Bayesian identification of linear dynamic systems

Darwish, M.A.H.

Published: 10/10/2017

Document Version
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

BAYESIAN IDENTIFICATION OF LINEAR DYNAMIC SYSTEMS

Synthesis of Kernels in the LTI Case and Beyond

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op

dinsdag 10 oktober 2017 om 16:00 uur

door

Mohamed Abdelmonim Hassan Darwish

geboren te Assiut, Egypte

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommisie is als volgt:

voorzitter:	Prof.dr.ir. A.B. Smolders
1^{e} promotor:	Prof.dr.ir. P.M.J. Van den Hof
copromotor:	Dr.ir. R. Tóth
leden:	Prof.dr.ir. J. Suykens (Katholieke Universiteit Leuven)
	Prof.dr. S. Weiland
	Prof.dr.ir. R. Pintelon (Vrije Universiteit Brussel)
	Dr. C. Rojas (Royal Institute of Technology, Stockholm)
	Dr.ir. T.J. Tjalkens

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Bayesian Identification of Linear Dynamic Systems

Synthesis of Kernels in the LTI Case and Beyond

Mohamed Abdelmonim Hassan Darwish

The research reported in this thesis is part of the research program of the Dutch Institute of Systems and Control (DISC). The author has successfully completed the educational program of the Graduate School DISC.

This research was financially supported by the Culture Affairs and Mission Sector, Ministry of Higher Education and Scientific Research, Government of Egypt.

A catalogue record is available from the Eindhoven University of Technology Library.
Bayesian Identification of Linear Dynamic Systems: Synthesis of Kernels in the LTI Case and Beyond /
by Mohamed A. H. Darwish. - Eindhoven: Technische Universiteit Eindhoven, 2017.
Proefschrift. - ISBN: 978-90-386-4352-6

This thesis was prepared using the LaTeX typesetting system.
Printed by: Gildeprint - the Netherlands.
Cover design: Sara H. El Sherif and Mohamed A. H. Darwish.

This thesis is dedicated to my beloved family

Summary

Bayesian Identification of Linear Dynamic Systems: Synthesis of Kernels in the LTI Case and Beyond

In the last few years, new avenues of data-driven modeling or so-called system identification have appeared due to the introduction of ideas stemming from the field of machine learning. One set of methodologies clustered around the so-called kernel based methods got serious attention in *Linear Time Invariant* (LTI) system identification due to their interpretation from the Bayesian point of view and their capability to realize an estimator that achieves regularization in *Reproducing Kernel Hilbert Spaces* (RKHSs). Such achievements are made possible via tailoring these learning techniques to dynamic systems by taking into account dynamic properties as stability. It has been shown that these new regularization based methods may outperform classical parametric approaches, i.e., maximum likelihood and prediction error methods, for the identification of stable LTI systems. The key feature of these learning approaches is that they circumvent the difficulties of model structure and model order selection and introduce a continuous optimization of the bias/variance trade-off based on a "nonparametric" form of the utilized model structure. The degrees of freedom of the estimation is kept restricted by incorporating prior knowledge of the unknown dynamic system, e.g., smoothness, stability, damping, resonance behavior, etc., through the kernel function that determines the hypothesis space for the estimation problem, i.e., which encodes the utilized "nonparametric" model structure. Hence, the choice of this kernel function is key to have a successful identification process in terms of a high accuracy model estimate.

The available kernel functions for the identification of impulse responses of LTI systems mainly focus on encoding smoothness and stability. These kernels came from static function estimation and ad hocly modified to enforce the decay of the estimated impulse responses without taking into account the dynamic aspects of these responses. Hence, it is essential to introduce kernels which are supported by system theory and allow for incorporating other dynamic properties, e.g., resonance behavior, into the kernel function. So, the resulting question is:

Research question 1: How to systematically synthesize kernel functions for linear systems that can encode/capture their dynamic behavior accurately?

On the other hand, nowadays, with the rapid advancement of technology, there is a growing need to accurately capture the increasingly *Nonlinear* (NL) and/or *Time-Varying* (TV) nature of new application designs. Moreover, the performance specifications of control systems, in terms of accuracy, reliability, robustness, energy consumption, etc., have been seriously increased. Thus, LTI modeling becomes insufficient to support model based control design methods such that the increasing performance specification can be fulfilled. On the other hand, dealing with such complicated systems, e.g., NL/TV, without any kind of structure is often found to be infeasible in practice in terms of modeling and control. Alternatively, advanced linear models have been introduced in the literature, e.g., *Piecewise Affine* (PWA), *Linear Time-Varying* (LTV), *Linear Parameter-Varying* (LPV) systems, etc. Among these models/classes, LPV systems, which can be seen as an intermediate step between the well-known LTI systems and the complicated NL/TV systems, have proven to provide an attractive framework to incorporate NL/TV phenomena with a wide representation capability of complex physical processes and at the same time preserve the linear structure of the representation, offering extensions of powerful LTI control approaches. However, the identification of the LPV model class is challenging task due to the difficulties associated with parameterizing the structural dependencies of the model on the so-called scheduling variable, denoted by p. This raises the following question:

Research question 2: How the promising approaches of Bayesian identification can be extended beyond the LTI case, i.e., towards LTV and LPV systems?

The main goal of this thesis is to address the above-mentioned two research questions. To answer these questions, this thesis focuses on presenting solutions for the following subgoals:

- Systematic utilization of the prior knowledge of the dynamic properties of the underlying system in the construction of kernels for Bayesian system identification.
- Automatic model structure selection and optimization of bias/variance trade-off of model estimates.
- For LTV and LPV systems:
  - Capturing structural dependencies directly from data.
  - Dealing with general noise scenarios in estimation.

To this end, this thesis includes the following contributions:

- Present a new class of kernel functions, which merges ideas from machine learning and system theory, suitable for the identification of LTI systems in both the time-domain, i.e., impulse response estimation, and the frequency-domain, i.e., transfer function estimation, in a Bayesian setting. This class of kernel functions is constructed based on Orthonormal Basis Functions (OBFs) that can be completely characterized via their generating poles. In this way, they offer a systematic approach to encode the expected dynamic properties of the system or flexibly adjust the model structure to it.

- Formulation of prediction error minimization for LTV and LPV systems from the view point of nonparametric Gaussian regression. The presented framework gives an interpretation of the estimation under a general noise scenario affecting the identification process, e.g., varying Box Jenkins model structures.
- A nonparametric Bayesian identification approach for series-expansion representation models of LTV and LPV systems, utilizing Output Error (OE) noise structure.
- A model structure learning approach for LPV models within the RKHS framework that is capable of determining the suitable dynamic order of the model (coefficient structure) and at the same time determine the underlying functional dependencies directly from data, with no prior parameterization of the p-dependent functions.

Contents

1 CHAPTER

Introduction

This thesis addresses data-driven modeling of finite dimensional, discrete-time and *Linear Dynamic Systems* (LDS). More specifically, utilizing the new developments in system identification stemming from the machine learning community, it addresses the open problems and associated challenges to efficiently deliver accurate linear models of the underlying physical process. To motivate the presented work and to formulate the intended research questions, first, in Section 1.1, the main principles of system identification will be reviewed. Section 1.2 motivates the interest in LDS and explores system classes beyond the classical *Linear Time-Invariant* (LTI) concept. Next, in Section 1.3 machine learning techniques that could be applied to dynamic systems are discussed. Then, the available approaches to data-driven modeling of LDS are overviewed in Section 1.4. The challenges and open problems of LDS identification are discussed in Section 1.5. Next, a brief discussion on the importance of *Orthonormal Basis Functions* (OBFs) as a promising tool for addressing some of these open problems are given in Section 1.6. The primary research questions and the subsequent subgoals are presented in Section 1.7. Finally, this chapter is ended with a brief overview of the contents and the main contributions.

1.1 Data-driven modeling of dynamic systems

Building a dynamic model based on first principle laws of physics, biology, chemistry, etc., requires detailed process knowledge from specialists, which might be even impossible to obtain if the required knowledge of first principles is missing. Such a modeling might also result in a highly complex mathematical description of the considered system with the need to perform dedicated experiments to estimate the model coefficients. An efficient alternative is to derive a mathematical model of the dynamic system on the basis of experimentally measured data. Such an approach is known as *data-driven* modeling or so-called *system identification*.

Table 1.1: The identification cycle

Step 1.	Experiment design, data collection, and data preprocessing.
Step 2.	Selection of model structure and parametrization.
Step 3.	Choice of the identification criterion.
Step 4.	Estimation of a model that is optimal with respect to the criterion.
Step 5.	Validation of the resulting model estimate.

The underlying process of system identification is often called the *identification cycle* (Ljung 1999), see Table. 1.1 for the involved steps and Figure 1.1 for a pictorial illustration. In the following, a brief overview of the involved steps is given in order to support the subsequent discussion.

1.1.1 Identification cycle

Step 1. Experiment design and data preprocessing

The purpose of this step is to choose an excitation signal which is used to actuate the considered system and produce an output that maximizes the information content about the underlying dynamics in the measured *Input/Output* (IO) data. Different aspects should be taken into account and reflected in the designed signal, e.g., sufficient number of data points, sampling frequency, levels of operation, etc. Two important requirements should be satisfied by the excitation signals: i) *persistency of excitation*, i.e., the input signals produce an output response which has enough information content to describe the dynamic relation of the system and have enough information content to distinguish between different models in the considered model class; ii) the practical applicability of the input signal, more specifically, white noise inputs are rather classical, exciting all frequencies, however, for a practical consideration, such input signals are often infeasible for physical actuation, and other exciting signals should be considered, e.g., random binary noises, frequency sweeps, multi-sines, etc.

Additionally, if the system is unstable or contains lightly damped dynamics it may be necessary to decide to conduct the experiment in closed loop. Also, data preprocessing such as removing outliers and the mean value of the signal and the attenuation of noises and disturbances should be performed before using the measured data for identification.

Dedicated experiments can also be used to detect and quantify the level of nonlinearity or time-variation exhibited by the system. These dedicated experiments can also be used to decide which model class is the most appropriate in terms of the utilization objective to capture the dynamics of the system.

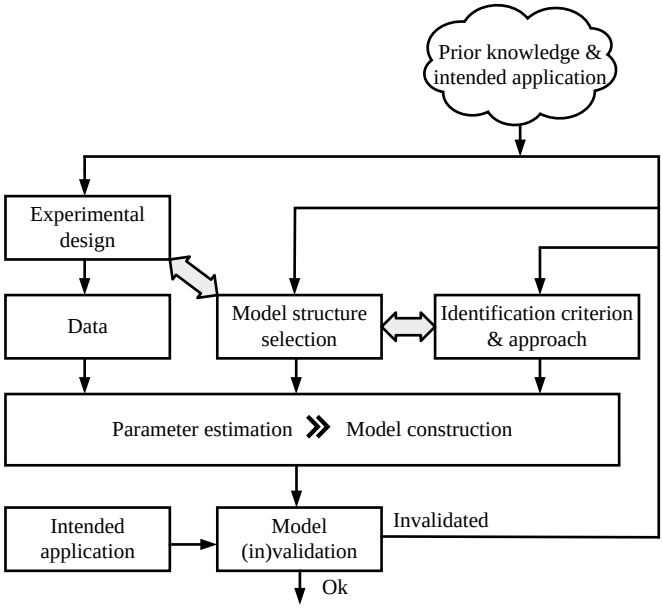


Figure 1.1: The identification cycle based on Ljung (1999).

Another important choice is the measurement setup itself used to gather experimental data. Such a choice has an impact on the modeling process and the utilization of the model afterwards. For example, for control applications with digital controllers, a *Zero Order Hold*[1] (ZOH) setup is typically used and hence the identification process is carried out with the aim to obtain *Discrete-Time* (DT) models. Differently, Band Limited (BL) setup is used for physical interpretation or to obtain models for the synthesis and tuning of controllers in CT and as a consequence CT models should be utilized in model estimation. Hence, these choices of data acquisition have consequences on the choice of the model structure, i.e., Step 2.

Step 2. Choice of the model structure

A model structure is a set of candidate models in which a suitable description of the system is searched for. This step is considered to be the most crucial step of the identification cycle. The selection of the model structure involves various aspects: i) representation form, e.g., *state-space*, IO, *series-expansion*, etc.; ii) parameterization and type of noise modeling. This includes postulating parametric models to describe both process and noise dynamics, where the size of the model set is im-

[1]The ZOH device is a signal hold instrument providing a *Continuous-Time* (CT) signal which is constant till the device is commanded to change it to a new value in a piecewise constant manner.

portant, e.g., the number of parameters and the order of the model structure, as it largely affects the well-known bias/variance trade-off of the estimation, see Section 1.1.2; iii) the complexity of the estimation algorithm that delivers the model estimates, where undesired local solutions of the estimation and non-uniqueness of the optimal solution need to be considered.

Step 3. Choice of the identification criterion

Various identification criteria can be considered. These can be seen as a mathematical formulation of the performance measure of the estimated models that define the user's purpose or expectation towards the model of the plant. The most common criterion is to identify the model that provides the best one-step-ahead predictions in terms of the smallest possible mean squared error between the measured outputs and the predictions.

Step 4. Model Estimation

Once both the model structure and identification criterion have been chosen, the next step is to solve the optimization problem associated with the identification criterion to get the model estimate.

Step 5. Model (in)validation

When a model has been estimated it should be evaluated to decide whether the model is "good enough" for the intended purpose of the user. To this end, another experimental data set, which is not used in the estimation step and is known as the *validation* data set, is often used with the estimated model to compute a predicted/simulated response and then compare it with the measurements based on the same identification criterion or various other measures of model quality. At this stage, it can be decided whether the estimated model is accepted or a refinement step is needed, which is accomplished by iterating and making more appropriate choices, e.g., different model order, till the obtained model passes the validation test.

1.1.2 Which system class and which parameterization?

Aiming at estimating a dynamic model that "best" describes the underlying physical process, two important decisions have to be taken. Firstly, the choice of a suitable system class, i.e., *Linear* (L), *Nonlinear* (NL) or *Time-Invariant/Time-Varying* (TI/TV) dynamics. Such a selection largely depends on the dynamic behavior of the process that we are dealing with. Moreover, the selected class affects the whole identification cycle, specifically, the representation capability, computational complexity, expected estimation accuracy, etc. Secondly, for the chosen system class, a parameterized model needs to be postulated, where the structure and the order of

that model are crucial choices, as has been discussed in the previous section. The size of the parameterization introduces the well-known dilemma of bias/variance trade-off. More specifically, by reducing model order, which often decreases the number of the to be estimated parameters, the bias error increases while the variance error is decreased. On the other hand, by increasing model order, i.e., introducing more parameters, the bias error decreases at the expense of increasing the variance.

In this thesis, we focus on data-driven modeling of *Linear Dynamic Systems* (LDS) aiming at achieving an optimal trade-off automatically in terms of bias/variance which leads to a fundamentally different look at system identification as in the previously discussed classical setting. In the next section, we motivate our interest in LDS, discussing the importance of *Linear Time-Invariant* (LTI) systems and advanced linear model structures.

1.2 Why Linear Dynamic Systems?

Automatic control has been known and used for more than 2000 years (Bissell 2009). It is mainly about employing a "*feedback*" concept to guarantee the stability and performance expectations of the underlying physical process. Motivated by the observation that many physical systems exhibit linear behavior at the desired operating points and that feedback linearizes the dynamics of the "*closed-loop*" system, LTI models have been used extensively in practice to describe the dynamic behavior of the considered systems. The LTI framework has become a mature field in terms of both modeling and control design with vast industrial experience accumulated through the years.

Nowadays, with the rapid advancement of technology, there is a growing need to accurately capture the increasingly NL and/or TV nature of new application designs, e.g., wafer scanner in *lithography*, modern process control, automotive applications, etc., see Figure 1.2 for some of these applications. Moreover, the performance specifications of control systems, in terms of accuracy, reliability, robustness, energy consumption, etc., have seriously increased. For example, the moving stages of wafer scanner machines in lithography require fast and accurate positioning in the nanometer scale. As a result, LTI modeling becomes insufficient to support model based control design methods such that the increasing performance specification can be fulfilled. On the other hand, dealing with such complicated systems, e.g., NL/TV, without any kind of structure is often found to be infeasible in practice in terms of modeling and control. Alternatively, advanced linear models have been introduced in the literature that are aiming at describing complex NL/TV behavior via a linear structure, e.g., *Switched Affine* (SA) (Djemai and Defoort 2015), *Piecewise Affine* (PWA) (Paoletti et al. 2007), *Linear Time-Varying* (LTV) (Marcovitz 1964), *Linear Parameter-Varying* (LPV) (Tóth 2010) systems, see Figure 1.3.

In general, these linear system classes can be classified into two main categories according to the nature of switching-time/parameter variation: i) models

Figure 1.2: Some examples of modern applications that exhibit NL/TV dynamics. Upper left: wafer scanners used in lithography. Upper right: automobiles. Lower left: aircrafts, e.g., F-16 Fighting Falcon. Lower right: modern process control.

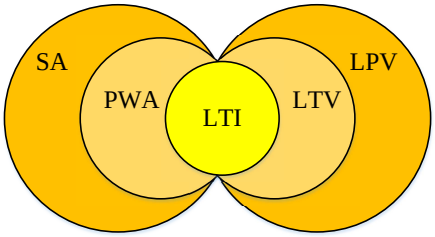


Figure 1.3: The linear system classes in the literature. These classes are aiming at describing complex NL/TV dynamic behavior via a linear structure.

that are defined as collection of linear/affine models connected by switches that are indexed by an additional discrete-valued variable, the so-called discrete state, i.e., SA systems. The class of PWA systems are considered as a special case of SA systems, where the discrete state is determined by a polyhedral partition of state-input domain. These models are equivalent to several classes of hybrid systems (Paoletti et al. 2007) with a wide scope of applications, e.g., regime switching in power electronics (Aguilera et al. 2014); econometrics (Hamilton 1990); control applications (Yin et al. 2009), etc.; ii) models that have a time/parameter variation, i.e., LTV/LPV models, respectively, due to a physical phenomenon or a scheduling parameter that varies smoothly as a function of time. Some applications include aeroplane dynamics during take off and landing (Dimitriadis and Cooper 2001); control of crane dynamics (Abdel-Rahman et al. 2003), etc.

In the sequel, we consider the second class of systems, i.e., LTV/LPV system classes. More specifically, in the sequel, we use LDS to refer to LTI, LTV and LPV system classes. Among these models/classes, LTV/LPV systems, which can be seen as an intermediate step between the well-known class of LTI systems and complicated NL/TV systems, have proven to provide an attractive framework to

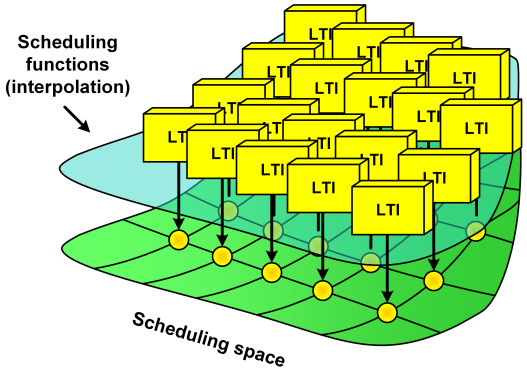


Figure 1.4: The mechanism of gain-scheduling (Tóth 2008): interpolation of local LTI models/controllers of the plant to approximate the global behavior on the entire operation regime, i.e. scheduling space.

incorporate NL/TV phenomena with a wide representation capability of complex physical processes and at the same time preserve the linear structure of the representation, offering extensions of powerful LTI control approaches. Note that, LPV systems are a natural extension of LTV systems. More specifically, for the LPV system class, the trajectory of the scheduling signal is assumed to be not known in advance, whereas for LTV system class the scheduling signal, i.e., time, follows a known linear trajectory. Hence, the available techniques to deal with LPV systems can be easily applied to the special case of LTV systems.

In many practical situations NL/TV systems can be well-approximated by local LTI models that describe the behavior of the plant around some operating points. Moreover, engineers working in industry prefer the application of LTI control design as they have a vast experience to deal with such systems due to the available attractive approaches of control design, e.g., optimal and robust control. Inspired by these observations, the concept of LPV systems have been introduced in Shamma and Athans (1992) through the idea of gain-scheduling. More specifically, the NL system is linearized at some operating points resulting in a collection of LTI models of the plant. An interpolation function, known as *scheduling function* that depends on the current operating point, is used to deliver a global model that describes the entire operating regime, see Figure 1.4 for a pictorial illustration of the above-mentioned mechanism (Tóth 2008). An external and measurable signal, known as the *scheduling signal* and it is denoted by p, is used to describe the change of the operating point. This interpolation of local linear dynamics is the so-called local approach to obtain an LPV model of the plant which approximates the original NL behaviour. Alternatively, it is possible to capture the whole behaviour of the NL system (without an approximation) by an LPV model through the concept of embedding. The latter approach is called global modeling.

In mathematical terms, the considered class of systems, i.e., LDS, can be char-

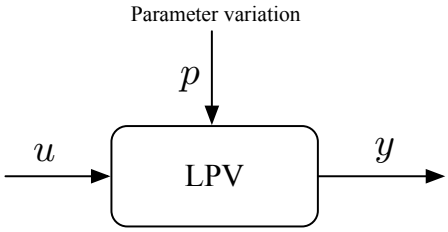


Figure 1.5: Input-Output signal flow of an LPV system. The dynamic relationship between the input and the output of the system is linear, but depends on an exogenous signal, the so-called scheduling signal.

acterized in DT[2] with an *Infinite Impulse Representation* (IIR). More specifically, for LTI systems the dynamic relation between the input signal u and output signal y is a linear dynamic relation, which can be read as

$$y(t) = \sum_{i=0}^{\infty} \mathrm{g}_i q^{-i} u(t), \tag{1.1}$$

where t is the discrete time, q denotes the forward time shift operator, i.e.,$q^{-i}u(t)= u(t-i)$, $y : \mathbb{Z} \to \mathbb{R}^{n_{\mathbb{Y}}}$, where $\mathbb{Z}$ is the set of integers and $\mathbb{R}$ is the set of real numbers, $u : \mathbb{Z} \to \mathbb{R}^{n_{\mathbb{U}}}$, and the coefficients $\{\mathrm{g}_i\}_{i=0}^{\infty}$ are known as the Markov parameter matrices. If the dynamic relation itself is dependent on time or an external signal, the so-called *scheduling signal* p, then the resulting system classes are known as LTV or LPV, respectively. For LPV systems, which is a generalization of LTV systems as has been discussed before, the dynamic relation can be characterized as a convolution in terms of u and p, which in DT can be read as

$$y(t) = \sum_{i=0}^{\infty} \mathfrak{g}_i(p(t)) q^{-i} u(t), \tag{1.2}$$

where $p : \mathbb{Z} \to \mathbb{P} \subset \mathbb{R}^{n_{\mathbb{P}}}$. Furthermore, the coefficients $\mathfrak{g}_i(p)$ are functions of the scheduling variable that define the varying linear dynamic relation u and y. See Figure 1.5 for the schematic view of such a relation. If the functions $\mathfrak{g}_i$ are assumed to be dependent only on the instantaneous value of the scheduling signal, i.e., $\mathfrak{g}_i : \mathbb{P} \to \mathbb{R}^{n_{\mathbb{Y}} \times n_{\mathbb{U}}}$, then the underlying LPV system is said to have a *static dependence* on p, otherwise, if $\mathfrak{g}_i$ are assumed to be dependent on the past values of p, i.e., $\mathfrak{g}_i : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_{\mathbb{Y}} \times n_{\mathbb{U}}}$, then the dependence is called *dynamic*. If the scheduling signal p is replaced by t, the resulting system is known as LTV. For a constant scheduling signal, i.e., $p(t) = \bar{\mathrm{p}}$ with $\bar{\mathrm{p}} \in \mathbb{P}$ being a constant $\forall t \in \mathbb{Z}$, (1.2) becomes equivalent to the LTI system (1.1), where each $\mathfrak{g}_i(p)$ is constant and corresponds to the i-th Markov parameter of that LTI system, i.e., g_i. Thus, LPV systems can

[2]Note that, for the DT model to be equivalent to its original CT model all continuous free signals (e.g. input signals) of the system are required to be generated by an ideal ZOH, i.e., they need to be piecewise constant. For LPV systems, this condition is required to hold for both the input and the scheduling signals (in a synchronized manner) Tóth et al. (2010).

be seen to be similar to LTI systems, but their signal behavior is different due to the variation of the g_i parameters.

The LPV concept has been proven to be successful in many applications. It can be said that control design for such class of systems has become a mature field, offering many approaches, e.g., (Scherer 1996), that guarantee stability, optimal performance and robustness over the entire operating regime expressed in terms $\mathbb{P}$. Moreover, from the representation capability, it has been shown that many NL systems can be converted into an LPV form, e.g., (Abbas et al. 2014; Donida et al. 2009). Motivated by these developments, LPV systems have found their way to industry and gained popularity, e.g., in process control (Bachnas et al. 2014), aerospace applications (Marcos and Balas 2004), CD players (Dettori and Scherer 2001), wafer scanners (Wassink et al. 2005), to name a few. With such advanced control design methods, accurate models are needed to support the available methods. A lot of work has been done regarding modeling and identification of LPV systems aiming at delivering accurate models. However, the developments in that area/direction are still lagging behind control design with many open problems and challenges that need to be taken into account due to the difficulties associated with parameterizing the structural dependency of the model on p.

1.3 The appearance of machine learning

Machine learning is concerned with the design of techniques that are able to automatically extract information and learn structure from data. An interesting subarea of machine learning is *regression*[3], which is about estimating (learning) an unknown function from a given set of observations. The classical approach to deal with this problem, i.e., regression, is by postulating a finite-dimensional hypothesis space, i.e., utilize a parametric model that depends on a finite-dimensional vector of parameters that are needed to be estimated from data. For instance, postulate a model that consists of a linear combination of a predefined set of basis functions. Such an approach has some difficulties, e.g., the choice of an appropriate set of basis functions and their number. These difficulties are circumvented by formulating the problem as function estimation, possibly in an infinite-dimensional space, i.e., by employing high order flexible models, the so-called nonparametric models (Bishop 2006). The basic idea is to determine the right complexity of the model, i.e., to determine both the basis functions and their effective number, from data and *high-level* prior knowledge about the unknown function, e.g., smoothness. Such knowledge is more relaxed and easier to be available than imposing a specific structure on the model, which might be restrictive and limits the representation capability of the model. The question now is how can we encode such prior knowledge about the unknown function into the estimation problem?

The modern nonparametric approaches mainly use *regularization* techniques introduced extensively in the so-called *inverse* problem literature (Tikhonov and

[3]Regression was originally developed in the field of statistics and has been extensively studied in other fields, like machine learning.

Arsenin 1977; Bertero 1989) in conjunction with *Reproducing Kernel Hilbert Spaces* (RKHSs) (Aronszajn 1950; Schölkopf and Smola 2002), to control the flexibility of the employed nonparametric models and ensure a well-posed solution. It has been shown that for every RKHS, there is one and only one positive definite function, known as *reproducing kernel* or simply kernel function, denoted by K, (Aronszajn 1950). Moreover, it can be also shown that every function of the considered RKHS inherits properties such as smoothness and integrability of the associated kernel function. Hence, instead of characterizing a whole space, it is enough to design a positive definite kernel function K that encodes the desired properties of the function to be estimated and hence such properties will be inherited by all the functions within the resulting RKHS. The design of the structure of K is concerned with choosing a parameterized form of it with some unknown parameters, the so-called *hyperparameters*, denoted by β, which can express a wide variety of properties, but at the same time restrict the high degree of freedom by encoding the expected properties, e.g., smoothness. Furthermore, it is important that the associated restrictions are sensitive to the choice of β, i.e., β can be efficiently used to decrease the RKHS associated with K towards a set capturing the properties of the unknown function. At the same time, β must be also low dimensional such that its optimization can be efficiently performed. Such an estimation approach, i.e., regularization in RKHS, has also a statistical interpretation in a Bayesian setting, i.e., in *Gaussian Process Regression* (GPR) (Rasmussen and Williams 2006). More specifically, the unknown function is assumed to be a particular realization of a zero-mean Gaussian process with a certain covariance function that is identical to the kernel function K associated with the considered RKHS. This Bayesian setting provides an efficient technique to estimate the unknown hyperparameters, i.e., β, that parameterize the kernel function K from data, i.e., by maximizing the marginal likelihood (Rasmussen and Williams 2006; MacKay 2003). This provides automatic model structure selection whose efficiency depends on the choice of K (Pillonetto and Chiuso 2015).

Such achievements can be extended to the identification of dynamic systems to deal with the issues related to model order/structure selection. For instance, the impulse response of an LTI system, i.e., g_i in (1.1), can be seen as a function and estimated by regularization techniques from machine learning. However, the above-mentioned approaches are mainly suited for estimating nonlinear functions, where the underlying relation is static and the utilized kernel functions focus on imposing smoothness of the hypothesis space. Hence, the available kernel functions in machine learning, e.g., *Gaussian kernel* (Rasmussen and Williams 2006), *spline kernel* (Wahba 1990), etc., are not directly suited for dynamic systems identification, in terms of estimating their impulse responses, as the dynamic relationship should be taken into account. For example, in addition to smoothness, other dynamic properties should be encoded via the kernel function, e.g., stability in terms of the decay of the impulse response, resonance behavior, etc.

In the next section, parametric identification of LDS and the recently introduced regularization approaches to system identification of LDS is reviewed.

1.4 Data-driven modeling of LDS systems

As has been explained in Section 1.1, the model structure selection is the most crucial step in the identification cycle as it has a significant effect on the final model estimates. This step involves many challenging decisions that need to be taken, e.g., selection of model structure, model order, etc. At this stage, two main streams of identification techniques can be distinguished, namely parametric and nonparametric techniques. In parametric approaches, a finite-dimensional model is chosen that depends on a prior chosen set of parameters. On the other hand, nonparametric methods that originate from machine learning and use models of possibly infinitely order, where the number of effective parameters is flexible and is decided from data. Note that, the latter methods can be seen as an extension of the "classical" nonparametric methods, which are known in system identification for long time (Ljung 1999), e.g., frequency response and impulse response estimation.

In this section, we discuss the available identification approaches to identify LDS, specifically, in the LTI, LTV and LPV system classes, from both parametric and nonparametric point of views.

1.4.1 Data-driven modeling of LTI systems

Time-domain

Data-driven modeling of LTI systems is a well-established field (Ljung 1999; Söderström and Stoica 1989; Pintelon and Schoukens 2012), where the main stream estimation methods are:

1. Subspace approaches which are based on the numerically robust *Singular Value Decomposition* (SVD) and *Least-Squares* (LS) techniques. Handling *Multi-Input Multi-Output* (MIMO) systems is straightforward with these methods and various numerically efficient implementation of these schemes are available (van Overschee and de Moor 1996; Chiuso 2007);

2. Set membership and worst-case identification (Milanese and Vicino 1991; Helmicki et al. 1991), which have been introduced motivated by the need to deliver models with hard bounds on model errors, i.e., to make use of such models in robust control (Zhou et al. 1995). However, these techniques might lead to conservative results (Hjalmarsson and Ljung 1994);

3. *Maximum Likelihood / Prediction Error Minimization* (PEM) (Ljung 1999; Söderström and Stoica 1989) approaches. Such methods provide a well-understood framework for consistency and stochastic interpretation of the model estimates and offer a large class of plant and noise models.

PEM approaches have been considered as the dominant parametric technique for identifying LTI systems for a long time (Ljung 1999; Söderström and Stoica 1989). These methods typically include the following steps:

1. Postulate a finite-dimensional parametric model structure;

2. Estimate parametric models of different orders by the minimization of the quadratic loss-function, denoted in the sequel by ℓ_2, of the prediction error.

3. Choose the model order, i.e., one model among the set of estimated candidates, based on any of the available model validation techniques as *Cross-Validation* (CV) (Ljung 1999), *Akiake Information Criterion* (AIC) (Akaike 1974), *Bayesian Information Criterion* (BIC) (Schwarz 1978), etc.

Theoretical properties of estimators obtained after model selection, the so-called *Post Model Selection Estimators* (PMSEs), are generally hard to study (Leeb and Pötscher 2005). Sample properties of PMSEs, such as impulse response estimators or predictors, when tested on experimental data, may depart sharply from those predicted by "standard" statistical theory, i.e., without model selection, which suggests that PEM should be asymptotically efficient for Gaussian innovations. See e.g., (Pillonetto et al. 2011a, Section 6) where PEM approaches, equipped with model validation techniques have proven to deliver unsatisfactory results for short and noisy observations. The choice of the model order is related to the well-known bias/variance dilemma, i.e., low model order leads to under-modeling and accordingly biased estimate and increasing the model order will lead to over-parameterized models, which leads to estimates with high variance. As a result, the obtained model will perform poorly when used to predict a new unseen input.

A different approach to deal with the bias/variance dilemma is to resort to regularization. The main idea behind regularization is that instead of minimizing the variance of unbiased models, we allow for biased models with reduced variance so to arrive at a smaller *Mean Squared-Error* (MSE). Some approaches are, e.g., ℓ_1/*Least Absolute Shrinkage and Selection Operator* (LASSO) (Tibshirani 1996), nuclear norm (Hjalmarsson et al. 2012) and *Non-Negative Garotte* (NNG) (Breiman 1995). However, the tuning of the regularization parameter in real world applications of these methods is often found to be a difficult task. In Rojas et al. (2013), an approach, the so-called SPARSEVA is proposed, which provides an automatic tuning of the amount of regularization to ensure consistency of the regularized estimator. However, ℓ_1 regularization is employed to perform parameter selection/order selection rather than optimizing the bias/variance trade-off. So, this points towards automatization of classical model order selection into a single step approach.

In the last few years, a new avenue of system identification have appeared due to the introduction of ideas stemming from machine learning, i.e., regularization techniques, known as kernel-based methods, see Section 1.3. Such achievements are made possible by tailoring these learning techniques to dynamic systems by taking into account dynamic properties as stability into the kernel function. It has been shown that these new regularization based methods may outperform classical parametric approaches, i.e., PEM methods, in the identification of stable LTI systems (Pillonetto et al. 2011a; Pillonetto and De Nicolao 2010; Pillonetto et al. 2014; Chen et al. 2012). The key feature of these learning approaches is that they

circumvent the difficulties of model structure and model order selection and introduce a continuous optimization of the bias/variance trade-off based on a nonparametric form of the utilized model structure. The degrees of freedom of the estimation is kept restricted by incorporating prior knowledge of the unknown dynamic system, e.g., smoothness, stability, damping, resonance behavior, etc., through the kernel function that determines the hypothesis space for the estimation problem, i.e., which encodes the utilized nonparametric model structure.

More specifically, in Pillonetto and De Nicolao (2010), a new kernel-based approach for stable LTI system identification has been introduced, where the impulse response is modeled as a realization of a *Gaussian Process* (GP), whose covariance function includes information on *Bounded-Input Bounded-Output* (BIBO) stability in addition to smoothness. The underlying paradigm, i.e., Bayesian approaches to identification, is much older, see, e.g., (Kitagawa and Gersch 1984, 1985, 1996) where the main interest has been the spectral analysis of time series and (Goodwin et al. 2002, 1992) where a similar approach has been proposed to quantify the under-modeling error, i.e., bias error. However, the real difference compared to these previous Bayesian methods is that in Pillonetto and De Nicolao (2010); Pillonetto and De Nicolao (2011), a probabilistic prior is formulated directly on the unknown impulse response. It has been shown that the minimum variance estimate belongs to an RKHS, whose kernel function coincides with the covariance function of the considered GP, see Figure 1.6 for a schematic view that describes the main difference between the classical parametric and regularized nonparametric approach to system identification (Pillonetto et al. 2011b). These achievements have been realized by introducing a new class of priors, i.e., kernel/covarinace function, the so-called *Stable Spline* (SS) kernels, which is a modified version of the well-known *Spline* kernel function (Wahba 1990) so that it includes information on BIBO stability. In Pillonetto et al. (2011a), this kernel-based approach has been extended to the PEM setting, where a one-step-ahead predictor form of the estimator has been formulated in the nonparametric setting. More specifically, the resulting "optimal" form of the one-step-ahead predictor can be seen as a system with two inputs (past outputs and inputs of the predicted system) and one output (output predictions). Therefore, estimating the predictor for LTI systems boils down to estimating two impulse responses.

Since the introduction of the SS kernel, it has become evident that the structure of the kernel function and its representation capability to encode a wide range of expected dynamic properties are the keys to further improve the efficiency of these methods. Accordingly, many kernel structures have been introduced in the literature to embed various prior knowledge, e.g., *Diagonal kernel* (DI), *Diagonal Correlated* (DC), *Tuned Correlated* (TC) (Chen et al. 2012), *Rank-1* kernel known as *Output Error* (OE) kernel (Chen et al. 2013), constructive state-space models induced kernels (Chen and Ljung 2014), to name a few. Moreover, in (Chen and Ljung 2015b,c; Chen and Ljung 2016), two different methods of designing kernel functions suitable for impulse response estimation are presented from a machine learning and system theory perspectives. It is worth to mention that the abovementioned kernels are considered to be single structure kernels, and hence not suitable to describe the dynamics of complicated systems with distinct modes,

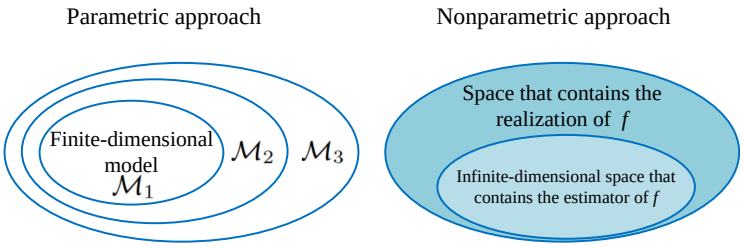


Figure 1.6: Left: Parametric approach to system identification. $\{\mathcal{M}_i\}_{i=1}^3$ denote finite-dimensional spaces of different complexity. Model order is typically chosen by criteria such as AIC or BIC requiring the solution of a (possibly) nonlinear optimization problem for each postulated model and relying upon likelihood functions which are only asymptotically exact. Right: Nonparametric approach to system identification using GPR. The unknown system is defined by a realization from a zero-mean Gaussian random field f whose covariance (kernel) encodes the prior knowledge, e.g., smoothness and stability. Model order selection is replaced by estimation of few hyperparameters that parameterize the kernel, obtained by optimizing a likelihood function that is exact, irrespective of the sample size, and accounts for the uncertainty of f. Once such parameters are determined, the minimum variance estimate of f is available in closed form and belongs to a (generally infinite-dimensional) RKHS (Pillonetto et al. 2011b).

i.e., slow and fast modes. Thus, multiple structure kernels have been introduced in (Chiuso et al. 2014; Chen et al. 2014), that handle such systems with multiple and distinct time constants. In (Marconato et al. 2016, 2017), a different approach to the above-mentioned Bayesian setting is introduced, where the prior knowledge is injected at the cost function level instead of including such knowledge in the kernel/covariance function itself. This allows to model low-pass, band-pass and high-pass systems, and systems with one or more resonances. In Chen and Ljung (2015a), regularization methods for impulse response estimation have been extended to the more general *Orthonormal Basis Functions* (OBFs) model structure estimation, where the generating poles of the OBFs are considered as hyperparameters and are tuned within the considered Bayesian setting , i.e., by maximizing the marginal likelihood, see Figure 1.7 for an overview of the available methods for LTI systems identification.

There have been significant efforts spent on understanding and analyzing these kernel functions to give more insights into the representation capability of various dynamic properties, i.e., stability, over-damped, under-damped, multiple distinct time constants, resonance behavior, etc., (Chen et al. 2016; Carli et al. 2017; Chen and Ljung 2015c,b; Dinuzzo 2015; Chen et al. 2015). Such an effort has resulted in characterizing the RKHSs associated with these kernels, i.e., like DC, TC, with well-understood spectral decomposition, i.e., eigenvector-eigenvalue decomposition that generalizes the eigenvector-eigenvalue decomposition of a positive-definite matrix. Moreover, efficient algorithms for implementing these kernel-

based methods for impulse response estimation have been studied in Carli et al. (2012); Chen and Ljung (2013). More specifically, in Carli et al. (2012), an efficient algorithm for the large scale scenario, i.e., data rich situation, has been proposed. However, this method mainly works for a family of kernel functions that have a well-defined spectral decomposition, e.g., SS kernels, whereas the approach proposed in Chen and Ljung (2013) relies on a QR factorization technique to efficiently and accurately evaluate the cost function involved in the marginal likelihood optimization. Furthermore, the presented approach in Chen and Ljung (2013) can deal with both large data sets and possibly ill-conditioned computations.

Frequency-domain

It has been shown that identification in time- and frequency-domain can be seen as equivalent problems as the involved data carry the same information (Schoukens et al. 2004). However, such information is represented differently, hence, it may be easier to access it in one domain than in the other. Therefore, in the following, we also cover frequency-domain identification of LTI systems.

Nonparametric estimation of *Transfer Functions* (TFs) of LTI systems provides valuable information about the dynamics of the system under consideration that can be used further to obtain an accurate parametric model through model validation and model selection (Ljung 1999; Pintelon and Schoukens 2012). The evaluation of the TF on the unit circle will be called *Frequency Response Function* (FRF) and it has been studied extensively in the literature (Antoni and Schoukens 2007; Pintelon and Schoukens 2012), to name a few.

One main challenge in the data-driven estimation of FRF is the transient effect, which is due to the fact that the input and output signals are not periodic or their periodicity does not match the length of the measurement window. As a result, most of the available approaches try suppressing the transient effect in different ways. More specifically, via spectral analysis as in Schoukens et al. (2006), or via a frequency-dependent smoothing procedure that is applied to the *Empirical Transfer Function Estimate* (ETFE) (Stenman et al. 2000). More recent approaches are estimating both the FRF and the transient simultaneously (Pintelon and Schoukens 2012), e.g., the *Local Polynomial Method* (LPM) (Pintelon et al. 2010a,b), which uses a local polynomial smoother, and the *Local Rational Method* (LRM) (McKelvey and Guérin 2012), which uses a local rational function as a smoother. Note that both methods, i.e., LPM and LRM, provide a set of local models centered around the bins of the used *Discrete Fourier Transform* (DFT), where the interpolation between the DFT bins is still an open question. As a consequence, the stability of their estimates is not defined. Alternatively, inspired by new developments in nonparametric estimation of LTI impulse response models in the time-domain (Pillonetto et al. 2014; Chen et al. 2012), regularized frequency domain estimates of both the FRF and the transient effects within the GPR framework has been introduced in Lataire and Chen (2016). More specifically, both the FRF and the transient are assumed to be a realization of a zero-mean real/complex GP (Schreier and Scharf 2010) with a certain covariance (kernel) function that encodes the relevant prior

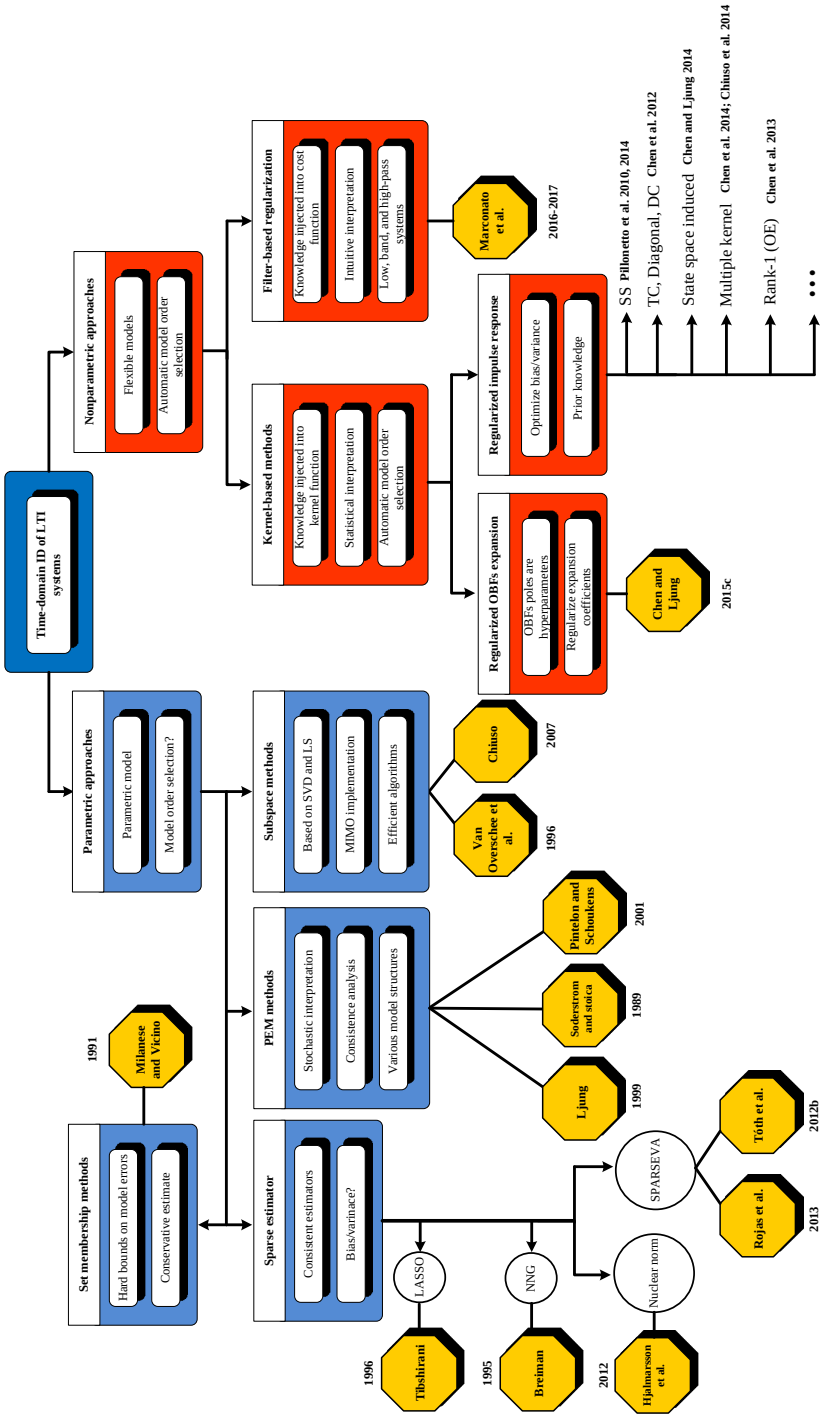


Figure 1.7: Available methods of LTI identification in the time-domain with some representative references.

knowledge about the system, e.g., smoothness and stability, etc. The direct formulation of the estimation problem in the frequency domain offers many advantages: i) it allows the estimation to be performed in a limited frequency band; ii) it allows for an efficient implementation for continuous-time systems.

One critical aspect that is related to the aforementioned approach is the design of the kernel function. It has to be flexible enough to describe a wide range of dynamic properties, e.g., stability, resonance behavior, damping, etc., and, at the same time, are parameterized by a small number of hyperparameters. In Lataire and Chen (2016), the kernels from the time-domain, e.g., DC, SS kernels, have been formulated in the frequency-domain. Moreover, it has been shown that for both of these kernels the resulting estimates are stable, i.e., all poles of the estimated FRF lie inside the unit circle.

1.4.2 Data-driven modeling of LTV systems

In the following, we briefly overview identification of LTV systems as a special case of LPV systems.

Significant efforts have been spent on developing efficient identification approaches for LTV systems from data. Differently from the LTI case, special attention has to be paid to both the parameterization of the dynamics and time-variation, see the introduction of (Pintelon et al. 2015) for the classification of the available techniques for LTV system identification.

Parametric identification techniques in the time-domain include, e.g., (Spiridonakos and Fassois 2009; Tsatsanis and Giannakis 1993; Verhaegen and Yu 1995), and in the frequency-domain, e.g., (Lataire and Pintelon 2011), where systems are usually described by differential (or difference) equations with time-dependent coefficients. Nonparametric techniques have also been introduced in this setting, e.g., (Lataire et al. 2012) and the references therein.

Alternatively, in Pillonetto (2008), the estimation of *Continuous-Time* (CT)-LTV state space models in the time-domain has been formulated as a regularization in RKHS. However, a matrix differential equation must be solved at each iteration of the algorithm. Whereas, in Lataire et al. (2017), a kernel-based approach has been proposed to identify CT-LTV models, where the time-varying coefficients are identified nonparametrically. Furthermore, to avoid approximating time derivatives, a mixed time- and frequency-formulation is adopted. Moreover, Bayesian approaches for LTI models have been successfully extended to the LTV case. More specifically, a SS-based estimator to identify LTV systems has been considered in Pillonetto and Aravkin (2014), where an additional hyperparameter has been included that plays the role of a forgetting factor which can be optimized in the considered Bayesian setting via maximizing the marginal likelihood. An online identification approach of TV systems in Bayesian setting, that extends the recently introduced kernel-based methods from the LTI framework (Pillonetto and De Nicolao 2010), has been introduced in Prando et al. (2016). A forgetting factor is used to track the TV nature of these systems. Moreover, to cope with the real-time constraints associated with updating the model hyperparameters in such a

scenario, i.e., online identification, these hyperparameters are updated with a one gradient step in the marginal likelihood optimization.

1.4.3 Data-driven modeling of LPV systems

Motivated by the need for accurate and low-complexity LPV models to exploit the available control synthesis approaches in practice, a wide range of LPV identification approaches have been developed in the past years (Tóth 2010). In general, for such a procedure to be successful, two main ingredients are needed: data containing measured information about the dynamics of the system, and prior knowledge in the form of assumptions on the expected behavior of the system. One of the most important of these is the selected model structure and the corresponding model set within which the identification method should find an estimate of the plant. Most identification methods dedicated to LPV modeling in the literature a priori assume a given suitable model set and focus on the estimation process whether the problem is formulated in a state-space form (e.g., (van Wingerden and Verhaegen 2009; Sznaier and Mazzarro 2003)), series-expansion including IIR, expansion in terms of OBFs (Tóth 2010) or IO representation (e.g., (Bamieh and Giarré 2002; Laurain et al. 2010; Piga et al. 2015)). However, the selection of this suitable set is rather complicated in practice as it is outmost desired to estimate an accurate model of the real system in terms of the utilization objective using as few parameters as possible (*parsimony principle*). Note that accuracy of estimated models is often also affected by the size of the parametrization in terms of the achievable limit on the variance of the model estimate. In this respect, the LPV modeling problem exhibits two main challenging issues: (i) the classical questions of determining the "suitable" dynamic order of the model, input delay and noise structure; (ii) to determine the underlying functional dependency of the coefficients on p such that they have the least possible complexity for adequately representing the variation of the dynamics.

In the following, we focus on LPV-IO models, where almost all the existing parametric approaches are formulated in DT and static dependence on p is assumed. Most of the well-known methods of LPV-IO identification are summarized in the following discussion.

Traditionally, the problem of estimating an LPV model on the basis of data is addressed within the parametric setting. In this setting, the underlying functional dependency of the coefficients on p is parameterized in terms of a priori chosen set of basis functions, e.g., polynomials, trigonometric functions, etc. The next step is to estimate the associated parameters with the assumed model structure. Two main categories can be distinguished: i) the local approaches, that rely on the gain-scheduling concept, where a set of LTI models, the so-called "frozen models", are identified at constant scheduling trajectories and then interpolated to deliver a global model (Zhu and Xu 2008; Zhu and Ji 2009; Bachnas et al. 2014), and ii) the global approaches where a parameterized LPV model structure is identified directly based on a global data set with varying scheduling trajectory. Various approaches that fall within the second category, i.e., the global approaches, are summarized below. Set membership methods from the LTI framework have been

also applied to LPV systems, where the noise is treated as deterministic uncertainty (Belforte et al. 2005; Cerone and Regruto 2008). In the PEM framework, which has been successfully extended to the LPV case (Tóth et al. 2012a), all the well-known noise model structures from the LTI literature are also formulated for LPV systems, e.g., *Auto-Regressive models with eXogenous input* (ARX), OE, *Auto-Regressive Moving Average model with eXternal signal* (ARMAX), *Box Jenkins* (BJ) model structures, and series-expansion models including IIR and OBFs models. Within this framework, identification boils down to perform a linear regression in case of ARX and OBFs model structures, when the underlying dependencies are parameterized linearly in terms of a priori chosen basis functions (Bamieh and Giarré 2002; Giarré et al. 2006; Wei 2006; Tóth et al. 2009b, 2011a). Moreover, in case of LPV-OBFs model structure, a suitable set of OBFs has to be chosen that has a wide representation capability over the whole scheduling domain. To cope with the latter issue, in Tóth et al. (2009a, 2011a), a selection scheme of the basis functions known as *Fuzzy Kolmogorov c-means clustering* (FKcM) algorithm, which is a joint application of the *Kolmogorov n-width* (KnW) theory (Oliveira e Silva 1996) and *Fuzzy c-Means* (FcM) clustering (Jain and Dubes 1988), has been proposed that is capable of asymptotically estimate the optimal set of OBFs, based on the availability of a collection of pole locations that are obtained from the local linear behavior of the LPV system.

In case of more general noise model structures, e.g., OE, ARMAX and BJ, a nonlinear optimization problem is needed to be solved to get the model estimate. Such an optimization can be solved by employing for instance *gradient-based* minimization approaches (Zhao et al. 2012). However, such approaches can have serious issues with local minima and computational complexity. A pseudo linear regression is an alternative to the computationally expensive nonlinear optimization (Tóth et al. 2011a). However, the statistical analysis of the results in that case, i.e., pseudo linear regression, is difficult due to the iterative nature of the procedure (Tóth et al. 2012a). Alternatively, *Refined Instrumental Variable/ Simplified Refined Instrumental Variable* (RIV/SRIV) methods (Tóth et al. 2012c; Laurain et al. 2010) have been extended to deal with LPV-BJ models, where the noise model is assumed to have LTI dynamics.

If the number of basis functions that are needed to parameterize the coefficient functions are not known a priori, a possible solution is the use of a model structure where the functional dependencies are parameterized in terms of an extensive set of basis functions such that the structure is capable of explaining a rich set of possible dynamics, where the appropriate sub-structure is decided from data. Such a decision is commonly achieved by employing model structure selection tools AIC, BIC, CV, etc. These tools can be seen as imposing a sparsity pattern on the parameters, because they determine a model sub-structure (where the estimated model should be found), by forcing some of the parameters of the overall model to be zero. Due to the size of parametrization in the LPV case, application of these selection tools on real-world-sized problems quickly becomes infeasible. Hence, to achieve structural selection via enforcing a sparsity pattern, ℓ_1 regularization based sparse estimators and shrinkage methods methods such as the NNG (Breiman 1995), the LASSO (Tibshirani 1996) and SPARSEVA (Rojas

et al. 2013) have been extended and successfully applied in the LPV case, see Tóth et al. (2009c, 2012b). Although, these methods are capable of achieving model structure selection in terms of the main challenging issues associated with LPV modeling, their efficiency strongly depends on adequate *a priori* selection of the basis functions which is left to rest on the shoulders of the user, see Figure 1.8 for an overview of the available parametric approaches for the identification of LPV-IO models.

Similar to the LTI case, this points towards automatization of classical model order selection, where such a selection is performed together with detecting structural dependency from data. One approach to avoid the dilemma of parameterizing the coefficient function is to resort to nonparametric methods. In (van der Maas et al. 2015, 2016a), (van der Maas et al. 2016b), a nonparametric FRF modeling of LPV systems has been presented, where the behavior of an LPV system is assumed to be a smooth function of the frequency as well as the scheduling variable. Moreover, in Hsu et al. (2008), a nonparametric approach based on dispersion function method has been proposed, where no prior knowledge of the underlying dependencies is required. However, such an approach does not allow for the incorporation of prior knowledge (if available), which might result in a more accurate estimate, into the estimation problem. Furthermore, the considered noise model is restricted to the LPV-ARX model structure. Alternatively, different regularization techniques, have been used to identify nonparametric models of LPV-IO systems. Specifically, the so-called kernel-based methods, which offer attractive approaches to capture the underlying dependencies directly from data without specifying any parameterization in terms of fixed basis functions. In these methods, a kernel function is introduced that acts as a basis generator driven by observed data. The main kernel-based approaches of LPV nonparametric identification in the literature are: i) *Least Squares-Support Vector Machine* (LS-SVM) methods (Vapnik 1998; Suykens et al. 2002), where the considered models are restricted to LPV-ARX noise models, e.g., (Tóth et al. 2011b; Piga and Tóth 2013; Duijkers et al. 2014). An extension of these methods in case of the presence of uncertainty in the scheduling signal has been introduced in Abbasi et al. (2014). Furthermore, to preserve the attractive properties of these approaches and in the same time overcome the drawbacks in the estimation of LPV models in a general noise setting, an IV-LS-SVM method has been introduced in Laurain et al. (2012). It has been shown that such an extension, i.e., IV-LS-SVM, results in unbiased estimates for a general noise setting on the expense of increasing the variance; ii) GP methods (Rasmussen and Williams 2006), where a confidence quantification of the estimate is available in addition to an automatic way to tune the unknown hyperparameters, that parameterize the kernel function, from data via marginal likelihood optimization (Golabi et al. 2014, 2017). However, the considered models are restricted to LPV-ARX noise models. An extension of these methods in case of the presence of uncertainty in the scheduling signal, i.e., additive noise on p, has been introduced in Abbasi et al. (2015). See Figure 1.9 for a schematic view of the available nonparametric methods of LPV-IO models.

The above-mentioned kernel-based methods have been successfully extended to LPV state-space models, under the assumption that the states are measurable,

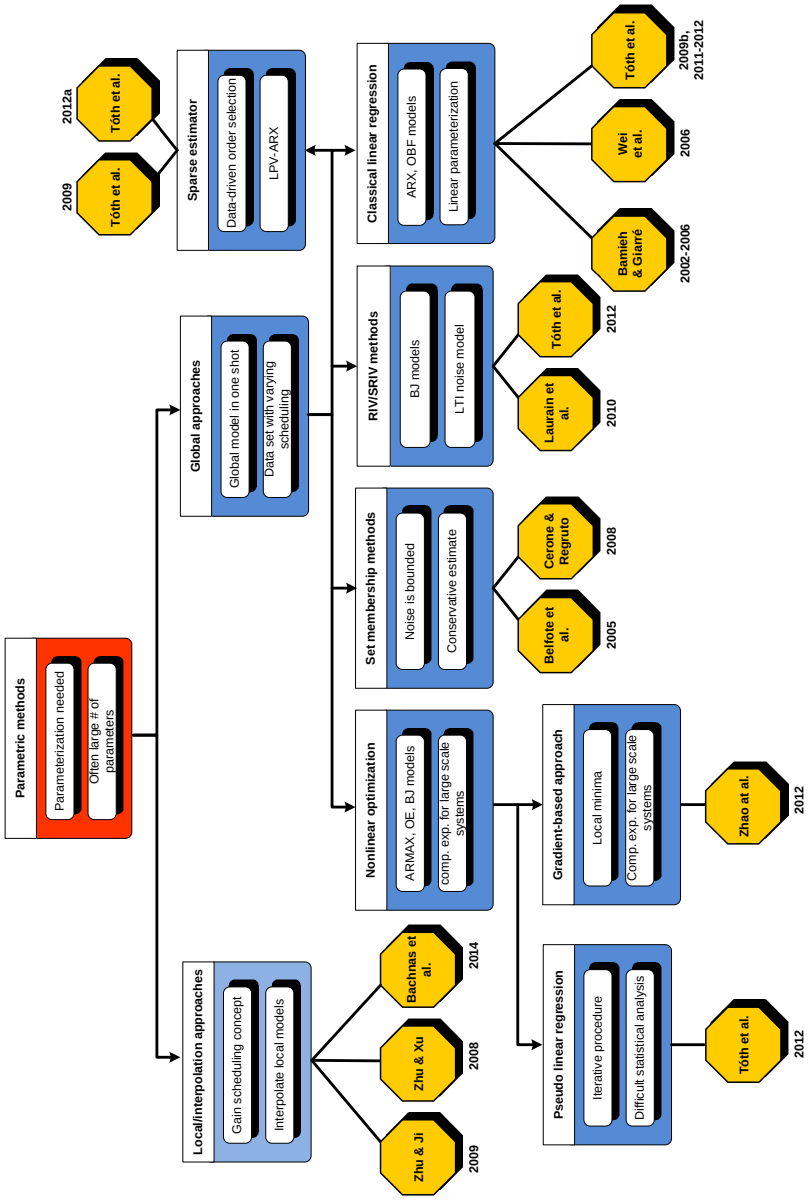


Figure 1.8: Available parametric methods of LPV-IO identification with some representative references.

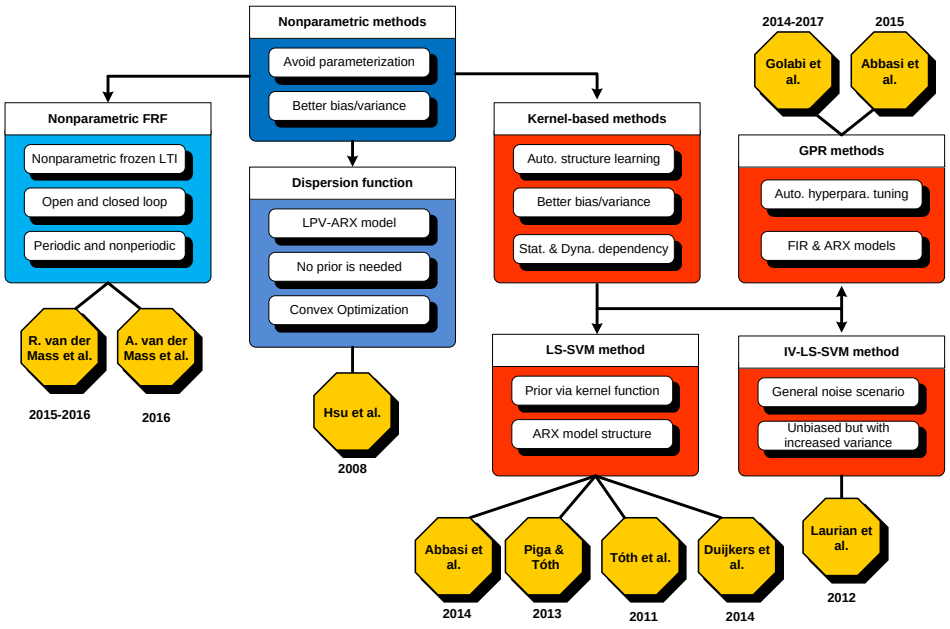


Figure 1.9: Available methods of nonparametric identification of LPV-IO models with some representative references.

offering a solution for estimating the dependency structure from data, e.g., LS-SVM (Lopes dos Santos et al. 2014; Rizvi et al. 2015b), IV-LS-SVM (Rizvi et al. 2015a). Recently, the restriction of a known state signal has been overcome via the application of Kernelized canonical correlation analysis (Rizvi et al. 2017).

1.5 Challenges and open problems

In the previous section, we have discussed the classical/parametric approaches to identify LDS based on observed data and we have seen the challenges associated with these approaches as model structure and order selection. Moreover, we have discussed machine learning techniques that have been extended to identification of LDS to tackle the above-mentioned challenges. However, to have a successful identification process based on these machine learning techniques, the most crucial step is the design of an appropriate kernel function that has a simple structure and at the same time has the capability to represent a wide range of expected dynamic properties of the unknown system. Such a design task becomes even more challenging for advanced linear models, i.e., LPV-IO models. The most crucial challenges and open problems are collected into the following list:

1. The available kernel functions for identifying LTI systems mainly focus on encoding smoothness and stability. These kernel functions have been bor-

rowed from static function estimation, where the main focus is to encode smoothness, then they have been modified to enforce stability constraints without taking into account the dynamic aspects of the to be captured system. Hence, it is essential to introduce kernels for linear systems which are supported by system theory and allow for the incorporation of other dynamic properties, e.g., resonance behavior, into the kernel function.

2. The available nonparametric estimators for LPV systems only consider a restrictive noise model structure, i.e., the LPV-ARX model structure. A computational efficient extension to the more general $p-$dependent noise model, i.e., LPV-BJ setting, is missing in the literature.

3. Most of the kernel-based methods, e.g., GP methods, for linear systems are restricted to IO and IIR models. However, as an extension of IIR models, OBFs model structure enjoys a wide representation capability, but have some issues associated with their identification from data: i) the choice a suitable set of OBFs; ii) parameterizing the coefficient functions; iii) guaranteeing the convergence of the estimated expansion; iv) handling the possible dynamic dependency on the scheduling signal. In line with the concepts in Chen and Ljung (2015a), it is essential to investigate how to extend kernel-based methods to such model structures both in the TV and PV cases.

4. In LPV nonparametric identification, kernel-based methods offer a solution for estimating the underlying structural dependency directly from data. However, the classical problem of selecting the model structure, i.e., model order, number of coefficient functions, delay, etc., has not been addressed like in the LTI case leaving the complexity/accuracy trade-off open.

1.6 Perspectives of OBFs based kernels

OBFs are a complete orthonormal basis functions for the space $\mathscr{RH}_2$, which is the Hilbert space of complex functions that are square integrable on the unit circle and analytic outside of it (Heuberger et al. 2005; Ninness and Gustafsson 1997). Their correspondence, i.e., the inverse $Z-$transform of these functions, in time-domain span a complete orthonormal basis for $\mathcal{R}\ell_2$ which is the space of squared summable real-valued sequences. These OBFs are generated by a cascaded network of *all-pass* functions, which are completely characterized, modulo the sign, by their generating poles. The spaces spanned by OBFs are RKHSs with a well-defined reproducing kernel, which is directly defined by the OBFs.

The OBFs provide a systematic way to represent dynamic systems with a long history of analysis in system theory (Ninness and Gustafsson 1997; Heuberger et al. 1995; Patwardhan et al. 2006; Nalbantoglu et al. 2003). There have already been few attempts to introduce OBFs based kernels for impulse response estimation in the Bayesian setting, e.g., (Chen and Ljung 2015a). However, the proposed OBFs based kernels do not perform well compared, e.g., with the TC kernel, as shown in (Chen and Ljung 2015a, Section V). Moreover, the formulation

of these kernels in the frequency-domain has been also given (Chen and Ljung 2015a, Equation 17). However, the stability of the estimated FRF by using such kernels has not been discussed. Furthermore, the problem of choosing the proper number of basis functions to be used has not been addressed, which hampers the utilization of this idea in the Bayesian estimator.

In this thesis, we show how the attractive properties of OBFs can be used to fill in the gap between data-driven modeling of dynamic systems and machine learning approaches.

