University of Bergamo
Department of Engineering and applied sciences
Doctoral Thesis
XXXV Cycle

Learning to filtering: a comparison of data-driven solutions to the filtering design problem

Author:
Luca Maurelli

Supervisor:
Prof. Fabio Previdi

Co-supervisors:
Prof. Mirko Mazzoleni

September 30, 2022

"Data are an implicit model."

Simone Formentin
seminar on "Learning to control: history and challenges of direct data-driven design"
University of Bergamo, Italy
May 23, 2022

UNIVERSITY OF BERGAMO

Abstract

Department of Engineering and Applied Sciences

Doctor of Philosophy

Learning to filtering: a comparison of data-driven solutions to the filtering design problem

by Luca Maurelli

Considering dynamical systems, the problem of designing a good filter to estimate a variable of interest is a key topic in the control system community. Since the 1900s, its theory and practice has been studied and performed by experts of all fields, starting from analog frequency-selective filters, spanning to the development of the stochastic filtering theory of Kolmogorov-Wiener, and culminating into the celebrated Kalman filter for non-stationary linear systems. In recent years, also statistical methods have been employed to tackle down the computational burden when deriving approximant Bayesian solutions in the non-linear case. In all these methods, the key assumption is that a-priori information about the system under analysis is known, enabling a model-based design of the filter. In contrast, approaches that rely on experimental data have seen an always increasing interest. Nowadays, the standard solution to the filtering problem is based on a two-step approach: (i) first a data-driven system identification is performed, (ii) then a filter is designed based on the identified model. In this case, the main idea is that a model of the system is not available and a data-driven methodology is used to estimate the system. This thesis aims to further expand the knowledge about data-driven methods to the filtering design problem. In the first part, a review of the history and the stochastic Kalman filter theory is presented in details. The second part focuses on personal contributions. Firstly, the new direct data-driven methodology derived from a parametric-statistical framework is proposed by means of a work of review and reformulation of the literature. In the direct solution available data are used to identify directly a model of the filter, skipping the system identification step and solving the practical problem of estimating the unknown noise covariance matrices from data. The standard two-step approach and the new direct one-step paradigm are compared, highlighting the flaws of each. Then, practical methodologies to implement both solutions with respect to a common steady-state LTI system framework are explored. The thesis end with experimental tests performed in simulation applied on an univariate dynamical system with no exogenous input. The tests are then expanded to a multivariate system with exogenous input taken from one example of a industrial application.

Acknowledgements

I would like to thank all the people that supported me during the doctoral years.

First, my sincere gratitude goes to my supervisor Prof. Fabio Previdi for giving me the opportunity to get involved in research and industrial projects, for always being a role model in complicated social and working scenarios, and for sharing invaluable advices.

At the same time, I warmly thank Proff. Mirko Mazzoleni and Simone Formentin, who were always available in these years to answer my questions, help me guiding my education, and from whom I learned a lot to be a better researcher.

Also, I would like to thank all the members of my research group, the Control System and Automation Laboratory of the University of Bergamo. My former desk-mate and friend Matteo Scandella who shared with me one meal a week, even if costly, at the restaurant ‘L’Anonimo’. I am in gratitude with all the colleagues and friends with whom I had daily exchanges of ideas and opinions, Prof. Antonio Ferramosca, Davide Previtali, Nicholas Valceschini, Leandro Pitturelli, Marco Polver, and Davide Cesani. I thank also former Ph.D. students Michele Ermidoro and Gabriele Maroni who shared their experience, new Ph.D. students Beatrice and Pietro, all the visiting students and professors, the undergraduate students of the laboratory, and in general all the professionals with who I had the pleasure to collaborate with.

Finally I thank my family, my parents Silvia and Edoardo who has always supported me, my dear sister Mara, my lovely partner Giorgia, and my close friends.

Thank you for sharing positive vibes. This would not have been the same without you.

Contents

List of Figures

List of Tables

List of Algorithms

List of Symbols

Notation	**Description**
k	is the time index for discrete-time dynamical systems.
$\mathcal{S}$	is the dynamical system.
$\boldsymbol{f}(\cdot)$	is the transition equation of the system $\mathcal{S}$.
$\boldsymbol{h}_y(\cdot)$	is the measurement equation of the system $\mathcal{S}$.
z or z^{-1}	is the z-operator of the Z transform. It also used to denote the lag operator in discrete time domain.
$\boldsymbol{G}_0(z)$	is the signal model of data generation model, that is the system $\mathcal{S}$.
$\boldsymbol{H}_0(z)$	is the noise model of data generation model, that is the system $\mathcal{S}$.
$\mathcal{M}(\theta)$	is the parameterized model of the system $\mathcal{S}$.
$\boldsymbol{G}(z,\theta)$	is the parameterized signal model of $\mathcal{M}(\hat{\theta})$.
$\boldsymbol{H}(z,\theta)$	is the parameterized noise model of $\mathcal{M}(\hat{\theta})$.
$\boldsymbol{x}_k \in \mathbb{R}^{n_x \times 1}$	is the state of the system at time k. It is a hidden variable.
$\boldsymbol{x}_i^k \in \mathbb{R}$	is an alternative notation for the i-th element of the state of the system at time k.
$n_{\boldsymbol{x}} \in \mathbb{N}$	is the number of elements of the state variable (its dimension).
$\boldsymbol{u}_k \in \mathbb{R}^{n_u \times 1}$	is the input of the system at time k. It is a measurable variable.
$n_{\boldsymbol{u}} \in \mathbb{N}$	is the number of elements of the input variable (its dimension).
$\boldsymbol{y}_k \in \mathbb{R}^{n_y \times 1}$	is the output of the system $\mathcal{S}$ at time k. It is a measurable variable.
$n_{\boldsymbol{y}} \in \mathbb{N}$	is the number of elements of the output variable (its dimension).
$\boldsymbol{w}_k \in \mathbb{R}^{n_w \times 1}$	is the process noise of the system $\mathcal{S}$ at time k affecting the state equation
$n_{\boldsymbol{w}} \in \mathbb{N}$	is the number of elements of the process noise variable (its dimension).
$\boldsymbol{v}_k \in \mathbb{R}^{n_y \times 1}$	is the measurement noise of the system $\mathcal{S}$ at time k affecting the output equation
	Time variant
$\boldsymbol{A}_k \in \mathbb{R}^{n_x \times n_x}$	is the time-variant state matrix, also known as system matrix, of the system $\mathcal{S}$ at time k.
$\boldsymbol{B}_k \in \mathbb{R}^{n_x \times n_u}$	is the time-variant state-input matrix of the system $\mathcal{S}$ at time k.
$\boldsymbol{C}_k \in \mathbb{R}^{n_y \times n_x}$	is the time-variant output-state matri of the system $\mathcal{S}$ at time k.
$\boldsymbol{D}_k \in \mathbb{R}^{n_y \times n_u}$	is the time-variant output-input matrix, also known as feedthrough matrix, of the system $\mathcal{S}$ at time k.
$\boldsymbol{G}_k \in \mathbb{R}^{n_x \times n_w}$	is the time-variant state-process matrix, also known as noise shaping matrix, of the system $\mathcal{S}$ at time k.

Notation	**Description**
$\boldsymbol{Q}_k \in \mathbb{R}^{n_w \times n_w}$	is the time-variant process noise covariance matrix of the system $\mathcal{S}$ at time k.
$\boldsymbol{R}_k \in \mathbb{R}^{n_v \times n_v}$	is the time-variant measurement noise covariance matrix of the system $\mathcal{S}$ at time k.
	Time invariant
$\boldsymbol{A} \in \mathbb{R}^{n_x \times n_x}$	is the time-invariant state matrix, also known as system matrix, of the system $\mathcal{S}$.
$\boldsymbol{B} \in \mathbb{R}^{n_x \times n_u}$	is the time-invariant state-input matrix of the system $\mathcal{S}$.
$\boldsymbol{C} \in \mathbb{R}^{n_y \times n_x}$	is the time-invariant output-state matrix of the system $\mathcal{S}$.
$\boldsymbol{D} \in \mathbb{R}^{n_y \times n_u}$	is the time-invariant output-input matrix, also known as feedthrough matrix, of the system $\mathcal{S}$.
$\boldsymbol{G} \in \mathbb{R}^{n_x \times n_w}$	is the time-invariant state-process matrix, also known as noise shaping matrix, of the system $\mathcal{S}$.
$\boldsymbol{x}_0 \sim \mathcal{G}(\bar{\boldsymbol{x}}_0, \boldsymbol{P}_0)$	is the initial condition of the state $\boldsymbol{x}$ of the system $\mathcal{S}$ at time $k = 0$.
$\bar{\boldsymbol{x}}_0$	is the mean value of the initial condition $\boldsymbol{x}_0$ of the system $\mathcal{S}$ at time $k = 0$.
$\boldsymbol{P}_0 \in \mathbb{R}^{n_x \times n_x}$	is the covariance of the initial condition $\boldsymbol{x}_0$ of the system $\mathcal{S}$ at time $k = 0$.
$\bar{\boldsymbol{x}}_k \in \mathbb{R}^{n_x \times 1}$	is the mean value of the state $\boldsymbol{x}$ of the system $\mathcal{S}$ at time k.
$\boldsymbol{P}_k \in \mathbb{R}^{n_x \times n_x}$	is the covariance of the state $\boldsymbol{x}$ of the system $\mathcal{S}$ at time k.
$\hat{\boldsymbol{x}}_{k\|k} \in \mathbb{R}^{n_x \times 1}$	is the filtered estimate of the state $\boldsymbol{x}$ of the system $\mathcal{S}$ at time k given data up to time k.
$\hat{\boldsymbol{x}}_{k\|k-1} \in \mathbb{R}^{n_x \times 1}$	is the (one-step ahead) predicted estimate of the state $\boldsymbol{x}$ of the system $\mathcal{S}$ at time k given data up to time $k - 1$.
$\hat{\boldsymbol{x}}_k \in \mathbb{R}^{n_x \times 1}$	is a compact notation for $\hat{\boldsymbol{x}}_{k\|k}$.
$\hat{\boldsymbol{x}}_k^- \in \mathbb{R}^{n_x \times 1}$	is a compact notation for $\hat{\boldsymbol{x}}_{k\|k-1}$.
$\boldsymbol{K}_k \in \mathbb{R}^{n_x \times n_y}$	is the Kalman gain at time k.
$\boldsymbol{K}_\infty \in \mathbb{R}^{n_x \times n_y}$	is the steady-state Kalman gain.
$\boldsymbol{L} \in \mathbb{R}^{n_x \times n_y}$	is a steady-state, stabilizing, not optimal, filter gain.
$\hat{\boldsymbol{\epsilon}}_k^- \in \mathbb{R}^{n_y \times 1}$	is the output (one-step ahead) prediction error, i.e. the innovation $\boldsymbol{y}_k - \hat{\boldsymbol{y}}_k^-$.
$\boldsymbol{\epsilon}_k^- \in \mathbb{R}^{n_y \times 1}$	is a compact notation for the innovation $\boldsymbol{y}_k - \hat{\boldsymbol{y}}_k^-$.
$\hat{\boldsymbol{\delta}}_k^- \in \mathbb{R}^{n_x \times 1}$	is the state (one-step ahead) prediction error, i.e. the residual $\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^-$.
$\boldsymbol{\delta}_k \in \mathbb{R}^{n_x \times 1}$	is a compact notation for the residual $\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^-$.
$\boldsymbol{P}_\epsilon(\tau)$	is the autocorrelation function of the residual at time lag τ.
$\hat{\boldsymbol{P}}_\epsilon(\tau)$	is the estimated autocorrelation function of the residual at time lag τ.
$\hat{\boldsymbol{A}}_{\mathrm{IV}}$	is the estimation of matrix A of the standard solution using the instrumental variable method.
$\hat{C}_{\mathrm{IV}}$	is the estimation of matrix C of the standard solution using the instrumental variable method.
	Operators
$\otimes$	is the Kronecker product.
$\boldsymbol{P}^\mathsf{T}$	is transpose of the matrix $\boldsymbol{P}$.
$\boldsymbol{P}^{-1}$	is inverse of the matrix $\boldsymbol{P}$.
$\boldsymbol{P}_s$	is columnwise stacking of the matrix $\boldsymbol{P}$.

Notation	**Description**
	Datasets
$\mathcal{Z}$	is a generic dataset used for the identification of the system.
$\mathcal{D}$	is a generic dataset used for the identification of the data-driven solutions to the filtering design problem.
$\mathcal{D}_{\text{ID}}$	is the dataset (partition of $\mathcal{D}$) used for the identification of filter.
$\mathcal{D}_{\text{VL}}$	is the dataset (partition of $\mathcal{D}$) used for the validation of filter.
N_{ID}	is number of data used for the identification of the filter.
N_{VL}	is number of data used for the validation of the filter.
N	is the total number of available data for the identification and validation of the filter.
	Bayesian derivation probabilities
$\boldsymbol{x}_{0:k}$	is the sequence of state variables (with the initial condition) until time k, i.e. $\boldsymbol{x}_0, \boldsymbol{x}_1, \dots, \boldsymbol{x}_k$
$\boldsymbol{y}_{1:k}$	is the sequence of output variables until time k, i.e. $\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_k$
$\mu(\boldsymbol{x}_0)$	is the prior of the initial condition $\boldsymbol{x}_0$ of the system $\mathcal{S}$ at time $k = 0$.
$f(\boldsymbol{x}_{k+1}\|\boldsymbol{x}_k)$	is the transition probability from the given state $\boldsymbol{x}_k$ to the state $\boldsymbol{x}_{k+1}$, i.e. the dynamics of the system $\mathcal{S}$
$g(\boldsymbol{y}_k\|\boldsymbol{x}_k)$	is the observation probability of the output $\boldsymbol{y}_k$ given the state $\boldsymbol{x}_k$.
$\mu(\boldsymbol{x}_{0:k})$	is the prior of the system $\mathcal{S}$ at time k.
$p(\boldsymbol{y}_{1:k}\|\boldsymbol{x}_{0:k})$	is the likelihood or data distribution of the system $\mathcal{S}$ at time k.
$p(\boldsymbol{x}_{0:k}\|\boldsymbol{y}_{1:k})$	is the posterior distribution of the system $\mathcal{S}$ at time k.
$\mathscr{X}_k$	is a compact notation for $\boldsymbol{x}_{0:k}$
$\mathscr{Y}_k$	is a compact notation for $\boldsymbol{y}_{1:k}$
	Direct filtering
$n_z \in \mathbb{N}$	is the number of elements of the desidered variable (its dimension). Here the assumption $n_z = n_x$ is taken.
$z_k \in \mathbb{R}^{n_z \times 1}$	is the desired variable of the system at time k. It is a measurable variable for a limited amount of time.
$\boldsymbol{C}_2 \in \mathbb{R}^{n_z \times n_x}$	is the time-invariant desidered variable-state matrix of the system $\mathcal{S}$. Here the assumption $\boldsymbol{C}_2 = \boldsymbol{I}$ is taken.
$\mathcal{F}$	is the true model of the direct filter.
$\mathcal{H}$	is the hypothesis set of the solution to the direct filter containing all feasible parameterized models $\mathcal{M}(\boldsymbol{\theta}_{\text{DF}})$.
$\mathcal{M}(\boldsymbol{\theta}_{\text{DF}})$	is the (generic) feasible model of the direct filter parameterized by $\boldsymbol{\theta}_{\text{DF}}$.
$\boldsymbol{G}(z^{-1}; \boldsymbol{\theta}_{\text{DF}})$	is the signal model part of the direct filter model $\mathcal{M}(\boldsymbol{\theta}_{\text{DF}})$
$\boldsymbol{H}(z^{-1}; \boldsymbol{\theta}_{\text{DF}})$	is the noise model part of the direct filter model $\mathcal{M}(\boldsymbol{\theta}_{\text{DF}})$.
$\boldsymbol{\Theta}_{\text{DF}} \subset \mathbb{R}^d$	is the set of feasible parameter vectors of the direct filter solution. Here $\mathbb{R}^d$ is a compact notation for $\mathbb{R}^{n_x \times (n_x + n_y + n_u)}$.
$\boldsymbol{\theta}_{\text{DF}} \in \boldsymbol{\Theta}_{\text{DF}}$	is the (generic) parameter vector of the direct filter solution.
$J_{N_{\text{ID}}}(\boldsymbol{\theta}_{\text{DF}})$	is the objective function of the direct filter optimization problem.
$\hat{\mathcal{F}}$	is the identified direct filter $\mathcal{M}(\hat{\boldsymbol{\theta}}_{\text{DF}})$.
$\hat{\boldsymbol{\theta}}_{\text{DF}}$	is the identified parameter vector of the direct filter.
$\boldsymbol{A}^{[0]} \in \mathbb{R}^{n_x \times n_x}$	is a notation for $\boldsymbol{I}$
$\boldsymbol{A}^{[1]} \in \mathbb{R}^{n_x \times n_x}$	is a notation for $(\boldsymbol{I} - \boldsymbol{K}_\infty \boldsymbol{C})\boldsymbol{A}$
$\boldsymbol{B}_y^{[0]} \in \mathbb{R}^{n_x \times n_y}$	is a notation for $\boldsymbol{K}_\infty$

Notation	Description
$\boldsymbol{B}_{\boldsymbol{u}}^{[0]} \in \mathbb{R}^{n_x \times n_u}$	is a notation for $\mathbf{0}$
$\boldsymbol{B}_{\boldsymbol{u}}^{[1]} \in \mathbb{R}^{n_x \times n_u}$	is a notation for $(\boldsymbol{I} - \boldsymbol{K}_\infty \boldsymbol{C})\boldsymbol{B}$
$\boldsymbol{A}(z^{-1})$	is a matrix polynomial formed by $\boldsymbol{A}^{[0]} + \boldsymbol{A}^{[1]} z^{-1}$
$\boldsymbol{B}_{\boldsymbol{y}}(z^{-1})$	is a matrix polynomial formed by $\boldsymbol{B}_{\boldsymbol{y}}^{[0]}$
$\boldsymbol{B}_{\boldsymbol{u}}(z^{-1})$	is a matrix polynomial formed by $\boldsymbol{B}_{\boldsymbol{u}}^{[0]} + \boldsymbol{B}_{\boldsymbol{u}}^{[1]} z^{-1}$
$\boldsymbol{\theta}_0 \in \mathbb{R}^{n_x \times (n_x + n_y + n_u)}$	is a notation for the true parameter vector of the direct filter formed by $\left[\boldsymbol{A}^{[1]}, \boldsymbol{B}_{\boldsymbol{y}}^{[0]}, \boldsymbol{B}_{\boldsymbol{u}}^{[1]}\right]^\top$
$\boldsymbol{G}_0(z^{-1}; \boldsymbol{\theta}_0)$	is the signal model part of the true direct filter model.
$n_{\boldsymbol{\theta}_0} \in \mathbb{N}$	is the number of elements of the (free) true parameter vector of the direct filter: $n_{\boldsymbol{x}} \cdot (n_{\boldsymbol{x}} + n_{\boldsymbol{y}} + n_{\boldsymbol{u}})$.
$\boldsymbol{\rho}_k \in \mathbb{R}^{n_x}$	is the noise process $\boldsymbol{\delta}_k + \boldsymbol{e}_k$.

Introduction

Context of the thesis

The world is full of applications where there is a growing interest in collecting data and events. In fact, when correctly exploited, data provide some sort of insight on the process under analysis. Indeed, there is a widespread tendency to take advantage of this new piece of information in order to derive practical solutions to specific problems or, more abstractly, to enhance a preexisting knowledge base.

