\mathbf{H}) = \frac{1}{N}\left\{\sum_{i=-k}^{k} |y(t+i)|^2 - \mathbf{p}_{m|k}^{\mathrm{H}}(t)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right\}. \tag{4.43}$$

As a consequence, one can find locally the best estimator by choosing $\mathbf{H}$ which minimizes the following function

$$\widehat{\mathcal{L}}(\mathbf{H}, t) = N \log \widehat{\sigma}_e^2(t|\mathbf{H}) - \log[\det(\mathbf{H})] + \log\{\det[\mathbf{S}_{m|k}(t|\mathbf{H})]\} + C_1, \quad C_1 \in \mathbb{R} \tag{4.44}$$

In practice, optimization can be done by means of parallel estimation, similar to the approach based on leave-one-out cross-validation.

Note that, one can again replace $\mathbf{S}_{m|k}(t|\mathbf{H})$ with its expected value $\mathbf{\Phi} \otimes \mathbf{I}_m + \mathbf{G}(\mathbf{H})$ (which is justified for k sufficiently large). Under the assumption (A3.2) one obtains

$$\det[\mathbf{S}_{m|k}(t|\mathbf{H})] \cong \det\{\mathbf{I}_{mn} + [\mathbf{D}^{\mathrm{H}} \otimes \mathbf{f}_{m|k}(0)][\mathbf{D} \otimes \mathbf{f}_{m|k}^{\mathrm{H}}(0)]\} = \det[\mathbf{I}_l + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{D}\mathbf{D}^{\mathrm{H}}]$$
$$= \prod_{j=1}^{n}[1 + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\lambda_j(\mathbf{H})], \tag{4.45}$$

where $\lambda_j(\mathbf{H})$, $j = 1, \ldots, n$ denotes the i-th eigenvalue of the matrix $\mathbf{H}$. The penultimate transition follows from the so-called Weistein-Aronszajn identity [98].

Remark

In the real-valued case the expression for the probability density function of a multivariate Gaussian random variable is different, hence the likelihood function $L(\mathbf{H}, \sigma_e^2, t)$ can be expressed as

$$L(\mathbf{H}, \sigma_e^2, t) = \sqrt{\frac{[\mathbf{f}_{m|k}^{\mathrm{T}}(0)\mathbf{f}_{m|k}(0)]^n \det(\mathbf{H})}{(2\pi\sigma_e^2)^N \det[\mathbf{S}_{m|k}(t|\mathbf{H})]}} \exp\left\{-\frac{1}{2\sigma_e^2}\left[\sum_{i=-k}^{k} y^2(t+i) - \mathbf{p}_{m|k}^{\mathrm{T}}(t)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right]\right\}, \tag{4.46}$$

therefore it is typical to work with a -2 log-likelihood function

$$\mathcal{L}(\mathbf{H}, \sigma_e^2, t) = -2\log L(\mathbf{H}, t) = N \log \sigma_e^2 - \log[\det(\mathbf{H})] + \log\{\det[\mathbf{S}_{m|k}(t|\mathbf{H})]\}$$
$$+ \frac{1}{\sigma_e^2}\left\{\sum_{i=-k}^{k} y^2(t+i) - \mathbf{p}_{m|k}^{\mathrm{T}}(t)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right\} + D, \quad D \in \mathbb{R}, \tag{4.47}$$

one can again find the maximum likelihood estimate of the noise variance

$$\widehat{\sigma}_e^2(t|\mathbf{H}) = \frac{1}{N}\left\{\sum_{i=-k}^{k} y^2(t+i) - \mathbf{p}_{m|k}^{\mathrm{T}}(t)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H})\right\}, \tag{4.48}$$

and use it to obtain the formula for the function to minimize

$$\widehat{\mathcal{L}}(\mathbf{H}, t) = N \log \widehat{\sigma}_e^2(t|\mathbf{H}) - \log[\det(\mathbf{H})] + \log\{\det[\mathbf{S}_{m|k}(t|\mathbf{H})]\} + D_1, \quad D_1 \in \mathbb{R}. \tag{4.49}$$

4.4   Regularization matrix design based on full (statistical) knowledge about parameter changes

It will be shown here that with some statistical knowledge about the parameter changes, one can derive the formula for the optimal regularization matrix. Since here, we are interested in the optimal regularization matrix, there is no point in introducing the regularization constant μ, so we shall again adopt the problem formulation presented in (4.17).

To perform the derivation, one needs assumptions (A2.5), (A3.1), and one additional assumption.

> (A4.1) $\{\boldsymbol{\theta}(t)\}$ is a sequence, independent of $\{u(t)\}$ and $\{e(t)\}$, with known correlation matrix $\mathrm{E}[\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\mathrm{H}}(t)] = \mathbf{Q}_0 > 0$.

Define the MSE of RLBF estimates as

$$M_{m|k}^{\mathrm{RLBF}}(\mathbf{H}) = \mathrm{E}\left\{ \left[\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) - \boldsymbol{\theta}(t)\right]^{\mathrm{H}} \left[\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) - \boldsymbol{\theta}(t)\right] \right\}, \tag{4.50}$$

where the avereging in this section is not only over $\{u(t)\}$, and $\{e(t)\}$, but also over $\{\boldsymbol{\theta}(t)\}$. Now, the following theorem will be proven

> **Theorem 4.2** Under assumptions (A2.5), (A3.1) and (A4.1), the matrix that minimizes the MSE (4.50) takes the form
>
> $$\mathbf{H}_{m|k}^{\mathrm{opt}} = \frac{\sigma_e^2}{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)l_{m|k}^{\mathrm{LBF}}} \mathbf{Q}_0^{-1}. \tag{4.51}$$

Proof

The proof proceeds along the same lines as the reasoning presented in [72]. The only difference is that here the proof is given for complex-valued basis functions. First, note that the equation (4.1) can be expressed as (under (A2.1))

$$M_{m|k}^{\mathrm{RLBF}}(\mathbf{H}) = \mathrm{Tr}\left\{ \mathbf{F}_{m|k}(0)\mathrm{E}\{[\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) - \boldsymbol{\alpha}(t)][\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) - \boldsymbol{\alpha}(t)]^{\mathrm{H}}\}\mathbf{F}_{m|k}^{\mathrm{H}}(0) \right\}. \tag{4.52}$$

Moreover, it holds that

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) - \boldsymbol{\alpha}(t) &= \left[\mathbf{S}_{m|k}^{-1}(t|\mathbf{H}) \sum_{i=-k}^{k} w_k(i)\boldsymbol{\psi}_{m|k}(t,i)\boldsymbol{\psi}_{m|k}^{\mathrm{H}}(t,i) - \mathbf{I}_{mn}\right] \boldsymbol{\alpha}(t) + \mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\boldsymbol{\zeta}_{m|k}(t) \\
&= [\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\mathbf{P}_{m|k}(t) - \mathbf{I}_{mn}]\boldsymbol{\alpha}(t) + \mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\boldsymbol{\zeta}_{m|k}(t) \\
&= -\mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\mathbf{G}(\mathbf{H})\boldsymbol{\alpha}(t) + \mathbf{S}_{m|k}^{-1}(t|\mathbf{H})\boldsymbol{\zeta}_{m|k}(t)
\end{aligned} \tag{4.53}$$

where

$$\boldsymbol{\zeta}_{m|k}(t) = \sum_{i=-k}^{k} w_k(i)\boldsymbol{\psi}_{m|k}(t,i)e^*(t+i). \tag{4.54}$$

The last transition follows from the fact that $\mathbf{S}_{m|k}(t|\mathbf{H}) = \mathbf{P}_{m|k}(t) + \mathbf{G}(\mathbf{H})$.

Under the assumptions mentioned at the beginning of this section, the matrix $\mathbf{P}_{m|k}(t)$ converges to its expected value in a mean squared sense (as it was shown in 2.3.1), which justifies the following approximation, valid for sufficiently large values of k

$$\mathbf{S}_{m|k}^{-1}(t|\mathbf{H}) = [\mathbf{P}_{m|k}(t) + \mathbf{G}(\mathbf{H})]^{-1} \cong [\sigma_u^2\mathbf{I}_{mn} + \mathbf{G}(\mathbf{H})]^{-1} = [\sigma_u^2\mathbf{I}_{mn} + \mathbf{B}^{\mathrm{H}}\mathbf{B}]^{-1}. \tag{4.55}$$

The above approximation leads to

$$
\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) - \boldsymbol{\alpha}(t) \cong -[\mathbf{I}_{mn} + \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1}\widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}\boldsymbol{\alpha}(t) + \frac{1}{\sigma_u^2}[\mathbf{I}_{mn} + \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1}\boldsymbol{\zeta}_{m|k}(t)
$$
$$
= [\mathbf{I}_{mn} + \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}]^{-1}\left[\frac{\boldsymbol{\zeta}_{m|k}(t)}{\sigma_u^2} - \widetilde{\mathbf{B}}^{\mathrm{H}}\widetilde{\mathbf{B}}\boldsymbol{\alpha}(t)\right]
$$
(4.56)

where $\widetilde{\mathbf{B}}^{\mathrm{H}} = \mathbf{B}/\sigma_u$. It is easy to check that using the lemma 4.1. yields

$$
\mathbf{F}_{m|k}(0)[\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{RLBF}}(t|\mathbf{H}) - \boldsymbol{\alpha}(t)] \cong [\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}\left[\frac{1}{\sigma_u^2}\sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)\boldsymbol{\varphi}(t+i)e^*(t+i) - \widetilde{\mathbf{H}}\boldsymbol{\theta}(t)\right]
$$
$$
= [\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}\left[\frac{1}{\sigma_u^2}\boldsymbol{\xi}_{m|k}(t) - \widetilde{\mathbf{H}}\boldsymbol{\theta}(t)\right],
$$
(4.57)

where $\boldsymbol{\xi}_{m|k}(t) = \sum_{i=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)\boldsymbol{\varphi}(t+i)e^*(t+i)$, and $\widetilde{\mathbf{H}} = \frac{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)}{\sigma_u^2}\mathbf{H}$. Hence

$$
M_{m|k}^{\mathrm{RLBF}}(\mathbf{H}) \cong \mathrm{Tr}\left\{[\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}\mathrm{E}\left\{\left[\frac{1}{\sigma_u^2}\boldsymbol{\xi}_{m|k}(t) - \widetilde{\mathbf{H}}\boldsymbol{\theta}(t)\right]\left[\frac{1}{\sigma_u^2}\boldsymbol{\xi}_{m|k}(t) - \widetilde{\mathbf{H}}\boldsymbol{\theta}(t)\right]^{\mathrm{H}}\right\}[\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}\right\}
$$
$$
= \mathrm{Tr}\left\{[\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}\mathrm{E}\left[\frac{1}{\sigma_u^4}\boldsymbol{\xi}_{m|k}(t)\boldsymbol{\xi}_{m|k}^{\mathrm{H}}(t) - \frac{1}{\sigma_u^2}\boldsymbol{\xi}_{m|k}(t)\boldsymbol{\theta}^{\mathrm{H}}(t)\widetilde{\mathbf{H}} - \frac{1}{\sigma_u^2}\widetilde{\mathbf{H}}\boldsymbol{\theta}(t)\boldsymbol{\xi}_{m|k}^{\mathrm{H}}(t) + \right.\right.
$$
$$
\left.\left. + \widetilde{\mathbf{H}}\boldsymbol{\theta}(t)\boldsymbol{\theta}^{\mathrm{H}}(t)\widetilde{\mathbf{H}}\right][\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}\right\}.
$$
(4.58)

Note that

$$
\mathrm{E}[\boldsymbol{\xi}_{m|k}(t)\boldsymbol{\theta}^{\mathrm{H}}(t)]\widetilde{\mathbf{H}} = \widetilde{\mathbf{H}}\mathrm{E}[\boldsymbol{\theta}(t)\boldsymbol{\xi}_{m|k}^{\mathrm{H}}(t)] = 0,
$$
(4.59)

because noise is independent of the input signal. It also holds that

$$
\frac{1}{\sigma_u^4}\mathrm{E}[\boldsymbol{\xi}_{m|k}(t)\boldsymbol{\xi}_{m|k}^{\mathrm{H}}(t)] = \frac{1}{\sigma_u^4}\mathrm{E}\left\{\sum_{i=-k}^{k}\sum_{j=-k}^{k} h_{m|k}^{\mathrm{fLBF}}(i)[h_{m|k}^{\mathrm{fLBF}}(j)]^* e^*(t+i)e(t+j)\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+j)\right\}
$$
$$
= \frac{1}{\sigma_u^4}\sum_{i=-k}^{k}\mathrm{E}[|h_{m|k}^{\mathrm{fLBF}}(i)|^2|e(t+i)|^2\boldsymbol{\varphi}(t+i)\boldsymbol{\varphi}^{\mathrm{H}}(t+i)] = \frac{\sigma_e^2}{\sigma_u^2}\sum_{i=-k}^{k}|h_{m|k}^{\mathrm{fLBF}}(i)|^2\mathbf{I}_n.
$$
(4.60)

It means that

$$
M_{m|k}^{\mathrm{RLBF}}(\mathbf{H}) \cong \mathrm{Tr}\left\{[\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}\left[\frac{\sigma_e^2}{\sigma_u^2 l_{m|k}^{\mathrm{fLBF}}}\mathbf{I}_n + \widetilde{\mathbf{H}}\mathbf{Q}_0\widetilde{\mathbf{H}}\right][\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}\right\}
$$
$$
= a\mathrm{Tr}\left\{[\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}[\mathbf{I}_n + \widetilde{\mathbf{H}}\widetilde{\mathbf{Q}}_0\widetilde{\mathbf{H}}][\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}\right\},
$$
(4.61)

where

$$
a = \frac{\sigma_e^2}{\sigma_u^2 l_{m|k}^{\mathrm{fLBF}}}
$$
$$
\widetilde{\mathbf{Q}}_0 = \frac{1}{a}\mathbf{Q}_0.
$$
(4.62)

According to the theorem 1 from [72] the matrix which minimizes the $[\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}[\mathbf{I}_n + \widetilde{\mathbf{H}}\widetilde{\mathbf{Q}}_0\widetilde{\mathbf{H}}][\mathbf{I}_n + \widetilde{\mathbf{H}}]^{-1}$ in a matrix sense is given by

$$
\widetilde{\mathbf{H}} = \widetilde{\mathbf{Q}}_0^{-1}.
$$
(4.63)

Note also that $\mathbf{X} \leq \mathbf{Y} \iff \mathrm{Tr}(\mathbf{X}) \leq \mathrm{Tr}(\mathbf{Y})$, therefore it holds that the matrix minimizing the MSE is given by

$$
\mathbf{H}_{m|k}^{\mathrm{opt}} = \frac{\sigma_u^2}{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)}\widetilde{\mathbf{Q}}_0^{-1} = \frac{\sigma_e^2}{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)l_{m|k}^{\mathrm{LBF}}}\mathbf{Q}_0^{-1} \quad \text{q.e.d.}
$$
(4.64)

Remark 1

When parameter trajectories are mutually uncorrelated, the optimal regularization matrix is diagonal

$$\mathbf{H}_{m|k}^{\text{opt}} = \frac{\sigma_e^2}{\mathbf{f}_{m|k}^{\text{H}}(0)\mathbf{f}_{m|k}(0)l_{m|k}^{\text{LBF}}} \text{diag}\left\{ \frac{1}{\text{var}[\theta_1(t)]}, \dots, \frac{1}{\text{var}[\theta_n(t)]} \right\}. \tag{4.65}$$

Note that such a matrix is structurally very similar to the regularization matrix emerging in the reweighted regularization, described previously in the section about the formulation for the ℓ^1 penalty.

Remark 2

From (4.51) the conclusion can be drawn that the regularization has the most to offer when the measurement noise is strong (σ_e^2 is growing), and plays marginal role for high values of SNR and for growing k (because the equivalent length of the analysis window grows for growing k). This confirms the intuition that the regularization is of the utmost help in situations, in which the variance component of the MSE dominates.

4.5   Fast regularized local basis function (fRLBF) method

Just as the fLBF estimator can be seen as an efficient alternative for the LBF method, one can also construct the regularized version of the fLBF method - the fast regularized local basis function (fRLBF) algorithm, which hopefully improves the accuracy of the estimates and makes it similar to the accuracy of estimates provided by the RLBF method. Another advantage of the fRLBF method over the RLBF (other than the computational cost) is its flexibility. The fLBF method allows one to analyze and track each parameter trajectory separately, similarly, the fRLBF allows one to design regularization for each parameter separately. Besides, when developing the fRLBF method, the same considerations apply as in the case of the RLBF estimator. Hence, the first decision one should make, is what penalty to use when defining the fRLBF method. We will again start with the ℓ^2 penalty because it is more common and easier to work with.