1.7 Research questions and goals

In the previous part, it has been discussed that machine learning approaches, when tailored to dynamic systems identification by including the stability constraint, can provide a better bias/variance trade-off and could, in many cases, outperform classical approaches. However, kernel functions that can systematically describe other dynamic properties are missing from the literature. This results in the following problem statement:

- Research question 1 -

How to systematically synthesize kernel functions for linear systems that can encode/capture their dynamic behavior accurately?

On the other hand, we have discussed that LTI modeling becomes insufficient to support model-based control techniques under the need to address NL/TV behavior in recent applications. We have also mentioned that such complex behavior can be described with advanced linear models, i.e., LPV models that can be seen as an intermediate step between LTI and NL/TV systems. However, identification of LPV model class is a challenging task due to the difficulties associated with parameterizing the structural dependencies of the model on the so-called scheduling variable and selecting the model structure/order, number of coefficient functions, delay, dealing with general noise scenarios, etc. This raises the following question:

- Research question 2 -

How the promising approaches of Bayesian identification can be extended beyond the LTI case, i.e., towards LTV and LPV systems?

The main goal of this thesis is to address the above-mentioned two research questions. To answer these questions, this thesis focuses on presenting solutions for the following subgoals:

1. Systematic utilization of the prior knowledge of the dynamic properties of the underlying LTI system, e.g., stability, resonance behavior, etc., in construction of kernels for Bayesian system identification. More specifically, investigate how OBFs based kernels can support machine learning-based approaches in dynamic system identification both in the time- and frequency-domains.
2. Investigate how to extend the Bayesian methods for LTI system identification to LPV models under a PEM setting and how to handle general noise scenarios.
3. Investigate how kernel-based methods can be extended to the identification of series-expansion models, e.g, LPV-IIR and LPV-OBFs model structures, to tackle the challenges associated with the identification of such models.
4. Investigate how to jointly reconstruct the scheduling-variable dependencies and the model order (coefficient structure) directly from data, with no prior parametrization of the p-dependent functions.

In the following, these subgoals are explained in more details.

Subgoal 1

Kernel-based methods provide an attractive framework for identification of LTI systems both in the time- and frequency-domain. However, constructing a kernel function that can, systematically, describe a wide range of dynamic properties with a low-dimensional parameterization is still missing from the literature. With OBFs, the dynamic properties are directly encoded via the generating poles of these basis. However, when utilizing the OBFs model structure or constructing a kernel function based on OBFs for the purpose of system identification, we face two issues: i) the choice of an appropriate set of OBFs, that has representation capability of the underlying system; ii) the choice of an effective number of these basis functions. These two issues are completely related to each other, i.e., with a "wrong" choice of the basis, a long expansion is needed while with a "well-chosen" basis, a short expansion is sufficient to achieve the same prediction capability. We are aiming at having a data-driven approach to decide on both issues.

Subgoal 2

It has been discussed that Bayesian methods for impulse response estimation has been extended to the PEM setting in the LTI case, i.e., to estimate a nonparametric model for the "optimal" predictor. Such an approach is followed to avoid model structure selection and to deal with general noise scenarios. In the LPV case, next to the similar questions of model order and noise structure selection, the question of model parameterization becomes even more involved as it also includes the question how to parameterize the dependency of the model on the scheduling

variable. To make correct decisions regarding the latter problem a lot of prior knowledge is needed which is not often available in many practical applications. Hence, it is important to investigate how the Bayesian approach of the LTI case can be extended to the LPV case delivering completely structure-free learning of the dynamic relation that includes the data-driven estimation of the dependency structure on the scheduling variable. Such a setting, if properly formulated, can provide an efficient way to handle LPV model estimation even under a general BJ noise scenario.

Since we are aiming at Bayesian estimation of LPV models in a predictor form, it must be investigated how to design a kernel function that encodes the prior knowledge about the considered predictor, e.g., stability, possible class of structural dependencies on p. Furthermore, in order to present solutions to practical situation, which involve MIMO systems, the above-mentioned investigations should be performed in the MIMO setting.

Subgoal 3

In the LTI case, OBFs model structures offer an attractive methodology to accomplish data-driven modeling and even further enhancements of the Bayesian methods as discussed in Subgoal 1. To have a successful identification process with these models, a suitable set of OBFs needs to be estimated. The Bayesian framework provides an approach to estimate these OBFs from data and at the same time keep the variance of the estimated expansion coefficients low. Since utilization of OBFs model structures is also attractive in the LPV case due to their wide representation capability, it is highly relevant to investigate how the Bayesian approach to identify LTI-OBFs model structures can be extended to the LPV-OBFs model structures. By that, we are also aiming at tackling the problems associated with the parametric identification of these model structures. This involves synthesis of a suitable kernel function that can encode: i) expected structural dependency of the expansion coefficients on p; ii) stability of the underlying system, i.e., to guarantee the convergence of the estimated expansion.

Subgoal 4

The main motivation for Subgoals 2 and 3 is to avoid the problem of model order, noise structure selection and parameterization of the coefficient dependencies by employing Bayesian approaches to obtain a nonparametric estimate of the underlying dynamic relation of the data-generating system. In the LTI case, the resulting impulse response models can be directly and efficiently realized in other representation forms like IO or state-space models to be further utilized, e.g., for control synthesis. However, in the LPV case, the involved realization and model reduction theories are too complex and applicable for only low truncation orders of the involved IIRs. Hence, it becomes a question how to achieve nonparametric identification of LPV models, e.g., in an IO form directly, where the underlying

dependencies of the coefficients are estimated as functions, but at the same time tackle model structure selection directly from data.

Note that in the LPV literature there has been already attempts to achieve structure selection or nonparametric learning of the coefficient dependencies but this has never been accomplished in a joint fashion. Hence, it is important to investigate how regularized methods can be extended to the identification of LPV-IO models to resolve the above mentioned challenge in a nonparametric setting.

1.8 Overview of the contents and results

Next to this introduction chapter and the conclusions in Chapter 7, this thesis consists of five chapters. Figure 1.10 presents an overview of the chapters and the relation among them.

The introductory Chapter 2 is devoted to introduce key system theoretic notions for LTI systems. Moreover, we give a brief introduction to the Hilbert and Hardy spaces that are related to LTI systems. The second part of this chapter is devoted to the concept of OBFs, their definitions in the time- and frequency-domain and the spaces spanned by them. Finally, we briefly discuss the classical PEM setting of identifying LTI models.

Chapter 3 is devoted to the introduction of kernel-based methods in machine learning. First, we define the regression problem from both its classical and regularization points of view. We also introduce the statistical interpretation of kernel-based methods from the Bayesian perspective. We then provide a brief discussion on the numerical implementation and computational complexity of these methods. Finally, we discuss the connection between regularization in RKHS and GPR. The importance of this chapter is that it introduces the basic concepts from machine learning required for developing new tools for dynamic system identification.

In Chapter 4, we introduce a novel class of kernel functions based on OBFs, which are able to describe a wide range of dynamic properties in a systematic way. First, we discuss the modifications that are needed to be accomplished in kernel-based methods to make them applicable to dynamic system identification, i.e., imposing the stability constraint in the kernel function. Then, we introduce OBFs based kernels both in time-domain to identify impulse response models and in frequency-domain to identify the FRF of a stable LTI system. This provides a direct answer for Subgoal (1). This chapter is based on the papers Darwish et al. (2017d, 2015b, 2017c).

In Chapter 5, first, we briefly review the classical PEM framework for LPV systems. Then, we extend Bayesian identification of LTI systems under a PEM setting to LPV systems. More specifically, we consider a MIMO LPV data-generating system affected by $p-$dependent noise dynamics, i.e., an LPV-BJ setting. First, we follow a Bayesian approach to identify the one-step-ahead predictor in a GPR setting, where the one-step-ahead predictor can be seen as a summation of two sub-predictors associated with the input and output signals. The main contribution is

the design of suitable kernel functions that can encode the prior knowledge about these sub-predictors taking into account their specific dynamic and dependency structures. This provides a direct answer for Subgoal (2). Finally, the presented Bayesian approach from the previous part is extended to series-expansion models, i.e., LPV-IIR and LPV-OBFs models. More specifically, we summarize the identification of LPV-OBFs models and discuss the associated challenges with identifying these model structures, e.g., the choice of a suitable set of OBFs and the need to guarantee the convergence of the estimated expansion. Then, we show how such challenges can be tackled in a Bayesian setting. This provides direct answers for Subgoals (2) and (3). This chapter is based on the papers Darwish et al. (2017a, 2015a, 2017b).

In Chapter 6, we formulate a unified framework for the identification of LPV-IO models in an RKHS setting, where both model order and structural dependencies are estimated from data. This provides a direct answer for Subgoal (4). This chapter is based on the paper Laurain et al. (2017).

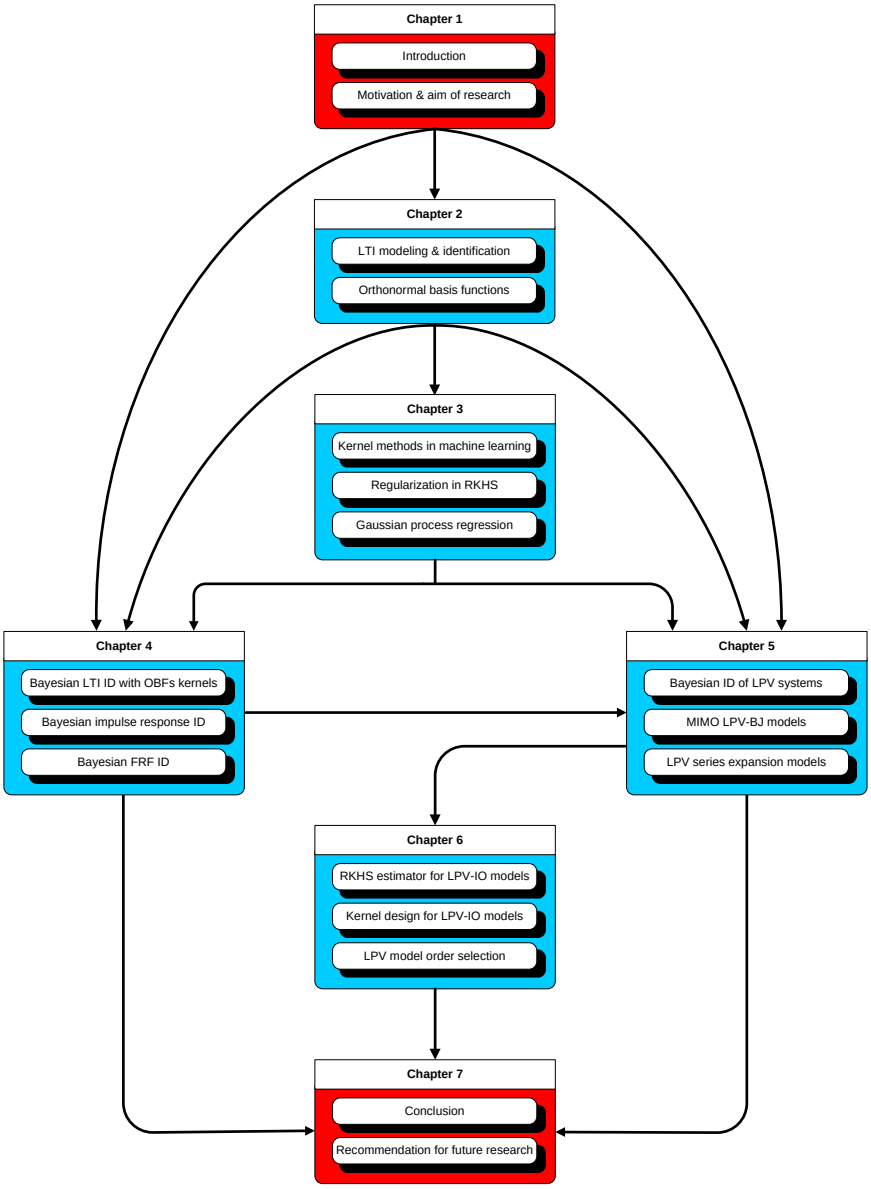


Figure 1.10: Structure of the thesis.

2 CHAPTER

LTI Systems and OBFs

This chapter is devoted to the introduction of representation, modeling and identification of LTI systems and of a general class of OBFs. Section 2.1 presents the class of LTI systems along with the corresponding representation forms, while the related Hilbert and Hardy spaces are introduced in Section 2.2. In Section 2.3, a brief introduction to a general class of OBFs is given followed by the classical approach for modeling and identification of LTI systems in Section 2.4.

2.1 LTI systems

A dynamic system, in a mathematical sense, can be seen as a mapping or an operator that assigns an output signal to a certain input signal (or variables that can be treated as input signals), in the sense that the output is completely determined by the input and initial conditions. See Figure 2.1 for a visual description, where u is the input, y is the output and $\mathcal{G}$ denotes a dynamic system. In this section, we consider the class of LTI systems, which is considered the simplest dynamic systems. Systems within this class have been successfully used in an enormous number of engineering applications to describe or approximate a wide range of physical phenomena. More specifically, we focus on *Finite Dimensional* LTI (FD-LTI) systems, denoted in the sequel by $\mathcal{F}$, whose time-invariant signal relations can be described by a real-rational and proper transfer function.

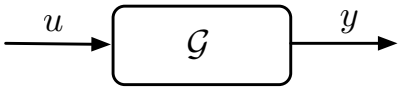


Figure 2.1: Block diagram of the dynamic system $\mathcal{G}$.

2.1.1 Representations of LTI systems

An FD-LTI system can be represented in different representations (forms), e.g., rational transfer function, impulse response, state-space, etc., with different degrees of representation efficiency, i.e., in terms of the required number of coefficients to represent a system in the considered domain with a certain level of accuracy. In the following, we present different representations of FD-LTI systems.

Consider a DT FD-LTI system $\mathcal{F}$ as

$$y(t) = G(q)u(t), \tag{2.1}$$

with $y : \mathbb{Z} \to \mathbb{R}^{n_\mathbb{Y}}$ and $u : \mathbb{Z} \to \mathbb{R}^{n_\mathbb{U}}$ such that the system $\mathcal{F}$ is represented by the transfer operator $G(q)$ and

$$G(q) = \sum_{k=0}^{\infty} \mathfrak{g}(k) q^{-k}, \tag{2.2}$$

where $\mathfrak{g} = \{\mathfrak{g}(k)\}_{k=0}^{\infty}$ is the (im)pulse response sequence of the system with $\mathfrak{g}(0)$, $\mathfrak{g}(1), \ldots$ being known as the impulse response/Markov coefficients. Such a response is equal to the response (output) of the system for a unit pulse input at zero. Accordingly, the IO mapping (2.1) can be written as

$$y(t) = \sum_{k=0}^{\infty} \mathfrak{g}(k) u(t-k) = (\mathfrak{g} \circledast u)(t), \tag{2.3}$$

where $(\mathfrak{g} \circledast u)(t)$ denotes the convolution between the impulse response $\mathfrak{g}$ and the input u at time t. Let $G(z) : \mathbb{C} \to \mathbb{C}$, with $z \in \mathbb{C}$ being the $Z-$variable and $\mathbb{C}$ being the complex plane. The transfer function of the system $\mathcal{F}$ is defined as

$$G(z) = \mathscr{Z}\{\mathfrak{g}(k)\} = \sum_{k=0}^{\infty} \mathfrak{g}(k) z^{-k}, \tag{2.4}$$

which is the $Z-$transform of $\mathfrak{g}$ with a corresponding *Region Of Convergence*[1] (ROC). Such a function, i.e., $G(z)$, for the considered class of FD-LTI systems is a rational transfer function that can be expressed as a ratio of finite order polynomials of z. It is called real-rational, if the coefficients of the numerator and denominator polynomials are real. It is called proper if $\lim_{|z|\to\infty} G(z) < \infty$ and strictly proper if in addition $\lim_{|z|\to\infty} G(z) = 0$, which implies that $\mathfrak{g}(0) = 0$. Substitution of z by $e^{\mathrm{j}\omega}$, with $\mathrm{j} = \sqrt{-1}$, gives the frequency response of the DT system for $\omega \in [-\pi, \pi]$, i.e., the FRF denoted by $G(e^{\mathrm{j}\omega})$.

In case of causal[2] DT-FD-LTI system $\mathcal{F}$, i.e., a system represented by a proper $G(z)$, a state-space representation of $\mathcal{F}$ is also available:

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{2.5}$$

[1] The region of convergence is the set of points in $\mathbb{C}$ for which the summation associated with the $Z-$transform converges.

[2] A causal system is a system where the output depends on the past and current inputs, but not on future inputs.

where the 4-tuple (A, B, C, D) are matrices with appropriate dimensions. Eq. (2.5) is called a state-space realization of the system $\mathcal{F}$. Given the realization (A, B, C, D) of $\mathcal{F}$, the corresponding transfer function G can be obtained as:

$$G(z) = D + C(zI - A)^{-1}B, \tag{2.6}$$

while the corresponding impulse response coefficients satisfy

$$\mathfrak{g}(0) = D,\ \mathfrak{g}(k) = CA^{k-1}B,\ \forall k > 0. \tag{2.7}$$

2.1.2 Stability

In this section, we introduce the notion of stability of DT-FD-LTI systems.

Definition 2.1 (Stability of DT-FD-LTI systems) (Pearson 1999) *A causal DT-FD-LTI system $\mathcal{F}$ with minimal state-space realization (A, B, C, D) and transfer function representation $G(z)$ with an IIR $\{g(k)\}_{k=1}^{\infty}$ is asymptotically stable if and only if one of the following equivalent conditions is satisfied:*

1. *For a minimal state-space realization*[3] *(A, B, C, D), all the eigenvalues of A are strictly inside the unit circle.*
2. *All poles of $G(z)$, i.e., the roots of the denominator (or common denominator in the MIMO) of $G(z)$, are strictly inside the unit circle.*
3. *The impulse response function denoted by g satisfies:*

$$\begin{aligned} \textit{SISO} &: \sum_{k=0}^{\infty} |g(k)| < \infty, \\ \textit{MIMO} &: \max_{i \in \{1,\ldots,n_{\mathbb{Y}}\}} \sum_{j=1}^{n_{\mathbb{U}}} \sum_{k=0}^{\infty} |g_{i,j}(k)| < \infty. \end{aligned} \tag{2.8}$$

Such a stability properly implies that the system is a *Bounded-Input Bounded-Output* (BIBO) stable system, i.e., it produces a uniformly bounded output when a uniformly bounded input is applied.

2.2 The related Hilbert and Hardy spaces

In this section, we discuss some fundamental notions that are related to *functional analysis*, which are useful for the later discussion.

[3] Note that state-space realization of a TF $G(z)$ is generally non unique and can result in (2.5) with various state dimensions $\dim(x)$. A minimal realization is one of these equivalent realizations with the least possible state dimension.

2.2.1 Metric, normed linear and inner product spaces

Definition 2.2 (Metric space) (Young 1988) *A metric space $(\mathscr{X}, d)$ is a set $\mathscr{X}$ together with an assigned metric function $d : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$, that determines the "distance" between any two elements in that space, and has the following properties:*

1. **Positive**: $d(x_1, x_2) \geq 0$ *for all* $x_1, x_2 \in \mathscr{X}$,
2. **Nondegenerate**: $d(x_1, x_2) = 0$ *if and only if* $x_1 = x_2$,
3. **Symmetric**: $d(x_1, x_2) = d(x_2, x_1)$ *for all* $x_1, x_2 \in \mathscr{X}$,
4. **Triangle inequality**: $d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$ *for all* $x_1, x_2, x_3 \in \mathscr{X}$.

An intuitive example of a metric space is $\mathbb{R}$ with the associated metric $|x_1 - x_2|$, for $x_1, x_2 \in \mathbb{R}$. The two interesting special cases of metric spaces: i) normed linear space; and ii) inner product space, where both of them are linear (vector) spaces.

Definition 2.3 (Linear (vector) space) *A linear (vector) space is a collection of objects, the so-called vectors, which can be added together, and multiplied by constants, i.e., scaled. The result of these actions, i.e., addition and scaling, is always an element in that space.*

Definition 2.4 (Normed linear space) *A (complex) normed linear space $(\mathscr{H}_{\mathrm{N}}, \|\cdot\|)$ is a linear (vector) space with a function $\|\cdot\| : \mathscr{H}_{\mathrm{N}} \to \mathbb{R}$ called a norm that satisfies the following properties:*

1. **Positive**: $\|x\| \geq 0$ *for all* $x \in \mathscr{H}_{\mathrm{N}}$,
2. **Nondegenerate**: $\|x\| = 0$ *if and only if* $x = 0$,
3. **Multiplicative**: $\|\varsigma x\| = |\varsigma|\|x\|$ *for all* $x \in \mathscr{H}_{\mathrm{N}}$ *and* $\varsigma \in \mathbb{C}$,
4. **Triangle inequality**: $\|x_1 + x_2\| \leq \|x_1\| + \|x_2\|$ *for all* $x_1, x_2 \in \mathscr{H}_{\mathrm{N}}$.

Moreover, a normed linear space $(\mathscr{H}_{\mathrm{N}}, \|\cdot\|)$ is a metric space, where the metric d is defined by $d(x_1, x_2) = \|x_1 - x_2\|$ for $x_1, x_2 \in \mathscr{H}_{\mathrm{N}}$.

Definition 2.5 (Inner product space) *An inner product space $(\mathscr{H}_{\mathrm{I}}, \langle\cdot,\cdot\rangle)$ is a linear (vector) space with a function $\langle\cdot,\cdot\rangle : \mathscr{H}_{\mathrm{I}} \times \mathscr{H}_{\mathrm{I}} \to \mathbb{C}$ called an inner product that satisfies the following properties:*

1. **Positive**: $\langle x, x\rangle \geq 0$ *for all* $x \in \mathscr{H}_{\mathrm{I}}$,
2. **Nondegenerate**: $\langle x, x\rangle = 0$ *if and only if* $x = 0$,
3. **Multiplicative**: $\langle \varsigma x_1, x_2\rangle = \varsigma\langle x_1, x_2\rangle$ *for all* $x_1, x_2 \in \mathscr{H}_{\mathrm{I}}$ *and* $\varsigma \in \mathbb{C}$,
4. **Symmetric**: $\langle x_1, x_2\rangle = \langle x_2, x_1\rangle^*$ *for all* $x_1, x_2 \in \mathscr{H}_{\mathrm{I}}$, *where* z^* *denotes the complex conjugate of the complex number* z.

5. **Distributive**: $\langle x_1 + x_2, x_3\rangle = \langle x_1, x_3\rangle + \langle x_2, x_3\rangle$ *for all* $x_1, x_2, x_3 \in \mathscr{H}_{\mathrm{I}}$.

Moreover, an inner product space $(\mathscr{H}_{\mathrm{I}}, \langle\cdot,\cdot\rangle)$ *is a normed linear space with the norm defined by* $\|x\| = \sqrt{\langle x, x\rangle}$.

It is worth to emphasize that

$$\text{inner product space } \Rightarrow \text{normed linear space } \Rightarrow \text{metric space.}$$

However, the converse implications are not true. Next, we give the definition of Cauchy sequences and the notion of complete metric space.

Definition 2.6 (Cauchy sequence) *Let* $\{x_k\}_{k=1}^{\infty}$ *be a sequence in a metric space* $(\mathscr{X}, \mathrm{d})$. *The sequence* $\{x_k\}_{k=1}^{\infty}$ *is said to be Cauchy if for every* $\varsigma > 0$ *there exists an integer* $k_\varsigma \in \mathbb{N}$, *where* $\mathbb{N}$ *is the set of natural numbers (positive integers), such that* $\mathrm{d}(x_i, x_j) < \varsigma$ *whenever* $i, j > k_\varsigma$.

Definition 2.7 (Complete metric space) *If every Cauchy sequence in a metric space* $\mathscr{X}$ *converges to an element of* $\mathscr{X}$, *then* $\mathscr{X}$ *is said to be complete.*

Definition 2.8 (Banach space) *A Banach space is a complete normed linear space.*

Definition 2.9 (Hilbert space) *A Hilbert space is a complete inner product space.*

Definition 2.10 (Orthonormal basis of a Hilbert space) *A sequence* $\{\phi_k\}_{k=1}^{\infty}$ *in a Hilbert space* $\mathscr{H}$, *equipped with an inner product denoted by* $\langle\cdot,\cdot\rangle_{\mathscr{H}}$, *is said to be a complete orthonormal basis if the following conditions are satisfied:*

- $\langle \phi_i, \phi_l\rangle_{\mathscr{H}} = \begin{cases} 0, & \text{for } i \neq l \\ 1, & \text{for all } i = l \geq 1. \end{cases}$
- *For any* $f \in \mathscr{H}$, $f(\cdot) = \sum_{k=1}^{\infty} c_k \phi_k(\cdot)$, *where* $c_k = \langle f, \phi_k\rangle_{\mathscr{H}}$ *are the expansion coefficients of* f *under the basis* $\{\phi_k\}_{k=1}^{\infty}$.

In the following, we introduce some related Hilbert spaces that are important for the development of the subsequent results.

2.2.2 Sequence-related Hilbert spaces

Denote by $\ell_2(\mathbb{Z})$, the Hilbert space of squared summable complex scalar sequences $h : \mathbb{Z} \to \mathbb{C}$, i.e., "finite energy sequences" that satisfy $\sum_{k=-\infty}^{\infty} |h(k)|^2 < \infty$, equipped with the inner product between any two elements $f, h \in \ell_2(\mathbb{Z})$ as

$$\langle f, h\rangle_{\ell_2} = \sum_{k=-\infty}^{\infty} f(k) h^*(k).$$

As causal sequences[4] are of special interest, an interesting subspace of $\ell_2(\mathbb{Z})$ is $\ell_2(\mathbb{N})$, which is the space of causal sequences $h : \mathbb{N} \to \mathbb{C}$ of finite energy, i.e., $\sum_{k=1}^{\infty} |h(k)|^2 < \infty$. Another interesting subspace of $\ell_2(\mathbb{N})$ is $\mathcal{R}\ell_2(\mathbb{N})$, which contains only squared summable, real and causal sequences. Moreover, $\mathcal{R}\ell_1(\mathbb{N})$, which is a Banach space, is the subspace of absolutely summable real sequences, i.e., $\sum_{k=1}^{\infty} |h(k)| < \infty$, equipped with the norm:

$$\|h\|_{\ell_1} = \sum_{k=1}^{\infty} |h(k)|.$$

The importance of the space $\mathcal{R}\ell_1(\mathbb{N})$ comes from the fact that the impulse response $\mathfrak{g}$ of all DT-FD-LTI stable and causal systems satisfies the necessary and sufficient condition $\sum_{k=1}^{\infty} |\mathfrak{g}(k)| < \infty$ in the SISO case, see Definition 2.1, hence it belongs to $\mathcal{R}\ell_1(\mathbb{N})$, which implies that it belongs to $\mathcal{R}\ell_2(\mathbb{N})$ since $\mathcal{R}\ell_1(\mathbb{N}) \subset \mathcal{R}\ell_2(\mathbb{N})$. However, the converse is not true, i.e., a square summable sequence does not need to be absolutely summable as $\mathcal{R}\ell_2(\mathbb{N}) \not\subset \mathcal{R}\ell_1(\mathbb{N})$.

Example 2.1 (The connection between $\mathcal{R}\ell_1(\mathbb{N}), \mathcal{R}\ell_2(\mathbb{N})$) *Consider the following harmonic series $h(k) = \frac{1}{k}$, $k \in \mathbb{N}$. Since $\sum_{k=1}^{\infty} |h(k)|^2 = \frac{\pi^2}{6} < \infty$, hence, $h \in \mathcal{R}\ell_2(\mathbb{N})$. However, $\sum_{k=1}^{\infty} |h(k)| = \infty$, which means that $h \notin \mathcal{R}\ell_1(\mathbb{N})$.*

2.2.3 Function-related Hilbert spaces

We denote by $L_2(\mathbb{J})$, where $\mathbb{J}$ is the unit circle, the Hilbert space of square integrable scalar complex functions on $\mathbb{J}$, i.e., $\frac{1}{2\pi}\int_{-\pi}^{\pi} |F(e^{\mathrm{j}\omega})|^2 d\omega < \infty$, equipped with the inner product

$$\langle F_1, F_2 \rangle_{L_2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} F_1(e^{\mathrm{j}\omega}) F_2^*(e^{\mathrm{j}\omega}) d\omega, \tag{2.9}$$

where $F_1, F_2 \in L_2(\mathbb{J})$. A very important subspace of $L_2(\mathbb{J})$ is the Hardy space $\mathscr{H}_2(\mathbb{E})$ defined below, with $\mathbb{E}$ being the exterior of the unit circle.

Definition 2.11 (The Hardy space over $\mathbb{E}$) (Heuberger et al. 2005) *Denote by $\mathscr{H}_2(\mathbb{E})$ the Hardy space of complex functions $F : \mathbb{C} \to \mathbb{C}$, which are analytic[5] on $\mathbb{E}$ and squared integrable on $\mathbb{J}$. $\mathscr{H}_2(\mathbb{E})$ is equipped with an inner product that is defined as*

$$\langle F_1, F_2 \rangle_{\mathscr{H}_2} = \frac{1}{2\pi \mathrm{j}} \oint_{\mathbb{J}} F_1(z) F_2^*(1/z^*) \frac{dz}{z}, \tag{2.10}$$

where $F_1, F_2 \in \mathscr{H}_2(\mathbb{E})$.

[4] A sequence h is causal if $h(k) = 0$ for $k < 0$.

[5] A complex function is said to be analytic on a region R if it is complex differentiable at every point in R. Moreover, if a complex function is analytic on a region R, it is infinitely differentiable in R.

Moreover, $\mathscr{RH}_2(\mathbb{E})$ is a subspace of $\mathscr{H}_2(\mathbb{E})$ which contains all functions that have a real-valued impulse responses. Another interesting subspace is $\mathscr{RH}_{2-}(\mathbb{E})$ which is a subspace of $\mathscr{RH}_2(\mathbb{E})$ which contains all real, proper, rational and finite-order transfer functions which are analytic on $\mathbb{E}$ and square integrable on $\mathbb{J}$, i.e., these transfer functions are stable in the sense that their impulse responses belong to $\mathcal{R}\ell_1(\mathbb{N})$.

2.2.4 Isomorphism between the considered spaces

For Hilbert spaces, an isomorphism is a one to one mapping that preserves inner products and hence norms. It is enough to define a one to one mapping that carries each element of a complete orthonormal basis of the first Hilbert space to a unique element of a complete orthonormal basis of the second Hilbert space.

The spaces $\ell_2(\mathbb{Z}), \ell_2(\mathbb{N})$ are isomorphic to $L_2(\mathbb{J}), \mathscr{H}_2(\mathbb{E})$, respectively, i.e., every $f \in \ell_2(\mathbb{N})$ corresponds to one and only one function $F \in \mathscr{H}_2(\mathbb{E})$ and vice versa. Such an isomorphism is defined through the following z-transform:

$$F(z) = \mathscr{Z}\{f\} = \sum_{k=1}^{\infty} f(k) z^{-k}, \tag{2.11}$$

which holds for all $z \in \mathbb{C}$ in the corresponding ROC. Equivalently, this isomorphism can be established via the following *Discrete-Time Fourier Transform* (DTFT) denoted by $\mathscr{F}$, where the Fourier transform of a sequence $f \in \ell_2(\mathbb{N})$ can be seen as the z-transform of that sequence evaluated on $\mathbb{J} \subset$ ROC, due to the norm definition (2.10), i.e.,

$$F(e^{\mathrm{j}\omega}) = \mathscr{F}\{f\} = \sum_{k=1}^{\infty} f(k) e^{-\mathrm{j}\omega k}, \tag{2.12}$$

where $F \in \mathscr{H}_2(\mathbb{E})$.

Similarly, an isomorphism is presented between the subspaces $\mathscr{RH}_2(\mathbb{E})$ and $\mathcal{R}\ell_2(\mathbb{N})$ and is defined through the same mapping as detailed in (2.11) and (2.12). It is worth to mention that the inverse $z-$transform, denoted by $\mathscr{Z}^{-1}$, of any $F \in \mathscr{H}_2(\mathbb{E})$ is defined as

$$f(k) = \mathscr{Z}^{-1}\{F(z)\} = \frac{1}{2\pi\mathrm{j}} \oint_{\mathbb{J}} F(z) z^{k-1}\, dz, \tag{2.13}$$

with $k \in \mathbb{N}$ and the inverse DTFT, denoted by $\mathscr{F}^{-1}$, of any $F \in \mathscr{H}_2(\mathbb{E})$ is defined as

$$f(k) = \mathscr{F}^{-1}\{F(e^{\mathrm{j}\omega})\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(e^{\mathrm{j}\omega}) e^{\mathrm{j}\omega k} d\omega, \tag{2.14}$$

where $f \in \ell_2(\mathbb{N})$. Figure 2.2 shows the above-mentioned connection between different spaces from different domains. At the right side of the figure, the sequences related spaces are shown, whereas the functions-related spaces are shown at the left side. The mapping between both domains is established by the z-transform, Fourier transform and their inverses.

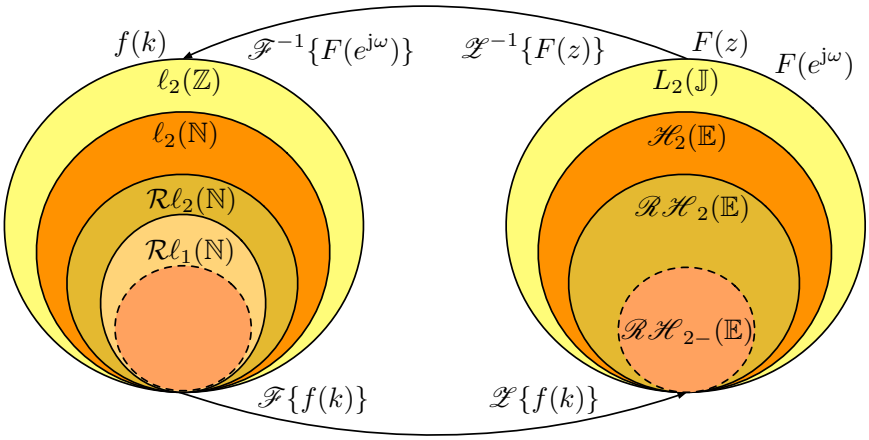


Figure 2.2: The isomorphisms between some related Hilbert and Hardy spaces.

Example 2.2 (Standard or canonical complete orthonormal basis)

- $\mathcal{R}\ell_2(\mathbb{N})$: $\phi_i(k) = \delta_{ik}, i \in \mathbb{N}$, *where* δ_{ik} *is the Kronecker delta function, i.e., it is equal to one if* $i = j$ *and equal to zero otherwise.*
- $\mathscr{RH}_2(\mathbb{E})$: $\phi_i(z) = z^{-i}, \quad i \in \mathbb{N}$.

2.2.5 Why Hilbert spaces are interesting?

Let $\{\phi_k\}_{k=1}^{\infty}$ be an orthonormal basis of a Hilbert space $\mathscr{H}$ with $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ as the inner product defined on it, $\mathcal{U}_n = \mathrm{Span}\,\{\phi_k\}_{k=1}^{n}$, and $\mathcal{U}_n^{\perp}$ is its orthogonal complement[6], i.e., $\mathcal{U}_n^{\perp} = \mathrm{Span}\,\{\phi_k\}_{k=n+1}^{\infty}$. According to the projection theorem (Young 1988), the direct sum of $\mathcal{U}_n$ and $\mathcal{U}_n^{\perp}$ is $\mathscr{H}$ itself. Hence, each element $x \in \mathscr{H}$ can be written as

$$x = \underbrace{\sum_{i=1}^{n} \langle x, \phi_i \rangle_{\mathscr{H}}\, \phi_i}_{\in \mathcal{U}_n} + \underbrace{\sum_{i=n+1}^{\infty} \langle x, \phi_i \rangle_{\mathscr{H}}\, \phi_i}_{\in \mathcal{U}_n^{\perp}}. \tag{2.15}$$

Note that the best approximation of $x \in \mathscr{H}$ on $\mathcal{U}_n$ is the projection of x onto $\mathcal{U}_n$, where the error of the approximation belongs to $\mathcal{U}_n^{\perp}$.

Remark 2.1 *It is noticeable that the approximation error depends on the number of terms* n *in the truncated expansion and the basis functions. With proper selection of the basis, i.e.,* $\{\phi_k\}_{k=1}^{\infty}$*, we can tune how fast the coefficients converge to zero and accordingly further influence the error term which is described by the orthogonal complement.*

[6] $\mathcal{U}_n^{\perp}$ is composed by all elements of $\mathscr{H}$ that are simultaneously orthogonal to all elements of $\mathcal{U}_n$.

As mentioned in Section 2.1.1, a series expansion representation, in terms of basis functions, of FD-LTI systems is a well-known representation. Regarding IIR models, any $G \in \mathscr{RH}_{2-}(\mathbb{E})$ can be written as a linear combination of the orthonormal basis of $\mathscr{RH}_{2-}(\mathbb{E})$, in particular the pulse basis $\phi_i(z) = z^{-i}$, as shown in (2.4). An approximation can be obtained by truncating the IIR representation (2.4) to the $n-$th order expansion, which is known as the *Finite Impulse Response* (FIR) representation:

$$\hat{G}(z) = \sum_{k=1}^{n} \mathrm{g}(k) z^{-k}. \tag{2.16}$$

The quality of this approximation depends on the relative magnitude of the impulse response coefficients that are not included in the finite expansion. For instance, in case of systems that exhibit slow dynamics, i.e., slow decaying impulse response, a high order FIR representation, associated with a large number of coefficients, is required to get a good approximation. However, from the utilization perspective, estimating such a large number of parameters is, in general, unattractive as the variance of the parameter estimates grows with the number of estimated parameters. A possible solution to cope with this problem is to utilize an efficient basis of $\mathscr{RH}_{2-}(\mathbb{E})$ instead of the pulse basis used in the FIR representation, i.e., basis that result in a faster decay of the expansion and hence reduce the required number of parameters to be estimated. Next, we introduce how such basis can be generated and discuss their attractive properties in system approximation.

2.3 Orthonormal basis functions

2.3.1 All-pass functions

A special set of functions in $\mathscr{H}_2(\mathbb{E})$, the so-called all-pass functions or *inner functions*, are of great importance in many areas of system and control theory, e.g., (Vidyasagar 1985), signal processing, e.g., (Regalia et al. 1988), network synthesis, e.g., (Deprettere and Dewilde 1980), etc.

Definition 2.12 (All-pass function) *A function $\mathcal{H} \in \mathscr{H}_2(\mathbb{E})$ is called all-pass, if*

$$\mathcal{H}(z)\mathcal{H}^*(1/z^*) = 1, \ \forall z \in \mathbb{C}. \tag{2.17}$$

Such a function if rational, is completely determined, modulo the sign, by its poles $\{\lambda_i \in \mathbb{D}\}_{i=1}^{n}$ with $\mathbb{D}$ is the unit disc, and it can be written as:

$$\mathcal{H}(z) = \pm \prod_{i=1}^{n} \frac{1 - \lambda_i^* z}{z - \lambda_i}, \tag{2.18}$$

which is known as a *Blaschke product* (Vidyasagar 1985). In the context of the work presented in this thesis, all-pass functions are the main building blocks in the construction of general rational orthonormal basis as will be shown in the next section.

2.3.2 General class of OBFs

Since the goal is to define an efficient basis to be used to represent DT-FD-LTI systems, in the following, we shall introduce rational OBFs which constitutes a complete basis for $\mathscr{H}_2(\mathbb{E})$ and some of its subspaces, e.g., $\mathscr{RH}_2(\mathbb{E})$ and $\mathscr{RH}_{2-}(\mathbb{E})$. Moreover, the definition of such basis will also be discussed in the time-domain with the related sequence spaces.

Takenaka-Malmquist basis

Let $\mathcal{H}_0 \equiv 1$ and $\{\mathcal{H}_i\}_{i=1}^{\infty}$ be a sequence of DT stable inner functions with McMillan degrees[7] $\{n_i\}_{i=1}^{\infty}$ and let (A_i, B_i, C_i, D_i) be minimal balanced state-space representations of $\mathcal{H}_i$ (Skogestad and Postlethwaite 1996). Let $\{\lambda_1, \lambda_2, \dots\} \subset \mathbb{D}$, denote the collection of all poles of the inner functions $\mathcal{H}_1, \mathcal{H}_2, \dots$ satisfying the completeness (Szász) condition[8] $\sum_{k=1}^{\infty}(1-|\lambda_k|) = \infty$. Then, the scalar elements of the sequence of vector functions

$$\mathcal{V}_i(z) = (zI - A_i)^{-1}B_i \prod_{l=0}^{i-1} \mathcal{H}_l(z),\ i > 0, \tag{2.19}$$

constitute a complete orthonormal basis for $\mathscr{H}_2(\mathbb{E})$, where each element $[\mathcal{V}_i]_j$ is orthonormal in $\mathscr{H}_2(\mathbb{E})$ with respect to the entire sequence. These scalar elements can be written as

$$\{\breve{\psi}_k(z)\}_{k=1}^{\infty} = \{[\mathcal{V}_i]_j\}_{i=1,j=1}^{\infty,n_i} = \frac{\sqrt{1-|\lambda_k|^2}}{z-\lambda_k} \prod_{l=1}^{k-1} \frac{1-\lambda_l^* z}{z-\lambda_l},\ k = \left(\sum_{l=0}^{i-1} n_l\right) + j, \tag{2.20}$$

where $[\mathcal{V}_i]_j$ denote the j-th element of $\mathcal{V}_i$, which are known as *Takenaka-Malmquist functions* (Heuberger et al. 2005). Note that, in the general case, in the sense that there are no further restrictions on the poles, such basis have complex-valued impulse responses. In order to guarantee that the associated impulse responses with the considered basis are real-valued, i.e., that the basis belong to $\mathscr{RH}_2(\mathbb{E})$, the complex poles should appear in complex conjugate pairs.

There are some interesting special cases of the class of OBFs generated by (2.20) that can be visualized in Figure 2.3. These special cases will be briefly discussed below (see Heuberger et al. (2005) for a detailed overview on these classes).

Hambo basis

The special case when all $\mathcal{H}_i = \mathcal{H}_{\mathrm{b}},\ \forall i > 0$, where $\mathcal{H}_{\mathrm{b}} \in \mathscr{RH}_{2-}(\mathbb{E})$ is an inner function with McMillan degree $n_{\mathrm{g}} > 0$, are known as the *Hambo basis*, also known

[7] The McMillan degree of a transfer function G(z) is defined as the state dimension of the minimal realization of G(z).

[8] This condition means that the sequence of the generating poles cannot converge too fast to the unit circle.

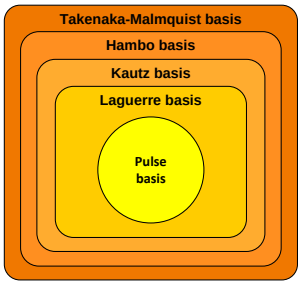


Figure 2.3: Classification of orthonormal basis functions (Tóth 2008).

as *Generalized* OBFs (GOBFs). Let $(A_\mathrm{b}, B_\mathrm{b}, C_\mathrm{b}, D_\mathrm{b})$ be a minimal balanced state-space realization of $\mathcal{H}_\mathrm{b}(z)$. Such a function is completely determined by its poles $\Lambda_{n_\mathrm{g}} = \{\lambda_1, \ldots, \lambda_{n_\mathrm{g}}\} \in \mathbb{D}$ with Λ_{n_g} containing real poles and/or complex conjugate pole pairs. The class of *Hambo basis* is obtained by cascading identical n_g-th order all-pass functions and can be written in a vector form as:

$$\mathcal{V}_i(z) = \mathcal{V}_1(z)\mathcal{H}_\mathrm{b}^{i-1}(z), \quad \text{for } i > 1, \tag{2.21}$$

where $\mathcal{V}_1(z) = (zI - A_\mathrm{b})^{-1}B_\mathrm{b}$ and I is the identity matrix with an appropriate size. Let $[\mathcal{V}_1]_j$ denote the j-th element of $\mathcal{V}_1$. Then, the GOBFs consists of the functions

$$\breve{\Psi}_{n_\mathrm{g}} = \left\{\breve{\psi}_k\right\}_{k=1}^{\infty} = \left\{[\mathcal{V}_1]_j \mathcal{H}_\mathrm{b}^i\right\}_{j=1,i=0}^{n_\mathrm{g},\infty}, \quad \text{with } k = i \cdot n_\mathrm{g} + j. \tag{2.22}$$

These functions, i.e., (2.22), constitute a complete orthonormal basis for $\mathscr{RH}_2(\mathbb{E})$.

Kautz basis

When $\mathcal{H}_i = \mathcal{H}_\mathrm{b}$, $\forall i > 0$ with $n_\mathrm{g} = 2$, the resulting OBFs are called *2-parameter Kautz functions*. Such functions can be considered to be adequate (in terms of the truncation concept of Section 2.2.5) for the expansion of transfer functions with dominant second order modes:

$$\begin{aligned}
\breve{\psi}_{2k-1}(z) &= \frac{\sqrt{1-\mathfrak{c}^2}(z-\mathfrak{b})}{z^2+\mathfrak{b}(\mathfrak{c}-1)z-\mathfrak{c}}\left(\frac{-\mathfrak{c}z^2+\mathfrak{b}(\mathfrak{c}-1)z+1}{z^2+\mathfrak{b}(\mathfrak{c}-1)z-\mathfrak{c}}\right)^{k-1} \\
\breve{\psi}_{2k}(z) &= \frac{\sqrt{(1-\mathfrak{c}^2)(1-\mathfrak{b}^2)}}{z^2+\mathfrak{b}(c-1)z-\mathfrak{c}}\left(\frac{-\mathfrak{c}z^2+\mathfrak{b}(\mathfrak{c}-1)z+1}{z^2+\mathfrak{b}(\mathfrak{c}-1)z-\mathfrak{c}}\right)^{k-1},
\end{aligned} \tag{2.23}$$

where $\mathfrak{b}, \mathfrak{c} \in (-1, 1)$. Note that (2.23) corresponds to a repeated complex pair $\lambda, \lambda^* \in \mathbb{D}$.

Laguerre basis

In case $\mathcal{H}_i = \mathcal{H}_\mathrm{b}$, $\forall i > 0$ with $n_\mathrm{g} = 1$ the basis are called *Laguerre functions*. As this type of functions in $\mathscr{RH}_2(\mathbb{E})$ have only a real repeated pole λ, therefore it can

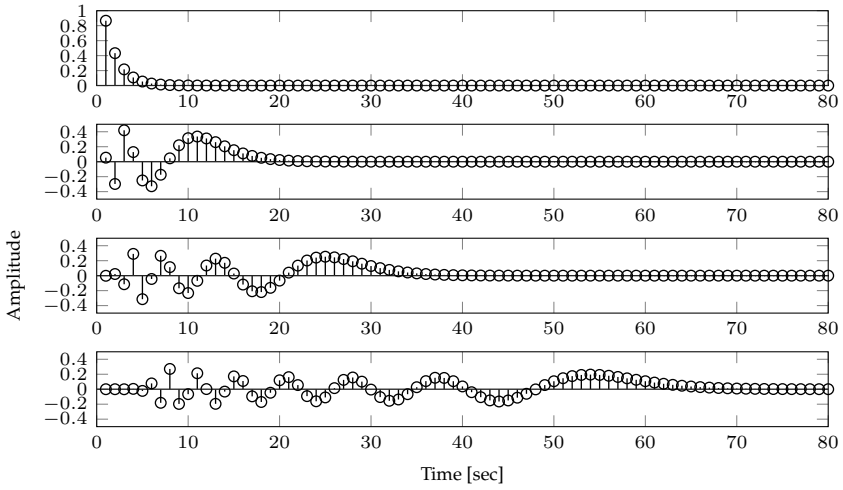


Figure 2.4: Laguerre basis functions, $\breve{\psi}_1, \breve{\psi}_5, \breve{\psi}_{10}, \breve{\psi}_{20}$, $\lambda = 0.5$.

provide an adequate basis for a $F \in \mathscr{RH}_2(\mathbb{E})$ with a dominant first-order mode.

$$\breve{\psi}_k(z) = \frac{\sqrt{1-\lambda^2}}{z-\lambda}\left(\frac{1-\lambda z}{z-\lambda}\right)^{k-1}, \quad \lambda \in (-1,1), \tag{2.24}$$

where the parameter λ is known as the Laguerre parameter or generating pole. The impulse response of Laguerre basis functions exhibits an exponential decay as shown in Figure 2.4 for $\lambda = 0.5$.

Pulse basis

In the case when $\mathcal{H}_i = z^{-1}, \forall i > 0$, the corresponding basis are called *pulse functions*. Such functions are utilized in the "well-known" impulse response representation of LTI systems, see Section 2.1.1.

2.3.3 OBFs model structure

Since $\breve{\Psi}_{n_g}$ in (2.22) constitutes a complete basis of $\mathscr{RH}_2(\mathbb{E})$, any $G \in \mathscr{RH}_2(\mathbb{E})$ can be decomposed as[9]

$$\begin{aligned} G(z) &= \sum_{i=0}^{\infty}\sum_{j=1}^{n_g} \breve{c}_{ij}[\mathscr{V}_1]_j \mathcal{H}_{\mathrm{b}}^i(z) \\ &= \sum_{k=1}^{\infty} c_k \breve{\psi}_k(z), \end{aligned} \tag{2.25}$$

[9]There are various forms of series-expansion defined in the MIMO case. Here we take the simplest form of expansion in terms of scalar basis, see Heuberger et al. (2005) for more details.

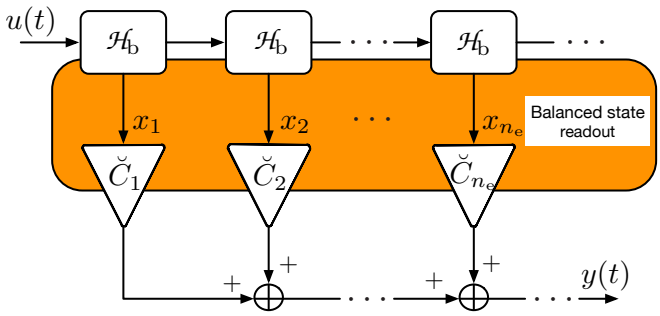


Figure 2.5: IO relation when the OBFs parameterization (2.25) is used with $\breve{C}_i = [\breve{c}_{i1} \ \cdots \ \breve{c}_{in_\mathrm{g}}]$.

which is the generalization of expansion in terms of the well-known *pulse basis functions*, i.e., $\{z^{-k}\}_{k=1}^{\infty}$, used in the impulse response model structure (2.4). It can be shown that the rate of convergence of this series expansion is bounded by $\rho_\mathrm{b} = \max_k |\mathcal{H}_\mathrm{b}(\breve{\lambda}_k^{-1})|$, the so-called decay rate, where $\{\breve{\lambda}_k\}$ are the poles of $G(z)$ (Oliveira e Silva 1996). In the best case, i.e., when the poles of G are the same (with multiplicity) as the poles of $\mathcal{H}_\mathrm{b}$, only the terms with $i = 0$ in (2.25) are non-zero. When such a model representation, i.e., (2.25), is used to describe the dynamics of an LTI system, the IO relation is illustrated in Figure 2.5, where the signals $\{x_i\}$ are the state variables of the balanced state-space realization of $\mathcal{H}_\mathrm{b}$. More specifically, $x_i = (zI - A_\mathrm{b})^{-1}B_\mathrm{b}\mathcal{H}_\mathrm{b}^{i-1}(z)u$. In practice, only a finite number of extensions of $\mathcal{H}_\mathrm{b}$, i.e., n_e, is used

$$\breve{\Psi}_{n_\mathrm{g}}^{n_\mathrm{e}} = \left\{\breve{\psi}_k\right\}_{k=1}^{n_\mathrm{g}n_\mathrm{e}} = \left\{[\mathscr{V}_1]_j\mathcal{H}_\mathrm{b}^i\right\}_{j=1,i=0}^{n_\mathrm{g},n_\mathrm{e}}, \quad \text{with } k = i \cdot n_\mathrm{g} + j, \tag{2.26}$$

like in FIR models, where $\left\{z^{-k}\right\}_{k=1}^{n}$ are used as basis functions. It can be shown that there exists a $\varsigma > 0$ such that all expansion coefficients $\breve{c}_{ij} \in \mathbb{R}$ satisfy[10]:

$$|\breve{c}_{ij}| \leq \varsigma\rho_\mathrm{b}^{n_\mathrm{g}(i+1)+j}. \tag{2.27}$$

In contrast with FIR structures, the OBFs parameterization uses a broad class of basis functions with *Infinite Impulse Representation*. Therefore, OBFs parameterization can achieve an arbitrary low modeling error with a relatively small number of parameters due to the faster convergence of the series representation than in the FIR case, which in system identification results in decreased variance of the final model estimate (Heuberger et al. 1995; Tóth et al. 2009a).

Since we are interested also in impulse response estimation based on time-domain data, it is more convenient to define the corresponding OBFs in the time-domain. Denote by

$$\Psi = \{\psi_k\}_{k=1}^{\infty}, \tag{2.28}$$

[10]Note that, such an upper bound is only true in the SISO case. However, in the MIMO case, it is still valid for each output channel.

$\psi_k(t) = \mathscr{Z}^{-1}\{\breve{\psi}_k(z)\}$, i.e., the correspondent of $\{\breve{\psi}_k\}_{k=1}^{\infty}$ in the time-domain. It is an important result that $\{\psi_k\}_{k=1}^{\infty}$ is a complete basis of $\mathcal{R}\ell_2(\mathbb{N})$ (Oliveira e Silva 1995), hence any impulse response $\mathrm{g} \in \mathcal{R}\ell_2(\mathbb{N})$ associated with a $G(z) \in \mathscr{RH}_2(\mathbb{E})$ can be written as

$$\mathrm{g}(t) = \sum_{k=1}^{\infty} c_k \psi_k(t). \tag{2.29}$$

Note also that the expansion coefficients c_k are the same as in (2.25) and decay to zero according to (2.27).

Next, we derive a bound for the Takenaka-Malmquist basis (Heuberger et al. 2005), which will be useful later[11].

Proposition 2.1 (Magnitude bound of OBFs) *Consider the Takenaka Malmquist basis which is defined as in* (2.20) *with pole locations* $\{\lambda_i\}_{i=1}^{\infty} \subset \mathbb{D}$*, which are assumed to appear as real or complex conjugate pairs, being the generating poles locations of* $\{\breve{\psi}_k\}_{k=1}^{\infty}$ *and* $\{\psi_k\}_{k=1}^{\infty}$ *are their associated impulse responses. It holds that*

$$\|\psi_k\|_{\ell_1} \leq 2k\kappa, \tag{2.30}$$

where κ *is a constant that depends on the generating poles.*

Proof: See Appendix A.1. □

2.4 Modeling and identification of LTI systems

System identification is about building mathematical models for dynamic systems based on experimentally measured IO data record. The identification cycle summarized in Table 1.1 gives an overview of the underlying procedure. Two crucial steps involved in that cycle are the choices of an appropriate model set and the identification criterion. The former describes the set in which the suitable description of the system is sought, while the latter defines the aimed performance of the model. The importance of the model set comes from the fact that it directly influences the maximum achievable accuracy or quality of the identified model in terms of the user-defined criterion. The model set should be as large as possible in order to contain as many candidate models as possible, which reduces the structural/bias error of the optimal model in the set. At the same time, the number of parameters of the model should be kept as small as possible, because the variability of the identified models increases with increasing number of parameters. Such conflicting objectives correspond to the well-known bias/variance trade-off.

In this section, a brief introduction of DT *prediction error identification* is given, based on Ljung (1999) and Heuberger et al. (2005). Note that the remaining part of this section is largely based on Tóth (2010).

[11] Note that the derived bound holds true for all of the subclasses of the Takenaka-Malmquist basis, see Figure 2.3.

2.4.1 Identification setting

In the following, the black-box setting of Ljung (1999) is adopted as a framework. In this setting, we are aiming at identifying an unknown system without the use of a prior structural information, but under the assumption that the underlying, so called *data-generating system*, is an LTI discrete-time SISO system:

$$y = G_0(q)u + v, \tag{2.31}$$

where $G_0 \in \mathscr{RH}_{2-}(\mathbb{E})$, u is a *quasi-stationary* signal, and v is a stationary stochastic process (see Ljung (1999) for a definition of these properties). Furthermore v satisfies

$$v = H_0(q)e, \tag{2.32}$$

with H_0 is a monic[12], rational transfer function[13] such that $H_0, H_0^{-1} \in \mathscr{RH}_{2-}(\mathbb{E})$[14] and e is a zero-mean white noise process with variance σ_e^2. Figure 2.6 shows the block diagram of the data generating system under such a setting. Assume furthermore that a data sequence $\mathcal{D}_N = \{u(t), y(t)\}_{t=1}^N$, generated by (2.31), is available. Under the given assumptions, the so called *one-step ahead prediction* of $y(t)$ based on $\{y(t-1), y(t-2), \ldots\}$ and $\{u(t), u(t-1), \ldots\}$ is

$$\hat{y} := (1 - H_0^{-1}(q))y + H_0^{-1}(q)G_0(q)u, \tag{2.33}$$

where due to the monic nature of H_0, only information on $\{y(t-1), y(t-2), \ldots\}$ and $\{u(t), u(t-1), \ldots\}$ are needed to compute $\hat{y}(t)$. In prediction error identification, a parameterized model $(G(q,\theta), H(q,\theta))$ is hypothesized, where $\theta \subset \Theta$ represents the parameter vector that contains the real-valued coefficients of the model, and $\Theta \in \mathbb{R}^{n_\theta}$ is the allowed parameter space. This model structure leads to the one-step ahead predictor:

$$\hat{y}_\theta := (1 - H^{-1}(q,\theta))y + H^{-1}(q,\theta)G(q,\theta)u. \tag{2.34}$$

Then, in the prediction error setting, we would like to choose θ such that the resulting $\hat{y}_\theta$ is a good approximation of y, i.e. the so called *prediction error*

$$\epsilon(t,\theta) := y(t) - \hat{y}_\theta(t), \tag{2.35}$$

is minimized. This is commonly performed by the minimization of the scalar-valued prediction error or the so-called "least squares" LS identification criterion

$$\mathcal{W}_N(\theta, \mathcal{D}_N) = \frac{1}{N}\sum_{t=1}^{N} \epsilon^2(t,\theta), \tag{2.36}$$

[12]Monicity implies that $\lim_{z\to\infty} H_0(z) = 1$.

[13]In the LTI case and due to the linearity of G_0, it is possible to lump many different sources of disturbances into v, e.g. process noise, uncontrollable inputs, etc., which suggests to assume the power spectrum of the noise process to be a rational function, i.e., v is a filtered zero-mean white noise process.

[14]Note that due to the definition of H_0 to be monic it does not belong to $\mathscr{RH}_{2-}(\mathbb{E})$ as a direct feedthrough is required to augment $\mathscr{RH}_{2-}(\mathbb{E})$ resulting in $\mathscr{R}\hat{\mathscr{H}}_{2-}(\mathbb{E}) = \{\varsigma + H\}$, where $\varsigma \in \mathbb{R}$, $H \in \mathscr{RH}_{2-}(\mathbb{E})$. However, since it is not relevant to distinguish these two sets most of the time, we will use the same notation, i.e., $\mathscr{RH}_{2-}(\mathbb{E})$, for both of them.

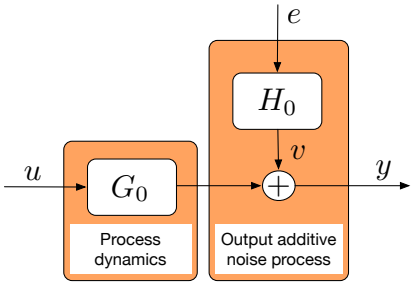


Figure 2.6: Data generating system in the LTI prediction error framework.

Table 2.1: Black-box model structures

	ARX	ARMAX	OE	FIR	BJ
$G(q,\theta)$	$\frac{R_{\mathrm{B}}(q^{-1},\theta)}{R_{\mathrm{A}}(q^{-1},\theta)}$	$\frac{R_{\mathrm{B}}(q^{-1},\theta)}{R_{\mathrm{A}}(q^{-1},\theta)}$	$\frac{R_{\mathrm{B}}(q^{-1},\theta)}{R_{\mathrm{F}}(q^{-1},\theta)}$	$R_{\mathrm{B}}(q^{-1},\theta)$	$\frac{R_{\mathrm{B}}(q^{-1},\theta)}{R_{\mathrm{F}}(q^{-1},\theta)}$
$H(q,\theta)$	$\frac{1}{R_{\mathrm{A}}(q^{-1},\theta)}$	$\frac{R_{\mathrm{C}}(q^{-1},\theta)}{R_{\mathrm{A}}(q^{-1},\theta)}$	1	1	$\frac{R_{\mathrm{C}}(q^{-1},\theta)}{R_{\mathrm{D}}(q^{-1},\theta)}$

resulting in

$$\hat{\theta}_N = \arg\min_{\theta\in\Theta} \mathcal{W}_N\left(\theta, \mathcal{D}_N\right), \tag{2.37}$$

based on the available data record $\mathcal{D}_N$. For other options regarding the identification criterion, see Ljung (1999). Optimization of the identification criterion according to (2.37) is generally a non-convex optimization problem for which iterative (e.g., gradient-based) algorithms have to be applied. This also implies that convergence to a global optimum can not be easily guaranteed. However, in specific cases of parametrization, the optimization reduces to a convex problem with an analytical solution, as will be shown later.

2.4.2 Model structures

One advantage offered by the *prediction error* framework is the various black-box model structures available for the parametrization of $G(q,\theta)$ and $H(q,\theta)$ (see Table 2.1 for these model structures, where the two transfer operators $G(q,\theta)$ and $H(q,\theta)$ are parameterized in terms of ratio of polynomials $R_{\mathrm{A}},\ldots,R_{\mathrm{F}}$ in the backward time-shift operator q^{-1}). The parameter vector θ of these model structures contains the collection of the coefficients of the polynomials. Commonly, the denominator polynomials are assumed to be monic to ensure uniqueness of the parametrization. Every model structure or parametrization induces a set of predictor models, commonly called the *model set*:

$$\left\{(G,H) \in \mathscr{RH}_{2-}\left(\mathbb{E}\right) \times \mathscr{RH}_{2-}\left(\mathbb{E}\right) \mid \theta \in \Theta \subset \mathbb{R}^{n_\theta}\right\}. \tag{2.38}$$

This concept allows us to distinguish the following situations:

- The data generating system $(G_0(q), H_0(q))$ is in the model set, i.e., an exact representation of the data generating system can be found by the chosen model structure.
- $(G_0(q), H_0(q))$ is not in the model set, and hence no exact representation of the system exists by the model structure.

When the main focus is given to the process dynamics of the system, i.e., the deterministic part of the data generating system G_0, it is more convenient to deal with the set of IO models:

$$\{G \in \mathscr{RH}_{2-}(\mathbb{E}) \mid \theta \in \Theta \subset \mathbb{R}^{n_\theta}\}. \tag{2.39}$$

This leads to situations where the process dynamics G_0 can be or can not be captured within the chosen model set. Different model structures offer different properties, namely *linear-in-the-parameter* and *independent parameterization* of the process and noise dynamics:

- For ARX and FIR model structures, the expression of the output predictor (2.34) is linear in the unknown parameters θ, i.e. both the terms $(1 - H^{-1}(q,\theta))$ and $H^{-1}(q,\theta)G(q,\theta)$ are polynomials, which has the major benefit that the LS criterion can be minimized by solving a set of linear equations.
- For FIR, OE, and BJ model structures, G and H are independently parameterized, hence they can be estimated independently in the sense that, we might be able to consistently estimate G even if H is misspecified.

It is particularly attractive to consider a FIR model structure as it satisfies both properties, i.e., linear-in-the-parameter and independent parameterization of the process and noise dynamics.

2.4.3 Identification with OBFs

As explained in the previous section, the FIR model structure enjoys two attractive properties: linear-in-the-parameter property and independent parameterization of the process and noise models. However, such a model structure, i.e., FIR, has a major drawback. More specifically, its capability to efficiently capture the dynamics of physical systems is limited as it generally requires a large number of parameters especially for slow systems, i.e., where the impulse response becomes "long", which leads to increased variance of the estimated model.

In order to retain the above-mentioned attractive properties and at the same time increase the representation capability of such models, OBFs can be used instead of the pulse basis (2.4), which can effectively reduce the required number of

parameters and accordingly reduce the variance of the estimated model. Next, we consider the following model structure:

$$G(q,\theta)=\sum_{i=1}^{n} c_i\breve{\psi}_i(q), \quad H(q,\theta)=1, \tag{2.40}$$

where $\{\breve{\psi}_i\}_{i=1}^{n}$ are orthonormal basis functions in $\mathscr{RH}_{2-}(\mathbb{E})$ with pole locations Λ_{n}. The unknown series-expansion coefficients of (2.40) are collected into the parameter vector $\theta=[c_1\ \cdots\ c_n]^\top\subset\mathbb{R}^n$.

2.4.4 Linear regression

The linear-in-the-parameter property holds for ARX, FIR and OBFs model structures. Hence, the LS problem (2.37) becomes a convex optimization problem with the analytic solution:

$$\hat{\theta}_N=\left[\frac{1}{N}\Upsilon_N^\top\Upsilon_N\right]^{-1}\left[\frac{1}{N}\Upsilon_N^\top Y_N\right]. \tag{2.41}$$

where $Y_N=[y(1)\ \cdots\ y(N)]^\top$ is the collection of the measured output samples and $\Upsilon_N=[\gamma_{\mathrm{r}}(1)\ \cdots\ \gamma_{\mathrm{r}}(N)]^\top$ contains the regressor vector γ_{r} that describes the data relation according to the one-step-ahead predictor: $\hat{y}_\theta(k)=\gamma_{\mathrm{r}}^\top(k)\theta$. For the ARX case with $\deg(R_{\mathrm{A}})=n_{\mathrm{a}}$ and $\deg(R_{\mathrm{B}})=n_{\mathrm{b}}$, the regressor vector is

$$\gamma_{\mathrm{r}}^\top(k)=[y(k-1)\ \cdots\ y(k-n_{\mathrm{a}})\ u(k-1)\ \cdots\ u(k-n_{\mathrm{b}})],$$

while in the FIR case with $\deg(R_{\mathrm{B}})=n_{\mathrm{b}}$, the regressor vector becomes

$$\gamma_{\mathrm{r}}^\top(k)=[u(k-1)\ \cdots\ u(k-n_{\mathrm{b}})],$$

and finally for the OBFs case, the regression vector becomes

$$\gamma_{\mathrm{r}}^\top(k)=\left[\ (\breve{\psi}_1(q)u)(k)\ \cdots\ (\breve{\psi}_n(q)u)(k)\right],$$

containing filtered versions of the input signal rather than delayed versions of u or y.