In signal processing theory data that evolve through time are called signals. In many engineering applications signals can be observed and measured directly by sensors, thus the related information is accessible. However, practical problems affect sensors in a way that the provided raw information cannot be relied on. Some examples include:

1. Sensors are by construction inaccurate. Their particular noisy nature makes the required piece of information to be hidden by an undesired disturbance.
2. The desired information cannot be collected at all. This happens, for instance, when sensors or the related data acquisition experiments are too expensive for the application on hand. Another case is the particular difficult nature of the experiments, e.g. the devices are exposed by critical environmental conditions that could either damage their functionality or prevent their acquisition capabilities. Common examples involve corrosion in chemical processes or deterioration due to thermal conditions in extreme working scenarios.
3. Lastly, there are also times where interesting physical quantities cannot be accessed directly, recalling the need to estimate them.

In all the mentioned cases, filtering refers to the technique or the algorithm that resolves the difficulty by estimating the signal of interest. Instead, the filtering design problem refers to the difficulty of finding the estimator that implements the filtering process. In general, filtering has been extensively studied over the last ten decades and, also due to its spread usage, plays a crucial role in control systems and signal processing communities.

Research and new contribution of the thesis

This thesis aims to expand the knowledge about data-driven solutions for the filtering design problem. Given the long history of filtering in the last century, the thesis first reviews its historical developments, with a focus on the stochastic setting. The overview includes the derivation of the state-of-the-art filtering solution in a LTI setting, e.g. the Kalman filter, and a brief overview of key concepts derived from intuition about the filtering problem in general. Then, the document contains three different contributions: **a)** the first one is a contribution of review and reformulation of the literature. In particular it introduces the standard data-driven approach and the new “direct” approach, in a common unifying framework, **b)** the second and third contributions are more practical, dealing with the derivation of the routines and related ingredients which implement the mentioned approaches, and **c)** the third and larger contribution is experimental, comparing the results of the studied approaches by means of Monte-Carlo simulations in a LTI univariate and multivariate case. The aforesaid contributions are now reviewed in more details.

The first contribution deals with framing the classical solutions to the filtering design problem derived through their history in a model-based paradigm. In more details, most of the methodologies dealing with the filtering problem take for granted that a perfect model of the system is available to the designer of the filter, which is used as ingredient to build the filter from it. This is not the case in practice, allowing the introduction to the data-driven two-step paradigm in the design of filtering solutions. In this setting, first a model of the system is identified by informative data, and, then, a filter is design based on the estimated model. In this context, a common misconception is that available system identification techniques permit the access to the full knowledge of the system, i.e. to the deterministic and the stochastic components of the system by means of estimating a proper state-space realization. Again, this is a misbelief in standard system identification procedures where the selection of the stochastic properties of the model of the system, namely the characteristics of the noise entering the state equation and the noise entering the measurement equations, is left to the designer of the filter. As a consequence, when the stochastic properties are not fine-tuned the performance of the filter worsen. Moreover, theoretical insights derived in the model-based paradigm may fail to be applied when the practical situation enforces different assumptions. To this end, an interesting area of research is found in deriving a new “direct” data-driven solution to the filtering design problem, where the sequential two-step approach described above is summarized in one-step, i.e. the direct design of the filter from an informative dataset. It is interesting to note that the direct term resembles in analogy with the direct design of controllers in the *Identification for Control* research field. Anyway, the introduction of the direct solution aims to bypass the identification of the system, allowing the estimation of the filter directly from data. Within this context, the questioning of pros and cons when developing the mentioned paradigms further motivate and fuel the research. For instance, some interesting questions do include:

- What is the impact w.r.t. filtering performance of the standard data-driven filtering solution when multiple sequential estimation routines take place?

- The problem of selecting a proper model structure and model complexity for the filter is naturally derived from the model of the system in the standard solution. What about the direct solution? Does the new solution bypass this difficulty or is the problem shifted to choosing a proper model structure and model complexity directly for the filter?
- When the real system is (is not) in the hypothesis set of the model family can the data-driven direct approach driven by data offer a valid alternative w.r.t. a filtering performance point of view?

The second contribution deals with the development of a practical implementation of the data-driven solution to the filtering design problem based on the two-step approach. In the unifying mathematical framework composed by a state-space representation of a LTI system, the acquisition of a informative input-output dataset for the design of the filter contains also measurements of the variable of interested (the one to be filtered), which are available for a limited amount of time. In other words, measurements of desirable variable are available for the design of the filter but not for its use later on. Following this rationale, the available dataset is used to first estimate a model of the system exploiting the new knowledge encoded in the samples of the desired variable. In this context, the deterministic components of the system, matrices $\boldsymbol{A}, \boldsymbol{C}$, are estimated by means of instrumental variable least squares routines derived specifically to this problem. Finally, also the noise components of the system, noise covariance matrices $\boldsymbol{Q}, \boldsymbol{R}$, are estimates as well from available data by means of a state-of-the-art auto least square solution based on the autocorrelation function of pseudo-innovations. In the end, a filter is designed from the estimated deterministic and stochastic components of the system following the standard Kalman filter theory.

Alongside the previous contribution, also the third contribution investigates with the practical implementation of the data-driven direct solution to the filtering design problem. In particular, the solution is based on the one-step approach, hence a direct solution. In this settings the available informative dataset is exploited in the design of filter directly, skipping the modeling of the system. In the aforementioned unifying LTI framework, when considering the steady-state case of the filter, the problem of choosing directly: **i)** a proper filter structure, and **ii)** a proper filter complexity, is tackled by the author proposing as a reference the same properties of the well-known best linear unbiased estimator, the Kalman filter. Additionally, it is observed how the new method does not require the sequential routines needed for the estimation or tuning of the required ingredients for the identification of the system components.

To end the thesis, a large part of research time and efforts has been dedicated to the experimental comparison of the mentioned filtering solutions. In particular, the LTI and steady-state case is considered when deriving the filters in order to have an unifying framework where different design paradigms can be compared in a fair manner. The first example under analysis includes an univariate system with no exogenous input. Later on, a second example compares the solutions with different layers of added complexity, namely: **i)** a multivariate system, e.g. the desired variable to be filtered is multivariate, and **ii)** to derive a design of a direct filter when an exogenous input is considered. Empirical results and insights are highlighted, giving attentions to the advantages or flaws of each solution. For instance, discussions deal with: **a)** missing constraints in the optimization routines of different filtering solutions to enforce their assumptions, **b)** numerical problems related to the scaling of dimensionality, and **c)** the role of estimation variance and its impact when performing multiple sequential estimation. In the end, results indicate that a direct solution is viable and, in some cases, its filtering performance can be superior to the standard solution.

Structure of the thesis

The remainder of the thesis is organized in two parts:

- Part I introduces the state of the art found in literature about the filtering design problem.
- Part II focuses to the research and new contributions developed during the doctoral study to the filtering design problem.

In particular, the two parts are organized in dedicated sections as follows:

- Chapter 1 reviews the developments of the filtering design problem highlighting the theoretical and technological advancements through its history.

- Chapter 2 tries to build some intuition to the filtering design problem in a framed academic pattern. The aim is to summarize the core concepts that were used through the historical developments reviewed in the previous chapter in order to have a deeper understanding of the problem in hand. The focus in primary on stochastic filtering and the Bayesian intuition.

- Chapter 3 is dedicated to a a complete derivation of the multivariate Kalman Filter in discrete-time. The filter serves as a baseline to the researched data-driven solutions.

- Chapter 4 introduces the first contribution of the thesis by giving context to the different existing paradigms that can be used to derive solutions to the filtering design problem. Model-based and data-driven paradigms are compared and their differences are highlighted. In the end, the new data-driven "direct" solution is introduced.

- Chapter 5 explains the second contribution of the thesis, that is the standard solution to the filtering design from data problem. This solution is divided the system identification step and the filter design step. Then, practical problems and misconceptions are reviewed and solved by exploiting researched methods.

- Chapter 6 explains the third contribution of the thesis, that is the direct solution to the filtering design from data problem. In this setting, the focus of the design is to bypass the common problems of the standard solution.

- Chapter 7 compares the derived data-driven solution in the last two chapters by means of simulated data in the context of different classes of LTI systems.

PART I

State of the art on the filtering design problem

CHAPTER 1

An historical review of the filtering problem

The following chapter is dedicated to the overview of the historical developments of the filtering problem in the last century. In particular, the presentation of the chapter is divided in:

- Section 1.1 introduces the filtering term and its modern meaning;
- Section 1.2 reviews the first derivation of filters with frequency-selective properties;
- Section 1.3 gives an overview of the stochastic theory applied to the filtering theory. Their merging culminates in the well-known Kalman theory.
- Section 1.4 is focused on the Bayesian interpretation of the filtering theory, born with the research field dedicated to derive approximant solutions in non-linear and non-Gaussian settings.

1.1 Meaning of the filtering term

Filtering in one form or another has been alongside human history for a very long time. In [1] a natural example is given: "for many centuries many attempts have been performed to remove the more visible of the impurities from the water, i.e. water purification, through a filtering process". For context, nowadays the Merriam-Webster online dictionary[1] gives a first meaning for the noun filter as:

> "a porous article or mass (as of paper or sand) through which a gas or liquid is passed to separate out matter in suspension".

In the context of engineering and modern sciences, usage of the word filter is extended to abstract entities named signals that represent physical quantities, for instance an electrical voltage. Nevertheless, there is still the notion of something passing a barrier, in the form perhaps of an electric circuit. In this sense, filtering can thus be understood by means of *signal processing*.

[1] Merriam-Webster is one of the oldest and most trusted English dictionary and thesaurus.

It is easy to think of engineering situations in which filtering of signals might be desired. For instance, in many control systems the control is derived by feedback, which involves processing measurements derived from the system. Frequently, these measurements contain random inaccuracies or are contaminated by unwanted signals. Following this rationale, filtering is thus necessary in order to make the control close to that desired.

1.2 Frequency selective filters

Early 1900s As described in the historical review of [1] and here adapted, the first filters were originally seen as circuits with *frequency-selective* behavior. Later, this kind of filtering became known as *Classical Filtering*. A fundamental example of these circuits is the series or parallel tuned circuit in electrical engineering, exploited in practice for the first time in 1900 by the Italian radio pioneer Guglielmo Marconi [51]. As another example, the "wave trap" was a crucial ingredient in early crystal sets. More sophisticated versions of this same idea are seen in the IF (*intermediate frequency*) strip of most radio receivers; here, tuned circuits, coupled by active elements such as transformers and amplifiers, are used to shape a pass-band of frequencies which are amplified, and a stop-band where attenuation occurs.

1920s-1940s More sophisticated collections of tuned circuits are necessary for many applications, and as an understandable result, caused the maturing of an extensive body of filter design theory. Some of its landmarks are constant k and m-derived filters invented in the early 1920s respectively by American engineers Campbell and Zobel [56, 178], and, later in the 1930s, Butterworth filters [55] by the British physicist Stephen Butterworth, Chebyshev filters, and elliptical filters by the German mathematician William Cauer and the American engineer Sidney Darlington [5, 64]. In the following years, there has been an extensive development of numerical algorithms for filter design. In this context, filters are designed to meet given specifications on amplitude and phase response. There are also constraints imposed on the filter structure. For instance, they include impedance levels, types of components, number of components, etc.

Nonlinear filtering ideas have also started to be applied to different applications for many years. These include, for example, **a)** the AM envelope detector, which is a combination of a diode and a low-pass filter, **b)** the automatic gain control (AGC) circuit using a low-pass filter and a nonlinear element, and **c)** the phase-locked-loop used for FM reception.

During these years, the notion of a filter as a device processing continuous-time signals and possessing frequency-selective behavior has been stretched by two major developments [1].

Among these developments, the first one is digital filtering [12, 22], made possible by new innovations in integrated circuit technology. Digital circuit components such as analog-to-digital and digital-to-analog converters, shift registers, read-only memories, and microprocessors began to appear alongside circuit modules used in classical filters. Even though the goals of digital and classical filtering are the same, how to achive these goals gives space to different solutions. In digital filtering, the minimization of the active element count, the size of inductors, the dissipation of the reactive elements, or the termination impedance mismatch is no longer sought. Instead, new objectives such as the word length, the round-off error, the number of wiring operations in construction, and the processing delay need to be minimized. Aside from the possible cost benefits, important advantages are observed by the fact that filter parameters can now be set and maintained to a high order of precision, thereby achieving filter characteristics that could not normally be obtained reliably with classical filtering, and that parameters can be easily reset or made adaptive with little extra cost. Again, some digital filters, incorporating microprocessors, can be time-shared to perform many simultaneous tasks effectively.

The second major development to the filtering problem consists in advancement in the theory. These innovations are extensively explored in Section 1.3.

1.3 Stochastic filtering theory

1940s-1960s A major development to the filtering problem came with the application of statistical ideas largely spurred by developments in theory. As described in the historical review of [1]:

> “The classical approaches to filtering postulate, at least implicitly, that the useful signals lie in one frequency band and unwanted signals, normally termed noise, lie in another, though on occasions there can be overlap. The statistical approaches to filtering, on the other hand, postulate that certain statistical properties are possessed by the useful signal and unwanted noise. Measurements are available of the sum of the signal and noise, and the task is still to eliminate by some means as much of the noise as possible through processing of the measurements by a filter.”

Stochastic filtering theory finds its roots in the early 1940s due to the pioneering work done by the American mathematician Norbert Wiener [217, 33], and the Russian mathematician Andrey Nikolayevich Kolmogorov [103, 212] about stochastic processes whose statistical properties do not change with time, i.e. *stationary processes*. For these processes it proved possible to relate the statistical properties of the useful signal and unwanted noise with their frequency domain properties. There is, thus, a conceptual link with classical filtering of Section 1.2. A significant aspect of the statistical approach is the definition of a measure of suitability or performance of a filter. Roughly the best filter is that which, on the average, has its output closest to the correct or useful signal. By constraining the filter to be linear and formulating the performance measure in terms of the filter impulse response and the given statistical properties of the signal and noise, it generally transpires that a unique impulse response corresponds to the best value of the measure of performance or suitability.

The statistical development of the theory later proved significant for the great activity of the next decades in signal estimation [94, 200]. It was also one of the greatest factor in bridging the gap between different fields such as communication theory and control theory due to the innovative statistical point of view, as acknowledged by Shannon's own dedication at the end of his magnificent papers [154, 155] in 1948 founding Information Theory:

> "Credit must also be given to Professor Norbert Wiener, whose elegant solution of the problems of prediction and filtering has considerably influenced writer's thinking in this field."

As reviewed also in [59], during the late 1940s and early 1950s, many credits should be also attributed to earlier works that lead the development of the stochastic filtering theory. These include, for instance, the publications of Bode and Shannon [52], Zadeh and Ragazzini [172, 173], Levinson [111], Swerling [32], and many others.

1960s As already noted, the stationary assumption for the underlying signal and noise processes is crucial to the *Kolmogorov-Wiener theory*. It was not until the late 1950s and early 1960s that a new theory arose that did not require the stationarity assumption. The theory emerged because of the inadequacy of the *Kolmogorov-Wiener* theory for coping with certain applications in which non stationarity of the signal and/or noise was intrinsic to the problem. The new theory soon acquired the name *Kalman theory* due to the publication of the *Kalman filter* (KF) [98] in 1960 and the subsequent *Kalman-Bucy filter* in 1961 [99], which describe the optimal recursive solution to the linear filtering problem. Under the LQG (Linear-Gaussian-Quadratic) assumption, Kalman filter was originally derived with the orthogonal projection method. It is worth noting that one reason for its success is that the Kalman filter can be understood and applied with very much lighter mathematical machinery than the Wiener filter.

It is also valuable to mention that the stationary theory was normally developed and thought of in frequency-domain terms, while the nonstationary theory was naturally developed and thought of in time-domain terms. Thus, the contact point between the two theories initially seemed slight. However, there is substantial contact, if for no other reason than that a stationary process is a particular type of nonstationary process. Indeed, despite the mathematical simplicity and generality, the Kalman filter [99] actually contains the Wiener filter as its limiting special case. In other words, in a stationary situation, the Kalman filter is precisely the Wiener filter for stationary least-squares smoothing [59], that is a time-variant Wiener filter [41]. At the present time, rapprochement of the two filtering theories is now easily achieved [189, 1, 94] through a connection between the computational tools used for each approach. In brief, the fundamental computational tool used in the Wiener theory is the study of the *spectral factorization* [151], while for the Kalman theory, it is the study of the *Riccati equation* [190, 191].

As noted above, the Kalman filtering theory was developed at a time when applications called for it, and the same comment is really true of the Wiener filtering theory. It is also pertinent to note that the problems of implementing Kalman filters and the problems of implementing Wiener filters were both consistent with the technology of their time [1]. Wiener filters were implementable with amplifiers and time-invariant network elements such as resistors and capacitors, while Kalman filters could be implemented with digital integrated circuit modules. The point of contact between the two recent streams of development, digital filtering and statistical filtering, comes when one is faced with the problem of implementing a discrete-time Kalman filter using digital hardware.

Without any exaggeration, it seems fair to say that the Kalman filter (and its numerous variants) have dominated the adaptive filter theory for decades in signal processing and control areas. Bearing in mind that the Kalman filter is limited by its assumptions, numerous non-linear filtering methods along its line have been proposed and developed to overcome its limitation. This fact led ultimately to fruitful research areas, still active today, of *Bayesian derivation* or *Bayesian interpretation* of filtering, explored in Section 1.4, due to natural way of deriving non-linear and non-Gaussian filters with optimal performance requirements.

1970s-1980s In the early 1970s, the study of filtering theory and in particular of the Kalman filter continued. A great contribution was due to Kailath [96], who used the innovation approach developed by Wold and Kolmogorov to reformulate the Kalman filter with the tool of *martingales theory*. This resulted in the rewriting the Kalman Filter into the well-known *innovation representation*. In this context, from an innovations point of view, the Kalman filter can be seen as a whitening filter or, in other words, the innovations process is defined as a white Gaussian noise process [96, 16]. Also, the innovations concept can be used straightforwardly in deriving non-linear filtering solutions, see for instance [37]. In particular, again from a innovations point of view, one of the criteria to justify the optimality of the solution to a nonlinear filtering problem is to check how white the pseudo-innovations are, with the property that the more white the more optimal.

At the same time, in the early 1970s, another interesting area of research was born due to the practical problems of applying the Kalman filtering theory, namely the assumptions of knowing the noise covariance matrices. Practical solutions to the estimation of the unknown parameters in conjunction with the design of a Kalman filter fall under the term of *Adaptive Filtering* due to the early works by Raman Mehra [123, 207] and Pierre Bélanger [47]. Adaptive filtering can also be seen as a way to adapt for process dynamics which are not modeled in the process model by means of time-varying parameters, allowing a flexible solution.

1.4 Bayesian filtering

> "The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening."
> — Thomas Bayes (1702-1761), [45]

1.4.1 Bayesian Theory

The historical survey about *Bayesian* theory is an adaptation from [59]. *Bayesian* theory was originally established by the British researcher Thomas Bayes in the posthumous publication "Essay towards solving a problem in the doctrine of chances" in 1763 [45]. The well-known Bayes theorem describes the fundamental probability law governing the process of logical inference. However, Bayesian theory has not gained its deserved attention in the early days until its modern form was rediscovered by the French mathematician Pierre-Simon de Laplace in "Théorie analytique des probailités", see the historical review [7].