4.5.1   Formulation for the ℓ^2 penalty

The formulation of the fRLBF estimator is very straightforward

$$\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t) = \arg\min_{\boldsymbol{\alpha}_j}\left\{ \sum_{i=-k}^{k} w_k(i)|\widetilde{\theta}_j(t+i) - \mathbf{f}_{m|k}^{\text{H}}(i)\boldsymbol{\alpha}_j|^2 + \mu_j||\boldsymbol{\alpha}_j||_{\mathbf{G}_j}^2 \right\}$$

$$\widehat{\theta}_{j,m|k}^{\text{fRLBF}}(t) = \mathbf{f}_{m|k}^{\text{H}}(0)\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t), \quad j = 1, \dots, n. \tag{4.66}$$

It is easy to show that

$$\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t) = (\mathbf{I}_m + \mu_j\mathbf{G}_j)^{-1}\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fLBF}}(t), \quad j = 1, \dots, n, \tag{4.67}$$

the regularized estimates of basis coefficients can be obtained by the linear transformation of the fLBF estimates of basis coefficients, and in the ridge regression case, when the matrix $\mathbf{G}_j = \mathbf{I}_m$, $j = 1, \dots, n$, the regularized estimates can be obtained by uniform "shrinking" of the fLBF estimates

$$\widehat{\theta}_{j,m|k}^{\text{fRLBF}}(t) = \frac{\widehat{\theta}_j^{\text{fLBF}}(t)}{1 + \mu_j}. \tag{4.68}$$

Obviously, one can also design the estimator in a similar way to the RLBF one, which can be useful, e.g. when one wants to apply the conclusions from the section about the regularization matrix design to the fRLBF estimator.

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fRLBF}}(t) = \arg\min_{\boldsymbol{\alpha}}\left\{ \sum_{i=-k}^{k} w_k(i)||\widetilde{\boldsymbol{\theta}}(t+i) - \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}||^2 + \mu||\boldsymbol{\alpha}||_{\mathbf{G}}^2 \right\}$$

$$\widehat{\boldsymbol{\theta}}_{m|k}^{\text{fRLBF}}(t) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fRLBF}}(t), \tag{4.69}$$

then

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fRLBF}}(t) = (\mathbf{I}_{mn} + \mu\mathbf{G})^{-1}\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t). \tag{4.70}$$

These two descriptions are equivalent when $\mathbf{G} = \text{bl diag}\{\mu_1\mathbf{G}_1, \ldots, \mu_n\mathbf{G}_n\}$ and $\mu = 1$.

4.5.2 Formulation for the ℓ^1 penalty

The formulation for the ℓ^1 penalty for each system parameter is defined as

$$\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t) = \underset{\boldsymbol{\alpha}_j}{\arg\min}\left\{ \sum_{i=-k}^{k} w_k(i)|\widetilde{\theta}_j(t+i) - \mathbf{f}_{m|k}^{\text{H}}(i)\boldsymbol{\alpha}_j|^2 + \mu_j||\boldsymbol{\alpha}_j||_1 \right\}$$
$$\widehat{\theta}_{j,m|k}^{\text{fRLBF}}(t) = \mathbf{f}_{m|k}^{\text{H}}(0)\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t), \quad j = 1, \ldots, n. \tag{4.71}$$

Unlike the RLBF case, there is a closed-form solution to this problem and it is given by [100]

$$\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t) = \text{sign}[\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t)] \odot [|\widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t)| - \mu_j], \quad j = 1, \ldots, n, \tag{4.72}$$

where both operations sign($\mathbf{x}$) and $|\mathbf{x}|$ are here applied elementwise - to every element of the vector of basis function coefficients, and $\odot$ denotes the elementwise multiplication (the Hadamard product).

This means that the ℓ^1 penalty with the fRLBF can be practically useful, however, it still might be hard to optimize the value of μ_j. Therefore, as in the RLBF case, one might want to use the reweighting technique to approximate the solution to the problem with the ℓ^1 penalty by the solution to the problem with the ℓ^2 penalty. Applying the reweighting technique boils down to solving the problem written down as

$$\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t,p) = \underset{\boldsymbol{\alpha}_j}{\arg\min}\left\{ \sum_{i=-k}^{k} w_k(i)|\widetilde{\theta}_j(t+i) - \mathbf{f}_{m|k}^{\text{H}}(i)\boldsymbol{\alpha}_j|^2 + \mu_j||\boldsymbol{\alpha}_j||_{\mathbf{G}_j(p)}^2 \right\}$$
$$\widehat{\theta}_{j,m|k}^{\text{fRLBF}}(t,p) = \mathbf{f}_{m|k}^{\text{H}}(0)\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t,p), \quad j = 1, \ldots, n, \, p \in \mathbb{N}_+ \tag{4.73}$$

where

$$\mathbf{G}_j(p) = \text{diag}\{|\widehat{\alpha}_{j,1m|k}^{\text{fLBF}}(t,p-1)|, \ldots, |\widehat{\alpha}_{j,mm|k}^{\text{fLBF}}(t,p-1)|\}^{-1}, \quad j = 1, \ldots, n, \, p \in \mathbb{N}_+,$$

and $\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t,0) = \widehat{\boldsymbol{\alpha}}_{m|k}^{\text{fLBF}}(t)$.

4.5.3 Hyperparameter optimization with the LOOCV-based approach

As in the RLBF case, one can use the LOOCV or empirical Bayes approach to find the regularization matrix that provides the estimates of best accuracy. Again, we will assume a slightly different problem formulation, in which the regularization constant is a part of the regularization matrix. The LOOCV criterion will be derived only from the decentralized description because it allows one to obtain final formulae in a concise form. More precisely, we will start with the following problem description

$$\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t|\mathbf{G}_j) = \underset{\boldsymbol{\alpha}_j}{\arg\min}\left\{ \sum_{i=-k}^{k} w_k(i)|\widetilde{\theta}_j(t+i) - \mathbf{f}_{m|k}^{\text{H}}(i)\boldsymbol{\alpha}_j|^2 + ||\boldsymbol{\alpha}_j||_{\mathbf{G}_j}^2 \right\}$$
$$\widehat{\theta}_{j,m|k}^{\text{fRLBF}}(t|\mathbf{G}_j) = \mathbf{f}_{m|k}^{\text{H}}(0)\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\text{fRLBF}}(t|\mathbf{G}_j), \quad j = 1, \ldots, n. \tag{4.74}$$

As it was explained before, in the cross-validation approach one allows several estimators equipped with different settings to run in parallel, and at each time instant, one chooses the one which minimizes the following cost function

$$J_j(t|\mathbf{G}_j) = \sum_{i=-L}^{L} |\varepsilon_{o,j}^{\text{fRLBF}}(t+i|,\mathbf{G}_j)|^2, \tag{4.75}$$

where $\varepsilon_{o,j}^{\mathrm{fRLBF}}(t|\mathbf{G}_j) = y(t) - [\widehat{\boldsymbol{\theta}}_{o,m|k}^{\mathrm{fRLBF}}(t|\mathbf{G}_j)]^{\mathrm{H}}\boldsymbol{\varphi}(t)$ the leave-one-out interpolation error, sometimes also called deleted residuals. They can be obtained as a result of estimation with an excluded central observation from the analysis window

$$\widehat{\boldsymbol{\alpha}}_{o,j}^{\mathrm{fRLBF}}(t|\mathbf{G}_j) = \arg\min_{\boldsymbol{\alpha}_j}\left\{\sum_{\substack{i=-k \\ i\neq 0}}^{k} w_k(i)|\widetilde{\theta}_j(t+i) - \mathbf{f}_{m|k}^{\mathrm{H}}(i)\boldsymbol{\alpha}_j|^2 + ||\boldsymbol{\alpha}_j||_{\mathbf{G}_j}^2\right\}$$

$$\widehat{\theta}_{o,j}^{\mathrm{fRLBF}}(t|\mathbf{G}_j) = \mathbf{f}_{m|k}^{\mathrm{H}}(0)\widehat{\boldsymbol{\alpha}}_{o,j}^{\mathrm{fRLBF}}(t|\mathbf{G}_j), \quad j = 1,\ldots,n. \tag{4.76}$$

Note that exactly the same derivation can be made as in (4.22) - (4.24) to find the formula for the interpolation errors $\varepsilon_{o,j}^{\mathrm{fRLBF}}(t|\mathbf{G}_j)$, one only needs to set $\mathbf{S}_{m|k}(t) = \mathbf{I}_m + \mathbf{G}_j$, $\boldsymbol{\psi}_{m|k}(t,0) = \mathbf{f}_{m|k}(0)$ and $y^*(t) = \widetilde{\theta}_j(t)$ to obtain

$$\widehat{\theta}_{o,j}^{\mathrm{fRLBF}}(t|\mathbf{G}_j) = \frac{1}{1 - \beta_{j,m|k}(\mathbf{G}_j)}[\widehat{\theta}_{j,m|k}^{\mathrm{fRLBF}}(t|\mathbf{G}_j) - \beta_{j,m|k}(\mathbf{G}_j)\widetilde{\theta}_j(t)] \tag{4.77}$$

where

$$\beta_{j,m|k}(\mathbf{G}_j) = \mathbf{f}_{m|k}^{\mathrm{H}}(0)(\mathbf{I}_m + \mathbf{G}_j)^{-1}\mathbf{f}_{m|k}(0), \tag{4.78}$$

and can be computed prior to estimation.

4.5.4   Adaptive choice based on the empirical Bayes approach

Centralized approach

As in the case of the RLBF case, to derive the formulae based on the empirical Bayes approach, we will assume the modified version of the centralized formulation (4.69), in which the regularization constant μ is a part of the regularization matrix $\mathbf{G}$

$$\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fRLBF}}(t|\mathbf{H}) = \arg\min_{\boldsymbol{\alpha}}\left\{\sum_{i=-k}^{k} w_k(i)||\widetilde{\boldsymbol{\theta}}(t+i) - \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}||^2 + ||\boldsymbol{\alpha}||_{\mathbf{G}(\mathbf{H})}^2\right\}$$

$$\widehat{\boldsymbol{\theta}}_{m|k}^{\mathrm{fRLBF}}(t|\mathbf{H}) = \mathbf{F}_{m|k}(0)\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fRLBF}}(t|\mathbf{H}). \tag{4.79}$$

When the weighting sequence is uniform $w_k(i) \equiv 1$, $i \in I_k$ and the signals are Gaussian and circular, one can use the Bayesian interpretation to optimize the value of $\mathbf{G}$. Note that minimizing the cost function from (4.69) is equivalent to maximizing the following expression

$$\exp\left\{-\frac{1}{\sigma_z^2}\sum_{i=-k}^{k}||\widetilde{\boldsymbol{\theta}}(t+i) - \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}||^2\right\} \times \exp\left\{-\frac{1}{\sigma_z^2}||\boldsymbol{\alpha}||_{\mathbf{G}(\mathbf{H})}^2\right\}. \tag{4.80}$$

The first term can be attributed to the conditional probability distribution of preestimates

$$p(\widetilde{\Theta}(t)|\boldsymbol{\alpha}) = \frac{1}{(\pi\sigma_z^2)^{Kn}}\exp\left\{-\frac{1}{\sigma_z^2}\sum_{i=-k}^{k}||\widetilde{\boldsymbol{\theta}}(t+i) - \mathbf{F}_{m|k}(i)\boldsymbol{\alpha}||^2\right\}, \tag{4.81}$$

where $\widetilde{\Theta}(t) = \{\widetilde{\boldsymbol{\theta}}(t+i), i \in I_k\}$. The second term corresponds to the prior distribution of basis function coefficients

$$\pi(\boldsymbol{\alpha}|\mathbf{H}) = \left[\frac{\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)}{\pi\sigma_z^2}\right]^n \det(\mathbf{H})\exp\left\{-\frac{1}{\sigma_z^2}\boldsymbol{\alpha}^{\mathrm{H}}\mathbf{G}(\mathbf{H})\boldsymbol{\alpha}\right\}. \tag{4.82}$$

The reasoning behind this formula is the same as behind (4.37).

Just as in the case of the RLBF estimator, one can compute the likelihood function as

$$
\begin{aligned}
L(\mathbf{H}, \sigma_z^2, t) &= \int_{\mathbb{C}^{mn}} p(\widetilde{\Theta}(t)|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}|\mathbf{H})d\boldsymbol{\alpha} \\
&= \frac{[\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)]^n \det(\mathbf{H})}{(\pi\sigma_z^2)^{Kn+n}} \times \\
&\times \int_{\mathbb{C}^{mn}} \exp\left\{-\frac{1}{\sigma_z^2}\left\{\boldsymbol{\alpha}^{\mathrm{H}}[\mathbf{I}_{mn}+\mathbf{G}(\mathbf{H})]\boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathrm{H}}\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fLBF}}(t) - [\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{H}}\boldsymbol{\alpha}\right.\right. \\
&\left.\left. + \sum_{i=-k}^{k}||\widetilde{\boldsymbol{\theta}}(t+i)||^2\right\}\right\}d\boldsymbol{\alpha} \\
&= \frac{[\mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)]^n \det(\mathbf{H})}{(\pi\sigma_z^2)^{N_2}\det[\mathbf{I}_{mn}+\mathbf{G}(\mathbf{H})]} \exp\left\{-\frac{1}{\sigma_z^2}\left\{\sum_{i=-k}^{k}||\widetilde{\boldsymbol{\theta}}(t+i)||^2 - [\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{H}}\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fRLBF}}(t|\mathbf{H})\right\}\right\},
\end{aligned}
\tag{4.83}
$$

where $N_2 = Kn + n - nm$. The minus log-likelihood function is given by

$$
\mathcal{L}(\mathbf{H}, \sigma_z^2, t) = -\log[\det(\mathbf{H})] + N_2 \log\sigma_z^2 + \log\{\det[\mathbf{I}_l + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{D}\mathbf{D}^{\mathrm{H}}]\}+
$$
$$
+ \frac{1}{\sigma_z^2}\left\{\sum_{i=-k}^{k}||\widetilde{\boldsymbol{\theta}}(t+i)||^2 - [\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{H}}\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fRLBF}}(t|\mathbf{H})\right\} + C_2, \quad C_2 \in \mathbb{R}.
\tag{4.84}
$$

Note that using (4.45), one can write

$$
\log\{\det[\mathbf{I}_l + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\mathbf{D}\mathbf{D}^{\mathrm{H}}]\} = \sum_{j=1}^{n}\log[1 + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\lambda_j(\mathbf{H})].
\tag{4.85}
$$

Just as in the RLBF case, one can develop the formula for the maximum likelihood estimate of the preestimates variance

$$
\widehat{\sigma}_z^2(t|\mathbf{H}) = \frac{1}{N_2}\left\{\sum_{i=-k}^{k}||\widetilde{\boldsymbol{\theta}}(t+i)||^2 - [\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{H}}\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fRLBF}}(t|\mathbf{H})\right\}.
\tag{4.86}
$$

Putting these facts together, one should minimize the function

$$
\widehat{\mathcal{L}}(\mathbf{H}, t) = -\log[\det(\mathbf{H})] + N_2 \log\widehat{\sigma}_z^2(t|\mathbf{H}) + \sum_{j=1}^{n}\log[1 + \mathbf{f}_{m|k}^{\mathrm{H}}(0)\mathbf{f}_{m|k}(0)\lambda_j(\mathbf{H})] + C_3, \quad C_3 \in \mathbb{R}.
\tag{4.87}
$$

Note that $\widehat{\sigma}_z^2(t|\mathbf{H})$ is the only quantity that needs to be evaluated at each time step.

Remark 1

The final formula for the real-valued signals were given in [79] and looks very similar to (4.87)

$$
\widehat{\mathcal{L}}(\mathbf{H}, t) = -\log[\det(\mathbf{H})] + N_2 \log\widehat{\sigma}_z^2(t|\mathbf{H}) + \sum_{j=1}^{n}\log[1 + \mathbf{f}_{m|k}^{\mathrm{T}}(0)\mathbf{f}_{m|k}(0)\lambda_j(\mathbf{H})] + D_2, \quad D_2 \in \mathbb{R},
\tag{4.88}
$$

where

$$
\widehat{\sigma}_z^2(t|\mathbf{H}) = \frac{1}{N_2}\left\{\sum_{i=-k}^{k}||\widetilde{\boldsymbol{\theta}}(t+i)||^2 - [\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_{m|k}^{\mathrm{fRLBF}}(t|\mathbf{H})\right\}.
\tag{4.89}
$$

Remark 2

In [79] the problem is defined differently, for a regularization matrix of a fixed structure one is looking for the optimal regularization constant μ (or μ_j). For such a formulation of a problem, one can find the closed-form formulas for the regularization constant, which minimizes the minus log-likelihood function. These formulas were given in [79].