From the point of view of numerical implementation, the matrix inversion required for the solution (2.41) is not computed directly, but via a QR-algorithm.

Remark 2.2 *For other model structures that do not have the linear-in-the-parameter property, e.g., OE and ARMAX model structures, the LS problem* (2.37) *does not lead to a convex optimization problem with analytic solution and a NL optimization, prone to local minima, is needed to obtain the model estimate.*

2.4.5 Validation in the prediction error setting

The final step in the identification cycle is the (in)validation of the estimated model as shown in Table 1.1. It is mainly about deciding if we accept the estimated model or not for the intended application. In the considered setting, i.e., prediction error, commonly a measured data record, known as the validation data set, is used with the estimated model to compute a simulated / predicted response and then compare it with the measurements. Various approaches are available to accomplish such a step including correlation analysis of the residual or employing an error measure that quantifies the difference between the measured y and the simulated/predicted output $\hat{y}$. Some popular measures are the following:

Definition 2.13 (Mean squared error) (Ljung 1999) *The Mean Squared Error (MSE) is the expected value of the squared estimation error :*

$$\mathrm{MSE} := \mathcal{E}\{(y - \hat{y})^2\}, \tag{2.42}$$

where $\mathcal{E}$ is the expectation operator. The MSE is often computed in a sampled form:

$$\widehat{\mathrm{MSE}} := \frac{1}{N}\sum_{k=1}^{N}\big(y(k) - \hat{y}(k)\big)^2. \tag{2.43}$$

It is worth to mention that the MSE is equal to the LS criterion (2.36) evaluated for the predicted $\hat{y}$. Hence, a high value indicates invalidity of the model.

Definition 2.14 (Best fit percentage) (Ljung 2006) *The Best Fit Rate (BFR) percentage is defined as*

$$\mathrm{BFR} := 100\% \cdot \max\left(1 - \frac{\|y - \hat{y}\|_2}{\|y - \bar{y}\|_2}, 0\right), \tag{2.44}$$

where $\bar{y}$ is the mean of y.

The BFR percentage is a relative measure, often used in the identification toolbox of Matlab to indicate the validity of the identified model.

Definition 2.15 (Variance accounted for) *The Variance Accounted For (VAF) percentage is the percentage of the output variation that is explained by the model:*

$$\mathrm{VAF} := 100\% \cdot \max\left(1 - \frac{\mathrm{var}\,\{y - \hat{y}\}}{\mathrm{var}\,\{y\}}, 0\right). \tag{2.45}$$

The VAF measure describes how much of the output variation is explained by the model, disregarding possible bias of the estimates.

2.5 Summary

In this chapter, we have introduced some representation forms of LTI systems and the related notion of stability. This notion is essential to the derivations in Chapter 4. Furthermore, we have discussed some interesting Hilbert spaces that are related to LTI systems, in terms of both impulse response and transfer function representations. The importance of the Hilbert space, in general, is also highlighted. In the second part of this chapter, we have discussed the OBFs, how to generate them and how to use them to efficiently represent dynamic models. This provides the necessary background and notations for our investigations in Chapter 4 for the use of OBFs to construct a rich class of kernels for LTI system identification. Finally, the classical parametric identification approach in the PEM setting for LTI systems has been also briefly introduced. Such a discussion is important for two reasons: the first is that the PEM framework will be mainly used in the subsequent chapters for estimating linear models; the second is to characterize the associated challenges with the existing parametric approaches.

3 CHAPTER

Kernel Methods in Machine Learning

This chapter is devoted to an overview of kernel-based methods in machine learning. These methods are the backbone of the approaches developed in the subsequent chapters. We start by defining the general regression problem in Section 3.1, explaining the classical approaches for regression and the associated model order selection issue and how it can be solved with the kernel based methods. In Section 3.2, we introduce the concept of regularization in RKHSs to give a unified interpretation of various kernel-based methods. This is followed by an overview on GPR in Section 3.3, where the considered regularization approach is investigated from a Bayesian perspective. Section 3.4 provides a brief discussion on the numerical implementation and computational complexity associated with the GPR approach. In Section 3.5, the connection between RKHSs and GPR in case of white measurement noise with Gaussian distribution is given.

3.1 Regression problem

Regression is the problem of estimating (learning) an unknown function g in a functional relationship $y = g(x)+v$ from a set of observations, i.e., IO measurements of x and y. Here v represents a noise process, which is often considered to be white noise in many applications of regression based estimation. Such a problem arises frequently in many fields, e.g., reinforcement learning, control theory, statistics, etc. It also serves as the basic estimation concept in black-box estimation.

3.1.1 Generating Model

In the standard regression problem, we assume that a set of observations, i.e., $\mathcal{D}_N = \{Y_N, X_N\}$, $Y_N = [y_1 \ \cdots \ y_N]^\top$, $X_N = [x_1 \ \cdots \ x_N]^\top$, is available, generated

by the original relation/data-generating system as follows[1]

$$y_i = g(x_i) + \epsilon_i, \tag{3.1}$$

where $x_i \in \mathbb{R}$ is the input sequence, $y_i \in \mathbb{R}$ is the output, $g : \mathbb{R} \to \mathbb{R}$ is an unknown (non)linear function and $\epsilon_i \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ is an independent Gaussian additive noise with σ_ϵ^2 being the noise variance. Our goal is to reconstruct the function g that describes the observed data and to provide a good prediction of the function value at a new input location x, i.e., for an arbitrary new pair (x, y), the predicted value $g(x)$ should be close to y in the MSE sense. In the next section, the classical parametric approach for this estimation problem is given and problems associated with such an approach and regularization techniques that cope with them are explained.

3.1.2 Parametric approach

The classical approach to reconstruct the function g from the available noisy measurements is to use a parametric model $g_\theta : \mathbb{R} \to \mathbb{R}$ that depends on a vector of parameters $\theta \in \mathbb{R}^{n_\theta}$, e.g., a finite dimensional polynomial model $g_\theta(x) = \theta_1 + \theta_2 x + \theta_3 x^2$. Next, the well-known classical LS method, that dates back to Gauss, can be used to obtain an estimate of θ by minimizing the following quadratic cost (loss) functional:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - g_\theta(x_i))^2. \tag{3.2}$$

An analytic solution of (3.2) is available and a global minimum w.r.t. θ is guaranteed when a linear-in-the-parameter model is postulated, i.e.,

$$g_\theta(x) = \sum_{i=1}^{n_\theta} \theta_i \phi_i(x), \tag{3.3}$$

where $\{\phi_i\}_{i=1}^{n_\theta}$ are predefined basis functions. However, with such parameterization, a fixed structure is imposed upon the function g. Moreover, the number of parameters becomes fixed and is determined in advance regardless of the number of data points N. This immediately introduces the challenge of choosing the appropriate model complexity determined, e.g., by n_θ. Typically, such a selection is performed by model validation techniques as CV, AIC, BIC, etc., see Ljung (1999), Söderström and Stoica (1989) for more details. The complexity of the postulated model largely affects the final model estimates. Its choice is related to the well-known *bias/variance* trade-off: i) Low n_θ leads to under-modeling and accordingly biased estimate; ii) Increasing n_θ will lead to over-parameterized models, which leads to estimates with high variance possibly due to data interpolation (the model fits the noise). As a result, the obtained model will perform poorly when used to predict at a new unseen input. This issue can be also explained in

[1] For the sake of simplicity, in the sequel, we will restrict the scope to the LTI case.

the light of ill-posedness in the sense of (Hadamard 1923): the solution of (3.2) can become highly sensitive to small perturbations of the data y_i.

The basic idea now is to use flexible models, i.e., high order models, possibly also infinite-dimensional, and in the same time to have a "well-posed" solution, in the sense of (Hadamard 1923). A problem is considered to be well-posed if the solution: i) exists; ii) unique; and ii) depends continuously on data and parameters. Such a solution can be found via *nonparametric*[2] regression, which is mainly about determining the required complexity of the model from data and using *high-level* assumptions, e.g., smoothness, which are more relaxed than imposing a specific structure on the model. The modern nonparametric approaches to accomplish such a task mainly use *regularization* techniques introduced extensively in the *inverse* problem literature (Tikhonov and Arsenin 1977; Bertero 1989) in conjunction with RKHSs (Aronszajn 1950; Schölkopf and Smola 2002). In the remaining of this chapter, two regularization techniques, namely regularization in RKHSs and Gaussian regression, are overviewed as they are essential for the developed theory in the subsequent chapters.

3.2 Regularization in RKHSs

RKHSs provide an attractive framework to treat in a unified way many regularization methods, namely, *kernel-based* methods including *smoothing splines* (Wahba 1990), *Regularization Networks* (RN) (Poggio and Girosi 1990), SVM (Suykens et al. 2002; Vapnik 1998) and GPR (Rasmussen and Williams 2006). It has been successfully applied in *statistics* (Wahba 1990), *approximation theory* (Poggio and Girosi 1990), *computer vision* (Bertero et al. 1988) and introduced to the machine learning community in Girosi (1998).

3.2.1 The concept of the regularization network

Regularization in RKHSs is one of the most popular approaches for nonparametric regression, where the unknown function can be obtained by minimizing a regularized functional over a Hilbert space $\mathscr{H}$. For instance, a RN (Poggio and Girosi 1990) is

$$\hat{g} = \underset{g \in \mathscr{H}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{N} (y_i - g(x_i))^2}_{\text{"data-fit"}} + \gamma \underbrace{\|g\|_{\mathscr{H}}^2}_{\text{"regularizer"}} . \tag{3.4}$$

One can see that the objective function consists of two contradicting terms: i) The quadratic loss term that measures the adherence to the observed data; ii) The regularizer in terms of the squared norm of the Hilbert space $\mathscr{H}$ (hypothesis space), that guarantees the well-posedness of (3.4) by penalizing undesired behavior and restricts/governs complexity of g. Finally, γ is the regularization parameter that

[2]Nonparametric does not mean that there are no parameters in the model, but it implies that the number of parameters is flexible and grows with the number of data points.

controls the trade-off between the two contradictory terms, i.e., for small values of γ the "data-fit" term becomes dominant and the probability to overfit the data becomes high, whereas for large values of γ the "regularizer" term becomes dominant and a simple model is expected to be obtained with a potentially large bias error.

Denote by $\mathscr{X}$ the function domain: this is a non-empty set known as the *input space* in the machine learning community. A basic requirement for the Hilbert space $\mathscr{H}$ is that every function in $\mathscr{H}$ has to be well-defined pointwise for any $x \in \mathscr{X}$. Moreover, we assume that pointwise evaluations are continuous, linear and bounded over $\mathscr{H}$, i.e.,

$$\forall x \in \mathscr{X},\ \exists\, \varsigma_x < \infty,\ \text{such that } |g(x)| \leq \varsigma_x \|g\|_{\mathscr{H}}. \tag{3.5}$$

Before jumping into the details of regularization in RKHSs, we give a brief overview of some required basics related to kernel functions and RKHSs.

3.2.2 Kernel functions and RKHSs

Let us first recall the definition of a Mercer kernel. Note that, in the sequel, we use this definition both w.r.t. real- and complex-valued spaces. Therefore, in this section, we will introduce the general case of complex-valued spaces and the concept of real-valued ones follows directly.

Definition 3.1 (Mercer kernel) (Schölkopf and Smola 2002) *Let $\mathscr{X}$ be a metric space. A complex-valued function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$ is called a Mercer kernel if it is continuous, symmetric*[3] *and satisfies $\sum_{i,j=1}^{m} a_i a_j^* K(x_i, x_j) \geq 0$ for any finite set of points $\{x_1, \ldots, x_m\} \subset \mathscr{X}$ and $\{a_1, \ldots, a_m\} \subset \mathbb{C}$. If K satisfies all the stated conditions, but not continuity, it is called* positive definite.

Definition 3.2 (Reproducing kernel) *Let $\mathscr{H}$ be a Hilbert space of complex-valued functions on $\mathscr{X}$ with inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$. A complex-valued function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$ is a reproducing kernel for $\mathscr{H}$ if*

1. *$\forall x \in \mathscr{X}, K_x = K(x, \cdot) \in \mathscr{H}$, where K_x is the so-called kernel section centered at x;*
2. *The reproducing property holds, such that*

$$f(x) = \langle f(\cdot), K(x, \cdot) \rangle_{\mathscr{H}},\ \forall x \in \mathscr{X}, \forall f \in \mathscr{H}.$$

A Hilbert space of complex-valued functions which possesses a reproducing kernel is called an RKHS (Wahba 1990). Moreover, due to the Moore-Aronszajn theorem (Aronszajn 1950), there is a one-to-one correspondence between an RKHS $\mathscr{H}$ and its reproducing kernel K, i.e., to every positive definite kernel K, there is a unique RKHS $\mathscr{H}$ with K as its reproducing kernel and vice versa.

[3] Symmetric here means that $K(x_1, x_2) = K(x_2, x_1)$.

Definition 3.3 (RKHS) *Let K be a positive definite kernel function and $\mathscr{H}$ is the associated RKHS. Then, $\mathscr{H}$ is defined to be the completion of* $\mathrm{Span}\{K_x := K(x,\cdot) : x \in \mathscr{X}\}$, *i.e., the functions in $\mathscr{H}$ can be written as*

$$\mathscr{H} = \left\{ f : \mathscr{X} \to \mathbb{C} \mid f(\cdot) = \sum_{i=1}^{\infty} a_i K_{x_i}(\cdot), x_i \in \mathscr{X},\ a_i \in \mathbb{C},\ \|f\|_{\mathscr{H}} < +\infty \right\}, \quad (3.6)$$

where $\|f\|_{\mathscr{H}} = \sqrt{\langle f, f\rangle_{\mathscr{H}}}$ is the norm in $\mathscr{H}$ induced by the inner product defined in $\mathscr{H}$ as

$$\langle f, f'\rangle_{\mathscr{H}} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_i b_j^* K(x_i, x_j),$$

for $f = \sum_{i=1}^{\infty} a_i K_{x_i}$ and $f' = \sum_{j=1}^{\infty} b_j K_{x_j}$.

Since the reproducing kernel K completely characterizes the associated RKHS $\mathscr{H}$, in the sequel we shall denote that RKHS as $\mathscr{H}_K$ and its inner product as $\langle\cdot,\cdot\rangle_K$ with the associated norm $\|\cdot\|_K$. It is worth to emphasize that Definition 3.3 implies that all $f \in \mathscr{H}_K$ inherit their properties from the kernel, e.g., the continuity of K implies the continuity of all $f \in \mathscr{H}_K$ (Cucker and Smale 2001). It can be said that this property is the most crucial one and considered to be the main advantage of RKHSs-based estimators. High-level assumptions, e.g., smoothness, integrability, etc., can be easily encoded in $\mathscr{H}_K$ via the associated kernel function K. Another interesting property of RKHSs in the context of regularization methods is that, they allow to easily obtain a closed-form and unique solution of problem (3.4), even if the employed RKHS is an infinite-dimensional space. This comes from the following *Representer Theorem* (Kimeldorf and Wahba 1970; Schölkopf et al. 2001; Argyriou and Dinuzzo 2014; Suykens et al. 2002).

Theorem 3.1 (Representer Theorem) *If $\mathscr{H}_K$ is an RKHS, with K the associated kernel function, the solution of* (3.4) *for $\mathscr{H} = \mathscr{H}_K$ is unique and given by*

$$\hat{g}(\cdot) = \sum_{i=1}^{N} c_i K_{x_i}(\cdot), \quad (3.7)$$

where $c = [c_1\ \cdots\ c_N]^\top$ is defined by

$$c = (\mathcal{K} + \gamma I_{\mathrm{N}})^{-1} Y_N,$$

with I_{N} being the $N \times N$ identity matrix and $\mathcal{K}$ is the kernel matrix whose (i,j)-th entry is $K(x_i, x_j)$.

The similarity between the solution of the classical parametric approach and the regularized estimate can be easily seen from (3.3) and (3.7): both of them are a linear combination of some basis functions. However, a fundamental difference is that in (3.3) the basis functions are predefined and their number is independent of the size of data set. On the other hand, the number of the basis functions in (3.7), i.e., kernel sections at the input data $K_{x_i}(\cdot)$, is not fixed a priori and depends on the size of the data vector.

To illustrate the strength of the introduced regularization approach, i.e., regularization in RKHSs, over the classical approaches, we consider a simple simulation example. Consider the following unknown nonlinear function to be estimated

$$g_0(x) = \exp(\sin(8x)),\ 0 \leq x \leq 1. \tag{3.8}$$

Our goal is to reconstruct this function from a set of $N = 100$ noisy measurements generated according to

$$y_i = g_0(x_i) + \epsilon_i,\ \text{with } x \sim \mathcal{U}(0,1), \tag{3.9}$$

where $\mathcal{U}$ denotes a uniform distribution on the interval $[0,1]$ and $\epsilon_i \sim \mathcal{N}\left(0, 0.2^2\right)$, i.e., additive white Gaussian noise with variance $\sigma_\epsilon^2 = 0.2^2$. We adopt (3.4) with the *Gaussian kernel* that encodes the smoothness of the unknown function, see (3.14) for the exact definition of that kernel. Figure 3.1 shows three cases for different values of the regularization parameter γ:

- $\gamma \gg \gamma_{\text{opt}}$ as shown in the upper part of the figure: too large γ results in a large bias. The estimate is too smooth and unable to follow the data due to overweighting of the regularization term.
- $\gamma \ll \gamma_{\text{opt}}$ as shown in the middle part of the figure: too low γ results in an estimate with large variance due to the overweighting of the data-fit term. Hence, a too flexible model is obtained that overfits the measurements.
- $\gamma = \gamma_{\text{opt}}$ as shown in the lower part of the figure: The optimal regularization parameter is obtained by a so-called *Oracle* estimator that makes use of the knowledge of the true function to find the optimal γ that minimizes the MSE, achieving an optimal bias/variance trade-off.

It is worth to mention that the Oracle estimator is not implementable in reality. An attractive approach to tune γ to obtain a well-balanced bias/variance trade-off will be detailed in the next section in the light of the Bayesian interpretation of the considered estimator (Carlin and Louis 2000; Maritz and Lwin 1989). It can be seen from the above discussion that the choice of γ replaces the model order selection in the classical parametric approach. However, tuning γ can be done in a continuous manner, while in the classical framework, we choose a model from a set of discrete number of candidate models.

3.2.3 Orthonormal basis viewpoint of kernels

Under certain conditions, Mercer's theorem allows to represent the kernel function K in terms of orthonormal eigenfunctions and eigenvalues (Mercer 1909; Hochstadt 1988). First, let us introduce some definitions before stating Mercer's theorem.
Let $K : \mathscr{X} \times \mathscr{X} \to \mathbb{C}$ be a Mercer kernel where $\mathscr{X}$ is a metric space and not neces-

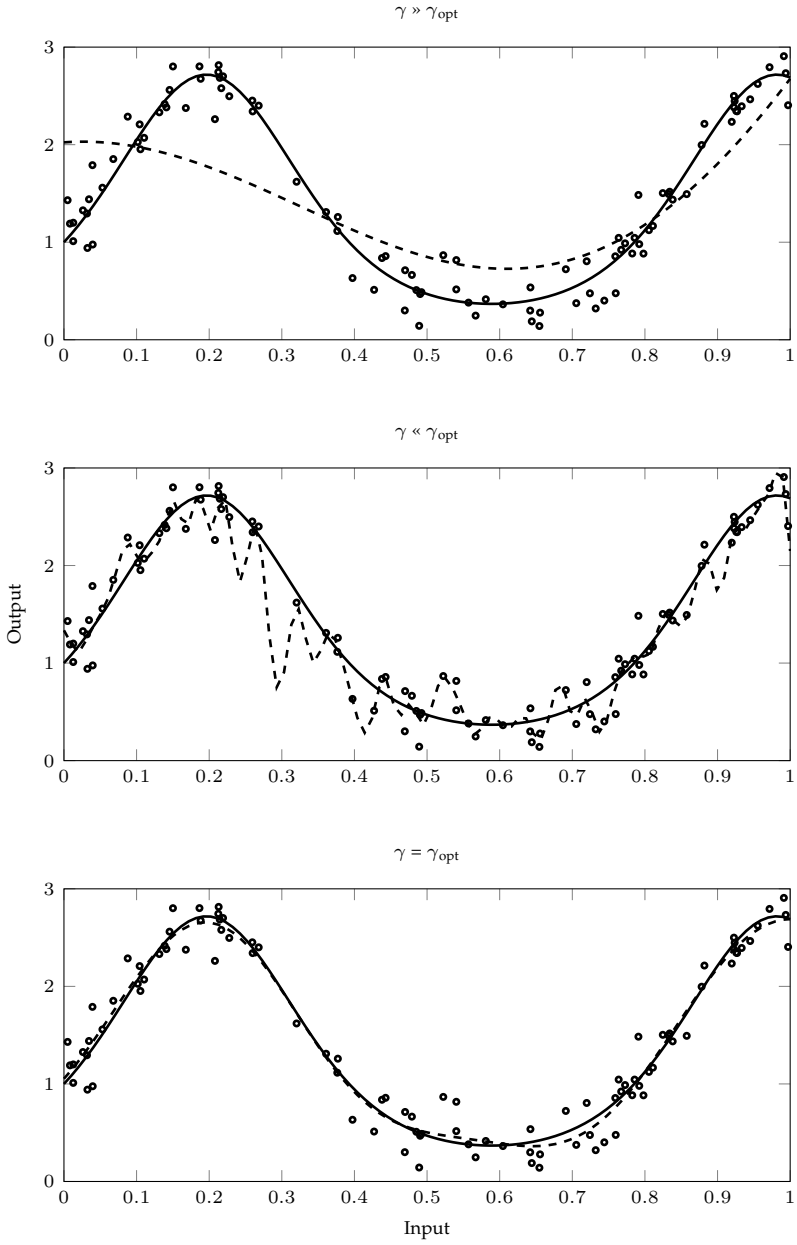


Figure 3.1: Controlling complexity with regularization in RKHSs: true underlying function (solid line), noisy data (o) and estimate (dashed line).

sarily compact[4]. Let μ be a nondegenerate Borel measure[5] on $\mathscr{X}$, meaning that for every nonempty open set $V \subset \mathscr{X}, \mu(V) > 0$. Assume a sequence of compactness structure for $\mathscr{X}$: $\mathscr{X} = \bigcup_{i=1}^{+\infty} \mathscr{X}_i$, where $\mathscr{X}_1 \subset \mathscr{X}_2 \subset \cdots \subset \mathscr{X}_i \subset \cdots$ and each $\mathscr{X}_i$ is compact with finite measure, i.e., $\mu(\mathscr{X}_i) < +\infty$. Moreover, any compact subset of $\mathscr{X}$ is contained in $\mathscr{X}_i$ for some i.
For a given kernel K and $\phi \in L_{2,\mu}(\mathscr{X})$, i.e., the Hilbert space of squared integrable functions on $\mathscr{X}$ for metric μ, we define the integral operator on $L_{2,\mu}(\mathscr{X})$:

$$\mathfrak{L}_K(\phi(x)) \triangleq \int_{\mathscr{X}} K(x, x')\phi(x')d\mu(x'),\ x \in \mathscr{X}. \tag{3.10}$$

To guarantee that $\mathfrak{L}_K(\phi) \in L_{2,\mu}(\mathscr{X})$, one can assume that the kernel is squared integrable, i.e.,

$$\int_{\mathscr{X}} \int_{\mathscr{X}} K^2(x, x')d\mu(x)d\mu(x') < +\infty \tag{3.11}$$

which also guarantees that the integral operator $\mathfrak{L}_K$ is bounded and compact[6].

Theorem 3.2 (Mercer's theorem) (Sun 2005) *Consider a Mercer kernel $K(x, x')$ defined on a not necessarily compact metric space $\mathscr{X}$ with a nondegenerate Borel measure μ. assume that*

A1) *$K_x \in L_{2,\mu}(\mathscr{X})$ for every $x \in \mathscr{X}$.*

A2) *K is squared integrable, i.e.,* (3.11) *is satisfied.*

Then, the following results hold:

R1) *$\mathfrak{L}_K$ is a positive, bounded and compact operator.*

R2) *$\mathfrak{L}_K$ has at most countably many positive eigenvalues $\left\{\breve{\lambda}_i\right\}_{i=1}^{\infty}$, such that $\breve{\lambda}_1 \geq \breve{\lambda}_2 \geq \cdots > 0$, and corresponding eigenfunctions (eigenvectors) $\{\varphi_i\}_{i=1}^{\infty}$ with $\varphi_i \in L_{2,\mu}(\mathscr{X})$,*

[4]A metric space $\mathscr{X}$ is called compact if and only if it is complete and totally bounded. It is said to be totally bounded if and only if for every real number $\varsigma > 0$, there exists a finite collection of open balls in $\mathscr{X}$ of radius ς whose union contains $\mathscr{X}$.

[5]A finite Borel measure on $\mathscr{X}$ is a map $\mu : \mathscr{B}(\mathscr{X}) \to [0, \infty)$ such that $\mu(\emptyset) = 0$ and

$$\text{if } \mathscr{X}_1, \mathscr{X}_2, \ldots \in \mathscr{B}(\mathscr{X}) \text{ are mutually disjoint } \Rightarrow \mu(\cup_{i=1}^{\infty} \mathscr{X}_i) = \sum_{i=1}^{\infty} \mu(\mathscr{X}_i),$$

where $\mathscr{B}(\mathscr{X})$ is known as the Borel σ-algebra (σ-field) and is the smallest σ-algebra in $\mathscr{X}$ that contains all open subsets of $\mathscr{X}$.

[6]Recall that a linear operator $\mathfrak{L} : \mathscr{H} \to \mathscr{H}$ where $\mathscr{H}$ is a Hilbert space, is called bounded if there exists a $M > 0$ such that for all $f \in \mathscr{H}$, where M is independent of f, $\|\mathfrak{L}(f)\|_{\mathscr{H}} \leq M\|f\|_{\mathscr{H}}$. It is said to be compact if for every bounded sequence $\{f_i\}_{i=1}^{\infty}$ in $\mathscr{H}$, the sequence $\{\mathfrak{L}(f_i)\}_{i=1}^{\infty}$ contains a convergent subsequence.

$$\langle \varphi_j, \varphi_k \rangle_{L_{2,\mu}} = \begin{cases} 1, & \text{if } j = k; \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, $\mathfrak{L}_K(\varphi_k(x)) = \breve{\lambda}_k \varphi_k(x)$.

R3) $K(x, x') = \sum_{k=1}^{\infty} \breve{\lambda}_k \varphi_k(x) \varphi_k(x')$ *where the series converges absolutely and uniformly on* $\bar{\mathscr{X}} \times \bar{\mathscr{X}}'$ *with* $\bar{\mathscr{X}}, \bar{\mathscr{X}}'$ *being any compact subsets of* $\mathscr{X}$.

R4) $\left\{ \sqrt{\breve{\lambda}_k}\, \varphi_k \right\}_{k=1}^{\infty}$ *forms an orthonormal basis for* $\mathscr{H}_K$, *the associated RKHS with* K.

Lemma 3.1 (Wahba 1990, Lemma 1.1.1 page 4). *Let* K *be a Mercer kernel that satisfies* (3.11). *If*

$$a_i = \int_{\mathscr{X}} f(x) \varphi_i(x)\, d\mu(x),$$

then $f = \sum_{i=1}^{\infty} a_i \varphi_i \in \mathscr{H}_K$ *if and only if*

$$\sum_{i=1}^{\infty} a_i^2 / \breve{\lambda}_i < \infty.$$

Furthermore, $\|f\|_K^2 = \sum_{i=1}^{\infty} a_i^2 / \breve{\lambda}_i$.

As a result of the above discussion, the RKHS $\mathscr{H}_K$ associated with a Mercer kernel K can be equivalently defined as (Rasmussen and Williams 2006)

$$\mathscr{H}_K = \left\{ f : \mathscr{X} \to \mathbb{C} \mid f(x) = \sum_{i=1}^{\infty} a_i \varphi_i(x), \text{ with } \sum_{i=1}^{\infty} a_i^2 / \breve{\lambda}_i < +\infty \right\}. \tag{3.12}$$

This means that any function $f \in \mathscr{H}_K$ can be represented as a linear combination of the orthonormal basis of the kernel K. Moreover, the inner product $\langle f, f' \rangle_K$ for any $f, f' \in \mathscr{H}_K$ with $f = \sum_{i=1}^{\infty} a_i \varphi_i$ and $f' = \sum_{i=1}^{\infty} b_i \varphi_i$ can be represented as

$$\langle f, f' \rangle_K = \sum_{i=1}^{\infty} a_i b_i^* / \breve{\lambda}_i.$$

It is worth to mention that the spectral representation in (3.12), which is related to the series expansion given in Item R3 of Theorem 3.2, is not unique since the eigen-decomposition depends on the measure μ. However, all such spectral forms associated with K lead to the same RKHS.

3.3 Gaussian process regression

In this section, GPR is briefly introduced. More specifically, important concepts of the Bayesian inference mechanism within the GP framework are introduced, i.e., concepts of *prior*, *likelihood* and *posterior*. Let us start by introducing the definition of a stochastic process (Kamen and Heck 2007).

Definition 3.4 (Stochastic process) *A stochastic process is $S(t, \mathcal{R})$, where $t \in \mathcal{I}$ represents the index set (time), and $\mathcal{R}$ is the outcomes of the sample space, realizations of functions (signals or time sequences), respectively.*

One can distinguish the following:

- $S(t, \mathcal{R})$ is a collection of realized functions (signals), i.e., $\{s_1(t), s_2(t), \ldots\}$;
- $S(t, r_i)$ is a particular realized function (signal), i.e., $s_i(t)$;
- $S(t_0, \mathcal{R})$ is a random variable, a possible collection of function (signal) values at time t_0, i.e., $\{s_1(t_0), s_2(t_0), \ldots\}$, for a given $r_i \in \mathbb{R}$;
- $S(t_0, r_i)$ is a particular value of s_i at index (time) t_0.

A probability distribution/multivariate distribution describes scalar/vector random variables, respectively, while a stochastic process extends such concept to functions. A GP is a distribution over functions and a generalization of the Gaussian distribution to an infinite-dimensional function space, i.e., by considering a function as an infinitely long vector with each entry to represent the function value at a certain input. It offers an attractive framework that is capable of capturing functional relations consistently with finite observations.

Definition 3.5 (Gaussian process) (Rasmussen and Williams 2006) *A GP is a collection of random variables, for which every finite number of these variables have a joint Gaussian distribution.*

As in the Gaussian distribution case, a GP is completely characterized by its mean and covariance functions

$$m(x) = \mathcal{E}\left\{f(x)\right\}, \qquad K(x, x') = \operatorname{cov}\left(f(x), f(x')\right),$$

respectively, where $f(x)$ is a real function and $K(x, x')$ is known as the kernel function and it specifies the covariance between any two function values at x and x'. In the sequel, we shall write the Gaussian process as

$$f(x) \sim \mathcal{GP}(m(x), K(x, x')).$$

3.3.1 Bayesian inference

In the standard GP regression, a data set $\mathcal{D}_N$ is assumed to be generated according to (3.1), see Section 3.1.1 for more details on the generating model. Bayesian inference is a 3-step approach (Rasmussen and Williams 2006; Bishop 2006), namely: i) specify a prior distribution on the unknown function g; ii) observe the data, i.e., $\mathcal{D}_N$; iii) compute the posterior distribution on g. Indeed, the posterior is a refinement of the prior based on the incorporated evidence from the observation. Next, we briefly discuss the above-mentioned three steps of the Bayesian inference within the GP framework.

Prior

With GPR, we consider the unknown function g as a random function and postulate a prior over it that represents our beliefs and the high-level assumptions about g, e.g., smoothness, stability, etc. More specifically, we assume a zero-mean GP prior on the unknown function g, i.e.,

$$g(x) \sim \mathcal{GP}(0, K_\beta(x, x')), \tag{3.13}$$

with K_β to denote the covariance/kernel function parameterized by an unknown hyperparameter vector β. For instance, a well-known choice for the kernel function that encodes both smoothness and stationarity of the unknown function is the *Radial Basis Function* (RBF), also known as *Squared Exponential* (SE) (Rasmussen and Williams 2006)

$$K_\beta(x, x') = \beta_\alpha^2 \exp\left(-\frac{1}{2}(x - x')^\top \Gamma^{-1}(x - x')\right), \tag{3.14}$$

where $x, x' \in \mathbb{R}^{n_\mathbb{X}}$, β_α^2 is a scaling parameter that represents the signal variance and $\Gamma = \operatorname{diag}\left(\left[\beta_{\mathrm{w}1}^2 \cdots \beta_{\mathrm{w}n_\mathbb{X}}^2\right]\right)$ is a diagonal matrix of squared characteristic length-scales $\{\beta_{\mathrm{w}i}\}_{i=1}^{n_\mathbb{X}}$. In this case, the so-called hyperparameter vector β, which characterizes the kernel, consists of β_α^2 and $\{\beta_{\mathrm{w}i}\}_{i=1}^{n_\mathbb{X}}$. The effect of the characteristic length-scale parameter, i.e., $\beta_{\mathrm{w}i}$, can be, informally, understood as the distance one has to move in the input space before the function value can significantly change. Figure 3.2 shows the effect obtained by varying β_w. It is worth to emphasize that the values of $\beta_{\mathrm{w}i}$ give the relevant importance of the associated inputs, i.e., if $\beta_{\mathrm{w}i}$ is very small, then x_i has a strong effect on the predicted output and vice versa. This can be used to automatically determine the relevant inputs to be included in the input regressor from data. The optimal choice of the covariance/kernel func-

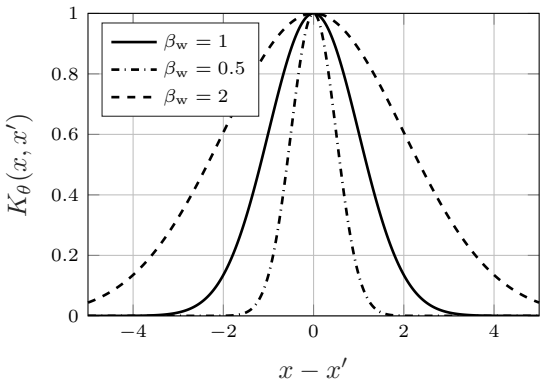


Figure 3.2: The covariance function K_β in (3.14) for $n_\mathbb{X} = 1$ and with $\beta_\alpha = 1$ and different values of β_w, i.e., $\beta_\mathrm{w} = 0.5, 1, 2$.

tion is problem dependent. A large variety of kernel functions is introduced in the literature. In general, the covariance functions can be classified into two major

groups, namely, stationary and non-stationary covariance functions (Rasmussen and Williams 2006, Chapter 4).

A stationary covariance function is a function of the distance between the inputs, i.e., $(x - x')$, and thus is invariant to translations in the input space. Moreover, if the covariance function is a function of only $|x - x'|$, then it is called *isotropic*. A well-known example is the SE covariance function, introduced in (3.14). It is infinitely differentiable and has squared integrable derivatives of all order (Rasmussen and Williams 2006), i.e., it is a smooth function. Other common stationary covariance functions are *exponential covariance*, *rational quadratic covariance*, *Matérn covariance*, *periodic covariance*, *Cubic Spline covariance* (CS), etc.

Non-stationary covariance functions might also be interesting in some cases, e.g., to describe the change of the underlying function behavior over time. Some common non-stationary covariance functions include, e.g., *linear covariance*, *polynomial covariance*, *neural network covariance*, etc. We refer the interested reader for more detailed exposition on this topic to Schölkopf and Smola (2002) and Hofmann et al. (2008).

Posterior

The prior is refined by incorporating evidences from the observations $\mathcal{D}_N$ which results in the posterior distribution over g. As can be seen by applying Bayes' Theorem

$$\underbrace{\mathrm{p}(g \mid X_N, Y_N, \beta)}_{\text{posterior}} = \frac{\overbrace{\mathrm{p}(Y_N \mid g, X_N, \beta)}^{\text{likelihood}} \times \overbrace{\mathrm{p}(g \mid \beta)}^{\text{prior}}}{\underbrace{\mathrm{p}(Y_N \mid X_N, \beta)}_{\text{marginal likelihood}}}, \tag{3.15}$$

where p denotes a *Probability Density Function* (PDF) and $\mathrm{p}(y \mid x)$ denotes that the distribution of y is conditioned on x. The likelihood in (3.15), i.e., $\mathrm{p}(Y_N \mid g, X_N, \beta)$, encodes the assumed noise model, i.e., if we assume additive independent and identical distributed (i.i.d.) Gaussian noise, the observations y_i will be conditionally independent given X_N, then the likelihood can be written as

$$\begin{aligned} \mathrm{p}(Y_N \mid g, X_N, \beta) &= \prod_{i=1}^{N} \mathrm{p}(y_i \mid g(x_i), \beta) \\ &= \prod_{i=1}^{N} \mathcal{N}\left(y_i \mid g(x_i), \sigma_\epsilon^2\right) \\ &= \mathcal{N}\left(Y_N \mid g(X_N), \sigma_\epsilon^2 I_N\right). \end{aligned} \tag{3.16}$$

Furthermore, $\mathrm{p}(Y_N \mid X_N, \beta)$ is known as the *Marginal Likelihood* (ML) or *evidence*, which is the likelihood of the hyperparameter β given the data $\mathcal{D}_N$ after marginalizing out the unknown function g, more details will follow in the next subsection. Finally, the prior, i.e., $\mathrm{p}(g \mid \beta)$, gives/delivers our beliefs about the unknown function.

Figure 3.3 illustrates the involved operations in Bayes' Theorem (3.15). More specifically, we start by the prior knowledge and by absorbing the available information in the data record (likelihood), i.e., simply multiply the prior distribution by the likelihood distribution and then normalize the result of the multiplication, the prior knowledge can be further improved and the posterior distribution is obtained. For a given value of the hyperparameter vector β, the GP prior assump-

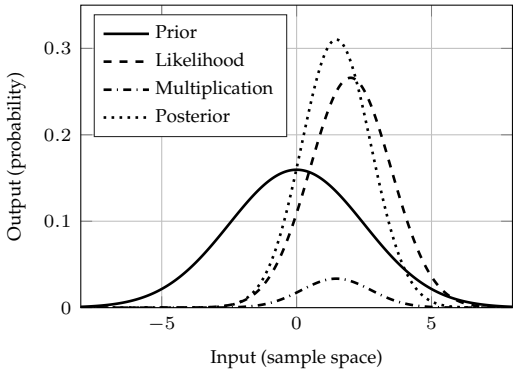


Figure 3.3: The prior distribution $\mathcal{N}\left(0, 2.5^2\right)$ and the likelihood or measurement distribution $\mathcal{N}\left(2, 1.5^2\right)$. To obtain the posterior distribution according to (3.15), we can simply multiply both the prior and likelihood distributions and then normalize the result.

tion together with the Gaussian assumption on the noise, i.e., $\epsilon(k) \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ and accordingly the likelihood is Gaussian, results in a GP posterior[7]. As a consequence, the associated estimation problem has an analytic form. The GP posterior distribution is used to make predictions about g at an arbitrary input $x_* \in \mathbb{R}^{n_\mathbb{x}}$ as will be shown later along with the expressions of the mean and covariance of the resulting GP posterior.

Tuning model complexity: choosing the hyperparameters

As mentioned earlier, Bayesian inference can be seen as a three-level scheme as shown in Figure 3.4, where at the right most of the figure we see the observation level, i.e., the available data record $\mathcal{D}_N$, Level 1 simply represents the unknown function g that we are interested in estimating in terms of its distribution given the data, Level 2 is the choice of hyperparameters that specify the distribution of the unknown functions values, e.g., how much these values are correlated, and Level 3 represents the possible different models, e.g., by considering various covariance functions, hence model structure selection is needed. However, in this thesis, we only consider a two-level inference scheme, i.e., Level 1 and 2 as we only consider a single covariance function that is chosen a priori. In the light of the considered two-level inference scheme, a critical step in the GPR framework is to design the

[7] Although, the likelihood is a finite probability distribution, the GP posterior is infinite dimensional because the GP prior is an infinite dimensional object (Deisenroth 2010).

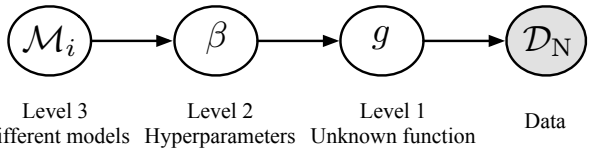


Figure 3.4: Graphical model of the three-level scheme in Bayesian inference.

kernel function that encodes the expected high-level assumptions about the unknown function. Such a step involves parameterizing the kernel function with a few number of parameters known as hyperparameters and at the same time making it flexible enough to describe a wide range of expected properties with such simple parameterization. Accordingly, the next step is to tune the hyperparameters from the observed data. This is a critical step, since these hyperparameters control the bias/variance trade-off and hence the "complexity" of the estimated function versus its accuracy.

In the following, we focus on hyperparameters estimation. To study the effect of varying the hyperparameters on GP prediction, consider 40 data points (red circles in Figure 3.5), generated from a GP with the SE kernel (3.14) using $(\beta_\alpha, \beta_{\mathrm{w}}, \sigma_\epsilon) = (1, 1, 0.1)$. Figure 3.5 (left part) shows twice the standard deviation σ of the prediction with GP posterior (used for prediction) calculated using the true hyperparameters, i.e., the 2σ bound which corresponds to a 95% confidence internal. It is interesting to note that uncertainty increases for the input values away from the training point due to the lack of information in these regions. Now, let us change the value of the hyperparameters of the posterior GP to $(\beta_\alpha, \beta_{\mathrm{w}}, \sigma_\epsilon) = (1, 0.3, 0.04)$. The effect of the new situation can be seen in the middle part of Figure 3.5. The model becomes more complex and overfits the data. By changing the hyperparameters to $(\beta_\alpha, \beta_{\mathrm{w}}, \sigma_\epsilon) = (1, 3, 0.9)$, the model becomes simple and unable to follow the data as can be seen in the right part of Figure 3.5.

Now, we have two options to go forward, w.r.t. tuning the hyperparameters: i) a fully Bayesian approach; ii) or a so-called empirical Bayes approach. In the fully Bayesian approach, a hyper-prior $\mathrm{p}(\beta)$, i.e., an assumed distribution of β, is placed on β and then it is integrated out:

$$
\begin{aligned}
\mathrm{p}(g) &= \int_\beta \mathrm{p}(g \mid \beta)\mathrm{p}(\beta)\, d\beta \\
\mathrm{p}(Y_N \mid X_N) &= \int_\beta \int_g \mathrm{p}(Y_N \mid X_N, g, \beta)\mathrm{p}(g \mid \beta)\mathrm{p}(\beta)\, dg\, d\beta \\
&= \int_\beta \mathrm{p}(Y_N \mid X_N, \beta)\mathrm{p}(\beta)\, d\beta,
\end{aligned}
$$

which is analytically intractable due to the complex expression resulting from $\mathrm{p}(Y_N \mid X_N, \beta)$, which is involved in the integration. One way to deal with such situation is to employ numerical approximation methods, e.g., *Monte-Carlo* (MC) methods (Svensson et al. 2015), which are computationally expensive. Alternatively, the empirical Bayes approach (Maritz and Lwin 1989; Carlin and Louis

2000) can be utilized. The main idea behind the empirical Bayes approach is to focus on obtaining a good point estimate $\hat{\beta}$ of β and then to condition our inference on that value instead of marginalizing out the hyperparameters. This approach will be detailed below. The posterior on the hyperparameters is given by

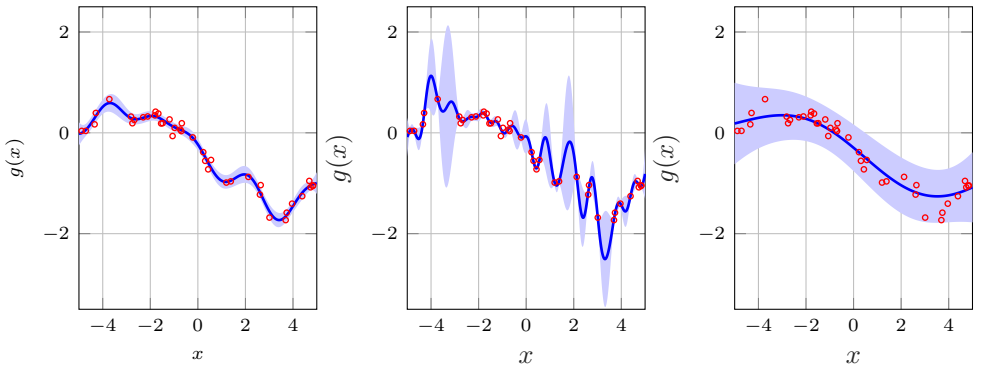


Figure 3.5: Left part: Data (red circle) is generated from a GP with $(\beta_\alpha, \beta_\mathrm{w}, \sigma_\epsilon) = (1, 1, 0.1)$ along with the 95% confidence interval (shaded area). Middle part: complex model with large variance when using $(\beta_\alpha, \beta_\mathrm{w}, \sigma_\epsilon) = (1, 0.3, 0.04)$. Right part: Simple model with large bias when using $(\beta_\alpha, \beta_\mathrm{w}, \sigma_\epsilon) = (1, 3, 0.9)$.

$$\mathrm{p}(\beta \mid X_N, Y_N) = \frac{\mathrm{p}(Y_N \mid \beta, X_N) \times \mathrm{p}(\beta)}{\mathrm{p}(Y_N \mid X_N)}, \tag{3.17}$$

where $\mathrm{p}(\beta)$ is the hyper-prior on β. A popular approach to learn the hyperparameters from data is to choose the hyper-prior $\mathrm{p}(\beta)$ to be flat, i.e., any values of β is equally possible a priori, which makes the posterior over β to be proportional to the marginal likelihood in (3.15), i.e., $\mathrm{p}(\beta \mid X_N, Y_N) \propto \mathrm{p}(Y_N \mid X_N, \beta)$. Hence, the *Maximum a Posterior (MAP)* estimate of the hyperparameters β equals the maximum marginal likelihood estimate. Therefore, the hyperparameters can be found by maximizing the marginal likelihood of the output w.r.t. to β (MacKay 1999)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \ \log \mathrm{p}(Y_N \mid X_N, \beta), \tag{3.18}$$

where the log-marginal likelihood function is

$$\begin{aligned}
\log \mathrm{p}(Y_N \mid X_N, \beta) &= \log \int_g \mathrm{p}(Y_N \mid g, X_N, \beta) \mathrm{p}(g \mid \beta) \, dg \\
&= -\frac{N}{2} \log(2\pi) - \underbrace{\frac{1}{2} Y_N^\top \left(\mathcal{K}_\beta + \sigma_\epsilon^2 I_\mathrm{N} \right)^{-1} Y_N}_{\text{“data-fit” term}} \\
&\qquad - \underbrace{\frac{1}{2} \log \det(\mathcal{K}_\beta + \sigma_\epsilon^2 I_\mathrm{N})}_{\text{complexity term}}.
\end{aligned} \tag{3.19}$$

Employing (3.18) to tune the hyperparameters leads to an automated trade-off between data-fit and model complexity (Pillonetto and Chiuso 2015; Rasmussen and Williams 2006; MacKay 2003). This can be seen as a manifest of *Occam's razor* principle to use the simplest model that explains the data (under the given prior assumptions) (Rasmussen and Williams 2006, Section 5.2). However, the price to be paid is that maximizing the marginal likelihood (3.18) is a nonlinear and non-convex optimization problem, prone to local minima. Other tuning methods to tune the hyperparameters are available, e.g., C_p statistics (Hastie et al. 2009), CV (Ljung 1999), *Predicted Residual Sums of Squares* (PRESS) (Wang and Cluett 1996), *Generalized Cross-Validation* (GCV) (Golub et al. 1979), *Stein's Unbiased Risk Estimator* (SURE) (Stein 1981). It is worth to mention that the superiority of maximizing the marginal likelihood over these classical tuning methods has been investigated in Pillonetto and Chiuso (2015), showing that it can better balance data fit and model complexity.

Univariate prediction

Given a test point $x_* \in \mathbb{R}^{n_\mathbb{X}}$, the joint distribution of the observed values, i.e., Y_N, and the function value at the arbitrary test location, i.e., $g(x_*)$, under the GP prior (3.13), is[8]

$$\begin{bmatrix} Y_N \\ g(x_*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathcal{K}(X_N, X_N) + \sigma_\epsilon^2 I & \kappa(X_N, x_*) \\ \kappa(x_*, X_N) & K(x_*, x_*) \end{bmatrix}\right) \tag{3.20}$$

where $\mathcal{K}(X_N, X_N)$ is the kernel matrix and its (i,j)-th element $[\mathcal{K}]_{ij} = K(x_i, x_j)$ and $\kappa(x_*, X_N) = \kappa^\top(X_N, x_*) = \begin{bmatrix} K(x_*, x_1) & \cdots & K(x_*, x_N) \end{bmatrix}$ denotes a vector of covariances between the test point and the training points. From (3.20) and by applying the well-known Gaussian identities (Rasmussen and Williams 2006), the predictive marginal distribution of $g(x_*)$ conditioned on the observations is Gaussian and is given by

$$\mathrm{p}(g(x_*) \mid X_N, Y_N, \beta) \sim \mathcal{N}\left(\hat{g}_{x_*}, \mathrm{cov}(g(x_*) \mid Y_N, X_N)\right) \tag{3.21}$$

where

$$\hat{g}_{x_*} = \kappa(x_*, X_N)\left[\mathcal{K}(X_N, X_N) + \sigma_\epsilon^2 I_\mathrm{N}\right]^{-1} Y_N = \sum_{i=1}^{N} c_i K(x_i, x_*), \tag{3.22}$$

is the minimum variance estimate of $g(x_*)$, i.e., it is equal to $\mathcal{E}\left\{g(x_*) \mid X_N, Y_N, x_*\right\}$, and c_i denotes the i-th element of

$$c = \left(\mathcal{K}(X_N, X_N) + \sigma_\epsilon^2 I_\mathrm{N}\right)^{-1} Y_N.$$

[8]Note that the dependency on the hyperparameter β is dropped since we replace β with the estimated value, e.g., maximizing the marginal likelihood. This is known also as the empirical Bayes method (Carlin and Louis 2000).

Furthermore, the associated covariance with the estimate $\hat{g}(x_*)$ in (3.22) is given as

$$\begin{aligned}\operatorname{cov}(g(x_*) \mid Y_N, X_N) &\triangleq \mathcal{E}\left\{g(x_*)g^\top(x_*) \mid X_N, Y_N\right\} \\ &= K(x_*, x_*) - \mathcal{k}(x_*, X_N)\left[\mathcal{K}(X_N, X_N) + \sigma_\epsilon^2 I_{\mathrm{N}}\right]^{-1}\mathcal{k}(X_N, x_*),\end{aligned} \tag{3.23}$$

which defines the uncertainty of the estimation.

To gain more insights into the Bayesian inference technique within the GPR framework, the left part of Figure 3.6 shows a few samples from a GP prior distribution that favors smooth functions by employing the Gaussian kernel. The shaded area represents a 2σ bound computed pointwise, i.e., at each input location x, . By observing some data, i.e., the red circles in the right part of Figure 3.6, the prior uncertainty about the unknown function has been significantly reduced as shown in the figure. Note that the dashed lines now represent samples from the GP posterior distribution that has been computed based on the observations. As mentioned, the shaded area represent the pointwise 2σ bound and it indicates the expected accuracy of predicting the unknown function value at previously unseen test inputs. More specifically, for test inputs that are located in the well-presented regions by training inputs, the uncertainty becomes small. On the other hand, for the test inputs that are located away from these regions, the GP posterior falls back to the GP prior, where no information is available, which can be easily seen at the left corner of the right panel of Figure 3.6.

Note that the uncertainty at the measurements, i.e., training points, is zero, since while developing the figure, we have assumed that the noise is not present. If the measurements are noisy, then the uncertainty will be accordingly increased at all the data points including the training points themselves.

Remark 3.1 *In the above discussion, we have considered making a prediction at a deterministic test input x_*, which will be also the case along the whole thesis. In case of making a prediction at uncertain test input, i.e., the input has a probability distribution, the mapping of such distribution through a nonlinear function, i.e., the GP posterior, results in a non-Gaussian and non-unimodal predictive distribution. As a result, an approximation has to be performed, e.g.,* moment matching *where the exact predictive distribution is approximated by a Gaussian distribution that possesses the same mean and covariance as the exact predictive distribution (Deisenroth 2010).*

Multivariate prediction

In the previous subsections, we have considered a univariate prediction, i.e., $x_* \in \mathbb{R}^{n_{\mathbb{X}}}, y_* \in \mathbb{R}$. However, in case of multivariate prediction, i.e., $y_* \in \mathbb{R}^{n_{\mathbb{Y}}}$, the correlation between different function values should be considered, To do so, the covariance function $K(x, x')$ can be replaced by a *covariance matrix*

$$K(x, x') = \begin{bmatrix} K_{11}(x, x') & \cdots & K_{1n_{\mathbb{Y}}}(x, x') \\ \vdots & \ddots & \vdots \\ K_{n_{\mathbb{Y}}1}(x, x') & \cdots & K_{n_{\mathbb{Y}}n_{\mathbb{Y}}}(x, x') \end{bmatrix}, \tag{3.24}$$

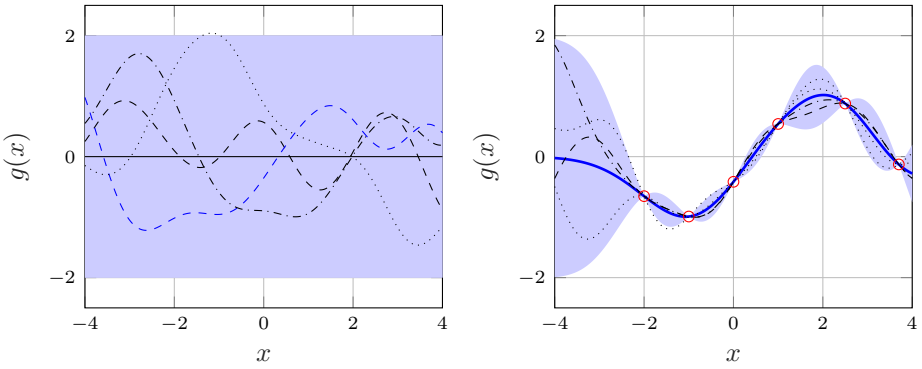


Figure 3.6: Left part: Samples from the GP prior distribution. The mean is shown with a solid line. With no observed data, the prior uncertainty about the unknown function is constant everywhere. Right part: Samples from the GP posterior distribution after observing 6 data points (red circles). The mean of the posterior is also shown with a solid line. The posterior uncertainty is significantly reduced due to the added information from the measurements. Such uncertainty largely depends on the locations of the training data points.

where $K_{i,j}$ denotes the covariance between the i-th and j-th output channels. However, by assuming that the function values $g_1(x_*), \ldots, g_{n_\mathbb{Y}}(x_*)$ are conditionally independent given an input x_*, then the off-diagonal entries in (3.24) become 0. Indeed, the output values from different output dimensions can be only correlated via x and since we are considering a deterministic input, then the target outputs are independent, otherwise, i.e., if the input x is uncertain, then the off-diagonal entries should be taken into account. So, in case of multivariate prediction with a deterministic test input x_*, $n_\mathbb{Y}$ independent GP models are trained with the same training inputs X_N and with different training outputs $Y_{\mathrm{N},i} = [y_i(1) \cdots y_i(N)]^\top$, $i = 1, \ldots, n_\mathbb{Y}$.

3.4 Numerical implementation

The overall algorithm for regularized estimation consists mainly of two parts:

- Hyperparameters estimation: This step involves the minimization of the cost function, i.e., the log-likelihood (3.18) whose single evaluation requires $\mathcal{O}(N^3)$ operations;
- Computation of the function estimate: This step requires $\mathcal{O}(N^3)$ operations.

It is well-known that training and predicting with GP becomes problematic when the data set size becomes large as the computational burden becomes prohibitively expensive.

3.4.1 Hyperparameters optimization methods

There are many available options for the optimization of the log-likelihood function.

Deterministic optimization methods, using e.g., gradient-based methods (Nelles 2001) like *conjugate gradient* and *trust region* methods (Zhang and Leithead 2005; Rasmussen and Williams 2006), require the computation of the partial derivatives of the objective function:

$$\frac{\partial \mathrm{p}(Y_N \mid X_N, \beta)}{\partial \beta_i} = -\frac{1}{2}\mathrm{tr}\left(\mathcal{K}_\beta^{-1}\frac{\partial \mathcal{K}_\beta}{\partial \beta_i}\right) + \frac{1}{2}Y_N^\top \mathcal{K}_\beta^{-1}\frac{\partial \mathcal{K}_\beta}{\partial \beta_i}\mathcal{K}_\beta Y_N. \tag{3.25}$$

where $\mathrm{tr}(\cdot)$ is the trace operator. The computation of the partial derivative involves the computation of the inverse of the covariance matrix of size $N \times N$ during every iteration with the computational complexity $\mathcal{O}(N^3)$. However, the results of such methods depend heavily on the initial values of the hyperparameters. This becomes even more relevant in case of multidimensional systems with many local optima associated with the objective function. Alternatively, stochastic optimization methods can be employed to deal with the multiple optima issue, e.g., genetic algorithm, differential evolution and particle swarm optimization. A drawback of these methods is that the computational complexity becomes a prohibitive factor in a large scale situation, see (Kocijan 2016, Section 2.4.2) for more details.

3.4.2 Numerical implementation

GP modeling has a noticeable drawback related to computational implementation, since it involves expensive operations as matrix inversion and calculation of the log determinant of the covariance matrix, which restricts the number of data points that can be handled within this framework. As a result, an efficient implementation is needed rather than plain computation of the above-mentioned operations. A common and practical approach is to employ Cholesky decomposition of $\mathcal{K}$ to compute the objective function (3.19) and its derivative (3.25) to avoid the direct inversion of the covariance matrix. This approach is summarized in Algorithm 1.

Algorithm 1, which tackles the GPR implementation using Cholesky decomposition instead of directly inverting the covariance matrix, provides a fast, efficient and numerically stable implementation. The computational complexity is significantly decreased: i) Cholesky factorization in Line 4 costs $\mathcal{O}(N^3/6)$; ii) Solving the two triangular systems in Line 5 costs $\mathcal{O}(N^2/2)$; iii) $\mathbf{L}\backslash\mathcal{k}$ in Line 10 costs $\mathcal{O}(N^2/2)$ for each test input. Note that, for large N, it may not be possible to represent the determinant, which is needed to compute the cost function, i.e., the log-likelihood. However, such a problem can be tackled via representing the log determinant with the Cholesky decomposition as shown in Line 11.

There are many other approaches to deal with large data records, e.g.,

Algorithm 1 Numerically efficient implementation of the GPR approach.

Require: X_N (inputs), Y_N (outputs), K (covariance function), x_* (test input), σ_ϵ^2 (noise variance).

1: **Hyperparameter tuning:**
2: Initialize β.
3: Compute $\mathcal{K}_\beta$.
4: Compute $\mathbf{L} = \text{Cholesky}(\mathcal{K}_\beta + \sigma_\epsilon^2 I_\text{N})$.
5: Solve $\mathbf{L}\gamma_t = Y_N$ for γ_t and $\mathbf{L}^\top \gamma_{t'} = \gamma_t$ for $\gamma_{t'}$ to get $\gamma_{t'} = \mathcal{K}_\beta^{-1} Y_N$.
6: Compute $\log \mathrm{p}(Y_N \mid X_N, \beta)$ and $\frac{\partial \mathrm{p}(Y_N|X_N,\beta)}{\partial \beta}$ using $\gamma_{t'}$
7: Repeat: Change β and go to Step 3 until $-\log \mathrm{p}(Y_N \mid X_N, \beta)$ is minimized.
8: **Function estimation:**
9: Compute the prediction mean $g(x_*) = \mathcal{k}^\top \gamma_{t'}$
10: Compute the prediction variance $\text{cov}(g(x_*)) = K(x_*, x_*) - (\mathbf{L}\backslash\mathcal{k})^\top (\mathbf{L}\backslash\mathcal{k})$.
11: Compute log-likelihood $\log \mathrm{p}(Y_N \mid X_N, \beta) = -\frac{1}{2} Y_N^\top \gamma_{t'} - \sum_i \log \mathbf{L}_{ii} - \frac{N}{2} \log 2\pi$

Return: $g(x_*)$, $\text{cov}(g(x_*))$ and $\log \mathrm{p}(Y_N \mid X_N, \beta)$.

1. sparse GP learning (Quiñonero-Candela and Rasmussen 2005): These approaches are aiming at reducing the computational burden associated with training and prediction by finding a low-rank approximation of $\mathcal{K}$;

2. online GP learning including evolving GP (Petelin and Kocijan 2011): For these methods, the structure of the GP model and the associated hyperparameters are adapted online and the computational complexity is kept controlled by keeping the size of the informative data set restricted according to the considered application;

3. *Local* GP (LGP) (Nguyen-Tuong et al. 2010): Inspired by locally weighted regression, a method for speeding-up the training and prediction process has been presented, where the training data is partitioned into local regions and an independent Gaussian process model is learnt for each region. The number of data points in the local models is limited, where insertion and removal of data points can be treated in a principled manner. The prediction for a query point is performed by weighted average.

3.5 The connection between GPR and RKHSs

The connection between regularized function estimation in RKHSs and Bayesian estimation of continuous-time GP was initially studied in Kimeldorf and Wahba (1970) in the context of spline regression (Wahba 1990). More specifically, the regularization network (3.4) has a statistical interpretation, where the function g is assumed to be a particular realization of a zero-mean GP with a prior covariance proportional to K, i.e., the reproducing kernel associated with $\mathscr{H}_K$, and that function is assumed to be independent of an additive white Gaussian measurement noise. The latter setting has been investigated in details in Section 3.3, where it has been shown that the predicted function value at a test input, i.e., $\hat{g}(x_*)$ is the

posterior mean and hence the minimum variance estimate of $g(x_*)$. Moreover, such estimate has a closed-form as given in (3.22) and it coincides with the expression obtained by the representer theorem (3.7). Next, we show the details of such a connection in the Gaussian measurement noise case, which is of special interest for our purposes. The connection in the non-Gaussian case is not discussed here, since it is not relevant to the work in the subsequent chapters, however, a detailed discussion on that case can be found in Aravkin et al. (2015). The following discussion depends largely on Aravkin et al. (2015).

Let us start by setting up the necessary assumptions.

Assumption 3.1 *Given a data set $\mathcal{D}_N = \{x_i \in \mathscr{X}, y_i \in \mathbb{R}\}_{i=1}^N$, which is generated according to* (3.1)*, i.e., in the presence of an additive measurement noise and a covariance function K on $\mathscr{X} \times \mathscr{X}$ that satisfies Definition 3.1 such that for any finite sequence of points $\{x_i\}_{i=1}^n$, the vector $[g(x_1) \ \cdots \ g(x_n)]^\top$ is a zero-mean Gaussian random variable with the covariance between any two elements $\operatorname{cov}(g(x_i), g(x_j)) = K(x_i, x_j)$, i.e., g is a zero-mean Gaussian random field on $\mathscr{X}$.*

Assumption 3.2 *For the given data set, i.e., $\mathcal{D}_N$, the corresponding loss function denoted by $\mathcal{V}$ is a function of $(y_i - g(x_i))$. For a given positive scalar σ_{e}, we have*

$$p(Y_N \mid g) \propto \prod_{i=1}^{N} \exp\left(-\frac{\mathcal{V}(y_i - g(x_i))}{2\sigma_{\mathsf{e}}^2}\right). \tag{3.26}$$

Finally, the measurement noise is a random variable $e_i = y_i - g(x_i)$ independent of the random function g.

The special case, where the loss function is $\mathcal{V}(y_i - g(x_i)) = (y_i - g(x_i))^2$, is equivalent to the situation of Gaussian measurements noise, i.e., $\{e_i\}$ are i.i.d. Gaussian random variables with variance σ_{e}^2. In Assumption 3.2, it has been assumed that g and e are independent, which means that $g(x)$ and Y_N are jointly Gaussian for any $x \in \mathscr{X}$. As a result, the posterior $\mathrm{p}(g(x) \mid Y_N)$ is also Gaussian. The posterior mean, i.e., $\mathcal{E}\{g(x \mid Y_N)\}$, and variance, i.e., $\operatorname{cov}(g(x) \mid Y_N)$, can be easily obtained, as given in (3.22), (3.23), respectively, which shows that in the considered Gaussian case the minimum variance estimate coincides with $\hat{g}$ in (3.7) obtained with the representer theorem as the solution of the regularization network (3.4). The following proposition summarizes the above discussion.

Proposition 3.1 *Suppose the unknown function g satisfies Assumption 3.1 and $p(Y_N \mid g)$ satisfies Assumption 3.2 with the loss function $\mathcal{V}(y_i - g(x_i)) = (y_i - g(x_i))^2$. Then, the minimum variance estimate of $g(x)$ given the observation Y_N is given by* (3.4)*, with $\gamma = \sigma_{\mathsf{e}}^2$ and $\mathscr{H}$ is the RKHS induced by K.*

3.6 Summary

In this chapter, some preliminaries on kernel-based methods in machine learning have been introduced to provide the required background for the next chapters.

First, we have defined the regression problem along with the data-generating model in Section 3.1. As a next step, the classical approach to tackle such problem based on the LS approach is given, followed by a discussion that clarifies the difficulties associated with the selection of the required model complexity within the classical approaches and the importance of the resulting bias/variance trade-off. This has brought us to specific regularization approaches, i.e., kernel-based methods, that cope with these issues and can be handled within a unified framework of RKHSs. The regularization in RKHSs including the definition of the kernel function and the associated estimation problem has been introduced in Section 3.2. Moreover, attractive properties, i.e., analytic solution via the representer theorem and inclusion of high-level assumptions and prior knowledge via the kernel function, of such estimators have been also discussed.

To tune the unknown hyperparameters associated with the kernel function, an empirical Bayes approach can be utilized, hence, the statistical interpretation, i.e., Bayesian inference within GPR framework, of the considered regularization approach has been introduced in Section 3.3. Based on such an interpretation, a maximum marginal likelihood approach to tune the unknown hyperparameters has been thoroughly discussed. The issues associated with the implementation and computational complexity of these methods, including tuning the hyperparameters and estimating the unknown function, has been briefly discussed in Section 3.4. Finally, the connection between both regularization in RKHSs and the Bayesian estimator of a Gaussian random field has been given in Section 3.5.