Bayesian inference devoted to applying Bayesian statistics to statistical inference, has become one of the important branches in statistics, and has been applied successfully in statistical decision, detection and estimation, pattern recognition, and machine learning. In many scenarios, the solutions gained through Bayesian inference are viewed as *optimal*. Not surprisingly, Bayesian theory was also studied in the filtering literature.

1960s One of the first exploration of iterative Bayesian estimation is found in the 1960s, in the early stages of the history of the Kalman filter [98], by Ho and Lee' paper [91], where the principle and procedure of Bayesian filtering were specified. The Kalman filter was soon discovered to belong to the class of Bayesian filters [91, 92, 14, 19]. The corresponding Bayesian smoothers were soon developed, see [110, 145, 146]. An interesting historical detail [25] is that while Kalman and Bucy were formulating the linear theory in the United States, the Russian mathematician Ruslan Stratonovich was doing the pioneering work on the probabilistic (Bayesian) approach in Russia [14, 213]. The iterative application of Bayes rule to sequential parameter estimation was discussed in [161], where it was termed as "Bayesian learning". Bayesian approach to optimization of adaptive systems is discussed in [60, 115]. Bucy and Senne [54] also explored the point-mass approximation method in the Bayesian filtering framework.

1.4.2 Monte-Carlo methods

1990s In recent decades, Monte Carlo techniques have been rediscovered independently in statistics, physics, engineering, econometrics, biology, and many others, see for instance [180]. Originally, the research was started in the 1940s and 1950s during World World II, in Los Alamos Laboratory by John von Neumann, Stanislaw Ulam, Niick Metropolis, and others [124]. As explained by the authors:

> "The method is essentially, a statistical approach to the study of differential equations, or more generally, of integro-differential equations that occur in various branches of the natural sciences."

Later, in the 1990s, many new Monte Carlo methodologies (e.g. Bayesian bootstrap, hybrid Monte Carlo, quasi Monte Carlo) have since been rejuvenated and developed. Again, this active research field was aided once again by technological advancements, namely to the ever increasing raw computational power in the mid-1990s, which led to the feasibility of the approach. Indeed, the Monte Carlo technique is a kind of stochastic sampling approach aiming to tackle the complex systems which are analytically intractable. The power of Monte Carlo methods is that they can attack the difficult numerical integration problems. One of the attractive merits of sequential Monte Carlo approaches lies in the fact that they allow online estimation by combining the powerful Monte Carlo sampling methods with Bayesian inference, at an expense of reasonable computational cost. In particular, the sequential Monte Carlo approach has been used in parameter estimation and state estimation, for the latter of which it is sometimes called *Particle Filter*, a term coined in 1996 by Del Moral [68]. In particular, the basic idea of particle filter is to use a number of independent random variables called particles, sampled directly from the state space, to represent the posterior probability, and update the posterior by involving the new observations; the "particle system" is properly located, weighted, and propagated recursively according to the Bayesian rule. Thus, there is an intuitive genetic mutation-selection type mechanism that is generally found in Feynman-Kac models and their interacting particle interpretation models that can be used, for instance, in nonlinear filtering [8]. In other words, as stated by Del Moral:

> "These stochastic models are increasingly used to sample from complex high-dimensional distributions. They approximate a given target probability distributions by a large cloud of random samples termed particles. Practically, the particles evolve randomly around the space independently and to each particle is associated a positive potential function. Periodically we duplicate particles with high potentials at the expense of particles with low potentials which die."

For these reasons, and because various sampling variations has been developed in the literature, other terms found in the literature do include, for instance, "bootstrap filter", "genetic filter", and many others. In retrospect, the earliest idea of Monte Carlo method used in statistical inference can be found in earlier works, but the formal establishment describing the heuristic of particle filters is found in the journal article given by Gordon, Salmond and Smith in 1993 [84], who introduced certain novel resampling technique to the formulation. Instead, the first conference article presenting the heuristic of particle filters is from Kitagawa in 1996 [102] (a revisited version of a previous seminar work from 1993). The mathematical foundations, and the performance analysis of the discrete generation particle models are rather recent. The first rigorous study in this field seems to be the article published in 1996 on the applications of particle methods to non-linear estimation problems from De Moral [68], and later works, see for instance its personal review[2]. Instead, recent textbook about Bayesian non-linear filtering such as SMC (*Sequential Monte Carlo*) and particle filters can be found in [8, 9, 11, 25], in the tutorial papers [69, 211], and various review papers [59]. See also the references about SMC and particle filters from the resources list of Arnaud Doucet[3].

In conclusion, the historical development of filtering solutions in a stochastic framework and, later on, in a Bayesian framework is summarized by the chronological table in Table 1.1 (see in [59]) where different classes of methods, types of solutions and some comments on their properties are specified.

[2] The review of the literature work of De Moral can be found in a dedicated page of the author at the following link: "`https://people.bordeaux.inria.fr/pierre.delmoral/simulinks.html`".

[3] The resource list of Arnaud Doucet is available at the following link: "`https://www.stats.ox.ac.uk/~doucet/smc_resources.html`".

Table 1.1: A historical review of the filtering design problem in the stochastic setting, see [59].

Author(s)	**Method(s)**	**Type of solution**	**Comment**
Kolmogorov (1941) [103]	innovations	exact	linear, stationary
Wiener (1942) [33]	spectral factorization	exact	linear, stationary, infinite memory
Levinson (1946) [111]	lattice filter	approximate	linear, stationary, finite memory
Bode & Shannon (1950) [52]	innovations, whitening	exact	linear, stationary
Zadeh & Ragazzini (1950) [173]	innovations, whitening	exact	linear, non-stationary
Kalman (1960) [98]	orthogonal projections	exact	LQG, non-stationary, discrete
Kalman & Bucy (1961) [99]	recursive Riccati equation	exact	LQG, non-stationary, continuous
Stratonovich (1960) [162]	conditional Markov process	exact	nonlinear, non-stationary
Kushner (1967) [105]	PDE	exact	nonlinear, non-stationary
Zakai (1969) [175]	PDE	exact	nonlinear, non-stationary
Handschin & Mayne (1969) [86]	Monte Carlo	approximate	nonlinear, non-Gaussian, non-stationary
Bucy & Senne (1971) [54]	point-mass, Bayes	approximate	nonlinear, non-Gaussian, non-stationary
Kailath (1970) [96]	innovations	exact	linear, non-Gaussian, non-stationary
Beneš (1981) [48]	Beneš	exact solution of Zakai eqn.	nonlinear, finite-dimensional
Daum (1986) [65]	Daum, virtual measurement	exact solution of FPK eqn.	nonlinear, finite-dimensional
Gordon, Salmond, & Smith (1993) [103]	bootstrap, sequential Monte Carlo	approximate	nonlinear, non-Gaussian, non-stationary
Julier & Uhlmann (1997) [198]	unscented transformation	approximate	nonlinear, (non)-Gaussian, derivative-free

CHAPTER 2

Building intuition about filtering

> “I learned very early the difference between knowing the name of something and knowing something.”
>
> — Richard Feymann (1918-1988)

The following chapters tries to aim with the important aspect of developing a simple and natural *intuition* about what filtering really is and how to naturally and historically the filtering design problem was solved. The ideal and desirable result about the process of forming some intuition would be to ultimately have a deeper understanding of the few core concepts around filtering. Later on, these highlighted concepts are shown to be translatable to the design of more complex filters when using formal mathematical and statistical languages. It is also worth noting that the goal of forming a filtering intuition goes further than this, wishing that this imprinting could be also exploited in the derivation of alternate and/or new paradigms to the filtering design problem. For the understanding of subsequent chapters, this chapter may be omitted for experts in the field as the intuition discussion is detached from the expertise.

Returning to the content of the chapter, the main topic are stochastic filters which can be framed in a Bayesian perspective. Using simple terms, Bayesian refers to the management of probability knowledge where Bayesian probability determines what is likely to be true based on past information.

It is respectful to mention that the work herein takes inspiration from [18], where the original content is modified and adapted in order to make it more suitable for the current presentation. Nonetheless, the herein discussion is only the starting part of the full content. Therefore, please refer to [18] for a full reference.

In particular, the discussion of the chapter is divided in:

1. Section 2.1 starts the chapter by introducing some general concepts through a simple example about scales.

2. Section 2.2 presents a simple Bayesian filter known as the *g-h filter*. In the end, this filter is shown to be special case of the classical Kalman filter.

2.1 A simple example about noisy scales

Imagine a world without scales - the devices people stand on to weigh themselves. Then, scales are invented and the first time it is tried, for instance on yourself, the result is announced to be 80 kg. A second measurement, though, may give the result 82 kg. The problem here is that sensors are inaccurate. This is the motivation behind a huge body of work in filtering which provided solutions that have been developed over the last century, as seen in Chapter 1. These solutions were developed by asking very basic, fundamental questions into the nature of what it is know and how it is known.

The following example is an attempt to follow that journey of discovery, forming an intuition about filtering.

Another scale In the example above, it is possible to use another scale. Unfortunately, the new scale is inaccurate too: the first scale (A) reads 80kg, the second scale (B) reads 90kg. What can be concluded about the weight? There are some possible choices:

- Believe only scale (A), and assign 80kg to the weight estimate.
- Believe only scale (B), and assign 90kg to the weight estimate.
- Assign a new weight estimate less then both scales (A) and (B).
- Assign a new weight estimate greater then both scales (A) and (B).
- Assign a new weight estimate between scales (A) and (B).

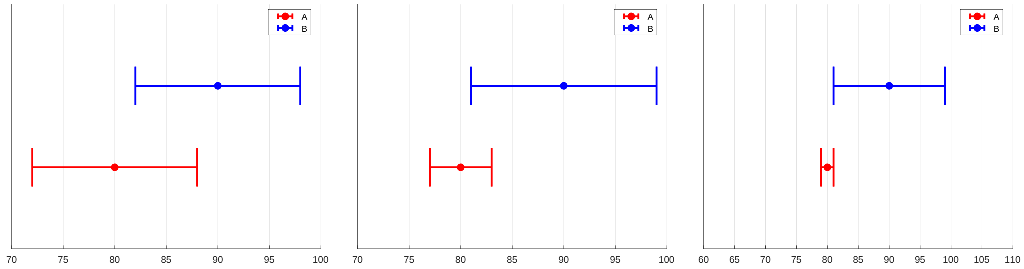
The first two choices are plausible, but there is no reason to favor one scale over the other. In other words, there is no reason to choose to believe (A) instead of (B). The third and fourth choices are irrational. The scales are admittedly not very accurate, but there is no reason at all to choose a number outside of the range of what they both measured. The final choice is the only reasonable one. If both scales are inaccurate, and as likely to give a result above the actual weight as below it, more often than not the answer is somewhere between (A) and (B).

In the case of multiple many readings, some of the times both scales will read too low, sometimes both will read too high, and the rest of the time they will straddle the actual weight. By choosing a number between (A) and (B) the effect of the worst measurement is mitigated. For example, suppose the actual weight is 95kg. 80kg is a big error. Instead, by choosing a weight between 80kg and 90kg the estimate will be better than 80kg. The same argument holds if both scales returned a value greater than the actual weight.

For now, it is clear that the best estimate is the average of (A) and (B):

$$\frac{A+B}{2} = \frac{80+90}{2} = 75\text{kg}$$

The measurements of (A) and (B) with an assumed error of 8kg are plotted on Fig. 2.1a. The measurements falls between 80kg and 90kg so the only weight that makes sense must lie within 80 and 90kg.



(a) Same error (b) Different error (c) Different error, limit

Figure 2.1: Measurements of scales (A) and (B) and relatives errors

So 75kg looks like a reasonable estimate, but there is more information to take advantage of. The only weights that are possible lie in the intersection between the error bars of (A) and (B). For example, a weight of 81kg is impossible because scale (B) could not give a reading of 90kg with a maximum error of 8kg. Likewise a weight of 89kg is impossible because scale (A) could not give a reading of 80kg with a maximum error of 8kg. In this example the only possible weights lie in the range of 82kg and 88kg.

First Example In a search of a better weight estimate, consider now the case that (A) is three times more accurate than (B).

By inspecting again the five possibilities of choosing an estimate, it still makes no sense to choose a number outside the range of (A) and (B). It perhaps seems more compelling to choose (A) as estimate — after all, it is known to be more accurate, why not use it instead of (B)? Can (B) possibly improve knowledge over (A) alone? The answer, perhaps counter intuitively, is yes, it can.

Second Example First, consider a second example in Fig. 2.1b with the same measurements of $A = 80$kg and $B = 90$kg, instead the error of (A) is 3kg and the error of (B) is 3 times as much, 9kg. Again, the overlap of the error bars of (A) and (B) is the only possible true weight. Notice that this overlap is smaller than the error in (A) alone. More importantly, in this case it is observable that the overlap does not include 80kg or 95kg. If only the measurement from (A) was used because it is more accurate than (B), then the estimate would be 80kg. Instead, if the estimate is the average between (A) and (B), then the result would be 85kg. Neither of those weights are possible given the knowledge of the accuracy of the scales. By including the measurement of B the estimate is somewhere between 81kg and 83kg, the limits of the intersections of the two error bars.

Third Example As a limit example, consider a third case in Fig. 2.1c, where it is assumed scale (A) is accurate to 1kg, and scale (B) is accurate to 9kg. Readings from both scales are taken, again $A = 80$kg, and $B = 90$kg. What should be the estimate of the weight?

It is shown that the only possible weight is 81kg. This example highlights an important result: **two relatively inaccurate sensors are able to deduce an extremely accurate result**. In other words, the take-home message is to never throw information away, no matter how poor it is. The filters in literature were studied and developed in this way, allowing the inclusion of all possible sources of information to form the best possible estimate.

Returning to the examples, what if there is only one scale and multiple readings are taken? In this case it can be shown, in a simulation where sampling of random measurements is possible, that taking the average of a large number of weights will be very close to the actual weight.

The simulation makes one assumption that probably is not true — that the scale is as likely to read 80kg as 85kg for a true weight of 85kg, which in reality almost never true. Real sensors are more likely to get readings nearer the true value, and are less and less likely to get readings the further away from the true value it gets. In other words they can be modeled as being Gaussian distributed.

Consider now the case of measuring the body weight of a person once a day. For instance, the readings are 85kg, 81kg, and then 84kg. Did the person gain weight, lose weight, or is this all just noisy measurements?

There are many available explanations. The first measurement was 85kg, and the last was 84kg, implying a 1kg loss. But if the scale is only accurate to 10kg, that is explainable by noise. The person could have actually gained weight; maybe the weight on day one was 82kg, and on day three it was 86kg. It is possible to get those weight readings with that weight gain. The scale may suggest losing weight, but in reality one may actually gain weight. In Fig. 2.2, the measurements are plotted

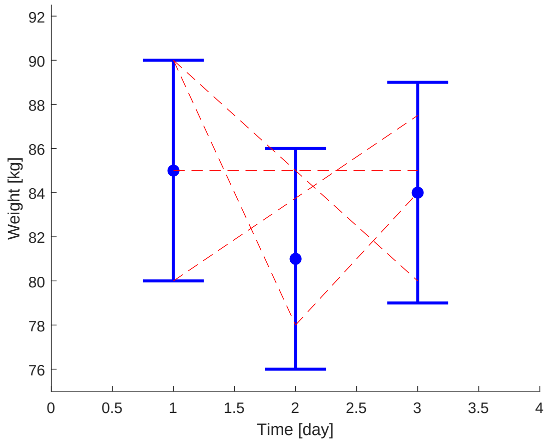
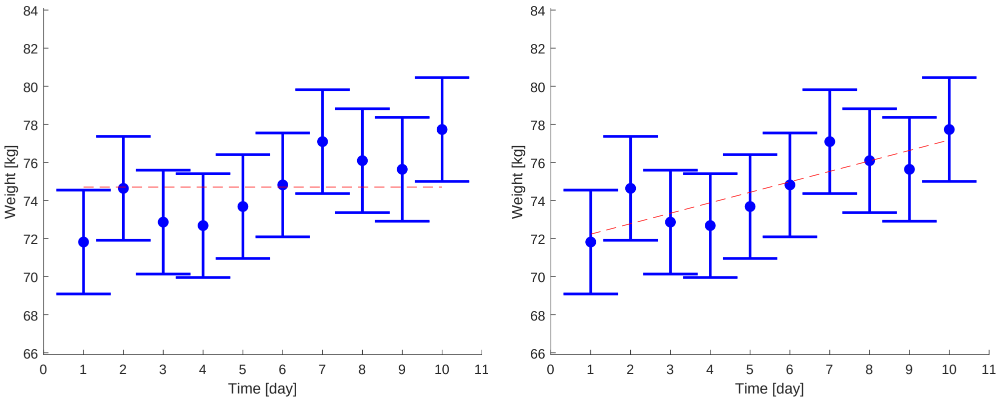


Figure 2.2: An example of readings of body weight in blue and related plausible hypotheses in red. Instead, the weight estimate is depicted in black. The variable notation is a common one when dealing with filters.

along with the error bars, and then some possible weight gain/losses that could be explained by those measurements in dotted red lines. As shown there is an extreme range of weight changes that could be explained by these three measurements. In fact, there are an infinite number of choices. To ease the problem, recall that the case in hand is about measuring a human's weight. There is no reasonable way for a human to weight 90kg on day 1 and 80kg on day 3. It is also impossible to lose 15kg in one day only to gain it back the next (assuming no amputations or other trauma has happened to the person). In other words, the behavior of the physical system under analysis should influence how the measurements are interpreted. In the case of a rock, all the variance would have been attributed to noise. Instead, in the case of weighing a cistern fed by rain and used for household chores, there might be a believe such weight changes are real. Suppose now to take a different scale



(a) Constant trend hypothesis

(b) Upward trend hypothesis

Figure 2.3: Example of daily measurements of the weight of a person.

and get a measurement each day as in Fig. 2.3. The idea is to compare two different hypotheses. The first assumption is that the weight did not change. Figure 2.3a tests that assumption by agreeing on averaging the measurements: the result does not look very convincing. In fact, there is no horizontal line that could be drawn that is inside all of the error bars. The second assumption is of gaining weight. Figure 2.3b tests that assumption by doing a least square fit: the result looks better. Notice now the hypothesis lies very close to each measurement, whereas in the previous plot the hypothesis was often quite far from the measurement. It seems far more likely to be true that the person gained weight than he did not gain any weight. In particular, suppose to gain weight about 1kg a day as a result of a high calorie diet, for instance. The idea is to make use of such information, if it was available.