Decentralized approach

One can also approach the optimization problem starting from the decentralized description (4.74) instead of (4.79). This leads to the conditional probability

$$p(\widetilde{\Theta}_j(t)|\boldsymbol{\alpha}_j) = \frac{1}{(\pi\sigma_{z_j}^2)^K} \exp\left\{ -\frac{1}{\sigma_{z_j}^2} \sum_{i=-k}^{k} ||\widetilde{\theta}_j(t+i) - \mathbf{f}_{m|k}^{\mathrm{H}}(i)\boldsymbol{\alpha}_j||^2 \right\}, \quad j = 1,\ldots,n, \qquad (4.90)$$

where $\widetilde{\Theta}_j(t) = \{\widetilde{\theta}_j(t+i), i \in I_k\}$, $j = 1,\ldots,n$, and

$$\pi(\boldsymbol{\alpha}_j|\mathbf{G}_j) = \frac{1}{(\pi\sigma_{z_j}^2)^m} \det(\mathbf{G}_j) \exp\left\{ -\frac{1}{\sigma_{z_j}^2} \boldsymbol{\alpha}_j^{\mathrm{H}} \mathbf{G}_j \boldsymbol{\alpha}_j \right\}, \quad j = 1,\ldots,n. \qquad (4.91)$$

The likelihood function is defined as

$$\begin{aligned}
L_j(\mathbf{G}_j, \sigma_{z_j}^2, t) &= \int_{\mathbb{C}^m} p(\widetilde{\Theta}_j(t)|\boldsymbol{\alpha}_j)\pi(\boldsymbol{\alpha}_j|\mathbf{G}_j)d\boldsymbol{\alpha}_j \\
&= \frac{\mu_j^m \det(\mathbf{G}_j)}{(\pi\sigma_{z_j}^2)^{K+m}} \int_{\mathbb{C}^m} \exp\left\{ -\frac{1}{\sigma_{z_j}^2}\left\{ \boldsymbol{\alpha}_j^{\mathrm{H}}(\mathbf{I}_m + \mathbf{G}_j)\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_j^{\mathrm{H}}\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fLBF}}(t) - [\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{H}}\boldsymbol{\alpha}_j \right.\right. \\
&\qquad\left.\left. + \sum_{i=-k}^{k} |\widetilde{\theta}_j(t+i)|^2 \right\}\right\}d^m\boldsymbol{\alpha}_j \\
&= \frac{\det(\mathbf{G}_j)}{(\pi\sigma_{z_j}^2)^K \det(\mathbf{I}_m + \mathbf{G}_j)} \times \\
&\quad \times \exp\left\{ -\frac{1}{\sigma_{z_j}^2}\left\{ \sum_{i=-k}^{k} |\widetilde{\theta}_j(t+i)|^2 - [\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{H}}\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fRLBF}}(t|\mathbf{G}_j) \right\}\right\}, \quad j = 1,\ldots,n.
\end{aligned}$$
$$(4.92)$$

The minus log-likelihood function is then defined as

$$\begin{aligned}
\mathcal{L}_j(\mathbf{G}_j, \sigma_{z_j}^2, t) &= -\log[\det(\mathbf{G}_j)] + K \log\sigma_{z_j}^2 + \log[\det(\mathbf{I}_m + \mathbf{G}_j)] \\
&\quad + \frac{1}{\sigma_{z_j}^2}\left\{ \sum_{i=-k}^{k} |\widetilde{\theta}_j(t+i)|^2 - [\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{H}}\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fRLBF}}(t|\mathbf{G}_j) \right\} + C_4, \quad C_4 \in \mathbb{R}, j = 1,\ldots,n.
\end{aligned}$$
$$(4.93)$$

One can now find the maximum likelihood estimate of the variance of preestimate noise as

$$\widehat{\sigma}_{z_j}^2(t|\mathbf{G}_j) = \frac{1}{K}\left\{ \sum_{i=-k}^{k} |\widetilde{\theta}_j(t+i)|^2 - [\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{H}}\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fRLBF}}(t|\mathbf{G}_j) \right\}, \qquad (4.94)$$

which leads to a function that should be minimized to find locally the best regularization matrix

$$\widehat{\mathcal{L}}_j(\mathbf{G}_j, t) = -\log[\det(\mathbf{G}_j)] + K \log\widehat{\sigma}_{z_j}^2(t|\mathbf{G}_j) + \log[\det(\mathbf{I}_m + \mathbf{G}_j)] + C_5, \quad C_5 \in \mathbb{R}, j = 1,\ldots,n. \qquad (4.95)$$

Remark 1

Just as in the previous examples, for real-valued signals, one typically uses the -2 log-likelihood function of a form

$$\widehat{\mathcal{L}}_j(\mathbf{G}_j, t) = -\log[\det(\mathbf{G}_j)] + K \log\widehat{\sigma}_{z_j}^2(t|\mathbf{G}_j) + \log[\det(\mathbf{I}_m + \mathbf{G}_j)] + D_3, \quad D_3 \in \mathbb{R}, j = 1,\ldots,n, \qquad (4.96)$$

where

$$\widehat{\sigma}_{z_j}^2(t|\mathbf{G}_j) = \frac{1}{K}\left\{ \sum_{i=-k}^{k} |\widetilde{\theta}_j(t+i)|^2 - [\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fLBF}}(t)]^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_{j,m|k}^{\mathrm{fRLBF}}(t|\mathbf{G}_j) \right\}. \qquad (4.97)$$

Remark 2

In [23] it was shown that one can also use the generalized cross-validation (GCV) to choose the value of the regularization constant μ_j and that there exists a closed-form solution to the GCV optimization problem. However, this approach will be skipped here because it was based on a special case when the weighting is uniform $w_k(i) \equiv 1, \ i \in I_k$ and the regularization matrix $\mathbf{G}_j = \mathbf{I}_m$.

Chapter 5

Computer simulations

5.1 Introduction

All results presented in this chapter were obtained for either real- or complex-valued simulated self-interference UWA channel, characterized by the equation

$$y(t) = \sum_{j=1}^{20} \theta_j^*(t) u(t - j + 1) + e(t).$$ (5.1)

The parameter trajectories are lowpass filtered white Gaussian noises, either real or complex. The fast parameter changes correspond to the cut-off angular frequency of $\omega_0 = 0.006\pi$, while to obtain the slow parameter changes the cut-off angular frequency $\omega_0 = 0.001\pi$ was used. The variance of system parameters was decreasing exponentially to reflect the power absorption loss present in real UWA communication systems

$$\text{var}[\theta_j(t)] = 0.69^{j-1}, \quad j = 1, \dots, 20.$$ (5.2)

The input signal was random white binary $u(t) = \pm 1$ for a real-valued system, and random circular binary $u(t) = \pm 1 \pm i$ for a complex-valued system. The noise $\{e(t)\}$ was either white Gaussian or circular white Gaussian, in the case of a complex-valued system. All of the results, presented in this chapter were averaged over one long data realization (10^5 samples). To avoid boundary problems, data generation started 1000 samples before $t = 1$ and ended 1000 samples after $t = 10^5$.

Fragments of one of the simulated trajectories are shown in figure 5.1.

Figure 5.2 depicts real and imaginary parts of a typical frozen impulse response of a simulated underwater system.

5.2 LBF

We shall start by giving a comparison of the accuracy of estimates obtained using the LBF method with different weighting sequences: rectangular, cosinusoidal, and Hann (see 2.6.5). To make this comparison fair, for each weighting sequence, and the number of basis functions m, the half-width of the analysis window k was chosen in a way that assures the euivalent number of observations $l_{m|k}^{\text{LBF}}$ for different algorithms is approximately the same. Because, the variance component of the MSE depends on this quantity, this corresponds to comparing bias components resulting from using different weighting. Table 5.1 shows the half-widths k of the analysis window, for different number of basis functions, different lengths K_R of the analysis window with rectangular weighting and different types of weighting.

The comparison for different speeds of parameter changes and different SNRs were presented in tables 5.2 and 5.3.

As can be seen from tables 5.2 and 5.3, applying the non-rectangular weighting tends to improve the identification results at a cost of slight increase in the processing delay.

An example of LBF estimates obtained for SNR equal to 10 dB, $m = 5$, $k = 200$, polynomial basis functions and rectangular weighting sequence are presented in figure 5.3. Figure 5.4 presents an example of an estimate of a frozen impulse response obtained for the settings listed above.
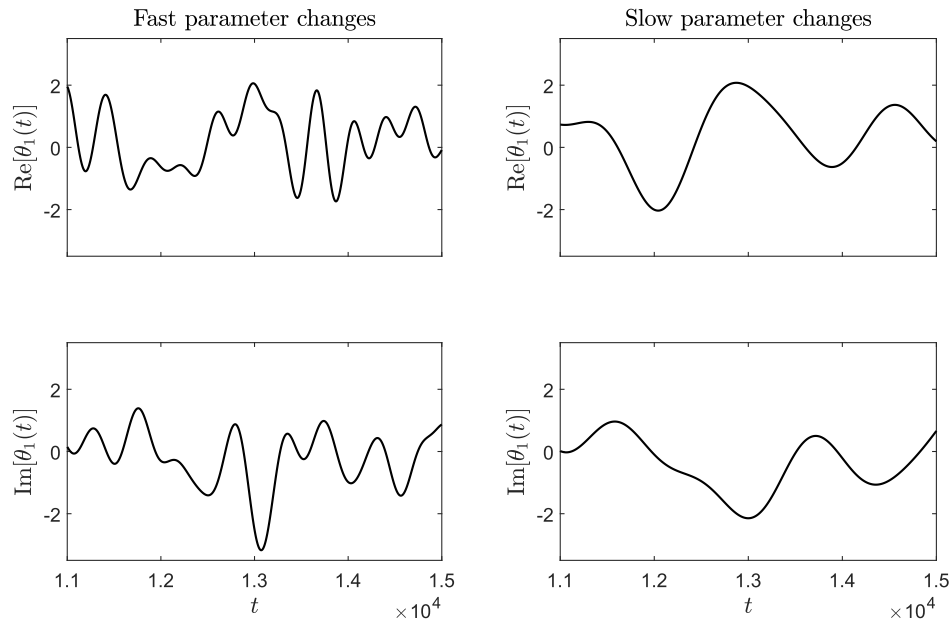
Figure 5.1: Real (upper plots) and imaginary (lower plots) parts of a typical simulated trajectory. On the left there is an example of fast parameter changes, on the right there is an example of slow parameter changes.
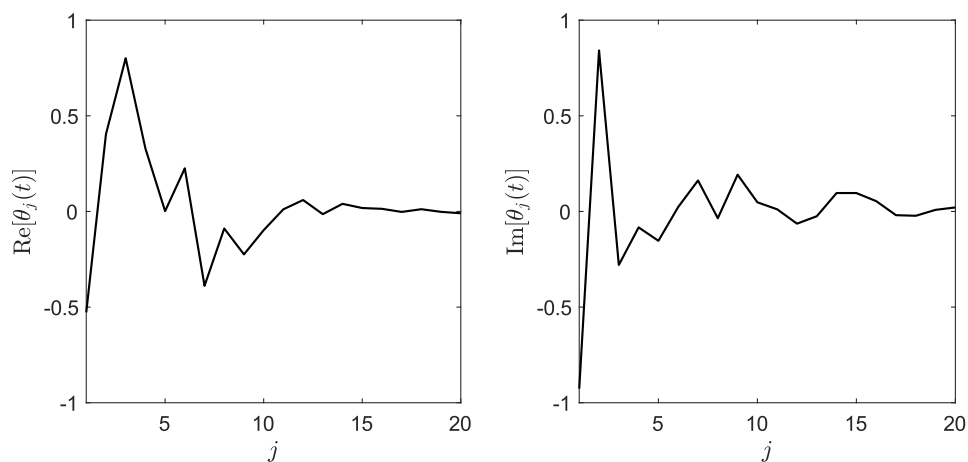
Figure 5.2: Real and imaginary parts of a typical frozen impulse response of a simulated underwater system.

Table 5.1: Value of k needed to obtain approximately the same equivalent number of observations $l_{m|k}^{\text{LBF}}$ as for the rectangular weighting of length K_R, for different weighting sequences and different numbers of basis functions.

	Cosinusoidal			Hann		
$K_R\backslash m$	1	3	5	1	3	5
201	124	113	109	151	130	121
401	247	226	218	301	259	242
801	494	452	436	601	516	482

Table 5.2: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for the fast parameter changes and for different weighting sequences.

SNR [dB]	$m \backslash k_R$	Rectangular			Cosinusoidal			Hann		
		201	401	801	201	401	801	201	401	801
10	1	-3.90	3.00	5.51	-5.02	1.77	4.85	-5.03	1.60	4.68
	3	-8.96	-7.14	3.75	-9.11	-8.06	2.62	-9.19	-8.05	2.23
	5	-6.32	**-10.26**	-1.91	-6.56	**-10.37**	-3.07	-6.68	**-10.41**	-3.33
30	1	-4.43	2.95	5.49	-5.70	1.71	4.83	-5.70	1.54	4.67
	3	-23.58	-8.56	3.71	-24.73	-9.86	2.56	-24.73	-9.83	2.16
	5	**-26.31**	-24.90	-2.19	**-26.56**	-25.79	-3.45	**-26.68**	-25.67	-3.72
50	1	-4.43	2.95	5.49	-5.71	1.71	4.83	-5.71	1.54	4.67
	3	-24.99	-8.57	3.71	-26.61	-9.88	2.56	-26.57	-9.85	2.16
	5	**-45.55**	-26.32	-2.19	**-45.89**	-27.56	-3.45	**-45.96**	-27.37	-3.73

Table 5.3: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for the slow parameter changes and for different weighting sequences.

| SNR [dB]	$m \backslash k_R$	Rectangular			Cosinusoidal			Hann		
		201	401	801	201	401	801	201	401	801
10	1	-11.63	-9.62	-1.66	-12.17	-10.73	-2.93	-12.25	-10.72	-2.96
	3	-9.08	-12.60	**-14.48**	-9.20	-12.66	**-14.85**	-9.28	-12.68	**-14.85**
	5	-6.34	-10.38	-13.77	-6.58	-10.47	-13.81	-6.70	-10.51	-13.82
30	1	-16.75	-10.66	-1.74	-18.55	-12.10	-3.03	-18.78	-12.09	-3.07
	3	-29.06	-31.48	-20.02	-29.19	-31.88	-21.43	-29.27	-31.92	-21.42
	5	-26.34	-30.38	**-33.44**	-26.58	-30.47	**-33.57**	-26.70	-30.51	**-33.57**
50	1	-16.86	-10.67	-1.74	-18.71	-12.12	-3.04	-18.95	-12.11	-3.07
	3	-47.71	-37.80	-20.14	-48.25	-39.57	-21.60	-48.35	-39.66	-21.59
	5	-46.34	**-50.36**	-44.29	-46.58	**-50.45**	-45.64	-46.70	**-50.49**	-45.63
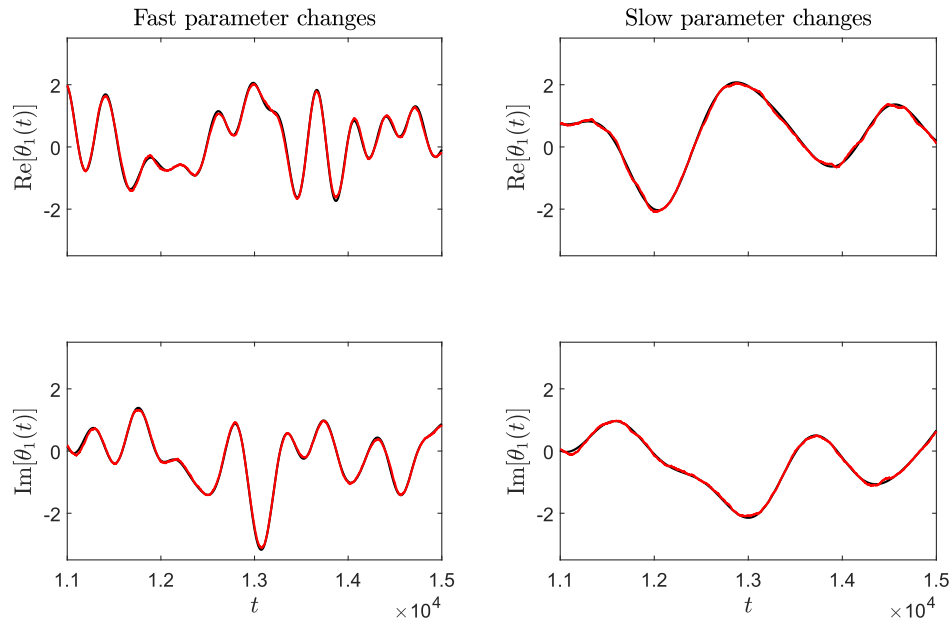
Figure 5.3: LBF estimates (red lines) obtained for SNR equal to 10 dB, $m = 5$, $k = 200$, polynomial basis functions and rectangular weighting sequence, superimposed on true parameter trajectories (black lines).