In the next chapter, we will discuss how to make the kernel-based methods practical for dynamic system identification. Then, we will investigate how to combine both concepts from machine learning community (Chapter 3) and system theory for representing dynamic systems (Chapter 2) to design more efficient regularized methods for LTI systems identification in both the time and the frequency domains.

4 CHAPTER

Bayesian Identification of LTI systems: An OBFs approach

This chapter is aiming at addressing Subgoal 1, i.e., systematic construction of a kernel function that can describe a wide range of dynamic properties with a low-dimensional parameterization for the identification of LTI systems both in the time- and frequency-domain. This chapter is organized as follows. In Section 4.1, the transition from machine learning to dynamic systems identification is introduced, where we show how the approaches presented in Chapter 3 can be utilized for such a task, and also discuss the available kernel functions for LTI system identification. This is followed by motivating the need for a new class of kernels to describe the dynamic properties of LTI systems. In Section 4.2, the OBFs based kernels in time-domain are presented and used to construct a well-designed RKHS for impulse response estimation. Finally, in Section 4.3, the frequency-domain formulation of these kernels is introduced and directly applied for frequency-domain estimation of LTI dynamics. Monte-Carlo simulations show that OBFs-based kernels perform well compared with the existing kernel functions, e.g., the TC and DC kernels, especially for slow systems with dominant poles close to the unit circle. Moreover, the capability of Kautz basis to model resonant systems is also shown.

4.1 From machine learning to system identification

In Chapter 3, kernel-based methods in machine learning have been discussed. The unknown function g is considered to be, in general, a static function, i.e., the underlying relation between the input and output does not depend on time. However, for dynamic systems, such a relation which can correspond to various representation forms of the system is dynamic and is dependent on time, more specifically, it depends on the actual past trajectory of these signals, i.e., the input

and output signals, as dynamic systems exhibit memory. Moreover, in the static case, we utilize prior knowledge about the unknown function, e.g., smoothness, but for dynamic systems, there are other properties that are needed to be taken into account in addition to smoothness, such as stability to restrict the search space for the model estimate to the actual assumptions of the identification setting. Accordingly, the presented approaches in Chapter 3 need to be adapted before being applied to system identification; otherwise, unsatisfactory results may be obtained as shown later in this chapter. Furthermore, it is also important to consider which representation form of the to-be-captured system is most suitable to use machine learning methods for system identification.

4.1.1 Problem statement

Data-generating system

Consider a SISO DT-FD-LTI stable data-generating system

$$y(t) = G_0(q)u(t) + v(t), \tag{4.1}$$

where $v(t)$ is an additive noise process. In the following and for the sake of simplicity, $v(t)$ is assumed to be a white Gaussian noise process with variance σ_{e}^2, independent of the input u. The case when v is colored can be handled in a straightforward way as shown in (Pillonetto et al. 2011a, Section 5.3). The transfer operator $G_0(q)$ can be represented as (see Chapter 2):

$$G_0(q) = \sum_{k=1}^{\infty} \mathrm{g}(k) q^{-k}, \tag{4.2}$$

where $G_0 \in \mathscr{RH}_{2-}(\mathbb{E})$ and $\mathrm{g} = \{\mathrm{g}(k)\}_{k=1}^{\infty}$ is the impulse response of the system. Note that, it is assumed without loss of generality that $G_0(q)$ does not have a feedthrough term, i.e., $\mathrm{g}(0) = 0$. The corresponding frequency response can be defined as

$$G_0(e^{\mathrm{j}\omega}) = \sum_{k=1}^{\infty} \mathrm{g}(k) e^{-\mathrm{j}\omega k}. \tag{4.3}$$

Note that this representation form of the underlying system is well suited for the function estimation concept of the machine learning methods: i) g is a function of time; ii) data appears in a linear convolution structure; iii) no explicit choice of model order or parameterization is needed as the IIR can express all stable systems with arbitrary finite order; and iv) relationship or the corresponding function estimation concept does not change fundamentally given time- or frequency-domain data, see Section 4.3.

Given N data points $\mathcal{D}_N = \{u(t), y(t)\}_{t=1}^{N}$, generated by (4.1), our goal is to find an estimate $\hat{\mathrm{g}}$ of g that is as good as possible, in the sense that the MSE of such an estimate, i.e., $\hat{\mathrm{g}}$, is minimized. Accordingly, an estimate $\hat{G}_N(e^{\mathrm{j}\omega})$ of the frequency response $G_0(e^{\mathrm{j}\omega})$ can be obtained.

Classical parametric approach

It has been discussed in Chapter 2 that one popular approach to deal with the problem of identifying LTI systems based on IO data is to follow the prediction error framework. The starting point is to choose (postulate) a parameterized model structure $G(q,\theta)$. Then, based on a given IO data of length N, we obtain an estimate $\hat{\theta}_N$ of θ by minimizing the squared prediction error. This in fact yields the model estimate $\hat{G}_N(e^{\mathrm{j}\omega}) = G(e^{\mathrm{j}\omega}, \hat{\theta}_N)$. As previously discussed, among the available model structures, see Table 2.1, the truncated impulse response model, i.e., FIR, offers an attractive model structure that enjoys the linear-in-the-parameters property. This property ensures that the parameters can be estimated via a convex optimization problem according to the PEM estimation concept. Specifically, consider a truncated IIR model, i.e., FIR of order n,

$$G(q,\theta) = \sum_{k=1}^{n} \mathfrak{g}(k) q^{-k}, \;\; \theta = [\mathfrak{g}(1) \; \cdots \mathfrak{g}(n)]^\top. \tag{4.4}$$

By writing the model as

$$y(t) = \gamma_{\mathrm{r}}^\top(t)\theta, \quad \gamma_{\mathrm{r}}^\top(t) = [u(t-1) \; \cdots \; u(t-n)], \tag{4.5}$$

or equivalently in a vector form

$$Y_N = \Upsilon_N \theta + \mathcal{V}_N, \tag{4.6}$$

where

$$Y_N = [y(1) \; \cdots \; y(N)]^\top,$$
$$\Upsilon_N = [\gamma_{\mathrm{r}}(1) \; \cdots \; \gamma_{\mathrm{r}}(N)]^\top,$$
$$\mathcal{V}_N = [v(1) \; \cdots \; v(N)]^\top,$$

and by following the well-known LS solution, we obtain:

$$\hat{\theta}_N^{\mathtt{LS}} = \left[\hat{\mathfrak{g}}^{\mathtt{LS}}(1) \; \cdots \; \hat{\mathfrak{g}}^{\mathtt{LS}}(n)\right]^\top = \operatorname*{argmin}_{\theta} \; \mathcal{W}_N(\theta), \tag{4.7}$$

where

$$\mathcal{W}_N(\theta) = \|Y_N - \Upsilon_N\theta\|_2^2 = \sum_{t=1}^{N} \left(y(t) - \gamma_{\mathrm{r}}^\top\theta\right)^2. \tag{4.8}$$

The analytic solution of (4.7) is given by

$$\hat{\theta}_N = \left[\frac{1}{N}\Upsilon_N^\top\Upsilon_N\right]^{-1}\left[\frac{1}{N}\Upsilon_N^\top Y_N\right]. \tag{4.9}$$

Remark 4.1 *The issue of unknown initial conditions required to form the regressors γ_{r}, i.e., the unknown inputs $\{u(k)\}_{k=-n+1}^{0}$, can be dealt with in various ways (Ljung 1999):*

- Non-windowed *approach, where the first n data points in the data set are not used and the summation in* (4.8) *starts from $n+1$;*
- Pre-windowed *approach, where the unknown inputs are assumed to be zero;*
- *Estimating the transient, i.e., the effect of initial condition, as an additional FIR model with impulsive input.*

In the following, the *Pre-windowed* approach is considered for the sake of notational simplicity.

Bias/variance trade-off

A well-known measure to quantify the quality of the resulting estimates is the frequency wise MSE, or the expected spectral error, which in the SISO case is (Ljung 1999)

$$\mathcal{M}_N(\omega) = \mathcal{E}\left\{\left|\hat{G}_N(e^{\mathrm{j}\omega}) - G_0(e^{\mathrm{j}\omega})\right|^2\right\}, \tag{4.10}$$

where the expectation $\mathcal{E}$ is taken w.r.t. the noise process v. Such an expression of the MSE can be divided into two parts, the bias $\mathcal{B}_N(\omega)$ and the variance parts $\mathcal{V}_N(\omega)$:

$$\mathcal{B}_N(\omega) = \mathcal{E}\left\{\hat{G}_N(e^{\mathrm{j}\omega})\right\} - G_0(e^{\mathrm{j}\omega}), \tag{4.11}$$

$$\mathcal{V}_N(\omega) = \mathcal{E}\left\{\left|\hat{G}_N(e^{\mathrm{j}\omega}) - \mathcal{E}\{\hat{G}_N(e^{\mathrm{j}\omega})\}\right|^2\right\}, \tag{4.12}$$

and the MSE can be equivalently written as

$$\mathcal{M}_N(\omega) = \mathcal{V}_N(\omega) + |\mathcal{B}_N(\omega)|^2. \tag{4.13}$$

When the model becomes flexible, i.e., more complex with more parameters, the bias term $\mathcal{B}_N$ decreases and the variance term $\mathcal{V}_N$ increases and vice versa. Instead of aiming at an unbiased estimate, which may be associated with high variance, it is often useful to allow some bias to reduce the variance in order to further reduce the MSE. Indeed, this is the main idea employed by the regularization approaches detailed in Chapter 3. It has been discussed at the beginning of this section that these approaches should be adapted to be applied to dynamic system identification, which is the topic of the next section.

4.1.2 Regularization techniques for dynamic system identification

Next, the regularization approaches presented in Chapter 3 are adapted to be applicable in case of dynamic system identification in order to tackle the issues associated with LTI system identification, e.g., model order selection, and to optimize the bias/variance trade-off (Pillonetto and De Nicolao 2010; Pillonetto et al. 2014).

Regularized estimation of IIR models

The main idea is to employ the regularization network (3.4), where the unknown function is considered to be the impulse response g of the system and $\mathscr{H}_K$ is constructed such that it is the RKHS of impulse responses of LTI systems induced by a kernel $K : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$. These spaces are assumed to contain only causal functions that represent impulse responses g of stable LTI systems. Denote $\mathscr{L}_i[\mathrm{g}]$ the functional that is defined as:

$$\begin{aligned}\mathscr{L}_i[\mathrm{g}] =&(\mathrm{g} \circledast u)(t_i),\\ =&\sum_{k=1}^{\infty} \mathrm{g}(k)u(t_i - k).\end{aligned} \tag{4.14}$$

The estimation of g given $\mathcal{D}_N$ is then accomplished by:

$$\hat{\mathrm{g}} = \underset{\mathrm{g}\in\mathscr{H}_K}{\operatorname{argmin}} \sum_{i=1}^{N} \big(y_i - \mathscr{L}_i[\mathrm{g}]\big)^2 + \gamma\|\mathrm{g}\|_K^2, \tag{4.15}$$

Note that (4.15) coincides with (3.4) except that $\mathscr{L}_i[\mathrm{g}]$ replaces g in terms of characterizing the "data-fit". Moreover, it is known that (3.4) admits a finite-dimensional solution according to Theorem 3.1. If the linear functionals $\mathscr{L}_i$ are continuous on $\mathscr{H}_K$, i.e.,

$$\forall i,\ \exists\, \varsigma_i < \infty :\ \big|\mathscr{L}_i[\mathrm{g}]\big| \leq \varsigma_i\|\mathrm{g}\|_K,\ \forall \mathrm{g} \in \mathscr{H}_K,$$

then, (4.15) also admits a finite-dimensional representation according to the following theorem.

Theorem 4.1 (Representer theorem for system identification) (Pillonetto et al. 2014) *If $\mathscr{H}_K$ is an RKHS, with K the associated kernel function and all $\mathscr{L}_i$ are continuous linear functionals on $\mathscr{H}_K$, the solution of* (4.15) *is unique and given by*

$$\hat{g}(\cdot) = \sum_{i=1}^{N} c_i\mathscr{L}_i\big[K_i(\cdot)\big], \tag{4.16}$$

where $c = [c_1\ \cdots\ c_N]^\top$ is given by

$$c = (\mathcal{K}^{\mathrm{o}} + \gamma I_{\mathrm{N}})^{-1}Y_N,$$

and the (i, j)-th entry of $\mathcal{K}^{\mathrm{o}}$, the so-called output kernel matrix, is given by:

$$\begin{aligned}K^{\mathrm{o}}(i, j) =&\big(u \circledast (u \circledast K)(t_i)\big)(t_j),\\ =&\sum_{l=1}^{\infty}\left(\sum_{k=1}^{\infty} K(k, l)u(t_i - k)\right)u(t_j - l).\end{aligned} \tag{4.17}$$

Note that the solution (4.16) is similar to the expression obtained in the static case, i.e., (3.7), but with replacing the kernel sections, i.e., K_i, see Definition 3.2, with their convolution with the input u, i.e., $\mathscr{L}_i[K_i(\cdot)] = \sum_{k=1}^{\infty} K(\cdot, k)u(t_i - k)$.

Regularized estimation of FIR models

For IIR models, the function domain is $\mathscr{X} = \mathbb{N}$; whereas, for FIR models of order n, i.e., $\theta = [\mathrm{g}(1) \ \cdots \mathrm{g}(n)]^\top$ in (4.4):

$$\mathscr{X} = \{1, 2, \ldots, n\}$$

$$K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$$

$$\mathscr{L}_i[\theta] = \gamma_{\mathrm{r}}^\top(i)\theta,$$

where γ_{r} is the regressor and is defined in (4.5). Since K is a positive definite kernel, this results in $\mathcal{K} \in \mathbb{R}^{n\times n}$, which is a symmetric and positive definite kernel matrix. As a result, there is a unique $\mathscr{H}_K$ of real-valued functions over the domain $\mathscr{X}$ with a finite number of kernel sections, i.e., the columns of $\mathcal{K}$. Accordingly, any $\mathrm{g}_\theta \in \mathscr{H}_K$ can be written as a linear combination of the kernel sections:

$$\mathrm{g}_\theta(\cdot) = \sum_{i=1}^{n} c_i K(i, \cdot), \tag{4.18}$$

with $\|\mathrm{g}_\theta\|_K^2 = c^\top \mathcal{K} c$, where $c = [c_1 \ \cdots \ c_n]^\top$. From (4.18), it can be easily seen that $\theta = \mathcal{K}c$ and $\|\mathrm{g}_\theta\|_K^2 = \theta^\top \mathcal{K}^{-1}\theta$. Then, (4.15) becomes equivalent to the following *Regularized* LS (ReLS) problem:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \ \|Y_N - \Upsilon_N\theta\|_2^2 + \gamma\theta^\top \mathcal{K}^{-1}\theta, \tag{4.19}$$

and the solution $\hat{\theta}$ that represents the *Regularized* FIR (RFIR) can be obtained by some algebraic manipulation:

$$\begin{aligned}\hat{\theta} &= \left(\Upsilon_N^\top\Upsilon_N + \gamma\mathcal{K}^{-1}\right)^{-1}\Upsilon_N^\top Y_N,\\ &= \mathcal{K}\Upsilon_N^\top\left(\Upsilon_N\mathcal{K}\Upsilon_N^\top + \gamma I_{\mathrm{N}}\right)^{-1} Y_N.\end{aligned} \tag{4.20}$$

By noting that $\mathscr{L}_i[K_i(\cdot)]$ in (4.16) is simply $\mathcal{K}\Upsilon_N^\top$ and $c = (\Upsilon_N\mathcal{K}\Upsilon_N^\top + \gamma I_{\mathrm{N}})^{-1}Y_N$, where

$$\mathcal{K}^{\mathrm{o}} = \Upsilon_N\mathcal{K}\Upsilon_N^\top, \tag{4.21}$$

a similar expression for $\hat{\theta}$ can be obtained using Theorem 4.1.

Connection with the Bayesian estimator of GP

A similar connection between the RKHS estimator (4.15) and the Bayesian estimator of continuous-time GP, as shown in Section 3.5, can be established (Pillonetto et al. 2014). As previously explained, such a connection provides a meaningful probabilistic interpretation in a Bayesian framework, which gives an efficient way to tune the unknown hyperparameters of the kernel function from data by maximizing the marginal likelihood. More specifically, based on the notation in (4.14),

the data-generating system can be written as

$$y(t) = \mathscr{L}_t[\mathtt{g}] + e(t), \tag{4.22}$$

where

A1 $e(t)$ is a zero-mean Gaussian noise process with variance $\sigma_{\mathtt{e}}^2$, independent of u.

A2 The impulse response $\mathtt{g}$ is modeled as a zero-mean GP on $\mathbb{N}$ with covariance K and being independent of $e(t)$.

Collect the noiseless outputs at time $t = 1, \dots, N$ in a vector $\chi = \begin{bmatrix}\mathscr{L}_1[\mathtt{g}] & \cdots & \mathscr{L}_N[\mathtt{g}]\end{bmatrix}$, which is a multivariate zero-mean Gaussian vector. As $\operatorname{cov}(\chi_i, \chi_j) = \mathscr{L}_i\big[\mathscr{L}_j[K]\big]$, the covariance of χ is K^{o} (4.17). From A1-A2, it can be concluded that $\mathtt{g}, Y_N$ are jointly Gaussian and the posterior $\mathtt{p}(\mathtt{g}(\cdot) \mid Y_N)$ is also Gaussian and its minimum variance estimate coincides with (4.16). The above discussion can be summarized in the following proposition, which is similar to Proposition 3.1.

Proposition 4.1 (Pillonetto et al. 2014) *Consider* (4.22) *under A1-A2 with prior knowledge reflected by the covariance function K. The minimum variance estimate of $g(\cdot)$ given the observations Y_N is $\hat{g}(\cdot)$ and is given by* (4.15) *with $\gamma = \sigma_{\mathtt{e}}^2$ and with $\mathscr{H}_K$ being the RKHS associated with K.*

As a result of such a connection, an estimate of the hyperparameters that parameterize the kernel function, i.e., β, is obtained by marginal likelihood optimization as follows:

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmax}}\ \log \mathtt{p}(Y_N \mid \beta),\\ &= \underset{\beta}{\operatorname{argmax}}\ -\frac{N}{2}\log(2\pi) - \frac{1}{2}Y_N^\top \left(\mathcal{K}_\beta^{\mathrm{o}} + \sigma_{\mathtt{e}}^2 I_{\mathrm{N}}\right)^{-1} Y_N - \frac{1}{2}\log\det(\mathcal{K}_\beta^{\mathrm{o}} + \sigma_{\mathtt{e}}^2 I_{\mathrm{N}}).\end{aligned} \tag{4.23}$$

4.1.3 RKHSs of impulse responses

Stable RKHSs

In order to have a successful identification, the RKHS, which is used as a hypothesis space for the estimation problem, needs to be appropriately designed. Having the one-to-one correspondence between an RKHS $\mathscr{H}_K$ and its reproducing kernel K, a kernel function K can be designed such that it encodes all relevant prior knowledge that will be automatically reflected on the resulting RKHS. The kernels used in the machine learning community for nonlinear function estimation include smoothness information, e.g., Gaussian and CS[1] kernels, but cannot be

[1] See (Pillonetto et al. 2014, Section 10.4) for more details on spline kernels.

applied directly for impulse response estimation as they do not include any stability constraint on the impulse response. Hence, the variance of the estimates is expected to be significant as the hypothesis space is unnecessary large. To show the role of the stability constraint in LTI system identification, consider the following example. Let us pick a system from the data set `S1D1` given in Section 4.2.5, with the same identification setting, where the main goal is to reconstruct the impulse response function from a set of noisy measurements. We adopt the estimator (4.15), which boils down to (4.19) in case of RFIR estimation. The kernel matrix, i.e., $\mathcal{K}$, is constructed based on CS and Gaussian kernels. The hyperparameters associated with these kernels are tuned with marginal likelihood maximization (4.23). Figure 4.1 shows the impulse response reconstructions, where it can be easily seen that both kernels do not perform well on the considered system. These kernels result in many oscillations in the estimates as both of them do not include information on the stability of the impulse response, i.e., it is inevitable decay to zero. This can be further understood by looking at the resulting kernel matrix

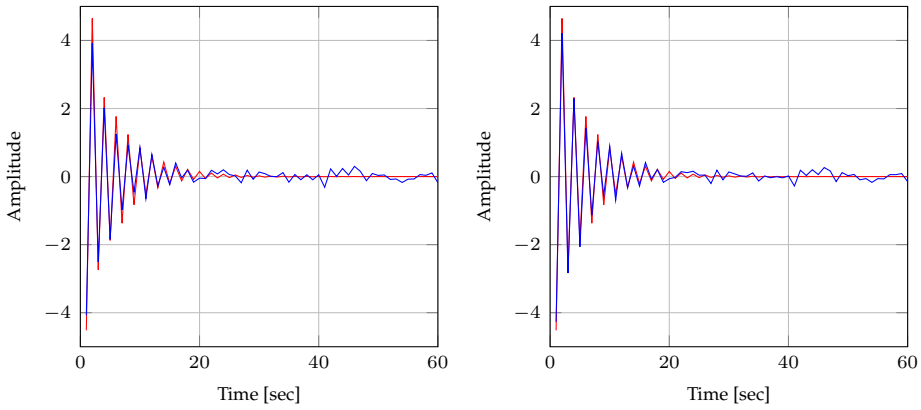


Figure 4.1: RFIR estimation with kernels that do not include a stability constraint. The true impulse response is given in red and the estimated response is given in blue. Left part: CS kernel. Right part: Gaussian kernel.

associated with these kernels. Figure 4.2 displays the scaled images of the kernel matrix constructed from CS, left part, and Gaussian kernel, right part. Because of the lack of information on impulse response stability, the diagonal elements of the kernel matrix do not decay to zero. Indeed, in case of a CS kernel, these elements increase and in case of a Gaussian kernel these elements have a steady state behavior. Such a behavior is not what we should expect from a stable impulse response. Moreover, the left part of Figure 4.3 shows the kernel sections associated with the CS kernel, i.e., columns of the kernel matrix, which are not decaying to zero. It has been discussed that the final estimate can be represented as a linear combination of these kernel sections, then the final estimate is also not going to decay to zero. The right part of Figure 4.3 shows randomly generated realizations from a GP with the CS kernel as its covariance, which are deviating from zero as time progresses. Therefore, a different prior, i.e., kernel function, should be employed that includes, not only information on smoothness, but also on BIBO-stability of

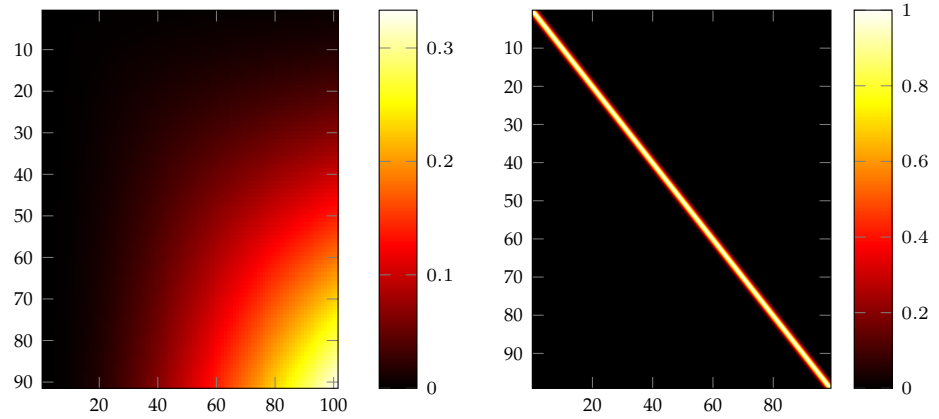


Figure 4.2: Scaled image of the kernel matrix $\mathcal{K}$ constructed with: Left part: CS kernel. Right part: Gaussian kernel. Note that the resulting image is an $m \times n$ grid of pixels where m and n are the number of columns and rows of $\mathcal{K}$, respectively. Each element of $\mathcal{K}$ specifies the color for a pixel of the image according to the color map shown on the right of each figure.

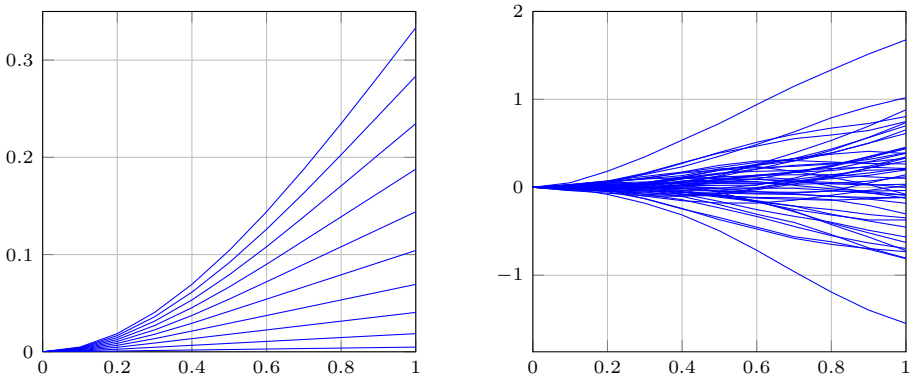


Figure 4.3: Left part: kernel sections of the CS kernel $K_{x_i}(\cdot)$ for $x_i = 0.1, \ldots, 1$. Right part: realization from a GP with the CS kernel as its covariance.

the to be estimated impulse response.

The necessary and sufficient condition for an LTI system $\mathcal{F}$ to be BIBO stable is that its impulse response g be in the Banach space $\mathcal{R}\ell_1(\mathbb{N})$, see Definition 2.1. Therefore, in a system identification scenario, the impulse response should be searched for in an RKHS contained in $\mathcal{R}\ell_1(\mathbb{N})$. Such RKHSs are known as stable RKHSs (Dinuzzo 2015; Pillonetto et al. 2014).

Definition 4.1 (Stable RKHSs) (Pillonetto et al. 2014) *Let $\mathscr{H}_K$ be the RKHS of real-valued functions on the domain $\mathscr{X} = \mathbb{N}$, induced by a kernel K. Then, $\mathscr{H}_K$ is said to be a stable RKHS, and the associated K is called stable, if $\mathscr{H}_K \subset \mathcal{R}\ell_1(\mathbb{N})$.*

Given a kernel K, it is often impossible to check the condition in Definition 4.1, as it is hard to understand which functions are contained in the associated RKHS. For instance, only recently and by means of sophisticated mathematical argument, an explicit characterization of the RKHS associated with the well-known Gaussian kernel has been obtained (Steinwart et al. 2006).

Instead, constructing K to directly guarantee stability of $\mathscr{H}_K$ is an easier task to accomplish. A necessary, but not sufficient condition for stability of K or $\mathscr{H}_K$ is that all kernel sections should be in $\mathcal{R}\ell_1(\mathbb{N})$, i.e., all $K_i(\cdot) \in \mathcal{R}\ell_1(\mathbb{N})$. Such a condition is not sufficient because $\mathscr{H}_K$ also contains all the Cauchy limits of linear combinations of kernel sections. For instance, the Gaussian kernel has stable kernel sections; however, it is not a stable kernel.

Now, let us introduce the space $\ell_\infty(\mathbb{N})$ of all bounded and real sequences defined on $\mathbb{N}$, i.e.,

$$\ell_\infty(\mathbb{N}) = \{h = \{h_i\}_{i=1}^{\infty}, \text{ such that } \|h\|_\infty < \infty\},$$

$$\|h\|_\infty = \sup_i |h_i|.$$

A necessary and sufficient condition for K to be associated with a stable $\mathscr{H}_K$ is summarized in the following theorem.

Theorem 4.2 (RKHSs stability) (Chen and Ljung 2015c) *Let $\mathscr{H}_K$ be the RKHS induced by $K : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$. It holds that*

$$\mathscr{H}_K \subset \mathcal{R}\ell_1(\mathbb{N}) \Leftrightarrow \sum_{i=1}^{\infty} \left| \sum_{j=1}^{\infty} u(j)K(i,j) \right| < \infty, \ \forall u \in \ell_\infty(\mathbb{N}). \tag{4.24}$$

Corollary 4.1 *Let $\mathscr{H}_K$ be the RKHS induced by K. Then,*

$$\mathscr{H}_K \subset \mathcal{R}\ell_1(\mathbb{N}) \Leftarrow \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |K(i,j)| < \infty. \tag{4.25}$$

In addition, considering only nonnegative-valued kernels denoted by K^+, i.e.,

$$K^+(i,j) \geq 0, \forall i,j \in \mathbb{N},$$

or diagonal kernels denoted by K^{d}, i.e.,

$$K^{\mathrm{d}}(i,j) = 0, \forall i \neq j,$$

condition (4.25) *becomes also necessary:*

$$\mathscr{H}_K \subset \mathcal{R}\ell_1(\mathbb{N}) \Leftrightarrow \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} K^{+}(i,j) < \infty. \tag{4.26}$$

and

$$\mathscr{H}_K \subset \mathcal{R}\ell_1(\mathbb{N}) \Leftrightarrow \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \left|K^{\mathrm{d}}(i,j)\right| < \infty. \tag{4.27}$$

Theorem 4.2 and Corollary 4.1 allow for the assessment of the stability or instability of an RKHS $\mathscr{H}_K$ given its reproducing kernel K without the need to characterize $\mathscr{H}_K$ itself. Next, we give an example of a kernel function, which is unstable, i.e., the corresponding $\mathscr{H}_K$ is not fully contained in $\mathcal{R}\ell_1(\mathbb{N})$.

Example 4.1 (RKHS of Gaussian kernel is not stable) *Consider a Gaussian kernel*

$$K(i,j) = \exp\left(-\frac{(i-j)^2}{\beta_{\mathrm{w}}}\right).$$

Note that K is a nonnegative-valued kernel. Such a kernel includes smoothness information, but by utilizing (4.26)*, it can be easily concluded that it is not a stable kernel and the resulting RKHS is not contained in $\mathcal{R}\ell_1(\mathbb{N})$, since*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \exp\left(-\frac{(i-j)^2}{\beta_{\mathrm{w}}}\right) = \infty.$$

Next, we focus on stable kernel functions that are suitable for impulse response estimation.

Kernel structures for impulse response estimation

For impulse response estimation, the kernel function K should reflect reasonable assumptions about the impulse response. See Figure 4.4 for various dynamic responses, which represent a possible prior knowledge that we are aiming to encode via the kernel function. For example, if the system is exponentially stable, the impulse response coefficients g_i should decay exponentially, which can be expressed by a stable kernel, and if the impulse response is smooth, neighboring values should have a positive correlation (Pillonetto et al. 2014). For this purpose, it is useful to recall that the optimal kernel of the estimation problem (4.15) is given by (Chen et al. 2012, Theorem 1):

$$K_{\mathrm{opt}}(i,j) = \mathrm{g}_i \mathrm{g}_j, \tag{4.28}$$

where $i, j \in \mathbb{N}$ and $\mathrm{g} = \{\mathrm{g}_i\}_{i=1}^{\infty}$ is the true impulse response. The term "optimal" means that such a kernel minimizes the MSE of the estimated impulse response.

More specifically, let $\hat{\mathrm{g}}$ denote the estimated impulse response, K is an arbitrary kernel function, then it holds that:

$$\text{MSE}(\hat{\mathrm{g}}) \text{ based on } K \;\geq\; \text{MSE}(\hat{\mathrm{g}}) \text{ based on } K_{\text{opt}}.$$

Even if (4.28) is impossible to be used in practice since the true impulse response is unknown, it provides a guideline to design suitable kernel functions for impulse response estimation. For instance, let the kernel mimic the behavior of the optimal kernel. Moreover, the prior knowledge of the true impulse response should be utilized in the design of the kernel function. The left part of Figure 4.5 shows a scaled image of the optimal kernel matrix, i.e., with $\mathcal{K}$ constructed with the kernel given in (4.28). Such an image gives an idea how the behavior of the kernel that describes the impulse response of a stable LTI system should look like. In the

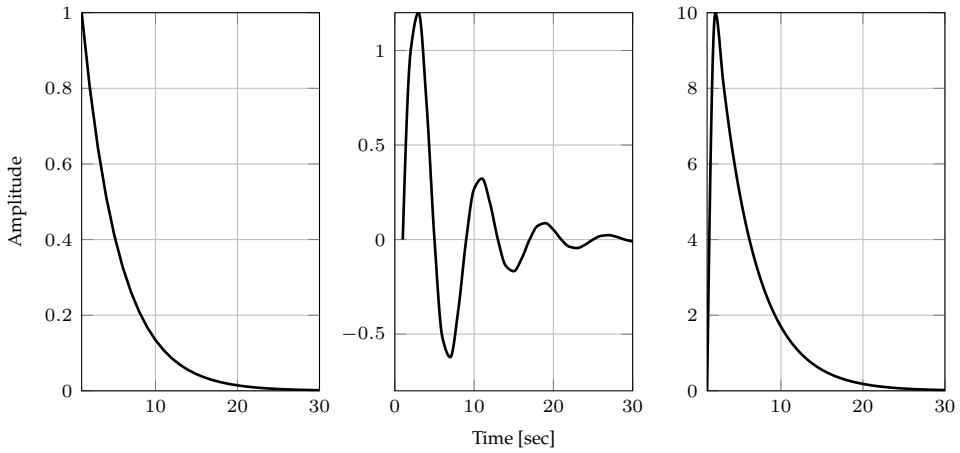


Figure 4.4: Common prior knowledge in impulse response estimation. Left part: stable over-damped. Middle part: stable under-damped. Right part: multiple, distinct time constants.

literature, many kernel structures have been introduced to embed various forms of prior knowledge or taken assumptions on the behavior/distribution of the expected impulse responses. In the following discussion, we give an overview of these kernel functions.

***Diagonal kernel* (DI)** (Chen et al. 2012):

$$K(i,j) = \begin{cases} \beta_1 \beta_2^i, \; i = j; \\ 0, \qquad \text{otherwise}; \end{cases} \quad , \; \beta_1 \geq 0, \; 0 \leq \beta_2 < 1, \tag{4.29}$$

where β_2 expresses the exponential decay rate. Note that the correlation between the impulse response entries at different time instants is not encoded in such a kernel.

***Stable Spline* (SS)** (Pillonetto and De Nicolao 2010):

$$K(i,j) = \begin{cases} \beta_1 \frac{\beta_2^{2i}}{2}\left(\beta_2^j - \frac{\beta_2^i}{3}\right), & i \geq j; \\ \beta_1 \frac{\beta_2^{2j}}{2}\left(\beta_2^i - \frac{\beta_2^j}{3}\right), & i < j; \end{cases}, \ \beta_1 \geq 0, \ 0 \leq \beta_2 < 1, \tag{4.30}$$

where β_2 expresses the exponential decay rate.

***Diagonal Correlated* (DC)** (Chen et al. 2012):

$$K(i,j) = \beta_1 \beta_2^{|i-j|} \beta_3^{\frac{i+j}{2}}, \ \beta_1 \geq 0, \ -1 < \beta_2 < 1, \ 0 \leq \beta_3 < 1, \tag{4.31}$$

where β_2 describes the correlation between impulse response entries at different time instants and β_3 accounts for the exponential decay rate.

***Tuned Correlated* (TC)** (Chen et al. 2012):

$$K(i,j) = \beta_1 \min\left(\beta_2^i, \beta_2^j\right), \ \beta_1 \geq 0, \ 0 \leq \beta_2 < 1, \tag{4.32}$$

where such a kernel is a special case of the DC kernel obtained by substituting $\beta_2 = \sqrt{\beta_3}$ in (4.31). It is also known as *first order stable spline kernel* (Pillonetto and De Nicolao 2010).

By including the stability constraint into the kernel function, the hypothesis space utilized in the estimation, i.e., the RKHS associated with the kernel, becomes a function space, where all elements are impulse responses that decay to zero. For example, for the SS kernel (4.30), a scaled image of the kernel matrix constructed with such a kernel is given in the right part of Figure 4.5, where it can be easily seen that it behaves very similar to the optimal situation, i.e., the left part of the same figure. Moreover, the left and right parts of Figure 4.6 show the kernel sections and a randomly generated realization of a GP associated with the SS kernel, respectively. By adopting the SS kernel for the example given in the previous subsection, a better result is obtained due to the inclusion of the stability constraint, see Figure 4.7, where the estimated impulse response is very close to the true one without the undesired oscillations that have been observed in Figure 4.1 when unstable kernels have been employed. There are many other kernel functions, e.g., the *Rank-1* kernel known also as the OE kernel (Chen et al. 2013), constructive state-space model induced kernels (Chen and Ljung 2014). Moreover, in (Chen and Ljung 2015c,b; Chen and Ljung 2016), two different methods of designing kernel functions suitable for impulse response estimation are presented from a machine learning perspective and system theory perspective. It is worth to mention that the above-mentioned kernels are considered to be single structure kernels, whereas multiple structure kernels have been introduced in (Chiuso et al. 2014; Chen et al. 2014), that handle systems with multiple and distinct time constants. Although, the above-mentioned kernels guarantee the stability of the estimated impulse response, there are other interesting dynamic properties that could be included besides stability, e.g., resonance behavior. In the following, we present an advanced kernel structure that can, with a simple parameterization, represent a wide range of dynamic properties in a systematic way.

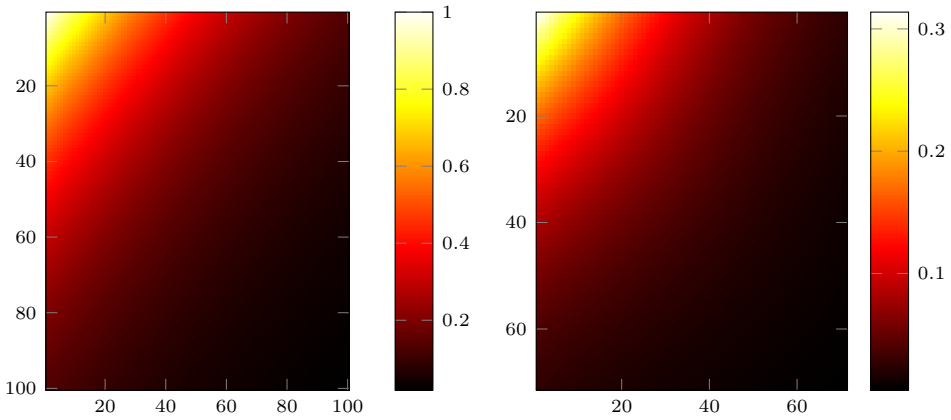


Figure 4.5: Scaled image of the kernel matrix constructed with: Left part: optimal kernel, i.e., $\mathtt{g}\mathtt{g}^\top$. Right part: SS kernel as an example of a stable kernel. Note that the resulting image is an $m \times n$ grid of pixels where m and n are the number of columns and rows of the kernel matrix, respectively. Each element of the kernel matrix specifies the color for a pixel of the image according to the color map shown on the right of each figure.

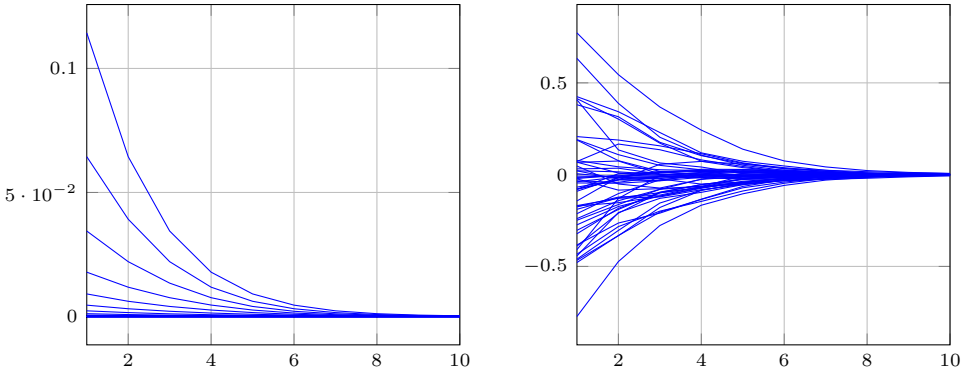


Figure 4.6: Left part: kernel sections of the SS kernel $K_{x_i}(\cdot)$ for $x_i = 0.1, \ldots, 1$. Right part: realization from a GP with the SS kernel as its covariance.

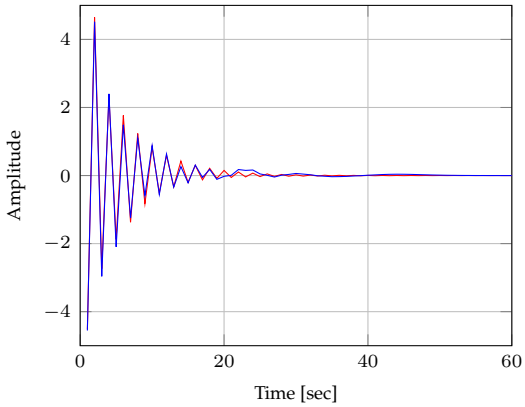


Figure 4.7: RFIR estimation with an SS kernel that includes the stability constraint. The true impulse response is given in red and the estimated is given in blue.

4.2 Bayesian identification with OBFs kernels

In this section, we give a systematic way to construct kernel functions for impulse response estimation based on OBFs that are capable of describing a wide range of dynamic properties and result in a well-designed RKHS and hence, improve the accuracy of the estimates by achieving a better bias/variance trade-off compared with existing kernels.

4.2.1 RKHS associated with OBFs in the time-domain

A fundamental result on RKHSs:

Proposition 4.2 (Unique kernel of an RKHS) (Aronszajn 1950) *Let $\mathscr{H}$ be a separable*[2] *Hilbert space of real-valued functions over $\mathscr{X}$ with orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$. Then,*

$$\mathscr{H} \text{ is an RKHS } \Leftrightarrow \sum_{i=1}^{\infty} |\phi_i(x)|^2 < \infty, \forall x \in \mathscr{X}.$$

The unique kernel K that is associated with $\mathscr{H}$ is

$$K(x, x') = \sum_{i=1}^{\infty} \phi_i(x)\phi_i(x'). \tag{4.33}$$

Consider $\mathcal{R}\ell_2(\mathbb{N})$ and its standard orthonormal basis, see Example 2.2. Using the above result, i.e., Proposition 4.2, it is immediate to conclude that $\mathcal{R}\ell_2(\mathbb{N})$ is an RKHS with a kernel given by the infinite-dimensional identity matrix, i.e., $K(i, j) = \delta_{ij}$. Interestingly, it has been discussed in Section 2.3.3 that the OBFs

[2] A Hilbert space is said to be separable if it has a basis with at most a countable number of elements.

$\Psi = \{\psi_k(t)\}_{k=1}^{\infty}$ constitute a complete orthonormal basis of $\mathcal{R}\ell_2(\mathbb{N})$. Accordingly, the simplest kernel that can be built using these OBFs is given by

$$K_{\Psi}(i,j) = \sum_{k=1}^{\infty} \psi_k(i)\psi_k(j), \tag{4.34}$$

which represents the formulation of the OBFs based kernel in time-domain and is a reproducing kernel for the RKHS space spanned by Ψ, i.e., $\mathcal{R}\ell_2(\mathbb{N})$.

The OBFs, with all of their variants, i.e., Takenaka-Malmquist, GOBFs, Laguerre and Kautz, are generated by a cascaded network of all-pass functions that are completely determined, modulo the sign, by their generating poles, which makes these basis a perfect candidate to represent the dynamic properties of LTI systems via the generating poles of the OBFs. Embedding such a representation capability into the regularization framework gives an attractive approach to be investigated for fulfilling Subgoal 1. Indeed, regularization techniques offer an attractive framework for estimation problems with a controlled bias/variance tradeoff and a systematic way to include prior knowledge via the kernel function. By combining such an attractive framework with the OBFs based kernels that are constructed from the OBFs generated by a set of poles, the prior knowledge of the dynamic properties of the unknown systems can be encoded.

4.2.2 OBFs kernels based IIR estimation

In the following, we assess the usefulness of the OBFs based kernel introduced in the previous subsection for estimating IIR models. More specifically, the presented class of kernels is assessed from two different points of view, i.e., system theory and machine learning perspectives.

System theory perspective

Now, we assess the OBFs based kernels from a system theory perspective, in the sense that we start from a system theoretic representation of LTI systems in terms of OBFs expansion and combine that with the optimal kernel (4.28).

From (2.29) and using (4.28), it follows that the optimal kernel in terms of the OBFs sequence $\Psi = \{\psi_k(t)\}_{k=1}^{\infty}$ is given by:

$$\begin{aligned} K_{\Psi}(i,j) &= \mathfrak{g}(i)\mathfrak{g}(j), && (4.35)\\ &= \sum_{k=1}^{\infty} c_k\psi_k(i) \sum_{l=1}^{\infty} c_l\psi_l(j), && (4.36)\\ &= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} c_k\psi_k(i)c_l\psi_l(j). && (4.37) \end{aligned}$$

In the Bayesian setting, $\mathfrak{g}$ is assumed to be a particular realization of a Gaussian random process. This corresponds to the assumption that $\{c_k\}_{k=1}^{\infty}$ is a sequence

of independent random variables with zero-mean and variance $\sigma_{\mathsf{c}}^2(k)$, i.e., $c_k \sim \mathscr{N}(0, \sigma_{\mathsf{c}}^2(k))$, then, by taking the expectation, we get

$$K_\Psi(i,j) = \mathcal{E}\left\{\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} c_k c_l \psi_k(i)\psi_l(j)\right\} \tag{4.38}$$

$$= \sum_{k=1}^{\infty} \sigma_{\mathsf{c}}^2(k)\psi_k(i)\psi_k(j). \tag{4.39}$$

It is well-known that the expansion coefficients $\{c_k\}_{k=1}^{\infty}$ satisfy $\sum_{k=1}^{\infty}|c_k|^2 < \infty$, i.e., $\{c_k\}_{k=1}^{\infty} \in \mathcal{R}\ell_2(\mathbb{N})$. Such a space is a rich space that contains the impulse responses of possibly infinite-dimensional and time-varying systems. However, we are interested only in FD-LTI systems with impulse responses belonging to $\mathcal{R}\ell_1(\mathbb{N})$, and hence the expansion coefficients must satisfy a more restrictive condition $\sum_{k=1}^{\infty}|c_k| < \infty$. One possible way to impose such a behavior in the kernel definition is to characterize the decay of the expansion coefficients by an exponential term. This will become more clear in the next subsection.

Machine learning perspective

In this section, we analyze the OBFs based kernel from a machine learning perspective. More specifically, we assess the suitability of the RKHSs constructed from OBFs for impulse response estimation.

Given a sequence of OBFs $\Psi = \{\psi_k(\cdot)\}_{k=1}^{\infty}$, it is shown that Ψ span an RKHS with the reproducing kernel given by (4.34). Since Ψ contains an orthonormal basis in $\mathcal{R}\ell_2(\mathbb{N})$, from Proposition 4.2, it can be easily seen that $K_\Psi(i,j) = \delta_{ij}$. If the system to be identified is stable, this kernel will perform poorly (this coincides with the conclusion in (Chen and Ljung 2015a, Section V)): in fact, the optimal structure (4.28) suggests that the diagonal elements of the kernel should decay to zero, instead of being constant. In addition, the off-diagonal elements should be different from zero. Moreover, the Bayesian interpretation of regularization, as described, e.g., in (Pillonetto et al. 2014, subsection 4.3), also supports the same conclusions from a Bayesian perspective. The estimator (4.15) can in fact be seen as the minimum variance estimator of the impulse response when the latter, i.e., the impulse response, is a zero-mean Gaussian process, independent of the noise, with covariance proportional to K. When (4.34) is adopted, g becomes proportional to a stationary white noise. But the variability of a stable impulse response is expected to decay to zero as time progresses. One problem related to (4.34) is that it defines a kernel which is not stable according to Definition 4.1. Indeed, the kernel defined by the OBFs Ψ, i.e., K_Ψ, leads to an RKHS as a hypothesis space given by $\mathcal{R}\ell_2(\mathbb{N})$. However, $\mathcal{R}\ell_2(\mathbb{N}) \not\subset \mathcal{R}\ell_1(\mathbb{N})$, and hence the kernel is not stable.

In view of the above results from both machine learning and system theory perspectives, to include the stability constraint, we consider the kernel construction

$$K_\Psi^{\mathrm{s}}(i,j) = \beta_\alpha \sum_{k=1}^{\infty} \mathfrak{D}_k(\beta_{\mathrm{d}})\,\psi_k(i)\psi_k(j), \tag{4.40}$$

where $\mathfrak{D}_k(\beta_{\rm d})$ is a decay term that weighs the OBFs and converges to zero as $k \to \infty$, $\beta_{\rm d}$ is considered to be a hyperparameter that is responsible for magnitude scaling of the basis $\{\psi_k\}_{k=1}^{\infty}$ associated with (4.40). The decay term, i.e., $\mathfrak{D}_k(\beta_{\rm d})$, with $\beta_{\rm d}$ tuned by marginal likelihood optimization also acts as an automatic way to select the number of significant basis functions that are needed to construct the kernel. In absence of more sophisticated prior information, as in the case of many practical scenarios, monotonically decreasing weights, e.g.,

$$\mathfrak{D}_k(\beta_{\rm d}) = k^{-\beta_{\rm d}}, \;\; \beta_{\rm d} > 0, \tag{4.41}$$

or

$$\mathfrak{D}_k(\beta_{\rm d}) = \beta_{\rm d}^{-k}, \;\; \beta_{\rm d} > 1, \tag{4.42}$$

are effective and able to well guard against ill-conditioning of the system identification problem. This is also supported from the view point of system theory, where it is known that the decay rate of the expansion can be always upper bounded by an exponential decay. However, depending on the available knowledge, other parameters can be introduced in the decay term that describe more complicated shapes for the weights. Similarly, when a prior information is available, this can support the choice of the basis functions. This fits in the framework developed in this chapter, e.g., if the number of resonance peaks is known, we can use such information to decide the number of complex pairs/real poles that should be considered for GOBFs generating $K_{\Psi}^{\rm s}$. The other hyperparameters, besides $\beta_{\rm d}$, are the scale factor β_{α} and the poles used to generate the sequence $\psi_k(\cdot)$ collected in a vector $\beta_{\rm p}$.

The following proposition provides guarantees on the stability of the kernel constructed using the more general Takenaka-Malmquist OBFs. Note that, we prove the stability under a general class of OBFs and hence the results hold for the special cases, e.g., GOBFs, Laguerre and Kautz basis.

Proposition 4.3 (Stability of the OBFs based kernels) *Consider the kernel* (4.40)*, which is built using the general OBFs basis, with all generating poles assumed to be uniformly away from the unit circle, i.e.,* $\exists \varsigma > 0$ *s.t.* $\lambda_i \leq \varsigma < 1$ *for all* i*. Then, the kernel is stable if* $\mathfrak{D}_k(\beta_{\rm d}) = k^{-\beta_{\rm d}}$ *and* $\beta_{\rm d} > 3$ *or* $\mathfrak{D}_k(\beta_{\rm d}) = \beta_{\rm d}^{-k}$ *and* $\beta_{\rm d} > 1$*.*

Proof: See Appendix A.2. □

This allows to introduce an identification scheme for regularized impulse response estimation with the OBFs based kernel (4.40) as summarized in Algorithm 2.

4.2.3 Regularized OBFs expansion estimation

Overview

Instead of using OBFs for kernel construction, in Chen and Ljung (2015a), a regularization based estimation of the OBFs expansion (ROBFs) has been investigated. More specifically, consider the data-generating system (4.1), and let the transfer

Algorithm 2 Regularized IIR estimation with OBFs based kernel (4.40).

Require: A data record $\mathcal{D}_N = \{u(t), y(t)\}_{t=1}^{N}$.

1: Estimate the noise variance σ_{e}^2 with a low bias and high order ARX or FIR model. The estimate is denoted as $\hat{\sigma}_{\mathsf{e}}^2$.
2: **Hyperparameter tuning:** Solve (4.23) to get the empirical Bayes estimate $\hat{\beta}$ for $\beta = \left[\beta_\alpha \ \beta_{\mathrm{d}} \ \beta_{\mathrm{p}}^\top\right]^\top$.
3: **Function estimation:** Impulse response estimation: with $\beta = \hat{\beta}$ and $\gamma = \hat{\sigma_{\mathsf{e}}}^2$ compute the estimate of the impulse response via (4.20).
4: **Return:** Estimated impulse response $\hat{\theta}$.

operator $G_0(q)$ be represented by a series expansion representation in terms of the OBFs $\{\psi_i(t)\}_{i=1}^{\infty}$ or equivalently, the frequency domain representation related operator form $\{\breve{\psi}_i(q)\}_{i=1}^{\infty}$, see Section 2.4.3. Furthermore, by considering the first n_ψ terms in the expansion, i.e., $\{\psi_i(t)\}_{i=1}^{n_\psi}$, (4.1) can be written as

$$y(t) = \sum_{i=1}^{n_\psi} c_i(\psi_i \circledast u)(t) + v(t) \tag{4.43}$$

$$= \sum_{i=1}^{n_\psi} c_i \breve{\psi}_i(q)u(t) + v(t). \tag{4.44}$$

The model structure given by (4.43), for a given basis set, renders the identification problem a linear regression problem for estimating the expansion coefficients $\{c_i\}_{i=1}^{n_\psi}$, see Section 2.4.4 for more details. However, by extending the ReLS approach for FIR, see (4.19), to the OBFs model structure, we gain the following:

- The generating poles of the OBFs utilized in the model structure can be treated as a hyperparameters , which can be estimated by the empirical Bayes method, i.e., maximizing the marginal likelihood.
- A bias/variance trade-off can be demonstrated, which is exploited to reduce the MSE of the final model estimate.

Now, we estimate the expansion coefficients $c = \left[c_1 \cdots c_{n_\psi}\right]^\top$ by minimizing the following ReLS criterion:

$$\hat{c} = \underset{c\in\mathbb{R}^{n_\psi}}{\operatorname{argmin}} \ \|Y_N - \Upsilon_N(\beta_{\mathrm{p}})c\|_2^2 + \gamma c^\top \mathcal{K}_c^{-1}(\beta_{\mathrm{c}})c,$$

$$= \underset{c\in\mathbb{R}^{n_\psi}}{\operatorname{argmin}} \ \sum_{t=1}^{N}\left(y(t) - \sum_{k=1}^{n_\psi} c_k(\psi_k \circledast u)(t)\right)^2 + \gamma c^\top \mathcal{K}_c^{-1}(\beta_{\mathrm{c}})c, \tag{4.45}$$

where $\mathcal{K}_c$ is the regularization matrix on the coefficients $\{c_i\}_{i=1}^{n_\psi}$ and Υ_N is the regression matrix which can be constructed as follows:

$$\gamma_{\mathrm{r}}^\top(t) = \left[\breve{\psi}_1(q)u(t) \ \cdots \ \breve{\psi}_{n_\psi}(q)u(t)\right]$$

$$\Upsilon_N = [\gamma_{\mathrm{r}}(1) \ \cdots \ \gamma_{\mathrm{r}}(N)]^\top .$$

Now, let us discuss the design of the suitable kernel structure that encodes the prior knowledge about the expansion coefficients. Actually, the behavior of the expansion coefficients depends largely on the utilized basis functions. For instance, if we consider a Laguerre basis, it has been suggested in Wahlberg (1991) to assume that $c \in \mathcal{R}\ell_1(\mathbb{N})$, i.e.,

$$\sum_{i=1}^{\infty} |c_i| < \infty.$$

In this case, the expansion coefficients can be regarded as the impulse response of a stable LTI system and accordingly we can use any of the kernel structures presented in Section 4.1.3 for regularized FIR estimation to regularize the expansion coefficients. For sophisticated basis, e.g., general Takenaka-Malmquist basis, the behavior of the expansion coefficients will accordingly be more involved. However, upon the availability of more prior knowledge, this can be embedded in the kernel structure. For example, as suggested in Chen and Ljung (2015a), an adapted version of the DC kernel can be used to represent the slower convergence rate in the Laguerre model when the dominant poles are very close to the unit circle:

$$K_c(i,j) = \beta_1 \beta_2^{|i-j|} \mathcal{D}(i+j), \tag{4.46}$$

where $\mathcal{D}(\cdot)$ is a nonnegative function that decays slower than the exponential function of the DC kernel in (4.31). Note that, we denote the regularization matrix by $\mathcal{K}_c(\beta_{\mathrm{c}})$ to indicate that the kernel function depends on the hyperparameter vector β_{c}, which describes the behavior of the expansion coefficients, e.g., β_1, β_2 in (4.46). In such cases and in the absence of detailed prior knowledge, we can still use the kernel recommended as before.

In (4.45), we denote the regression matrix as $\Upsilon_N(\beta_{\mathrm{p}})$ to indicate that it depends on β_{p}, which is a vector containing the unknown generating poles. These poles along with the hyperparameter that controls the decay rate of the expansion coefficients, i.e., β_{c}, can be estimated by maximizing the marginal likelihood.

Connection to regularized FIR estimation with unstable OBFs kernels

It has been shown in (Chen and Ljung 2015a, Section V) that the regularized FIR estimation with the OBFs based kernel K_Ψ (4.34), is a special ill-defined case of the ROBFs estimation problem (4.45) pointing out that using OBFs in defining a kernel function in the time-domain in a naive way is not advisable. More specifically, consider (4.15), which can be written in case of FIR as

$$\hat{\mathrm{g}} = \underset{\mathrm{g} \in \mathscr{H}}{\operatorname{argmin}} \ \sum_{i=1}^{N} (y_i - (\mathrm{g} \circledast u)(t_i))^2 + \gamma \|\mathrm{g}\|_K^2. \tag{4.47}$$

Furthermore, by utilizing the OBFs kernel K_Ψ in (4.34), we can define the corresponding RKHS $\mathscr{H}_{K_\Psi}$ as:

$$\begin{aligned}\mathscr{H}_{K_\Psi} &= \mathrm{Span}\{\psi_1, \psi_2, \ldots, \psi_{n_\psi}\} \\ &= \{\mathrm{g} \mid \mathrm{g}(t) = \sum_{k=1}^{n_\psi} c_k \psi_k(t), c_k \in \mathbb{R}\},\end{aligned} \tag{4.48}$$

with

$$\|\mathrm{g}\|_K^2 = \sum_{i=1}^{n_\psi} c_i^2.$$

Now, substituting $\|\mathrm{g}\|_K^2 = \sum_{i=1}^{n_\psi} c_i^2$ and $\mathrm{g}(t) = \sum_{i=1}^{n_\psi} c_i \psi_i(t)$ into (4.47), will turn the estimation problem to estimate the expansion coefficients c instead of g

$$\hat{c} = \underset{c \in \mathbb{R}^{n_\psi}}{\mathrm{argmin}} \sum_{t=1}^{N} \left(y(t) - \sum_{k=1}^{n_\psi} c_k (\psi_k \circledast u)(t) \right)^2 + \gamma c^\top I_{\mathrm{n}_\psi} c, \tag{4.49}$$

where the regularizer $c^\top c = \|c\|_2^2$ corresponds to a Ridge regression of c. However, the resulting Ridge regression indicates that the regularization matrix is the identity matrix, which in our case does not reflect the prior knowledge that the expansion coefficients should be absolutely summable and also this will not guarantee the stability of the resulting impulse response estimate. This is completely in agreement with the conclusion that has been drawn in Section 4.2.1 that the kernel defined in (4.48) will not work for impulse response estimation as it does not encode the stability constraints.

Connection to regularized FIR estimation with stable OBFs kernels

In the following, the connection of the stable OBFs based kernel K_Ψ^{s} defined in (4.40) with the ROBFs approach is shown. By considering an $n_\psi-$truncated kernel representation[3] of K_Ψ^{s}, i.e., $K_\Psi^{\mathrm{s}}(i,j) = \beta_\alpha \sum_{k=1}^{n_\psi} \mathfrak{D}_k(\beta_{\mathrm{d}}) \psi_k(i) \psi_k(j)$, and letting $\mathfrak{D}_k(\beta_{\mathrm{d}}) = d_k$ for $k = 1, \ldots, n_\psi$, the associated RKHS $\mathscr{H}_{K_\Psi^{\mathrm{s}}}$ can be written as

$$\begin{aligned}\mathscr{H}_{K_\Psi^{\mathrm{s}}} &= \mathrm{Span}\{\psi_1, \psi_2, \ldots, \psi_{n_\psi}\} \\ &= \{\mathrm{g} \mid \mathrm{g}(t) = \sum_{k=1}^{n_\psi} c_k \psi_k(t), c_k \in \mathbb{R}\},\end{aligned} \tag{4.50}$$

with

$$c_k = \langle \mathrm{g}, \psi \rangle_{K_{\Psi_k}^{\mathrm{s}}},$$

[3] If we consider that $\hat{\mathrm{g}}$ to be the estimate with the infinite kernel representation, i.e., $n_\psi = \infty$ and $\hat{\mathrm{g}}_{n_\psi}$ is the estimate with the $n_\psi-$truncated representation of the kernel, then, the following result holds $\lim_{n_\psi \to \infty} \|\hat{\mathrm{g}} - \hat{\mathrm{g}}_{n_\psi}\|_{K_\Psi^{\mathrm{s}}} = 0$, see (Pillonetto and Bell 2007, Theorem 7).

and

$$\|\mathrm{g}\|^2_{K^{\mathrm{s}}_{\Psi}} = \sum_{k=1}^{n_\psi} c_k^2/\mathrm{d}_k = c^\top \mathcal{K}_c^{-1} c$$

where $c = [c_1 \ \cdots \ c_{n_\psi}]^\top$ and $[\mathcal{K}_c]_{ij} = \mathrm{d}_j \delta_{ij}$.

As $\mathrm{g}(\cdot) = \sum_{k=1}^{n_\psi} c_k \psi_k(\cdot)$ and $\|\mathrm{g}\|^2_{K^{\mathrm{s}}_{\Psi}} = c^\top \mathcal{K}_c^{-1} c$, (4.47) can be written as

$$\hat{c} = \underset{c \in \mathbb{R}^{n_\psi}}{\operatorname{argmin}} \sum_{t=1}^{N} \left(y(t) - \sum_{k=1}^{n_\psi} c_k (\psi_k \circledast u)(t) \right)^2 + \gamma c^\top \mathcal{K}_c^{-1} c, \tag{4.51}$$

This gives that (4.51) is identical with (4.45). Although they are identical from the optimization point of view, conceptually they are different. The approach in Chen and Ljung (2015a) utilizes the Bayesian approach to regularize the estimation of the expansion coefficients. On the other hand, the approach presented in this chapter uses the OBFs to construct a kernel function that results in a stable RKHS directly in the time-domain, which can be used for impulse response estimation and provides a better understanding of that space. Both approaches consider the generating poles as hyperparameters and tune them with the marginal likelihood maximization.

4.2.4 Hyperparameter tuning and computational complexity

In case of the OBFs based kernel defined in (4.40), the hyperparameters that are needed to be estimated from data are the scaling parameter β_α, the decay parameter β_{d} and the generating poles. Note that in case of the Laguerre-based kernel, only one real pole, i.e., λ in (2.24), is needed to generate the full sequence of basis. For the Kautz-based kernel, two conjugate complex poles defined by $\mathfrak{b}$ and $\mathfrak{c}$ in (2.23) are required to generate that sequence. Hence, the estimation of these hyperparameters following the empirical Bayes approach can be accomplished by solving the optimization problem (4.23).

In regularized impulse response estimation, the overall algorithm mainly consists of two steps (Chen and Ljung 2013):

1. Hyperparameters estimation: This step involves the minimization of a cost function (4.23) for which a single evaluation for the cost function is $\mathcal{O}(N^3)$.
2. Impulse response estimation: The computational complexity of this step is $\mathcal{O}(N^3)$.

In (Carli et al. 2012), a new computational strategy has been proposed which may reduce significantly the computational load and extend the practical applicability of this methodology to large-scale scenarios. The proposed algorithm (Carli et al. 2012, Algorithm 2) is mainly developed for SS kernels and exploits the spectral decomposition of these kernels (Pillonetto and De Nicolao 2010). With this approach, the computational complexity now scales as $\mathcal{O}(n_\psi^3)$, where n_ψ is the number of the

eigenfunctions. Moreover, it can effectively compute the marginal likelihood with $\mathcal{O}(N^2 n_\psi)$ for a single evaluation of the cost, see (Carli et al. 2012, Table 1). This algorithm is directly applicable for kernels that have a spectral decomposition, like the OBFs based kernel. Moreover, the effectiveness of this algorithm depends on the impulse response to be estimated, which is not known a priori, and if it can be approximated with a few number of eigenfunctions (Carli et al. 2012, Page 5). This motivates also the use of OBFs as eigenfunctions, offering a wide range of basis selection which if properly chosen, a few number of basis is needed to get a high approximation accuracy. In Kondo et al. (2017), a new hyperparameters estimation algorithm is presented for the regularized least squares problem in the empirical Bayesian approach arising from FIR model identification, which is purposed for OBFs based kernel. Such an algorithm consists of two steps. More specifically, first divide the decision variables into two groups, namely the variables associated with the decay term and the generating poles of the utilized OBFs. Then, alternately minimizing with respect to each group. It is shown that *difference of convex functions* programming is effectively applicable in the algorithm because the search space is shown to be bounded.

4.2.5 Numerical simulation

In this section, the performance of the presented OBFs based kernels in the considered Bayesian identification setting is assessed on *Monte Carlo* (MC) based simulation studies using randomly generated discrete-time LTI systems.

Simulation studies

By using the setting of (4.1) as the data-generating system, five simulation studies have been accomplished for the following scenarios:

1. `S1D1`: fast systems, $\mathcal{D}_N$ with $N = 500$, *Signal-to-Noise Ratio* (SNR)=10dB.
2. `S1D2`: fast systems, $\mathcal{D}_N$ with $N = 375$, SNR=1dB.
3. `S2D1`: slow systems, $\mathcal{D}_N$ with $N = 500$, SNR=10dB.
4. `S2D2`: slow systems, $\mathcal{D}_N$ with $N = 375$, SNR=1dB.
5. `S3`: oscillatory systems, $\mathcal{D}_N$ with $N = 400$, SNR=10dB.