Following the rationale, the first measurement is around 71.8kg as seen in Fig. 2.4. Because there is no other information, the information is accepted as an estimate. Now, if the weight of today is around 71.8kg, what will it be tomorrow? By making use of the gaining weight diet assumption, 1kg a day, the prediction turns out to be around 72.8kg. The predictions could go on for the following days but there is also the information given by the scales, e.g. the measurements of the next days. Indeed the next day, the scale displays 74.6kg. Notice now that the prediction does not match the measurement as expected. If the prediction was always exactly the same as the measurement, it would not be capable of adding any information to the filter, and there would be no reason to ever measure. If estimates are only formed from the measurement then the prediction will not affect the result. If estimates are only formed from the prediction then the measurement will be ignored. In order to exploit both information there is the need to blend the prediction and measurement together. Using the same reasoning as before in the scale example, the only thing that makes sense is to choose a number between the prediction and the measurement. For example, an estimate of 75kg makes no sense, nor does 71kg. The estimates should lie between 72.8kg, i.e. the prediction, and 74.6kg, i.e. the measurement. The

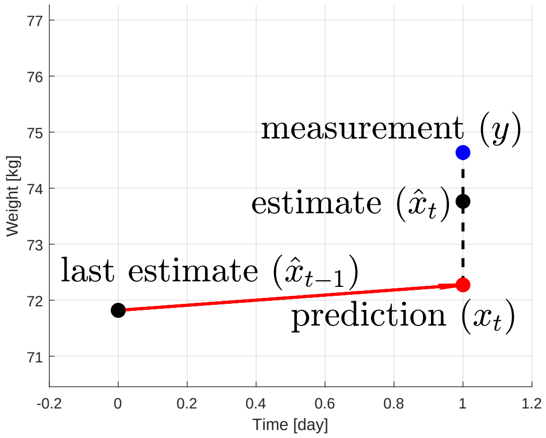


Figure 2.4: An example of readings of body weight in blue and related plausible hypotheses in red.

important concept is agreeing that **when presented two values with errors, the estimate should be formed part way between the two values**. Moreover, it does not matter how those values were generated. In the start of the examples there were two measurements, but now there is one measurement and one prediction. In order words, by replacing an inaccurate scale with an inaccurate weight prediction based on human physiology makes no difference: the reasoning of having two pieces of data with a certain amount of noise and how to combine them is the same, and hence the math is the same in both cases.

Should the estimate be half way between the measurement and prediction? In general, it seems like there might be known that the prediction is more or less accurate compared to the measurements. Probably the accuracy of the prediction differs from the accuracy of the scale. Recalling the scales example, when scale (A) was much more accurate than scale (B), the solution was to scale the answer to be closer to (A) than (B) as seen in Fig. 2.4. Without making any assumption, let the scale factor be random, for instance:

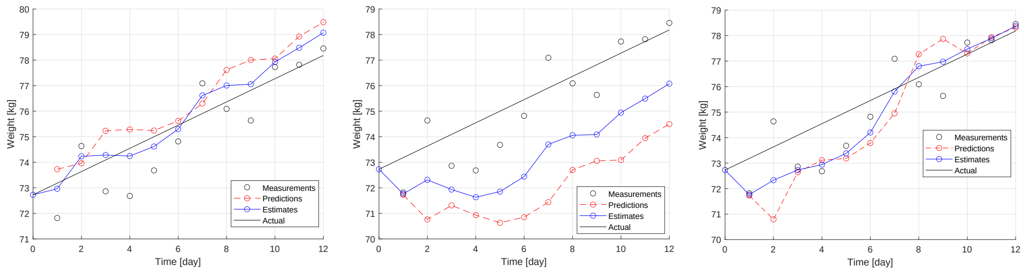
$$\text{scale factor} = \frac{4}{10}$$

Then, the estimate will be four tenths the measurement and the rest will be from the prediction. In other words, a belief is expressed here, a belief that the prediction is somewhat more likely to be correct than the measurement. The computation is as follows:

$$\text{estimate} = \text{prediction} + \frac{4}{10}(\text{measurement} - \text{prediction})$$

The difference between the measurement and prediction is called the residual, which is depicted by the black vertical line in Fig. 2.4. Smaller residuals imply better performance.

By coding it, as shown in Algorithm 2.1, the results are depicted in Fig. 2.5a when tested against the series of weights from above. Note that weight gain has unit of kg/day, so in this case it is added a time step, which is set to 1 day. The simulated weight data correspond to a true starting weight of 72.7kg, and a weight gain of 1kg per day. In other words on the first day (day zero) the true weight is 72.7kg, on the



(a) A first example when the weight gain is assumed correctly as the conclusion.

(b) A second example when the weight gain is assumed wrongly.

(c) A third example when also the rate gain is self-adjusted.

Figure 2.5: Some examples of filter implementations about weighting a person. Respectively in blue, red, and black, the estimates of the filter starting at day 0 with the initial guess of 72.7kg, the predictions made from the previous day's weight, and in black the actual weight gain of the person being weighted.

second day (day one, the first day of weighing) the true weight is 73.7kg, and so on. There is also the need to guess for the initial weight. For now it is assumed to be 72.7kg. Notice in Fig. 2.5a that the estimates on each day is part way between the related prediction and measurement. The estimates are not a straight line, but they are straighter than the measurements and somewhat close to the created trend line. Also, it seems to get better over time. In this case, the results of the filter may seem silly: the data look good if it is assumed the conclusion, i.e. that the weight gain is around 1 kg/day. Consider now the case when the initial guess is bad, that is when the gain weight is assumed to be -1 kg/day, i.e. a weight loss. The resulting plot in Fig. 2.5b shows that the estimates quickly divert from the measurements. The problem is that a filter that requires to correctly guess a rate of change is not very useful. Even if the initial guess is correct, the filter fails as soon as that rate of change changes. Thus, instead of leaving the weight gain at the initial guess of 1 kg

Algorithm 2.1: A first simple example about implementing a filter

```
Data:
measurement = [71.8, 74.6, 72.9, 72.7, 73.7, 74.8, 77.1, 76.1, 75.6, 77.7, 77.8,
  78.5]
Initialization:
timeStep = 1
scaleFactor = 4/10
initialEstimate = 72.7
forall measurement do
    Predict step:
    prediction = estimate + gainRate * timeStep
    Correct step:
    estimate = prediction + scaleFactor * (measurement - prediction)
```

it can be computed from the existing measurements and estimates as follows:

$$\text{newGain} = \text{oldGain} + \text{rateScale}\left(\frac{\text{measurement} - \text{prediction}}{\text{timeStep}}\right)$$

Again, the expression is the the same as before, the one used to combine two values together. The scale factor in this case is chosen randomly again, for the sake of the example let it be $\frac{1}{3}$. Note that in the above example, working with a rate (kg/day), the expression needs to be adjusted to incorporate the time information through the chosen time step. The result in Fig. 2.5c looks good. Even if the initial guess of the weight gain is poor, i.e. a weight loss of 1kg/day, it takes the filter several days to accurately predict the weight, but once it does that it starts to accurately track the weight. There was no methodology for choosing the scaling factors of the weight measurement and the gain rate (actually, they are poor choices for this problem), but otherwise the expression are derived from very reasonable assumptions.

Algorithm 2.2: A second example about implementing a filter with self-adjusting gain rate

```
Data:
measurement=[71.8, 74.6, 72.9, 72.7, 73.7, 74.8, 77.1, 76.1, 75.6, 77.7, 77.8, 78.5]
Initialization:
timeStep = 1
weightScale = 4/10
rateScale = 1/3
initialWeight = 72.7
forall measurement do
    Predict step:
    weight = weight + gainRate * timeStep
    gainRate = gainRate
    Correct step:
    residual = measurement - weight
    gainRate = gainRate + gainScale * residual/timeStep
    weight = weight + weightScale * residual
```

2.2 The complementary filter

The algorithm derived in the Section 2.1 is known as the *Complementary Filter*, or *g-h Filter*, or *α-β Filter*. In particular g and h refer to the two scaling factors used in the example, respectively, the scaling used for the measurement (weight in the example), and the scaling used for the change in measurement over time (kg/day in the example).

This filter is the basis for a huge number of filters, including the Kalman filter. In other words, the Kalman filter can be seen as a form of the *g-h* filter and viceversa in special cases. So are other filters, like the Least Squares filter or the Benedict-Bordner filter for instance. Each filter has a different way of assigning values to g and h, but otherwise the algorithms are identical. For example, the Benedict-Bordner filter assigns a constant to both g and h, constrained to a certain range of values. Instead, the Kalman filter will vary g and h dynamically at each time step.

To summarize, even if the math may look profoundly different, the algorithm will be exactly the same. Some key insights from the previous examples are here summarized:

- Multiple data points, even if inaccurate, are more accurate than one data point — do not throw away information.
- Always choose a number part way between two data points to create a more accurate estimate.
- Predict the next measurement and rate of change based on the current estimate and how much it is thought it will change.
- The new estimate is then chosen as part way between the prediction and next measurement scaled by how accurate each is.

In order to have a better understand of the algorithm different problem domains are hereafter explored:

1. Consider the problem of trying to track a train on a track. The track constrains the position of the train to a very specific region. Furthermore, trains are large and slow. It takes many minutes for them to slow down or speed up significantly. Following this rationale, it makes sense to suppose that by knowing the state of the system at a given time instant, e.g. for instance the position and velocity of the train, it is possible to be confident in predicting the future state of the system, for instance at the next second. This assumption is important when related to the scaling versus the measurement. Again, if the measurement suggests a rapidly change of position and velocity, it makes sense to suppose that measurement is inaccurate. In other words, when designing a filter for this problem, the idea is to have very high weighting to the prediction versus the measurement.
2. Now consider the problem of tracking a thrown ball. It is known that a ballistic object moves in a parabola in a vacuum when in a gravitational field. But a ball thrown on Earth is influenced by air drag, so it does not travel in a perfect parabola. Baseball pitchers, for instance, take advantage of this fact when they throw curve balls. Consider the case when the tracking of the ball inside a stadium is performed using computer vision. The accuracy of the computer vision tracking might be modest, but predicting the ball's future positions by assuming that it is moving on a parabola is not accurate either. In this case, probably the design a filter should give roughly equal weight to the measurement and the prediction.
3. Now consider trying to track a helium party balloon in a hurricane. There is no legitimate model that would allow to predict the balloon's behavior except over very brief time scales. In this case the design of the filter would emphasize the measurements over the predictions.

To summarize, it is possible to tune the parameters of the filter, as done in a quality manner in the previous examples, to have better filter performance. Moreover, a particular choice might perform well in one situation, but very poorly in another. Even when understanding the effect of g and h it can be difficult to choose proper values. In fact, it is extremely unlikely any chosen values will be optimal for any given problem. **Filters are designed, not selected ad hoc**. The Kalman filter, as explained later in Chapter 3 will do this in an optimal way by developing a very powerful form of probabilistic reasoning about filtering — namely *Bayesian Filtering*. This kind of class of filters, also the complementary filter as well as other filters, can be summarized as in Algorithm 2.3. There is also the case when, for instance, the model of the system does not take into account something, e.g. the acceleration in the previous examples. This is called the *lag error* or systemic error of the system. It is a fundamental property of g-h filters. The take-home point is that **the filter is only as good as the mathematical model used to express the system**. Anyway, the fundamental idea is to blend somewhat inaccurate measurements with somewhat inaccurate models of how the systems behaves to get a filtered estimate that is better than either information source by itself.

Algorithm 2.3: How to implement a complementary filter

```
Initialization:
Initialize the state of the filter
Initialize the belief in the state

Predict step:
Use system behavior to predict state at the next time step
Adjust belief to account for the uncertainty in prediction

Correct step:
Get a measurement and associated belief about its accuracy
Compute residual between estimated state and measurement
New estimate is somewhere on the residual line
```

CHAPTER 3

The discrete-time multivariate Kalman Filter

This chapter briefly reviews the literature about Kalman theory applied for the estimation of a discrete-time LTI multivariate dynamical system. In particular, the derivation of the filtering equations is proposed by means of a powerful Bayesian perspective. It is respectful to mention that the herein content and the logical order of presentation is inspired and adapted from [1]. In conclusion, this chapter is organized as follow:

1. Section 3.1 introduces the mathematical framework used for the derivation of the discrete time multivariate Kalman filter. In this setting, common convention are used and description of the deterministic and noise components of the system are provided;
2. Section 3.2 is dedicated to the Bayesian inference applied to the state estimation problem. A Bayesian filtering approach is derived leading to the well-known prediction and correction equations.

3.1 Mathematical framework

In the following section the mathematical framework under analysis will be introduced. In particular it consists of:

- A signal model description
- A noise model description

3.1.1 Description of the system model

In the following chapter attention will be given primarily to discrete-time systems, e.g. systems where the underlying system equations are difference equations. In particular, the class of discrete-time systems under analysis has the following characteristics: **a)** Linear, **b)** Time variant, and **c)** Finite dimensional , which are described by state-space equations:

$$\mathcal{S}: \begin{cases} \boldsymbol{x}_{k+1} = \boldsymbol{A}_k \boldsymbol{x}_k + \boldsymbol{B}_k \boldsymbol{u}_k + \boldsymbol{G}_k \boldsymbol{w}_k & \text{(3.1a)} \\ \boldsymbol{y}_k = \boldsymbol{C}_k \boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{u}_k + \boldsymbol{v}_k & \text{(3.1b)} \end{cases}$$

Time indexing The subscript k is a time argument. The initial time at which the system starts operating cab be assumed to be finite. Then by shifting of the time origin, it is assumed that Eq. (3.1) hold for $\forall k$ s.t. $k > 0$. Additionally, without loss of generality, successive time instants are denoted by the integer k.

State, input and output signals The set $\{(\boldsymbol{x}_k, k)$ s.t. $k > 0\}$ is denoted by the symbol $\{\boldsymbol{x}_k\}$. Instead, $\boldsymbol{x}_k$ in Eq. (3.1) is the value of the *system state* at time k. Under ideal circumstances, $\boldsymbol{y}_k = \boldsymbol{C}_k \boldsymbol{x}_k$ would be the corresponding system output, but in this case an additive measurement noise process $\boldsymbol{v}_k$ is added. The input process to the system is $\{\boldsymbol{w}_k\}$, and like $\{\boldsymbol{v}_k\}$, it is a noise process. Further details of $\{\boldsymbol{v}_k\}$ and $\{\boldsymbol{w}_k\}$ will be given shortly, as will some motivation for introducing the whole model in Eq. (3.1). Note that, according to the multivariate case under analysis, an appropriate vector processes notation will be used on processes $\{\boldsymbol{x}_k\}$, $\{\boldsymbol{y}_k\}$, $\{\boldsymbol{w}_k\}$, $\{\boldsymbol{v}_k\}$.

3.1.2 Description of the noise model

In order to derive a filtering theory it is mandatory to cast a probabilistic structure on the noise processes affecting the system, i.e. the process noise $\{\boldsymbol{w}_k\}$ and the measurement noise $\{\boldsymbol{v}_k\}$. Following this rationale, the assumption related to the stochastic part of the system are:

Assumption 3.1. $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{v}_k\}$ are individually white processes, i.e. $\forall k, l$ s.t. $k \neq l$, $\boldsymbol{v}_k$ and $\boldsymbol{v}_l$ are independent random variables, and $\boldsymbol{w}_k$ and $\boldsymbol{w}_l$ are independent random variables.

Assumption 3.2. $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{v}_k\}$ are individually gaussian processes of zero mean and known (bounded) auto-covariance, i.e. $\boldsymbol{v}_k \sim \mathcal{G}(\boldsymbol{0}, \boldsymbol{R}_k)$ and $\boldsymbol{w}_k \sim \mathcal{G}(\boldsymbol{0}, \boldsymbol{Q}_k)$.

Assumption 3.3. $\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{v}_k\}$ are independent processes.

Assumptions 3.1 and 3.2 means that, for instance, the joint pdf of $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_k$ for arbitrary k is gaussian. The computation of the joint pdf is simply the product of the individual densities due to the whiteness of $\boldsymbol{v}_k$, guaranteed by Asm. 3.1, and can be carried out by knowing the mean and covariance values since they determine the joint pdf completely.

In order to fulfill Asms. 3.1 to 3.3 the process mean and the process covariance, that are respectively the set of values of $\mathbb{E}[\boldsymbol{v}_k]$ and $\mathbb{E}[\boldsymbol{v}_k \boldsymbol{v}_l^\top]$, $\forall k, l$, need to be specified. In other words, the process mean is:

$$\mathbb{E}[\boldsymbol{w}_k] = \mathbf{0} \in \mathbb{R}^{n_w}, \quad \forall k \tag{3.2a}$$
$$\mathbb{E}[\boldsymbol{v}_k] = \mathbf{0} \in \mathbb{R}^{n_y}, \quad \forall k \tag{3.2b}$$

while the process covariance is:

$$\mathbb{E}[\boldsymbol{w}_k \boldsymbol{w}_l^\mathsf{T}] = \boldsymbol{Q}_k \delta_{k,l}, \quad \forall k, l \tag{3.3a}$$
$$\mathbb{E}[\boldsymbol{v}_k \boldsymbol{v}_l^\mathsf{T}] = \boldsymbol{R}_k \delta_{k,l}, \quad \forall k, l \tag{3.3b}$$

by noting that:

$$\mathbb{E}[\boldsymbol{w}_k \boldsymbol{w}_l^\mathsf{T}] \overset{3.1}{=} \mathbb{E}[\boldsymbol{w}_k]\mathbb{E}[\boldsymbol{w}_l^\mathsf{T}] \overset{3.2}{=} \mathbf{0}_{n_w \times n_w}, \quad \forall k, l \text{ s.t. } k \neq l \tag{3.4}$$
$$\mathbb{E}[\boldsymbol{v}_k \boldsymbol{v}_l^\mathsf{T}] \overset{3.1}{=} \mathbb{E}[\boldsymbol{v}_k]\mathbb{E}[\boldsymbol{v}_l^\mathsf{T}] \overset{3.2}{=} \mathbf{0}_{n_y \times n_y}, \quad \forall k, l \text{ s.t. } k \neq l \tag{3.5}$$

where $\delta_{k,l}$ is the *Kronecker's delta*[1], $\mathbb{E}[\boldsymbol{w}_k \boldsymbol{w}_k^\mathsf{T}] = \boldsymbol{Q}_k$ and $\mathbb{E}[\boldsymbol{v}_k \boldsymbol{v}_k^\mathsf{T}] = \boldsymbol{R}_k$ are non-negative definite matrices. Instead, the cross-covariance of the processes is:

$$\mathbb{E}[\boldsymbol{v}_k \boldsymbol{w}_l^\mathsf{T}] \overset{3.3}{=} \mathbf{0}_{n_y \times n_w}, \quad \forall k, l \tag{3.6}$$

3.1.3 Description of the initial condition

An initial condition for the difference equation in Eq. (3.1) needs to be specified. From the practical point of view, if it is impossible to measure the state $\boldsymbol{x}_k$, exactly for arbitrary k, it is unlikely that a measure of the initial condition $\boldsymbol{x}_0$ will be available. This leads to the adoption of a random initial condition for the system. In particular, we it is assumed that the uncertainty about the initial condition $\boldsymbol{x}_0$ is described as well by a random variable Gaussian distributed:

Assumption 3.4. The initial condition $\boldsymbol{x}_0$ is a Gaussian random variable with known mean $\bar{\boldsymbol{x}}_0$ and known covariance $\boldsymbol{P}_0$:

$$\boldsymbol{x}_0 \sim \mathcal{G}(\bar{\boldsymbol{x}}_0, \boldsymbol{P}_0) \tag{3.7}$$

where

$$\mathbb{E}[\boldsymbol{x}_0] = \bar{\boldsymbol{x}}_0 \tag{3.8a}$$
$$\mathbb{E}[(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0)(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0)^\mathsf{T}] = \boldsymbol{P}_0 \tag{3.8b}$$

Moreover, it is assumed that the initial condition is independent from the noise processes:

Assumption 3.5. The initial condition $\boldsymbol{x}_0$ is independent from noise processes

[1] The *Kronecker's delta* $\delta_{k,l}$ is 1 for $k = l$ and 0 for $k \neq l$.