In general, using bell-shaped weighting sequences can improve the results [70], however, in some examples, like the one given here, the results are comparable. In tables 5.4 and 5.5 the MSE scores for different types of basis functions were presented. It is clear that in this experimental scenario, tested basis functions yield very similar results. This can be attributed to the fact that parameter trajectories can be only approximated by linear combinations of basis functions, and all of them have similar modeling capabilities.

Table 5.4: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for the fast parameter changes and different types of basis functions.

SNR [dB]	$m \backslash k$	Polynomial			Sinusoidal			Complex exponential			Prolate spheroidal		
		100	200	400	100	200	400	100	200	400	100	200	400
10	1	-3.90	3.00	5.51	-3.90	3.00	5.51	-3.90	3.00	5.51	-3.90	3.00	5.51
	3	-8.96	-7.14	3.75	-8.75	-8.26	3.48	-6.43	**-9.74**	2.10	-8.27	**-9.25**	3.18
	5	-6.32	**-10.26**	-1.91	-5.95	**-10.00**	-3.32	-3.87	-8.25	-9.55	-5.06	-9.17	-6.74
30	1	-4.43	2.95	5.49	-4.43	2.95	5.49	-4.43	2.95	5.49	-4.43	2.95	5.49
	3	-23.58	-8.56	3.71	-26.88	-10.37	3.43	-12.51	-15.04	2.01	-20.75	-12.44	3.12
	5	**-26.31**	-24.90	-2.19	-25.92	**-29.52**	-3.76	-13.22	**-18.56**	-13.05	-19.83	**-23.53**	-7.99
50	1	-4.43	2.95	5.49	-4.43	2.95	5.49	-4.43	2.95	5.49	-4.43	2.95	5.49
	3	-24.99	-8.57	3.71	-31.30	-10.40	3.43	-12.64	-15.14	2.01	-21.55	-12.49	3.12
	5	**-45.55**	-26.32	-2.19	**-44.05**	-39.03	-3.76	-20.35	**-28.51**	-19.66	-31.98	**-37.29**	-12.00

Tables 5.6, 5.7 show the comparison of the MSE scores for the LBF with polynomial basis functions and uniform weighting, and the LBF with KL functions, the number of which was chosen adaptively, according to (2.95). This table also presents the lowest theoretical MSE score, attainable if the KL functions could be chosen without any restrictions.

Tables 5.8, and 5.9 present the comparison of the MSE scores obtained for the LBF estimator with polynomial basis functions and uniform weighting, for the time-updated recursive least squares (TU-RLS) algorithm (which, until recently, was considered the state-of-the-art in UWA communication applications [19], [46]), and for the multi-resolution wavelets (MW) estimator, presented in [47], [106], with settings recommended in [47]. Because the MW algorithm was developed for the real-valued systems, these tables present results obtained for the real-valued version of the system described at the beginning of this chapter. The MW algorithm is also based on a localized identification method, however, unlike the LBF, it allows one to find the estimates for the entire
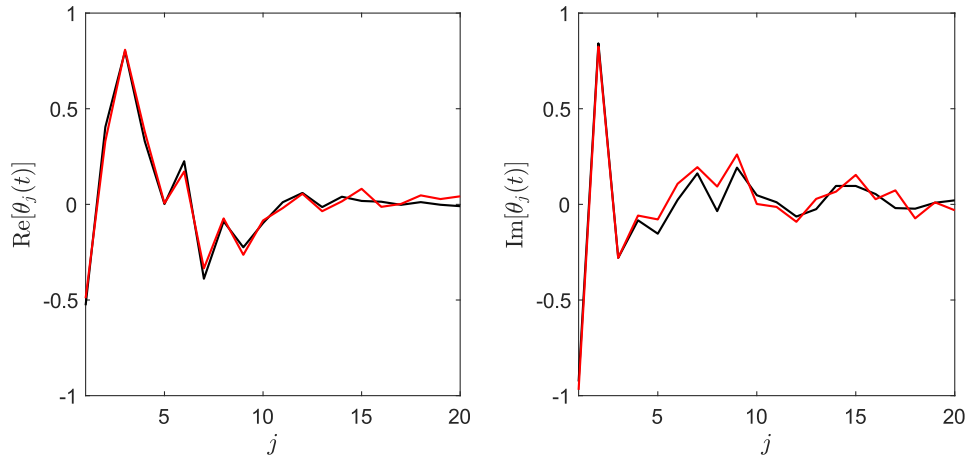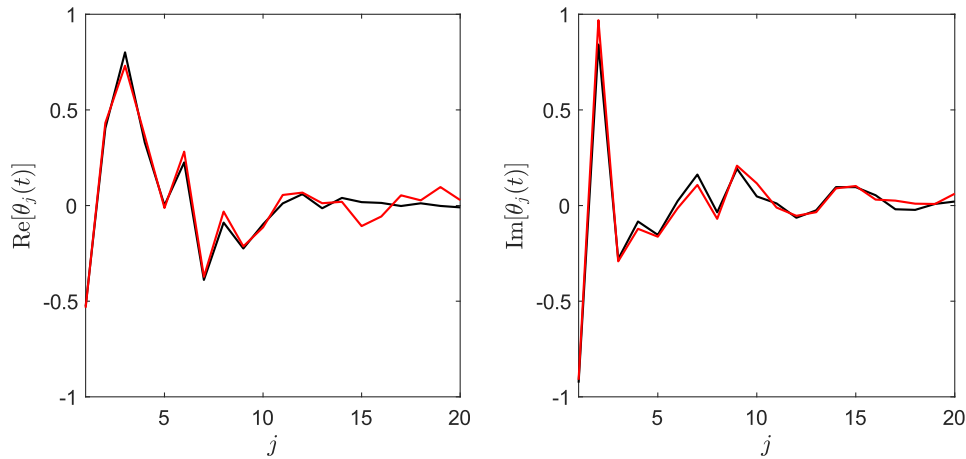
Figure 5.4: LBF estimate (red lines) of a system frozen impulse response obtained for SNR equal to 30 dB, $m = 5$, $k = 200$, polynomial basis functions and rectangular weighting sequence, superimposed on a true frozen impulse response (black lines).

Table 5.5: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for slow parameter changes and different types of basis functions.

SNR [dB]	$m\backslash k$	Polynomial			Sinusoidal			Complex exponential			Prolate spheroidal		
		100	200	400	100	200	400	100	200	400	100	200	400
10	1	-11.63	-9.62	-1.66	-11.63	-9.62	-1.66	-11.63	-9.62	-1.66	-11.63	-9.62	-1.66
	3	-9.08	-12.60	**-14.48**	-8.78	-12.30	**-15.17**	-7.40	-10.48	**-12.53**	-8.43	-11.85	**-15.07**
	5	-6.34	-10.38	-13.77	-5.96	-10.02	-13.42	-4.27	-8.35	-11.57	-5.16	-9.22	-12.61
30	1	-16.75	-10.66	-1.74	-16.75	-10.66	-1.74	-16.75	-10.66	-1.74	-16.75	-10.66	-1.74
	3	-29.06	-31.48	-20.02	-27.62	-30.03	-24.94	-19.49	-17.98	-16.62	-24.54	-24.39	-26.49
	5	-26.34	-30.38	**-33.44**	-25.94	-29.97	**-33.31**	-19.01	-19.70	**-20.61**	-23.13	-24.49	**-26.77**
50	1	-16.86	-10.67	-1.74	-16.86	-10.67	-1.74	-16.86	-10.67	-1.74	-16.86	-10.67	-1.74
	3	-47.71	-37.80	-20.14	-33.71	-33.75	-25.34	-20.19	-18.17	-16.67	-26.72	-25.18	-27.07
	5	-46.34	**-50.36**	-44.29	-44.51	-47.22	**-48.01**	-30.73	-30.45	**-31.42**	-40.99	-39.34	**-42.08**

Table 5.6: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for the fast parameter changes, obtained for polynomial and KL basis functions.

SNR [dB]	$m\backslash k$	Polynomial			KL			Lower bound		
		100	200	400	100	200	400	100	200	400
10	1	-3.90	3.00	5.51						
	3	-8.96	-7.14	3.75	-8.68	-10.15	**-11.22**	-10.43	-11.62	**-12.09**
	5	-6.32	**-10.26**	-1.91						
30	1	-4.43	2.95	5.49						
	3	-23.58	-8.56	3.71	-27.16	-29.57	**-30.40**	-29.60	-30.65	**-31.46**
	5	**-26.31**	-24.90	-2.19						
50	1	-4.43	2.95	5.49						
	3	-24.99	-8.57	3.71	-46.14	-48.35	**-49.99**	-48.16	-49.73	**-51.05**
	5	**-45.55**	-26.32	-2.19						

Table 5.7: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for slow parameter changes, obtained for polynomial and KL basis functions.

SNR [dB]	$m\backslash k$	Polynomial			KL			Lower bound		
		100	200	400	100.00	200.00	400.00	100	200	400
10	1	-11.63	-9.62	-1.66						
	3	-9.08	-12.60	**-14.48**	-11.60	-10.76	**-15.12**	-13.33	-14.24	**-15.60**
	5	-6.34	-10.38	-13.77						
30	1	-16.75	-10.66	-1.74						
	3	-29.06	-31.48	-20.02	-29.00	-32.25	**-33.53**	-30.68	-32.97	**-34.25**
	5	-26.34	-30.38	**-33.44**						
50	1	-16.86	-10.67	-1.74						
	3	-47.71	-37.80	-20.14	-48.76	-45.06	**-53.06**	-50.08	-52.08	**-53.87**
	5	-46.34	**-50.36**	-44.29						

analysis window. As mentioned in Chapter 2, such estimators typically provide estimates of the highest accuracy on the center of the analysis interval. Hence, to make the comparison fair, the overlap-add technique, described in [73], was applied. The last three columns show the result of applying techniques of adaptive selection of the number of basis functions m and the length of the analysis window $2k + 1$. The LOOCV 0 denotes the adaptive choice based on the exact formula (2.125), and LOOCV denotes the results obtained for its simplified version (2.127).

Tables below present also the TU-RLS results for the settings which guarantee the best achievable estimation accuracy.

Table 5.8: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for the fast parameter changes and different identification algorithms.

SNR [dB]	$m\backslash k$	100	200	400	FPE	LOOCV 0	LOOCV
10	1	-6.86	0.09	2.53			
	3	-11.86	-10.09	0.88	-9.27	-12.69	-9.31
	5	-9.27	**-13.19**	-4.87			
	TU-RLS	-7.90					
	MW	-7.09	-10.35	**-13.06**	X	X	X
30	1	-7.38	0.04	2.51			
	3	-26.45	-11.51	0.83	-29.26	-29.15	-29.26
	5	**-29.26**	-27.85	-5.15			
	TU-RLS	-13.83					
	MW	-18.18	-25.55	**-30.47**	X	X	X
50	1	-7.38	0.04	2.51			
	3	-27.88	-11.52	0.83	-48.49	-48.49	-48.49
	5	**-48.49**	-29.29	-5.16			
	TU-RLS	-14.15					
	MW	-25.84	-30.09	**-47.22**	X	X	X

The LBF algorithm outperforms the TU-RLS algorithm in terms of estimation accuracy, and yields results that are almost always better than the results provided by the MW algorithm. It is also worth noting that all adaptive algorithms yield estimates of accuracy comparable to the accuracy of the estimates provided by the best estimator incorporated in the parallel scheme, and substantially better than the results provided by most of the algorithms involved in the computations. The best results achievable for the MW algorithm are very similar to the best results yielded by the LBF algorithm, however, the MW estimator was working with more than 30 basis functions.

Table 5.9: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for slow parameter changes and different idnetification algorithms.

SNR [dB]	$m \backslash k$	100	200	400	FPE	LOOCV 0	LOOCV
10	1	-14.64	-12.68	-4.81			
	3	-12.03	-15.58	**-17.50**	-9.33	-16.21	-9.37
	5	-9.33	-13.35	-16.74			
	TU-RLS	-10.76					
	MW	-8.41	-11.79	**-13.85**	X	X	X
30	1	-19.78	-13.75	-4.90			
	3	-32.02	-34.50	-23.20	-29.33	-35.44	-29.37
	5	-29.33	-33.35	**-36.44**			
	TU-RLS	-20.55					
	MW	-21.14	-27.19	**-31.13**	X	X	X
50	1	-19.89	-13.76	-4.90			
	3	-50.73	-40.88	-23.33	-49.33	-52.98	-49.35
	5	-49.33	**-53.33**	-47.46			
	TU-RLS	-22.74					
	MW	-27.06	-32.34	**-47.97**	X	X	X

5.3 fLBF

In this part of the chapter, we demonstrate some results for the fLBF method. Firstly, we give the comparison between the LBF and fLBF without debiasing, with adaptive debiasing and debiasing using the nominal delay (in the table this is called "fixed debiasing"). The results are presented in tables 5.10 and 5.11. The comparison was carried out for polynomial basis functions and rectangular weighting sequence. The fLBF estimates were based on EWLS preestimates.

Table 5.10: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for the fast parameter changes. The comparison was obtained for LBF and fLBF algorithms with and without debiasing.

SNR [dB]	$m \backslash k$	LBF			fLBF			fLBF + adaptive debias			fLBF + fixed debias		
		100	200	400	100	200	400	100	200	400	100	200	400
10	1	-3.90	3.00	5.51	-4.87	2.75	5.41	-5.20	2.67	5.38	-4.99	2.75	5.41
	3	-8.96	-7.14	3.75	-7.78	-6.96	3.59	-8.13	-7.39	3.51	-8.12	-7.18	3.59
	5	-6.32	**-10.26**	-1.91	-6.05	**-8.67**	-1.94	-6.30	**-9.11**	-2.18	-6.28	**-9.08**	-1.99
30	1	-4.43	2.95	5.49	-5.84	2.67	5.39	-6.30	2.59	5.36	-5.99	2.67	5.39
	3	-23.58	-8.56	3.71	-16.18	-8.91	3.52	-21.88	-9.74	3.43	-19.52	-9.28	3.52
	5	**-26.31**	-24.90	-2.19	-15.64	**-16.37**	-2.36	-20.19	**-22.69**	-2.63	-18.51	**-19.94**	-2.41
50	1	-4.43	2.95	5.49	-5.85	2.67	5.39	-6.32	2.59	5.36	-6.01	2.67	5.39
	3	-24.99	-8.57	3.71	-16.44	-8.94	3.52	-23.02	-9.77	3.43	-20.13	-9.31	3.52
	5	**-45.55**	-26.32	-2.19	-16.00	**-16.59**	-2.36	-21.35	**-23.78**	-2.64	-19.26	**-20.47**	-2.41

As it can be seen from these tables, the fLBF algorithm can provide estimates of similar accuracy as the estimates produced by the LBF method, especially when one incorporates debiasing. Obviously, the MSE scores for the fLBF estimates are slightly lower, which is a price for a lower computational burden. However, it was also shown that this discrepancy becomes visible only when the number of parameters n grows [70], [78]. It is also worth noticing, that the results provided by the fLBF algorithm tend to get better quickly when one increases m and k [63], [78], which has a very slight impact on the computational costs, whereas increasing the values of these two parameters in the LBF method results in a substantially higher computational burden.

An example of fLBF estimates obtained for EWLS-based preestimates, SNR equal to 10 dB, $m = 5$, $k = 200$, polynomial basis functions and rectangular weighting sequence are shown in figures 5.3, 5.4.

Tables 5.12, and 5.13 show the comparison of the MSE scores for the fLBF algorithm based

Table 5.11: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for slow parameter changes. The comparison was obtained for LBF and fLBF algorithms with and without debiasing.