Each Scenario 1) to 4) corresponds to 100 randomly generated (by the `drss` Matlab function) 30-th order discrete-time LTI systems for G_0. The fast systems have all poles inside $0.95\mathbb{J}$ and the slow systems have at least one pole in the ring $\mathbb{J} - 0.95\mathbb{J}$, i.e., slow dominant poles. These systems are used to generate data sets for a white u, with $u \sim \mathcal{N}(0, 1)$ and v being additive white Gaussian noise. The variance of v is set such that the SNR $= 1$ or 10dB for various Monte Carlo experiments, where SNR is defined as

$$10 \log_{10} \left(\frac{\sum_{k=1}^{N} \breve{y}^2(k)}{\sum_{k=1}^{N} v^2(k)} \right), \tag{4.52}$$

where $\breve{y}(k)$ denotes the noise-free system output, i.e., $\breve{y}(k) = G_0(q)u(k)$. Whereas, Scenario 5) has been generated as reported in (Chiuso et al. 2014, Section VI), but with only one dominant complex conjugate pole pair.

Identification setting

In all of the five scenarios, we estimate FIR models, i.e., the $n-$truncated impulse response of (4.2) or equivalently θ in (4.19), with $n = 125$ and with the following estimators:

1. RFIR-TC: regularized impulse response estimation, where the impulse response coefficients are estimated by solving (4.15) and regularized with the TC kernel (4.32).
2. RFIR-OBFs-L, -K or -G: regularized impulse response estimation, where the impulse response coefficients are estimated by solving (4.15) and regularized with the OBFs based kernel (4.40) constructed with three different classes of basis functions; Laguerre with one real pole, Kautz with one complex conjugate pair or GOBFs with two real poles where the generating inner function is a 2^{nd} order one, respectively.
3. ROBFs-L, -K or -G: regularized OBFs expansion estimation, where the expansion coefficients are estimated by solving (4.45) and regularized with the diagonal kernel (4.29), i.e., DI kernel, with three different classes of basis functions used in the expansion; Laguerre with one real pole, Kautz with one complex conjugate pair or GOBFs with two real poles, respectively.

Furthermore, two different scenarios are considered for the number of basis functions to construct the OBFs based kernel and the OBFs model structure, i.e., 40 basis and 100 basis. This is essential to show the effectiveness of the presented approach to control the flexibility offered by employing large number of basis. The performance index that is used to measure the quality of the impulse response estimation with different estimators is BFR of the estimated impulse response $\hat{\mathrm{g}}_k$

$$\text{BFR} = 100\%.\left(1 - \sqrt{\frac{\sum_{k=1}^{125}|\mathrm{g}_k - \hat{\mathrm{g}}_k|^2}{\sum_{k=1}^{125}|\mathrm{g}_k - \bar{\mathrm{g}}|^2}}\right),\ \ \bar{\mathrm{g}} = \frac{1}{125}\sum_{k=1}^{125}\mathrm{g}_k, \tag{4.53}$$

where, $\{\mathrm{g}_k\}$ are the true coefficients values. The hyperparameters have been estimated by solving (4.23). Note that, the approach proposed in Chen and Ljung (2013), i.e., QR factorization, is employed in this example to solve the optimization problem to tune the unknown hyperparameters and to estimate the unknown impulse response.

Identification results

The average model fits over the considered five data sets estimated with TC kernel are reported in Table 4.1, whereas the average model fits in case of (RFIR-OBFs)/ROBFs-L, (RFIR-OBFs)/ROBFs-K and (RFIR-OBFs)/ROBFs-G are reported

in Tables 4.2, 4.3 and 4.4, respectively. Moreover, for each case, the results are given with both 40 basis and 100 basis. The highest average model fit over the RFIR-OBFs alternatives and the ROBFs alternatives are highlighted in bold. It is worth to mention that in case of RFIR-OBFs estimators both decay terms, i.e., (4.41) and (4.42), are implemented and the one that gives the highest fit is shown in the results as they performed the same on average. For illustration, the distributions

Table 4.1: Average of the BFR of the estimated FIR models with TC kernel.

RFIR-TC	S1D1	S1D2	S2D1	S2D2	S3
	90.82	77.25	84.08	63.61	86.01

Table 4.2: Average of the BFR of the estimated FIR models with Laguerre basis.

RFIR-OBFs-L	S1D1	S1D2	S2D1	S2D2	S3
40 basis	91.85	78.97	85.71	67.88	85.31
100 basis	91.90	79.20	87.86	68.88	87.67
ROBFs-L	S1D1	S1D2	S2D1	S2D2	S3
40 basis	91.90	78.92	85.70	69.28	83.41
100 basis	91.90	79.18	88.13	69.73	88.93

Table 4.3: Average of the BFR of the estimated FIR models with Kautz basis.

RFIR-OBFs-K	S1D1	S1D2	S2D1	S2D2	S3
40 basis	91.91	79.08	87.29	70.35	93.40
100 basis	91.92	78.99	88.44	70.83	**93.70**
ROBFs-K	S1D1	S1D2	S2D1	S2D2	S3
40 basis	91.89	78.64	87.50	70.52	**93.72**
100 basis	91.98	78.93	88.74	71.33	93.64

of the model fits over the five data sets with TC and the estimates with RFIR-OBFs, ROBFs, which are highlighted in bold, are shown by boxplots in Fig. 4.8 to 4.10.

Discussion

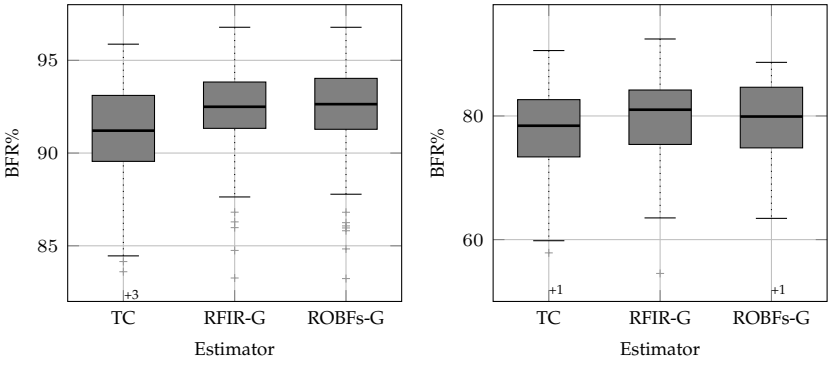
Next, we give some insights that can be obtained from the results.

1. In general, RFIR-OBFs with all its alternatives perform better than RFIR-TC, because RFIR-OBFs estimators employ kernels that inherently describe dynamic properties, e.g., resonance behavior, damping, etc., via the generating poles of the OBFs, rather than only focusing on smoothness and stability.

2. For the resonating system, i.e., S3: RFIR-OBFs-L/ROBFs-L have difficulties to deal with such systems, which is well-known that for a system with resonance behavior, a long Laguerre expansion is needed to get good accuracy.

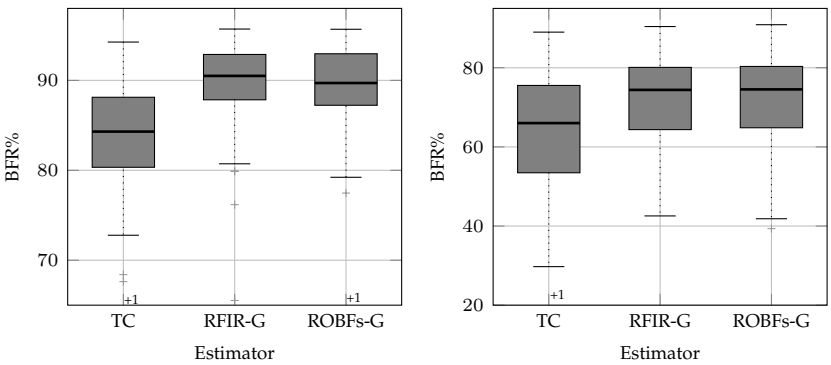
Table 4.4: Average of the BFR of the estimated FIR models with GOBFs basis.

RFIR-OBFs-G	S1D1	S1D2	S2D1	S2D2	S3
40 basis	92.07	**79.54**	87.74	69.82	83.53
100 basis	**92.21**	79.52	**89.26**	**71.10**	88.76
ROBFs-G	S1D1	S1D2	S2D1	S2D2	S3
40 basis	92.18	**79.27**	87.37	71.12	83.40
100 basis	**92.31**	79.18	**89.42**	**72.38**	89.20



(a) S1D1 (b) S1D2

Figure 4.8: Boxplot for model fits over S1D1, S1D2. The shown estimators are those highlighted in bold in the tables. Note that TC and RFIR-G denote RFIR-TC and RFIR-OBFs-G, respectively.



(a) S2D1 (b) S2D2

Figure 4.9: Boxplot for model fits over S2D1, S2D2. The shown estimators are those highlighted in bold in the tables. Note that TC and RFIR-G denote RFIR-TC and RFIR-OBFs-G, respectively.

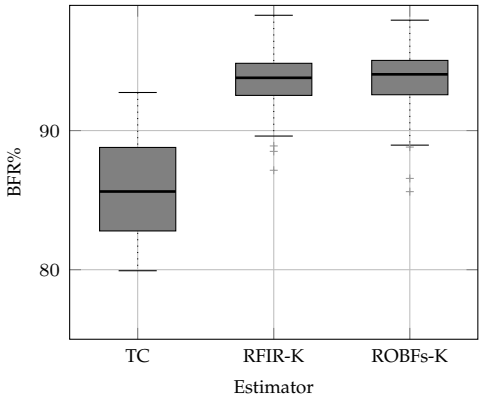


Figure 4.10: Boxplot for models fit over S3. The shown estimators are those highlighted in bold in the tables. Note that TC and RFIR-K denote RFIR-TC and RFIR-OBFs-K, respectively.

This can be easily seen from the poor performance in case of 40 basis compared with TC kernel. However, when increasing the number of basis to 100, the results improve a lot due to the employed long expansion and on top of that the regularization that reduces the variance.

3. In case of the resonance behaviour in systems in S3, Kautz basis perform significantly better compared to other estimators. This is due to the fact that the Kautz basis is generated by two repeated conjugate complex poles that are tuned by marginal likelihood optimization, which is proven to be a numerically robust approach to accomplish this.

4. Due to the regularization acting on the estimation problem, in most of the cases we gain from increasing the number of basis functions, which is not the case in classical identification due to the increased variance resulting from increasing the number of the expansion coefficients.

5. For slow systems, i.e., S2D1 and S2D2: RFIR-OBFs estimators show a significant improvement over the TC kernel, especially RFIR-OBFs-G, which gives the best performance for the first four data sets, i.e., S1D1, ..., S2D2. This is due to the fact that, if the basis functions are properly chosen, the OBFs offer a more compact model structure which results in a better RKHS as a hypothesis space.

6. The results of RFIR-OBFs and ROBFs are pretty close to each other due to the conceptual equivalence provided in Section 4.2.3. Note that the difference in the results is due to numerical issues, i.e., in case of RFIR-OBFs we directly estimate a 125 length impulse response, whereas in case of ROBFs we only estimate as many expansion coefficients as the number of basis functions and then reconstruct a 125 length impulse response of the estimated OBFs model.

4.3 Bayesian frequency domain identification with OBFs based kernels

In this section, we look at the problem of estimating the FRF of a stable LTI system $\mathcal{F}$, i.e., $G_0(e^{\mathrm{j}\omega})$, directly in the frequency-domain including the handling of the transient effect within the Bayesian setting. To achieve this, we will also use the formulation of the presented OBFs based kernels in the frequency-domain.

4.3.1 Problem statement

Consider a SISO DT-FD-LTI stable data-generating system, which is given in (4.1) with its impulse response denoted by $\mathrm{g} = \{\mathrm{g}(k)\}_{k=1}^{\infty}$. Assume that, we measure $y(t)$ at $t = 0, 1, \ldots, N-1$, i.e., $\mathcal{D}_N = \{u(t), y(t)\}_{t=0}^{N-1}$. Denote by $\Omega = e^{\mathrm{j}\omega}$ the frequency variable for $\omega \in \mathbb{R}$, then Ω_k for $k \in \mathbb{Z}$ is defined as

$$\Omega_k = e^{\mathrm{j}\omega_k} = e^{\frac{\mathrm{j}2\pi k}{N}}. \tag{4.54}$$

It is worth to mention that for $k \in \mathbb{Z}$, Ω_k corresponds to the k-th bin of an N-point DFT, which is defined below.

Definition 4.2 (N-point DFT) *The N-point DFT, at frequency bin k, of a sampled signal $x_{\mathrm{s}}(\tau), \tau = 0, 1, \ldots, N-1$ is given by*

$$X_{\mathrm{s}}(k) = \frac{1}{\sqrt{N}} \sum_{\tau=0}^{N-1} x_{\mathrm{s}}(\tau) e^{\frac{-2\mathrm{j}\pi k\tau}{N}}, \; k \in \mathbb{Z}. \tag{4.55}$$

Accordingly, denote $U_{\mathrm{s}}(k), Y_{\mathrm{s}}(k), V_{\mathrm{s}}(k)$ the N-point DFT at frequency bin k of $u(t)$, $y(t)$, $v(t)$, respectively.

In order to get the frequency-domain equivalent of (4.1), we first introduce the following remark.

Remark 4.2 *For a discrete-time, windowed signal $u(t)$, i.e., $u(t) = 0$ for $t < 0$ and $t \geq N$, the* Discrete-Time Fourier Transform *(DTFT) $U_{\mathrm{s}}(e^{\mathrm{j}\omega_k}) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} u(t) e^{-\mathrm{j}\omega_k t}, k \in \{0, \ldots, N-1\}$, is equal to its N-point DFT.*

Moreover, the DTFT of the impulse response g gives the FRF $G(\Omega)$, i.e., $G(\Omega) = \mathscr{F}\{\mathrm{g}(t)\}$. Now, the frequency-domain representation of (4.1) is

$$Y_{\mathrm{s}}(k) = \underbrace{G(\Omega_k) U_{\mathrm{s}}(k) + \mathfrak{T}(\Omega_k)}_{Y_{\mathrm{s}0}(k)} + V_{\mathrm{s}}(k), \tag{4.56}$$

where the spectrum $Y_{\mathrm{s}0}(k)$ is the measurement noise-free output and $\mathfrak{T}(\Omega_k)$ is the transient, which depends on the difference $u(t) - u(t+N)$ for $t < 0$ and on the impulse response of the system (Lataire and Chen 2016, Lemma 1).

Given the exact input and measured output $N-$point DFT spectra $U_{\rm s}(k)$ and $Y_{\rm s}(k)$, we are aiming at estimating $G(\Omega)$ and $\mathfrak{T}(\Omega)$.

4.3.2 Bayesian frequency-domain identification

In the Bayesian approach to system identification within the GPR framework, the unknown function to be estimated is assumed to be a realisation of a zero-mean GP with a certain covariance (kernel) function that encodes our priori knowledge about it. Given observed data of joint Gaussian processes and *a prior* mean and covariance, the goal is to obtain the *a posteriori* mean and covariance, which can be used for prediction of the unknown function at arbitrary input values.

For the LTI system (4.56), it holds true that the FRF takes both real (at 0 Hz and at the Nyquist frequency) and complex values. As a result, it is not possible to model it as a real or as a complex GP. Hence, both the FRF and the transient have to be defined as a *Real/Complex* GP (RCGP) (Lataire and Chen 2016, Section 2). More specifically, a RCGP $\eta(k)$ is defined as

$$\eta(k) \sim \mathcal{RCGP}(m, K_{\rm cov}, K_{\rm rel}) \mid \mathbb{K}_{\mathcal{R}}, \tag{4.57}$$

where $m, K_{\rm cov}, K_{\rm rel}$ are the mean, covariance, and relation functions, respectively, and $\mathbb{K}_{\mathcal{R}}$ is a set of indices that indicates where $\eta(k)$ is real, i.e., $\mathbb{K}_{\mathcal{R}} = \{0, \pm N/2, \pm 2(N/2), \ldots\}$. Following the Bayesian approach within the GPR framework, the FRF $G(\Omega_k)$ and the transient $\mathfrak{T}(\Omega_k)$ are assumed to be independent of each other and are assumed to be RCGPs over $k \in \mathbb{R}$:

$$G(\Omega_k) \sim \mathcal{RCGP}(0, \beta_{\rm G} K_{\rm cov}, \beta_{\rm G} K_{\rm rel}) \mid \mathbb{K}_{\mathcal{R}}, \tag{4.58}$$

$$\mathfrak{T}(\Omega_k) \sim \mathcal{RCGP}(0, \beta_{\mathfrak{T}} K_{\rm cov}, \beta_{\mathfrak{T}} K_{\rm rel}) \mid \mathbb{K}_{\mathcal{R}}, \tag{4.59}$$

where $\beta_{\rm G} \geq 0, \beta_{\mathfrak{T}} \geq 0$, $K_{\rm cov}, K_{\rm rel}$ are well-defined covariance and relation functions, respectively. Once $K_{\rm cov}, K_{\rm rel}$ are defined, the *Maximum a Posteriori* (MAP) estimates $\hat{G}$ and $\hat{\mathfrak{T}}$ of the FRF and the transient, respectively, can be directly computed (Lataire and Chen 2016). Such an estimator is denoted in the following by GPTF.

4.3.3 Kernel functions in the frequency-domain

A natural way to construct the kernel function for FRF estimation is to utilize the duality between the FRF and impulse response function (Schoukens et al. 2004), i.e., $G(e^{\mathrm{j}\omega}) = \sum_{\tau=0}^{\infty} \mathfrak{g}(\tau) e^{-\mathrm{j}\omega\tau}$, and the linearity of the Fourier transform, to derive the corresponding covariance and relation functions in the frequency-domain. More specifically, if the impulse response function $\mathfrak{g}$ is assumed to be a realization of a zero-mean GP, as has been introduced in the previous section, with covariance $\mathrm{cov}(\mathfrak{g}(i), \mathfrak{g}(j)) = \beta_G K(i,j), i, j = 0, 1, \ldots$, then, $G(e^{\mathrm{j}\omega})$ is a RCGP with mean function

$$\mathcal{E}\{G(e^{\mathrm{j}\omega})\} = \mathscr{F}\big\{\mathcal{E}\{\mathfrak{g}(t)\}\big\} = 0, \tag{4.60}$$

and covariance and relation functions ($k, l \in \mathbb{R}$):

$$\beta_G K_{\mathrm{cov}}(e^{\mathrm{j}\omega_k}, e^{\mathrm{j}\omega_l}) = \mathcal{E}\{G(e^{\mathrm{j}\omega_k})G^*(e^{\mathrm{j}\omega_l})\}, \tag{4.61}$$

$$\beta_G K_{\mathrm{rel}}(e^{\mathrm{j}\omega_k}, e^{\mathrm{j}\omega_l}) = \mathcal{E}\{G(e^{\mathrm{j}\omega_k})G(e^{\mathrm{j}\omega_l})\} \tag{4.62}$$

$$= \beta_G K_{\mathrm{cov}}(e^{\mathrm{j}\omega_k}, e^{-\mathrm{j}\omega_l}), \tag{4.63}$$

where

$$K_{\mathrm{cov}}(e^{\mathrm{j}\omega_k}, e^{\mathrm{j}\omega_l}) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} K(i,j)e^{-\mathrm{j}\omega_k i}e^{\mathrm{j}\omega_l j}. \tag{4.64}$$

In Lataire and Chen (2016), the authors make use of such an approach to define kernel functions used in the time-domain based literature for FRF estimation, e.g., DC and SS kernels, see the previous section for more details. For the sake of space, the formulation of these kernels in the frequency domain can be found in (Lataire and Chen 2016, Equations 55,56). Furthermore, it has been proven that these kernels guarantee stability of the resulting model estimates. A sufficient condition on the kernel function to guarantee the stability of the estimated FRF is to satisfy the condition in (Lataire and Chen 2016, Property 7) or equivalently, the corresponding impulse response of the estimated FRF must be absolutely summable, see Proposition 2.1.

Remark 4.3 *Regarding the kernel function for the transient $\mathfrak{T}(\Omega)$, it has been shown in (Lataire and Chen 2016, Section 5.3) that a computational convenient way is to assume $G(\Omega)$ and $\mathfrak{T}(\Omega)$ have the same kind of covariance function, but with different scaling hyperparameters β_G and $\beta_{\mathfrak{T}}$, respectively.*

The aforementioned kernels can describe stability and smoothness of the estimated FRF. However, as recommended in Lataire and Chen (2016), kernels that are able to describe other dynamic properties would be beneficial in the frequency-domain identification, e.g., resonance behavior, damping, etc., but keeping a simple structure of the kernel function. In Section 4.2, we have introduced a class of kernels for regularized impulse response estimation and it has been shown that they correspond to a systematic construction of efficient kernels that are able to describe a wide range of dynamic properties with simple parameterization. Fortunately, the OBFs based kernels introduced in Section 4.2 have a direct representation in the frequency-domain. Next, we show how to make use of the frequency-domain formulation of the OBFs based kernel to estimate the FRF and the transient directly in the frequency-domain.

4.3.4 OBFs based kernels in the frequency-domain

It is well-known that the space spanned by the OBFs $\breve{\Psi}$, i.e., $\mathscr{RH}_2(\mathbb{E})$, is an RKHS (Ninness et al. 1999) with the following reproducing kernel

$$K_{\mathrm{cov}}(e^{\mathrm{j}\omega_k}, e^{\mathrm{j}\omega_l}) = \sum_{i=0}^{\infty} \breve{\psi}_i(e^{\mathrm{j}\omega_k})\breve{\psi}_i^*(e^{\mathrm{j}\omega_l}), \tag{4.65}$$

and with the following well-defined inner product

$$\langle f_1, f_2 \rangle_{\mathscr{RH}_2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f_1(e^{\mathrm{j}\omega}) f_2^*(e^{\mathrm{j}\omega}) d\omega, \tag{4.66}$$

for any $f_1, f_2 \in \mathscr{RH}_2(\mathbb{E})$.

Now, let us look into the details of the stability of the estimated FRF based on the OBFs based kernel in (4.65). Similarly to the reasoning presented in Section 4.1.2, the space spanned by the OBFs, which is used to construct the kernel and used as a hypothesis space for the estimation problem should be restricted to a subset of $\mathscr{RH}_2(\mathbb{E})$, i.e., $\mathscr{RH}_{2-}(\mathbb{E})$. Because when the space spanned by the OBFs is utilized, the resulting estimate will be in $\mathscr{RH}_2(\mathbb{E})$, which means that the impulse response function corresponding to the estimated FRF will belong to $\mathcal{R}\ell_2(\mathbb{N})$. Accordingly the stability condition that the impulse response should be absolutely summable, i.e., $\hat{\mathrm{g}} \in \mathcal{R}\ell_1(\mathbb{N})$, is not satisfied as $\mathcal{R}\ell_2(\mathbb{N}) \not\subset \mathcal{R}\ell_1(\mathbb{N})$.

In order to guarantee the stability of the estimated FRF and at the same time tackle the problem of determining the right number of basis functions, similarly, as has been done in the time-domain, we include a decay term that weighs the OBFs

$$K_{\mathrm{cov}}(e^{\mathrm{j}\omega_k}, e^{\mathrm{j}\omega_l}) = \sum_{i=0}^{\infty} \mathfrak{D}_i(\beta_{\mathrm{d}}) \breve{\psi}_i(e^{\mathrm{j}\omega_k}) \breve{\psi}_i^*(e^{\mathrm{j}\omega_l}), \tag{4.67}$$

where the decay term $\mathfrak{D}_i(\beta_{\mathrm{d}}) \to 0$ as $i \to \infty$ and β_{d} is considered to be a hyperparameter that determines the decay rate of the expansion in (4.67). The decay term, i.e., $\mathfrak{D}_i(\beta_{\mathrm{d}})$, with β_{d} tuned by marginal likelihood optimization acts as an automatic way to select the number of significant basis functions that is needed to construct the kernel. Possible choices for monotonically decreasing weights are

$$\mathfrak{D}_i(\beta_{\mathrm{d}}) = i^{-\beta_{\mathrm{d}}}, \ \beta_{\mathrm{d}} \geq 0, \tag{4.68}$$

$$\mathfrak{D}_i(\beta_{\mathrm{d}}) = \beta_{\mathrm{d}}^i, \ 0 \leq \beta_{\mathrm{d}} < 1, \tag{4.69}$$

Note that the relation function K_{rel} can be constructed accordingly via (4.67) and (4.61)-(4.62).

4.3.5 Hyperparameters tuning

The kernel function defined above, i.e., the OBFs based kernel (4.67), depends on some unknown hyperparameters that need to be tuned from the observed data. These hyperparameters are the scaling parameters β_{G} and $\beta_{\mathfrak{T}}$, the noise variance σ_{e}^2, the parameter β_{d} that determines the decay rate of the expansion and a vector β_{p} of the generating poles of the OBFs. Denote by β the vector of the hyperparameters, i.e., $\beta = [\ \beta_{\mathrm{G}} \quad \beta_{\mathfrak{T}} \quad \sigma_{\mathsf{e}}^2 \quad \beta_{\mathrm{d}} \quad \beta_{\mathrm{p}}^\top\ ]^\top$. One popular approach to tune β within the Bayesian framework is by maximising the log marginal likelihood, i.e.,

$\log \mathrm{p}(Y_{\mathrm{s}}(\mathcal{K} \mid \beta))$ of the output spectrum (Rasmussen and Williams 2006)

$$\log\left(Y_{\mathrm{s}}(\mathcal{K} \mid \beta)\right) = -\frac{1}{2} Y_{\mathrm{s}}(\mathcal{K})^{H} \Gamma_{Y_{\mathrm{s}}}^{-1}(\beta) Y_{\mathrm{s}}(\mathcal{K}) - \frac{1}{2}\log\left|\Gamma_{Y_{\mathrm{s}}}(\theta)\right| - \frac{n_{\mathrm{r}}}{2}\log 2\pi - n_{\mathrm{c}}\log\pi, \quad (4.70)$$

where $\mathcal{K} = \{k_1, k_2, \ldots, k_{\mathrm{n}}\} \subset \{0, \ldots, N/2\}$ is the set DFT-frequency indices that lie in the frequency band of interest, $\Gamma_{Y_{\mathrm{s}}}$ is the augmented covariance matrix which can be constructed from the covariance and relation functions, i.e., K_{cov} and K_{rel} (4.67) and (4.61)-(4.62), see (Lataire and Chen 2016, Equation 36) for constructing $\Gamma_{Y_{\mathrm{s}}}$, n_{r} is the number of frequencies where the FRF has real values and n_{c} is the number of frequencies where the FRF has complex values.

4.3.6 Simulation studies

In this section, the presented OBFs based kernel function, formulated directly in the frequency domain, is tested and compared to the existing kernels, e.g., the DC kernel, for FRF estimation. A challenging system is considered to show the capability of the presented kernel to model a wide range of dynamic properties, specifically resonance behavior, with a simple kernel structure.

Considered system

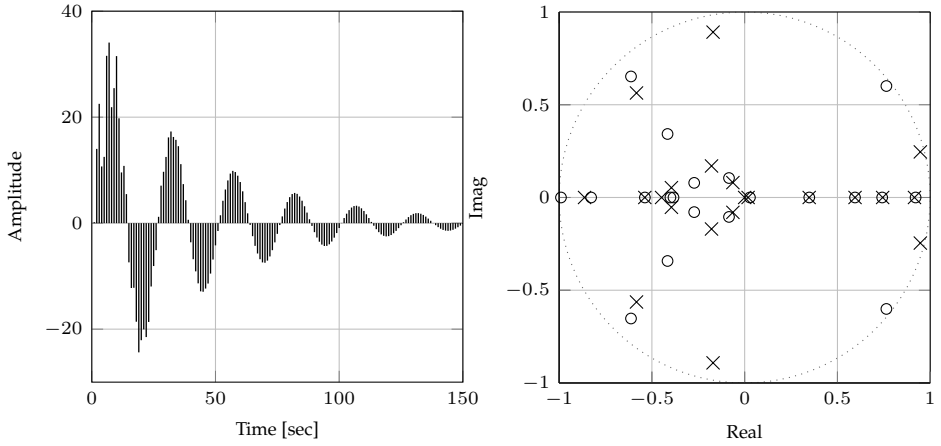
We consider a randomly generated 20-th order, LTI and discrete-time system G generated by the `drss` Matlab function. The sampling period T_{s} is $1\,s$. We make sure that there are two dominant complex conjugate pole pairs. These dominant poles are located at $0.95 \pm j0.25, -0.17 \pm j0.89$, see Figure 4.11 for the impulse response and the pole/zero plot of the generated system.

Identification settings

MC simulations of 100 runs are performed, where at each run a new realisation of the input $u(t)$ and the noise $v(t)$ are utilised according to (4.1). The considered system is used to generate a data set of length $N = 512$ for each MC run using a zero-mean white, Gaussian and periodic input u and an additive white Gaussian noise v. The variance of v is chosen such that the SNR corresponding to two estimation scenarios is 10dB or 40dB.

The considered estimators are:

- GPTF with DC kernel;
- GPTF with OBFs based kernel, specifically, with GOBFs based kernel where the poles of the inner function $\mathcal{H}_{\mathrm{b}}$ are $\{\lambda_1, \lambda_1^*, \lambda_2, \lambda_2^*\}$, which are considered as hyperparameters;



(a) Impulse response. (b) Poles (×) and zeros (o).

Figure 4.11: The impulse response and pole/zero plot of the considered system.

- Parametric model identified with the Identification Toolbox of Matlab (2016a), more specifically, an OE model with the true order of the system, i.e., using the command `oe(20,20)`. We will call this estimator as an Oracle estimator, in the sense that it knows the true model structure and order.

For each MC run, the hyperparameters of the GPTF estimators, for both DC and OBFs kernels, are tuned by maximising the marginal likelihood (4.70). The estimation is performed on a limited frequency band, i.e., from $\omega = 0.1$ rad/s to $\omega = 3$ rad/s. For the GPTF estimator, 241 frequency domain samples in the mentioned range were used, whereas the OE model was estimated based on the whole data record.

Results and discussion

The performance measure that is used to determine the quality of the estimated FRF with different estimators is the averaged MSE over all frequencies in the band of interest, i.e.,

$$\text{MSE} = \frac{1}{100}\sum_{i=1}^{100}\left(\frac{1}{N}\sum_{k=1}^{N}|\hat{G}_i(\Omega_k) - G(\Omega_k)|^2\right), \tag{4.71}$$

where $\hat{G}_i$ is the estimated FRF at the MC run i, which is calculated on a more dense frequency grid, i.e., 966 frequencies, but within the same frequency band as the training data set.

The averaged MSE for the estimates over all the frequencies in the considered frequency band is summarized in Table 4.5. It can be seen from the table that

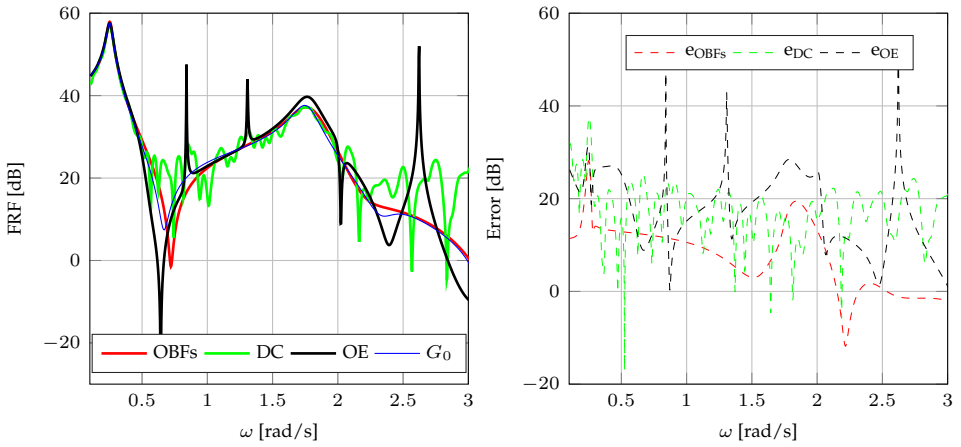
Table 4.5: Averaged MSE of all estimates (in dB) for different SNR scenarios.

Estimator	10dB	40dB
GPTF (DC)	46.45	-0.28
GPTF (GOBFs)	33.46	-7.05
OE (20-th order)	50.94	6.89

GPTF estimators perform better than the parametric estimator, even though the latter makes use of more data points and more importantly it makes use of the true model structure. Moreover, the GPTF estimator with the GOBFs based kernel shows a significant improvement with respect to the GPTF estimate with the DC kernel. The main reason is that the complex conjugate poles included in the GOBFs based kernel are better at modeling the resonance behavior and result in a smoother estimate. To visualise such results, the left parts of Figure 4.12, 4.13, show the estimated FRF and the true function at the validation set of frequencies of one MC run within the considered frequency band for both cases of SNR of 10dB and 40dB, respectively. The right parts of the figures show the error associated with the employed estimators in dB. From these figures, it can be easily seen that the OBFs based kernel performs well compared to the DC kernel based estimator and can deliver an acceptable estimate even in the high frequency range.

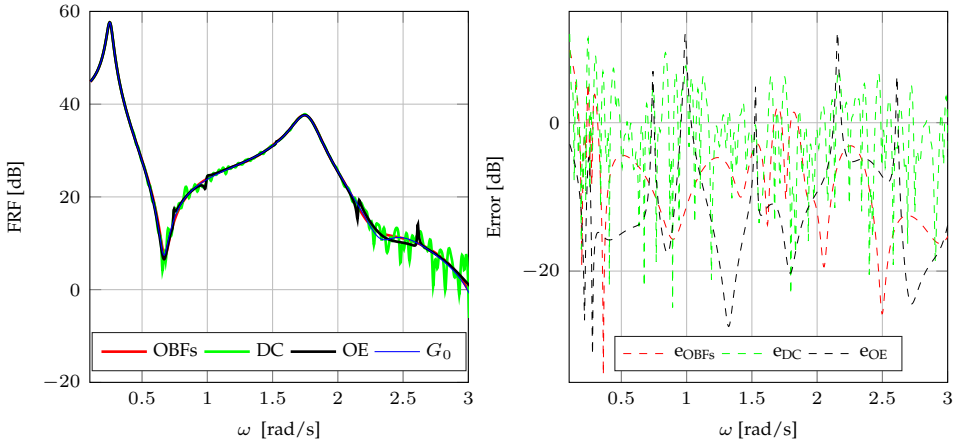
4.4 Summary

In this chapter, Bayesian identification of stable LTI systems has been discussed. First, we have overviewed the classical approach to do that. Then, we have recaped the modifications that are needed to adapt the approaches from machine learning to be applied to dynamic system identification. Existing kernel functions for LTI impulse response identification have been also discussed and the need for a new class of kernels has been motivated. The answer of that has been found in constructing kernel functions based on OBFs where the dynamic prior knowledge can be systematically encoded via the generating poles of the used OBFs. Such kernels have been modified to guarantee the stability of the estimates by including a decay term that automatically selects the number of significant basis that should be used in the kernel construction. The connection between regularized impulse response estimation with OBFs based kernels and regularized OBFs series expansion based model estimation has been given and the equivalence between them has been proven theoretically as well as empirically. Moreover, we have shown that OBFs based kernels have a direct representation in the frequency-domain and hence provide a powerful tool that can represent dynamic properties in that domain.



(a) (b)

Figure 4.12: (a) Plot of one MC realisation of the FRF of the data-generating system with the results of various estimators in case of SNR=10dB. (b) The frequency wise magnitude of the error (in dB) associated with the considered estimators.



(a) (b)

Figure 4.13: (a) Plot of one MC realisation of the FRF of the data-generating system with the results of various estimators in case of SNR=40dB. (b) The frequency wise magnitude of the error (in dB) associated with the considered estimators.

5 CHAPTER

Bayesian Identification of LPV Systems

This chapter aims at addressing Subgoals 2 and 3. More specifically, the Bayesian identification approach presented in Section 3.3 is extended to efficiently estimate, both in the stochastic and computational sense, MIMO LPV models under general noise model structure of the BJ type. The approach is based on the estimation of the one-step-ahead predictor form of general LPV-BJ structures, where the sub-predictors associated with the input and output signals are captured as asymptotically stable IIR models. These IIR sub-predictors are identified in a completely nonparametric sense, where not only the coefficients are estimated as functions, but also the whole time evolution of the impulse response w.r.t. the scheduling signal. In the resulting Bayesian setting, the estimate of the one-step-ahead predictor is a realization from a zero-mean Gaussian random field, where the covariance function is a multidimensional Gaussian kernel that encodes both the possible structural dependencies and the stability of the predictor. As a next step, the developed approach is extended to the identification of series-expansion models, e.g., LPV-IIR and LPV-OBFs model structures. This chapter is organized as follows: a brief introduction to LPV prediction error identification is presented in Section 5.2. The Bayesian identification of LPV-IO models is developed in Section 5.3, where a suitable kernel function is designed to encode our prior knowledge about such models. Finally, the extension of the developed Bayesian approach to LPV series expansion models is given in Section 5.4.

5.1 Introduction

Identification of LPV-IO models have gained popularity, as PEM methods have been successfully extended to LPV models, providing a well-understood frame-

work for consistency and stochastic interpretation of the estimates together with low computational complexity of the resulting identification procedures (Tóth et al. 2012a). Moreover, the PEM framework offers a large class of noise and plant models, see Tóth (2010) for an overview. Although, LPV-IO models offer a variety of process and noise representations, where the BJ model is the most general form, PEM identification of BJ models leads to a nonlinear optimization problem (Tóth et al. 2012a), which is sensitive to local minima. Alternatively, the IV method provides an attractive approach that deals with the general noise scenario and avoids nonlinear optimization (Laurain et al. 2010). Another important issue in the identification of LPV-IO models is capturing the structural dependency on the scheduling signal. In the parametric case, the structural dependency is characterized by using a pre-specified set of basis functions, which need either a significant prior knowledge of the underlying system or tedious repetitive execution of methods to synthesize an acceptable basis (Tóth et al. 2012a). In addition, the choice of the number of these basis represents a challenge, as it directly introduces a bias/variance trade-off, i.e., by using a smaller number of basis functions, the under-modeling (bias) error will increase while increasing their number results in an increase of the variance of the estimated models.

Alternatively, the so-called nonparametric methods offer an attractive approach to capture the underlying dependencies directly from data without specifying any parameterization in terms of fixed basis functions. This work is inspired by the recent advances in nonparametric identification of LTI models in a PEM setting (Pillonetto et al. 2011a) and in the design of optimal kernels (Pillonetto et al. 2014). Here, we aim at formulating a nonparametric estimator of the one-step-ahead predictor for an LPV-BJ model, preserving both the generality of the noise class and the asymptotic optimality of PEM. More specifically, we consider the one-step-ahead predictor as the summation of two sub-predictors associated with the input and output signals, where these sub-predictors are captured as asymptotically stable LPV-IIR models. These LPV-IIR sub-predictors are identified in a nonparametric sense, where not only the coefficients are estimated as functions, but also the whole time evolution of the impulse response.

We follow a Bayesian approach for the nonparametric estimation by modeling the sub-predictors as a realization of a zero-mean Gaussian random field, which can be completely characterized by its covariance (kernel) function that implicitly acts as a basis generator to describe both the functional dependencies and the time evolution of the impulse response of the sub-predictors. To this end, we introduce a multidimensional Gaussian kernel which encodes:

1 the possible structural dependencies on the scheduling signal by using RBF;

2 the stability of the predictor by including a decay term, which models the vanishing influence of the past input-scheduling-output pairs on the predicted output.

Two different kernel formulations are presented for the LPV setting, namely the DI-like and TC-like kernels, where the TC-like kernel is able to describe the correlation between different coefficient functions associated with different time in-

stants. The hyperparameters that parameterize the kernel can be efficiently estimated from data by maximizing the marginal likelihood w.r.t. the observations (MacKay 2003). The main contribution in this chapter includes the kernel function formulation for the MIMO case.

Interestingly, the developed approach for the identification of LPV-IO models can be extended to the identification of LPV series-expansion models. More specifically, a nonparametric approach is presented for the identification of a general class of LPV models, i.e., LPV-IIR and LPV-OBFs models that can be handled in a similar setting. Following a nonparametric Bayesian approach, most of the challenges and problems associated with identification of such models are tackled. The parameterization of the coefficients functions are avoided and the convergence of the series-expansion is guaranteed. This is done by considering the estimation of the expansion as a function estimation problem, where it is modeled as a zero-mean Gaussian random field with a multidimensional covariance function that encodes the prior knowledge about the structural dependencies of the coefficient functions and the convergence of the expansion.

For LPV-OBFs models, the considered model structure is obtained by using a fixed OBFs structure, where the expansion coefficients are assumed to depend on a measurable and time-varying scheduling signal. The generating poles of the OBFs structure are considered as hyperparameters and are tuned by maximizing the marginal likelihood. Moreover, the presented approach is directly applicable in case the scheduling signal can not be kept fixed and the identification has to be done based on a time-varying scheduling signal, i.e., LTI models at different constant scheduling signal can not be obtained, hence the FKcM approach or any other method that relies on the optimization of the basis using such local information can not be applied.

5.2 Prediction error identification of LPV systems

5.2.1 Impulse response representation of LPV systems

Consider an LPV system $\mathcal{S}$, represented in a filter form of a discrete-time IO representations (difference equation), which is defined in the MIMO case as:

$$\underbrace{\sum_{i=0}^{n_{\mathrm{a}}} \mathfrak{a}_i(p,t)q^{-i}}_{A(p,t,q^{-1})} y(t) = \underbrace{\sum_{j=0}^{n_{\mathrm{b}}} \mathfrak{b}_j(p,t)q^{-j}}_{B(p,t,q^{-1})} u(t), \tag{5.1}$$

where $u : \mathbb{Z} \to \mathbb{U} = \mathbb{R}^{n_{\mathbb{U}}}$ is the input, $y : \mathbb{Z} \to \mathbb{Y} = \mathbb{R}^{n_{\mathbb{Y}}}$ is the output, $p : \mathbb{Z} \to \mathbb{P}$ is the so-called scheduling variable, which is assumed to be known exactly[1], with compact range $\mathbb{P} \subset \mathbb{R}^{n_{\mathbb{P}}}$, the matrix functions $\mathfrak{a}_i(p,t) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_{\mathbb{Y}} \times n_{\mathbb{Y}}}$ and $\mathfrak{b}_j(p,t) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_{\mathbb{Y}} \times n_{\mathbb{U}}}$ are shorthand notations for $\mathfrak{a}_i(p,t) =$

[1] In the LPV framework, the scheduling variable is always assumed to be measurable or available based on the exact modeling approach taking w.r.t. the physical system, for details see Tóth (2010).

$\mathfrak{a}_i(p(t),\dots,p(t-i))$ and $\mathfrak{b}_j(p,t) = \mathfrak{b}_j(p(t),\dots,p(t-j))$. These functions are assumed to be smooth and bounded functions on $\mathbb{P}$ to guarantee the well-posedness of (5.1). Furthermore, the p-dependent operators $A(p,t,q^{-1})$ and $B(p,t,q^{-1})$ are matrix polynomials in q^{-1} of degree $n_\mathrm{a}, n_\mathrm{b} \geq 0$, respectively, with varying coefficients $\mathfrak{a}_i, \mathfrak{b}_j$. Next, the notion of LPV stability is described, which is relevant to the subsequent discussion.

Definition 5.1 (Asymptotic stability of LPV systems) *an LPV system $\mathcal{S}$ represented, e.g., in a filtered form* (5.1)*, is called globally asymptotically stable, if for all trajectories of $\{u(t), y(t), p(t)\} \in (\mathbb{U}, \mathbb{Y}, \mathbb{P})^{\mathbb{Z}}$ satisfying the system equation* (5.1) *with $u(t) = 0$ for $t \geq 0$ it holds that $\lim_{t\to\infty} |y(t)| = 0$.*

A computational approach to check the asymptotic LPV-IO stability of Definition 5.1 can be found in Wollnack et al. (2017).

Definition 5.2 (BIBO stability of LPV systems) *an LPV system $\mathcal{S}$ represented, e.g., in a filtered form* (5.1)*, is called globally BIBO stable, if for all trajectories of $\{u(t), y(t), p(t)\} \in (\mathbb{U}, \mathbb{Y}, \mathbb{P})^{\mathbb{Z}}$ satisfying the system equation* (5.1)*, all bounded input trajectories will result in bounded output trajectories, i.e., in the ℓ_k-norm for every $1 \leq k < \infty$ we have*

$$\sum_{t=0}^{\infty} \|u(t)\|_{\ell_k} < \infty \quad \Rightarrow \quad \sum_{t=0}^{\infty} \|y(t)\|_{\ell_k} < \infty.$$

It is worth to mention that just like in the LTI case asymptotic stability of LPV systems implies BIBO stability.

In Tóth (2010), it has been shown that, in discrete-time, the dynamic relation of an asymptotically stable, according to Definition 5.1, LPV system $\mathcal{S}$, represented in (5.1) can be described as a convolution involving p and u, corresponding to an LPV-IIR form. This convolution is given as

$$\begin{aligned} y(t) &= \sum_{k=0}^{\infty} \mathfrak{g}_k(p,t)u(t-k), \\ &= \big((G(q) \diamond p)u\big)(t) \triangleq \left(\sum_{k=0}^{\infty} \mathfrak{g}_k(p,t)q^{-k}u\right)(t) \end{aligned} \tag{5.2}$$

where $G(q)\diamond p = A^\dagger(p,t,q^{-1})B(p,t,q^{-1})$ is the transfer operator, with $A^\dagger(p,t,q^{-1})$ denoting the unilateral inverse (Darwish et al. 2017b, Lemma 2) of $A(p,t,q^{-1})$ under the condition that A is monic, i.e., $\mathfrak{a}_0(p,t) \equiv I$. The symbol $\diamond$ is used to express that the dynamic relationship is dependent on the trajectory of p. The so-called impulse response coefficients $\mathfrak{g}_k$ are functions of $p(t)$ and its time shifted instances $p(t-1), p(t-2), \dots, p(t-k)$ and assumed to be smooth and bounded on $\mathbb{P}$. Note that $\mathfrak{g}_k(p,t) : \mathbb{P} \times \dots \times \mathbb{P} \to \mathbb{R}^{n_\mathbb{Y} \times n_\mathbb{U}}$ is a shorthand notation for $\mathfrak{g}_k(p,t) = \mathfrak{g}_k(p(t),\dots,p(t-k))$. The description (5.2) is known as the LPV-IIR and can be seen as a *series-expansion representation* of $\mathcal{S}$ in terms of the so-called pulse basis $\{q^{-k}\}_{k=0}^{\infty}$. It can be proven that for any asymptotically stable LPV system $\mathcal{S}$, the

expansion (5.2) is convergent (Tóth 2010). Note that for a constant scheduling signal, i.e., $p(t) = \bar{\mathrm{p}}\ \forall t \in \mathbb{Z}$ with $\bar{\mathrm{p}} \in \mathbb{P}$ being a constant, (5.2) becomes equivalent to the IIR of an LTI system, where $\mathfrak{g}_k(\bar{\mathrm{p}}, \ldots, \bar{\mathrm{p}})$ corresponds to the k-th Markov parameter of that LTI system. Due to the asymptotic stability of the considered system, the IIR in (5.2) can be truncated to a finite expansion

$$y(t) \approx \sum_{k=0}^{n} \mathfrak{g}_k(p,t) q^{-k} u(t), \tag{5.3}$$

which is known as an LPV-FIR model of order n, where the truncation error, i.e., $\sum_{k=n+1}^{\infty} \mathfrak{g}_k(p,t) q^{-k} u(t)$, can be made arbitrary small by choosing n large enough to capture the dominant dynamics of the system. By using the LPV-IIR representation, the classical LTI-PEM framework has been successfully extended to the LPV case in Tóth (2010); Tóth et al. (2012a), where such a setting allows for sophisticated analysis of LPV-IO models. In the following, we present the data-generating system, which is an essential concept of the LPV-PEM framework.

5.2.2 Data-generating system

Similar to the LTI-PEM case, the LPV data-generating system is considered as a discrete-time deterministic p-dependent filter G_0 whose output is affected by an additive stochastic noise process $v : \mathbb{Z} \to \mathbb{Y}$, which is assumed to be a quasi-stationary with a bounded power spectral density (Tóth 2010), see Figure 5.1. Consider a MIMO LPV data-generating system described in DT that can be described by the following difference equations:

$$A_0(p,t,q^{-1})\breve{y}(t) = B_0(p,t,q^{-1})u(t), \tag{5.4a}$$
$$D_0(p,t,q^{-1})v(t) = C_0(p,t,q^{-1})e(t), \tag{5.4b}$$
$$y(t) = \breve{y}(t) + v(t), \tag{5.4c}$$

where $\breve{y} : \mathbb{Z} \to \mathbb{Y} = \mathbb{R}^{n_\mathbb{Y}}$ is the noiseless output. Similar to (5.1), the p-dependent operators $A_0(p,t,q^{-1})$ and $B_0(p,t,q^{-1})$ that describe the process model (5.4a) are matrix polynomials in q^{-1} of degree n_a and n_b, respectively:

$$A_0(p,t,q^{-1}) = I_{n_\mathbb{Y}} + \sum_{i=1}^{n_\mathrm{a}} \mathfrak{a}_i^\mathrm{o}(p,t) q^{-i}, \tag{5.5a}$$

$$B_0(p,t,q^{-1}) = \sum_{j=0}^{n_\mathrm{b}} \mathfrak{b}_j^\mathrm{o}(p,t) q^{-j}, \tag{5.5b}$$

where the matrix functions $\mathfrak{a}_i^\mathrm{o}(p,t) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_\mathbb{Y} \times n_\mathbb{Y}}$ and $\mathfrak{b}_j^\mathrm{o}(p,t) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_\mathbb{Y} \times n_\mathbb{U}}$ are shorthand notations for $\mathfrak{a}_i^\mathrm{o}(p,t) = \mathfrak{a}_i^\mathrm{o}(p(t), \ldots, p(t-i))$ and $\mathfrak{b}_j^\mathrm{o}(p,t) = \mathfrak{b}_j^\mathrm{o}(p(t), \ldots, p(t-j))$. These functions are assumed to be smooth and bounded functions on $\mathbb{P}$. The resulting transfer operator that describes the process

part, i.e., the deterministic part, is given as

$$G_0(p,t,q^{-1}) = A_0^{\dagger}(p,t,q^{-1})B_0(p,t,q^{-1}), \tag{5.6a}$$

$$= \sum_{k=0}^{\infty} \mathfrak{g}_k^{\circ}(p,t)q^{-k} \tag{5.6b}$$

where $A_0^{\dagger}$ denotes the unilateral left inverse (Darwish et al. 2017b, Lemma 2), of the monic A_0 and $\mathfrak{g}_k^{\circ}$ are the impulse response coefficients associated with the LPV-IIR of the process part and are assumed to be bounded w.r.t. all $p \in \mathbb{P}$. Remember that, such an IIR is guaranteed to be convergent under asymptotic

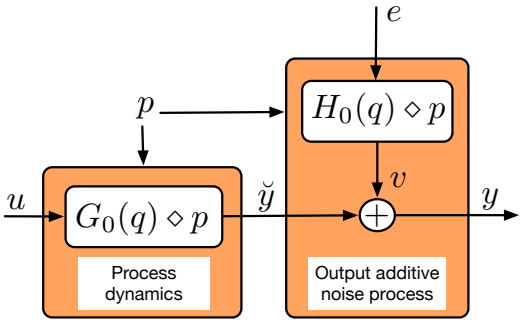


Figure 5.1: LPV concept of the data-generating system.

stability of the true underlying system. In a similar fashion, the noise model relation (5.4b), i.e., the stochastic part, characterized by $C_0(p,t,q^{-1})$ and $D_0(p,t,q^{-1})$, which are monic matrix polynomials (5.7) in q^{-1}, is defined as

$$C_0(p,t,q^{-1}) = I_{\mathrm{n_Y}} + \sum_{i=1}^{n_{\mathrm{c}}} \mathfrak{c}_i^{\circ}(p,t)q^{-i}, \tag{5.7a}$$

$$D_0(p,t,q^{-1}) = I_{\mathrm{n_Y}} + \sum_{j=1}^{n_{\mathrm{d}}} \mathfrak{d}_j^{\circ}(p,t)q^{-j}, \tag{5.7b}$$

where $\mathfrak{c}_i^{\circ}(p,t) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_{\mathrm{Y}} \times n_{\mathrm{Y}}}$ and $\mathfrak{d}_j^{\circ}(p,t) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_{\mathrm{Y}} \times n_{\mathrm{Y}}}$ are the coefficient function matrices of degree n_{c} and n_{d}, respectively. The resulting transfer operator that describes the stochastic part is given as

$$H_0(p,t,q^{-1}) = D_0^{\dagger}(p,t,q^{-1})C_0(p,t,q^{-1}), \tag{5.8a}$$

$$= I_{\mathrm{n_Y}} + \sum_{k=1}^{\infty} \mathfrak{h}_k^{\circ}(p,t)q^{-k}, \tag{5.8b}$$

where $D_0^{\dagger}$ denotes the unilateral left inverse of the polynomial D_0, and $\mathfrak{h}_k^{\circ}$ are assumed to be bounded functions. Hence, the noise process v can be described by

$$v(t) = H_0(p,t,q^{-1})e(t), \tag{5.9}$$

where H_0 represents a convergent and monic LPV-IIR, i.e., it corresponds to an asymptotically stable[2] LPV filter with $H_0(\cdot,\cdot,\infty) = I_{\mathrm{n}_\mathbb{Y}}$, and $e : \mathbb{Z} \to \mathbb{Y}$ is a white noise process with normal (Gaussian) distribution, i.e., $e(t) \sim \mathcal{N}(0, \Sigma_{\mathsf{e}})$ with covariance $\Sigma_{\mathsf{e}} \in \mathbb{R}^{n_\mathbb{Y} \times n_\mathbb{Y}}$.

5.2.3 The IIR form of the one-step-ahead predictor

In classical identification approaches, one is interested in finding the process G_0 and noise dynamics H_0, e.g., see (Ljung 1999),

$$y(t) = G_0(p,t,q^{-1})u(t) + H_0(p,t,q^{-1})e(t), \tag{5.10}$$

where the process and noise models are defined in (5.6)-(5.8). Similar to the LTI case (Ljung 1999), under the assumption of the existence of a stable inverse of the H_0, the representation (5.10) can be formulated based on the trajectory of u, p, y and the current value of e as[3]

$$y(t) = \big(I_{\mathrm{n}_\mathbb{Y}} - H_0^\dagger(p,t,q^{-1})\big)y(t) + H_0^\dagger(p,t,q^{-1})G_0(p,t,q^{-1})u(t) + e(t). \tag{5.11}$$

Assume furthermore that a data sequence $\mathcal{D}_N = \{u(t), p(t), y(t)\}_{t=1}^N$, generated by (5.10), is available. Under the given assumptions, the so-called *one-step ahead prediction* of $y(t)$ based on $\{y(t-1), y(t-2), \ldots\}$, $\{p(t), p(t-1), \ldots\}$ and $\{u(t), u(t-1), \ldots\}$ is the conditional expectation of (5.11) w.r.t. the past data, providing that

$$\hat{y}(t) = \big(I_{\mathrm{n}_\mathbb{Y}} - H_0^\dagger(p,t,q^{-1})\big)y(t) + H_0^\dagger(p,t,q^{-1})G_0(p,t,q^{-1})u(t). \tag{5.12}$$

Following a similar reasoning as in the LTI case, a parameterized model is hypothesized i.e., $(G_\theta(p,t,q^{-1}), H_\theta(p,t,q^{-1}))$, where $\theta \subset \Theta$ represents the parameter vector that comes from a real-valued parameterization of coefficients of the model, i.e., $\{\mathfrak{a}_i^\circ, \ldots, \mathfrak{d}_i^\circ\}$ according to (5.4), and $\Theta \in \mathbb{R}^{n_\theta}$ is the allowed parameter space. This model structure leads to the one-step ahead predictor:

$$\hat{y}_\theta(t) = \big(I_{\mathrm{n}_\mathbb{Y}} - H_\theta^\dagger(p,t,q^{-1})\big)y(t) + H_\theta^\dagger(p,t,q^{-1})G_\theta(p,t,q^{-1})u(t). \tag{5.13}$$

Now again, we are looking for an estimate of θ such that $\hat{y}_\theta$ is a "good" approximation of y, in the sense that the prediction error

$$\epsilon(t,\theta) := y(t) - \hat{y}_\theta(t), \tag{5.14}$$

is minimized, which can be performed, just like in the LTI case, by the minimization of the scalar-valued LS identification criterion (2.36).

Under the assumption that, in (5.12), both $\left(I_{\mathrm{n}_\mathbb{Y}} - H_0^\dagger\right)$ and $H_0^\dagger G_0$ correspond to stable LPV filters, the one-step-ahead predictor of the system given in (5.4) can

[2]The asymptotic stability of H_0 is a necessary assumption in the classical PEM setting, otherwise, the power spectrum density of the noise process would not be bounded resulting in an ill-posed estimation problem of G_0. It is possible to define PEM and estimation under noise scenarios which can not modeled under this condition, however, such considerations are beyond the scope of this thesis.

[3]The interested reader is referred to Tóth et al. (2012a) for a detailed proof.

be written as:

$$\hat{y}(t \mid t-1) = \sum_{i=1}^{\infty} \mathfrak{h}_{y,i}(p,t) q^{-i} y(t) + \sum_{j=0}^{\infty} \mathfrak{h}_{u,j}(p,t) q^{-j} u(t), \tag{5.15}$$

where, $\mathfrak{h}_{y,i}(p,t) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_\mathbb{Y} \times n_\mathbb{Y}}$ and $\mathfrak{h}_{u,j}(p,t) : \mathbb{P} \times \ldots \times \mathbb{P} \to \mathbb{R}^{n_\mathbb{Y} \times n_\mathbb{Y}}$ are real bounded and smooth[4] matrix coefficient functions in the scheduling signal p.

With the IIR representation it is possible to formulate a predictor, which is only based upon the past input-scheduling-output signals. It is worth to mention that the IIR predictor form (5.15) also exists for LPV state-space models represented with a specific innovation noise structure, making representation (5.15) also attractive for identifying such LPV-state space models, e.g., (van Wingerden and Verhaegen 2009; Proimadis et al. 2015).

This representation is similar to the LTI case, where the one-step-ahead predictor is written as a summation of the output of two IIRs. In the LTI case, estimation of such a predictor boils down to the estimation of two impulse responses with LS, whereas in the LPV case and due to the difficulties associated with parameterizing the structural dependency on p, such an estimation task becomes challenging. More specifically, one may go for a simple parameterization of $(\mathfrak{a}_i^o, \ldots, \mathfrak{d}_i^o)$ in (5.5)-(5.7), which results in a nonlinear estimation problem, prone to local minima. A direct parameterization of $\mathfrak{h}_{y,i}, \mathfrak{h}_{u,j}$ in (5.15) may lead to a simple estimation problem, but requires a complicated basis with large number of parameters to be estimated from data, which leads to estimates with large variance. Moreover, the "optimal" basis that are required to parameterize $\mathfrak{h}_{y,i}, \mathfrak{h}_{u,j}$ are rarely known a priori. The question is how to utilize the simplicity of the IIR form (5.15), but overcome the issues associated with its identification, i.e., how to avoid the need for large parameterization of $\mathfrak{h}_{y,i}, \mathfrak{h}_{u,j}$ and at the same time to optimize the bias/variance trade-off of the estimates? A solution can be found in the regularization framework, see Chapter 3. More specifically, the functional dependencies $\mathfrak{h}_{y,i}, \mathfrak{h}_{u,j}$ are estimated nonparametrically, where a regularization is introduced to keep the variance of the estimates low by allowing a small amount of estimation bias.

5.3 Bayesian identification of LPV-IO models

In this section, the Gaussian regression framework of Chapter 3, will be applied to the LPV-IIR representation of the one-step- ahead predictor (5.15). First, the identification of (5.11) will be formulated as a function estimation problem as shown in Chapter 3. Second, an appropriate kernel K will be designed for the estimation of the unknown structural dependencies on p.

[4] It can be shown that $\mathfrak{h}_{y,i}$ and $\mathfrak{h}_{u,j}$ are polynomial functions of $(\mathfrak{a}_i^o, \ldots, \mathfrak{d}_i^o)$ and due to the assumed asymptotic stability they decay to the zero function.

5.3.1 GP regression model

The covariance on the noise e is assumed to be diagonal, i.e., $\Sigma_{\mathtt{e}} = \operatorname{diag}\big([\sigma_{\mathtt{e}_1}^2 \ \cdots \ \sigma_{\mathtt{e}_{n_{\mathbb{Y}}}}^2]\big)$. Hence, the $\nu-$th output channel of (5.15) can be written as:

$$\big[y(t)\big]_\nu = \sum_{\iota=1}^{n_{\mathbb{Y}}}\sum_{i=1}^{\infty}\big[\mathfrak{h}_{y,i}(p,t)\big]_{\nu,\iota}q^{-i}\big[y(t)\big]_\iota + \sum_{\iota=1}^{n_{\mathbb{U}}}\sum_{i=1}^{\infty}\big[\mathfrak{h}_{u,i}(p,t)\big]_{\nu,\iota}q^{-i}\big[u(t)\big]_\iota + \big[e(t)\big]_\nu, \tag{5.16}$$

where $\big[\mathfrak{h}_{y,i}(\cdot)\big]_{\nu,\iota}$ denotes the (ν,ι)-th element of the matrix function $\mathfrak{h}_{y,i}(\cdot)$ and $\big[y(t)\big]_\nu$ is the ν-th element of the vector $y(t)$. Eq. (5.16) can be written as

$$\big[y(t)\big]_\nu = \underbrace{\underbrace{\sum_{\iota=1}^{n_{\mathbb{Y}}}\mathfrak{f}_{\nu,\iota}^{\mathrm{y}}\left(x^{(t)}\right)}_{\mathfrak{f}_\nu^{\mathrm{y}}} + \underbrace{\sum_{\iota=1}^{n_{\mathbb{U}}}\mathfrak{f}_{\nu,\iota}^{\mathrm{u}}\left(x^{(t)}\right)}_{\mathfrak{f}_\nu^{\mathrm{u}}}}_{\mathfrak{f}_\nu = \left[\hat{y}(t|t-1)\right]_\nu} + \big[e(t)\big]_\nu, \tag{5.17}$$

where $\mathfrak{f}_\nu^{\mathrm{y}}$, $\mathfrak{f}_\nu^{\mathrm{u}}$ represent the sub-predictors that form the one-step-ahead predictor $\mathfrak{f}_\nu$, and

$$\begin{aligned}\mathfrak{f}_{\nu,\iota}^{\mathrm{y}}\left(x^{(t)}\right) &= \sum_{i=1}^{\infty}\big[\mathfrak{h}_{y,i}(p,t)\big]_{\nu,\iota}q^{-i}\big[y(t)\big]_\iota,\\ \mathfrak{f}_{\nu,\iota}^{\mathrm{u}}\left(x^{(t)}\right) &= \sum_{i=1}^{\infty}\big[\mathfrak{h}_{u,i}(p,t)\big]_{\nu,\iota}q^{-i}\big[u(t)\big]_\iota,\end{aligned} \tag{5.18}$$

which, under the stability assumption of the data-generating system, represent convergent IIRs with $\mathfrak{f}_{\nu,\iota}^{\mathrm{y}}(\cdot) : \mathbb{P}\times\ldots\times\mathbb{P}\times\mathbb{Y}\times\ldots\times\mathbb{Y}\to\mathbb{R}$ and $\mathfrak{f}_{\nu,\iota}^{\mathrm{u}}(\cdot) : \mathbb{P}\times\ldots\times\mathbb{P}\times\mathbb{U}\times\ldots\times\mathbb{U}\to\mathbb{R}$. It is worth to remind the reader that $x^{(t)} = \left\{u^{(t)}, p^{(t)}, y^{(t-1)}\right\}$ is the shorthand notation of the past measurements till time t, e.g., $u^{(t)} = \{u(k)\}_{k\le t}$.

From (5.17), the identification of the one-step-ahead predictor $\mathfrak{f}_\nu$ can be considered as a standard GP regression model. More specifically, by following the Bayesian setting within the GP framework detailed in Chapter 3, $\mathfrak{f}_{\nu,\iota}^{\mathrm{y}}(\cdot), \mathfrak{f}_{\nu,\iota}^{\mathrm{u}}(\cdot)$ are assumed to be a particular realization from a zero-mean Gaussian random field, i.e.,

$$\mathfrak{f}_{\nu,\iota}^{\mathrm{y}}(\cdot) \sim \mathcal{GP}\left(0, K_{\nu,\iota}^{\mathrm{y}}\right), \quad \mathfrak{f}_{\nu,\iota}^{\mathrm{u}}(\cdot) \sim \mathcal{GP}\left(0, K_{\nu,\iota}^{\mathrm{u}}\right), \tag{5.19}$$

respectively, where $\mathfrak{f}_{\nu,\iota}^{\mathrm{y}}(\cdot), \mathfrak{f}_{\nu,\iota}^{\mathrm{u}}(\cdot)$ can be completely defined by their covariances $K_{\nu,\iota}^{\mathrm{y}}, K_{\nu,\iota}^{\mathrm{u}}$. In the Bayesian setting, these covariance functions encode the prior knowledge and assumptions about the to be estimated functional dependency. Hence, in order to have a successful identification, the kernel function needs to be appropriately designed for the problem at hand.

5.3.2 Kernel design for LPV-subpredictors

First of all, within the LPV framework, the relation between the input and the output is assumed to be linear, but with coefficients $\mathfrak{a}_i^{\mathrm{o}}, \ldots, \mathfrak{d}_i^{\mathrm{o}}$ in (5.5)-(5.7) that are

dependent on the scheduling variable p. In many situations, the functional dependencies consist of a p-independent (LTI) part and a p-dependent part, which should be represented in the kernel. In addition, the kernel should explicitly include the stability of the one-step-ahead predictor. To conclude, the kernel functions $K^{\mathrm{y}}_{\nu,\iota}, K^{\mathrm{u}}_{\nu,\iota}$ should be parameterized to

B1 Describe possible structural dependencies on p.

B2 Encode asymptotic stability of the predictor.

B3 Take the LTI part into account.