$\{\boldsymbol{w}_k\}$ and $\{\boldsymbol{v}_k\}$, that is the *causality assumption*.

$$\mathbb{E}[(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0)\boldsymbol{w}_k^\top] = \mathbf{0}_{n_x \times n_w}, \quad \forall k \tag{3.9}$$

$$\mathbb{E}[(\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0)\boldsymbol{v}_k^\top] = \mathbf{0}_{n_x \times n_v}, \quad \forall k \tag{3.10}$$

Assumption on the noise processes and the initial state are:

$$\begin{bmatrix} \boldsymbol{w}_k \\ \boldsymbol{v}_k \\ \boldsymbol{x}_0 \end{bmatrix} \overset{\text{i.i.d}}{\sim} \mathcal{G}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \bar{\boldsymbol{x}}_0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{Q} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{P}_0 \end{bmatrix}\right) \tag{3.11}$$

Without loss of generality, to simplify the analysis the static sub-system is not considered, i.e. $\boldsymbol{D} = \mathbf{0}$.

3.1.4 Gaussian and Markov properties of the system

Important properties of the random process $\{\boldsymbol{x}_k\}$ are derived considering **Asms. 3.1** to 3.5.

The transition equation First, consider the *transition equation* regarding the state equation in Eq. (3.1a):

$$\boldsymbol{x}_k = \boldsymbol{\Phi}_{k,0}\boldsymbol{x}_0 + \sum_{i=0}^{k-1}\left[\boldsymbol{\Phi}_{k,i+1}(\boldsymbol{B}_i\boldsymbol{u}_i + \boldsymbol{G}_i\boldsymbol{w}_i)\right] \tag{3.12a}$$

$$= \boldsymbol{\Phi}_{k,0}\boldsymbol{x}_0 + \sum_{i=0}^{k-1}[\boldsymbol{\Phi}_{k,i+1}\boldsymbol{B}_i\boldsymbol{u}_i + \boldsymbol{\Phi}_{k,i+1}\boldsymbol{G}_i\boldsymbol{w}_i], \quad \forall k \tag{3.12b}$$

where $\boldsymbol{\Phi}_{k,l}$ is the transition matrix defined as:

$$\boldsymbol{\Phi}_{k,l} \equiv \boldsymbol{A}_{k-1}\boldsymbol{A}_{k-2}\dots\boldsymbol{A}_l, \quad \forall k,l \text{ s.t. } k > l \tag{3.13a}$$

$$\boldsymbol{\Phi}_{k,k} \equiv \boldsymbol{I}, \quad \forall k \tag{3.13b}$$

$$\boldsymbol{\Phi}_{k,l}\boldsymbol{\Phi}_{l,m} = \boldsymbol{\Phi}_{k,m}, \quad \forall k,l,m \text{ s.t. } k \geq l \geq m \tag{3.13c}$$

Considering Eq. (3.12), the state variable $\boldsymbol{x}_k$ has the following properties:

Remark 3.1. $\boldsymbol{x}_k$ is expressed as a linear combination of the jointly gaussian random vectors $\boldsymbol{x}_0$ and $\boldsymbol{w}_0, \dots, \boldsymbol{w}_{k-1}$. Note that the fact that the variables are individually **a)** gaussian, and **b)** independent implies their jointly density is also gaussian. Now since linear transformations of gaussian random variables preserve their gaussian character, it follows that $\boldsymbol{x}_k, \forall k$ is a gaussian random variable.

Remark 3.2. The second property is that $\{\boldsymbol{x}_k\}, \forall k$ is a gaussian random process. This property is simply an extension of the first, see Rem. 3.1.

Remark 3.3. The last property is that $\{\boldsymbol{x}_k\}, \forall k$ is a Markov random process. In other words, the probability density of $\boldsymbol{x}_k$ conditioned on $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{k-1}$ is simply the probability density of $\boldsymbol{x}_k$ conditioned on $\boldsymbol{x}_{k-1}$:

$$p(\boldsymbol{x}_k \mid \boldsymbol{x}_1, \dots, \boldsymbol{x}_{k-1}) = p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}) \tag{3.14}$$

The Markov property is a consequence of these factors: **a)** the whiteness of $\boldsymbol{w}_k$, and **b)** the causality of the system in Eq. (3.1).

Instead, regarding the output variable $\boldsymbol{y}_k$ it can be said that:

Remark 3.4. $\{\boldsymbol{y}_k\}$ is a gaussian process as well, for the same reasons as $\{\boldsymbol{x}_k\}$. Also, $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$ are jointly gaussian. Instead, $\{\boldsymbol{y}_k\}$ is no longer a Markov process, due to the fact that $\{\boldsymbol{y}_k\}$ is not white. In other words, the correlation between two measurements $\boldsymbol{y}_{k_1}$ and $\boldsymbol{y}_{k_2}$ with $|k_2 - k_1| > 1$ is not equal to zero, e.g. the two measurements convey more information jointly about $\boldsymbol{y}_k$.

3.1.5 Propagation of the statistics

As seen in Section 3.1.4, $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$ are jointly gaussian processes. Therefore, their probabilistic properties are entirely determined by their means and covariances. In particular propagation of these statistics can be derived directly from the difference equations in Eq. (3.1), considering 1-step ahead propagation, or more in general, by using the *transition equation* in Eq. (3.12). Without loss of generality, to simplify the following computations, the deterministic sub-system is not considered, i.e. $\boldsymbol{B}_k = \boldsymbol{0}$. If not, the following calculations can easily be extended with the presence of an exogenous deterministic input. In the end, the propagation of the statistics of the stochastic part of the system is as follows.

From the difference equations Regarding $\{\boldsymbol{x}_k\}$, the computation of the mean is as follows:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_{k+1}] &\overset{(3.1a)}{=} \mathbb{E}[\boldsymbol{A}_k\boldsymbol{x}_k + \boldsymbol{G}_k\boldsymbol{w}_k] && (3.15a)\\
&\overset{(3.2a)}{=} \boldsymbol{A}_k\mathbb{E}[\boldsymbol{x}_k] + \boldsymbol{G}_k\cancel{\mathbb{E}[\boldsymbol{w}_k]} && (3.15b)\\
&= \boldsymbol{A}_k\mathbb{E}[\boldsymbol{x}_k] && (3.15c)
\end{aligned}$$

Regarding $\{\boldsymbol{y}_k\}$, the computation of the mean is as follows:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}_k] &\overset{(3.1b)}{=} \mathbb{E}[\boldsymbol{C}_k\boldsymbol{x}_k + \boldsymbol{v}_k] && (3.16a)\\
&\overset{(3.2b)}{=} \boldsymbol{C}_k\mathbb{E}[\boldsymbol{x}_k] + \cancel{\mathbb{E}[\boldsymbol{v}_k]} && (3.16b)\\
&= \boldsymbol{C}_k\mathbb{E}[\boldsymbol{x}_k] && (3.16c)
\end{aligned}$$

For ease of notation, set $\mathbb{E}[\boldsymbol{x}_k] = \bar{\boldsymbol{x}}_k$ Then, the computation of the covariance of $\boldsymbol{x}_k$ can be noted as:

$$\mathbb{E}[(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)^\top] = \boldsymbol{P}_k \tag{3.17}$$

Observe that the notation is consistent with the use of $\boldsymbol{P}_0$ in Eq. (3.8b). Regarding $\{\boldsymbol{x}_k\}$, the computation of the covariance is as follows:

$$
\begin{aligned}
\boldsymbol{P}_{k+1} &\overset{(3.17)}{=} \mathbb{E}[(\boldsymbol{x}_{k+1} - \bar{\boldsymbol{x}}_{k+1})(\boldsymbol{x}_{k+1} - \bar{\boldsymbol{x}}_{k+1})^\mathsf{T}] && (3.18a)\\
&\overset{(3.1a)}{=} \mathbb{E}[(\boldsymbol{A}_k\boldsymbol{x}_k + \boldsymbol{G}_k\boldsymbol{w}_k - \boldsymbol{A}_k\bar{\boldsymbol{x}}_k)(\boldsymbol{A}_k\boldsymbol{x}_k + \boldsymbol{G}_k\boldsymbol{w}_k - \boldsymbol{A}_k\bar{\boldsymbol{x}}_k)^\mathsf{T}] && (3.18b)\\
&\overset{(3.2a)}{=} \boldsymbol{A}_k\mathbb{E}[(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)^\mathsf{T}]\boldsymbol{A}_k^\mathsf{T} + \boldsymbol{G}_k\mathbb{E}[\boldsymbol{w}_k\boldsymbol{w}_k^\mathsf{T}]\boldsymbol{G}_k^\mathsf{T}\\
&\quad + \boldsymbol{A}_k\mathbb{E}[\boldsymbol{x}_k\boldsymbol{w}_k^\mathsf{T}]\boldsymbol{G}_k^\mathsf{T} + \boldsymbol{G}_k\mathbb{E}[\boldsymbol{w}_k\boldsymbol{x}_k^\mathsf{T}]\boldsymbol{A}_k^\mathsf{T} - \boldsymbol{A}_k\bar{\boldsymbol{x}}_k\cancel{\mathbb{E}[\boldsymbol{w}_k^\mathsf{T}]}\boldsymbol{G}_k^\mathsf{T} - \boldsymbol{G}_k\cancel{\mathbb{E}[\boldsymbol{w}_k]}\bar{\boldsymbol{x}}_k^\mathsf{T}\boldsymbol{A}_k^\mathsf{T}\\
&\overset{3.3}{=} \boldsymbol{A}_k\mathbb{E}[(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)^\mathsf{T}]\boldsymbol{A}_k^\mathsf{T} + \boldsymbol{G}_k\mathbb{E}[\boldsymbol{w}_k\boldsymbol{w}_k^\mathsf{T}]\boldsymbol{G}_k^\mathsf{T}\\
&\quad + \boldsymbol{A}_k\cancel{\mathbb{E}[\boldsymbol{x}_k\boldsymbol{w}_k^\mathsf{T}]}\boldsymbol{G}_k^\mathsf{T} + \boldsymbol{G}_k\cancel{\mathbb{E}[\boldsymbol{w}_k\boldsymbol{x}_k^\mathsf{T}]}\boldsymbol{A}_k^\mathsf{T} && (3.18c)\\
&\overset{(3.3a)\text{ and }(3.17)}{=} \boldsymbol{A}_k\boldsymbol{P}_k\boldsymbol{A}_k^\mathsf{T} + \boldsymbol{G}_k\boldsymbol{Q}_k\boldsymbol{G}_k^\mathsf{T} && (3.18d)
\end{aligned}
$$

Regarding $\{\boldsymbol{y}_k\}$, the computation of the covariance is as follows:

$$
\begin{aligned}
\mathbb{V}[\boldsymbol{y}_k, \boldsymbol{y}_k] &\equiv \mathbb{E}[(\boldsymbol{y}_k - \bar{\boldsymbol{y}}_k)(\boldsymbol{y}_k - \bar{\boldsymbol{y}}_k)^\mathsf{T}] && (3.19a)\\
&\overset{(3.1b),(3.16)}{=} \mathbb{E}[(\boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k - \boldsymbol{C}\bar{\boldsymbol{x}}_k)(\boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k - \boldsymbol{C}\bar{\boldsymbol{x}}_k)^\mathsf{T}] && (3.19b)\\
&\overset{(3.2b)}{=} \boldsymbol{C}_k\mathbb{E}[(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)^\mathsf{T}]\boldsymbol{C}_k^\mathsf{T} + \mathbb{E}[\boldsymbol{v}_k\boldsymbol{v}_k^\mathsf{T}]\\
&\quad + \boldsymbol{C}_k\mathbb{E}[\boldsymbol{x}_k\boldsymbol{v}_k^\mathsf{T}] + \mathbb{E}[\boldsymbol{v}_k\boldsymbol{x}_k^\mathsf{T}]\boldsymbol{C}_k^\mathsf{T} - \boldsymbol{C}_k\bar{\boldsymbol{x}}_k\cancel{\mathbb{E}[\boldsymbol{v}_k^\mathsf{T}]} - \cancel{\mathbb{E}[\boldsymbol{v}_k]}\bar{\boldsymbol{x}}_k^\mathsf{T}\boldsymbol{C}_k^\mathsf{T}\\
&\overset{3.3,3.5}{=} \boldsymbol{C}_k\mathbb{E}[(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)^\mathsf{T}]\boldsymbol{C}_k^\mathsf{T} + \mathbb{E}[\boldsymbol{v}_k\boldsymbol{v}_k^\mathsf{T}]\\
&\quad + \boldsymbol{C}_k\cancel{\mathbb{E}[\boldsymbol{x}_k\boldsymbol{v}_k^\mathsf{T}]} + \cancel{\mathbb{E}[\boldsymbol{v}_k\boldsymbol{x}_k^\mathsf{T}]}\boldsymbol{C}_k^\mathsf{T} && (3.19c)\\
&\overset{(3.17),(3.3a)}{=} \boldsymbol{C}_k\boldsymbol{P}_k\boldsymbol{C}_k^\mathsf{T} + \boldsymbol{R}_k && (3.19d)
\end{aligned}
$$

From the transition equation It can be shown that, using the *transition equation* in Eq. (3.12), that mean and covariance statistics of $\{\boldsymbol{x}_k\}$ and $\{\boldsymbol{y}_k\}$ are as follows:

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{x}_{k+1}] &= \mathbb{E}[\boldsymbol{\Phi}_{k,0}\boldsymbol{x}_0] && (3.20a)\\
&= \boldsymbol{\Phi}_{k,0}\bar{\boldsymbol{x}}_0 && (3.20b)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{y}_k] &= \mathbb{E}[\boldsymbol{C}_k\boldsymbol{x}_k] && (3.21a)\\
&= \boldsymbol{C}_k\boldsymbol{\Phi}_{k-1,0}\bar{\boldsymbol{x}}_0 && (3.21b)
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{P}_{k,l} &= \mathbb{E}[(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_l - \bar{\boldsymbol{x}}_l)^\mathsf{T}] && (3.22a)\\
&= \boldsymbol{\Phi}_{k,l}\boldsymbol{P}_l, \qquad k \geq l && (3.22b)\\
&= \boldsymbol{P}_k\boldsymbol{\Phi}_{l,k}^\mathsf{T}, \qquad k \leq l && (3.22c)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{V}[\boldsymbol{y}_k, \boldsymbol{y}_l] &\equiv \mathbb{E}[(\boldsymbol{y}_k - \bar{\boldsymbol{y}}_k)(\boldsymbol{y}_l - \bar{\boldsymbol{y}}_l)^{\mathsf{T}}] && \text{(3.23a)}\\
&= \boldsymbol{C}_k \boldsymbol{\Phi}_{k,l} \boldsymbol{P}_l \boldsymbol{C}_l^{\mathsf{T}} + \boldsymbol{R}_k \delta_{k,l}, \qquad k \geq l && \text{(3.23b)}\\
&= \boldsymbol{C}_k \boldsymbol{P}_k \boldsymbol{\Phi}_{l,k}^{\mathsf{T}} \boldsymbol{C}_l^{\mathsf{T}} + \boldsymbol{R}_k \delta_{k,l}, \qquad k \leq l && \text{(3.23c)}
\end{aligned}
$$

Remark 3.5 (Dropping the gaussian assumption). Until now, $\boldsymbol{x}_0$, $\boldsymbol{w}_k$, and $\boldsymbol{v}_k$ have been assumed gaussian. If this is not the case, but they remain described by their first order and second order statistics, then all the calculations still carry through in the sense that formulas for the mean and covariance of the $\{\boldsymbol{x}_k\}$, and $\{\boldsymbol{y}_k\}$ sequences are precisely as before. In other words:

- In the gaussian case, knowledge of the mean and covariance is sufficient to deduce pdfs of any order.
- In the non-gaussian case, knowledge of the mean and covariance does not provide other than incomplete information about higher order moments, let alone pdfs.

3.1.6 Estimation criteria

In this section, it is briefly illustrated how knowledge of the value taken by one random variable can give information about the value taken by a second random variable. Please refer to [1] for a full reference. In particular, it is noted how an estimate can be made of the value taken by this second random variable. The goal is to present an overview of the different estimation criteria. However, the introduction of the Kalman filter equations will be derived from a Bayesian interpretation in Section 3.2. To this end, please refer also to Appendix A where computation details are explored.

In the remainder of this section an upper-case letter is used to denote a random variable, and a lower-case letter to denote a value taken by that variable. In other words, if the random variable is $\boldsymbol{X}$ and the underlying probability space $\boldsymbol{\Omega}$ has elements $\boldsymbol{\omega}$, the symbol $\boldsymbol{x}$ will in effect be used in place of $\boldsymbol{X}(\boldsymbol{\omega})$ and the symbol $\boldsymbol{X}$ in place of $\boldsymbol{X}(\cdot)$, or the set of pairs $\{\boldsymbol{\omega}, \boldsymbol{X}(\boldsymbol{\omega})\}$ as $\boldsymbol{\omega}$ ranges over $\boldsymbol{\Omega}$.

For arbitrary densities

For arbitrary densities, the following concepts hold:

The conditional pdf

Given two vector random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, the knowledge that $\boldsymbol{Y} = \boldsymbol{y}$ modifies the a-priori information $p_{\boldsymbol{X}}(\boldsymbol{x})$ about $\boldsymbol{X}$ through the concept of conditional pdf:

$$
p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x} \mid \boldsymbol{y}) \equiv \frac{p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{y})}{p_{\boldsymbol{Y}}(\boldsymbol{y})} \tag{3.24}
$$

where it is assumed that $p_{\boldsymbol{Y}}(\boldsymbol{y}) \neq 0$.

The conditional pdf $p_{X|Y}(\boldsymbol{x} \mid \boldsymbol{y})$ with a particular value substituted for $\boldsymbol{y}$ and with $\boldsymbol{x}$ regarded as a variable sums up all the information which knowledge that $Y = \boldsymbol{y}$ conveys about $\boldsymbol{X}$.

Since it is a function rather than a single vector of real numbers, it makes sense to seek a simpler entity in order to form an estimate. For example, one estimate would be the value of $\boldsymbol{x}$ maximizing $p_{X|Y}(\boldsymbol{x} \mid \boldsymbol{y})$, that is the *maximum a posteriori estimate* (MAP).

However, it is helpful to introduce a different kind of estimate, namely the *minimum variance estimate*, more properly the *conditional minimum variance estimate*. Sometimes other names are used, such as *least-square estimate*, and *minimum mean square estimate* (MMSE).

The conditional mean estimate

The *conditional mean estimate* $\hat{\boldsymbol{x}}$ is:

$$\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{X}|Y = \boldsymbol{y}] \equiv \int_{-\infty}^{+\infty} \boldsymbol{x}\, p_{X|Y}(\boldsymbol{x} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \tag{3.25}$$

It can be shown that the conditional mean estimate in Eq. (3.25) is also the *conditional minimum variance estimate*:

$$\mathbb{E}\left[\|\boldsymbol{X} - \hat{\boldsymbol{x}}\|^2 | Y = \boldsymbol{y}\right] \leq \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{z}(\boldsymbol{y})\|^2 | Y = \boldsymbol{y}\right] \tag{3.26}$$

for all functions $\boldsymbol{z}$ of $\boldsymbol{y}$.

In particular, if the two variables are gaussian and independent, they are jointly gaussian. Moreover, another general implication is that since they are jointly gaussian, also their conditional pdf is gaussian.