SNR [dB]	$m\backslash k$	LBF			fLBF			fLBF + adaptive debias			fLBF + fixed debias		
		100	200	400	100	200	400	100	200	400	100	200	400
10	1	-11.63	-9.62	-1.66	-11.51	-10.40	-1.88	-11.56	-10.58	-1.96	-11.59	-10.46	-1.89
	3	-9.08	-12.60	**-14.48**	-8.35	-11.24	**-13.33**	-8.38	-11.28	**-13.43**	-8.40	-11.32	**-13.46**
	5	-6.34	-10.38	-13.77	-6.47	-9.36	-12.27	-6.50	-9.39	-12.31	-6.50	-9.41	-12.37
30	1	-16.75	-10.66	-1.74	-21.74	-12.38	-2.01	-23.57	-12.79	-2.09	-22.79	-12.48	-2.01
	3	-29.06	-31.48	-20.02	-24.12	-25.28	-20.18	-27.23	-30.09	-21.57	-26.25	-28.34	-20.88
	5	-26.34	-30.38	**-33.44**	-22.98	-24.65	**-25.63**	-25.11	-28.40	**-31.24**	-24.51	-27.15	**-29.08**
50	1	-16.86	-10.67	-1.74	-22.19	-12.41	-2.01	-24.35	-12.83	-2.09	-23.39	-12.51	-2.02
	3	-47.71	-37.80	-20.14	-26.10	-26.50	-20.37	-33.78	-36.92	-21.86	-30.26	-31.34	-21.11
	5	-46.34	**-50.36**	-44.29	-25.47	-26.37	**-26.63**	-30.84	-35.67	**-38.78**	-28.77	-30.98	**-31.78**
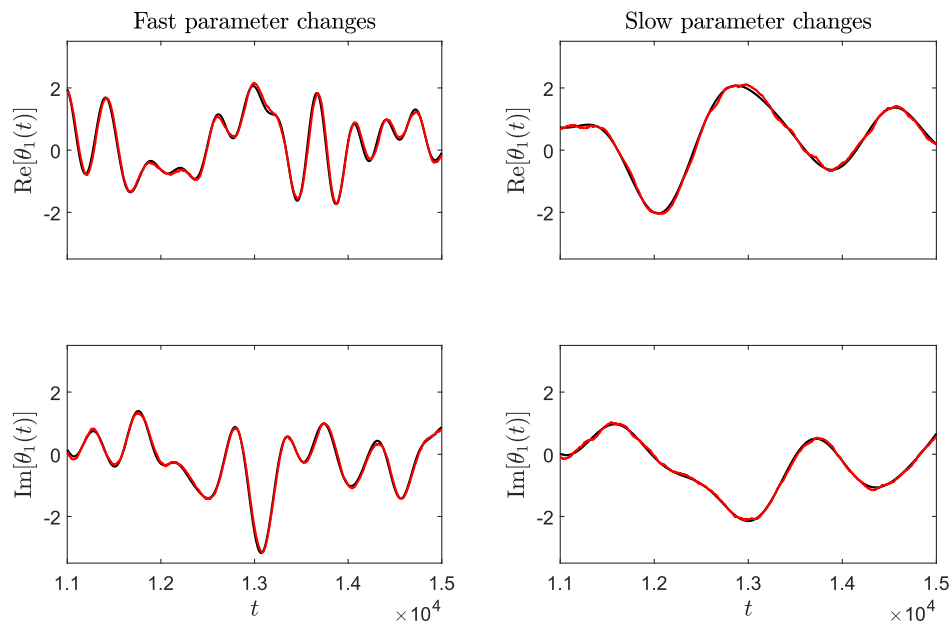
Figure 5.5: fLBF estimates (red lines) obtained for SNR equal to 10 dB, $m = 5$, $k = 200$, polynomial basis functions and rectangular weighting sequence, superimposed on true parameter trajectories (black lines).

Figure 5.6: fLBF estimate (red lines) of a system frozen impulse response obtained for SNR equal to 30 dB, $m = 5$, $k = 200$, polynomial basis functions and rectangular weighting sequence, superimposed on a true frozen impulse response (black lines).

on different types of preestimates. The EWLS-based and direct preestimates were obtained for $\lambda = 0.9$, and the LMS-based preestimates were computed for $\mu = 0.09$. The enhanced preestimates were obtained using the fLBF estimates calculated for EWLS-based preestimates. Finally, the E^2WLS-based and simplified E^2WLS-based preestimates were calculated for $\lambda = 0.8$. All results in these tables were obtained for $k = 200$, polynomial basis functions, and $w_k(i) \equiv 1$, $i \in I_k$.

Table 5.12: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for fast parameter changes. The comparison was obtained for the fLBF algorithm based on different types of preestimates.

SNR [dB]	m	EWLS	Direct	LMS	Enhanced	E^2WLS	Simplified E^2WLS
10	1	2.75	4.81	3.06	3.26	2.75	2.77
	3	-6.96	2.85	**-2.92**	-5.30	-7.14	-6.76
	5	**-8.67**	**2.50**	-1.92	**-9.26**	**-8.92**	**-8.21**
30	1	2.67	4.80	2.70	3.20	2.67	2.67
	3	-8.91	2.80	-8.27	-6.41	-9.45	-9.32
	5	**-16.37**	**2.42**	**-11.40**	**-19.04**	**-23.32**	**-22.41**
50	1	2.67	4.80	2.70	3.20	2.66	2.67
	3	-8.94	2.80	-8.38	-6.42	-9.49	-9.36
	5	**-16.59**	**2.42**	**-11.76**	**-19.42**	**-24.65**	**-23.69**

As expected, the results obtained using the bidirectional preestimates are better than these obtained using unidirectional ones. It is also worth noting that using the enhanced preestimates can substantially improve identification results if the starting estimates are of sufficient accuracy. In the next simulation, the performance of the FLBf with a polynomial basis was compared with the performance of this algorithm with the optimized impulse response. Tables 5.14 and 5.15 present the MSE scores for the optimal impulse response, described by (3.136) (called "optimal centralized") and suboptimal impulse response (3.137) (called "suboptimal centralized"), and their decoupled versions, i.e. the fLBF algorithm, for which the shape of the optimal impulse response was adjusted for each parameter trajectory separately. These algorithms are called "optimal decoupled" and "suboptimal decoupled" respectively.

Quite surprisingly, sometimes, the MSE score for the algorithm equipped with the optimized filter, designed using the centralized approach, yields results slightly worse than those, obtained from the algorithm using polynomial functions. This might be due to the fact that for the design of the optimal (and suboptimal) filter the average value of the preestimation variance was used.

Table 5.13: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for slow parameter changes. The comparison was obtained for the fLBF algorithm based on different types of preestimates.

SNR [dB]	m	EWLS	Direct	LMS	Enhanced	E^{2}WLS	Simplified E^{2}WLS
10	1	-10.40	2.58	**-6.65**	-8.14	-10.33	-10.01
	3	**-11.24**	2.26	-4.35	**-11.75**	**-10.97**	**-10.25**
	5	-9.36	**2.39**	-2.37	-9.69	-9.05	-8.35
30	1	-12.38	2.56	-12.34	-9.02	-12.51	-12.48
	3	**-25.28**	2.20	**-20.77**	**-28.00**	**-30.17**	**-29.40**
	5	-24.65	**2.31**	-18.81	-27.10	-28.49	-27.77
50	1	-12.41	2.56	-12.45	-9.04	-12.55	-12.52
	3	**-26.50**	2.20	**-23.18**	-30.32	-37.75	**-36.79**
	5	-26.37	**2.31**	-21.23	**-30.48**	**-37.29**	-36.43

Table 5.14: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for fast parameter changes. The results were obtained for different shapes of the filter impulse response, including impulse response associated with the polynomial basis (called here "fLBF"), the optimal and suboptimal filters designed for the entire system ("optimal centralized" and "suboptimal centralized" respectively) and optimal and suboptimal filters designed for each parameter separately ("optimal decoupled" and "suboptimal decoupled" respectively).

SNR [dB]	$m\backslash k$	fLBF			Optimal centralized			Suboptimal centralized			Optimal decoupled			Suboptimal decoupled		
		100	200	400	100	200	400	100	200	400	100	200	400	100	200	400
10	1	-4.87	2.75	5.41	-7.35	-7.80	**-8.06**	-7.36	-7.80	**-8.06**	-9.51	-9.79	**-9.95**	-9.51	-9.79	**-9.95**
	3	-7.78	-6.96	3.59												
	5	-6.05	**-8.67**	-1.94												
30	1	-5.84	2.67	5.39	-16.07	-16.29	**-16.46**	-16.08	-16.28	**-16.46**	-16.33	-16.42	**-16.49**	-16.33	-16.42	**-16.49**
	3	-16.18	-8.91	3.52												
	5	-15.64	**-16.37**	-2.36												
50	1	-5.85	2.67	5.39	-16.39	-16.58	**-16.74**	-16.40	-16.58	**-16.74**	-16.57	-16.64	**-16.70**	-16.57	-16.64	**-16.70**
	3	-16.44	-8.94	3.52												
	5	-16.00	**-16.59**	-2.36												

Table 5.15: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for slow parameter changes. The results were obtained for different shapes of the filter impulse response, including impulse response associated with the polynomial basis (called here "fLBF"), the optimal and suboptimal filters designed for the entire system ("optimal centralized" and "suboptimal centralized" respectively) and optimal and suboptimal filters designed for each parameter separately ("optimal decoupled" and "suboptimal decoupled" respectively).

SNR [dB]	$m\backslash k$	fLBF			Optimal centralized			Suboptimal centralized			Optimal decoupled			Suboptimal decoupled		
		100	200	400	100	200	400	100	200	400	100	200	400	100	200	400
10	1	-11.51	-10.40	-1.88	-10.48	-11.00	**-11.73**	-10.39	-11.01	**-11.71**	-12.34	-13.09	**-13.55**	-12.34	-13.09	**-13.55**
	3	-8.35	-11.24	**-13.33**												
	5	-6.47	-9.36	-12.27												
30	1	-21.74	-12.38	-2.01	-24.17	-25.03	**-25.33**	-24.16	-25.00	**-25.32**	-24.78	-25.29	**-25.50**	-24.78	-25.29	**-25.50**
	3	-24.12	-25.28	-20.18												
	5	-22.98	-24.65	**-25.63**												
50	1	-22.19	-12.41	-2.01	-26.15	-26.47	**-26.59**	-26.16	-26.48	**-26.59**	-26.37	-26.58	**-26.64**	-26.37	-26.58	**-26.64**
	3	-26.10	-26.50	-20.37												
	5	-25.47	-26.37	**-26.63**												

However, if some statistical knowledge about parameter changes is known, then this optimization approach can be successfully used to achieve satisfactory identification results without looking for the best combination of m and k. If the maximum acceptable communication delay is known, then the maximum value of k can be easily found, and it is easy to notice that the estimation results for the optimized filters tend to improve with growing k.

The last two tables - tab. 5.16 and 5.17 show the results after applying the adaptive techniques described in Section 3.9. It is a comparison of results yielded by the adaptive choice applied in the centralized manner based on the LOOCV criterion either for the system errors or preestimates errors, and based on the FPE criterion. The example of results obtained after using the sign test was presented in the last section of this chapter.

Table 5.16: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for fast parameter changes. The results were obtained for different strategies of adaptive choice of m and k - LOOCV based on system errors, preestimates errors or FPE criterion.

SNR [dB]	$m\backslash k$	100	200	400	CV, system errors	CV, preestimates errors	FPE
10	1	-4.87	2.75	5.41			
	3	-7.78	-6.96	3.59	-7.65	-7.79	-6.86
	5	-6.05	**-8.67**	-1.94			
30	1	-5.84	2.67	5.39			
	3	-16.18	-8.91	3.52	-16.42	-16.38	-16.01
	5	-15.64	**-16.37**	-2.36			
50	1	-5.85	2.67	5.39			
	3	-16.44	-8.94	3.52	-16.64	-16.60	-16.31
	5	-16.00	**-16.59**	-2.36			

Table 5.17: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for slow parameter changes. The results were obtained for different strategies of adaptive choice of m and k - LOOCV based on system errors, preestimates errors or FPE criterion.

SNR [dB]	$m\backslash k$	100	200	400	CV, system errors	CV, preestimates errors	FPE
10	1	-11.51	-10.40	-1.88			
	3	-8.35	-11.24	**-13.33**	-9.61	-9.44	-8.35
	5	-6.47	-9.36	-12.27			
30	1	-21.74	-12.38	-2.01			
	3	-24.12	-25.28	-20.18	-25.53	-25.59	-24.56
	5	-22.98	-24.65	**-25.63**			
50	1	-22.19	-12.41	-2.01			
	3	-26.10	-26.50	-20.37	-26.74	-26.67	-26.48
	5	-25.47	-26.37	**-26.63**			

As can be seen in tables 5.16 and 5.17, the adaptive methods can provide results that are substantially better then the results yielded by most of the algorithms incorporated in the parallel scheme, and comparable with the MSE scores obtained for the best of the algorithms. Sometimes they might even improve slightly the best of the results obtained from the candidate algorithms.

5.4 RLBF

Here we show the potential of the regularization. As was pointed out in the Remark 2 in Section 4.4, regularization brings benefits when the variance component of the MSE dominates. Therefore, we show here the comparison for a short analysis window ($k = 70$) and relatively high number of basis functions ($m = 5$). In tables below, the results were shown for the LBF algorithm and for the RLBF algorithm with the regularization matrix described by (4.51), where $\mathbf{Q}_0 = \text{diag}\{\gamma, \ldots, \gamma^{n-1}\}$, where $\gamma \in \{0.5, 0.7, 0.9\}$, none of which is exactly equal to the true value of γ. Table 5.18 shows also the results provided by the algorithms with adaptive choice of the regularization matrix - LOOCV with the exact value of $q_{m|k}(t)$ (A0), its approximation $\bar{q}_{m|k}(t)$ (A1), and the Bayesian approach with the exact formula (4.44) (B0) and with the approximation (4.45) (B1).

Table 5.18: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for fast (on the left) and slow (on the right) parameter changes. The results were obtained for $k = 70$, $m = 5$, the LBF and RLBF algorithms with the adaptive choise of a parameter γ, based on cross-validation (A0 and A1) and empirical Bayes approach (B0 and B1).

SNR [dB]	γ	RLBF	LBF	SNR [dB]	γ	RLBF	LBF
	0.5	-5.04			0.5	-5.03	
	0.7	**-7.06**			0.7	**-7.06**	
	0.9	-4.19			0.9	-4.20	
10	A0	-6.98	-3.37	10	A0	-6.99	-3.39
	A1	-7.01			A1	-7.01	
	B0	-4.15			B0	-4.18	
	B1	-4.15			B1	-4.18	
	0.5	-17.54			0.5	-17.53	
	0.7	**-23.68**			0.7	**-23.70**	
	0.9	-23.38			0.9	-23.40	
30	A0	-23.68	-23.37	30	A0	-23.70	-23.39
	A1	-23.68			A1	-23.69	
	B0	-18.82			B0	-18.80	
	B1	-18.80			B1	-18.78	
	0.5	-34.66			0.5	-34.67	
	0.7	**-43.36**			0.7	**-43.39**	
	0.9	-43.36			0.9	-43.39	
50	A0	-43.36	-43.36	50	A0	-43.39	-43.39
	A1	-43.36			A1	-43.39	
	B0	-34.68			B0	-34.69	
	B1	-34.68			B1	-34.69	

Results gathered in table 5.18 confirm the conclusions drawn in the previous chapter, that applying regularization brings the greatests benefits for low values of SNR. It is also clearly visible that the adaptive algorithms based on simplified formulas yield almost the same results as the algorithms based on the exact formulas.

5.5   fRLBF

In this section we compare the results yielded by the fLBF algorithm for $k = 100$ and $m = 5$ and its regularized version (fRLBF), characterized by (4.7) and (4.8). LOOCV and empirical Bayes (EB) rules (both in decentralized fashion) were applied to choose the value of $\mu \in \{0.001, 0.1, 0.5\}$. The results of this experiment were gathered in table 5.19.

Table 5.19: Comparison of the MSE scores [dB] averaged over one long realization of input\output data for fast (on the left) and slow (on the right) parameter changes. The results were obtained for $k = 100$, $m = 5$, the fLBF and fRLBF algorithms with the adaptive choise of a regularization constant, based on cross-validation and empirical Bayes approach.

SNR [dB]	μ	fRLBF	fLBF	SNR [dB]	μ	fRLBF	fLBF
10	0.001	-6.05	-6.05	10	0.001	-6.47	-6.47
	0.1	-6.65			0.1	-7.20	
	0.5	**-7.76**			0.5	**-8.90**	
	LOOCV	-7.58			LOOCV	-8.79	
	EB	-6.06			EB	-6.49	
30	0.001	-15.66	-15.64	30	0.001	-23.03	-22.98
	0.1	-15.77			0.1	-23.15	
	0.5	-12.71			0.5	-16.90	
	LOOCV	**-15.82**			LOOCV	**-23.32**	
	EB	-15.54			EB	-22.69	
50	0.001	-16.03	-16.00	50	0.001	**-25.55**	-25.47
	0.1	-16.04			0.1	-24.28	
	0.5	-12.76			0.5	-16.97	
	LOOCV	**-16.11**			LOOCV	-25.43	
	EB	-15.88			EB	-24.90	

The results presented in table 5.19 confirms that the greatest improvement can be achieved for low values of SNR. It can be also seen, that the too small or too high values of the regularization constant can substantially deteriorate the estimates quality. Fortunately, proposed adaptive schemes always provide results that are either better or comparable with the best results yielded by all algorithms incorporated in the parallel estimation scheme.