Next, we show how to design a kernel satisfying B1-B3 to identify the MIMO LPV-BJ system by employing the one-step-ahead predictor (5.15). From (5.17) and under the GP prior (5.19) of the IRR (5.18), for output channel ν, we collect the data in the vector $Y_\nu = \left[[y(1)]_\nu \cdots [y(N)]_\nu\right]^\top$. Under the assumption that $\mathcal{E}\left\{\mathfrak{f}^{\mathrm{y}}_{\nu,\iota}\mathfrak{f}^{\mathrm{u}}_{\nu,\iota'}\right\} = 0$ for all $\nu, \iota = 1, \ldots, n_{\mathbb{Y}}$ and $\iota' = 1, \ldots, n_{\mathbb{U}}$, the covariance of the output channel ν is given by $\mathcal{E}\left\{Y_\nu Y_\nu^\top\right\}$ and its $(\mathfrak{i},\mathfrak{j})-$th entry is described as follows:

$$\mathcal{E}\left\{\left[y(\mathfrak{i})\right]_\nu\left[y(\mathfrak{j})\right]_\nu\right\} = \sum_{\iota=1}^{n_{\mathbb{Y}}} K^{\mathrm{y}}_{\nu,\iota}\left(x^{(\mathfrak{i})}, x^{(\mathfrak{j})}\right) + \sum_{\iota=1}^{n_{\mathbb{U}}} K^{\mathrm{u}}_{\nu,\iota}\left(x^{(\mathfrak{i})}, x^{(\mathfrak{j})}\right) + \sigma^2_{\mathtt{e}_\nu}. \tag{5.20}$$

where $K^{\mathrm{y}}_{\nu,\iota}$ is defined as[5]

$$\begin{aligned} K^{\mathrm{y}}_{\nu,\iota}\left(x^{(\mathfrak{i})}, x^{(\mathfrak{j})}\right) &= \mathcal{E}\left\{\mathfrak{f}^{\mathrm{y}}_{\nu,\iota}\left(x^{(\mathfrak{i})}\right)\mathfrak{f}^{\mathrm{y}}_{\nu,\iota}\left(x^{(\mathfrak{j})}\right)\right\} \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left(\left[y(\mathfrak{i}-i)\right]_\iota \mathcal{Q}^{\mathrm{y}}_{\nu,\iota}\left(p^{(\mathfrak{i},i)}, p^{(\mathfrak{j},j)}\right)\left[y(\mathfrak{j}-j)\right]_\iota\right). \end{aligned} \tag{5.21}$$

In (5.21), $p^{(\mathfrak{i},i)}$ being the vector of past scheduling values starting from time $\mathfrak{i}$ till $\mathfrak{i}-i$, i.e., $p^{(\mathfrak{i},i)} = \left[p^\top(\mathfrak{i})\cdots p^\top(\mathfrak{i}-i)\right]^\top$ and

$$\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}\left(p^{(\mathfrak{i},i)}, p^{(\mathfrak{j},j)}\right) = \mathcal{E}\left\{\left[\mathfrak{h}_{y,i}(p,\mathfrak{i})\right]_{\nu,\iota}\left[\mathfrak{h}_{y,j}(p,\mathfrak{j})\right]_{\nu,\iota}\right\}.$$

It is obvious that we need to parameterize the kernel $\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}$ to encode the prior knowledge, i.e., B1-B3. Interestingly, due to the linearity of the addressed system class, ideas of kernel design for LTI systems; e.g., DI kernel (Chen et al. 2012), TC kernel (Pillonetto and De Nicolao 2010), etc., can be extended to the considered setting in this chapter. More specifically, in the following, we show how to design a DI-like and a TC-like kernel for LPV systems.

To this end, let us analyze the prior knowledge that is needed to be encoded into the kernel function in more details. First, to describe the underlying structural dependency on p represented in terms of the matrix coefficient functions $\mathfrak{h}_{y,i}, \mathfrak{h}_{u,j}$, i.e., Item B1, any positive definite kernel, e.g., polynomial, spline, etc., can be used.

[5]Note that $K^{\mathrm{u}}_{\nu,\iota}$ is defined in a similar fashion.

However, the appropriate choice is problem-dependent. In our case, $\mathfrak{h}_{y,i}, \mathfrak{h}_{u,j}$, are smooth matrix coefficient functions and hence the RBF kernel can be used efficiently to describe such a dependency. Secondly, to encode the asymptotic stability of the predictor or equivalently to guarantee the convergence of the estimated IIR, i.e., Item B2, a decay term should be included that models the vanishing influence of the past input-scheduling-output pairs on the predicted output, i.e., the effect of $\{y(k), u(k), p(k)\}$ over $y(t)$ decreases as $t - k \to \infty$. Thirdly, to take the LTI part into account, i.e., Item B3, the kernel function should be composed of two parts, namely a part to describe the LTI dynamics and a part to describe the $p-$dependent dynamics, e.g., the RBF. In view of the above discussion, a general formulation of a kernel function that encodes the prior knowledge about the underlying IIR $\mathfrak{f}^{\mathrm{y}}_{\nu,\iota}(\cdot)$ is

$$\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}\left(p^{(\mathrm{i},i)}, p^{(\mathrm{j},j)}\right) = \underbrace{\mathcal{Q}^{\mathrm{y,lin}}_{\nu,\iota}(i,j)}_{\text{linear part}} + \underbrace{\mathcal{Q}^{\mathrm{y,p}}_{\nu,\iota}\left(p^{(\mathrm{i},i)}, p^{(\mathrm{j},j)}\right)}_{p\text{-dependent part}}, \tag{5.22}$$

where

$$\mathcal{Q}^{\mathrm{y,lin}}_{\nu,\iota}(i,j) = \beta_1 r_1(\beta_2), \tag{5.23a}$$

$$\mathcal{Q}^{\mathrm{y,p}}_{\nu,\iota}\left(p^{(\mathrm{i},i)}, p^{(\mathrm{j},j)}\right) = \beta_3 r_2(\beta_4) \exp\left(-\frac{\left\|p^{(\mathrm{i},i)} - p^{(\mathrm{j},j)}\right\|_2^2}{\left[\beta^{\mathrm{y}}_{\mathrm{w}}(i,j)\right]^2_{\nu,\iota}}\right), \tag{5.23b}$$

with β_1, β_3 being scaling parameters, $r_1(\beta_2), r_2(\beta_4) \to 0$ as $i, j \to \infty$ to describe the decay of the expansion coefficient, i.e., to ensure that the IIR is convergent. The RBF part describes the possible structural dependency on p, where $\left[\beta_{\mathrm{w}}(i,j)\right]_{\nu,\iota}$ is the width of the RBF.

Due to the assumed stability of the resulting one-step-ahead predictor, the IIRs of the associated sub-predictors $\mathfrak{f}^{\mathrm{y}}_{\nu,\iota}(\cdot)$, $\mathfrak{f}^{\mathrm{u}}_{\nu,\iota}(\cdot)$ in (5.17) asymptotically decay to zero and the high-order terms of the expansion become insignificant. Hence, the one-step-ahead predictor can be arbitrary well approximated by truncating the corresponding infinite sum. The truncated one-step-ahead predictor for channel ν is given by

$$\bar{\mathfrak{f}}_\nu = \underbrace{\sum_{\iota=1}^{n_{\mathbb{Y}}} \bar{\mathfrak{f}}^{\mathrm{y}}_{\nu,\iota}(\cdot)}_{\bar{\mathfrak{f}}^{\mathrm{y}}_\nu} + \underbrace{\sum_{\iota=1}^{n_{\mathbb{U}}} \bar{\mathfrak{f}}^{\mathrm{u}}_{\nu,\iota}(\cdot)}_{\bar{\mathfrak{f}}^{\mathrm{u}}_\nu}, \tag{5.24}$$

with

$$\begin{aligned}
\bar{\mathfrak{f}}^{\mathrm{y}}_{\nu,\iota}(\cdot) &= \sum_{i=1}^{n_{\mathfrak{f}_{\mathrm{y}}}} \left[\mathfrak{h}_{y,i}(p,t)\right]_{\nu,\iota} q^{-i} \left[y(t)\right]_\iota, \\
\bar{\mathfrak{f}}^{\mathrm{u}}_{\nu,\iota}(\cdot) &= \sum_{i=1}^{n_{\mathfrak{f}_{\mathrm{u}}}} \left[\mathfrak{h}_{u,i}(p,t)\right]_{\nu,\iota} q^{-i} \left[u(t)\right]_\iota,
\end{aligned} \tag{5.25}$$

where $n_{\mathfrak{f}_{\mathrm{y}}}$ and $n_{\mathfrak{f}_{\mathrm{u}}}$ are large enough to capture the dominant dynamics of the sys-

tem. As a result, the covariance function (5.21) can be accordingly truncated to a finite order as

$$\bar{K}^{\mathrm{y}}_{\nu,\iota}\left(\bar{x}^{(\mathrm{i})},\bar{x}^{(\mathrm{j})}\right)=\mathcal{E}\left\{\bar{\mathfrak{f}}^{\mathrm{y}}_{\nu,\iota}\left(\bar{x}^{(\mathrm{i})}\right)\bar{\mathfrak{f}}^{\mathrm{y}}_{\nu,\iota}\left(\bar{x}^{(\mathrm{j})}\right)\right\}$$
$$=\sum_{i=1}^{n_{\mathfrak{f}_{\mathrm{y}}}}\sum_{j=1}^{n_{\mathfrak{f}_{\mathrm{y}}}}\left(\left[y(\mathrm{i}-i)\right]_{\iota}\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}\left(p^{(\mathrm{i},i)},p^{(\mathrm{j},j)}\right)\left[y(\mathrm{j}-j)\right]_{\iota}\right),\quad(5.26)$$

where $\bar{x}^{(\mathrm{i})}$ is the set of truncated past measurements, i.e., $\bar{x}^{(\mathrm{i})}=\left\{u^{(\mathrm{i},n_{\mathfrak{f}_{\mathrm{u}}})},p^{(\mathrm{i},n_{\mathfrak{f}})},\right.$ $\left.y^{(\mathrm{i},n_{\mathfrak{f}_{\mathrm{y}}})}\right\}$ and $n_{\mathfrak{f}}=\max(n_{\mathfrak{f}_{\mathrm{y}}},n_{\mathfrak{f}_{\mathrm{u}}})$ is the maximum truncation order. It is worth to remind that the truncated covariance $\bar{K}^{\mathrm{u}}_{\nu,\iota}\left(\bar{x}^{(\mathrm{i})},\bar{x}^{(\mathrm{j})}\right)$ is defined similarly to (5.26), but with truncation order $n_{\mathfrak{f}_{\mathrm{u}}}$.

In this case, i.e., truncated kernel representation, for the output channel ν the hyperparameters consists of the following items:

- $n_{\mathbb{Y}}(n_{\mathfrak{f}_{\mathrm{y}}}+4)$ kernel parameters of the output side sub-predictor. For each of the $n_{\mathbb{Y}}$ IRRs, $n_{\mathfrak{f}_{\mathrm{y}}}+4$ parameters are needed. Specifically, four parameters are needed to characterize the decay terms, i.e., two of them for the LTI part and the other two for the p-dependent part, in addition to the $n_{\mathfrak{f}_{\mathrm{y}}}$ characteristic length parameters that characterize the width of the RBF kernel.
- $n_{\mathbb{U}}(n_{\mathfrak{f}_{\mathrm{u}}}+4)$ kernel parameters of the input side sub-predictor., where we have $n_{\mathbb{U}}$ number of sub impulse responses associated with inputs. The details are the same as in the above point.
- The noise variance if considered as a hyperparameter[6].

As a result, the total number of hyperparameters is

$$\underbrace{\left(\overbrace{n_{\mathbb{Y}}\underbrace{(n_{\mathfrak{f}_{\mathrm{y}}}+4)}_{\text{For each sub IIR}}}^{\text{Output related sub IRRs}}+\overbrace{n_{\mathbb{U}}\underbrace{(n_{\mathfrak{f}_{\mathrm{u}}}+4)}_{\text{For each sub IIR}}}^{\text{Input related sub IIRs}}+\underbrace{1}_{\text{Noise variance}}\right)n_{\mathbb{Y}}}_{\text{For each output}},\quad(5.27)$$

which grows rapidly in $n_{\mathbb{Y}},n_{\mathbb{U}},n_{\mathfrak{f}_{\mathrm{y}}}$, and $n_{\mathfrak{f}_{\mathrm{u}}}$, potentially leading to a computational problems. However, further assumptions can be made to reduce the number of hyperparameters:

Assumption 5.1 *For the output channel ν: all the sub IIRs associated with the sub-predictor $\mathfrak{f}^{\mathrm{y}}_{\nu}$, i.e., $\mathfrak{f}^{\mathrm{y}}_{\nu,\iota}(\cdot)$ for $\iota=1,\ldots,n_{\mathbb{Y}}$, share the same decay rate, i.e., they share the same parameterization for r_1,r_2, in* (5.23) *with different scaling parameters. The same assumption can be taken for the IIRs associated with the sub-predictor $\mathfrak{f}^{\mathrm{u}}_{\nu}$, i.e., i.e., $\mathfrak{f}^{\mathrm{u}}_{\nu,\iota}(\cdot)$ for $\iota=1,\ldots,n_{\mathbb{U}}$.*

[6]Another possibility is to identify a high order ARX or FIR model and then use the sample variance of the residual as an estimate of the noise variance. However, this is more complicated in the LPV case, as usually it is not known what type of dependency structure should be considered for these models.

Assumption 5.2 *For every IIR* $\mathfrak{f}^{\mathrm{y}}_{\nu,\iota}(\cdot)$*, the kernel width is assumed to be the same for all coefficient functions within this IIR, i.e.,* $\left[\beta^{\mathrm{y}}_{\mathrm{w}}(i,j)\right]^2_{\nu,\iota}$ *are the same for all* i, j*, in* (5.23)*. The same holds true for* $\mathfrak{f}^{\mathrm{u}}_{\nu,\iota}(\cdot)$*.*

Under Assumptions 5.1-5.2, the total number of hyperparameters needed to be estimated from data is reduced to

$$\big(3\left(n_{\mathbb{Y}}+n_{\mathbb{U}}\right)+5\big)n_{\mathbb{Y}}. \tag{5.28}$$

For example, in case $n_{\mathbb{Y}}=2, n_{\mathbb{U}}=2, n_{\mathfrak{f}_{\mathrm{y}}}=n_{\mathfrak{f}_{\mathrm{u}}}=10$, the total number of the hyperparameters that are needed to be estimated is 114. However, by following Assumptions 5.1, 5.2, this number is reduced to 34. Based on the above considerations, now we have two different scenarios w.r.t. the relation between the coefficient functions associated with different time instants: i) being uncorrelated, a DI-like representation of the resulting kernel in (5.22) can be realized with the following choice

$$\begin{aligned}\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}\left(p^{(\mathrm{i},i)},p^{(\mathrm{j},j)}\right)=\left[\beta_1\right]_{\nu,\iota}\left(\left[\beta_2\right]_{\nu}\right)^{i}\delta_{i,j}+\\ \left[\beta_3\right]_{\nu,\iota}\left(\left[\beta_4\right]_{\nu}\right)^{i}\exp\left(-\frac{\left\|p^{(\mathrm{i},i)}-p^{(\mathrm{j},j)}\right\|_2^2}{\left[\beta^{\mathrm{y}}_{\mathrm{w}}(i,j)\right]^2_{\nu,\iota}}\right)\delta_{i,j},\end{aligned} \tag{5.29}$$

where $\delta_{i,j}$ is the Kronecker delta function w.r.t. (i,j); ii) being correlated, a TC-like modification of the above given kernel can be also given to take the correlation between the coefficient functions associated with different time instants into account:

$$\begin{aligned}\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}\left(p^{(\mathrm{i},i)},p^{(\mathrm{j},j)}\right)=\left[\beta_1\right]_{\nu,\iota}\left(\left[\beta_2\right]_{\nu}\right)^{\max(i,j)}\\ +\left[\beta_3\right]_{\nu,\iota}\left(\left[\beta_4\right]_{\nu}\right)^{\max(i,j)}\exp\left(-\frac{\left\|p^{(\mathrm{i},i)}-p^{(\mathrm{j},j)}\right\|_2^2}{\left[\beta^{\mathrm{y}}_{\mathrm{w}}(i,j)\right]^2_{\nu,\iota}}\right),\end{aligned} \tag{5.30}$$

where $\left[\beta_1\right]_{\nu,\iota},\left[\beta_3\right]_{\nu,\iota}$ are scaling parameters of the LTI and the $p-$dependent part of the $(\nu,\iota)-$th sub IIR, respectively, and $\left[\beta_2\right]_{\nu},\left[\beta_4\right]_{\nu}$ are the parameters that determine the decay rate of the IIRs associated with the ν-th output channel. See Figure 5.2 to visualize the difference between the DI and TC kernels in terms of representing the correlation between coefficient functions associated with different time instants. The left part of Figure 5.2 displays a scaled image of a kernel matrix constructed with the DI kernel, where it can be easily seen that only the diagonal elements are nonzeros, i.e., the elements associated with different time instants are assumed to be uncorrelated, whereas the right part of the figure display the scaled image in case of TC kernel, where not only the diagonal entries, but also the off-diagonal entries are nonzero, providing the kernel with the ability to express the correlation between functions associated with different time instants.

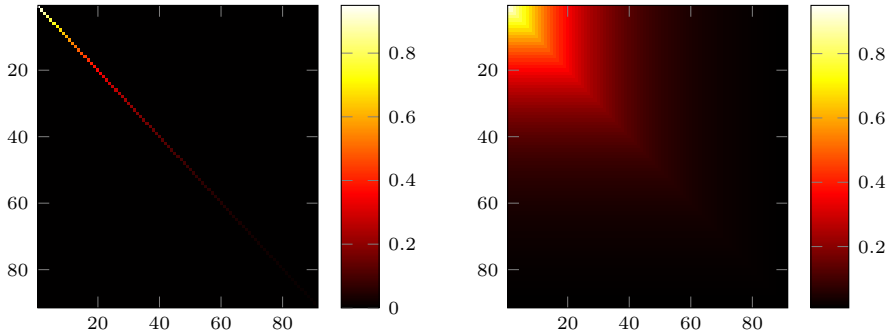


Figure 5.2: Scaled image of a kernel matrix constructed with: Left part: DI kernel. Right part: TC kernel. Note that the resulting image is an $m \times n$ grid of pixels where m and n are the number of columns and rows of the kernel matrix, respectively. Each element of kernel matrix specifies the color for a pixel of the image according to the color map shown on the right of each figure.

5.3.3 Estimation of the predictor from data

The last remaining item to discuss is the estimation of the predictor $\mathfrak{f}_\nu$ in (5.17) by the truncated model (5.24) from a given data set $\mathcal{D}_N = \{y(t), u(t), p(t)\}_{t=1}^{N}$. This is accomplished within the Gaussian regression framework, introduced in Chapter 3. First, let us define β to denote the vector of unknown hyperparameters related to the output channel ν. Furthermore, let $Y = \left[y^\top(1) \cdots y^\top(N)\right]^\top, Y' = \left[y^\top(n_\mathfrak{f}+1) \cdots y^\top(N)\right]^\top, U = \left[u^\top(1) \cdots u^\top(N)\right]^\top, Y'_\nu = \left[[y(n_\mathfrak{f}+1)]_\nu \cdots [y(N)]_\nu\right]^\top$, and $P = \left[p^\top(1) \cdots p^\top(N)\right]^\top$.

The minimum variance estimate of the predictor for output channel ν, i.e., $\bar{\mathfrak{f}}_\nu$ in (5.24) conditioned on a fixed β can be written as

$$\hat{\mathfrak{f}}_\nu(\bullet) = \mathcal{E}\left\{\bar{\mathfrak{f}}_\nu(\bullet) \mid Y, U, P, \beta\right\} = \sum_{t=n_\mathfrak{f}+1}^{N} c_{t-n_\mathfrak{f}} \bar{K}_\nu\left(\bullet, \bar{x}^{(t)}\right), \tag{5.31}$$

$$\bar{K}_\nu\left(\bullet, \bar{x}^{(t)}\right) = \sum_{\iota=1}^{n_\mathbb{Y}} \bar{K}^{\mathrm{y}}_{\nu,\iota}\left(\bullet, \bar{x}^{(t)}\right) + \sum_{\iota=1}^{n_\mathbb{U}} \bar{K}^{\mathrm{u}}_{\nu,\iota}\left(\bullet, \bar{x}^{(t)}\right),$$

where $\bar{x}^{(t)}$ is the set of truncated past measurements, i.e., $\bar{x}^{(t)} = \left\{u^{(t,n_\mathfrak{f})}, p^{(t,n_\mathfrak{f})}, y^{(t,n_\mathfrak{f})}\right\}$ and $c_{t-n_\mathfrak{f}}$ is the $(t-n_\mathfrak{f})$-th component of the vector

$$c = \left(\Sigma_\nu(\beta)\right)^{-1} Y'_\nu,$$

with $\Sigma_\nu(\beta) \in \mathbb{R}^{N-n_\mathfrak{f} \times N-n_\mathfrak{f}}$ being invertible[7] and given by

$$\left[\Sigma_\nu(\beta)\right]_{i,j} = \bar{K}_\nu\left(\bar{x}^{(n_\mathfrak{f}+i)}, \bar{x}^{(n_\mathfrak{f}+j)}\right) + \sigma_{\mathtt{e}_\nu}^2 \delta_{i,j}.$$

Now the minimum variance estimate obtained in (5.31) is conditioned on a fixed value of the hyperparameters vector β. In this work, we follow the approach of maximizing the marginal likelihood of the output w.r.t. β (MacKay 2003). More specifically, the log-marginal likelihood of the observations Y'_ν given β:

$$\log \mathrm{p}(Y'_\nu \mid U, P, \beta) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}Y_\nu'^{\top}\left(\Sigma_\nu(\beta)\right)^{-1}Y'_\nu - \frac{1}{2}\log\det\left(\Sigma_\nu(\beta)\right). \quad (5.32)$$

Then, an estimate for β is obtained by maximizing the marginal likelihood or equivalently

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \; -\log \mathrm{p}(Y'_\nu \mid U, P, \beta). \quad (5.33)$$

According to the empirical Bayes approach (Carlin and Louis 2000), the minimum variance estimate of the predictor, i.e., $\hat{\mathfrak{f}}_\nu(\bullet)$ in (5.31), is obtained by substituting the optimized $\hat{\beta}$ from (5.33). Moreover, such an estimate is associated with a probability level or confidence region which provide a quantification for the quality of the estimate via computing the variance of the prediction by applying (3.23)

$$\operatorname{cov}\left(\hat{\mathfrak{f}}_\nu(\bullet) \mid Y, U, P, \beta\right) = \bar{K}_\nu(\bullet,\bullet) - \kappa_\nu^{\top}\left[\Sigma_\nu(\hat{\beta})\right]^{-1}\kappa_\nu, \quad (5.34)$$

where $\kappa_\nu = \left[\bar{K}_\nu\left(\bullet, \bar{x}^{(n_\mathfrak{f}+1)}\right) \cdots \bar{K}_\nu\left(\bullet, \bar{x}^{(N)}\right)\right]^{\top}$.

5.3.4 Reconstruction of the individual coefficient functions

In the following, we illustrate how to recover the individual coefficient functions from the estimated one-step-ahead predictor.

By remembering (5.31) and the definition of $\bar{K}^{\mathrm{y}}_{\nu,\iota}\left(\bar{x}^{(\mathrm{i})}, \bar{x}^{(\mathrm{j})}\right)$ in (5.26). From (Pillonetto et al. 2011b, Theorem 4), it can be seen that the kernels $\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}$, $\mathcal{Q}^{\mathrm{u}}_{\nu,\iota}$ induce mutually orthogonal subspaces $\mathscr{H}_{\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}}$, $\mathscr{H}_{\mathcal{Q}^{\mathrm{u}}_{\nu,\iota}}$, where, $\mathscr{H}_{\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}}$, $\mathscr{H}_{\mathcal{Q}^{\mathrm{u}}_{\nu,\iota}}$, denote the associated RKHSs with $\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}$, $\mathcal{Q}^{\mathrm{u}}_{\nu,\iota}$, respectively. As a result, the minimum variance estimate of the individual coefficient functions, i.e., $[\hat{\mathfrak{h}}_{y,i}]_{\nu,\iota}, [\hat{\mathfrak{h}}_{u,j}]_{\nu,\iota}$, can be obtained as the orthogonal projection of $\hat{\mathfrak{f}}_\nu(\bullet) \in \mathscr{H}_{\bar{K}_\nu}$, where $\mathscr{H}_{\bar{K}_\nu}$ is the RKHS associated with $\bar{K}_\nu$, onto $\mathscr{H}_{\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}}$, $\mathscr{H}_{\mathcal{Q}^{\mathrm{u}}_{\nu,\iota}}$, respectively, as follows:

$$\left[\hat{\mathfrak{h}}_{y,i}(\bullet)\right]_{\nu,\iota} = \sum_{k=n_\mathfrak{f}+1}^{N} c_{k-n_\mathfrak{f}}\left[y(k-i)\right]_\iota \mathcal{Q}^{\mathrm{y}}_{\nu,\iota}\left(\bullet, p^{(k,i)}\right), \quad (5.35)$$

[7] Since the corresponding kernel function is positive semidefinite, the resulting Σ_ν becomes a symmetric and positive definite kernel matrix. Hence, its inverse exists.

and the corresponding covariance estimate is given by

$$\operatorname{cov}\left(\left[\hat{\mathfrak{h}}_{y,i}(\bullet)\right]_{\nu,\iota}\right) = \mathcal{Q}^{\mathrm{y}}_{\nu,\iota}(\bullet,\bullet) - \left(\mathcal{K}^{\mathrm{y}}_{\nu,\iota}\right)^{\top}\left(\Sigma_{\nu}(\beta)\right)^{-1}\mathcal{K}^{\mathrm{y}}_{\nu,\iota}, \tag{5.36}$$

where

$$\mathcal{K}^{\mathrm{y}}_{\nu,\iota} = \left[\mathcal{Q}^{\mathrm{y}}_{\nu,\iota}\left(\bullet, p^{(n_{\mathfrak{f}}+1,i)}\right)\left[y(n_{\mathfrak{f}}-i+1)\right]_{\iota} \cdots \mathcal{Q}^{\mathrm{y}}_{\nu,\iota}\left(\bullet, p^{(N,i)}\right)\left[y(N-i)\right]_{\iota}\right]^{\top}. \tag{5.37}$$

Such a covariance estimate provides a quantification of the uncertainties of the estimated coefficient functions by highlighting the regions that suffer from poor excitation. Hence, such information can be used to further improve the estimate. The minimum variance estimate of $\hat{\mathfrak{h}}_{u,i}(\bullet)$ and its associated covariance can be formulated in a similar fashion.

It is interesting to mention that the obtained nonparametric estimates of the one-step-ahead predictor, i.e., $\hat{\mathfrak{h}}_{y,i}, \hat{\mathfrak{h}}_{u,i}$ in (5.35), can be utilized to obtain a nonparametric estimates of the process and noise dynamics G_0, H_0, respectively. More specifically, a nonparametric estimate of $\mathfrak{g}^{\mathrm{o}}_k, \mathfrak{h}^{\mathrm{o}}_k$ in (5.6b)-(5.8b), respectively, can be realized via recursive relations that depend on the identified $\hat{\mathfrak{h}}_{y,i}, \hat{\mathfrak{h}}_{u,i}$ see (Darwish et al. 2017b, Section 5).

5.3.5 Numerical simulation

In this section, the performance of the presented nonparametric approach for the identification of LPV-BJ models based on their one-step-ahead predictor is shown by means of an extensive Monte-Carlo study.

Data-generating system

The considered data-generating system is a MIMO system with $n_{\mathbb{U}} = 2$, $n_{\mathbb{Y}} = 2$ and $n_{\mathbb{P}} = 2$ in the form of (5.4). The LPV-BJ data-generating system has a plant model order of $n_{\mathrm{a}} = n_{\mathrm{b}} = 2$ and a noise model order of $n_{\mathrm{c}} = n_{\mathrm{d}} = 2$. The matrix polynomials associated with the plant and noise models are given in details in Appendix B.

Identification setting

The one-step-ahead predictor is estimated using an identification data set with three different sizes $N = \{200, 500, 1000\}$ and the prediction performance of the estimated model is examined on a validation data set that contains $N_{\mathrm{val}} = 200$ samples. The identification and validation data sets are generated with independent realizations of a white noise input signal u with uniform distribution, i.e.,

$\big[u(t)\big]_\nu \sim \mathcal{U}(-1,1),\ \nu = 1,2$. The scheduling signals are given by

$$\big[p(t)\big]_\nu = 0.4\sin(0.035t + \frac{\nu\pi}{5}) + 0.25\nu + \mathcal{U}(-0.15, 0.15),\ \text{for } \nu = 1,2. \tag{5.38}$$

The variance of the white noise e driving the noise process is chosen such that the SNR ratio

$$\text{SNR}_{[y]_\nu} = 10\log\frac{\sum_{t=1}^{N}\big[y(t)\big]_\nu^2}{\sum_{t=1}^{N}\big[v(t)\big]_\nu^2},$$

is 20dB. To analyze the statistical properties of the presented identification approach, a Monte-Carlo study with $N_{\text{MC}} = 100$ runs is carried out. At each run, a new realization of the input u, the scheduling signal p and the noise e are taken.

The predicted output $\hat{y}$ from the estimated one-step-ahead predictor model is compared to the true output of the data-generating system by the BFR

$$\text{BFR} = \max\left(1 - \frac{\frac{1}{N}\sum_{t=1}^{N}\big\|y(t) - \hat{y}(t)\big\|_2}{\frac{1}{N}\sum_{t=1}^{N}\big\|y(t) - \bar{y}\big\|_2}, 0\right)\cdot 100\%, \tag{5.39}$$

where $\bar{y}$ defines the mean of the true output $y(t)$. Note that the definition in (5.39) characterizes the average performance over all output channels.

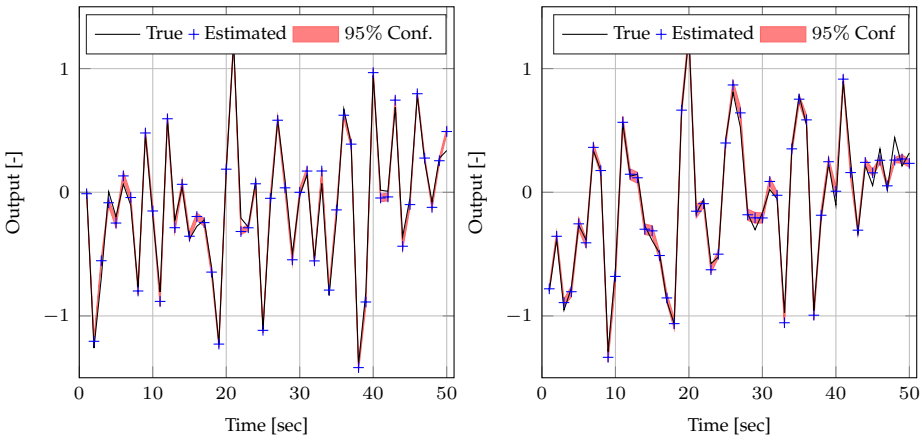
Identification results

In this section, the results of the identification of the one-step-ahead predictor of the data-generating system given in Section 5.3.5 under the identification setting detailed in Section 5.3.5 are discussed. The results have been obtained with a truncation order $n_{\mathfrak{f}_\text{y}} = n_{\mathfrak{f}_\text{u}} = n_\mathfrak{f} = 10$. The considered estimators are

1. Bayesian estimator with the DI-like kernel, i.e., (5.29).
2. Bayesian estimator with the TC-like kernel, i.e., (5.30).
3. Oracle estimator that knows the true underlying nonlinear functional dependencies[8] of $\mathfrak{h}_{y,i}, \mathfrak{h}_{u,j}$. With such knowledge, the Oracle estimator performs a LS estimate of a high-order ARX model with a truncation order of $n = 15$, which is chosen large enough to capture the dynamics of the system. It is worth to mention that the number of parameters is $(2n+1)n_\mathbb{Y}$, which in this case gives 62 parameters to be estimated via LS. In this case, the truncation order introduces a bias/variance trade-off which is not the case in the nonparametric estimation. Note that such an Oracle estimator is considered to establish a notion of the "best achievable" performance and it is not applicable in practice as the true underlying nonlinear functional dependencies are not known a priori.

[8] A recursive expressions to calculate such functions based on the true underlying coefficient functions of the process and noise dynamics (5.5)-(5.7) can be found in (Darwish et al. 2017b, Equation 12).

In case of the Bayesian estimator, the hyperparameters are estimated via solving (5.33). Figure 5.3 displays the first 50 samples of one realization of the true and the predicted output response on the validation data set with 95% confidence region for the one-step-ahead predictor estimated with TC-like kernel, truncation order of $n_{\mathfrak{f}} = 10$ and a data set with $N = 1000$ samples for both output channels. The figure shows that the presented Bayesian approach is able to identify an LPV model under the general noise model structure of BJ type. Table 5.1 gives the



(a) Output #1 (b) Output #2

Figure 5.3: The first 50 samples of the true and the predicted output response on the validation data set with 95% confidence region for the one-step-ahead predictor estimated with TC-like kernel, model order of $n_{\mathfrak{f}} = 10$ and a data set with $N = 1000$ samples.

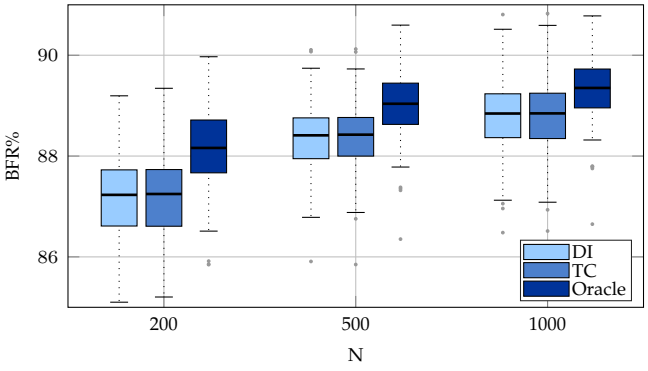


Figure 5.4: The BFR of the predicted response with respect to the validation data sets using the estimated models for the DI, TC kernels and the Oracle estimate, under various sizes of the identification data set $N = \{200, 500, 1000\}$.

mean and the standard deviation (std) of the BFR of the identified predictor over

Table 5.1: Mean and std of the BFR of the identified predictor on the validation data set over $N_{\mathrm{MC}} = 100$ Monte-Carol runs.

	Estimator	BFR [%]	
		Mean	Std
N=200, $n_{\mathfrak{f}}$=10	DI	87.18	0.8222
	TC	87.20	0.8170
	Oracle	88.11	0.8150
N=500, $n_{\mathfrak{f}}$=10	DI	88.36	0.6769
	TC	88.39	0.6804
	Oracle	89.01	0.7054
N=1000, $n_{\mathfrak{f}}$=10	DI	88.80	0.7173
	TC	88.80	0.7227
	Oracle	89.32	0.6828

$N_{\mathrm{MC}} = 100$ runs. The performance criterion is based on the predicted output of the identified predictor on a validation data set. To gain more insights into the performance of the considered estimators, Figure 5.4 gives the distribution of the model fit for different sample sizes shown by boxplots. It can be seen from Table 5.1 and Figure 5.4 that all the predictors benefit from the increasing the amount of samples in the estimation data set, which is evident in the increased BFR. Moreover, the performance of the presented Bayesian approach gets closer to the Oracle by increasing the size of the data set.

5.4 Bayesian identification of LPV series-expansion models

In this section, we investigate how to extend the Bayesian identification technique of LPV systems that has been presented in Section 5.3 to LPV series-expansion models.

5.4.1 LPV series-expansion by OBFs

As an extension of LTI series expansion representation, it can be proven that any DT asymptotically stable LPV system $\mathcal{S}$ has a series-expansion representation in terms of rational OBFs $\breve{\Psi} = \{\breve{\psi}_k\}_{k=1}^{\infty}$, which are the orthonormal basis for $\mathscr{RH}_{2-}(\mathbb{E})$. Such a representation is very attractive and provide a flexible model structure which can represent a given LPV system $\mathcal{S}$ globally on $\mathbb{P}$ (Tóth et al. 2009a). Based on the LTI transfer function theory, a pulse basis function $q^{-i},\ i > 0$ has a unique series-expansion in terms of $\breve{\Psi}$:

$$q^{-i} = \sum_{j=1}^{\infty} w_{i,j} \breve{\psi}_j(q), \tag{5.40}$$

where $w_{i,j} \in \mathbb{R},\ i,j = 1,\dots,\infty$ are the expansion coefficients. It is known that such a series-expansion is convergent. Substituting (5.40) into (5.2), the IO of $\mathcal{S}$, i.e., the process dynamic, in terms of the basis and the associated parameter dependent weighting functions is given by

$$\breve{y}(t) = \mathcal{W}_0(p,t)u(t) + \sum_{k=1}^{\infty} \mathcal{W}_k(p,t)\breve{\psi}_k(q)u(t), \tag{5.41}$$

where $\{\mathcal{W}_k(p,t)\}_{k=0}^{\infty}$ is a set of *coefficient functions* with dynamic dependence on p with $\mathcal{W}_0(p,t) = \mathfrak{g}_0(p,t)$ and

$$\mathcal{W}_k(p,t) = \sum_{l=1}^{\infty} w_{l,k}\mathfrak{g}_l(p,t), \tag{5.42}$$

for $k = 1,\dots,\infty$. It is worth mentioning that the coefficient functions $\mathcal{W}_k(p,t)$ can depend on arbitrary many shifted versions of $p(t)$, i.e., $\{p(t-l)\}_{l=0}^{\infty}$. This representation is known as the (Wiener) LPV-OBFs representation and can be seen as a generalization of (5.2), where general basis functions are used instead of pulse basis. In practice, for asymptotically stable systems it is always possible to find a finite $\breve{\Psi}_{n_\psi} \subset \breve{\Psi}_{\infty}, n_\psi \in \mathbb{N}$, such that the representation error of (5.41) is negligible as $\mathcal{W}_k$ also needs to converge to 0 due to the properties of (5.40) and the stability of the represented relation. Then, $\breve{\Psi}_{n_\psi}$ and their associated weighting functions provide an efficient representation of $\mathcal{S}$ and the IO map of $\mathcal{S}$ can be represented as

$$\breve{y}(t) \approx \mathcal{W}_0(p,t)u(t) + \sum_{k=1}^{n_\psi} \mathcal{W}_k(p,t)\breve{\psi}_k(q)u(t), \tag{5.43}$$

which is the generalization of the LPV-FIR (5.3). Since LPV-OBFs models can be seen as a generalization of LPV-IIR models, in the following, we will focus on the identification of LPV-OBFs models, utilizing a similar approach to what we used in the LTI case.

5.4.2 Parametric identification of LPV-OBFs models

Consider a SISO[9] LPV data-generating system $\mathcal{S}$ as described in Section (5.2.2), where the process dynamics is characterized as a series expansion model in terms of a given set of OBFs $\breve{\Psi}_{\infty} = \{\breve{\psi}_k\}_{k=1}^{\infty}$, which are a complete set of orthonormal basis in $\mathscr{RH}_{2-}(\mathbb{E})$, hence,

$$y(t) = \mathcal{W}_0(p,t)u(t) + \sum_{k=1}^{\infty} \mathcal{W}_k(p,t)\breve{\psi}_k(q)u(t) + e(t), \tag{5.44}$$

[9] In this section, we consider a SISO setting to simplify the discussion. The MIMO extension can be found in Section 5.2.2.

where e is a zero-mean white Gaussian noise process[10] with variance σ_{e}^2, i.e., $e \sim \mathcal{N}\left(0, \sigma_{\mathsf{e}}^2\right)$. It is worth to mention that the case of colored noise can be handled similarly as in Section 5.3.

Given a data set $\mathcal{D}_N = \{u(t), p(t), y(t)\}_{t=1}^N$, our goal is to approximate the underlying system $\mathcal{S}$ as good as possible by the LPV-OBFs model structure in (5.44). This boils down to the choice of the basis $\breve{\psi}_k$ and estimating the $p-$dependent expansion coefficient functions $\mathcal{W}_k$.

In classical LPV identification, given a set of basis functions $\breve{\Psi}_{n_\psi} = \{\breve{\psi}_k\}_{k=1}^{n_\psi}$, identification of LPV systems based on the representation of (5.44) simplifies to the estimation of the scheduling functions, i.e., $\mathcal{W}_k$. Under the assumption that the dependency of $\mathcal{W}_k$ on p is static, we can proceed to tackle the identification task by one of the following methods (Tóth et al. 2009a):

Local approach

The LPV-OBFs model structure also corresponds to the well known fact that an LPV system $\mathcal{S}$ can always be viewed as a collection of "local" behaviors $\mathfrak{F}_{\mathbb{P}} = \{\mathcal{F}_{\bar{\mathsf{p}}}\}_{\bar{\mathsf{p}} \in \mathbb{P}}$, where $\mathcal{S}$ is identical to the LTI system $\mathcal{F}_{\bar{\mathsf{p}}}$ for constant scheduling: $p(t) = \bar{\mathsf{p}} \in \mathbb{P}$, for all $t \in \mathbb{Z}$, and parameter dependent weighting functions $\mathfrak{W}_{\mathbb{P}} = \{\mathcal{W}_{\bar{\mathsf{p}}}(\cdot)\}_{\bar{\mathsf{p}} \in \mathbb{P}}$ that schedule between these local behaviors (Rugh and Shamma 2000). This principle can be used in the following way: As $\mathfrak{F}_{\mathbb{P}}$ corresponds to a subset of the LTI system space, therefore every $\mathcal{F}_{\bar{\mathsf{p}}} \in \mathfrak{F}_{\mathbb{P}}$ can be represented as a linear combination of the orthogonal basis of the LTI system space, denoted by $\breve{\Psi}_\infty = \{\breve{\psi}_k\}_{k=1}^\infty$, as $\mathcal{F}_{\bar{\mathsf{p}}} = \mathcal{W}_0 + \sum_{k=1}^\infty \mathcal{W}_k \breve{\psi}_k(q)$, where $\{\mathcal{W}_k\}_{k=0}^\infty$ is the set of coefficients.

Within the local approach, an LPV model is constructed by interpolating LTI models, i.e., $\mathcal{F}_{\bar{\mathsf{p}}} \in \mathfrak{F}_{\mathbb{P}}$ that are identified around some pre-chosen operating points, with the following LTI-OBFs model structure

$$y(t) = \sum_{k=0}^{n_\psi} r_{\bar{\mathsf{p}},k} \breve{\psi}_k(q) u(t) + e(t). \tag{5.45}$$

First, a functional dependency should be chosen, e.g., polynomial interpolation , where monomials of order n_{m} are employed as basis functions, for instance

$$r_{\bar{\mathsf{p}},k} = \sum_{j=0}^{n_{\mathrm{m}}} \theta_{k,j} \bar{\mathsf{p}}^j, \tag{5.46}$$

where $\theta_{k,j} \in \mathbb{R}$ are the parameters of the polynomials collected into the vector θ. These parameters can then be estimated by minimizing the quadratic loss ℓ_2 of the approximation error or other error measure w.r.t. the locally estimated expansion coefficients. Then, the resulting $r_{\bar{\mathsf{p}},k}$ is interpolated to obtain an estimate of $\{\mathcal{W}_k\}_{k=0}^{n_\psi}$ in (5.44) such that $\mathcal{W}_k(\bar{\mathsf{p}}) = r_{\bar{\mathsf{p}},k}$.

[10] Equation (5.44) corresponds to an OE model structure, where we have an independent parameterization of the process dynamics and the noise dynamics.

Global approach

The global approach aims at estimating, in one step, an LPV model of the considered system based only on a single data set with varying scheduling trajectory, such an estimation approach requires a linear parameterization of each expansion coefficient function, i.e., $\mathcal{W}_k$, in terms of a prior chosen set of basis $\{\phi_k\}_{k=1}^{n_\phi}$, $\phi_k : \mathbb{P} \to \mathbb{R}$

$$\mathcal{W}_k = \sum_{j=0}^{n_\phi} \theta_{k,j}\phi_j. \tag{5.47}$$

The problem then becomes a linear regression based on (5.44) in a PEM setting.

5.4.3 Associated challenges with LPV-OBFs models identification

Although a series-expansion model structure is very flexible offering several advantages, its identification from observed data is not an easy task. Many challenges are associated with the estimation of these models from which some is shared with the IO models considered in Section 5.3:

1. The selection of a suitable set of OBFs, i.e., $\{\breve{\psi}_k\}$;
2. The parameterization of the coefficient functions, i.e., $\{\mathcal{W}_k(p)\}$ to have a good bias/variance trade-off without requiring a lot of prior knowledge;
3. How to guarantee the convergence of the estimated expansion while estimating it, which in fact is required from the considered stability viewpoint;
4. How to deal with the dynamic dependency on the scheduling signal.

These challenges have been partially addressed in Tóth et al. (2009a). More specifically, a basis functions selection scheme, which is a joint application of the *Kolmogrov n-width* (KnW) theory (Oliveira e Silva 1996) and *Fuzzy c-Means* (FcM) clustering (Jain and Dubes 1988), has been proposed that is capable of asymptotically estimate the optimal set of OBFs based on local information. Then, a local or global approach can be followed to identify the coefficient functions by first parameterizing them linearly, e.g., polynomial basis, under the assumption that the functional dependencies are static, and then perform a linear regression in a LS prediction error setting. However, as a starting point, the aforementioned approach assumes that a collection of pole locations is available, obtained from local identification of the LPV system $\mathcal{S}$, which is not always directly feasible, for instance in systems where the output is the scheduling signal, see the DC motor example in Chapter 6. Moreover, the parameterization of the coefficient functions in terms of some basis functions, e.g., polynomial, requires prior knowledge and in many cases a complicated analysis of first-principles based models is needed, where the benefits obtained by data-driven modeling can be easily lost. Moreover, such a selection introduces a bias/variance trade-off, where the selection of the number of basis function is critical. This means that, in a general setting, the above-mentioned challenges still need to be further investigated, which is the topic of the next section.

5.4.4 Bayesian identification of LPV-OBFs models

In the following, we show how the developed approach in Section 5.3 can be extended to identify LPV-OBFs models within the Bayesian setting. Eq. (5.44) can be written as

$$y(t) = \overbrace{\sum_{k=0}^{\infty} \underbrace{\mathcal{W}_k(p,t)u_k^{\mathrm{f}}(t)}_{\mathfrak{f}_k\left(x^{(t)}\right)}}^{\mathfrak{f}\left(x^{(t)}\right)} + e(t), \tag{5.48}$$

where $x^{(t)} = \{u^{(t)}, p^{(t)}\}$ is a shorthand notation of the past measurements till time t, i.e., $u^{(t)} = \{u(k)\}_{k\leq t}$, $p^{(t)} = \{p(k)\}_{k\leq t}$, and $u_k^{\mathrm{f}}(t) = \breve{\psi}_k(q)u(t)$ is the filtered input u through the basis function $\breve{\psi}_k$ and $\mathfrak{f}$ represents a convergent series-expansion. It can be seen that the estimation of LPV-OBFs model from a given data set $\mathcal{D}_N = \{u(t), p(t), y(t)\}_{t=1}^{N}$ can be regarded as a standard GP regression problem (Rasmussen and Williams 2006). More specifically, by assuming that $\mathfrak{f}$ is a particular realization from a zero-mean Gaussian random field that can be completely defined by its covariance K, i.e.,

$$\mathfrak{f} \sim \mathcal{GP}\big(0, K\big), \tag{5.49}$$

which is the prior knowledge in the considered Bayesian setting. It can be easily seen that, from this point on, the design of a suitable kernel function K and the identification of the considered model structure can be performed following the same approach presented in Section 5.3. The main difference is that the basis functions utilized in the considered LPV-OBFs model structure are needed to be estimated. However, within the GPR approach, the generating poles can be considered as additional hyperparameters as we handled them in Section 4.2.3 and hence can be estimated by maximizing the marginal likelihood. In this way, the hyperparameters that parameterize the kernel function can be estimated according to the global approach, where only one data set is needed to do both tuning the hyperparameters and estimating the model. In this way, the basis functions can be estimated without the need to perform local experiments like in the FKcM algorithm, and accordingly the presented approach is suitable when maintaining a constant scheduling is not possible.

5.4.5 Simulation example

In this section, as a simulation example, we consider the following asymptotically stable DT LPV system $\mathcal{S}$ in an IO representation (5.1):

$$\sum_{i=0}^{5} \mathfrak{a}_i^{\mathrm{o}}(p(t))y(t-i) = \mathfrak{b}_1^{\mathrm{o}}(p(t))u(t-1), \tag{5.50}$$

where $p : \mathbb{Z} \to \mathbb{P}$ is the DT scheduling signal with $\mathbb{P} = [0.6, 0.8]$ and $\mathfrak{a}_0^{\mathrm{o}} = 0.58 - 0.1p, \mathfrak{a}_1^{\mathrm{o}}(p) = -\frac{511}{860} - \frac{48}{215}p^2 + 0.3(\cos(p) - \sin(p)), \mathfrak{a}_2^{\mathrm{o}}(p) = \frac{61}{110} - 0.2\sin(p), \mathfrak{a}_3^{\mathrm{o}}(p) =$

$-\frac{23}{85} + 0.2\sin(p), \mathfrak{a}_4^o(p) = \frac{12}{125} - 0.1\sin(p), \mathfrak{a}_5^o(p) = -0.003, \mathfrak{b}_1^o(p) = \cos(p)$.

By using constant scheduling signals with values $\{0.6; 0.6 + \delta; \ldots, 0.8\}$, where $\delta = 0.02$, 11 frozen local LTI representation of $\mathcal{S}$ are obtained, whose pole locations are shown in Figure 5.5. It can be easily seen that $\mathcal{S}$ exhibits significant changes in its dynamics at different constant p values. Our goal is to obtain an accurate but low complexity model for the considered system based on observed data.

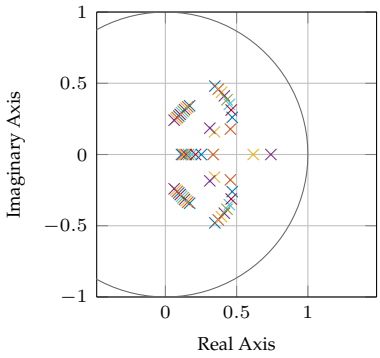


Figure 5.5: Pole locations of the 11 frozen LTI models associated with the LPV system $\mathcal{S}$ in (5.50).

Identification setting

A global approach is followed with a $N = 500$ sample long data record $\mathcal{D}_N$, which has been generated by uniform $u \in \mathcal{U}(-1, 1), p \in \mathcal{U}(0.6, 0.8)$ and with additive, white output noise e, whose variance is chosen such that the SNR ratio is 20dB. To analyze the statistical properties of the presented identification approach, a Monte-Carlo study with $N_{\text{MC}} = 100$ runs is carried out. At each run, new realizations of the input u, scheduling signal p and noise e are taken.

To capture such vast dynamics exhibited by the considered data-generating system, an LPV-OBFs model structure is estimated from data, i.e., $\mathcal{D}_N$, with the following estimators:

C1 Fully parametric: first obtain the optimal basis based on the frozen local LTI poles by the proposed approach in Tóth et al. (2009a), then parameterize the expansion coefficients with a 2-nd order polynomial. Finally, perform an LS fitting for a (Wiener)-LPV-OBFs model suggested in Tóth et al. (2009a);

C2 Semi nonparametric: the optimal basis are borrowed from C1 and only the expansion coefficients are estimated in a nonparametric setting from data, where the RBF kernel is used to describe the underlying structural dependency on p;

C3 Fully nonparametric: the approach presented in this chapter, where the generating poles of the basis and the expansion coefficients are jointly estimated from data in a Bayesian setting.

Furthermore, two different basis scenarios are considered:.

D1 Scenario 1: The GOBFs basis, where the generating inner function is of 8-th order with three conjugate complex poles pairs and two real poles. This scenario is considered with the estimators C1, C2;

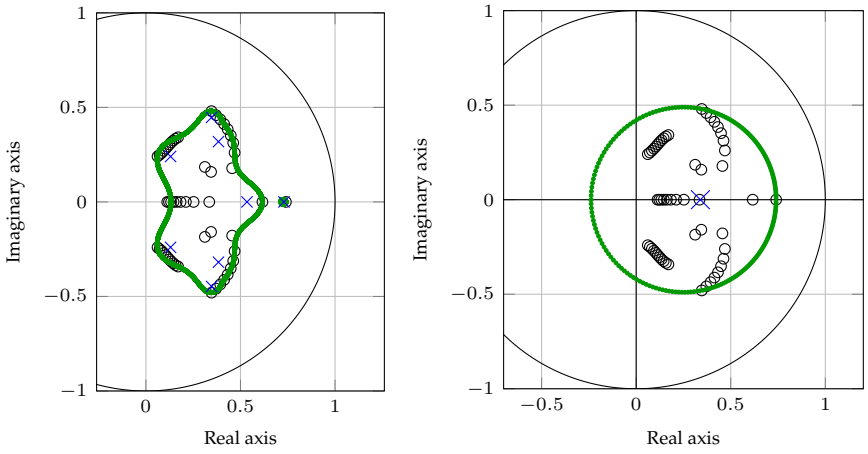
D2 Scenario 2: The Laguerre basis with one real pole. This scenario is considered with the estimators C1-C3.

The simulated output $\hat{y}$ of each of the identified models is compared to the true output of the data-generating system by the means of the BFR.

Identification results

Let us start with Scenario 1, where the estimators C1, C2 are compared. First, the result of running the FKcM algorithm to estimate the optimal set of basis based on the frozen LTI poles is shown in Figure 5.6(a). The selected number of clusters is 8, as recommended in Tóth et al. (2009a), corresponding to 3 complex conjugate poles and 2 real poles. The optimized generating poles of the basis are shown with crosses. From the figure and specifically from the Kolmogorov boundary[11] (solid green line), it can be easily seen that the optimized basis are very promising to be used to represent the underlying system. Second, to estimate the expansion coefficient functions, in C1 we parameterize them with a 2-nd order polynomial in $p(t)$ (static dependence) and by performing a LS fitting to identify the model, the average model fit is 82.25%. However, in C2 and with the optimal basis, the expansion coefficients are estimated in a nonparametric setting, and the average model fit is 88.09%. This shows the representation capability of the RBF utilized in the nonparametric estimate to describe the structural dependence of the coefficient functions with dynamic dependence compared to the 2-nd order polynomial basis with static dependence used in the parametric case. For illustration, the distributions of the model fits are shown by Boxplots in Figure 5.7(a). Now, let us analyze Scenario 2, where the considered three estimators, i.e., C1-C3, are compared. The used basis in this case is Laguerre basis, where the number of basis repetition is 10. By running the FKcM algorithm to estimate the optimal Laguerre basis, the optimal pole is found to be 0.3396, see Figure 5.6(b). From C1 and again by parameterizing the coefficients functions with a 2-nd order polynomial, the average model fit is 82.59%. From C2, where the coefficient functions are estimated in a nonparametric setting with the optimized pole obtained from C1, the average

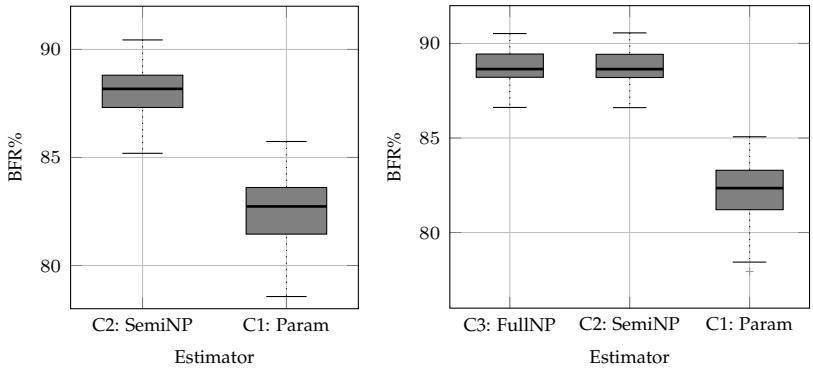
[11] Optimality of the optimized OBFs, specifically, the optimized poles of the generating inner function $\mathcal{H}_\mathrm{b}$ is ensured with a worst-case decay rate $\rho_\mathrm{b}^{n_\mathrm{e}+1}$, with n_e is the number of basis repetition, for systems with pole locations inside the regions defined by the Kolmogorov boundaries. The Kolmogorov boundary gives the boundaries of the regions $\{z \in \mathbb{C}, |\mathcal{H}_\mathrm{b}(z^{-1})| \leq \rho_\mathrm{b}\}$.



(a) Scenario D1, with GOBFs of order 8, $n_e = 1$

(b) Scenario D2, with Laguerre basis, $n_e = 10$

Figure 5.6: Results of FKcM clustering: sample poles (o), resulting cluster centers (x) and Kolmogorov boundaries (green line).



(a) Scenario 1

(b) Scenario 2

Figure 5.7: The distributions of the model fits.

model fit is 88.72%. Finally, with C3, where both the Laguerre pole and expansion coefficient functions are tuned simultaneously from data by maximizing the marginal likelihood, the average model fit is 88.73%. Such a close result to C2, where the optimal pole is used, is due to the efficiency of the empirical Bayes methods to tune the generating pole from data, which gives 0.3277 that is very close to the optimal pole location computed based on the Kolmogorov theory. As in Scenario 1, the distributions of the model fits are given by Boxplots in Figure 5.7(b).

5.5 Summary

In this chapter, we have presented a nonparametric identification approach for MIMO LPV-BJ models. Similar to the LTI case, it has been shown that the one-step-ahead predictor of such models is a summation of two sub-predictors associated with the input and output signals, where under the asymptotic stability of the data-generating system, these sub-predictors are shown to be convergent IIRs. To cope with issues associated with identifying such models, e.g., parameterization of matrix coefficient functions, a Bayesian nonparametric approach within the GP framework has been adopted. More specifically, the IIRs associated with the predictor are assumed to be realizations of zero-mean Gaussian random fields with a suitable designed kernel that encodes the expected prior knowledge on the predictor, e.g., stability, structural dependencies, etc. One of the main important contribution of this work is to show how to design such kernels to encode the expected prior knowledge about the predictor:

- Ensure the stability of the identified predictor;
- Encode possible structural dependencies;
- Take into account the LTI part as well as the $p-$dependent part of the model.

Two kernel formulation have been presented, i.e., DI-like and TC-like kernel. The hyperparameters of the kernels are tuned by maximizing their marginal likelihood over the observed data.

The developed approach has also been extended to series-expansion models, e.g., LPV-OBFs models, where it has been shown that such a setting could cope the challenges associated with identifying such models. More specifically, the choice of a suitable set of OBFs from data by maximizing the ML and guaranteeing the convergence of the identified series.

In this chapter, we have focused on identifying LPV systems in an IIR and series-expansion representation forms. In the next chapter, a different point of view is adopted. More specifically, the goal is to identify LPV models in an LPV-IO representation instead of the IIR form treated in this chapter, where a more parsimonious model can be obtained that allows to be further utilized in control design.

6 CHAPTER

Model Structure Learning for LPV-IO Identification

In this chapter, we return to the identification of LPV-IO models from a different view point by trying to retain the dynamic structure of the IO form, i.e., estimate the underlying relationship in an IO representation instead of the IIR form treated in Chapter 5. While the advantage of the Bayesian method of Chapter 5 was to avoid the problems of model order, noise structure selection and parameterization of the coefficient dependencies, the results of the estimation procedure was a process model G and a noise model H in an IIR form. This form is utilizable for prediction but for control design purposes a more parsimonious representation is needed in the form of an IO model. Due to the difficulties in LPV realization and model reduction theory, it is also advantageous to consider kernel based methods of LPV models directly in an IO form. This chapter aims at addressing this objective, which corresponds to Subgoal 4. More specifically, the main question is how to jointly reconstruct the scheduling-variable dependencies in IO models and at the same time choose the model order and the corresponding coefficient structure directly from data, with no prior parametrization. To this end, a unified learning framework for the identification of LPV-IO models in the RKHS setting is presented, where various kernel-based methods can be embedded.

6.1 Introduction

It has been discussed in Section 1.4.3 that the LPV modeling problem exhibits two main challenging issues: (i) the classical questions of determining the "suitable" dynamic order of the model, input delay and noise structure; (ii) to determine the underlying functional dependency of the coefficients on p such that

they have the least possible complexity for adequately representing the variation of the dynamics. Moreover, it has been shown that the available classical approaches to identify LPV systems have not been successful to fully address (i)-(ii), which points towards automatization of classical model order selection jointly with capturing structural dependency of the dynamic relation on the scheduling signal from data. Moreover, it has been discussed that sparse estimators are capable of achieving model structure selection from data, however, their performance strongly depends on adequate a priori selection of the basis functions. Alternatively, nonparametric approaches, e.g., kernel-based methods, offer a solution for estimating the dependency structure directly from data. However, the classical problem of selecting the model structure (i.e., model order, number of coefficient functions, delay etc.) has not been addressed jointly with estimating the dependency structure leaving the complexity/accuracy trade-off in terms of (i) open.

In this chapter, we are aiming at bridging the gap between sparse estimators and nonparametric estimators introduced for LPV identification. This allows to jointly reconstruct the scheduling-variable dependencies and the model order (coefficient structure) directly from data, with no prior parametrization of the p-dependent functions, resolving data-driven model structure selection in terms of (i) and (ii) in one step. In order to do that, a unified treatment of the sparse nonparametric estimation setting is introduced in an RKHS framework (Aronszajn 1950; Cucker and Smale 2001) where both the LS-SVM methods and GP formulations are directly included. Hence the derived results are directly applicable in both methodologies. For the sake of simplicity to show the underlying core ideas, the derivations are provided for a simple regression form, which assumes that the data generating system has an ARX structure. Note that the case of more general noise model structure can be handled by utilizing an IV formulation (Laurain et al. 2012). This chapter is organized as follows. Section 6.1 provides a general introduction to the considered problem, whereas the mathematical formulation of the problem is presented in Section 6.2 together with the considered model structure. In Section 6.3, the problem of identifying LPV-IO models from data formulated in the RKHS setting is presented. The resulting ℓ_2 regularized estimator is modified to include an ℓ_1 regularization term in Section 6.4, to enforce sparsity in the estimated model by detecting which coefficient functions are relevant. The resulting elastic-net function estimation problem is analyzed in the RKHS setting, proving the model structure selection capability of the method. In Section 6.5, the presented approach is compared to exiting solutions in terms of a detailed simulation study and also experimentally on the LPV identification of an unbalanced DC drive. Finally, the conclusions are presented in Section 6.6.

6.2 Problem Formulation

6.2.1 Data-generating system

To capture the IO relationship of a given LPV data-generating system $\mathcal{S}$, see Section 5.2.2, the so-called LPV-IO model is considered, which is commonly defined

in a filter form. In the SISO[1] case, the process dynamics of $\mathcal{S}$, according to (5.4a), are defined as

$$\breve{y}(t) = -\sum_{i=1}^{n_{\mathrm{a}}^{\mathrm{o}}} \mathfrak{a}_i^{\mathrm{o}}(p(t))q^{-i}\breve{y}(t) + \sum_{j=q^{\mathrm{o}}}^{n_{\mathrm{b}}^{\mathrm{o}}} \mathfrak{b}_j^{\mathrm{o}}(p(t))q^{-j}u(t), \tag{6.1}$$

where $u : \mathbb{Z} \to \mathbb{R}$, $\breve{y} : \mathbb{Z} \to \mathbb{R}$ are the measured input and noiseless output signals of the system, $q^{\mathrm{o}} \geq 0$ is the delay in the input channel, $p : \mathbb{Z} \to \mathbb{P}$ is the so-called scheduling variable, which ranges in a compact set $\mathbb{P} \subset \mathbb{R}^{n_{\mathbb{P}}}$ and assumed to be known exactly, $\mathfrak{a}_i^{\mathrm{o}}$ and $\mathfrak{b}_j^{\mathrm{o}}$ are coefficient functions dependent on $p(t)$, which are assumed to be smooth and bounded on $\mathbb{P}$.

For clarity of the exposition, in this chapter we assume that $\mathfrak{a}_i^{\mathrm{o}}(p(t))$ and $\mathfrak{b}_j^{\mathrm{o}}(p(t))$ have a static dependence[2] on p, i.e., $\mathfrak{a}_i^{\mathrm{o}}(p(t))$ and $\mathfrak{b}_j^{\mathrm{o}}(p(t))$ depend only on the instantaneous value of p at time t.

The most simple noise setting to be considered with the LPV process model (6.1) is the ARX form:

$$y(t) = -\sum_{i=1}^{n_{\mathrm{a}}^{\mathrm{o}}} \mathfrak{a}_i^{\mathrm{o}}(p(t))q^{-i}y(t) + \sum_{j=q^{\mathrm{o}}}^{n_{\mathrm{b}}^{\mathrm{o}}} \mathfrak{b}_j^{\mathrm{o}}(p(t))q^{-j}u(t) + e_{\mathrm{o}}(t), \tag{6.2}$$

where $y : \mathbb{Z} \to \mathbb{R}$ is the output and $e_{\mathrm{o}}(t)$ is a zero-mean white noise. According to (5.4) this means that $D_0 = A_0$, while $C_0 = 1$.

In order to simplify the notation, we will often use the following compact form to represent the data-generating LPV system (6.2):

$$y(t) = \mathfrak{f}^{\mathrm{o}}(x^{\mathrm{o}}(t), p(t)) + e_{\mathrm{o}}(t) = \sum_{i=1}^{n_{\mathrm{ab}}^{\mathrm{o}}} \mathfrak{f}_i^{\mathrm{o}}(p(t))x_i^{\mathrm{o}}(t) + e_{\mathrm{o}}(t), \tag{6.3}$$

where $n_{\mathrm{ab}}^{\mathrm{o}} = n_{\mathrm{a}}^{\mathrm{o}} + n_{\mathrm{b}}^{\mathrm{o}} - q^{\mathrm{o}} + 1$ and $x_i^{\mathrm{o}}(t)$ is the i-th component of the vector

$$x^{\mathrm{o}}(t) = \left[y(t-1) \ \cdots \ y(t-n_{\mathrm{a}}^{\mathrm{o}}) \ u(t-q^{\mathrm{o}}) \ \cdots \ u(t-n_{\mathrm{b}}^{\mathrm{o}})\right]^{\top}. \tag{6.4}$$

The following model structure is used to estimate (6.2)

$$y(t) = -\sum_{i=1}^{n_{\mathrm{a}}} \mathfrak{a}_i(p(t))q^{-i}y(t) + \sum_{j=q}^{n_{\mathrm{b}}} \mathfrak{b}_j(p(t))q^{-j}u(t) + e(t), \tag{6.5}$$

where $e(t)$ denotes the residual term, n_{a}, n_{b}, $q^{\mathrm{o}} \geq q \geq 0$ and not necessarily equal with $n_{\mathrm{a}}^{\mathrm{o}}$, $n_{\mathrm{b}}^{\mathrm{o}}$, q^{o}. Similarly to (6.3), the model (6.5) will often be represented in the

[1] For the sake of simplicity, in this chapter we assume a SISO system, however, the MIMO can be handled as has been introduced in Chapter 5.