Example 3.1: Conditional pdf of two jointly gaussian variables

Consider the relationship:

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{N} \tag{3.27}$$

with $\boldsymbol{X} \sim \mathcal{G}(\boldsymbol{0}, \Sigma_x)$ and $\boldsymbol{Y} \sim \mathcal{G}(\boldsymbol{0}, \Sigma_y)$. Suppose also that $\boldsymbol{Y} = \boldsymbol{y}$, e.g. a measurement is available. It can be shown that the conditional pdf can be derived as:

$$p_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) \equiv \frac{p_{X,Y}(\boldsymbol{x}, \boldsymbol{y})}{p_Y(\boldsymbol{y})} \tag{3.28a}$$

$$= \frac{p_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) p_X(\boldsymbol{x})}{p_Y(\boldsymbol{y})} \tag{3.28b}$$

$$\tag{3.28c}$$

For gaussian densities

Let X and Y be vector independent gaussian variables. The joint pdf is gaussian, i.e. $Z = [X^\top Y^\top]^\top \sim \mathcal{G}(\boldsymbol{\mu}_Z, \Sigma_Z)$ with statistics:

$$\boldsymbol{\mu}_Z = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}, \qquad \Sigma_Z = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \tag{3.29}$$

3.2 Bayesian derivation of the Kalman filter

The following section introduces the concepts of Bayesian inference in dynamical models. Specialized dynamical models term as Hidden Markov Models are introduced, where their constitutive properties help deriving efficient solutions to the filtering problem by means of a recursive Bayesian estimation. The concepts adhere introduced are taken from various references, for instance see [4, 59, 69, 25, 211, 31].

3.2.1 Hidden Markov Model

The general *State Space Model* in Eq. (3.1) has a compatible description using an *Hidden Markov Model*:

$$\boldsymbol{x}_0 \sim \mu(\boldsymbol{x}_0) \tag{3.30a}$$
$$\boldsymbol{x}_{k+1}|\boldsymbol{x}_k \sim f(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k) \tag{3.30b}$$
$$\boldsymbol{y}_k|\boldsymbol{x}_k \sim g(\boldsymbol{y}_k|\boldsymbol{x}_k) \tag{3.30c}$$

where three main assumptions are made, that are:

1. (time) Causality;
2. Markov property on process $\{\boldsymbol{x}_k\}$;
3. Conditional independence of the measurements;

In particular, the Markov property states that:

Markov property on process $\{\boldsymbol{x}_k\}$

The process $\{\boldsymbol{x}_k\}$, e.g. the set of states, form a Markov sequence. The Markov property means that $\boldsymbol{x}_k$ (and actually the whole future $\boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+2} \dots$) given $\boldsymbol{x}_{k-1}$ is independent of anything that has happened before the time step $k-1$:

$$p(\boldsymbol{x}_k|\boldsymbol{x}_{0:k-1}, \boldsymbol{y}_{1:k-1}) = f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) \tag{3.31}$$

Also the past is independent of the future given the present:

$$p(\boldsymbol{x}_{k-1}|\boldsymbol{x}_{k:T}, \boldsymbol{y}_{k:T}) = f(\boldsymbol{x}_{k-1}|\boldsymbol{x}_k), \qquad T > k \tag{3.32}$$

Instead, the conditional independence of the measurements property can be formulated as follows:

The conditional independence of the measurements property:

The herein property states that the current measurement $\boldsymbol{y}_k$ given the current state $\boldsymbol{x}_k$ is conditionally independent of the measurement and state histories. In other words:

$$p(\boldsymbol{y}_k|\boldsymbol{x}_{0:k}, \boldsymbol{y}_{1:k-1}) = g(\boldsymbol{y}_k|\boldsymbol{x}_k) \tag{3.33}$$

An equivalent graphical description of the Hidden Markov Model in Eq. (3.30) is given by the Bayesian network as shown in Fig. 3.1 where the properties of causality, Markov, and conditional independence of the measurements can be easily checked.

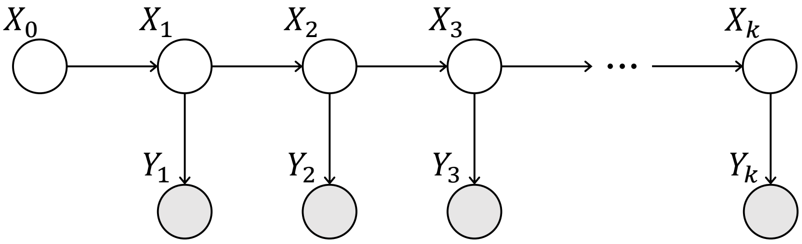


Figure 3.1: The Bayesian network used as a graphical model for the SSM in Eq. (3.30). Each random variable is encoded using a node, where the nodes that are filled (gray) corresponds to variables that are observed and nodes that are not filled (white) are hidden/latent/unobserved variables. The arrows encode the dependence among the variables.

3.2.2 Bayesian inference aim

Equations (3.30b) and (3.30c) define a Bayesian model in which Eq. (3.30b) defines the *prior distribution* $p(\boldsymbol{x}_{0:k})$ of the process of interest $\{\boldsymbol{x}_k\}$, that is:

$$p(\boldsymbol{x}_{0:k}) = \mu(\boldsymbol{x}_0) \prod_{i=0}^{k} p(\boldsymbol{x}_{i+1}|\boldsymbol{x}_{0:i}) \qquad \text{by chain rule} \tag{3.34a}$$

$$= \underbrace{\mu(\boldsymbol{x}_0)}_{\textbf{initial}} \prod_{i=0}^{k} \underbrace{f(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i)}_{\textbf{dynamics}} \qquad \text{by Markov on } \{\boldsymbol{x}_k\} \tag{3.34b}$$

The Markov property is highly desirable since it provides a memory efficient way of keeping track of the evolution of a dynamic phenomenon. Rather than keeping track of the growing full history of the process $\{\boldsymbol{x}_k\}$, it is sufficient to keep track of the present state $\boldsymbol{x}_k$ of the process. Hence, in a Markov process the current state contains everything all the information about the past and the present in order to predict the future. As a matter of fact, the prior (joint) distribution of all states can be written as in Eq. (3.34).

Instead, Eq. (3.30c) defines the *likelihood* or *data distribution* $p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_{0:k})$ as:

$$\begin{aligned} p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_{0:k}) &= \prod_{i=1}^{k} p(\boldsymbol{y}_i|\boldsymbol{y}_{1:i-1}, \boldsymbol{x}_{0:i}) && \text{by chain rule} && (3.35a) \\ &= \prod_{i=1}^{k} \underbrace{g(\boldsymbol{y}_i|\boldsymbol{x}_i)}_{\textbf{observation}} && \text{by conditionally independence} && (3.35b) \end{aligned}$$

In such a Bayesian context, inference about $\boldsymbol{x}_{0:k}$ given a realization of the observations $\boldsymbol{y}_{1:k}$ relies upon the *posterior* distribution $p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{1:k})$, that is:

$$\begin{aligned} p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{1:k}) &\equiv \frac{p(\boldsymbol{x}_{0:k}, \boldsymbol{y}_{1:k})}{p(\boldsymbol{y}_{1:k})} && \text{by definition} && (3.36a) \\ &= \frac{\overbrace{p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_{0:k})}^{\textbf{likelihood}} \overbrace{p(\boldsymbol{x}_{0:k})}^{\textbf{prior}}}{\underbrace{p(\boldsymbol{y}_{1:k})}_{\textbf{evidence}}} && \text{by Bayes' rule} && (3.36b) \\ &= \frac{p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_{0:k})p(\boldsymbol{x}_{0:k})}{\int p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_{0:k})p(\boldsymbol{x}_{0:k})\,\mathrm{d}\boldsymbol{x}_{0:k}} && \text{by marginalization} && (3.36c) \end{aligned}$$

For most non-linear non-Gaussian models, it is not possible to compute the distributions in Eq. (3.36) in closed-form and numerical methods are needed. In particular, the full posterior formulation has the serious disadvantage that each time a new measurement is obtained, the full posterior distribution would have to be recomputed. The problem arises particularly in the context of dynamic estimation, where measurements are typically obtained one at a time, and the the best possible estimate have to be recomputed after each measurement. When the number of time steps increases, the dimensionality of the full posterior distribution also increases, and in turns, the computational complexity of a single time step increases. Eventually, the computations becomes intractable and additional information or restrictive approximations are needed. For instance, by relaxing some conditions, e.g. by being satisfied with selected marginal distributions of the states, the computations become an order of magnitude lighter. Another useful relaxation is to restrict the class of dynamic models to probabilistic Markov sequences, as already shown in Eq. (3.30) where the Markov property ensures a memory efficient way to keep track of the available information. In this way, the computational complexity of a single time step is fixed, and optimal Bayesian filters can be implemented. The most favorable case is a linear Gaussian model as described in Eq. (3.1), where it will be shown that the full posterior $p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{0:k})$ is a gaussian distribution whose mean and covariance can be computed using Kalman techniques.

3.2.3 Filtering as a recursive Bayesian estimation

Filtering is a problem of characterizing the distribution of the state of the hidden Markov model at the present time, given the information provided by all of the observations received up to the present time. This can be thought of as a “tracking” problem: keeping track of the current “location” of the system given noisy observations. The term is sometimes also used to refer to the practice of estimating the full trajectory of the state sequence up to the present time given the observations received up to this time.

Regarding the filtering of the full trajectory of the hidden states, the problem of its estimation is solved through the unnormalized posterior distribution $p(\boldsymbol{x}_{0:k}, \boldsymbol{y}_{1:k})$ in Eq. (3.36a), for which:

$$
\begin{aligned}
p(\boldsymbol{x}_{0:k}|\boldsymbol{y}_{1:k}) &\equiv \frac{p(\boldsymbol{x}_{0:k}, \boldsymbol{y}_{1:k})}{p(\boldsymbol{y}_{1:k})} && \text{(3.37a)} \\
&\propto p(\boldsymbol{x}_{0:k}, \boldsymbol{y}_{1:k}) && \text{(3.37b)}
\end{aligned}
$$

up to some normalization constant. Equation (3.37) is particularly useful when the computation of the evidence model $p(\boldsymbol{y}_{1:k}) = \int p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_{0:k})p(\boldsymbol{x}_{0:k})\,\mathrm{d}\boldsymbol{x}_{0:k}$ is intractable.

Instead, for Bayesian networks as in Eq. (3.30) the unnormalized posterior distribution $p(\boldsymbol{x}_{0:k}, \boldsymbol{y}_{1:k})$ in Eq. (3.36a) satisfies:

$$
\begin{aligned}
p(\boldsymbol{x}_{0:k}, \boldsymbol{y}_{1:k}) &= p(\boldsymbol{y}_{1:k}|\boldsymbol{x}_{0:k})p(\boldsymbol{x}_{0:k}) && \text{(3.38a)} \\
&= \underbrace{\left(\prod_{i=1}^{k} \underbrace{g(\boldsymbol{y}_i|\boldsymbol{x}_i)}_{\textbf{observation}}\right)}_{\textbf{likelihood}} \underbrace{\left(\prod_{i=0}^{k} \underbrace{f(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i)}_{\textbf{dynamics}}\right) \underbrace{\mu(\boldsymbol{x}_0)}_{\textbf{initial}}}_{\textbf{prior}} && \text{(3.38b)}
\end{aligned}
$$

Equation (3.38) can also be written in a recursive pattern by making use of the notation $\mathcal{X}_k = \boldsymbol{x}_{0:k}$ and $\mathcal{Y}_k = \boldsymbol{y}_{1:k}$. Then Eq. (3.38) becomes:

$$
p(\mathcal{X}_k, \mathcal{Y}_k) = p(\mathcal{X}_{k-1}, \mathcal{Y}_{k-1})f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})g(\boldsymbol{y}_k|\boldsymbol{x}_k) \qquad \text{(3.39)}
$$

Consequently, the posterior $p(\mathcal{X}_k|\mathcal{Y}_k)$ in Eq. (3.36) satisfies the following recursion:

$$
\begin{aligned}
p(\mathcal{X}_k|\mathcal{Y}_k) &\equiv \frac{p(\mathcal{X}_k, \mathcal{Y}_k)}{p(\mathcal{Y}_k)} && \text{(3.40a)} \\
&= \frac{p(\mathcal{X}_{k-1}, \mathcal{Y}_{k-1})}{p(\mathcal{Y}_k)} f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})g(\boldsymbol{y}_k|\boldsymbol{x}_k) && \text{(3.40b)} \\
&= \frac{p(\mathcal{X}_{k-1}|\mathcal{Y}_{k-1})}{p(\mathcal{Y}_{k-1})} \frac{p(\mathcal{Y}_{k-1})}{p(\boldsymbol{y}_k|\mathcal{Y}_{k-1})} f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})g(\boldsymbol{y}_k|\boldsymbol{x}_k) && \text{(3.40c)} \\
&= \frac{p(\mathcal{X}_{k-1}|\mathcal{Y}_{k-1})}{p(\boldsymbol{y}_k|\mathcal{Y}_{k-1})} f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})g(\boldsymbol{y}_k|\boldsymbol{x}_k) && \text{(3.40d)}
\end{aligned}
$$

where

$$\begin{aligned} p(\boldsymbol{y}_k|\mathcal{Y}_{k-1}) &= \int p(\boldsymbol{y}_k|\boldsymbol{x}_k, \mathcal{Y}_{k-1})p(\boldsymbol{x}_k|\mathcal{Y}_{k-1})\,\mathrm{d}\boldsymbol{x}_k && (3.41a)\\ &= \int p(\boldsymbol{y}_k|\boldsymbol{x}_k, \mathcal{Y}_{k-1})\left(\int p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}, \mathcal{Y}_{k-1})p(\boldsymbol{x}_{k-1}|\mathcal{Y}_{k-1})\,\mathrm{d}\boldsymbol{x}_{k-1}\right)\mathrm{d}\boldsymbol{x}_k && (3.41b)\\ &= \int g(\boldsymbol{y}_k|\boldsymbol{x}_k)\left(\int f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})p(\boldsymbol{x}_{k-1}|\mathcal{Y}_{k-1})\,\mathrm{d}\boldsymbol{x}_{k-1}\right)\mathrm{d}\boldsymbol{x}_k && (3.41c)\\ &= \int p(\boldsymbol{x}_{k-1}|\mathcal{Y}_{k-1})f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})g(\boldsymbol{y}_k|\boldsymbol{x}_k)\,\mathrm{d}\boldsymbol{x}_{k-1:k} && (3.41d) \end{aligned}$$

In the literature, the recursion satisfied by the marginal posterior distribution $p(\boldsymbol{x}_k|\mathcal{Y}_k)$ is often presented instead of the joint posterior distribution $p(\mathcal{X}_k|\mathcal{Y}_k)$ of Eq. (3.40). By means of a straightforward marginalization, e.g. by integrating out $\mathcal{X}_{k-1}$, the marginal posterior distribution $p(\boldsymbol{x}_k|\mathcal{Y}_k)$ can be easily derived as:

$$\begin{aligned} p(\boldsymbol{x}_k|\mathcal{Y}_k) &= \int \frac{p(\mathcal{X}_{k-1}|\mathcal{Y}_{k-1})}{p(\boldsymbol{y}_k|\mathcal{Y}_{k-1})} f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})g(\boldsymbol{y}_k|\boldsymbol{x}_k)\,\mathrm{d}\mathcal{X}_{k-1} && (3.42a)\\ &= \frac{g(\boldsymbol{y}_k|\boldsymbol{x}_k)}{p(\boldsymbol{y}_k|\mathcal{Y}_{k-1})}\int f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})p(\mathcal{X}_{k-1}|\mathcal{Y}_{k-1})\,\mathrm{d}\mathcal{X}_{k-1} && (3.42b)\\ &= \frac{g(\boldsymbol{y}_k|\boldsymbol{x}_k)}{p(\boldsymbol{y}_k|\mathcal{Y}_{k-1})}\int f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})p(\boldsymbol{x}_{k-1}|\mathcal{Y}_{k-1})\,\mathrm{d}\boldsymbol{x}_{k-1} && (3.42c)\\ &= p(\boldsymbol{x}_k|\mathcal{Y}_{k-1})\frac{g(\boldsymbol{y}_k|\boldsymbol{x}_k)}{p(\boldsymbol{y}_k|\mathcal{Y}_{k-1})} && (3.42d) \end{aligned}$$

where

$$p(\boldsymbol{x}_k|\mathcal{Y}_{k-1}) = \int f(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})p(\boldsymbol{x}_{k-1}|\mathcal{Y}_{k-1})\,\mathrm{d}\boldsymbol{x}_{k-1} \qquad (3.43)$$

Equation (3.42) is usually known as *correction* step while Eq. (3.43) as *prediction* step.

3.2.4 Evolution of the marginal posterior distribution

As seen previously, the marginal posterior distributions are chained by the constitutive relationship of the dynamic of the state and by the new available information in the measurement in a recursive relationship. Thus, the filtering problem is solved by a recursive computation of the densities of interest by means of a forward computation, e.g. sequentially in time for from $k = 0$ to $k = n_k$. Other densities of interest may be computed in a similar manner, for instance densities regarding the prediction or smoothing problem, see Table 3.1. In general also backward computations, e.g. from $k = n_k$ to $k = 0$, or a composition of forward and backward computations may be used instead. At time $k+1$, the a-priori pdf $p(\boldsymbol{x}_{k+1}|\mathcal{Y}_k)$ relates to the distribution of the state at time $k+1$ from the past observations $\mathcal{Y}_k$ before the current observation $\boldsymbol{y}_{k+1}$ is made available. In other words, it should be considered as the time update from the a-posteriori pdf from the previous time sample using the evolution of the state. Assuming that the a-posteriori pdf $p(\boldsymbol{x}_k|\mathcal{Y}_k)$ is known, the evolution of the marginal posterior distribution is based on a two-step procedure:

1. The prediction step (estimation of the a-priori pdf at time k), that is the time

Table 3.1: Prediction, filtering, and smoothing densities of particular interest.

Aim	PDF
Marginal filtering	$p(\boldsymbol{x}_k \vert \mathcal{Y}_k)$
Joint filtering	$p(\boldsymbol{x}_{0:k} \vert \mathcal{Y}_k)$
Prediction	$p(\boldsymbol{x}_{k+1} \vert \mathcal{Y}_k)$
Fixed-step prediction	$p(\boldsymbol{x}_{k+T} \vert \mathcal{Y}_k), \quad T > k$
Marginal smoothing	$p(\boldsymbol{x}_k \vert \mathcal{Y}_N), \quad N > k$
Joint smoothing	$p(\boldsymbol{x}_{0:k} \vert \mathcal{Y}_N), \quad N > k$
Fixed-lag smoothing	$p(\boldsymbol{x}_{k-l+1:k} \vert \mathcal{Y}_k), \quad l < k$

update;

2. The correction step (estimation of a-posteriori pdf at time k), that is the observation update;

A graphical description of the evolution of the marginal posterior is shown in Fig. 3.2.