Remark

More results regarding regularized basis function methods can be found in [72], [74] and [79].

5.6   Lake experiment and sea simulations

This section presents the results published in [63] and [78]. The first experiment was simulating the self-interference cancellation in FD UWA communication, details can be found in [93]. The transmitter-receiver device, presented in figure 5.7, was placed inside an 8 m deep lake, 4 m under the surface. Transmitter was sending the known signal, while the receivers were recording its version contaminated by self-interference. Because in the real experiment, the true parameter trajectories are not available, the self-interference cancellation (SIC) factor [92] was used to measure the estimation accuracy. The SIC factor reflects the ratio between signal-to-interference ratios before and after self-interference cancellation. Hence, the higher the SIC score, the higher the accuracy of the estimates. Estimation was carried out for $n = 80$ system parameters.
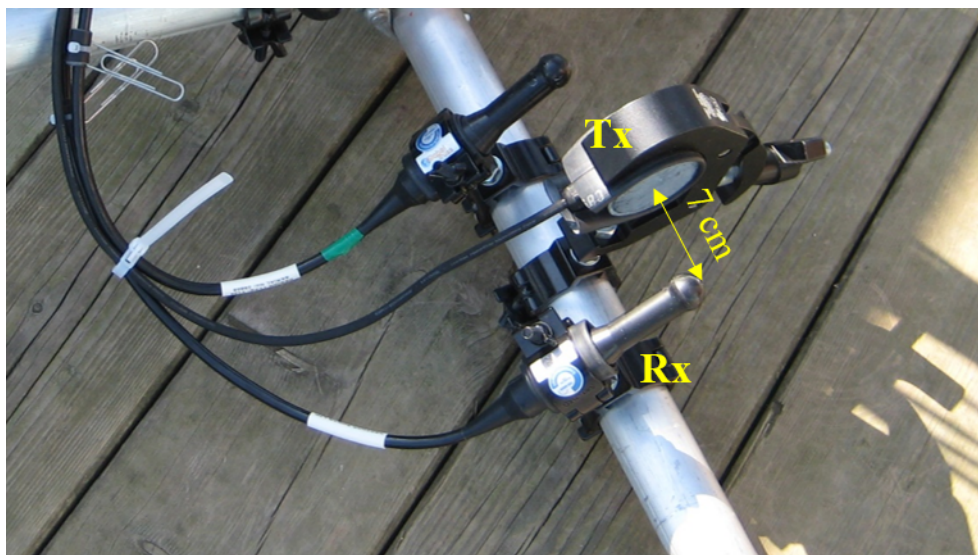
Figure 5.7: Photo of a transmitter and receiver antennas mounted on the frame [93].

Using the data gathered during the lake experiment, the fLBF algorithm was compared with the fLBF algorithm combined with the adaptive detection of static parameters (called adaptive fLBF), which was done using the sign test (see [78] for more details), the adaptively debiased fLBF (dfLBF), and the regularized version of the fLBF algorithm (fRLBF or dfRLBF). The regularization constant was chosen for each parameter separately using the empirical Bayes approach. The results were gathered in figure 5.8.

Figure 5.8 shows that all techniques proposed in the article [78] improve the identification results. The adaptive choice of time-varying parameters using the sign test increases the estimation accuracy because in practice only the parameters corresponding to the reflections from the moving objects like lake surface or underwater animals are time-varying, while the rest of them is constant (see figure 5.9). However, the static parameters can become time-varying and vice versa if the conditions change, e.g. if the far-end transmitter starts moving. This is the way the detection should be carried out for each time instant. Also, using regularization is helpful because of the sparse nature of the time-varying impulse response. As explained before, when the forgetting constant increases, so does the preestimation delay, hence, the adaptive debiasing also improves the estimation accuracy. Not surprisingly, a combination of these techniques provides the best estimation accuracy. In the article [63], the proposed improvements of the fLBF method (except for the regularization) were compared with the LBF method. Results are shown in figure 5.10.

The article [78] presents also the results for the more realistic simulated system. In this article, two situations were simulated: static transmitter and receiver, and static transmitter, moving receiver. Simulations were prepared using the Waymark simulator [50] and Bellhop acoustic toolbox [84] (for more details see [78]). Simulations were carried out in the "South Korean Sea" scenario (see [84]) - sea depth was set to 20 m, and the devices' depth was 10 m. In the first simulation, the distance between transmitter and receiver was set to 7 m and wind speed was equal to 10 m/s, in the second scenario the initial distance between communication devices was set to 50 m, the sea surface was assumed to be flat (zero wind speed) and the receiver was moving away at speed 2.5 m/s. Figure 5.11 presents the amplitudes of time-varying parameters for these simulated scenarios.

The estimation accuracy was measured using the mean squared deviation defined as

$$\text{MSD}(t) = \frac{||\widehat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}(t)||^2}{||\boldsymbol{\theta}(t)||^2}. \tag{5.3}$$

The MSD score, averaged over 5000 samples are depicted in the figure 5.12.
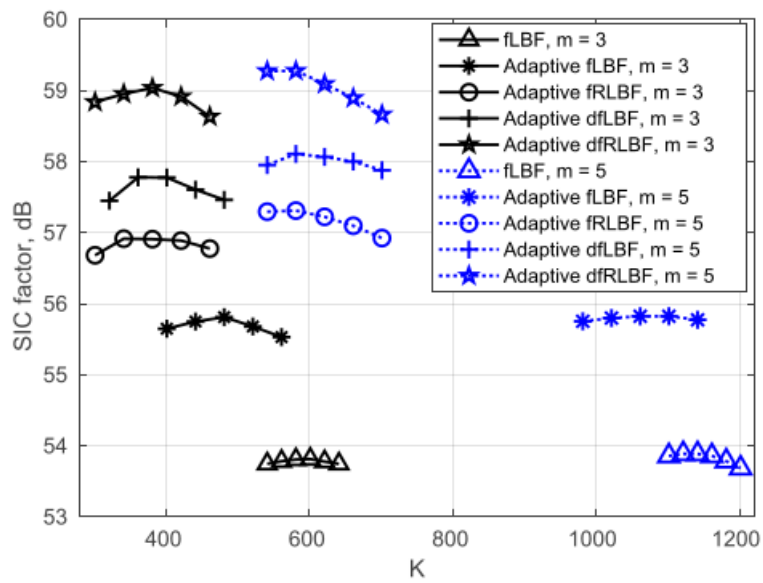
Figure 5.8: Comparison of a SIC factor obtained using fLBF, adaptively debiased fLBF (dfLBF) regularized fLBF (fRLBf or dfRLBF) using the empirical Bayes approach for choosing the regularization constant, and fLBF with the adaptive choice of static and time-varying parameters (Adaptive fLBF, dfLBF, fRLBF or dfRLBF) [78].
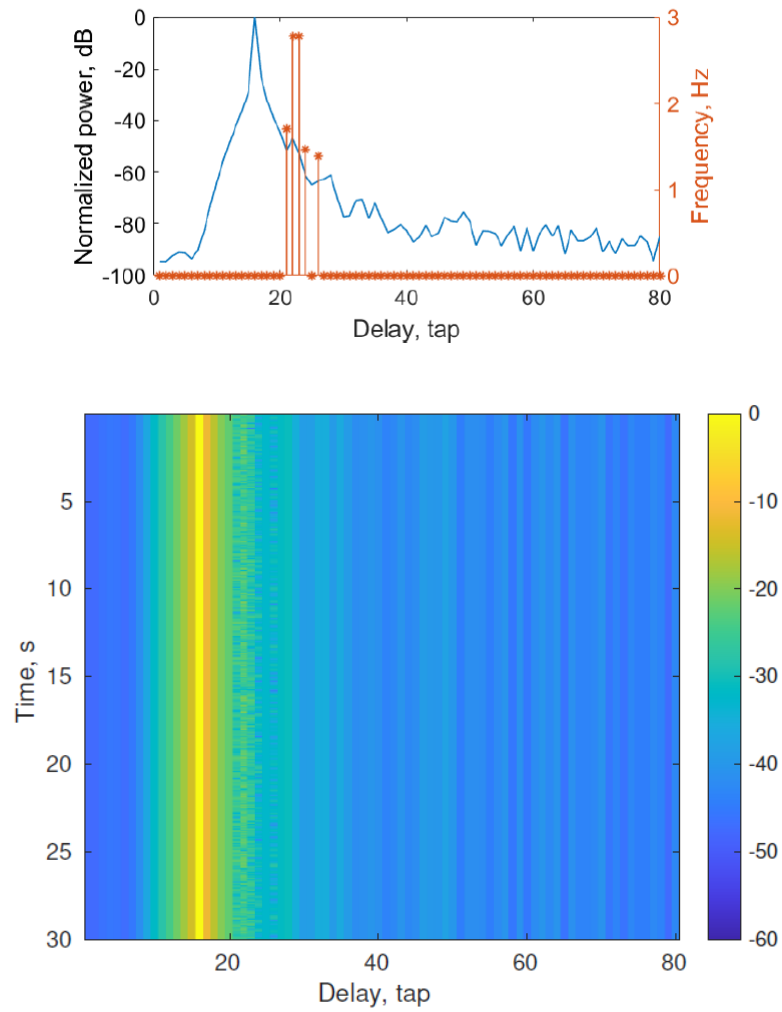
Figure 5.9: The power delay profile and highest frequencies for each of the system parameters (the upper plot), and the normalized amplitude of the time-varying impulse response in decibels (the lower plot) [78].
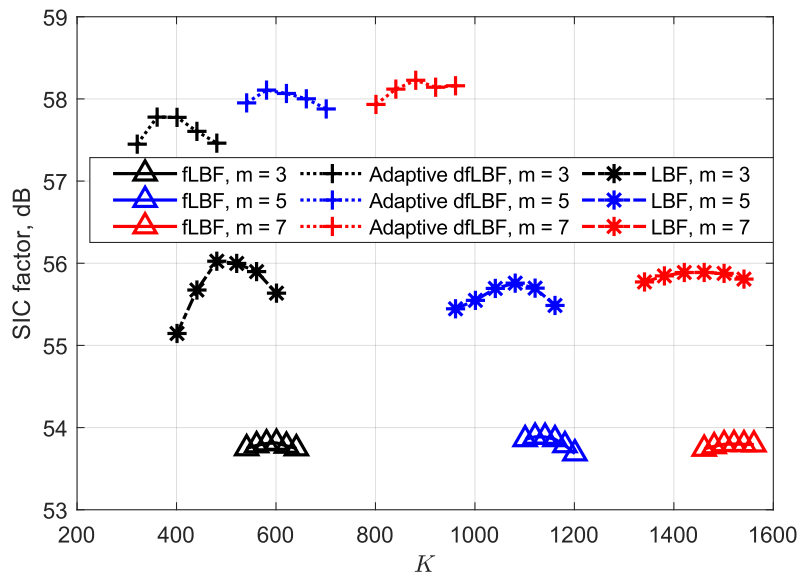
Figure 5.10: Comparison of a SIC factor obtained using fLBF, adaptively debiased fLBF with the adaptive choice of time-varying and static parameters, and LBF methods on the data gathered during the lake experiment [63].
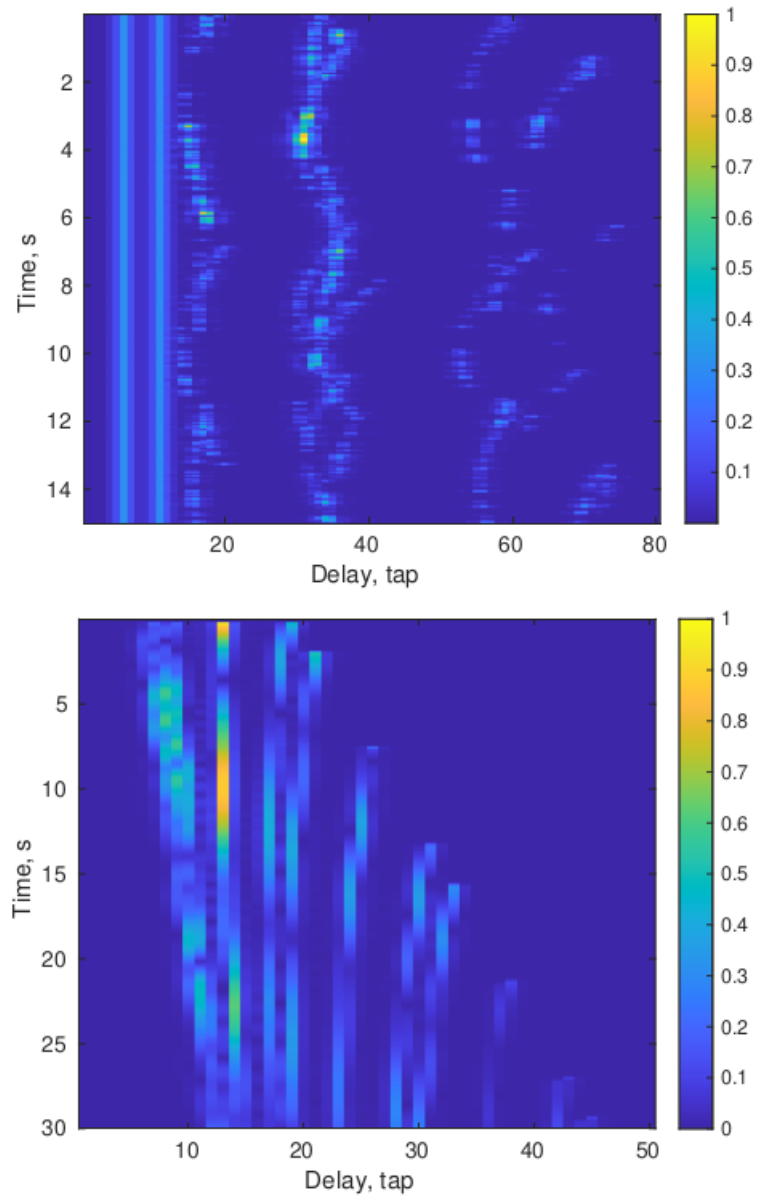
Figure 5.11: The normalized magnitudes of time-varying parameters for simulated sea scenarios with static communication devices (the upper plot), and static transmitter and moving receiver (the lower plot) [78].
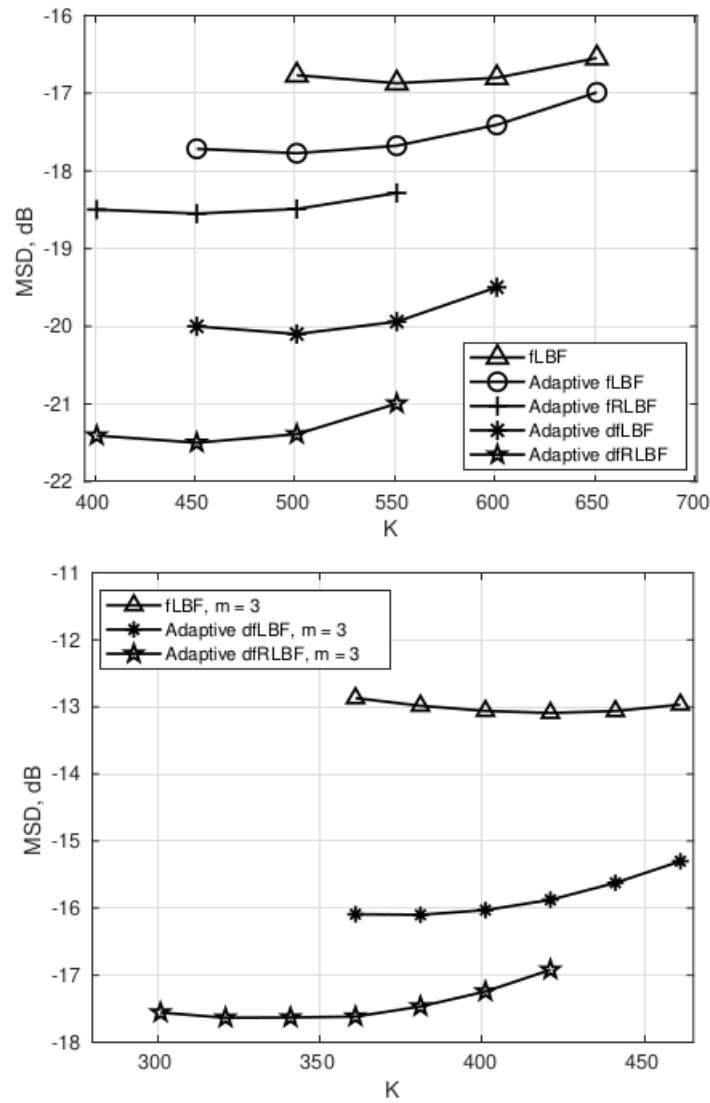
Figure 5.12: The MSD scores averaged over 5000 time samples obtained in the simulation of UWA communication in the sea, for static communication devices (the upper plot), and static transmitter and moving receiver (the lower plot) [78].