[2] If $\mathfrak{a}_i^{\mathrm{o}}$ and $\mathfrak{b}_j^{\mathrm{o}}$ have a dynamic dependence on p, i.e., they depend on the past values of the scheduling signal $p(t), p(t-1), \ldots$, the following discussion can be easily extended for that case as has been shown in Chapter 5.

following compact form:

$$y(t) = \mathfrak{f}(x(t), p(t)) + e(t) = \sum_{i=1}^{n_{\mathrm{ab}}} \mathfrak{f}_i(p(t)) x_i(t) + e(t), \tag{6.6}$$

where $n_{\mathrm{ab}} = n_{\mathrm{a}} + n_{\mathrm{b}} - q + 1$ and $x_i(t)$ is the i-th component of the vector

$$x(t) = \left[y(t-1) \ \cdots \ y(t-n_{\mathrm{a}}) \ u(t-q) \ \cdots \ u(t-n_{\mathrm{b}}) \right]^{\top}. \tag{6.7}$$

6.2.2 Problem statement

Our goal is to jointly reconstruct the scheduling variable dependencies and the model structure, i.e., model order, number of effective coefficient functions, delay, etc., directly from data. To this end, based on a finite record of input, output and scheduling parameter measurements, i.e., $\mathcal{D}_N = \{u(t), y(t), p(t)\}_{t=1}^{N}$, our goal is to tackle the following problems:

- E1 Enforce sparsity in the estimate of the functions $\mathfrak{a}_i(p(t)),\ i = 1, \ldots, n_{\mathrm{a}}$, and $\mathfrak{b}_j(p(t)),\ j = q, \ldots, n_{\mathrm{b}}$. In this way, the model that is "best suited" for the approximation of the underlying system is chosen directly from the data;
- E2 Estimate the possibly nonlinear functions $\mathfrak{a}_i(p(t))$ and $\mathfrak{b}_i(p(t))$, characterizing the estimated relationship, directly from data;

In the following, an RKHS estimator to tackle the above-mentioned problems, i.e., E1-E2, is presented.

6.3 RKHS estimator for LPV-IO models

In Chapter 3, we have discussed in depth kernel-based methods to estimate an unknown nonlinear function from data. These methods can be treated in a unified framework in terms of regularization in RKHS. The main issue that needs to be taken into account to have a successful identification process based on such techniques is the design of a suitable kernel function that encodes our priori knowledge about the unknown function. As we have discussed, a kernel function should be parameterized in terms of a few number of parameters, the so-called hyperparameters, and in the same time should offer a wide representation capability of the expected behavior. The design of a suitable kernel function for the considered model class, i.e., LPV-IO model, will be the topic of Subsection 6.3.1. Moreover, in Subsection 6.3.2, the presented approach will be used to tackle Goal E2, where an RKHS estimator together with the designed kernel function is used to reconstruct the structural dependency, i.e., the coefficient functions $\mathfrak{a}_i(p(t))$ and $\mathfrak{b}_i(p(t))$ from data. This will show the potential of the RKHS estimator to provide, under some conditions, an analytic solution to this problem by employing the well-known Representer Theorem, see Theorem 3.1 in Chapter 3.

6.3.1 Kernel choice for LPV-IO models

It has been discussed how the choice of kernel K equivalently defines a functional Hilbert space, i.e., an RKHS $\mathscr{H}_K$, and therefore a model class. In the considered LPV identification problem, given the model structure (6.6), the aim is to find a kernel K that is able to embed a function $\mathfrak{f}(x,p)$ with a specific structure $\mathfrak{f}(x,p) = \sum_{i=1}^{n_{\mathrm{ab}}} \mathfrak{f}_i(p)x_i$. Naturally, such a Hilbert space is not unique. For example any Hilbert space embedding functions of type $\sum_{i=1}^{n_{\mathrm{ab}}} \mathfrak{f}_i(p)\pi_i(x_i)$ with π_i a polynomial function would also embed $\mathfrak{f}(x,p)$. Nevertheless, it is important, just like in usual identification problems, to enforce a structure in $\mathscr{H}_K$ which represents $\mathfrak{f}$ with the least possible degrees of freedom to achieve better performance. It becomes then crucial to define a reproducing kernel K which accounts for the linear dependency of LPV models on the terms x_i.

Lemma 6.1 (Reproducing kernel for LPV-IO models) *Given the LPV-IO structure* $\mathfrak{f}(x,p)$ *in* (6.6)*, the function* $\mathfrak{f}(x,p)$ *is embedded in the RKHS* $\mathscr{H}_K$*, whose reproducing kernel* $K : \mathbb{R}^{n_{\mathrm{ab}}+n_{\mathbb{P}}} \times \mathbb{R}^{n_{\mathrm{ab}}+n_{\mathbb{P}}} \to \mathbb{R}$ *is defined as:*

$$K\big((x,p),(x',p')\big) = \sum_{i=1}^{n_{\mathrm{ab}}} x_i K_i(p,p')x'_i, \tag{6.8}$$

where each sub-kernel $K_i(p,p') : \mathbb{R}^{n_{\mathbb{P}}} \times \mathbb{R}^{n_{\mathbb{P}}} \to \mathbb{R}$ *defines an RKHS* $\mathscr{H}_{K_i}$ *embedding* $\mathfrak{f}_i(p) : \mathbb{R}^{n_{\mathbb{P}}} \to \mathbb{R}$.

Proof: It is well-known that the RKHS $\mathcal{L}_i$ embedding linear functions for variable $x_i \in \mathbb{R}$ is defined by 1-dimensional kernels $\mathcal{L}(x_i, x'_i) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that are equal to $x_i x'_i$. Let $K_i(p,p') : \mathbb{R}^{n_{\mathbb{P}}} \times \mathbb{R}^{n_{\mathbb{P}}} \to \mathbb{R}$ define an RKHS $\mathscr{H}_{K_i}$ embedding $\mathfrak{f}_i(p)$. Thanks to the *Aronszajn Theorem* on RKHS products (Aronszajn 1950) (see Appendix A.3.2), the function $\mathfrak{f}_i(p)x_i$ is embedded in the RKHS product $\mathscr{H}_{K_i} \otimes \mathcal{L}_i$, where $\otimes$ denotes the direct product. Then, by using the *Aronszajn Theorem* (Aronszajn 1950) on RKHS sums (see Appendix A.3.1), it follows directly that $\mathfrak{f}(x,p)$ is embedded in the RKHS given by:

$$\mathscr{H}_K = \sum_{i=1}^{n_{\mathrm{ab}}} \mathscr{H}_{K_i} \otimes \mathcal{L}_i, \tag{6.9}$$

and that the associated kernel defined in (6.8) reproduces $\mathscr{H}_K$, which ends the proof. □

It should be noted that, for the choice of the kernels K_i any positive definite kernel, e.g., linear, polynomial, rational, spline or wavelet kernels, can be used. Choosing the appropriate kernel highly depends on the problem at hand. More details on that topic can be found in Schölkopf and Smola (2002). For the LPV case, RBFs are typically chosen as kernels to describe the expected structural dependency on p. Such a choice is motivated by the fact that the coefficient functions $\mathfrak{a}_i(p(t))$ and $\mathfrak{b}_i(p(t))$ are assumed to be, in general, nonlinear and smooth functions and RBFs

are proven to perform well in these situations. See (3.14) of Chapter 3 for the exact definition of the RBF kernel.

6.3.2 Estimation of the coefficient functions from data

It is interesting to notice that in general, the operation $\mathscr{H}_{K_1} \otimes \mathscr{H}_{K_2}$ embeds all products between functions $\mathfrak{f}_1$ of $\mathscr{H}_{K_1}$ and $\mathfrak{f}_2$ of $\mathscr{H}_{K_2}$. After estimating such a product though, it is not possible to deduce directly the value of $\mathfrak{f}_1(x)$ and $\mathfrak{f}_2(x)$ separately. Nevertheless, in the LPV context, thanks to the linear nature of $\mathcal{L}_i$, it becomes possible to get a direct estimation of each of the functions $\mathfrak{f}_i(p)$, as described in Lemma 6.2. However, before presenting that Lemma, it is convenient to review the Representer Theorem from Chapter 3 and define the corresponding cost function in the considered situation of this chapter.

The main idea of regularization, when it is applied to the problem of identifying $\mathfrak{f}$ in (6.6) from a set of measurements $\{u_t, p_t, y_t\}_{t=1}^N$, is to have a loss function $\mathcal{V}$ that consists of two terms, i.e., a *"data-fit"* term denoted by $\mathscr{C}$ and a *"regularizer"* term denoted by $\mathscr{R}$ forming

$$\mathcal{V}(\mathfrak{f}) = \mathscr{C}\big(u_1, p_1, y_1, \mathfrak{f}(x_1, p_1), \ldots, u_N, p_N, y_N, \mathfrak{f}(x_N, p_N)\big) + \gamma \mathscr{R}\big(\|\mathfrak{f}\|_K\big), \tag{6.10}$$

where x is defined in (6.7) and $\gamma > 0$ is the regularization parameter which defines the trade-off between both contradicting terms. For the "data-fit" term, the use of a quadratic loss function is common, for example in the considered PEM setting. For the "regularizer", since $\mathscr{H}_K$ is a Hilbert space, a mathematically rigorous and elegant analysis of regularization methods is possible. In such a case, a typical regularizer is the squared norm in the Hilbert space, Hence, the most popular choices of $\mathscr{C}$ and $\mathscr{R}$ in the regression literature are:

$$\begin{aligned} \mathscr{C}\big(u_1, p_1, y_1, \mathfrak{f}(x_1, p_1), \ldots, u_N, p_N, y_N, \mathfrak{f}(x_N, p_N)\big) &= \sum_{t=1}^{N} \big(y(t) - \mathfrak{f}(x_t, p_t)\big)^2 \\ \mathscr{R}(\|\mathfrak{f}\|_K) &= \|\mathfrak{f}\|_K^2. \end{aligned} \tag{6.11}$$

Now, let us present the generalized representer theorem (Argyriou and Dinuzzo 2014; Schölkopf et al. 2001).

Theorem 6.1 (Generalized Representer Theorem) *For a given RKHS $\mathscr{H}_K$ with reproducing kernel $K\big((x,p),(x',p')\big)$, the minimizer of* (6.10) *for any positive $\mathscr{C}$ and any strictly monotonically increasing real-valued function $\mathscr{R}$ on $[0,\infty)$, can be represented as*

$$\hat{\mathfrak{f}}(x', p') = \sum_{k=1}^{N} c_k K\big((x_k, p_k), (x', p')\big). \tag{6.12}$$

The Representer Theorem indicates that using the regularized optimization criterion (6.10), the estimated function $\mathfrak{f}$ can be expressed as a finite sum of kernel

slices/sections centered on the available observations and computed at any point using the associated kernel.

In case $\mathscr{C}$ and $\mathscr{R}$ are chosen as in (6.11), the parameters $c = [c_1 \ \cdots \ c_N]^\top \in \mathbb{R}^N$ defining the estimated function $\hat{\mathfrak{f}}$ in (6.12) and minimizing the cost function $\mathcal{V}(\mathfrak{f})$ in (6.10) can be directly computed analytically as follows:

$$c = (\mathcal{K} + \gamma I_{\mathrm{N}})^{-1} Y_N, \tag{6.13}$$

where $Y_N = [y_1 \ \cdots \ y_N]^\top$, and $\mathcal{K}$ is the kernel matrix whose (i,j)-th entry is $K\big((x_i, p_i), (x_j, p_j)\big)$. Now, we can proceed with the Lemma that describes how the coefficient functions can be reconstructed from data.

Lemma 6.2 (Estimating the coefficient functions) *Let $\mathfrak{f}(x,p)$ be embedded in an RKHS $\mathscr{H}_K$ with reproducing kernel K as in* (6.8)*. If $\mathfrak{f}(x,p)$ has a representer $\mathfrak{f}(x',p') = \sum_{k=1}^{N} c_k K\big((x_k,p_k),(x',p')\big)$, then each subfunction $\mathfrak{f}_i(p')$ of $\mathfrak{f}(x',p')$ is represented as*

$$\mathfrak{f}_i(p') = \sum_{k=1}^{N} c_k x_{ki} K_i(p_k, p'). \tag{6.14}$$

Proof: Consider the computation of $\mathfrak{f}$ at a given point $\mathfrak{f}(x',p')$. Then, let x' be the specific point $x^{\varnothing_i} \in \mathbb{R}^{n_{\mathrm{ab}}}$, which is defined such that each of its component $x_j^{\varnothing_i}$, $i = 1, \ldots, n_{\mathrm{ab}}$, $j = 1, \ldots, n_{\mathrm{ab}}$ is given as

$$x_j^{\varnothing_i} = \delta_{ij} \tag{6.15}$$

In other words, $x^{\varnothing_i}$ contains only 0 except at its i-th component which is equal to 1. It becomes then clear that $\mathfrak{f}(x^{\varnothing_i}, p') = \sum_{j=1}^{n_{\mathrm{ab}}} \mathfrak{f}_j(p') x_j^{\varnothing_i} = \mathfrak{f}_i(p')$. In other words, the computation of the subfunction $\mathfrak{f}_i(p')$ can be expressed as the computation of $\mathfrak{f}$ at a specific point $[x^{\varnothing_i}, p']$. Hence, given the representer

$$\mathfrak{f}(x',p') = \sum_{k=1}^{N} c_k K\big((x_k,p_k),(x',p')\big),$$

it can be easily seen that $K\big((x_k,p_k),(x^{\varnothing_i},p')\big) = x_{ki} K_i(p_k,p')$ to end the proof. □

Based on the results in Lemma (6.2), the problem of estimating the coefficient functions $\mathfrak{f}_i(p)$ has been formulated in the general framework of the RKHS theory.

It is interesting to note that, by using an optimization criterion defined by choices for $\mathscr{C}$ and $\mathscr{R}$ as in (6.11), the estimation results obtained confirm the results previously obtained in Tóth et al. (2011b) from the LS-SVM and in Golabi et al. (2014) from the GP viewpoints.

6.4 LPV-IO model order selection

In this section, we extend the results presented in Section 6.3 to select the LPV model structure, defined in terms of the parameters n_{a}, n_{b} and delay q, directly from data. In particular, the main goal of this section is to enforce sparsity in the estimate of the function $\mathfrak{f}(x,p)$, i.e., Goal E1, where "enforce sparsity" should be read as "keep the number of nonzero functions $\mathfrak{f}_i(p)$ in (6.6) small". Such information, i.e., detecting the nonzero coefficient functions, can be used to determine the LPV model structure.

To this end, a regularization term that complements the traditional expression (6.10) is added. More specifically, the new cost function consists of three terms:

1. The "data-fit" term $\sum_{t=1}^{N}\big(y(t)-\mathfrak{f}(x(t),p(t))\big)^2$ that aims at fitting the measured data;
2. The "regularizer" term $\|\mathfrak{f}\|_K^2$ is employed to prevent overfitting;
3. A "sparsity" term, which is introduced to enforce sparsity in the estimate of the model. More specifically, this regularization term aims at shrinking the functions $\mathfrak{f}_i$ to the zero function in order to minimize the number of non-zero coefficient functions $\mathfrak{f}_i$ characterizing the chosen LPV model structure.

The "sparsity" term that we propose is $\big\|\|\mathfrak{f}_1\|_\infty \dots \|\mathfrak{f}_{n_{\mathrm{ab}}}\|_\infty\big\|_1$, where $\|\cdot\|_1$ is a convex approximation of the ℓ_0-pseudo norm[3] and $\|\mathfrak{f}_i\|_\infty$ gives the maximum absolute value of the function $\mathfrak{f}_i$ over $\mathbb{P}$. However, in practice, the infinity norm would require the computation of $\mathfrak{f}_i(p)$ at each point of $\mathbb{P}$ which is computationally infeasible. Instead, an approximation of the infinity norms $||\mathfrak{f}_i(p)||_\infty$ as the maximum absolute value of the function $\mathfrak{f}_i$ over n_χ points of the scheduling variable domain $\mathbb{P}$, can be employed

$$\overline{\mathcal{J}}_i = \max_{j=1,\dots,n_\chi} |\mathfrak{f}_i(\mathfrak{m}_j)|\,.$$

The set of n_χ nodes can be chosen among the measured points or simply as a randomly generated points belonging to $\mathbb{P}$ without loss of generality. Hereafter, we will refer to these points as nodes of $\mathbb{P}$. Nevertheless, the difference between the infinity norm and the proposed approximation scheme is expected to be small if the kernels K_i enforce a sufficient smoothness on $\mathfrak{f}_i(p)$ with respect to the spacing in the chosen gridding points. As a result of such an approximation, the "sparsity" term can be expressed as

$$\left\| \left[\, \overline{\mathcal{J}}_1 \;\cdots\; \overline{\mathcal{J}}_{n_{\mathrm{ab}}} \right] \right\|_1 .$$

To conclude, the new cost function can be written as:

$$\mathcal{V}(\mathfrak{f}) = \sum_{t=1}^{N}\big(y(t)-\mathfrak{f}(x(t),p(t))\big)^2 + \gamma\|\mathfrak{f}\|_K^2 + \gamma_{\mathrm{s}}\left\| \left[\, \overline{\mathcal{J}}_1 \;\cdots\; \overline{\mathcal{J}}_{n_{\mathrm{ab}}} \right] \right\|_1 , \tag{6.16}$$

[3]The ℓ_0-pseudo norm of a vector x characterizes the support of that vector, i.e., the number of nonzero elements. Minimization under a ℓ_0 objective is a nonconvex NP-hard problem.

where $\gamma_{\mathrm{s}} > 0$ is the hyperparameter controlling the effect of the new regularization term, $\{\mathfrak{m}_j\}_{j=1}^{n_\chi}$ is a set of (randomly generated) points belonging to the scheduling variable space $\mathbb{P} \subseteq \mathbb{R}^{n_\mathbb{P}}$. Thus, the estimation of the LPV model $\mathfrak{f}(x,p)$ in (6.6) can be formulated as the optimization problem:

$$\begin{aligned} \min_{\mathfrak{f}} \; & \sum_{t=1}^{N} \big(y(t) - \mathfrak{f}(x(t),p(t))\big)^2 + \gamma \|\mathfrak{f}\|_K^2 + \gamma_{\mathrm{s}} \left\| \left[\begin{array}{ccc} \overline{\mathcal{J}}_1 & \cdots & \overline{\mathcal{J}}_{n_{\mathrm{ab}}} \end{array} \right] \right\|_1 \\ \text{s.t.} \; & \\ & \overline{\mathcal{J}}_i = \max_{j=1,\ldots,n_\chi} |\mathfrak{f}_i(\mathfrak{m}_j)| \,. \end{aligned} \tag{6.17}$$

In order to produce an estimate of the function $\mathfrak{f}$ minimizing (6.17), it is important to derive a suitable representer of $\mathfrak{f}$ in the form of the following Theorem.

Theorem 6.2 (Representer Theorem for sparse LPV-IO models) *Let $\mathscr{H}_K$ be an RKHS embedding LPV-IO models* (6.6) *with $K\big((x,p),(x',p')\big)$ in the form of* (6.8)*, as the reproducing kernel with kernel slice*[4] *$K_{(x,p)}$. Then the minimizer of* (6.17) *can be expressed as a representer in the form:*

$$\hat{\mathfrak{f}}(\cdot) = \sum_{k=1}^{N} c_k K_{(x_k,p_k)}(\cdot) + \sum_{i=1}^{n_{\mathrm{ab}}} \left(\sum_{j=1}^{n_\chi} \overline{c}_j^i K_{(x^{\varnothing_i},\mathfrak{m}_j)}(\cdot) \right), \tag{6.18}$$

or equivalently

$$\hat{\mathfrak{f}}(x',p') = \sum_{k=1}^{N} c_k K\big((x_k,p_k),(x',p')\big) + \sum_{i=1}^{n_{\mathrm{ab}}} \left(\sum_{j=1}^{n_\chi} \overline{c}_j^i K_i\left(\mathfrak{m}_j,p'\right) x' \right), \tag{6.19}$$

where $x^{\varnothing_i}$ is as in (6.15)*.*

Proof: Our goal is to express the optimization criterion (6.16) in the form of (6.10) suited for applying the generalized Representer Theorem, Theorem 6.1. The cost function $\mathcal{V}(\mathfrak{f})$ in (6.16) can be split into two parts $\mathcal{V}(\mathfrak{f}) = \mathcal{V}_{\mathrm{data}}(\mathfrak{f}) + \mathcal{V}_{\mathrm{f}}(\mathfrak{f})$ with

$$\begin{aligned} \mathcal{V}_{\mathrm{data}}(\mathfrak{f}) = & \sum_{t=1}^{N} \big(y_t - \mathfrak{f}(x_t,p_t)\big)^2 + \gamma_{\mathrm{s}} \left\| \left[\begin{array}{ccc} \overline{\mathcal{J}}_1 & \ldots & \overline{\mathcal{J}}_{n_{\mathrm{ab}}} \end{array} \right] \right\|_1 \\ \text{s.t.} \quad & \overline{\mathcal{J}}_i = \max_{j=1,\ldots,n_\chi} |\mathfrak{f}_i(\mathfrak{m}_j)| \,. \end{aligned} \tag{6.20}$$

and

$$\mathcal{V}_{\mathrm{f}}(\mathfrak{f}) = \gamma \|\mathfrak{f}\|_K^2 \tag{6.21}$$

It is then clear that $\mathcal{V}_{\mathrm{data}}(\mathfrak{f})$ corresponds to a positive cost function of the following type $\mathscr{C}\left((\mathfrak{f}(x_t,p_t),x_t,p_t),\mathfrak{f}_i(\mathfrak{m}_j)\right)$, with $t = 1,\ldots,N$, $i = 1,\ldots,n_{\mathrm{ab}}$, $j = 1,\ldots,n_\chi$. Note that this function depends on the underlying coefficient functions $\mathfrak{f}_i$, while,

[4] The kernel slice/section is $K_{(x,p)}(\cdot) = K((x,p),\cdot)$, which is a real-valued function whose value at a given (x',p') is $K((x,p),(x',p'))$.

in order to apply the generalized Representer Theorem, the function $\mathscr{C}$ needs to depend on $\mathfrak{f}$.

In the same fashion as in Lemma 6.2, it is sufficient to notice that due to the linear structure, $\mathfrak{f}_i(\mathfrak{m}_j) = \mathfrak{f}(x^{\mathcal{O}_i}, \mathfrak{m}_j)$. Hence, $\mathcal{V}_{\mathrm{data}}(\mathfrak{f})$ can be written as a generic function in the form :

$$\begin{aligned}\mathcal{V}_{\mathrm{data}}(\mathfrak{f}) =& \mathscr{C}\big((\mathfrak{f}(x_t, p_t), x_t, p_t), (\mathfrak{f}(x^{\mathcal{O}_i}, \mathfrak{m}_j), x^{\mathcal{O}_i}, \mathfrak{m}_j)\big), \\ & \text{with} \quad t = 1, \ldots, N, \;\; i = 1, \ldots, n_{\mathrm{ab}}, \;\; j = 1, \ldots, n_\chi.\end{aligned} \tag{6.22}$$

$\mathcal{V}_{\mathrm{f}}(\mathfrak{f})$ is already in the form of an increasing function of $\|\mathfrak{f}\|_K$. Consequently, (6.17) is now clearly expressed as an optimization criterion in the form of (6.10) which allows direct application of the Representer Theorem, resulting in (6.18). In order to prove (6.19), it is sufficient to notice that $K_{(x^{\mathcal{O}_i}, \mathfrak{m}_j)}(x', p') = K\left((x^{\mathcal{O}_i}, \mathfrak{m}_j), (x', p')\right) = K_i\left(\mathfrak{m}_j, p'\right) x'$. □

It is worth to emphasize that (6.18) and (6.19) are equivalent. While equation (6.19) is useful in order to compute the value of $\mathfrak{f}$ at a new given point, equation (6.18) allows the derivation of $\|\mathfrak{f}\|_K$.

It is important to notice that in the expression of the representer the $\overline{c}$ coefficients are a consequence of the added constraint term for sparsity. The cost function $\mathcal{V}_{\mathrm{data}}(\mathfrak{f})$ in (6.20) does not only depend on the measured points (x_t, p_t) but also on the gridding points $(x^{\mathcal{O}_i}, \mathfrak{m}_j)$ which have been introduced for solving the order selection problem. Interestingly, even in the case $\mathfrak{m}_j = p_t$, the $\overline{c}^i$ terms cannot be removed and are needed to enforce the regularization separately on each coefficient function $\mathfrak{f}_i$.

Having defined an optimization criterion, an LPV kernel structure as well as a representer for the problem at hand, the order selection problem can now be defined as follows:

Order selection problem: Consider a data set $\mathcal{D}_N = \{u(t), y(t), p(t)\}_{t=1}^N$ measured from a data-generating system as expressed in (6.2) and a set of nodes $\{\mathfrak{m}_1, \ldots \mathfrak{m}_{n_\chi}\}$. Using the representer (6.18), with K defined in (6.8), estimate the associated coefficients $\{c_k\},\ k = 1, \ldots, N$ and $\{\overline{c}_j^i\}, i = 1, \ldots, n_{\mathrm{ab}},\ j = 1, \ldots, n_\chi$ which minimize (6.17).

The solution to this problem corresponds to a quadratic optimization problem. It can be solved using any optimization toolbox. For the sake of readability, the matricial quadratic optimization expression of (6.17) is reported in details in Laurain et al. (2017).

Using the provided representer, due to the linear part of the kernel, each coefficient function $\mathfrak{f}_i(p'), i = 1, \ldots, n_{\mathrm{ab}}$ can be computed at a given point by applying (6.19) and then computing the value of $\mathfrak{f}$ as $\mathfrak{f}(x^{\mathcal{O}_i}, p') = \mathfrak{f}_i(p')$, which reads as:

$$\mathfrak{f}_i(p') = \sum_{k=1}^{N} c_k x_{ki} K_i(p_k, p') + \sum_{j=1}^{n_\chi} \overline{c}_j^i K_i(\mathfrak{m}_j, p'). \tag{6.23}$$

Note that, because of the ℓ_1-penalty term $\gamma_{\mathrm{s}}\left\|\left[\begin{array}{ccc}\overline{\mathcal{J}}_1 & \cdots & \overline{\mathcal{J}}_{n_{\mathrm{ab}}}\end{array}\right]\right\|_1$ introduced in (6.16) to shrink the coefficient functions $\mathfrak{a}_i$ and $\mathfrak{b}_j$ to zero, the resulting estimates of $\mathfrak{a}_i$ and $\mathfrak{b}_j$ will be biased. To cope with that, a two-step procedure is employed to estimate the nonzero coefficient functions. More specifically, first, a *Sparse*-RKHS (S-RKHS) approach, i.e., by minimizing the objective function $\mathcal{V}(\mathfrak{f})$ in (6.16) with $\gamma_{\mathrm{s}} > 0$, should be used to select which coefficients play a role in the dynamic relation of the full model (6.5). Now, assume that the indices of the detected zero functions $\mathfrak{a}_i$ and $\mathfrak{b}_j$ are collected in the index sets $\mathcal{I}_{\mathrm{y}}$ and $\mathcal{I}_{\mathrm{u}}$, respectively, i.e., $\mathfrak{a}_i$, for $i \in \mathcal{I}_{\mathrm{y}}$ and $\mathfrak{b}_j$, for $j \in \mathcal{I}_{\mathrm{u}}$ are detected to be zero functions. As a second step, the zero coefficient functions are discarded in the description of the LPV model (6.27) and a lower-complexity LPV model is considered

$$y(t) = -\sum_{i=1, i\notin\mathcal{I}_{\mathrm{y}}}^{n_{\mathrm{a}}} \mathfrak{a}_i(p(t))q^{-i}y(t) + \sum_{j=q, j\notin\mathcal{I}_{\mathrm{u}}}^{n_{\mathrm{b}}} \mathfrak{b}_j(p(t))q^{-j}u(t) + e(t). \tag{6.24}$$

Then, the model (6.24) is re-identified with the LPV RKHS approach, i.e., by minimizing the objective function $\mathcal{V}(\mathfrak{f})$ in (6.16) with $\gamma_{\mathrm{s}} = 0$.

6.5 Case studies

The effectiveness of the developed RKHS approach is shown in this section on two case studies. The first one is a Monte-Carlo study based on a simulation example. The second example is an experimental case study addressing the identification of a DC motor with an unbalanced disc acting as a position dependent load.

6.5.1 Simulation example

As a simulation example, we consider the identification of an LPV system with a sparse dynamic relation using an overparameterized LPV-IO model.

Data-generating system

The LPV data-generating system is a *Multi-Input Single-Output* (MISO) system described by the difference equation

$$y_t = \mathfrak{a}_1^{\mathrm{o}}(p_t)y_{t-1} + \mathfrak{b}_{1,15}^{\mathrm{o}}(p_t)u_{1t-15} + \mathfrak{a}_{2,4}^{\mathrm{o}}(p_t)u_{2t-4} + \mathfrak{b}_{2,5}^{\mathrm{o}}(p_t)u_{2t-5} + e_{\mathrm{o}}(t), \tag{6.25}$$

where $e_{\mathrm{o}}(t)$ is a white noise process with Gaussian distribution $\mathcal{N}\left(0, \sigma_{\mathsf{e}}^2\right)$ and standard deviation $\sigma_{\mathsf{e}} = 0.3$. The coefficient functions $\mathfrak{a}_1^{\mathrm{o}}(p_t)$, $\mathfrak{b}_{1,15}^{\mathrm{o}}(p_t)$, $\mathfrak{b}_{2,4}^{\mathrm{o}}(p_t)$ and

$\mathfrak{b}^{\circ}_{2,5}(p_t)$ are described by the nonlinear maps:

$$\mathfrak{a}^{\circ}_1(p_t) = 0.9p_t^3, \tag{6.26a}$$

$$\mathfrak{b}^{\circ}_{1,15}(p_t) = 2\frac{\sin(2\pi p_t)}{2\pi p_t}, \tag{6.26b}$$

$$\mathfrak{b}^{\circ}_{2,4}(p_t) = \begin{cases} -1 & \text{if } p_t > 0.5; \\ -2p_t & \text{if } -0.5 \leq p_t \leq 0.5; \\ 1 & \text{if } p_t < -0.5, \end{cases} \tag{6.26c}$$

$$\mathfrak{b}^{\circ}_{2,5}(p_t) = 2p_t^2. \tag{6.26d}$$

The system is estimated from a data set $\mathcal{D}_N = \{u_{1t}, u_{2t}, y_t, p_t\}_{t=1}^N$ with $N = 600$ input, output, and scheduling variable measurements. To gather data, the input u_1 and the scheduling signal have been chosen to be white-noise sequences, independent of each other, both of them with uniform distribution $\mathcal{U}(-1, 1)$. The second input u_2 is a white noise process with Gaussian distribution $\mathcal{N}\left(0, \sigma^2_{\mathrm{u},2}\right)$ and standard deviation $\sigma_{\mathrm{u},2} = 1$. In order to provide representative results, a Monte-Carlo simulation of $N_{\mathrm{MC}} = 50$ runs is performed. At each run, new realizations of the noise, inputs and scheduling signal are considered. The average of the SNR over the Monte-Carlo simulation is equal to 13dB.

LPV model structure

The identification problem is formulated in the considered RKHS setting by using an overparameterized LPV model structure:

$$y_t = \sum_{i=1}^{n_{\mathrm{a}}} \mathfrak{a}_i(p_t) y_{t-i} + \sum_{j=1}^{n_{\mathrm{b},1}} \mathfrak{b}_{1,j}(p_t) u_{1t-j} + \sum_{j=1}^{n_{\mathrm{b},2}} \mathfrak{b}_{2,j}(p_t) u_{2t-j} + e(t). \tag{6.27}$$

with $n_{\mathrm{a}} = 20, n_{\mathrm{b},1} = 20$ and $n_{\mathrm{b},2} = 20$. According to the RKHS identification setting considered in this chapter, the dependence of the functions $\mathfrak{a}_i(p_t)$, $\mathfrak{b}_{1,j}(p_t)$ and $\mathfrak{b}_{2,j}(p_t)$ on the scheduling signal p is not specified.

Coefficient functions estimation and model order selection

For the sake of comparison, the LPV model (6.27) is identified first through the RKHS estimator given in Section 6.3 (or equivalently, minimizing $\mathcal{V}(\mathfrak{f})$ in (6.16) for $\gamma_{\mathrm{s}} = 0$). The RBF kernels are used for the kernels K_i, i.e.,

$$K_i(p, p') = \exp\left(-\frac{(p - p')^2}{\beta_{\mathrm{w}i}^2}\right).$$

The values of the hyperparameters γ (6.16) and $\beta_{\mathrm{w}i}$ are chosen through CV[5], that is by maximizing (with an exhaustive grid search) the BFR w.r.t. the simulated model response with a validation data set of length $N_{\mathrm{V}} = 200$. Furthermore, in order to simplify the tuning of the hyperparameters $\beta_{\mathrm{w}i}$, the same value of $\beta_{\mathrm{w}i}$ is set for all kernels K_i. The obtained values of γ and $\beta_{\mathrm{w}i}$ are $\gamma = 1$ and $\beta_{\mathrm{w}i} = 0.7$ for all i. The estimates of the functions $\mathfrak{a}_1$, $\mathfrak{b}_{1,15}$, $\mathfrak{b}_{2,4}$ and $\mathfrak{b}_{2,5}$ (representing the non-zero coefficients of the true data-generating system (6.25)) are reported in the left parts of Figure 6.1-6.4, where the mean of the estimated functions over the 50 Monte-Carlo runs is plotted together with the standard deviation.

As a second step, the LPV model structure is selected through the S-RKHS approach described in Section 6.4, thus minimizing the multi-criteria objective function $\mathcal{V}(\mathfrak{f})$ in (6.16), for $\gamma_{\mathrm{s}} > 0$. The interval $\mathbb{P} = [-1, 1]$ is gridded into $n_\chi = 11$ equidistant nodes $\mathfrak{m}_j$. Similar to the first situation, the RBF kernels K_i are used, and the values of the hyperparameters γ, γ_{s} and $\beta_{\mathrm{w}i}$ are tuned through CV and set equal to $\gamma = 0.01$, $\gamma_{\mathrm{s}} = 0.3$ and $\beta_{\mathrm{w}i} = 0.7$ for all i.

The maximum absolute values $\overline{\mathfrak{a}}_i$, $\overline{\mathfrak{b}}_{1,j}$ and $\overline{\mathfrak{b}}_{2,j}$ of the coefficients functions $\mathfrak{a}_i(p_t)$, $\mathfrak{b}_{1,j}(p_t)$ and $\mathfrak{b}_{2,j}(p_t)$ estimated via the RKHS estimator (i.e., with $\gamma_{\mathrm{s}} = 0$) and its sparse version are reported in Tables 6.1-6.6, which show the average and the standard deviation of $\overline{\mathfrak{a}}_i$, $\overline{\mathfrak{b}}_{1,j}$ and $\overline{\mathfrak{b}}_{2,j}$ over the 50 Monte-Carlo runs. It is important to highlight that $\overline{\mathfrak{a}}_i$, $\overline{\mathfrak{b}}_{1,j}$ and $\overline{\mathfrak{b}}_{2,j}$ are the maximum of $|\mathfrak{a}_i(\centerdot)|$, $|\mathfrak{b}_{1,j}(\centerdot)|$ and $|\mathfrak{b}_{2,j}(\centerdot)|$ over the whole interval $\mathbb{P} = [-1, 1]$, and not only over the chosen nodes. Results in Tables 6.1-6.6 show that the S-RKHS approach correctly detects the LPV model structure. In fact, the only coefficient functions with an (average) maximum absolute value greater than a threshold of 10^{-2} are $\mathfrak{a}_1$, $\mathfrak{b}_{1,15}$, $\mathfrak{b}_{2,4}$ and $\mathfrak{b}_{2,5}$, which are the nonzero coefficient functions defining the considered data-generating system in (6.25). It is also worth remarking that the true coefficient structure of the system is detected in 47 out of 50 Monte-Carlo runs, while in the other 3 runs, 5 nonzero functions were detected instead of 4.

Estimation of the nonzero coefficient functions

As it can be noticed from Tables 6.1-6.6, the estimated maximum values of $|\mathfrak{a}_1(\centerdot)|$, $|\mathfrak{b}_{1,2}(\centerdot)|$, $|\mathfrak{b}_{2,4}(\centerdot)|$ and $|\mathfrak{b}_{2,5}(\centerdot)|$ over the interval $\mathbb{P}$ are 0.27, 0.12, 0.67 and 0.75, respectively, while the corresponding true values are 0.9, 2, 1 and 2. This agrees with the discussion in Section 6.4 that the estimated coefficient functions will be biased due to the added ℓ_1-term to the cost function. Therefore, the two-step procedure presented in Section 6.4 is followed, where the coefficient functions with maximum absolute value smaller than a threshold of 10^{-2} are discarded and the remaining functions are re-estimated. The estimates of the nonzero coefficient functions $\mathfrak{a}_1(\centerdot)$, $\mathfrak{b}_{1,15}(\centerdot)$, $\mathfrak{b}_{2,4}(\centerdot)$ and $\mathfrak{b}_{2,5}(\centerdot)$ are plotted in the right parts of Figure 6.1-6.4, which show the mean estimate together with the standard deviation intervals computed over the 50 Monte-Carlo runs. The obtained results show that the nonlinear coefficient functions $\mathfrak{a}_1$, $\mathfrak{b}_{1,15}$, $\mathfrak{b}_{2,4}$ and $\mathfrak{b}_{2,5}$ are accurately estimated,

[5]Alternatively, an efficient approach to tune the unknown hyperparameters is to resort to the Bayesian interpretation of the considered approach, where maximizing the marginal likelihood can be utilized.

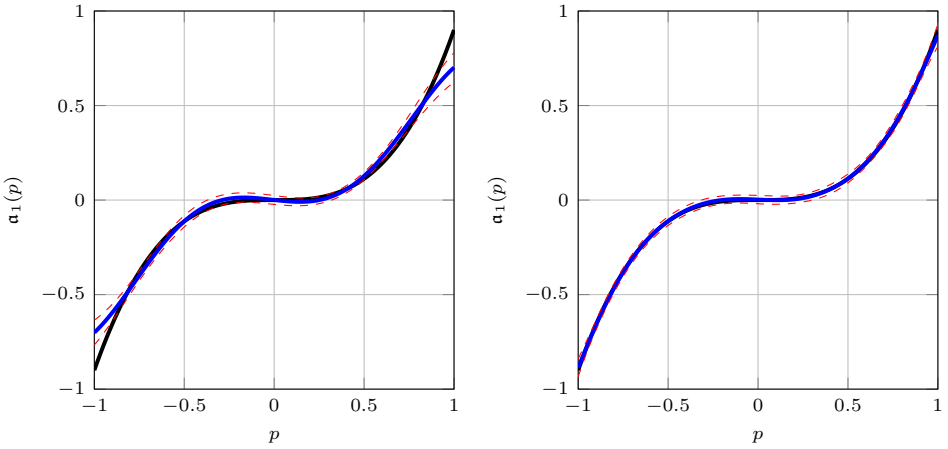


Figure 6.1: Example 1: coefficient functions $\mathfrak{a}_1(p(t))$: (left part) estimate via the RKHS estimator; (right part) estimate after model order selection. True function (solid black line), mean estimate (solid blue line) and the standard deviation intervals (dashed red line) over the 50 Monte Carlo runs.

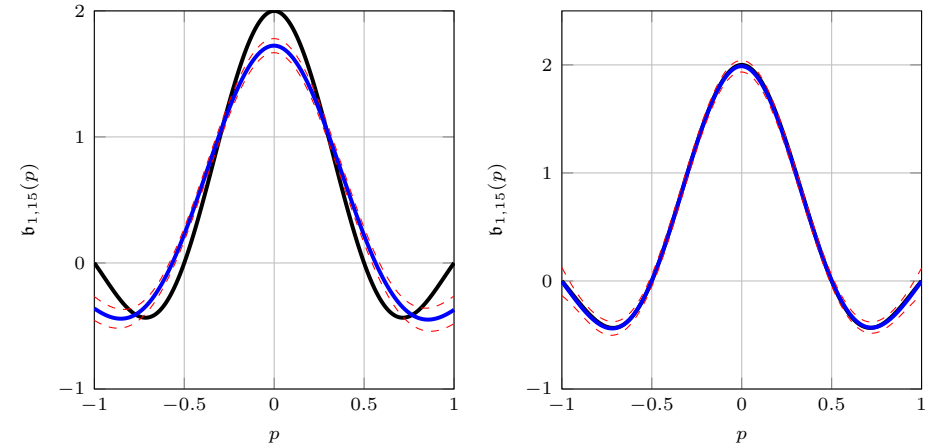


Figure 6.2: Example 1: coefficient functions $\mathfrak{b}_{1,15}(p(t))$: (left part) estimate via the RKHS estimator; (right part) estimate after model order selection. True function (solid black line), mean estimate (solid blue line) and the standard deviation intervals (dashed red line) over the 50 Monte Carlo runs.

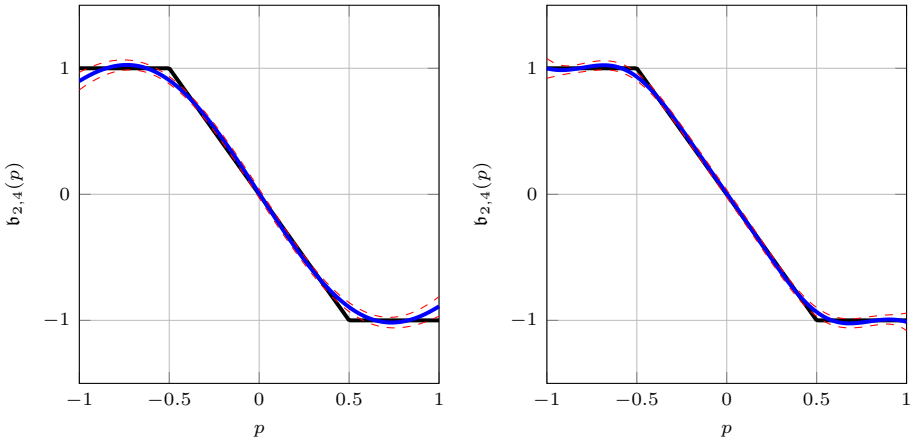


Figure 6.3: Example 1: coefficient functions $\mathfrak{b}_{2,4}(p(t))$: (left part) estimate via the RKHS estimator; (right part) estimate after model order selection. True function (solid black line), mean estimate (solid blue line) and the standard deviation intervals (dashed red line) over the 50 Monte Carlo runs.

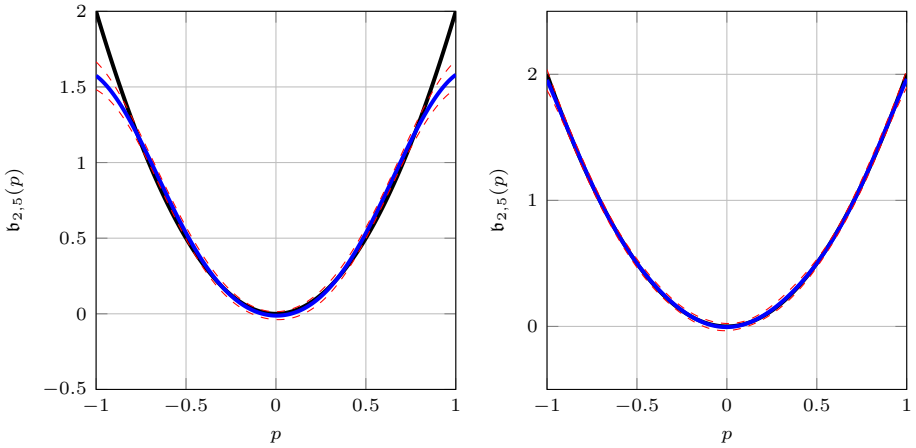


Figure 6.4: Example 1: coefficient functions $\mathfrak{b}_{2,5}(p(t))$: (left part) estimate via the RKHS estimator; (right part) estimate after model order selection. True function (solid black line), mean estimate (solid blue line) and the standard deviation intervals (dashed red line) over the 50 Monte Carlo runs.

Table 6.1: Example 1: average and standard deviation (over the 50 Monte-Carlo runs) of the maximum absolute value $\bar{\mathfrak{a}}_i$ of the coefficients functions $\mathfrak{a}_i(p_t)$, $i = 1, \dots, 10$. Comparison between the RKHS and S-RKHS estimators.

	True	Mean (RKHS)	Mean (S-RKHS)	Std (RKHS)	Std (S-RKHS)
$\bar{\mathfrak{a}}_1$	0.9	$7.4 \cdot 10^{-1}$	$2.7 \cdot 10^{-1}$	$5.7 \cdot 10^{-2}$	$0.7 \cdot 10^{-1}$
$\bar{\mathfrak{a}}_2$	0	$7.4 \cdot 10^{-2}$	$4.0 \cdot 10^{-4}$	$3.1 \cdot 10^{-2}$	$7.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_3$	0	$7.6 \cdot 10^{-2}$	$5.0 \cdot 10^{-4}$	$3.7 \cdot 10^{-2}$	$6.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_4$	0	$7.1 \cdot 10^{-2}$	$3.0 \cdot 10^{-4}$	$2.8 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_5$	0	$7.6 \cdot 10^{-2}$	$3.0 \cdot 10^{-4}$	$3.8 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_6$	0	$7.7 \cdot 10^{-2}$	$4.0 \cdot 10^{-4}$	$3.6 \cdot 10^{-2}$	$6.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_7$	0	$7.4 \cdot 10^{-2}$	$4.0 \cdot 10^{-4}$	$3.0 \cdot 10^{-2}$	$5.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_8$	0	$7.0 \cdot 10^{-2}$	$4.0 \cdot 10^{-4}$	$2.8 \cdot 10^{-2}$	$4.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_9$	0	$8.3 \cdot 10^{-2}$	$8.0 \cdot 10^{-4}$	$3.5 \cdot 10^{-2}$	$2.2 \cdot 10^{-3}$
$\bar{\mathfrak{a}}_{10}$	0	$7.6 \cdot 10^{-2}$	$7.0 \cdot 10^{-4}$	$3.5 \cdot 10^{-2}$	$2.0 \cdot 10^{-3}$

Table 6.2: Example 1: average and standard deviation (over the 50 Monte-Carlo runs) of the maximum absolute value $\bar{\mathfrak{a}}_i$ of the coefficients functions $\mathfrak{a}_i(p_t)$, $i = 11, \dots, 20$. Comparison between the RKHS and S-RKHS estimators.

	True	Mean (RKHS)	Mean (S-RKHS)	Std (RKHS)	Std (S-RKHS)
$\bar{\mathfrak{a}}_{11}$	0	$7.9 \cdot 10^{-2}$	$3.8 \cdot 10^{-4}$	$3.4 \cdot 10^{-2}$	$4.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_{12}$	0	$6.8 \cdot 10^{-2}$	$3.9 \cdot 10^{-4}$	$3.0 \cdot 10^{-2}$	$5.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_{13}$	0	$7.3 \cdot 10^{-2}$	$4.1 \cdot 10^{-4}$	$3.0 \cdot 10^{-2}$	$6.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_{14}$	0	$8.4 \cdot 10^{-2}$	$2.9 \cdot 10^{-4}$	$3.4 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_{15}$	0	$8.0 \cdot 10^{-2}$	$7.2 \cdot 10^{-4}$	$3.2 \cdot 10^{-2}$	$2.5 \cdot 10^{-3}$
$\bar{\mathfrak{a}}_{16}$	0	$6.3 \cdot 10^{-2}$	$4.5 \cdot 10^{-4}$	$3.5 \cdot 10^{-2}$	$6.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_{17}$	0	$7.4 \cdot 10^{-2}$	$4.6 \cdot 10^{-4}$	$3.2 \cdot 10^{-2}$	$7.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_{18}$	0	$6.8 \cdot 10^{-2}$	$3.4 \cdot 10^{-4}$	$3.3 \cdot 10^{-2}$	$3.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_{19}$	0	$6.8 \cdot 10^{-2}$	$3.5 \cdot 10^{-4}$	$3.2 \cdot 10^{-2}$	$3.0 \cdot 10^{-4}$
$\bar{\mathfrak{a}}_{20}$	0	$6.4 \cdot 10^{-2}$	$3.7 \cdot 10^{-4}$	$3.1 \cdot 10^{-2}$	$5.0 \cdot 10^{-4}$

Table 6.3: Example 1: average and standard deviation (over the 50 Monte-Carlo runs) of the maximum absolute value $\bar{\mathfrak{b}}_{1,i}$ of the coefficients functions $\mathfrak{b}_{1,i}(p_t)$, $i = 1, \dots, 10$. Comparison between the RKHS and S-RKHS estimators.

	True	Mean (RKHS)	Mean (S-RKHS)	Std (RKHS)	Std (S-RKHS)
$\bar{\mathfrak{b}}_{1,1}$	0	$1.2 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\bar{\mathfrak{b}}_{1,2}$	0	$1.3 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$5.1 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\bar{\mathfrak{b}}_{1,3}$	0	$1.3 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\bar{\mathfrak{b}}_{1,4}$	0	$1.2 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\bar{\mathfrak{b}}_{1,5}$	0	$1.2 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$4.9 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\bar{\mathfrak{b}}_{1,6}$	0	$1.1 \cdot 10^{-1}$	$2.0 \cdot 10^{-4}$	$4.5 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\bar{\mathfrak{b}}_{1,7}$	0	$1.2 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\bar{\mathfrak{b}}_{1,8}$	0	$1.3 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$5.8 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\bar{\mathfrak{b}}_{1,9}$	0	$1.4 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$5.5 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\bar{\mathfrak{b}}_{1,10}$	0	$1.2 \cdot 10^{-1}$	$2.0 \cdot 10^{-4}$	$5.1 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$

Table 6.4: Example 1: average and standard deviation (over the 50 Monte-Carlo runs) of the maximum absolute value $\overline{\mathfrak{b}}_{1,i}$ of the coefficients functions $\mathfrak{b}_{1,i}(p_t)$, $i = 11, \dots, 20$. Comparison between the RKHS and S-RKHS estimators.

	True	Mean (RKHS)	Mean (S-RKHS)	Std (RKHS)	Std (S-RKHS)
$\overline{\mathfrak{b}}_{1,11}$	0	$1.2 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$5.6 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{1,12}$	0	$1.2 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$5.2 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{1,13}$	0	$1.3 \cdot 10^{-1}$	$2.0 \cdot 10^{-4}$	$5.6 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{1,14}$	0	$1.3 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$5.5 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{1,15}$	2	1.72	$1.2 \cdot 10^{-1}$	$5.6 \cdot 10^{-2}$	$3.5 \cdot 10^{-2}$
$\overline{\mathfrak{b}}_{1,16}$	0	$1.3 \cdot 10^{-1}$	$2.0 \cdot 10^{-4}$	$5.5 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{1,17}$	0	$1.2 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$4.3 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{1,18}$	0	$1.3 \cdot 10^{-1}$	$1.0 \cdot 10^{-4}$	$5.3 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{1,19}$	0	$1.3 \cdot 10^{-1}$	$2.0 \cdot 10^{-4}$	$4.8 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{1,20}$	0	$1.2 \cdot 10^{-1}$	$2.0 \cdot 10^{-4}$	$4.3 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$

Table 6.5: Example 1: average and standard deviation (over the 50 Monte-Carlo runs) of the maximum absolute value $\overline{\mathfrak{b}}_{2,i}$ of the coefficients functions $\mathfrak{b}_{2,i}(p_t)$, $i = 1, \dots, 10$. Comparison between the RKHS and S-RKHS estimators.

	True	Mean (RKHS)	Mean (S-RKHS)	Std (RKHS)	Std (S-RKHS)
$\overline{\mathfrak{b}}_{2,1}$	0	$9.8 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$4.4 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,2}$	0	$8.8 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$3.8 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,3}$	0	$8.6 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$4.1 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,4}$	1	1.05	$6.7 \cdot 10^{-1}$	$3.7 \cdot 10^{-2}$	$3.9 \cdot 10^{-2}$
$\overline{\mathfrak{b}}_{2,5}$	2	1.63	$7.5 \cdot 10^{-1}$	$6.4 \cdot 10^{-2}$	$5.2 \cdot 10^{-2}$
$\overline{\mathfrak{b}}_{2,6}$	0	$1.1 \cdot 10^{-1}$	$1.2 \cdot 10^{-3}$	$4.8 \cdot 10^{-2}$	$1.2 \cdot 10^{-3}$
$\overline{\mathfrak{b}}_{2,7}$	0	$9.7 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$4.6 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,8}$	0	$8.4 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$4.1 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,9}$	0	$9.6 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$4.2 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,10}$	0	$9.7 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$3.2 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$

Table 6.6: Example 1: average and standard deviation (over the 50 Monte-Carlo runs) of the maximum absolute value $\overline{\mathfrak{b}}_{2,i}$ of the coefficients functions $\mathfrak{b}_{2,i}(p_t)$, $i = 11, \dots, 20$. Comparison between the RKHS and S-RKHS estimators.

	True	Mean (RKHS)	Mean (S-RKHS)	Std (RKHS)	Std (S-RKHS)
$\overline{\mathfrak{b}}_{2,11}$	0	$2.0 \cdot 10^{-4}$	$4.4 \cdot 10^{-2}$	$4.4 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,12}$	0	$9.4 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$4.7 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,13}$	0	$8.9 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$3.1 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,14}$	0	$9.6 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$3.5 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,15}$	0	$9.4 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$3.7 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,16}$	0	$9.6 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$3.3 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,17}$	0	$9.2 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$3.3 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,18}$	0	$8.7 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$3.7 \cdot 10^{-2}$	$1.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,19}$	0	$9.0 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$3.8 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$
$\overline{\mathfrak{b}}_{2,20}$	0	$9.4 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$	$4.6 \cdot 10^{-2}$	$2.0 \cdot 10^{-4}$

with a lower variance with respect to the RKHS estimator. The boxplots of the BFR on the validation dataset (used neither for training nor to tune the hyper-parameters γ, γ_{s} and β_{w_i}) obtained with the regularized and RKHS approach are also computed and reported in Figure 6.5. The obtained results clearly indicate that, thanks to an accurate reconstruction of the LPV model structure, the regularized method leads to higher BFR than the RKHS approach.

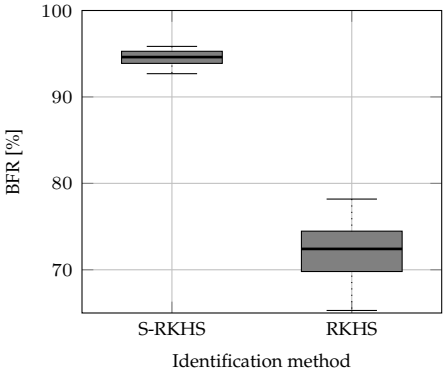


Figure 6.5: Example 1: boxplot of the Monte-Carlo simulation for the BFR obtained with the Sparse-RKHS (left) and standard RKHS (right) estimators.

6.5.2 Experimental example

As a second example, we consider an experimental study addressing the identification of a DC motor with an unbalanced disc shown in Figure 6.6. The motor has an additional mass, which is mounted on the disc attached to the rotor to make the mass distribution inhomogeneous, thus introducing nonlinear dynamics. The parameters characterizing the DC motor are reported in Table 6.7.

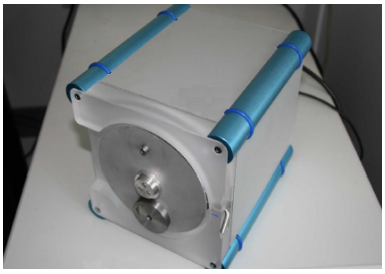


Figure 6.6: Example 2: DC motor with an unbalanced disc used as an experimental testbed.

Table 6.7: Example 2: physical parameters of the DC motor (Kulcsár et al. 2009).

Description	Value
Motor resistance	$9.5\ \Omega$
Motor inductance	$0.84 \cdot 10^{-3}$ H
Motor torque constant	$53.64 \cdot 10^{-3}$ N m/A
complete disc inertia	$2.2 \cdot 10^{-4}$ N m^2
Friction coefficient	$6.6 \cdot 10^{-5}$ N m s/rad
Additional mass	0.07 kg
Mass distance from the center	0.042 m

Identification setting

The goal is to estimate a model describing the relationship between the input voltage u [V] over the motor armature and the angular position y [rad] of the disc. To this aim, these variables are measured at a sampling time of 0.02 s, and a dataset with 500 samples is constructed. The first $N = 300$ samples are used to identify the model and tune the hyperparameters γ, $\gamma_{\rm s}$ and $\beta_{{\rm w}i}$ involved in the proposed RKHS based identification method, while the remaining 200 samples are used to assess the quality of the estimated model. The input voltage is chosen as a white noise sequence with uniform distribution in the interval $[-8, 8]$ V, filtered by a first order digital filter with a cutoff frequency of 1.6 Hz (the generated sequence of the input voltage u and of the angular position y of the disc are plotted in Fig. 6.7).

In formulating the RKHS identification method discussed in this chapter, the following LPV model structure is used:

$$y_t = \sum_{i=1}^{n_{\rm a}} \mathfrak{a}_i(p_{t-1}) y_{t-i} + \sum_{j=1}^{n_{\rm b}} \mathfrak{b}_j(p_{t-1}) u_{t-j} + e_t, \tag{6.28}$$

with $n_{\rm a} = 15$ and $n_{\rm b} = 15$. The output y_t (namely, the angular position of the disc) is used as a scheduling variable (i.e., $p_t = y_t$).

Identification results

The S-RKHS approach in Section 6.4 is used to identify the model (6.28), i.e., by minimizing (6.16) with the kernels K_i chosen as RBF kernels. The values of the hyperparameters, tuned through CV, are set to $\gamma = 0.01$, $\gamma_{\rm s} = 0.05$ and $\beta_{{\rm w}i} = 6$ for all i. The threshold is taken to be 10^{-2}. The two-stage approach is used, where first the optimization (6.16) is solved for $\gamma_{\rm s} = 0.05$, then the coefficient functions with maximum absolute values smaller than 10^{-2} are discarded and a reduced order model is re-estimated with the RKHS approach, i.e., with $\gamma_{\rm s} = 0$. The identified model is tested on the validation data set and a BFR of 88.16% is achieved. The true and the simulated outputs are plotted in Figure 6.8. The obtained results show the capability of the presented approach to accurately estimate the nonlinear dynamics of a real system from a relatively short data record.

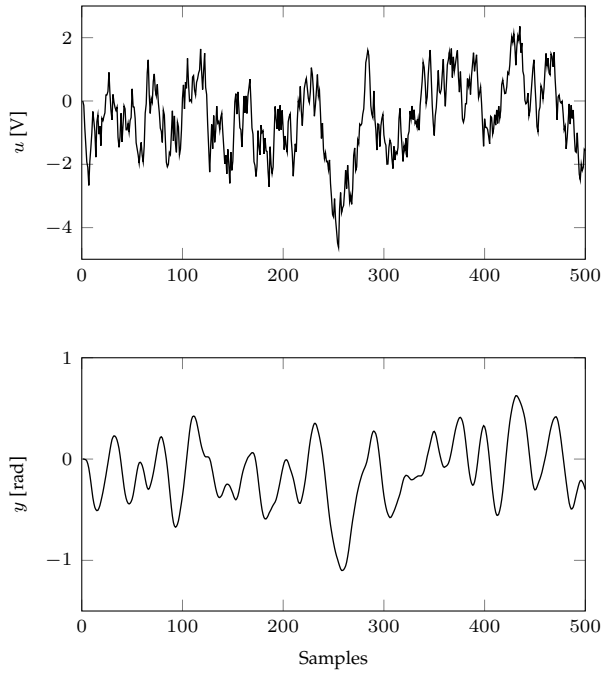


Figure 6.7: Example 2: input voltage u (upper plot) and angular position y of the disc (lower plot) used for the identification and validation of the DC motor with an unbalanced disc.

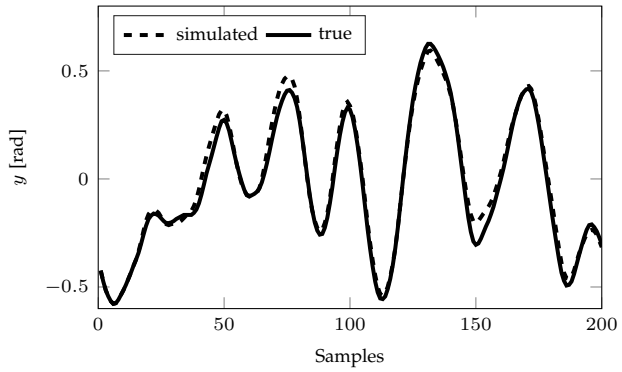


Figure 6.8: Example 2: true output (solid) vs. simulated model output (dashed) on a validation dataset.

6.6 Summary

In this chapter, Subgoal 4 has been addressed. More specifically, a unified framework in the RKHS setting for model structure learning of LPV-IO models has been formulated. In such a setting, kernel based methods, e.g., LS-SVM and GP, can be easily embedded. First, the problem has been formulated to estimate the coefficient functions from data. In addition to formulating a kernel function that is able to generate an RKHS that embeds the considered model structure, the utilized cost function has been modified to tackle the problem of model order selection from data by complementing it with a third term that enforces sparsity in the estimated coefficient functions. This provides an automatic way for learning structure of LPV-IO models from data. The effectiveness of the presented framework is tested on both simulation and experimental examples.

7 CHAPTER

Conclusions and Recommendations

In this thesis, we have aimed at utilizing new developments in system identification stemming from the machine learning community to address the open problems and associated challenges with data-driven modeling of *Linear Dynamic Systems* (LDS). This chapter provides the main conclusions of the presented research throughout the thesis. In addition, suggestions and research directions for future work are also given.

7.1 Conclusions

This thesis has been motivated by our interest to deliver accurate linear models of physical processes. These models include *Linear Time Invariant* (LTI) models and their extensions, e.g., *Linear Time Varying* (LTV) and *Linear Parameter Varying* (LPV) models. Such advanced linear models have proven to be capable of describing both *Nonlinear* (NL) and *Time-Varying* (TV) nature of physical systems. A lot of work has been done to tackle the associated challenges and problems with identifying these models. More specifically, the choice of the "right" structure and order of the model to be estimated from data to "best" describe the behavior of the considered process. Often we characterize performance in terms of the prediction error of the estimated models. By utilizing approaches borrowed from the machine learning community, many of the present problems have been circumvented. However, there are still many open questions to be addressed.

Based on the extensive literature overview that have been presented in Chapter 1, we have established two research questions to be answered as our primary research goal. Accordingly several subgoals have been formulated in Section 1.7, which had to be investigated to fulfill/answer the research questions.

In the following, we summarize the results of our investigations towards the considered subgoals.

Subgoal 1: Systematic utilization of prior knowledge

In the scope of our first research goal, we have investigated how to systematically construct kernel functions that are capable of describing a wide range of dynamic properties of LTI systems, e.g., stability, resonance behavior, damping, etc., both in the time- and frequency-domain. We have presented a class of kernel function that is based on *Orthonormal Basis Functions* (OBFs), where the prior knowledge can be encoded via the generating poles of the utilized OBFs.

To have a data-driven approach to decide on both the choice of the appropriate set of OBFs to be used and the effective number of these basis functions, a decay term has been introduced in the kernel construction. Such a decay term also guarantees that the resulting hypothesis spaces, i.e., the *Reproducing Kernel Hilbert Spaces* (RKHSs) associated with these kernel functions, are stable. More specifically, in the time- and frequency-domain, these hypothesis spaces contain only impulse responses and *Frequency Response Functions* (FRFs) of stable LTI systems, respectively.

It has been shown that by designing kernel functions that are supported by system theory, the capability of the kernel to encode a wide range of dynamic properties has been significantly improved. In the same time, such a representation capability can be achieved with a simple parameterization, where the unknown hyperparameters can be efficiently estimated by maximizing the marginal likelihood. Such an improvement in the kernel structure results in a better bias/variance trade-off compared to the previously used kernel functions that focus only on encoding smoothness and stability. Accordingly, the obtained estimates become more accurate in terms of minimizing the *Mean Squared Error* (MSE).

Subgoal 2: Bayesian PEM identification of LPV systems

In the scope of our second research goal, we have investigated how to cope with the issues associated with identifying LPV systems under a *Prediction Error Minimization* (PEM) setting, specifically, in case of general noise scenarios, i.e., an LPV-*Box Jenkins* (BJ) noise model structure. It has been shown that nonparametric identification within Bayesian setting provides a powerful tool to "efficiently" tackle these problems, e.g., the parameterization of the coefficient functions, model order and noise structure selection. As an extension of the LTI case, identification of LPV-BJ models has been formulated as obtaining nonparametric estimates of the one-step-ahead predictors of such models. It has been shown that the one-step-ahead predictor can be written as a summation of two sub-predictors associated with the input and output signals, which can be modeled as asymptotically stable *Infinite Impulse Representations* (IIRs). To account for all related aspects of these IIRs, specifically, the structure of dependency on the so-called scheduling signal p and the asymptotic stability, these IIRs are completely identified in a nonparametric sense: not only the coefficient functions are estimated as functions, but also the whole time evolution of the impulse response w.r.t. the scheduling signal.

To this end, it has been shown that in the Bayesian setting, the one-step-ahead predictor can be seen as a multi argument function and a statistical prior is pos-

tulated on it. Specifically, the prior of the one-step-ahead predictor is taken to be a zero-mean Gaussian random field, which can be completely characterized by its covariance/kernel function. A suitable kernel function has been designed that encodes the prior knowledge about the to-be-estimated function, i.e., the one-step-ahead predictor. Such a kernel function is a multidimensional Gaussian kernel that incorporates information on both possible structural dependency and the stability of the predictor.

Subgoal 3: Identification of series-expansion LPV models

Using the developed approach to identify LPV-BJ models within the Bayesian setting, where the associated sub-predictors of such models have been considered as convergent IIRs, the extension of Bayesian identification of LPV series-expansion models has become straightforward. More specifically, the identification of both LPV-IIR and LPV-OBFs model structures can be regarded as function estimation problem and hence the presented approach for Subgoal 2 can be extended to the considered case in Subgoal 3. Within Bayesian setting, the associated problems with identifying LPV series-expansion models has been tackled, e.g., parameterizing the coefficient functions and guaranteeing the convergence of the estimated series. Moreover, for LPV-OBFs models, the problem of selecting a proper set of OBFs from data has been dealt with by considering the generating poles of the utilized basis functions as hyperparameters and estimating them with maximizing the marginal likelihood. This can be seen also as an LPV extension of the Regularized OBFs approach that has been presented in Chapter 4 for LTI systems.