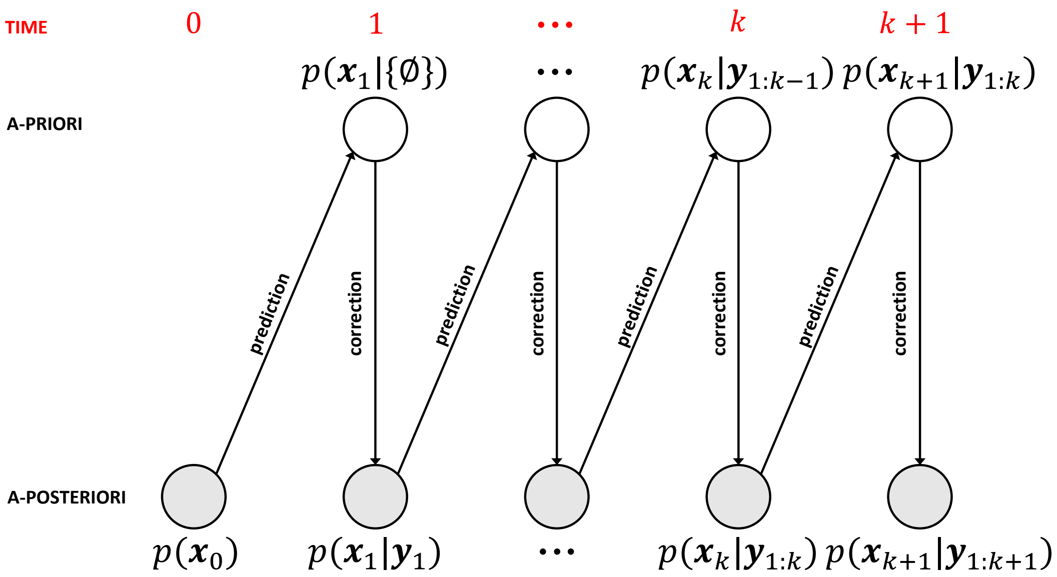

TIME
0
1
...
k
$k+1$
$p(\boldsymbol{x}_1 \vert \{\emptyset\})$
...
$p(\boldsymbol{x}_k \vert \boldsymbol{y}_{1:k-1})$
$p(\boldsymbol{x}_{k+1} \vert \boldsymbol{y}_{1:k})$
A-PRIORI
prediction
correction
prediction
correction
prediction
correction
prediction
correction
A-POSTERIORI
$p(\boldsymbol{x}_0)$
$p(\boldsymbol{x}_1 \vert \boldsymbol{y}_1)$
...
$p(\boldsymbol{x}_k \vert \boldsymbol{y}_{1:k})$
$p(\boldsymbol{x}_{k+1} \vert \boldsymbol{y}_{1:k+1})$

Figure 3.2: The evolution of the marginal posterior $p(\boldsymbol{x}_k \vert \mathcal{Y}_k)$ through time as a composition of the sequential prediction and correction steps.

3.2.5 The prediction step

Given the a-posteriori pdf at time k, i.e. $p(\boldsymbol{x}_k \vert \mathcal{Y}_k)$, the prediction makes use of the dynamic model of state evolution in Eq. (3.1a) to derive the a-priori pdf $p(\boldsymbol{x}_{k+1} \vert \mathcal{Y}_k)$. In other words, the prediction step implements the following transition:

$$p(\boldsymbol{x}_k \vert \mathcal{Y}_k) \rightarrow p(\boldsymbol{x}_{k+1} \vert \mathcal{Y}_k) \tag{3.44}$$

In practice, the the a-priori pdf $p(\boldsymbol{x}_{k+1} \vert \mathcal{Y}_k)$ at time step $k+1$ is derived by the a-posteriori pdf $p(\boldsymbol{x}_k \vert \mathcal{Y}_k)$ at the previous time step k by means of the *Chapman-*

Kolmogorov equation:

$$p(\boldsymbol{x}_{k+1}|\mathcal{Y}_k) = \int f(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\mathcal{Y}_k)\,\mathrm{d}\boldsymbol{x}_k \tag{3.45}$$

i.e. a straightforward marginalization over the nuisance variable $\boldsymbol{x}_k$.

3.2.6 The correction step

The *correction step* takes care of the *measurement update*, i.e. update the *a-priori* information with the new available information conveyed by the measurement in order to get the *a-posteriori* information. In terms of pdf, the correction step implements the following transition:

$$p(\boldsymbol{x}_k|\mathcal{Y}_{k-1}) \to p(\boldsymbol{x}_k|\mathcal{Y}_k) \tag{3.46}$$

In particular, the a-posteriori pdf can be derived when a new observation is made available to confirm or modify the a-priori pdf evaluated from all the observations up to time $k-1$. The observation $\boldsymbol{y}_k$ is accounted for using the Bayes relationship to derive the a-posteriori pdf:

$$p(\boldsymbol{x}_k|\mathcal{Y}_{k-1}) = \frac{g(\boldsymbol{y}_k|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\mathcal{Y}_{k-1})}{p(\boldsymbol{y}_k|\mathcal{Y}_{k-1})} \tag{3.47}$$

Once again, there is no guarantee that there is a closed form relationship of the a-posteriori pdf, except for the special case of linear Gaussian models.

3.2.7 The Kalman filter equations

The Kalman filter [98] is the closed form solution to the Bayesian filtering equations for the filtering model, where the dynamic and measurement models are linear Gaussian:

$$\boldsymbol{x}_0 \sim \mu(\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_0; \bar{\boldsymbol{x}}_0, \boldsymbol{P}_0) \tag{3.48a}$$

$$\boldsymbol{x}_{k+1}|\boldsymbol{x}_k \sim f(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k) = \mathcal{N}(\boldsymbol{x}_{k+1}; \boldsymbol{A}_k\boldsymbol{x}_k, \boldsymbol{G}_k\boldsymbol{Q}_k\boldsymbol{G}_k^\mathsf{T}) \tag{3.48b}$$

$$\boldsymbol{y}_k|\boldsymbol{x}_k \sim g(\boldsymbol{y}_k|\boldsymbol{x}_k) = \mathcal{N}(\boldsymbol{x}_k; \boldsymbol{C}_k\boldsymbol{x}_k, \boldsymbol{R}_k) \tag{3.48c}$$

In particular, the Kalman filter can be seen as a special case of the relationships for prediction and correction in Sections 3.2.5 and 3.2.6 where the pdfs are represented by the first and second order central moments, and there is a closed form for the statistical evolution of the state from the observations. More specifically, since at every step the pdfs are Gaussian, it is enough to derive the prediction and update equations for the statistics that completely characterize the distributions, e.g. the mean and the covariance. For instance, the scheme of prediction-correction relationship w.r.t. the estimate statistics is available in Fig. 3.3. As a matter of fact, there is the following pairing of the pdfs and their properties:

$$p(\boldsymbol{x}_k|\mathcal{Y}_k) \equiv \mathcal{N}(\boldsymbol{x}_{k|k}; \hat{\boldsymbol{x}}_{k|k}, \boldsymbol{P}_{k|k}) \tag{3.49}$$

$$p(\boldsymbol{x}_{k+1}|\mathcal{Y}_k) \equiv \mathcal{N}(\boldsymbol{x}_{k+1|k}; \hat{\boldsymbol{x}}_{k+1|k}, \boldsymbol{P}_{k+1|k}) \tag{3.50}$$

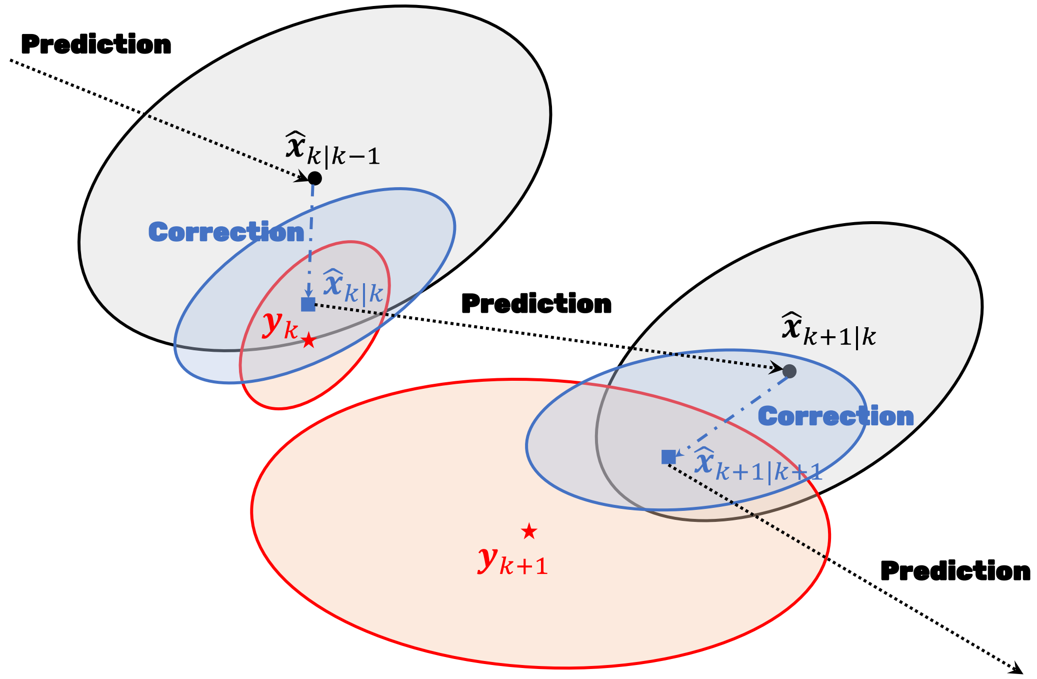


Figure 3.3: A scheme of the iterative prediction-correction relationship that takes place in the Kalman filtering theory.

The Bayesian filtering equations for the linear filtering model can be evaluated in closed form and the resulting distributions are Gaussian. Consider the following joint distribution:

$$
\begin{aligned}
p(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k | \mathcal{Y}_k) &= f(\boldsymbol{x}_{k+1} | \boldsymbol{x}_k) p(\boldsymbol{x}_k | \mathcal{Y}_k) && (3.51a)\\
&= \mathcal{N}(\boldsymbol{x}_{k+1}; \boldsymbol{A}_k \boldsymbol{x}_k, \boldsymbol{G}_k \boldsymbol{Q}_k \boldsymbol{G}_k^\top) \mathcal{N}(\boldsymbol{x}_{k|k}; \hat{\boldsymbol{x}}_{k|k}, \boldsymbol{P}_{k|k}) && (3.51b)\\
&= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_k \\ \boldsymbol{x}_{k+1} \end{bmatrix}; \begin{bmatrix} \hat{\boldsymbol{x}}_{k|k} \\ \boldsymbol{A}_k \hat{\boldsymbol{x}}_{k|k} + \boldsymbol{B}_k \boldsymbol{u}_k \end{bmatrix}, \begin{bmatrix} \boldsymbol{P}_{k|k} & \boldsymbol{P}_{k|k} \boldsymbol{A}_k^\top \\ \boldsymbol{A}_k \boldsymbol{P}_{k|k} & \boldsymbol{A}_k \boldsymbol{P}_{k|k} \boldsymbol{A}_k^\top + \boldsymbol{G}_k \boldsymbol{Q}_k \boldsymbol{G}_k^\top \end{bmatrix} \right) && (3.51c)
\end{aligned}
$$

Correction equations

The correction equation is given by:

$$
\begin{aligned}
p(\boldsymbol{x}_k | \mathcal{Y}_k) &= \mathcal{N}(\boldsymbol{x}_k; \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{C}_k \hat{\boldsymbol{x}}_{k|k-1} - \boldsymbol{D}_k \boldsymbol{u}_k), \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k \boldsymbol{S}_k \boldsymbol{K}_k^\top) && (3.52a)\\
&\equiv \mathcal{N}(\boldsymbol{x}_{k|k}; \hat{\boldsymbol{x}}_{k|k}, \boldsymbol{P}_{k|k}) && (3.52b)
\end{aligned}
$$

where $\boldsymbol{K}_k$ is known as the *Kalman gain*:

$$
\begin{aligned}
\boldsymbol{K}_k &= \boldsymbol{P}_{k|k-1} \boldsymbol{C}_k^\top (\boldsymbol{C}_k \boldsymbol{P}_{k|k-1} \boldsymbol{C}_k^\top + \boldsymbol{R}_k) && (3.53a)\\
&= \boldsymbol{P}_{k|k-1} \boldsymbol{C}_k^\top \boldsymbol{S}_k && (3.53b)
\end{aligned}
$$

with:

$$S_k = C_k P_{k|k-1} C_k^\mathsf{T} + R_k \tag{3.54}$$

where the statistics values for the mean and covariance are:

$$\begin{aligned}
\hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k(y_k - C_k \hat{x}_{k|k-1} - D_k u_k) && (3.55a)\\
P_{k|k} &= P_{k|k-1} - K_k S_k K_k^\mathsf{T} && (3.55b)\\
&= P_{k|k-1} - K_k S_k K_k^\mathsf{T} && (3.55c)
\end{aligned}$$

Prediction equations

The prediction equation is given by marginalization of x_k in Eq. (3.51). The computation of the integral has a closed form solution and is given by Eq. (A.5), which yields:

$$\begin{aligned}
p(x_{k+1}|\mathcal{Y}_k) &= \mathcal{N}(x_{k+1}; A_k \hat{x}_{k|k} + B_k u_k, A_k P_{k|k} A_k^\mathsf{T} + G_k Q_k G_k^\mathsf{T}) && (3.56a)\\
&\equiv \mathcal{N}(x_{k+1}; \hat{x}_{k+1|k}, P_{k+1|k}) && (3.56b)
\end{aligned}$$

where the statistics values for the mean and covariance are:

$$\begin{aligned}
\hat{x}_{k+1|k} &= A_k \hat{x}_{k|k} + B_k u_k && (3.57a)\\
P_{k+1|k} &= A_k P_{k|k} A_k^\mathsf{T} + G_k Q_k G_k^\mathsf{T} && (3.57b)
\end{aligned}$$

PART II

Contributions to the filtering design problem

CHAPTER 4

The model-based and data-driven paradigms to the filtering design problem

The following chapter contains a brief introduction to personal contributions and new research to the topic of the filtering design problem. It was written to present the standard design paradigm commonly found in the historical development of the filtering design problem in the overview in Chapter 1 as well as the new "direct" paradigm. As a consequence, it can be thought of as an extension of the already mentioned historical development. The chapter contains also trace of studies in other communities, such as the control community and the system identification community. For a deeper technical understanding, please refer also to the dedicated Chapter 5 that gives a detailed overview of the implementation of the standard solution and to the dedicated Chapter 6 that gives a detailed overview of the implementation of the direct solution.

This chapter is organized as follow:

1. Section 4.1 highlights the nature of the classical solutions to the filtering design problem, i.e. a model-based paradigm.
2. Section 4.2 introduces instead the data-driven paradigm to derive solutions to the filtering design problem. This is required in practice when the historical assumptions do not hold true. The section is mainly dedicated to two techniques, highlighting their differences: **a)** the standard solution, and **b)** the new "direct" solution.

4.1 Model-based paradigm to the filtering problem

So far in Chapter 1, solutions to the filtering design problem have been discussed in the case where the system under analysis is given. In order to have a better understanding, it is now considered the general mathematical framework in Eq. (4.1) that summarized the historical solutions derived in Chapter 1. In particular, considering the stochastic filtering theory, the framework involves a stochastic dynamical system $\mathcal{S}$ referred in discrete-time domain and described in state-space form. In general, the system is non-linear, non-stationary, and is corrupted by noise signals, see [14]:

$$\mathcal{S}: \begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{w}_k; k) && \text{(4.1a)} \\ \boldsymbol{y}_k &= \boldsymbol{h}_y(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{v}_k; k) && \text{(4.1b)} \end{aligned}$$

As already mentioned, in the derived solutions, the transition equation $\boldsymbol{f}(\cdot)$ (also known as diffusion equation, state equation) and the measurement equation $\boldsymbol{h}_y(\cdot)$ are thought to be known a-priori, i.e. they are fixed. In this context, the class of filtering design solutions can be categorized as *model-based*: once a model of the system is given, a filter is designed based only on the fixed model of the system. This kind of paradigm is, thus, referred to as the *model-based paradigm* to the *Filtering Design* (FD) problem.

4.2 Data-driven paradigm to the filtering problem

However, in most practical situations, this is not the case. As a matter of fact, this led to an alternative branch of research, still active nowadays, related to the so-called *data-driven* design solutions to the filtering problem. The main two solutions are further explored in the following sections: i) the so-called standard solution used in many applications nowadays, and ii) the direct solution, a new branch of research studied and developed in more details in this dissertation. A summarizing schematic of the two paradigms is given in Fig. 4.1 where high-level differences, namely the different steps taken and the intermediate ingredients, are highlighted.

4.2.1 The standard solution: a two-step approach

At the time of the writing of this dissertation, the *standard solution* to the data-driven design is based on an a two-steps approach, where the sequential steps to be performed are as follows:

1. First, a model of the system is estimated from available input-output data (measured through some experiments) using the best techniques. This step is known as the *Data-driven System Identification*, see [20, 30];

2. Secondly, a filter is designed from the identified model. This step is known as the *Model-based filter design*.

   Note that, as suggested by its name, the filter design step is model-based, founded on the "certainty equivalence principle", i.e. the identified model $\hat{\mathcal{S}} = \mathcal{M}(\hat{\theta}_N)$, parameterized by $\hat{\theta}_N$, is treated as if it represents the true system $\mathcal{S}$. Following this rationale, solutions to the FD problem are applied on $\hat{\mathcal{S}}$;

The described methodology is a sequential two-steps procedure known as the standard solution to the *Filter Design from Data* (FD2) problem. In particular, this solution is categorized as data-driven due to the fact that experimental data are used in the

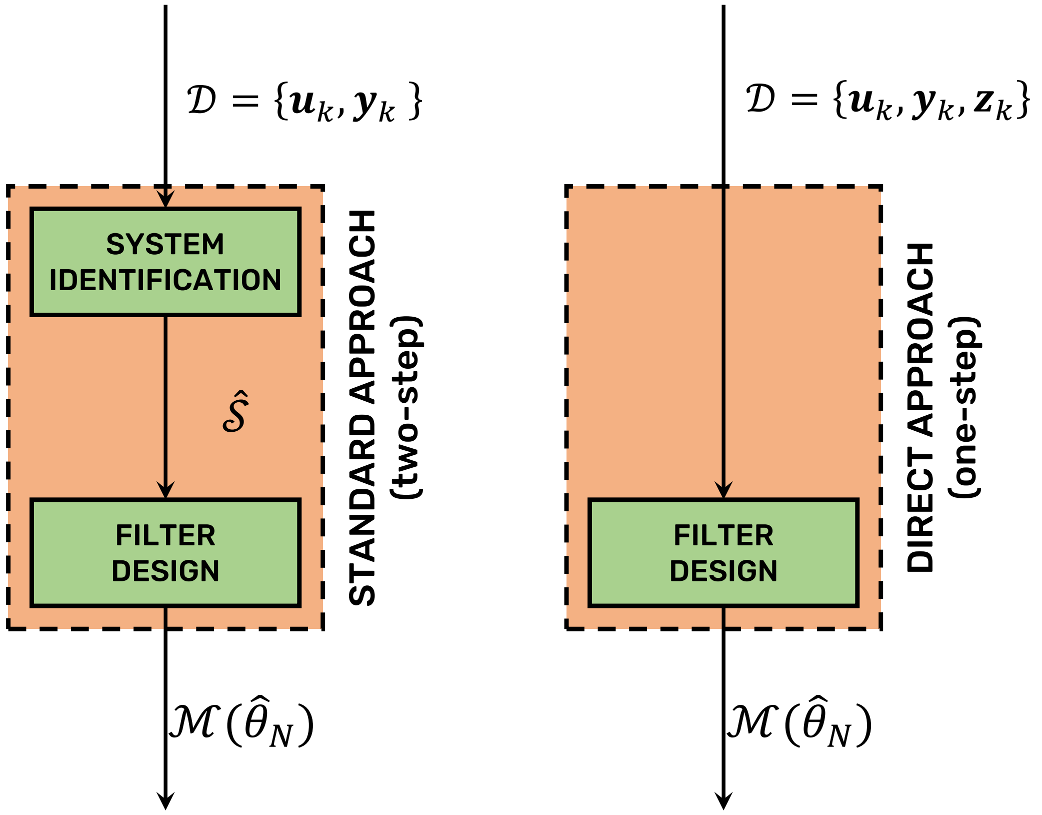

$\mathcal{D} = \{\boldsymbol{u}_k, \boldsymbol{y}_k\}$
SYSTEM IDENTIFICATION
$\hat{\mathcal{S}}$
FILTER DESIGN
STANDARD APPROACH (two-step)
$\mathcal{M}(\hat{\theta}_N)$
$\mathcal{D} = \{\boldsymbol{u}_k, \boldsymbol{y}_k, \boldsymbol{z}_k\}$
FILTER DESIGN
DIRECT APPROACH (one-step)
$\mathcal{M}(\hat{\theta}_N)$

Figure 4.1: A scheme of the data-driven filtering solution. On the left, the standard solution based on a two-step approach. On the right the direct solution based on a one-step approach.

system identification step even if the whole procedure is indeed hybrid, i.e. half data-driven and half model-based, because of the model-based filter design step. Therefore, it is interesting to study a pure data-driven paradigm, labeled as the “direct” solution, that is researched as the main contribution in this dissertation.