Chapter 6

Conclusions

This thesis summarizes the most recent developments of the basis function methods (LBF and fLBF). The theoretical analysis and results gathered in the previous chapter, as well as in the articles listed in the references, suggest that the LBF method is the state-of-the-art version of the basis function method for the identification of nonstationary processes. Conducted simulations confirm a few intuitive conclusions. First, applying a bell-shaped weighting sequence during identification can improve estimation accuracy by placing more emphasis on data close to the center of the analysis window. Second, when no prior knowledge about parameter changes is available, the use of "non-specific" basis functions like polynomials, or sine-cosines can provide estimates of satisfying quality. Third, when the basis functions are chosen based on the knowledge of statistics of parameter changes, the identification results can be further improved. Some other intuitive hopes turned out to be in vain. For example, one might have expected that the methods of adaptive choice of hyperparameters will result in estimates of higher quality than estimates provided by all the algorithms involved in parallel computations. Such methods evaluate the local measures of fit for different estimators, and at each time point, they choose the one that performs the best. As a consequence, one might expect that the final estimate - an amalgam of locally the best estimates will be of the highest quality of them all. However, numerous simulations show that it is rarely the case. In most cases, the adaptive methods yield estimates that are of comparable quality as the best estimates obtained from single algorithms and of noticeably higher quality than the worst of them. This conclusion holds for both LBF, fLBF estimators, and their regularized versions. Hence, it is a good method for robustifying estimation results against unknown, and possibly time-varying characteristics of parameter trajectories.

Another mild disappointment regards the fLBF method. Initial experiments, made for systems with a very low number of parameters (e.g. $n = 2$) indicated that the estimates provided by this method can be indistinguishable from the estimates yielded by the LBF algorithm. However further simulations for systems with more parameters have shown that even though the fLBF estimates still look convincingly accurate, they yield the mean square estimation errors that are noticeably higher than for the LBF estimates. This can be attributed to the fact that when n grows, one needs to use the value of forgetting constant that is closer to 1. This means that the causal preestimation algorithms will be characterized by higher delay and simultaneously the EWLS algorithm will be incapable of tracking fast parameter changes - more information about true parameter trajectories will be lost at this stage. The remedy for the first problem is quite simple - one can use either debiasing mechanisms or noncausal preestimation techniques. Quite interestingly, even for high values of n, the fLBF estimates are of higher accuracy than LBF estimates for low values of signal-to-noise ratio.

The aforementioned problems should not rule out the fLBF method completely. It proves very useful whenever a method computationally cheaper than LBF, but providing a similar accuracy of estimates, is needed. It allows also to use much higher number of basis functions with relatively short analysis windows - something that would be impossible for the LBF method due to the numerical conditioning requirements. The second case, where it proves to be useful, is when the speed or type of parameter changes differs for each system parameter. Then, as shown in [78], the ability of the fLBF method to process each parameter separately can provide many benefits.

Another novelty described in this thesis is a combination of regularization and basis function methods. First, it is worth noting that regularization can improve the conditioning of the least

squares problem in situations where the order of a hypermodel is close to or exceeds the length of the analysis interval. However, its main goal is to allow the incorporation of prior knowledge into the identification process. It can be achieved by a proper design of the regularization matrix. The regularization constant, the second important hyperparameter in regularization techniques, controls the ratio between information extracted from data and the one introduced in the form of prior knowledge. As pointed out previously, regularization provides the largest benefits when the variance component of the MSE dominates. It means that it can be particularly useful when the signal-to-noise ratio is low, or when the number of basis functions is high and simultaneously the processing delay is limited, which limits the length of the analysis window. Simulation results suggest also that the adaptive methods of choosing regularization hyperparameters are very important because a poorly designed regularization matrix or too high value of a regularization constant can deteriorate estimation results.

Contents

List of Figures

List of Tables

Appendix A

The Wirtinger derivative and the least squares problem for complex-valued vectors

In this appendix we discuss briefly the problem of optimization of real-valued functions of several complex variables. For a more rigorous mathematical treatment, see [31] and [89].

Let $z = x + \mathrm{i}y$ be a complex number where $x, y \in \mathbb{R}$. We call the function $f : \mathbb{C} \to \mathbb{C}$

$$f(z) = u(x, y) + \mathrm{i}v(x, y), \tag{A.1}$$

a function of a complex variable z, where $u : \mathbb{R}^2 \to \mathbb{R}$, $v : \mathbb{R}^2 \to \mathbb{R}$. The complex derivative of a function $f(z)$ at the point z_0 is defined as

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}, \tag{A.2}$$

provided that the limit exists. Note that the difference Δz can approach 0 in infinitely many ways. Note also that $\Delta z = \Delta x + \mathrm{i}\Delta y$. By computing the derivative in two different ways, first by setting $\Delta x = 0$, second by setting $\Delta y = 0$, one obtains the Cauchy-Riemann equations

$$\frac{\partial u}{\partial x}(x_0, y_0) = \frac{\partial v}{\partial y}(x_0, y_0) \ \wedge \ \frac{\partial u}{\partial y}(x_0, y_0) = -\frac{\partial v}{\partial x}(x_0, y_0). \tag{A.3}$$

When the complex derivative of a function $f(z)$ at the point z_0 exists, then the Cauchy-Riemann equations hold true. Hence, they can be used as a necessary condition for the existence of a complex derivative. Moreover, if the functions $u(x, y)$ and $v(x, y)$ have continuous partial derivatives in some neighborhood D of a point z_0 and additionally fulfill the Cauchy-Riemann conditions, then the complex derivative of a function $f(z)$ exists in every point of D [89]. A function, for which the complex derivative exists in its entire domain, is called holomorphic or analytic.

Although the definition of a complex derivative is very natural, its existence imposes serious limitations on the structure of the function $f(z)$. For instance, it is easy to show that a real-valued function of a complex variable, namely the function for which $v(x, y) \equiv 0$, is holomorphic only when it is a constant function. Unfortunately, many cost functions present in engineering applications are real-valued functions of a complex variable or several complex variables. This means that they are not complex-differentiable, i.e. differentiable in the sense described above. One solution to this problem is to interpret the function $f(z)$ as a two-dimensional function of real variables x and y, to compute the derivative with respect to the vector $[x, y]^{\mathrm{T}}$ and then to convert the solution to the complex domain, such approach was presented in [10]. However, there is a simpler approach, after the work of Wirtinger [109] often called the Wirtinger derivative [1], or less commonly called CR-calculus [20], [21].

One can interpret the function of a complex variable $f(z)$ as a function of two real variables $f(x, y)$. One can also note that

$$x = \frac{1}{2}(z + z^*), \quad y = \frac{1}{2}(z - z^*). \tag{A.4}$$

127

Hence, the function $f(x,y)$ depends on z and z^* which can be treated as independent variables. Provided that the partial derivatives exist, one can use the chain rule to obtain

$$
\begin{aligned}
\frac{\partial f}{\partial z}(z, z^*) &= \frac{1}{2}\left[\frac{\partial f}{\partial x}(x, y) - i\frac{\partial f}{\partial y}(x, y)\right] \\
\frac{\partial f}{\partial z^*}(z, z^*) &= \frac{1}{2}\left[\frac{\partial f}{\partial x}(x, y) + i\frac{\partial f}{\partial y}(x, y)\right].
\end{aligned}
\tag{A.5}
$$

Note that for real-valued functions

$$
\frac{\partial f}{\partial z^*}(z, z^*) = \left[\frac{\partial f}{\partial z}(z, z^*)\right]^*,
\tag{A.6}
$$

the partial derivative with respect to a complex conjugate of z is just a complex conjugate of a partial derivative with respect to z. As a consequence, one only needs to evaluate $\frac{\partial f}{\partial z}(z, z^*)$, because a real-valued function will have a stationary point at (z_0, z_0^*) when $\frac{\partial f}{\partial z}(z_0, z_0^*) = 0$ [1]. Another interesting observation is that the Cauchy-Riemann equations are equivalent to

$$
\frac{\partial f}{\partial z^*}(z, z^*) = 0,
\tag{A.7}
$$

which means that the holomorphic functions depend only on z - they are independent of z^*.

For the practitioner, the most important conclusions are that the rules for differentiation of functions of real variables apply in the Wirtinger calculus [89] and that the rules of the Wirtinger calculus can be easily extended to the functions of several complex variables [31]. This means that one can use the rules known from the calculus of functions of multiple real variables for solving the least squares problem with complex-valued vectors, which is demonstrated in the following example.

Example

The purpose of this example is to demonstrate the technique of solving the least-squares problems in a complex domain.

Consider the task of finding the $n \times 1$ complex-valued vector $\mathbf{z} = \mathbf{x} + i\mathbf{y}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ which minimizes the following cost function

$$
J(\mathbf{z}) = ||\mathbf{A}\mathbf{z} - \mathbf{b}||^2,
\tag{A.8}
$$

where $\mathbf{A} \in \mathbb{C}^{m \times n}$, $\mathbf{b} \in \mathbb{C}^m$, $\mathrm{rank}(\mathbf{A}) = n$, $m \geq n$ and $||\mathbf{z}|| = \sqrt{\mathbf{z}^H\mathbf{z}}$ denotes the ℓ^2 norm of a vector z. Naturally, $J: \mathbb{C}^n \to \mathbb{R}$ is a real-valued function, and hence is not complex-differentiable. The cost function (A.8) can be expressed as

$$
\begin{aligned}
J(\mathbf{x}, \mathbf{y}) &= (\mathbf{x} - i\mathbf{y})^T\mathbf{A}^H\mathbf{A}(\mathbf{x} + i\mathbf{y}) - (\mathbf{x} - i\mathbf{y})^T\mathbf{A}^H\mathbf{b} - \mathbf{b}^H\mathbf{A}(\mathbf{x} + i\mathbf{y}) + \mathbf{b}^H\mathbf{b} \\
&= \mathbf{x}^T\mathbf{A}^H\mathbf{A}\mathbf{x} + \mathbf{y}^T\mathbf{A}^H\mathbf{A}\mathbf{y} - \mathbf{x}^T\mathbf{A}^H\mathbf{b} - \mathbf{b}^H\mathbf{A}\mathbf{x} + i(\mathbf{x}^T\mathbf{A}^H\mathbf{A}\mathbf{y} - \mathbf{y}^T\mathbf{A}^H\mathbf{A}\mathbf{x} - \mathbf{b}^H\mathbf{A}\mathbf{y} + \mathbf{y}^T\mathbf{A}^H\mathbf{b}) \\
&\quad + \mathbf{b}^H\mathbf{b}.
\end{aligned}
\tag{A.9}
$$

From the above, it follows that the real derivatives are defined as

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{x}}(\mathbf{x}, \mathbf{y}) &= \mathbf{x}^T(\mathbf{A}^H\mathbf{A} + \mathbf{A}^T\mathbf{A}^*) - \mathbf{b}^T\mathbf{A}^* - \mathbf{b}^H\mathbf{A} + i(\mathbf{y}^T\mathbf{A}^T\mathbf{A}^* - \mathbf{y}^T\mathbf{A}^H\mathbf{A}) \\
\frac{\partial J}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y}) &= \mathbf{y}^T(\mathbf{A}^H\mathbf{A} + \mathbf{A}^T\mathbf{A}^*) + i(\mathbf{x}^T\mathbf{A}^H\mathbf{A} - \mathbf{x}^T\mathbf{A}^T\mathbf{A}^* - \mathbf{b}^H\mathbf{A} + \mathbf{b}^T\mathbf{A}^*).
\end{aligned}
\tag{A.10}
$$

As a consequence, the complex derivative can be written down as [31]

$$
\begin{aligned}
\frac{\partial J}{\partial \mathbf{z}}(\mathbf{z}, \mathbf{z}^*) &= \frac{1}{2}\left[\frac{\partial J}{\partial \mathbf{x}}(\mathbf{x}, \mathbf{y}) - i\frac{\partial J}{\partial \mathbf{y}}(\mathbf{x}, \mathbf{y})\right] \\
&= \frac{1}{2}\left[\mathbf{x}^T(\mathbf{A}^H\mathbf{A} + \mathbf{A}^T\mathbf{A}^*) + \mathbf{x}^T\mathbf{A}^H\mathbf{A} - \mathbf{x}^T\mathbf{A}^T\mathbf{A}^* - \mathbf{b}^T\mathbf{A}^* - \mathbf{b}^H\mathbf{A} - \mathbf{b}^H\mathbf{A} + \mathbf{b}^T\mathbf{A}\right] + \\
&\quad + \frac{1}{2}i\left[\mathbf{y}^T\mathbf{A}^T\mathbf{A}^* - \mathbf{y}^T\mathbf{A}^H\mathbf{A} - \mathbf{y}^T(\mathbf{A}^H\mathbf{A} + \mathbf{A}^T\mathbf{A}^*)\right] \\
&= \mathbf{x}^T\mathbf{A}^H\mathbf{A} - \mathbf{b}^H\mathbf{A} - i\mathbf{y}^T\mathbf{A}^H\mathbf{A} = \mathbf{z}^H\mathbf{A}^H\mathbf{A} - \mathbf{b}^H\mathbf{A}.
\end{aligned}
\tag{A.11}
$$

The stationary point can be obtained from

$$\frac{\partial J}{\partial \mathbf{z}}(\mathbf{z}, \mathbf{z}^*) = 0 \iff \mathbf{z}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}\mathbf{A} = \mathbf{b}^{\mathrm{H}}\mathbf{A} \iff \mathbf{z} = (\mathbf{A}^{\mathrm{H}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{H}}\mathbf{b}. \tag{A.12}$$

Since the cost function is convex, the solution is

$$\widehat{\mathbf{z}} = (\mathbf{A}^{\mathrm{H}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{H}}\mathbf{b}. \tag{A.13}$$

Note that this result can be also achieved after noting that the matrix $\mathbf{A}^{\mathrm{H}}\mathbf{A}$ is by definition positive definite, and incorporating the generalization of "square-completing" technique, described in Chapter 4 (see [98]).

References

[1] T. Adali, P. J. Schrier, and L. L. Scharf, "Complex-valued signal processing: the proper way to deal with impropriety". In: *IEEE Transactions on Signal Processing* vol. 59 (2011), pp. 5101–5125.

[2] H. Akaike, "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* vol. 19 (1974), pp. 716–723.

[3] H. Akaike, "Likelihood and the Bayes procedure". In: *Trabajos de Estadistica Y de Investigacion Operativa* vol. 31 (1980), pp. 143–166.

[4] H. Akaike, "Statistical predictor identification". In: *Annals of the Institute of Statistical Mathemathics* vol. 22 (1970), pp. 203–217.

[5] T. S. Alexander, *Adaptive Signal Processing, Theory and Applications*. Springer-Verlag, 1986.

[6] D. Allen, "The relationship between variable selection and data augmentation and a method for prediction". In: *Technometrics* vol. 16 (1974), pp. 125–127.

[7] B. Bamieh and L. Giarré, "Identification of Linear Parameter Varying models". In: *Proceedings of the 38th Conference on Decision and Control* (Phoenix, Arizona USA). 1999.

[8] M. S. Bani and B. L. Chalmers, "Best approximation in L^∞ via iterative Hilbert space procedures". In: *Journal of Approximation Theory* vol. 42 (1984), pp. 173–180.

[9] J. R. Bellegarda and D. C. Farden, "Constrained time-varying system modelling". In: *1988 American Control conference* (Atlanta, USA). 1988.

[10] A. van den Bos, "Complex gradient and Hessian". In: *IEEE Proc.-Vis. Image Signal Processing* vol. 141 (1994), pp. 380–382.

[11] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *Proc. Natl. Acad. Sci. USA* vol. 113 (2016), pp. 3932–3937.

[12] C. S. Burrus, J. A. Barreto, and I. W. Selesnick, "Iterative reweighted least-squares design of FIR filters". In: *IEEE Transactions on Signal Processing* vol. 42 (1994), pp. 2926–2936.

[13] E. Candes and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n". In: *The Annals of Statistics* vol. 35 (2007), pp. 2313–2351.