Subgoal 4: Model structure learning of LPV models

In the scope of our final research goal, we have investigated how to jointly reconstruct the scheduling-variable dependencies and the model order (coefficient structure) directly from data, with no prior parameterization of the p-dependent functions. More specifically, it has been shown that this problem can be formulated in the RKHS setting, where various regularization techniques can be embedded. In such a setting, the model estimate is the solution of an optimization problem that minimizes a three-term cost function. These three terms are the data-fit, the regularizer and the sparsity terms. The main goal of such a cost function is to obtain nonparametric estimates of the significant coefficient functions, where these functions are detected via the sparsity term that shrinks the insignificant coefficient functions to be "almost" zero. To this end, a suitable kernel function has been designed that embeds the considered LPV-IO models. Such a kernel function results in an RKHS that is used as a hypothesis space, where the model estimate is searched for.

Based on the above discussion, it can be concluded that the developed approaches of this thesis provide answers to the primary research questions: synthesis of kernel functions for linear models which are backed up by system theory to represent dynamic properties; extension of the promising Bayesian methods to advanced linear models, where open questions with nonparametric estimation in

a general noise setting, circumventing model structure selection problems have been addressed.

7.2 Recommendations for Future Research

This section recommends some research directions for future work as a continuation of the presented results. The suggestions are as follows:

- In Chapter 6, model structure learning of LPV-IO models has been formulated in the RKHS framework. Tuning the unknown parameters that include: the parameters that control the trade-off between various contradicting terms in the cost functions; the hyperparameters that parameterize the kernel functions, has been performed using *Cross Validation* (CV), which requires an additional validation data set. The resulting bias/variance trade-off and accordingly the accuracy of the estimated models are largely dependent on the tuned parameters. Hence, an efficient alternative of CV that does not need a a validation data set and can automatically balance bias/trade-off is needed. This requires the formulation of the problem of model structure learning of LPV-IO models from a Bayesian point of view under sparsity penalties, where maximizing the marginal likelihood can be employed to tune the unknown hyperparameters.
- In this thesis, nonparametric models of LPV systems have been obtained from data. However, the utilization of such models in control design is not investigated. Further research in that direction includes:
  - Based on the identified nonparametric models realize parametric models that can be used in the available LPV model-based control design methods;
  - Investigate a possible LPV control design method that can directly make use of the obtained nonparametric estimates, e.g., in a model predictive setting.
- In Chapter 4, kernel functions have been synthesized for LTI systems both in time- and frequency-domain based on OBFs. Furthermore, it has been shown how the resulting kernels provide a systematic approach to encode the expected dynamic properties of LTI systems. In order to extend such an approach to LPV systems, i.e., to construct kernel function based on OBFs that results in an RKHS that embed LPV systems, it is interesting to investigate if such orthonormal basis functions can be defined in a parameter-varying sense. Such a concept requires the full analysis of the properties of such basis, the resulting RKHS and the systems that can be embedded within such a setting. This would provide a full extension of the results of Chapter 4 to the LPV case.

A APPENDIX

Proofs

In this appendix, the proofs of the theories and lemmas of this thesis are presented. The proofs relay heavily on the notation and concepts introduced in the previous chapters.

A.1 Proof of Proposition 2.1

By following the same line of reasoning as in Wahlberg (1991): the $L_\infty(\mathbb{J})$-norm of the k-th Takenaka-Malmquist basis $\breve{\psi}_k$ is uniformly bounded

$$\sup_\omega \left| \frac{\sqrt{1-|\lambda_k|^2}}{e^{\mathrm{j}\omega}-\lambda_k} \prod_{i=1}^{k-1} \frac{1-\lambda_i^* e^{\mathrm{j}\omega}}{e^{\mathrm{j}\omega}-\lambda_i} \right| \leq \frac{\sqrt{1-|\lambda_k|^2}}{1-|\lambda_k|}.$$

Based on the fact that the $\ell_1(\mathbb{N})$-norm of the impulse response of a k-th order stable system is less than twice the nuclear norm of the associated Hankel operator (Glover et al. 1988, Section 2), and that nuclear norm is less than k times the $\mathscr{L}_\infty(\mathbb{J})$-norm (Glover et al. 1988, Theorem 2.1):

$$\|\psi_k\|_{\ell_1} \leq k2\frac{\sqrt{1-|\lambda_k|^2}}{1-|\lambda_k|}.$$

Let

$$\kappa = \sup_{\lambda \in \{\lambda_k\}_{k=1}^\infty} \frac{\sqrt{1-|\lambda|^2}}{1-|\lambda|}.$$

Accordingly,

$$\|\psi_k\|_{\ell_1} \leq k \cdot 2\kappa.$$

A.2 Proof of Proposition 4.3

The proof can be accomplished by the application of Corollary 4.1, see (4.25) and Proposition 2.1. From Corollary 4.1, one have to check that:

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|K_{\Psi}^{\mathrm{s}}(i,j)| = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|\sum_{k=1}^{\infty}\mathfrak{D}_k\left(\beta_{\mathrm{d}}\right)\psi_k(i)\psi_k(j)| < \infty.$$

To this end:

$$\begin{aligned}
&\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|\sum_{k=1}^{\infty}\mathfrak{D}_k\left(\beta_{\mathrm{d}}\right)\psi_k(i)\psi_k(j)| \\
&\leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\mathfrak{D}_k\left(\beta_{\mathrm{d}}\right)|\psi_k(i)||\psi_k(j)| \\
&= \sum_{k=1}^{\infty}\mathfrak{D}_k\left(\beta_{\mathrm{d}}\right)\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|\psi_k(i)||\psi_k(j)| \\
&= \sum_{k=1}^{\infty}\mathfrak{D}_k\left(\beta_{\mathrm{d}}\right)\underbrace{\sum_{i=1}^{\infty}|\psi_k(i)|}_{\|\psi_k\|_{\ell_1(\mathbb{N})}}\underbrace{\sum_{j=1}^{\infty}|\psi_k(j)|}_{\|\psi_k\|_{\ell_1(\mathbb{N})}} \\
&\leq 4\kappa^2\sum_{k=1}^{\infty}k^2\mathfrak{D}_k\left(\beta_{\mathrm{d}}\right),
\end{aligned}$$

where the last equation is obtained by the provided bound in Proposition 2.1. Hence, $\sum_{k=1}^{\infty}k^2\mathfrak{D}_k\left(\beta_{\mathrm{d}}\right) < \infty$ should be satisfied to guarantee the stability of the kernel.

In case of (4.41): $\sum_{k=1}^{\infty}k^2\mathfrak{D}_k\left(\beta_{\mathrm{d}}\right) = \sum_{k=1}^{\infty}k^{(2-\beta_{\mathrm{d}})}$, and with $\beta_{\mathrm{d}} > 3$, the series will have a convergent sum, which guarantees the stability of the kernel.

In case of (4.42): $\sum_{k=1}^{\infty}k^2\mathfrak{D}_k\left(\beta_{\mathrm{d}}\right) = \sum_{k=1}^{\infty}k^2\beta_{\mathrm{d}}^{-k}$, and with $\beta_{\mathrm{d}} > 1$, the series will have a convergent sum, which guarantees the stability of the kernel.

A.3 Aronszajn's Theorems (Aronszajn 1950)

A.3.1 Sum of kernels

If $K_i(x,x')$ is the reproducing kernel of the RKHS $\mathscr{H}_{K_i}$ with the norm $\|\,\boldsymbol{.}\,\|_{K_i}$, then $K(x,x') = \sum_{i=1}^{n}K_i(x,x')$ is the reproducing kernel of the RKHS $\mathscr{H}_K$ containing all functions $\mathfrak{f} = \sum_{i=1}^{n}\mathfrak{f}_i$ with $\mathfrak{f}_i \in \mathscr{H}_{K_i}$ and with the norm defined by $\|\mathfrak{f}\|_K^2 = \min\left[\sum_{i=1}^{n}\|\mathfrak{f}_i\|_{K_i}^2\right]$, the minimum taken for all the decompositions $\mathfrak{f} = \sum_{i=1}^{n}\mathfrak{f}_i$ with $\mathfrak{f}_i \in \mathscr{H}_{K_i}$. If all $\mathscr{H}_{K_i}$ are disjoint, and therefore do not include any common functions beside 0, then the norm in $\mathscr{H}_K$ is simply given by $\sum_{i=1}^{n}\|\mathfrak{f}_i\|_{K_i}^2$.

A.3.2 Product of kernels

The direct product of RKHS $\mathscr{H}_{K_1} \otimes \mathscr{H}_{K_2}$ defined by the reproducing kernels $K_1(x_1, x_1')$ and $K_2(x_2, x_2')$ and the norms $\| \cdot \|_{K_1}$ and $\| \cdot \|_{K_2}$ is an RKHS $\mathscr{H}_K$ defined by the reproducing kernel $K\big((x_1, x_1'), (x_2, x_2')\big) = K_1(x_1, x_1')K_2(x_2, x_2')$ with the norm $\|\{\mathfrak{f}_1, \mathfrak{f}_2\}\|_K = \|\mathfrak{f}_1\|_{K_1}^2 \|\mathfrak{f}_2\|_{K_2}^2$. $\mathscr{H}_K$ embeds all functions of type $\mathfrak{f}(x_1, x_2) = \sum_{i=1}^n \mathfrak{f}_1^{(i)}(x_1)\mathfrak{f}_2^{(i)}(x_2)$ with $\mathfrak{f}_1^{(i)}(x_1) \in \mathscr{H}_{K_1}$ and $\mathfrak{f}_2^{(i)}(x_2) \in \mathscr{H}_{K_2}$.

B Appendix

Description of the data generating system utilized in Section 5.3.5

In this appendix, we give the exact coefficient function matrices $\mathfrak{a}_i^{\circ}, \mathfrak{b}_j^{\circ}$ of the LPV process model utilized in Section 5.3.5, together with the coefficient function matrices $\mathfrak{c}_i^{\circ}, \mathfrak{d}_j^{\circ}$ of the considered noise dynamics, i.e., the corresponding full BJ model.

B.1 Coefficient functions of the process dynamics

$$\mathfrak{b}_0^{\circ}(p,t) = \begin{bmatrix} 1-\exp(-0.6p_1(t)) & 0.64-0.72\exp(0.7p_1(t)) \\ 0.3-0.4p_1^2(t)+0.5p_2(t) & 0.2+0.98\tan^{-1}(0.66p_2(t)) \end{bmatrix} \tag{B.1a}$$

$$\mathfrak{b}_1^{\circ}(p,t) = \begin{bmatrix} 0.24-0.32p_1^2(t)+0.4p_2(t-1) & 0.22\exp(0.4p_1(t-1)) \\ 0.16+0.9\tan^{-1}(0.63p_2(t)) & 0.22-0.5p_1^2(t)+0.45p_2(t-1) \end{bmatrix} \tag{B.1b}$$

$$\mathfrak{b}_2^{\circ}(p,t) = \begin{bmatrix} 0.16+0.64\tan^{-1}(0.8p_2(t-2)) & 0.14+0.7\tan^{-1}(0.6p_2(t-2)) \\ 0.64-0.64\exp(-0.6p_1(t-1)) & 0.17-0.32p_1^2(t)+0.32p_2(t-1) \end{bmatrix} \tag{B.1c}$$

$$\mathfrak{a}_1^{\circ}(p,t) = \begin{bmatrix} 0.2+0.12p_2^2(t-1) & 0 \\ 0 & 0.2+0.35\tan^{-1}(p_1(t))\cos(p_1(t-1)) \end{bmatrix} \tag{B.1d}$$

$$\mathfrak{a}_2^{\circ}(p,t) = \begin{bmatrix} 0.19+0.15\tan^{-1}(p_1(t-1))\cos(p_2(t-2)) & 0 \\ 0 & 0.17+0.11p_2^2(t-1) \end{bmatrix}. \tag{B.1e}$$

B.2 Coefficient functions of the noise dynamics

$$\mathfrak{d}_1^{\mathrm{o}}(p,t) = \begin{bmatrix} 0.3 + 0.3\sqrt{|(p_1(t))|} & 0 \\ 0 & 0.45 + 0.45\sin(p_2(t)) \end{bmatrix} \tag{B.2a}$$

$$\mathfrak{d}_2^{\mathrm{o}}(p,t) = \begin{bmatrix} 0.34 + 0.34\sin(p_2(t-1)) & 0 \\ 0 & 0.23 + 0.23\sqrt{|p_1(t-2)|} \end{bmatrix} \tag{B.2b}$$

$$\mathfrak{c}_1^{\mathrm{o}}(p,t) = \begin{bmatrix} 0.3 + 0.45p_1^3(t) + 0.3p_1^2(t-1) & 0 \\ 0 & 0.3 + 0.45p_2^2(t-1) \end{bmatrix} \tag{B.2c}$$

$$\mathfrak{c}_2^{\mathrm{o}}(p,t) = \begin{bmatrix} 0.24 + 0.36p_1^2(t-1) & 0 \\ 0 & 0.24 + 0.36p_2^3(t-2) + 0.24p_2^2(t-1) \end{bmatrix}. \tag{B.2d}$$

Bibliography

H.S. Abbas, R. Tóth, M. Petreczky, N. Meskin, and J. Mohammadpour. Embedding of nonlinear systems in a linear parameter-varying representation. In *Proc. of the 19th IFAC World Congress*, pages 6907–6913, Cape Town, South Africa, Aug. 2014.

F. Abbasi, J. Mohammadpour, R. Tóth, and N. Meskin. A support vector machine-based method for LPV-ARX identification with noisy scheduling parameters. In *Proc. of the 13th IEEE European Control Conference (ECC)*, pages 370–375, Strasbourg, France, June 2014.

F. Abbasi, J. Mohammadpour, R. Tóth, and N. Meskin. A Bayesian approach for model identification of LPV systems with uncertain scheduling variables. In *Proc. of the 54th IEEE Conference on Decision and Control (CDC)*, pages 789–794, Osaka, Japan, Dec. 2015.

E. M. Abdel-Rahman, A. H. Nayfeh, and Z. N. Masoud. Dynamics and control of cranes: A review. *Modal Analysis*, 9(7):863–908, 2003.

R.P. Aguilera, B.I. Godoy, J.C. Agüero, G.C. Goodwin, and J.I. Yuz. An EM-based identification algorithm for a class of hybrid systems with application to power electronics. *International Journal of Control*, 87(7):1339–1351, 2014.

H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19(6):716–723, 1974.

J. Antoni and J. Schoukens. A comprehensive study of the bias and variance of frequency-response-function measurements: Optimal window selection and overlapping strategies. *Automatica*, 43(10):1723–1736, 2007.

A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto. The connection between Bayesian estimation of a Gaussian random field and RKHS. *IEEE Trans. on Neural Networks and Learning Systems*, 26(7):1518–1524, 2015.

A. Argyriou and F. Dinuzzo. A unifying view of representer theorems. In *Proc. of the 31th Internatinal Conference on Machine Learning (ICML)*, pages 748–756, Beijing, China, June 2014.

N. Aronszajn. Theory of reproducing kernels. *Trans. of the American mathematical society*, 68(3):337–404, 1950.

A. A. Bachnas, R. Tóth, A. Mesbah, and J. Ludlage. A review on data-driven linear parameter-varying modeling approaches: A high-purity distillation column case study. *Journal of Process Control*, 24(4):272–285, 2014.

B. Bamieh and L. Giarré. Identification of linear parameter varying models. *International Journal of Robust and Nonlinear Control*, 12(9):841–853, 2002.

G. Belforte, F. Dabbene, and P. Gay. LPV approximation of distributed parameter systems in environmental modeling. *Environmental Modeling & Software*, 20(8): 1063–1070, 2005.

M Bertero. Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, 75:1–120, 1989.

M. Bertero, T. A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, 1988.

C. M. Bishop. *Pattern recognition and machine learning*. Springer-Verlag New York, 2006.

C. Bissell. A history of automatic control. In Shimon Y. Nof, editor, *Springer Handbook of Automation*, pages 53–69. Springer Berlin Heidelberg, 2009.

L. Breiman. Better subset regression using the nonnegative garotte. *Technometrics*, 37(4):373–384, 1995.

F. P. Carli, A. Chiuso, and G. Pillonetto. Efficient algorithms for large scale linear system identification using stable spline estimators. In *Proc. of the 16th IFAC Symposium on System Identification*, pages 119–124, Brussels, Belgium, July 2012.

F. P. Carli, T. Chen, and L. Ljung. Maximum entropy kernels for system identification. *IEEE Trans. on Automatic Control*, 62(3):1471 – 1477, 2017.

B. P. Carlin and T. A. Louis. *Bayes and empirical Bayes methods for data analysis*. CRC Press, 2000.

V. Cerone and D. Regruto. Set-membership identification of LPV models with uncertain measurements of the time-varying parameter. In *Proc. of the 47th IEEE Conference on Decision and Control (CDC)*, pages 4491–4496, Cancun, Mexico, Dec. 2008.

T. Chen and L. Ljung. Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, 49(7): 2213–2220, 2013.

T. Chen and L. Ljung. Constructive state space model induced kernels for regularized system identification. In *Proc. of the 19th IFAC World Congress*, pages 1047 – 1052, Cape Town, South Africa, Aug. 2014.

T. Chen and L. Ljung. Regularized system identification using orthonormal basis functions. In *Proc. of the 14th European Control Conference (ECC)*, pages 1291 – 1296, Linz, Austria, July 2015a.

T Chen and L Ljung. On kernel structure for regularized system identification (ii): a machine learning perspective. In *Proc. of the 17th IFAC symposium on system identification*, pages 1035–1040, Beijing, China, Oct. 2015b.

T Chen and L Ljung. On kernel structure for regularized system identification (ii): a system theory perspective. In *Proc. of the 17th IFAC symposium on system identification*, pages 1041–1046, Beijing, China, Oct. 2015c.

T. Chen and L. Ljung. On kernel design for regularized LTI system identification. *ArXiv e-prints: 1612.03542*, 2016.

T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and Gaussian processes–revisited. *Automatica*, 48(8):1525–1535, 2012.

T. Chen, A. Chiuso, G. Pillonetto, and L. Ljung. Rank-1 kernels for regularized system identification. In *Proc. of the 52nd IEEE Conference on Decision and Control (CDC)*, pages 5162–5167, Florence, Italy, Dec. 2013.

T. Chen, M. S. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto. System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Trans. on Automatic Control*, 59(11):2933–2945, 2014.

T. Chen, G. Pillonetto, A. Chiuso, and L. Ljung. Spectral analysis of the DC kernel for regularized system identification. In *Proc. of the 54th IEEE Conference on Decision and Control (CDC)*, pages 4017 – 4022, Osaka, Japan, Dec. 2015.

T. Chen, T. Ardeshiri, F. P. Carli, A. Chiuso, L. Ljung, and G. Pillonetto. Maximum entropy properties of discrete-time first-order stable spline kernel. *Automatica*, 66:34 – 38, 2016.

A. Chiuso. The role of vector autoregressive modeling in predictor-based subspace identification. *Automatica*, 43(6):1034–1048, 2007.

A. Chiuso, T. Chen, L. Ljung, and G. Pillonetto. On the design of multiple kernels for nonparametric linear system identification. In *Proc. of the 53rd IEEE Conference on Decision and Control (CDC)*, pages 3346–3351, Los Angeles, USA, Dec. 2014.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2001.

M. A. H. Darwish, P. B. Cox, G. Pillonetto, and R. Tóth. Bayesian identification of LPV Box-Jenkins models. In *Proc. of the 54th IEEE Conference on Decision and Control (CDC)*, pages 66–71, Osaka, Japan, Dec. 2015a.

M. A. H. Darwish, G. Pillonetto, and R. Tóth. Perspectives of orthonormal basis functions based kernels in Bayesian system identification. In *Proc. of the 54th IEEE Conference on Decision and Control (CDC)*, pages 2713–2718, Osaka, Japan, Dec. 2015b.

M. A. H. Darwish, A. A. Bachnas, G. Pillonetto, and R. Tóth. Bayesian identification of LPV series-expansion models. *To be submitted to Automatica*, 2017a.

M. A. H. Darwish, P. B. Cox, I. Proimadis, G. Pillonetto, and R. Tóth. Prediction-error identification of LPV systems: A nonparametric Gaussian regression approach. *Submitted to Automatica*, 2017b.

M. A. H. Darwish, J. Lataire, and R. Tóth. Bayesian frequency domain identification of LTI systems with OBFs kernels. In *Proc. of the 20th IFAC World Congress*, pages 6412–6417, Toulouse, France, July 2017c.

M. A. H. Darwish, G. Pillonetto, and R. Tóth. The quest for the right kernel in Bayesian impulse response identification: The use of OBFs. *Accepted for Automatica*, 2017d.

M. P. Deisenroth. *Efficient reinforcement learning using Gaussian processes*. PhD thesis, Karlsruhe Institute of Technology, 2010.

E. Deprettere and P. Dewilde. Orthogonal cascade realization of real multiport digital filters. *International Journal of Circuit Theory and Applications*, 8(3):245–272, 1980.

M. Dettori and C. W. Scherer. LPV design for a CD player: An experimental evaluation of performance. In *Proc. of the 40th IEEE Conference on Decision and Control (CDC)*, pages 4711–4716, Orlando, Florida, USA, Dec. 2001.

G. Dimitriadis and J. E. Cooper. Flutter prediction from flight flutter test data. *Journal of Aircraft*, 38(2):355–367, 2001.

F. Dinuzzo. Kernels for linear time invariant system identification. *SIAM Journal on Control and Optimization*, 53(5):3299–3317, 2015.

M. Djemai and M. Defoort, editors. *Hybrid Dynamical Systems: Observation and Control*. Lecture Notes in Control and Information Sciences, Vol. 457. Springer International Publishing, 2015.

F. Donida, C. Romani, F. Casella, and M. Lovera. Towards integrated modeling and parameter estimation: an LFT-Modelica approach. In *Proc. of the 15th IFAC Symposium on System Identification*, pages 1286–1291, Saint-Malo, France, July 2009.

R. Duijkers, R. Tóth, D. Piga, and V. Laurain. Shrinking complexity of scheduling dependencies in LS-SVM based LPV system identification. In *Proc. of the 53rd IEEE Conference on Decision and Control (CDC)*, pages 2561–2566, Los Angeles, CA, USA, Dec. 2014.

L. Giarré, D. Bauso, P. Falugi, and B. Bamieh. LPV model identification for gain scheduling control: An application to rotating stall and surge control problem. *Control Engineering Practice*, 14(4):351–361, 2006.

F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural computation*, 10(6):1455–1480, 1998.

K. Glover, R. F. Curtain, and J. R. Partington. Realisation and approximation of linear infinite-dimensional systems with error bounds. *SIAM Journal on Control and Optimization*, 26(4):863–898, 1988.

A. Golabi, N. Meskin, R. Tóth, and M. Mohammadpour. A Bayesian approach for estimation of LPV linear-regression models. In *Proc. of the 53rd IEEE Conference on Decision and Control (CDC)*, pages 2555–2560, Los Angeles, CA, USA, Dec. 2014.

A. Golabi, N. Meskin, R. Tóth, and J. Mohammadpour. A Bayesian approach for LPV model identification and its application to complex processes. *IEEE Trans. on Control Systems Technology*, (99):1–8, 2017.

G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

G. C. Goodwin, Michel. Gevers, and Brett. Ninness. Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Trans. on Automatic Control*, 37(7):913–928, 1992.

G. C. Goodwin, J. H. Braslavsky, and M. M. Seron. Non-stationary stochastic embedding for transfer function estimation. *Automatica*, 38(1):47–62, 2002.

J. Hadamard. *Lectures on Cauchy's problem in linear partial differential equations*. New Haven: Yale university press, 1923.

J. D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1):39 – 70, 1990.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009.

A. J. Helmicki, C. A. Jacobson, and C. N. Nett. Control oriented system identification: a worst-case/deterministic approach in $\mathcal{H}_\infty$. *IEEE Trans. on Automatic Control*, 36(10):1163–1176, 1991.

P. S. C. Heuberger, P. M. J. Van den Hof, and O. H. Bosgra. A generalized orthonormal basis for linear dynamical systems. *IEEE Trans. on Automatic Control*, 40(3): 451–465, 1995.

P. S. C. Heuberger, P. M. J. Van den Hof, and Bo Wahlberg. *Modeling and Identification with Rational Orthonormal Basis Functions*. Springer-Verlag, 2005.

H. Hjalmarsson and L. Ljung. A unifying view of disturbances in identification. In *Proc. of the 10th IFAC Symposium on System Identification*, pages 73–78, Copenhagen, Denmark, July 1994.

H. Hjalmarsson, J. S. Welsh, and C. R. Rojas. Identification of Box-Jenkins models using structured ARX models and nuclear norm relaxation. In *Proc. of the 16th IFAC Symposium on System Identification*, pages 322–327, Brussels, Belgium, July 2012.

H. Hochstadt. *Integral Equations*. John Wiley & Sons, Inc., 1988.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

K. Hsu, K. Poolla, and T. Vincent. Identification of structured nonlinear systems. *IEEE Trans. on Automatic Control*, 53(11):2497–2513, 2008.

A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.

E. W. Kamen and B. S. Heck. *Fundamentals of Signals and Systems Using the Web and Matlab (3rd ed.)*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 2007.

G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

G. Kitagawa and W. Gersch. A smoothness priors-state space modeling of time series with trend and seasonality. *Journal of the American Statistical Association*, 79(386):378–389, 1984.

G. Kitagawa and W. Gersch. A smoothness priors long AR model method for spectral estimation. *IEEE Trans. on Automatic Control*, 30(1):57–65, 1985.

G. Kitagawa and W. Gersch. *Smoothness Priors Analysis of Time Series*. Springer-Verlag New York, 1996.

J. Kocijan. *Modelling and Control of Dynamic Systems Using Gaussian Process Models*. Advances in Industrial Control. Springer International Publishing, Ljubljana, Slovenia, 2016.

T. Kondo, S. Yamaoka, and Y. Ohta. A hyper-parameter estimation algorithm in Bayesian system identification usinf OBFs-based kernels. In *Proc. of the 20th IFAC World Congress*, Toulouse, France, July 2017.

B. Kulcsár, J. Dong, J. W. van Wingerden, and M. Verhaegen. LPV subspace identification of a DC motor with unbalanced disc. In *Proc. of the 15th IFAC Symposium on System Identification*, pages 856–861, Saint-Malo, France, July 2009.

J. Lataire and T. Chen. Transfer function and transient estimation by Gaussian process regression in the frequency domain. *Automatica*, 72:217–229, 2016.

J. Lataire and R. Pintelon. Frequency domain weighted nonlinear least squares estimation of continuous-time, time-varying systems. *IET Control Theory & Applications*, 5(7):923–933, 2011.

J. Lataire, R. Pintelon, and E. Louarroudi. Non-parametric estimate of the system function of a time-varying system. *Automatica*, 48(4):666 – 672, 2012.

J. Lataire, R. Pintelon, D. Piga, and R. Tóth. Continuous-time linear time-varying system identification with a frequency-domain kernel-based estimator. *IET Control Theory & Applications*, 11(4):457–465, 2017.

V. Laurain, M. Gilson, R. Tóth, and H. Garnier. Refined instrumental variable methods for identification of LPV Box-Jenkins models. *Automatica*, 46(6):959–967, 2010.

V. Laurain, R. Tóth, W-X. Zheng, and M. Gilson. Nonparametric identification of LPV models under general noise conditions: An LS-SVM based approach. In *Proc. of the 16th IFAC Symposium on System Identification*, pages 1761–1766, Brussels, Belgium, July 2012.

V. Laurain, R. Tóth, D. Piga, and M. A. H. Darwish. Model structure learning for LPV-IO identification: An RKHS approach. *To be submitted to Automatica*, 2017.

H. Leeb and B. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59, 2005.

L. Ljung. *System Identification, theory for the user*. Prentice-Hall, 2nd edition, 1999.

L. Ljung. *System Identification Toolbox, for use with Matlab*. The Mathworks Inc., 2006.

P. Lopes dos Santos, T. P. Azevedo-Perdicoúlis, J. A. Ramos, S. Deshpande, D. E. Rivera, and J. L. Martins de Carvalho. LPV system identification using a separable least squares support vector machines approach. In *Proc. of the 53rd IEEE Conference on Decision and Control (CDC)*, pages 2548–2554, Los Angeles, CA, USA, Dec. 2014.

D. J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural computation*, 11(5):1035–1068, 1999.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

A. Marconato, M. Schoukens, and J. Schoukens. Filter interpretation of regularized impulse response modeling. In *Proc. of the 15th IEEE European Control Conference (ECC)*, pages 1655–1660, Aalborg, Denmark, June 2016.

A. Marconato, M. Schoukens, , and J. Schoukens. Filter-based regularisation for impulse response modelling. *IET Control Theory & Applications*, 11(2):194–204, 2017.

A. Marcos and G. J. Balas. Development of linear-parameter-varying models for aircraft. *Journal of Guidance, Control and Dynamics*, 27(2):218–228, 2004.

A. B. Marcovitz. *Linear time-varying discrete time systems*. The Franklin Institute, Lancaster, PA., 1964.

J. S. Maritz and T. Lwin. *Empirical Bayes Methods with Applications*. Chapman and Hall/CRC, 1989.

T. McKelvey and G. Guérin. Non-parametric frequency response estimation using a local rational model. In *Proc. of the 16th IFAC Symposium on System Identification*, pages 49 –54, Brussels, Belgium, July 2012.

J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Trans. of the Royal Society of London*, 209:415–446, 1909.

M. Milanese and A. Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty : An overview. *Automatica*, 27(6):997–1009, 1991.

V. Nalbantoglu, J. Bokor, G. Balas, and P. Gaspar. System identification with generalized orthonormal basis functions: an application to flexible structures. *Control Engineering Practice*, 11(3):245–259, March 2003.

O. Nelles. *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Springer Berlin Heidelberg, 2001.

D. Nguyen-Tuong, M. Seeger, and J. Peters. Real-time local GP model learning. In O. Sigaud and J. Peters, editors, *From Motor Learning to Interaction Learning in Robots*, pages 193–207. Springer Berlin Heidelberg, 2010.

B. Ninness, H. Hjalmarsson, and F. Gustafsson. Generalized Fourier and Toeplitz results for rational orthonormal bases. *SIAM Journal on Control and Optimization*, 37(2):429–460, 1999.

B. M. Ninness and F. Gustafsson. A unifying construction of orthonormal bases for system identification. *IEEE Trans. on Automatic Control*, 42(4):515–521, 1997.

T. Oliveira e Silva. Rational orthonormal functions on the unit circle and on the imaginary axis, with applications in system identification. Technical Report ver. 1.0 a, Automatic Control Group, Royal Institute of Technology, Stockholm, Sweden, 1995.

T. Oliveira e Silva. A n-width result for the generalized orthonormal basis function model. In *Proc. of the 13th IFAC World Congress*, pages 375–380, Sydney, Australia, July 1996.

S. Paoletti, A. Lj. Juloski, Ferrari-Trecate G, and R. Vidal. Identification of hybrid systems: A tutorial. *European Journal of Control*, 13(2):242–260, 2007.

S. C. Patwardhan, S. Manuja, S. Narasimhan, and S. L. Shah. From data to diagnosis and control using generalized orthonormal basis filters. part II: Model predictive and fault tolerant control. *Journal of Process Control*, 16(2):157–175, 2006.

R. K. Pearson. *Discrete-time dynamic models*. Oxford University Press, 1999.

D. Petelin and J. Kocijan. Control system with evolving Gaussian process models. In *Proc. of the IEEE Workshop on Evolving and Adaptive Intelligent Systems (EAIS)*, pages 178–184, Paris, France, Apr. 2011.

D. Piga and R. Tóth. LPV model order selection in an LS-SVM setting. In *Proc. of the 52nd IEEE Conference on Decision and Control (CDC)*, pages 4128–4133, Florence, Italy, Dec. 2013.

D. Piga, P. Cox, R. Tóth, and V. Laurain. LPV system identification under noise corrupted scheduling and output signal observations. *Automatica*, 53:329–338, 2015.

G. Pillonetto. Identification of time-varying systems in reproducing kernel Hilbert spaces. *IEEE Trans. on Automatic Control*, 53(9):2202–2209, 2008.

G. Pillonetto and A. Aravkin. A new kernel-based approach for identification of time-varying linear systems. In *Proc. of the 27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Reims, France, Sept. 2014.

G. Pillonetto and B. Bell. Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance. *Automatica*, 43(10):1698–1712, 2007.

G. Pillonetto and A. Chiuso. Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator. *Automatica*, 58:106–117, 2015.

G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

G. Pillonetto and G. De Nicolao. Kernel selection in linear system identification part I: A Gaussian process perspective. In *Proc. of the 50th IEEE Conference on Decision and Control and European Control Conference*, pages 4318–4325, Orlando, FL, USA, Dec. 2011.

G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47 (2):291–305, 2011a.

G. Pillonetto, M. H. Quang, and A. Chiuso. A new kernel-based approach for nonlinear system identification. *IEEE Trans. on Automatic Control*, 56(12):2825–2840, 2011b.

G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682, 2014.

R. Pintelon and J. Schoukens. *System identification, a frequency domain approach*. Wiley-IEEE Press, 2012.

R. Pintelon, J. Schoukens, G. Vandersteen, and K. Barbé. Estimation of nonparametric noise and FRF models for multivariable systems-part I: Theory. *Mechanical Systems and Signal Processing*, 24(3):573–595, 2010a.

R. Pintelon, J. Schoukens, G. Vandersteen, and K. Barbé. Estimation of nonparametric noise and FRF models for multivariable systems-part II: Extensions, applications. *Mechanical Systems and Signal Processing*, 24(3):596–616, 2010b.

R. Pintelon, E. Louarroudi, and J. Lataire. Nonparametric time-variant frequency response function estimates using arbitrary excitations. *Automatica*, 51:308–317, 2015.

T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

G. Prando, D. Romeres, and A. Chiuso. Online identification of time-varying systems: a Bayesian approach. In *Proc. of the 55th IEEE Conference on Decision and Control (CDC)*, pages 3775–3780, Las Vegas, USA, Dec. 2016.

I. Proimadis, HJ Bijl, and J. W. van Wingerden. A kernel based approach for LPV subspace identification. In *Proc. of the 1st IFAC Workshop on Linear Parameter-Varying Systems*, pages 97–102, Grenoble, France, Oct. 2015.

J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(Dec.): 1939–1959, 2005.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

P. A. Regalia, S. K. Mitra, and P. P. Vaidyanathan. The digital all-pass filter: a versatile signal processing building block. *Proceedings of the IEEE*, 76(1):19–37, 1988.

S.Z. Rizvi, J. Mohammadpour, R. Tóth, and N. Meskin. An IV-SVM-based approach for identification of state-space LPV models under generic noise conditions. In *Proc. of the 54th IEEE Conference on Decision and Control (CDC)*, pages 7380–7385, Osaka, Japan, Dec. 2015a.

S.Z. Rizvi, J. Mohammadpour, R. Tóth, and N. Meskin. A kernel-based approach to MIMO LPV state-space identification and application to a nonlinear process plant. In *Proc. of the 1st IFAC Workshop on Linear Parameter-Varying Systems*, pages 85–90, Grenoble, France, Oct. 2015b.

S.Z. Rizvi, J. Mohammadpour, F. Abbasi, R. Tóth, and N. Meskin. State-space LPV model identification using kernelized machine learning. *In print, Automatica*, 2017.

C. R. Rojas, R. Tóth, and H. Hjalmarsson. Sparse estimation of polynomial and rational dynamic models. *IEEE Trans. on Automatic Control*, 59:2962–2977, 2013. Special issue.

W. Rugh and J. S. Shamma. Research on gain scheduling. *Automatica*, 36(10): 1401–1425, 2000.

C. W. Scherer. Mixed $\mathcal{H}_2/\mathcal{H}_\infty$ control for time-varying and linear parametrically-varying systems. *International Journal of Robust and Nonlinear Control*, 6(9-10): 929–952, 1996.

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. Cambridge MA, USA: MIT Press, 2002.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *Computational Learning Theory*, pages 416–426. Springer Berlin Heidelberg, 2001.

J. Schoukens, R. Pintelon, and Y. Rolain. Time domain identification, frequency domain identification. equivalencies! differences? In *Proc. of the IEEE American Control Conference (ACC)*, pages 661–666, Boston, Massachusetts, USA, June 2004.

J. Schoukens, Y. Rolain, and R. Pintelon. Analysis of windowing/leakage effects in frequency response function measurements. *Automatica*, 42(1):27–38, 2006.

P. J. Schreier and L. L. Scharf. *Statistical signal processing of complex-valued data: the theory of improper and noncircular signals*. Cambridge University Press, 2010.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

J. S. Shamma and M. Athans. Gain scheduling: potential hazards and possible remedies. *IEEE Control Systems Magazine*, 12(3):101–107, 1992.

S. Skogestad and I. Postlethwaite. *Multivariable feedback control-analysis and design*. John Wiley and Sons, 1996.

T. Söderström and P. Stoica. *System identification*. Prentice-Hall, 1989.

M.D. Spiridonakos and S.D. Fassois. Parametric identification of a time-varying structure based on vector vibration response measurements. *Mechanical Systems and Signal Processing*, 23(6):2029 – 2048, 2009. Special Issue: Inverse Problems.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 9(6):1135–1151, 1981.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. on Information Theory*, 52(10):4635–4643, 2006.

A. Stenman, F. Gustafsson, D. E. Rivera, L. Ljung, and T. McKelvey. On adaptive smoothing of empirical transfer function estimates. *Automatica*, 8(11):1309–1315, 2000.

H. Sun. Mercer theorem for RKHS on noncompact sets. *Journal of Complexity*, 21 (3):337–349, 2005.

J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

A. Svensson, J. Dahlin, and T. B. Schön. Marginalizing Gaussian process hyperparameters using sequential Monte Carlo. In *Proc. of the 6th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 477–480, Cancun, Mexico, 2015.

M. Sznaier and C. Mazzarro. An LMI approach to control-oriented identification and model (in)validation of LPV systems. *IEEE Trans. on Automatic Control*, 48 (9):1619–1624, 2003.

R. Tibshirani. Regression shrinkage and selection with the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

A. Tikhonov and V. Arsenin. *Solutions of ill-posed problems*. Washington DC: Winston, Wiley, 1977.

R. Tóth. *Modeling and Identification of Linear Parameter-Varying Systems, an Orthonormal Basis Function Approach*. PhD thesis, Delft University of Technology, 2008.

R. Tóth. *Modeling and Identification of Linear Parameter-Varying Systems*. Lecture Notes in Control and Information Sciences, Vol. 403. Springer, Heidelberg, 2010.

R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof. Asymptotically optimal orthonormal basis functions for LPV system identification. *Automatica*, 45(6): 1359–1370, 2009a.

R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof. An LPV identification framework based on orthonormal basis functions. In *Proc. of the 15th IFAC Symposium on System Identification*, pages 1328–1333, Saint-Malo, France, July 2009b.

R. Tóth, C. Lyzell, M. Enqvist, P. S. C. Heuberger, and P. M. J. Van den Hof. Order and structural dependence selection of LPV-ARX models using a nonnegative garrote approach. In *Proc. of the 48th IEEE Conference on Decision and Control (CDC)*, pages 7406–7411, Shanghai, China, Dec. 2009c.

R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof. On the discretization of LPV state-space representations. *IET Control Theory & Applications*, 4:2082–2096, 2010.

R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof. LPV system identification using series-expansion models. In P. L. dos Santos, C. Novara, D. Rivera, J. Ramos, and T-P. Perdicoúlis, editors, *Linear Parameter-Varying System Identification: New Developments and Trends*, pages 259–294. World Scientific Publishing, Singapore, 2011a.

R. Tóth, V. Laurain, W-X. Zheng, and K. Poolla. Model structure learning: A support vector machine approach for LPV linear-regression models. In *Proc. of the 50th IEEE Conference on Decision and Control (CDC)*, pages 3192–3197, Orlando, Florida, USA, Dec. 2011b.

R. Tóth, P. S. C. Heuberger, and P. M. J. Van den Hof. Pediction error identification of LPV systems: present and beyond. In J. Mohammadpour and C. W. Scherer, editors, *Control of Linear Parameter Varying Systems with Applications*, pages 27–60. Springer, Heidelberg, 2012a.

R. Tóth, H. Hjalmarsson, and C. R. Rojas. Order and structural dependence selection of LPV-ARX models revisited. In *Proc. of the 51th IEEE Conference on Decision and Control (CDC)*, pages 6271–6276, Maui, Hawaii, USA, Dec. 2012b.

R. Tóth, V. Laurain, M. Gilson, and H. Garnier. Instrumental variable scheme for closed-loop LPV model identification. *Automatica*, 48(9):2314–2320, 2012c.

M. K. Tsatsanis and G. B. Giannakis. Time-varying system identification and model validation using wavelets. *IEEE Trans. on Signal Processing*, 41(12):3512–3523, 1993.

A. van der Maas, R. van der Maas, R. Voorhoeve, , and T. Oomen. Frequency response function identification of LPV systems: A 2d-LRM approach with application to a medical X-ray system. In *Proc. of the IEEE American Control Conference (ACC)*, pages 4598–4603, Boston, MA, USA, July 2016a.

R. van der Maas, A. van der Maas, and T. Oomen. Accurate frequency response function identification of LPV systems: A 2D local parametric modeling approach. In *Proc. of the 54th IEEE Conference on Decision and Control (CDC)*, pages 1465–1470, Osaka, Japan, Dec. 2015.

R. van der Maas, A. van der Maas, R. Voorhoeve, and T. Oomen. Accurate FRF identification of LPV systems: nD-LPM with application to a medical X-ray system. *IEEE Trans. on Control Systems Technology*, (99):1–12, 2016b.

P. van Overschee and B. de Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Application*. Kluwer Academic Press, 1996.

J. W. van Wingerden and M. Verhaegen. Subspace identification of bilinear and LPV systems for open- and closed-loop data. *Automatica*, 45(2):372–381, 2009.

V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

M. Verhaegen and X. Yu. A class of subspace model identification algorithms to identify periodically and arbitrarily time-varying systems. *Automatica*, 31(2): 201 – 216, 1995.

M. Vidyasagar. *Control System Synthesis: A Factorization Approach*. The MIT Press, Cambridge, Massachusetts, 1985.

G. Wahba. *Spline models for observational data*. Siam, Philadelphia, PA, USA, 1990.

B. Wahlberg. System identification using Laguerre models. *IEEE Trans. on Automatic Control*, 36(5):551–562, 1991.

L. Wang and W. R. Cluett. Use of PRESS residuals in dynamic system identification. *Automatica*, 32(5):781–784, 1996.

G. M. Wassink, M. van de Wal, C. Scherer, and O. Bosgra. LPV control for a wafer stage: beyond the theoretical solution. *Control Engineering Practice*, 13(2):231–245, 2005.

X. Wei. *Advanced LPV techniques for Diesel Engines*. PhD thesis, Johannes Kepler University, Linz, 2006.

S. Wollnack, H. S. Abbas, R. Tóth, and H. Werner. Fixed-structure LPV-IO controllers: An implicit representation based approach. *In print, Automatica*, 2017.

G. G. Yin, S. Kan, L. Y. Wang, and C. Z. Xu. Identification of systems with regime switching and unmodeled dynamics. *IEEE Trans. on Automatic Control*, 54(1): 34–47, 2009.

N. Young. *An introduction to Hilbert space*. Cambridge University Press, Cambridge, UK, 1988.

Y. Zhang and W.E. Leithead. Exploiting hessian matrix and trust-region algorithm in hyperparameters estimation of Gaussian process. *Applied Mathematics and Computation*, 171(2):1264 – 1281, 2005.

Yu Zhao, Biao Huang, Hongye Su, and Jian Chu. Prediction error method for identification of LPV models. *Journal of Process Control*, 22(1):180 – 193, 2012.

K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice-Hall, 1995.

Y. Zhu and G. Ji. LPV model identification using blended linear models with given weightings. In *Proc. of the 15th IFAC Symposium on System Identification*, pages 1674–1679, Saint-Malo, France, July 2009.

Y. Zhu and X. Xu. A method of LPV model identification for control. In *Proc. of the 17th IFAC World Congress*, pages 5018–5024, Seoul, Korea, July 2008.

List of Symbols

Operators

q	Forward time shift
$\circledast$	Convolution
$\mathscr{Z}$	Z-transformation
$\mathscr{F}$	Fourier-transformation
tr	Trace operator
z	z-domain variable
cov	Covariance operator
var	Variance operator
dim	Dimension
min	Minimum
max	Maximum
det	Determinant
$\bullet^{\mathrm{u}}, \bullet^{\mathrm{y}}$	Superscripts u, y denote the association with the input or output, respectively
$\hat{\bullet}$	The hat denotes an estimate of some quantity
$\bar{\bullet}$	Truncated representation of some quantity / mean value
$\mathcal{E}\{\bullet\}$	Mean operator
diag	Diagonal matrix operator
Span	Algebraic span
$\diamond$	p-dependent dynamic relation
$\langle \bullet, \bullet \rangle$	Inner product
$[\bullet]_{i,j}$	Element of the i-th row and j-th column
$[\bullet]_i$	The i-th element
$\bullet^\top$	Transposition
$\bullet^*$	Complex conjugate
$\bullet^{-1}$	Inverse

$\bullet^H$	Hermitian transpose
$\bullet^\dagger$	Unilateral inverse
$\bullet^\perp$	Orthogonal complement
$\mathfrak{L}$	Integral operator
$\otimes$	Direct product

Dynamic systems

$\mathcal{G}$	Dynamic system
$\mathcal{S}$	LPV system
$\mathcal{F}$	DT FD-LTI system

Signals and variables

t	Discrete time variable
u	Input signal
u^{f}	Filtered input signal
y	Output signal
$\breve{y}$	Noiseless output
p	Scheduling variable
$\bar{\mathrm{p}}$	Point of the scheduling space
ω	Frequency
Ω	Generalized frequency variable
θ	Parameters vector
θ_0	True parameter vector values
γ_{r}	Regressor
Υ	Regressor matrix

Spaces and fields

$\mathbb{Z}$	Set of integers
$\mathbb{N}$	Set of natural numbers
$\mathbb{C}$	Set of complex numbers
$\mathbb{R}$	Set of real numbers
$\mathcal{K}$	Set of DFT-frequency indices that lies in the frequency band of interest
Θ	Parameter vector space
$\mathscr{X}$	Generic set/domain
$\mathbb{U}$	Input domain
$\mathbb{Y}$	Output domain
$\mathbb{P}$	Scheduling signal domain

$\mathbb{D}$	Unit disk
$\mathbb{J}$	Unit circle
$\mathbb{E}$	Exterior of the unit disk
$L_2(\mathbb{J})$	Hilbert space of square integrable functions on $\mathbb{J}$
$\mathscr{H}$	Hilbert space
$\mathscr{H}_K$	RKHS associated with kernel K
$\mathscr{H}_2(\mathbb{E})$	Space of square integrable functions on $\mathbb{J}$ and analytic in $\mathbb{E}$
$\mathscr{RH}_2(\mathbb{E})$	Space of all $\mathscr{H}_2(\mathbb{E})$ functions that have real-valued impulse responses
$\mathscr{RH}_{2-}(\mathbb{E})$	Space of all $\mathscr{RH}_2(\mathbb{E})$ functions that are real, proper, rational and finite-order
$\mathscr{R}\hat{\mathscr{H}}_{2-}(\mathbb{E})$	Augmented $\mathscr{RH}_{2-}(\mathbb{E})$ space with a feedthrough term
$\ell_2(\mathbb{Z})$	Space of square summable sequences on $\mathbb{Z}$
$\ell_2(\mathbb{N})$	Space of all $\ell_2(\mathbb{Z})$ sequences that are causal
$\mathcal{R}\ell_2(\mathbb{N})$	Space of all $\ell_2(\mathbb{N})$ sequences that are real
$\mathcal{R}\ell_1(\mathbb{N})$	Space of absolutely summable real sequences
$\ell_\infty(\mathbb{N})$	Space of all bounded sequences on $\mathbb{N}$
$\mathcal{U}_n$	Generic subspace

Poles and Eigenvalues

λ	System pole
Λ	Pole set
$\breve{\lambda}$	Pole of the TF G / Eigenvalue

Dimensions

$n_{\mathbb{U}}$	Input dimension
$n_{\mathbb{Y}}$	Output dimension
$n_{\mathbb{P}}$	Scheduling dimension
$n_{\mathfrak{f}}$	Maximum truncation order
$n_{\mathfrak{f}_\mathrm{y}}$	Truncation order of the y-sided expansion
$n_{\mathfrak{f}_\mathrm{u}}$	Truncation order of the u-sided expansion
N	Data set length
n_g	Inner function dimension
n_e	Number of basis extension
n_θ	Parameter vector dimension
$n_{\mathbb{X}}$	Dimension of a signal x
n	Dimension or length
$n_\mathrm{a}, \ldots, n_\mathrm{d}$	Order of polynomials / Matrix polynomials
n_ψ	Number of considered basis

n_χ	Number of nodes of $\mathbb{P}$
n_{m}	Order of the monomials basis functions
n_ϕ	Number of basis functions

Functions

G	Transfer function/operator
$\mathcal{W}$	OBFs coefficient functions
$\mathfrak{a}^{\mathrm{o}}, \ldots, \mathfrak{d}^{\mathrm{o}}$	Matrix functions of p (true)
$\mathfrak{g}, \mathfrak{h}, \mathfrak{h}_{y,i}, \mathfrak{h}_{u,i}$	Matrix functions of p
$A_0, \ldots, D_0$	Matrix polynomial in q^{-1}
S	Stochastic process
$\mathcal{R}$	Outcomes of the sample space of a stochastic process
$R_{\mathrm{A}}, \ldots, R_{\mathrm{F}}$	Polynomial functions, classical model parameterization
$\mathcal{D}, \mathfrak{D}$	Nonnegative functions that decay to zero
$K, \mathcal{Q}$	Kernel function/kernel slice
$\mathcal{L}$	Linear kernel
$\mathcal{W}$	Criterion function
g	Generic unknown function
F	Complex function/transfer function
$\mathfrak{g}$	Impulse response function
$\mathscr{L}$	Impulse functional
$\mathcal{V}$	Loss function
$\mathscr{C}$	Data fit term of the loss function
$\mathscr{R}$	Regularizer term of the loss function
$\mathscr{l}_2$	Quadratic loss function
$\mathfrak{T}$	Transfer function of the transient effect
H	Transfer function associated with noise
$\mathcal{H}, \mathcal{H}_{\mathrm{b}}$	Inner function
ϕ	General basis function
φ	Eigenfunction
Ψ	OBFs set in the time domain
$\breve{\Psi}$	OBFs set in the frequency domain
ψ	Orthonormal basis function in the time domain
$\breve{\psi}$	Orthonormal basis function in the frequency domain
$\mathfrak{f}$	LPV function / predictor
$\mathcal{O}$	The complexity of an algorithm

Measures and Norms

d	Metric
$\|\cdot\|_n$	n-th vector norm
μ	Borel measure
ℓ_0	The support of a vector

Distributions

$\mathcal{U}$	Uniform distribution
$\mathcal{N}$	Normal distribution
p	Probability distribution
$\mathcal{GP}$	Gaussian process distribution
$\mathcal{RCGP}$	Real Complex Gaussian process distribution

Coefficients, constants and rates

j	Imaginary unit
ς	Arbitrary number
τ	Sampled signal index
v, e	Noise, stochastic process
ϵ	Noise/ error/ residual
$\mathcal{V}$	Noise signal vector
σ	Standard deviation
Σ_e, Σ_ν	Covariance matrix of the noise and output, respectively
$\mathcal{I}, \mathcal{I}_y, \mathcal{I}_\mathrm{y}$	Index set
$i, j, k, \iota, l, \mathrm{i}, \mathrm{j}, \nu$	Index variables
Y_N, Y	Output signal vector
U_N, U	Input signal vector
P	Scheduling signal vector
X_N, X	x vector
I_N	Identity matrix of size $N \times N$
$\mathfrak{b}, \mathfrak{c}$	Kautz basis unknowns
$\mathcal{k}$	Kernel vector
$\mathcal{K}$	Kernel matrix
$\mathcal{K}^\mathrm{o}$	Output kernel matrix
γ	Regularization parameter
γ_s	Regularization parameter for sparsity
Γ	Length scale matrix

β	Hyperparameter vector
β_{d}	Hyperparameter that describes the decay rate
β_{c}	Hyperparameter that describes the decay rate of the expansion coefficients
β_{p}	Hyperparameter vector of the OBFs generating poles
c	Expansion coefficient/representer coefficient
$\mathscr{V}$	Vector form of Takenaka-Malmquist basis
ρ_{b}	Convergence rate
κ	OBFs bound
q^{o}, q	True and parameterized delay in the input channel, respectively
$\mathfrak{m}_j$	Nodes of $\mathbb{P}$
$\overline{\mathcal{J}}_i$	Maximum absolute value of a coefficient function over the nodes of $\mathbb{P}$
m	Mean of a RCGP process
K_{cov}	Covariance of a RCGP process
K_{rel}	Relation of a RCGP process
$\mathbb{K}_{\mathcal{R}}$	Frequency indices for real values of TF
η	Realization of a RCGP process
β_α	Kernel scaling parameter
β_{w}	SE kernel width

Data sets

$\mathcal{D}_N$	Data record of length N

List of Abbreviations

AIC	Akiake Information Criterion
ARMAX	Auto-Regressive Moving Average model with eXternal signal
ARX	Auto-Regressive models with eXogenous input
BIBO	Bounded-Input Bounded-Output (stability)
BFR	Best Fit Rate
BIC	Bayesian Information Criterion
BJ	Box Jenkins (model)
CS	Cubic Spline (kernel)
CT	Continuous Time
CV	Cross Validation
DC	Diagonal Correlated (kernel)
DI	Diagonal (kernel)
DFT	Discrete Fourier Transform
DT	Discrete Time
DTFT	Discrete-Time Fourier Transform
ETFE	Empirical Transfer Function Estimate
FcM	Fuzzy c-Means
FD	Finite Dimensional
FIR	Finite Impulse Response
FKcM	Fuzzy Kolmogorov c-Max
FRF	Frequency Response Function
GCV	Generalized Cross-Validation
GOBFs	Generalized Orthonormal Basis Functions
GP	Gaussian Process

GPR	Gaussian Process Regression
GPTF	Gaussian Process Transfer Function
IIR	Infinite Impulse Representation
IO	Input/Output
IV	Instrumental Variable
LASSO	Least Absolute Shrinkage and Selection Operator
LDS	Linear Dynamic Systems
LGP	Local Gaussian Process
LPM	Local Polynomial Method
LPV	Linear Parameter-Varying
LRM	Local Rational Method
LS	Least-Squares (criterion)
LTI	Linear Time-Invariant
LTV	Linear Time-Varying
KnW	Kolmogorov n-width
MAP	Maximum a Posterior
MC	Monte Carlo
MIMO	Multi Input Multi Output
MISO	Multi Input Single Output
MSE	Mean Squared Error
NL	Nonlinear
NNG	Non-Negative Garotte
OBFs	Orthonormal Basis Functions
OE	Output Error (model)
PDF	Probability Density Function
PEM	Prediction Error Minimization
PMSEs	Post Model Selection Estimators
PRESS	Predicted Residual Sums of Squares
PV	Parameter Varying
PWA	Piecewise Affine
RBF	Radial Basis Function (kernel)
RIV	Refined Instrumental Variable
RKHSs	Reproducing Kernel Hilbert Spaces
ReLS	Regularized Least-Squares

RFIR	Regularized Finite Impulse Response
RCGP	Real/Complex Gaussian Process
RN	Regularization Networks
ROC	Region Of Convergence
ROBFs	Regularized Orthonormal Basis Functions
SA	Switched Affine
SE	Squared Exponential (kernel)
SISO	Single Input Single Output
SNR	Signal-to-Noise Ratio
SRIV	Simplified Refined Instrumental Variable
S-RKHS	Sparse Reproducing Kernel Hilbert Space
SS	Stable Spline (kernel)
SURE	Stein's Unbiased Risk Estimator
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TC	Tuned Correlated (kernel)
TF	Transfer Function
TI	Time Invariant
TV	Time Varying
VAF	Variance Accounted For
ZOH	Zero Order Hold

List of Publications

Journal papers

M. A. H. Darwish, G. Pillonetto, and R. Tóth. The quest for the right kernel in Bayesian impulse response identification: The use of OBFs. *Accepted for Automatica*, 2017.

M. A. H. Darwish, P. B. Cox, I. Proimadis, G. Pillonetto, and R. Tóth. Prediction-error identification of LPV systems: A nonparametric Gaussian regression approach. *Submitted to Automatica*, 2017.

V. Laurain, R. Tóth, D. Piga, and M. A. H. Darwish. Model structure learning for LPV-IO identification: An RKHS approach. *To be submitted to Automatica*, 2017.

M. A. H. Darwish, A. A. Bachnas, G. Pillonetto, and R. Tóth. Bayesian identification of LPV series-expansion models. *To be submitted to Automatica*, 2017.

M. A. H. Darwish and H. S. Abbas. DC motor speed and position control using discrete-time fixed-order H_∞ controllers. *International Journal on Information management*, 1(1):1–13, 2012.

M. A. H. Darwish, H. S. Abbas, A. I. Saleh, and M. M. M. Hassan. FLC implementation on a 8-bit microcontroller for DC motor speed and position control. *Journal of Engineering Sciences, Assiut University*, 39(2):405–423, 2011.

Conference proceedings

S. Chitraganti and M. A. H. Darwish. Performance evaluation aspects of an approximate nonlinear event-based state estimator by sequential Monte Carlo approach. In *Proc. of the 56th IEEE Conference on Decision and Control*, Melbourne, Australia, Dec. 2017.

M. A. H. Darwish, J. Lataire, and R. Tóth. Bayesian frequency domain identification of LTI systems with OBFs kernels. In *Proc. of the 20th IFAC World Congress*, pages 6412–6417, Toulouse, France, July 2017.

M. A. H. Darwish, P. B. Cox, G. Pillonetto, and R. Tóth. Bayesian identification of LPV Box-Jenkins models. In *Proc. of the 54th IEEE Conference on Decision and Control (CDC)*, pages 66–71, Osaka, Japan, Dec. 2015.

M. A. H. Darwish, G. Pillonetto, and R. Tóth. Perspectives of orthonormal basis functions based kernels in Bayesian system identification. In *Proc. of the 54th IEEE Conference on Decision and Control (CDC)*, pages 2713–2718, Osaka, Japan, Dec. 2015.

M. A. H. Darwish and H. S. Abbas. DC motor position control using discrete-time fixed-order H_∞ controllers. In *Proc. of the 1st International Conference on Innovative Engineering Systems*, pages 294–299, Alexandria, Egypt, Dec. 2012.

Peer-Reviewed Abstracts

M. A. H. Darwish, J. Lataire, R. Tóth, and P. M. J. Van den Hof. Bayesian frequency domain identification of LTI systems with OBFs kernels. In *36th Benelux Meeting on Systems and Control*, page 18, "Sol-Cress", Spa, Belgium, 2017.

M. A. H. Darwish, S. Chitraganti, T. B. Schön, R. Tóth, and P. M. J. Van den Hof. Maximum likelihood estimation of LPV-SS models: A sequential Monte-Carlo approach. In *35th Benelux Meeting on Systems and Control*, page 81, Soesterberg, The Netherlands, 2016.

M. A. H. Darwish, R. Tóth, and P. M. J. Van den Hof. Selecting shaping kernels in Bayesian identification of LTI systems: An orthonormal basis functions approach. In *34th Benelux Meeting on Systems and Control*, page 28, Lommel, Belgium, 2015.

M. A. H. Darwish, R. Tóth, and P. M. J. Van den Hof. Learning models and controllers from data: A philosophical perspective. In *33rd Benelux Meeting on Systems and Control*, page 68, Heijden, The Netherlands, 2014.

Acknowledgments

First of all, I would like to thank my daily supervisor Roland for his constant guidance and support during the past four and half years, starting with writing a Ph.D. project proposal till defending my thesis. Roland, you are really a special person who contributes significantly to my life, both from the academic and personal perspectives. I really enjoyed sharing my thoughts with you during our regular meetings. I have learned from you how to be critical and precise in presenting my ideas. Feeling safe, that nothing is going to be wrong, during my Ph.D. trajectory was the greatest thing you gave to me. Also, I would like to express my great gratitude to my promotor, Paul, it was an honor to be one of your students. I am always inspired by you as a teacher, researcher and team leader. With your critical thinking and "helicopter view", you could always make things more clear to me. Thanks also for your valuable suggestions which have improved a lot the readability of this thesis especially the introduction part.

A special thank goes to prof. G. Pillonetto, who contributed a lot to the results presented in this thesis. Thank you so much for being always available to answer my questions in details, which in many cases seemed very basic. I kindly appreciate the time you provided to read a chapter from my thesis. I am also grateful to John Lataire for the fruitful cooperation and for providing me with his code for Bayesian frequency domain identification.

In addition, I would like to convey my appreciations to dr. Hossam Abbas, who has the most prominent role in my life starting with supervising my M.Sc. project till introducing me to Roland who accepted me as one of his students based on the recommendation of dr. Abbas. I will not forget also when you waited for my first arrival at the station of Eindhoven. You were my guide who made the life easier for me and for my family.

I would like to thank the members of my Ph.D. committee prof.dr.ir. R. Pintelon, prof.dr.ir. J. Suykens, dr. C. Rojas, prof.dr. S. Weiland and dr.ir. T.J. Tjalkens, for the effort that they have spent during the summer vacation to read my thesis. Your comments have greatly improved my thesis. Special thanks goes to prof.dr.ir. R. Pintelon for the great discussion we had at the VUB and for the delicious lunch.

A special mention goes to Barbara who we sadly lost in 2016 - thank you for helping me a lot during the initial phase of my Ph.D. Diana and Lucia, thank you for taking care of the administration load away from us. Diana, we really enjoyed the Belgian chocolate that you brought to my family.

I would like to thank all my old and new colleagues for providing such a pleasant working atmosphere in the Control Systems group. I really enjoyed each moment I spent with you. Special thanks go to Koen, my office mate in Potentiaal and my new colleague at Bright Society for everything, especially the Dutch-English translation. Pepijn, you are a special colleague to me. You were always there when I needed you. I learned many things from you, especially, how to provide high-quality figures in Latex, which improved a lot the outlook of my work. Ioannis, what a great colleague you are. Thank you for the many discussions we had together and for the fruitful cooperation. Marcella, I will miss you so much. Thank you, Rian, for helping me to better understand OBFs and providing me with your code to obtain the optimal generating poles, Mohsin for the great discussion we had at our long coffee break, Esmaeil for the amazing discussions we always have, I really enjoy talking to you. Your thoughts always amaze me.

My appreciation is extended to my Egyptian friends and their families in The Netherlands, who have provided a great support for me and my family: Hossam Morsy, Ahmed Elkholy, Mohamed Ezz, Ahmed Shahin, Medo, Ahmed Omar. We never felt that we are alone without our family, you were always there at our tough times. Mahmoud Abdelrahim, thank you so much for the amazing squash games we had together.

I also owe a great deal to Leon, Marcel and Stijn the founders of Bright Society. Signing a contract for an industrial job nine months in advance significantly improved my performance, pushing me to focus only on writing my thesis without worrying about the future. You provide me with a great support to shape my future and to find my passion.

The financial support of the Culture Affairs and Mission Sector, Ministry of Higher Education and Scientific Research, Government of Egypt, is also gratefully acknowledged. Without such support, this research would not have happened.

My deepest gratitude goes to my parents for forgetting themselves and dedicating their life for us. My father, you are the greatest man I have ever known taking care of all the annoying details in our life. My Mother, your blessing and prayer for me is what keeps me going. I am also grateful to my lovely sisters for their support and kindness.

Finally, and most importantly, I would like to thank my beloved wife Shaimaa. Her support, encouragement, quiet patience and unwavering love were undeniably the bedrock upon which the past eight years of my life have been built. Thank you for giving me two handsome boys, Yassen and Ahmed. They are really the light of my life. Thank you for dedicating your life for me and for our sons.

Eindhoven, October 2017
Mohamed A. H. Darwish

Curriculum Vitae

Mohamed A. H. Darwish was born on January 12, 1984, in Assiut, Egypt. He received his B.Sc. in Electrical Engineering (Computers and Control) from Assiut University, Assiut, Egypt in 2005. His B.Sc. graduation project was entitled Design and Implementation of an Autonomous Robot. He stood the 1st among his colleagues and graduated with the highest distinction. He had the honor of receiving the Excellence award from both the Egyptian Engineering Syndicate and the Government of Egypt during the Science Day in 2005 and 2006, respectively. In 2006, he started at the same department as a Teaching Assistant, where he was responsible for teaching and supervising graduation projects. At the same time, he continued his studies at the same department and obtained his M.Sc. in Electrical Engineering (Systems and Control) in 2011. His M.Sc. thesis was entitled Microcontroller Implementation for DC Motor Speed and Position Control. In 2011, he received a promotion and became an Assistant Lecturer at the same department. In 2013, he wrote a research proposal together with dr. ir. Roland Tóth for the scholarships offered by the Ministry of Higher Education & Scientific Research, Egypt, for the Egyptian Universities staff to perform their Ph.D. at a prestigious foreign university. His proposal was accepted and he obtained a fully funded Ph.D. scholarship to continue his graduate studies toward a Ph.D. degree at the Eindhoven University of Technology.

Since September 2013, Mohamed has been working as a Ph.D. student on a project entitled Bayesian Identification of Linear Dynamic Systems: Synthesis of Kernels in the LTI Case and Beyond at the Control Systems research group, Eindhoven University of Technology, The Netherlands, under the supervision of prof. dr. ir. P. M. J. Van den Hof and dr. ir. Roland Tóth. His Ph.D. research has mainly focused on utilizing new developments stemming from the machine learning community to address the open problems and associated challenges of efficiently estimating accurate linear models of physical processes. During the period 2013-2016, he took graduate courses at the Dutch Institute of Systems and Control (DISC) and received the DISC certificate.

Mohamed married his lovely wife Shaimaa in 2010 and now they have two handsome boys Yassen and Ahmed.

His research interests include data-driven modeling of linear and nonlinear dynamic systems and machine learning.
E-mail: mohamed.a.h.darwish@gmail.com