[14] S. Chen, S. A. Billings, and W. Luo, "Orthogonal Least Squares Methods and their application to nonlinear system identification". In: *International Journal of Control* vol. 50 (1989), pp. 1873–1896.

[15] T. Chen, "On kernel design for regularized LTI system identification". In: *Automatica* vol. 90 (2018), pp. 109–122.

[16] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and gaussian process - revisited". In: *Automatica* 48 (2012), pp. 1525–1535.

[17] M. Ciołek, M. Niedźwiecki, and A. Gańcza, "Decoupled Kalman filter based identification of time-varying FIR systems". In: *IEEE Access* vol. 9 (2021), pp. 74622–74631.

[18] B. Efron, "Biased versus unbiased estimation". In: *Advances in Mathematics* vol. 16 (1975), pp. 259–277.

[19] T. H. Eggen, A. B. Baggeroer, and J. C. Preisig, "Communication over Doppler spread channels. Part I: Channel and receiver presentation". In: *IEEE Journal of Oceanic Engineering* vol. 25 (2000), pp. 62–71.

[20] J. Eriksson, E. Ollila, and V. Koivunen, "Essential statistics and tools for complex random variables". In: *IEEE Transactions on Signal Processing* vol. 58 (2010), pp. 5400–5408.

[21] J. Eriksson, E. Ollila, and V. Koivunen, "Statistics for complex random variables revisited". In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (Taipei, Taiwan). 2009.

[22] A. Gańcza and M. Niedźwiecki, "Regularized identification of fast time-varying systems - comparison of two regularization strategies". In: *Proc. of the 60th IEEE Conference on Decision and Control* (Austin, Texas, USA). 2021.

[23] A. Gańcza and M. Niedźwiecki, "Regularized identification of time-varying FIR system based on generalized cross-validation". In: *Proc. of the 29th European Signal Processing Conference* (Dublin, Irleand). 2021.

[24] A. Gańcza, M. Niedźwiecki, and M. Ciołek, "Regularized local basis function approach to identification of nonstationary systems"". In: *IEEE Transactions on Signal Processing* vol. 69 (2021), pp. 1665–1680.

[25] R. C. Geary, "Relative efficiency of count of sign changes for assessing residual autoregression in least squares regression". In: *Biometrika* vol. 57 (1970), pp. 123–127.

[26] W. Gersch and G. Kitagawa, "Smoothness priors transfer function estimation". In: *Automatica* vol. 25 (1989), pp. 603–608.

[27] G. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter". In: *Technometrics* vol. 21 (1979), pp. 215–223.

[28] G. Golub and C. van Loan, *Matrix Computations*. The John Hopkins University Press, 1996.

[29] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* vol. 41 (1983), pp. 899–911.

[30] D. M. Gruenbacher and D. R. Hummels, "A simple algortihm for generating Discrete Prolate Spheroidal Sequences". In: *IEEE Transactions on Signal Processing* vol. 42 (1994), pp. 3276–3278.

[31] R.C. Gunning and H. Rossi, *Analytic functions of several complex variables*. Providence: AMS Chelsea Publishing, 2009.

[32] L. Guo and L. Ljung, "Performance analysis of general tracking algorithms". In: *IEEE Transactions on Automatic Control* vol. 40 (1995), pp. 1388–1402.

[33] M. G. Hall, A. V. Oppenheim, and A. S. Willsky, "Time-varying parametric modeling of speech". In: *Signal Processing* vol. 5 (1983), pp. 267–285.

[34] T. Hastie and R. Tibshirani, "Varying-coefficient models". In: *Journal of the Royal Statistical Society B* vol. 55 (1993), pp. 757–796.

[35] S. Haykin, *Adaptive Filter Theory*. Pearson, 1996.

[36] A. E. Hoerl and W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems". In: *Technometrics* vol. 12 (1970), pp. 55–67.

[37] J. Z. Huang, C. O. Wu, and L. Zhou, "Varying-coefficient models and basis function approximations for the analysis of repeated measurements". In: *Biometrika* vol. 89 (2002), pp. 111–128.

[38] M. Kac and A. J. F. Siegert, "On the theory of noise in radio receivers with square-law detectors". In: *Journal of Applied Physics* vol. 18 (1947), pp. 383–397.

[39] T. Kailath, "A view of three decades of linear filtering theory". In: *IEEE Transactions on Information Theory* vol. 20 (1974), pp. 146–181.

[40] K. Karhunen, "Über lineare Methoden in der Wahrscheinlichkeitsrechnung". In: *Annales Academiae Scientarium Fennicae* vol. 37 (1947), pp. 1–79.

[41] G. Kitagawa and W. Gersch, "A smoothness priors time-varying AR coefficient modeling of nonstationary covariance time series". In: *IEEE Transactions on Automatic Control* vol. 30 (1985), pp. 48–56.

[42] D. D. Kosambi, "Statistics in function space". In: *Journal of the Indian Mathematical Society* vol. 7 (1943), pp. 76–88.

[43] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory.* London: MIT Press, 1984.

[44] H. J. Landau and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier analysis and uncertainty - III: the dimension of the space of essentially time- and band-limited signals". In: *Bell System Technical Journal* vol. 41 (1962), pp. 1295–1336.

[45] M. Lavielle, "Optimal segmentation of random processes". In: *IEEE Transactions on Signal Processing* vol. 46 (1998), pp. 1365–1373.

[46] M. R. Lewis, *Evaluation of Vector Sensor for Adaptive Equalization in Underwater Acoustic Communication.* PhD thesis, 2014.

[47] Y. Li et al., "Time-varying model identification for time-frequency feature extraction from EEG data". In: *Journal of Neuroscience Methods* vol. 196 (2011), pp. 151–158.

[48] J. Lin et al., "Optimal tracking of time-varying channels: a frequency domain approach for known and new algorithms". In: *IEEE Transactions on Selected Areas in Communications* vol. 13 (1995), pp. 141–154.

[49] L. A. Lipporace, "Linear estimation of nonstationary signals". In: *Journal of the Acoustical Society of America* vol. 58 (1975), pp. 1288–1295.

[50] C. Liu, Y. Zakharov, and T. Chen, "Doubly selective underwater acoustic channel model for a moving transmitter/receiver". In: *IEEE Transactions on Vehicular Technology* vol. 61 (2012), pp. 938–950.

[51] L. Ljung and T. Chen, "What can regularization offer for estimation of dynamical systems?" In: *Proc. of the 11th IFAC Workshop on Adaptation and Learning in Control and Signal Processing* (Caen, France). 2013.

[52] L. Ljung, T. Chen, and B. Mu, "A shift in paradigm for system identification". In: *International Journal of Control* vol. 93 (2020), pp. 173–180.

[53] L. Ljung, M. Morf, and D. Falconer, "Fast calculation of gain matrices for recursive estimation schemes". In: *International Journal of Control* vol. 27 (1978), pp. 1–19.

[54] M. Loéve, "Sur les fonctions aléatoires du second ordre". In: *Comptes rendus de l'Académie des Sciences* vol. 220 (1945), p. 380.

[55] A. Marconato, M Schoukens, and J. Schoukens, "Filter-based regularisation for impulse response modelling". In: *IET Control Theory and Applications* vol. 112 (2016), pp. 194–204.

[56] G. Matz and F. Hlawatsch, "Fundamentals of time-varying communication channels". In: *Wireless Communications over Rapidly Time-Varying Channels* (2011), pp. 1–63.

[57] S. L. Miller and D. G. Childers, *Probability and Random Processes. With Applications to Signal Processing and Communications.* Elsevier Academic Press, 2004, pp. 291–295.

[58] E. Moulines, P. Priouret, and F. Roueff, "On recursive estimation for time varying autoregressive processes". In: *The Annals of Statistics* vol. 33 (2005), pp. 2610–2654.

[59] B. Mu, T. Chen, and L. Ljung, "Asymptotic properties of generalized cross-validation estimators for regularized system identification". In: *IFAC-PapersOnLine* vol. 51 (2018), pp. 203–208.

[60] B. Mu, T. Chen., and L. Ljung, "On asymptotic properties of hyperparameter estimators for kernel-based regularization methods". In: *Automatica* vol. 94 (2018), pp. 381–395.

[61] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory". In: *IEEE Transactions on Information Theory* vol. 39 (1993), pp. 1293–1302.

[62] A. Y. Ng, "Feature selection, L_1 vs. L_2 regularization, and rotational invariance". In: *Proceedings of the 21st international conference on Machine learning* (Banff, Kanada). 2004.

[63] M. Niedźwiecki et al., "Adaptive identification of underwater acoustic channel with a mix of static and time-varying parameters". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Singapore, Singapore). 2022.

[64] M. Niedźwiecki, "Functional series modeling approach to identification of nonstationary stochastic system". In: *IEEE Transactions on Automatic Control* vol. 33 (1988), pp. 955–961.

[65] M. Niedźwiecki, "Functional series modeling identification of nonstationary stochastic systems - the clipping technique". In: *1st IAESTED International Symposium on signal Processing and its Applications* (Barton, Australia). 1987.

[66] M. Niedźwiecki, *Identification of Time-varying Processes.* New York: Wiley, 2000.

[67] M. Niedźwiecki, "Locally adaptive cooperative Kalman smoothing and its application to identification of nonstationary stochastic systems". In: *IEEE Transactions on Signal Processing* vol. 60 (2012), pp. 48–59.

[68] M. Niedźwiecki and M. Ciołek, "Akaike's final prediction error criterion revisited". In: *Proc. of the 40th International Conference on Telecommunications and Signal Processing* (Barcelona, Spain). 2017.

[69] M. Niedźwiecki and M. Ciołek, "Generalized Savitzky-Golay filters for identification of nonstationary systems". In: *Automatica* vol. 108.108477 (2019).

[70] M. Niedźwiecki, M. Ciołek, and A. Gańcza, "A new look at the statistical identification of nonstationary systems". In: *Automatica* vol. 118.109037 (2020).

[71] M. Niedźwiecki, M. Ciołek, and A. Gańcza, "Fast basis function estimators for identi cation of nonstationary stochastic processes". In: *Proc. of the 27th European Signal Processing Conference* (A Coruna, Spain). 2019.

[72] M. Niedźwiecki and A. Gańcza, "Optimally regularized local basis function approach to identification of time-varying systems". In: *61st IEEE Conference on Decision and Control* (Cancun, Mexico). 2022.

[73] M. Niedźwiecki, A. Gańcza, and M. Ciołek, "On the preestimation technique and its application to identification of nonstationary systems". In: *59th IEEE Conference on Decision and Control* (Jeju Island, South Korea). 2020.

[74] M. Niedźwiecki, A. Gańcza, and P. Kaczmarek, "Identification of fast time-varying communication channels using the preestimation technique". In: *Proc. of the 19th IFAC Symposium, System Identification: learning models for decision and control* (Padova, Italy). 2021.

[75] M. Niedźwiecki and L. Guo, "Nonasymptotic results for finite-memory WLS filters". In: *IEEE Transactions on Automatic Control* vol. 36 (1991), pp. 198–208.

[76] M. Niedźwiecki and T. Kłaput, "Fast recursive basis function estimators for identification of time-varying processes". In: *IEEE Transactions on Signal Processing* vol. 8 (200), pp. 1925–1934.

[77] M. Niedźwiecki et al., "A new method of noncausal identification of time-varying systems". In: *Proc. of the 24th IEEE Signal Processing, Algorithms, Architectures, Arrangements, and Applications* (Poznań, Poland). 2020.

[78] M. Niedźwiecki et al., "Adaptive identification of sparse underwater acoustic channels with a mix of static and time-varying parameters". In: *Signal Processing* vol. 200.108664 (2022).

[79] M. Niedźwiecki et al., "Application of regularized Savitzky-Golay filters to identification of time-varying systems"". In: *Automatica* vol. 9 (2021), pp. 74622–74631.

[80] J. P. Norton, "Identification by optimal smoothing using integrated randowm walks". In: *Proceedings of the IEEE* vol. 123 (1976), pp. 451–452.

[81] D. L. Phillips, "A technique for the numerical solution of certain integral equations of the first kind". In: *Journal of the ACM* vol. 9 (1962), pp. 84–97.

[82] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification". In: *Automatica* vol. 46 (2010), pp. 81–93.

[83] G. Pillonetto et al., "Kernel methods in system identification, machine learning and function estimation: A survey". In: *Automatica* vol. 50 (2014), pp. 657–682.

[84] M. B. Porter, *Bellhop3d user guide.* Tech. rep. Heat, Light, and Sound Research Inc., 2016.

[85] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications.* Pearson Prentice Hall, 2007.

[86]   C. R. Rao, *Linear Statistical Inference and Its Applications*. New York: Wiley, 2002.

[87]   T. Subba Rao, "The fitting of nonstationary time-series models with time-dependent parameters". In: *Journal of the Royal Statistical Society B* vol. 32 (1970), pp. 312–322.

[88]   I. Reed, "On a moment theorem for complex Gaussian processes". In: *IRE Transactions on Information Theory* vol. 8 (1962), pp. 194–195.

[89]   R. Remmert, *Theory of Complex Functions*. New York: Springer, 1991.

[90]   A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures". In: *Analytical Chemistry* vol. 36 (1964), pp. 1627–1639.

[91]   R. W. Schafer, "What is a Savitzky-Golay filter?" In: *IEEE Signal Processing Magazine* vol. 28 (2011), pp. 111–117.

[92]   L. Shen et al., "Adaptive filtering for full-duplex UWA systems with time-varying self-interference channel". In: *IEEE Access* vol. 8 (2020), pp. 187590–187604.

[93]   L. Shen et al., "BEM adaptive filtering for SI cancellation in full-duplex underwater acoustic systems". In: *Signal Processing* vol. 191 (2022).

[94]   L. Shen et al., "Digital self-interference cancellation for underwater acoustic systems". In: *IEEE Transactions on Circuits and Systems II: Express Briefs* vol. 67 (2020), pp. 192–196.

[95]   L. Shen et al., "Finite-window RLS algorithms". In: *Signal Processing* vol. 198.108599 (2022).

[96]   D. Slepian, "Prolate Spheroidal Wave Functions, Fourier analysis and uncertainty - V: the discrete case". In: *Bell System Technical Journal* vol. 57 (1978), pp. 1371–1430.

[97]   D. Slepian and H. O. Pollak, "Prolate Spheroidal Wave Functions, Fourier analysis and uncertainty - I". In: *Bell System Technical Journal* vol. 40 (1961), pp. 43–63.

[98]   T. Söderström and P. Stoica, *System Identification*. New York: Prentice-Hall, 1988.

[99]   M. Stojanovic and J. Preisig, "Underwater acoustic communication channels: Propagation models and statistical characterization". In: *IEEE Communication Magazine* vol. 47 (2009), pp. 84–89.

[100] R. Tibshirani, "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society B* vol. 58 (1996), pp. 267–288.

[101] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*. Winston/Wiley, 1977.

[102] R. Tóth, *Modeling and Identification of Linear Parameter-Varying Systems*. Springer, 2010.

[103] M. K. Tsatsanis and G. B. Giannakis, "Modelling and equalization of rapidly fading channels". In: *International Journal of Adaptive Contorl and Signal Processing* vol. 10 (1996), pp. 159–176.

[104] M. K. Tsatsanis and G. B. Giannakis, "Time-varying system identification and model validation using wavelets". In: *IEEE Transactions on Signal Processing* vol. 41 (1993), pp. 3512–3523.

[105] M. Udell and A. Townsend, "Why are big data matrices approximately low rank?" In: *SIAM Journal on Mathematics of Data Science* vol. 1 (2019), pp. 144–160.

[106] H. L. Wei, S. A. Billings, and J. J. Liu, "Time-varying parametric modelling and time-dependent spectral characterisation with applications to EEG signals using multiwavelets". In: *International Journal of Modelling, Identification and Control* vol. 9 (2010), pp. 215–224.

[107] H. L. Wei, J. J. Liu, and S. A. Billings, "Identification of time-varying systems using multi-resolution wavelet models". In: *International Journal of Systems Science* vol. 33 (2002), pp. 1217–1228.

[108] E. T. Whittaker, "On a new method of graduation". In: *Proceedings of the Edinburgh Mathematical Society* vol. 41 (1922), pp. 63–75.

[109] W. Wirtinger, "Zur formalen theorie der funktionen von mehr komplexen Veränlichen". In: *Mathematische Annalen* vol. 97 (1927), pp. 357–375.

[110] R. A. Wooding, "The multivariate distribution of complex normal variables". In: *Biometrika* vol. 43 (1956), pp. 212–215.

[111]   Y. Zakharov, G. P. White, and J. Liu, "Low-complexity RLS agorithms using dichotomous coordinate descent iterations". In: *IEEE Transactions on Signal Processing* vol. 56 (2008), pp. 3150–3161.