However, before attempting this journey, the standard methodology is analyzed in more detailed with the aim to highlight its flaws and its common misconceptions.

The data-driven system identification step It is now explored in more details the methodology of the first step of the standard solution to the data-driven filtering design problem, namely the data-driven system identification step. In this context, the desired idea is to obtain a model of the system. Again, this branch of research is known in literature as *System Identification* and was developed including different classes of methods depending on the made assumption for the domain of expertise. The most common approach, for instance, and the one that is used in this dissertation — hence the “data-driven” attached term — is to perform the modeling from experiments. In other words, the system is excited with some inputs and the outputs are observed; then, these data are used to identify the process that links inputs and outputs. This particular approach is known as *black-box* modeling, as it refers to a procedure that does not go into many details of what is actually happening inside the system. To summarize, first an optimal design experiment is performed in order to sample informative input and output data on the running system. Then, the collected data are exploited as ingredients in statistical methods to build mathematical models of the dynamical system. In the end, the unknown parameters of the system's model are estimated by means of statistical learning data-driven methods.

However, black-box system identification is not the only possible way to do modeling. Its opposite, when no experiments are performed and only the knowledge of experts in the field is used for the estimation of the model, i.e. modeling by "first principles", is instead termed *white-box* modeling. This is the case, for instance, when the system is not very complex or the physical laws that governs the analyzed process are well known and understood. In other applications, the two approaches are often used together (*grey-box* modeling): the physical approach is used to define a model from first principles, and the identification approach is used to fit its parameters, so that the model agrees with what it is observed [184].

The model-based filter design step Once the data-driven system identification step is performed, a model $\hat{\mathcal{S}}$ of the system $\mathcal{S}$ is available. Following this rationale and applying the "certainty equivalence principle" the identified model $\hat{\mathcal{S}} = \mathcal{M}(\hat{\theta}_N)$ is treated as if it represents the true system $\mathcal{S}$. Founding on this principle, common solutions to the filtering design (FD) problem as described in Chapter 1 can be, and are, applied on $\hat{\mathcal{S}}$. In the case of LTI systems and filters, with a LQG setting, Chapter 3 presents the BLUE (*Best Linear Unbiased Estimator*) solution to the filtering design problem, namely the Kalman Filter, under the assumption that the system is known, or it has been estimated from a a-priori system identification step following the rationale.

4.2.2 A brief historical review of system identification

The following section tries to present a brief overview of the system identification history from its roots to present days to complement the data-driven standard solution to the filtering design problem. It is interesting to note the developed and the expected to-be developing historical and technical entanglements between the control problems, the system identification problems, and the filtering problems in the community related to the analysis of dynamical systems. The herein references are taken mainly from different source of historical references, see [82, 153] and the mentioned authors' works.

Before 1965, parameter estimation techniques had been applied for some time to the control of systems with known structure but unknown (or poorly known) parameters.

1960s The roots of black-box identification in the control community were developed independently by Ho and Kalman, Silverman, and others, starting from the *deterministic realization theory* from 1965, see for instances [90, 188]. The theory tackle the problem to solve how to determine a finite-dimensional state-space realization from the infinite-dimensional representation of an LTI dynamical system fully described by its Markovian parameters (also known as impulse response parameters).

In other words, given the input-output (IO) model:

$$y_k = \sum_{i=1}^{\infty} \boldsymbol{H}_i u_{k-i} \tag{4.2}$$

described by its impulse response matrices $\boldsymbol{H}_k \in \mathbb{R}^{n_y \times n_u}$. The problem is to find a replacement for the infinite description:

$$H(z) = \sum_{i=1}^{\infty} \boldsymbol{H}_i z^{-i} \tag{4.3}$$

with a finite description $\boldsymbol{A} \in \mathbb{R}^{n_x \times n_x}, \boldsymbol{B} \in \mathbb{R}^{n_x \times n_u}, \boldsymbol{C} \in \mathbb{R}^{n_y \times n_x}$ so that

$$H(z) = \boldsymbol{C}(z\boldsymbol{I} - \boldsymbol{A})^{-1} - \boldsymbol{B} \tag{4.4}$$

and $\boldsymbol{A}$ has minimal dimension.

This problem can be divided into two parts, i.e. find the McMillan degree of the impulse response transfer function $\boldsymbol{H}(z)$, which is then the minimal dimension of $\boldsymbol{A}$, and compute the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$. The key observation used for solving this problem was that the Hankel matrix $\mathcal{H}$ when properly dimensioned versus the order of the LTI system, can be factorized through an SVD procedure into the product of an infinite observability matrix and infinite controllability matrix:

$$\mathcal{H} = \begin{bmatrix} \boldsymbol{H}_1 & \boldsymbol{H}_2 & \boldsymbol{H}_3 & \dots \\ \boldsymbol{H}_2 & \boldsymbol{H}_3 & \boldsymbol{H}_4 & \dots \\ \boldsymbol{H}_3 & \boldsymbol{H}_4 & \boldsymbol{H}_5 & \dots \\ \vdots & \vdots & \vdots & \end{bmatrix} \tag{4.5}$$

$$= \underbrace{\begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{C}\boldsymbol{A} \\ \boldsymbol{C}\boldsymbol{A}^2 \\ \vdots \end{bmatrix}}_{\text{Observability matrix}} \underbrace{\begin{bmatrix} \boldsymbol{B} & \boldsymbol{A}\boldsymbol{B} & \boldsymbol{A}^2\boldsymbol{B} & \dots \end{bmatrix}}_{\text{Controllability matrix}} \tag{4.6}$$

Again, the SVD of the Hankel matrix provides a basis of the column space observability matrix and row space of the controllability matrix of the LTI system. The knowledge of these spaces enable the estimation of a realization of the system matrices. Indeed, the Ho-Kalman realization method [90] is based on the following properties: if the McMillan degree of $H(z)$ is n_x, then:

1. $\text{rank}(\mathcal{H}) = n_x$
2. $\exists \boldsymbol{A} \in \mathbb{R}^{n_x \times n_x}, \boldsymbol{B} \in \mathbb{R}^{n_x \times n_u}, \boldsymbol{C} \in \mathbb{R}^{n_y \times n_x} \mid \boldsymbol{H}_k = \boldsymbol{C}\boldsymbol{A}^{k-1}\boldsymbol{B}, \quad \forall k \geq 1;$

At the same time of the state-space formulation, Åstrom and Bohlin [43] introduced the *maximum likelihood framework* for estimating the parameters of input-output models in ARMAX form, e.g. the following SISO model:

$$A(z^{-1})y_z = B(z^{-1})u_z + \lambda C(z^{-1})e_z \tag{4.7}$$

where e_t is a sequence of independent identically distributed zero-mean, unit-variance Gaussian random variables. The concepts and notation introduced in [43] have been with the *System Identification* theory for almost 60 years now. Some examples of household notation of the community found in are as follows:

- the residuals as $C(z^{-1})\epsilon_z = A(z^{-1})y_z - B(z^{-1})u_z$;
- the cost criterion as $V(\theta) = \frac{1}{2}\sum_{i=0}^{N}\epsilon_i^2$;
- the parameter estimate as $\hat{\theta}_N = \arg\min V(\theta)$;
- the white noise variance estimate as $\hat{\lambda}^2 = \frac{2}{N}V(\hat{\theta}_N)$;

The publication of [43] gave rise to activity in parametric identification and established the basis for the prediction-error framework. As stated in [82]:

> "The step from maximum likelihood to prediction error essentially consists of observing that, under the assumption of white Gaussian noise in the ARMAX model, maximization of the likelihood function of the observations is equivalent to minimizing the sum of the squared prediction errors. The prediction-error framework consists of adopting the minimization of a norm of the prediction errors as the criterion for parameter estimation, even when the probability distribution for the observations is unknown."

1970s In the early 1970s, the combination of deterministic realization theory based on the factorization of the Hankel matrix, with the theory of Markovian and innovations representations, gave rise to the stochastic theory of minimal realizations. The stochastic realization problem was studied intensively in in connection with innovations theory and spectral factorization theory [35, 196, 81].

1970s-1980s The years 1975—1985 saw frantic activity in system identification in the engineering community. The methods based on a prediction-error criterion together with input-output models completely took over the field, at the expense of methods based on realization theory. Their theoretical superiority over stochastic realization methods was based on the statistical properties of the parametric estimates: **a)** prediction-error methods are asymptotically efficient (since their covariance achieves the Cramér-Rao bound), and **b)** the asymptotic accuracy can also be evaluated. The main reason for the growing appeal of prediction-error methods, however, was that increased computer speed and the development of special purpose identification software made it more feasible to iteratively minimize a cost criterion over a range of possible model structures.

During this period, new "methods", i.e. new combinations of model structures and methods, appeared constantly in the scientific journals with claims about their supremacy over existing methods. To solve this problem, Ljung contributed in the field by doing a major clean-up task that was felt needed by means of separating two independent concepts: **a)** the choice of a parametric model structure, which provided a vehicle for computing predictions and hence parameter-dependent prediction errors, and **b)** and the choice of an identification criterion, which was a non-negative function of the prediction errors and hence of the parameter vector [116]. In this setup, all existing parametric identification methods could then be seen as particular cases of this prediction-error framework. In [116], the generic true SISO input-output model structure was introduced, i.e. the data-generation model:

$$\mathcal{S} : \quad y_k = G_0(z)u_k + H_0(z)\eta_k \tag{4.8}$$

and its parameterized model:

$$\mathcal{M}(\theta) : \quad y_k = G(z,\theta)u_k + H(z,\theta)e_k \tag{4.9}$$

where $G(z,\theta)$ and $H(z,\theta)$ are parameterized rational transfer functions and e_k and η_k are white noises. In particular, the parameter θ lives in the parameterized model set:

$$\mathcal{M} = \left\{ G(z,\theta), H(z,\theta), \theta \in \Theta \subset \mathbb{R}^d \right\} \tag{4.10}$$

where $\mathcal{M}$ denotes the *model family* as a whole by defining both the *model structure* (the model structural form, i.e. the model class) and the *model complexity* (the number of its parameters, i.e. the order). The particular problem of selecting a model family is termed *model selection.* Instead $\mathcal{M}(\theta) \in \mathcal{M}$ acts as a particular model or hypothesis given the parameter θ which belongs to the hypothesis space $\mathcal{M}$. In this framework, all commonly used model structures were special cases of the generic structure in Eq. (4.9). Ingredients for the estimation of the parameter θ are the parameter-dependent one-step-ahead prediction:

$$\hat{y}_{k|k-1}(\theta) = H^{-1}(z,\theta)G(z,\theta)u_k + \left[1 - H^{-1}(z,\theta)\right] y_k \tag{4.11}$$

and hence the one-step-ahead prediction error:

$$\epsilon(\theta) = y_k - \hat{y}_{k|k-1}(\theta) = H^{-1}(z,\theta)\Big[\big(G_0(z) - G(z,\theta)\big)u_k + H_0(z,\theta)e_k\Big] \tag{4.12}$$

Next, given an input-output set $\mathcal{Z} = \{\boldsymbol{u}_i, \boldsymbol{y}_i\}_{i=1}^N$ of N data and hence of N prediction errors, the criterion to be minimized can be defined as:

$$V_N(\theta, \mathcal{Z}) = \frac{1}{N}\sum_{i=1}^{N} l\big(\epsilon_i(\theta)\big) \tag{4.13}$$

where $l(\cdot)$ is a non-negative scalar-valued function. Minimizing $V(\theta, \mathcal{Z})$ with respect to θ over its domain Θ then yields the parameter estimate:

$$\hat{\theta}_N = \arg\min_{\theta \in \Theta} V_N(\theta, \mathcal{Z}) \tag{4.14}$$

Under some mild assumptions, the asymptotic case results in:

$$\hat{\theta}_N \to \theta^* \text{as } N \to \infty \tag{4.15}$$

where

$$\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}\Big[l\big(\epsilon_k(\theta)\big)\Big] \tag{4.16}$$

In other words, the procedure is shown to converge to the best approximation of the true system contained in the chosen family of models for large datasets.

This work culminated in the publication of the book [20] in 1987 by Ljung, which has become the standard reference on system identification, both as a theoretical basis and as a guide for applications. Its usefulness for applications has been greatly enhanced by the simultaneous production by Ljung in 1987 of the MATLAB *System Identification Toolbox* [206]. At the same time, Stoica and Sodestrom complemented the work by Ljung with [30], a book which adopted the same clear distinction between choice of model structure and choice of criterion; their contribution focused less on design issues but more on analysis and on alternative criteria, in particular criteria based on correlation methods and instrumental variables. In the end, under this unifying framework known as *Prediction Error Minimization* methods (PEM) different kind of models were grouped together. These included, for instance, time series and dynamical models expressed in their regression representation both having linear and non-linear characteristic: Auto-Regressive models (AR), Auto-Regressive models with eXogenous input (ARX), Output-Error models (OE), Auto-Regressive Moving Average models (ARMA), Auto-Regressive Moving Average with eXogenous input models (ARMAX), and non-linear variants, e.g. Non-linear Auto-Regressive Moving Average with eXogenous input models (NARMAX).

1990s-2000s The reasons for the emergence of subspace identification are to be found in the state of the art of identification of multivariable systems in the 1980s. Even though the manifold structure of MIMO systems had been extensively studied in the late 1970s, there were still open problem related to the practical identification of MIMO systems. The main concern was the parameterizations of multivariable systems through structure indices which resulted in ill-conditioned numerical procedures. On the other hand, other techniques such as singular value decomposition and least squares, which bypassed the need for estimating structure indices, were numerically stable and consequently more interesting. The development of subspace-based identification methods rose in this context, also thanks to the fact that in that framework the handling of MIMO systems causes no additional difficulty. In the early 1990s, several research pioneers contributed to breakthroughs in the mentioned problems, for instance Van Overschee and De Moor — introducing the N4SID approach [139], Verhaegen — introducing the MOESP approach [165, 166] and Larimore — presenting subspace techniques (ST) in the framework of Canonical Variate Analysis (CVA) [202]. Other studies include, for instance, references [61, 194, 62, 101, 143]. See also the valuable textbooks and reviews available in [183, 21, 142].

2010s-2020s In the present days, a revival of the bias-variance tradeoff concept and of the regularization term and its functioning in the system identification was made available by the growing interested in machine-learning. At the same time, novel system identification approaches, based on the kernel-methods, are also being studied in details. For a complete overview of both lines of active research, please refer to [184, 186].

4.2.3 A brief historical review of noise covariance matrices estimation

The following section presents a brief overview of the noise covariance matrices (CMs) estimation history from its roots to present days to complement the data-driven design solutions to the filtering problem. It is interesting to note the developed and the expected to-be developing historical and technical entanglements between the control problems, the system identification problems, and the filtering problems in the community related to the analysis of dynamical systems. The adhere references are taken mainly from different source of historical references, see [70, 72] and the mentioned authors' works.

As stated in the historical review and comparison of estimation methods in [72]:

> "Knowledge of a system model is a key prerequisite for many state estimation, signal processing, fault detection, and optimal control problems. The model is often designed to be consistent with random behaviour of the system quantities and properties of the measurements. While the deterministic part of the model often arises from mathematical modelling on the basis of physical, chemical, or biological laws governing the behaviour of the system, the statistics of the stochastic part are often difficult to find by the modelling and have to be identified using the measured data. Incorrect description of the noise statistics may result in significant worsening of estimation, signal processing, detection, or control quality or even in a failure of the underlying algorithms."

In this sense, in the last five decades, a great research interest has been focused on a design of the methods for the estimation of the properties of the stochastic part of the model. In particular, the following review focuses on state-space models in discrete-time even though attention has been devoted also to input-output models, both with recursive and batch processing methods. It is also important to highlight the interest of the methods estimating the covariance matrices (CMs) of noises in the state and measurement equation from a sequence of measured data, conforming to the framework of the data-driven paradigm of the filtering design problem.

In the literature, an extensive number of various noise CM estimation methods can be found, see for instance industrial applications with components of adaptive control and signal processing systems [40, 75, 136, 147]. The methods differ in assumptions related to the considered model, underlying ideas and principles, properties of the estimates, and number and essence of the design parameters. Traditionally, four groups are related to the noise CMs estimation methods, namely:

1. The correlation methods, see [47, 207, 136, 177] and many more publications [34, 36, 38, 39, 49, 192, 66, 70, 195, 71, 76, 78, 80, 109, 113, 114, 204, 117, 132, 141, 144, 148, 157, 218, 174, 176] extracted from the detailed review [72].

2. The maximum-likelihood methods (MLMs), see [179, 44, 46, 79, 197, 100, 152, 156, 28, 164, 167];

3. The covariance matching methods (CMMs), see [63, 67, 73, 85, 182, 108, 121, 122, 130, 131, 187, 160, 215, 168];

4. The Bayesian methods, see [39, 192, 77, 78, 87, 107, 203, 112, 205, 118–120, 140, 148, 150, 158, 159, 163, 169–171]

In the literature, several papers characterize the methods by their properties, advantages, and disadvantages. In particular, the mentioned four groups of available methods differ for their estimation approach, which ultimately can be:

- *Feedback methods*, where the unknown parameters of the noise CMs are estimated simultaneously with the unknown state by joint or dual estimation techniques. In this case, the augmented state vector enforces that a technique of nonlinear state estimation has to be applied since a relatively simple estimator with a linear structure with respect to the measurement, providing only two conditional moments of the state estimate, fails: the reason can be found in a missing (linear) correlation between the elements of the extended state. Following this rationale, the feedback noise CM estimation methods are covered mostly by the Bayesian methods and the CMMs. Two approaches can be identified within the Bayesian methods: (i) the joint estimation of the state and noise CMs using a nonlinear filter [77, 107, 205, 118, 119, 140, 150, 163, 169–171] and (ii) a multiple model approach [39, 192, 78, 87, 107, 203, 112, 205, 120, 148, 150, 158, 171] which provides CM estimates as a mixture of basis CMs. The CMMs utilize a (non)linear filter for the estimation of the state and aim for making the noise CM estimates and the state and measurement estimate errors consistent [63, 67, 73, 85, 182, 108, 121, 122, 130, 131, 187, 160, 215, 168]; An example of the architecture of these methods is available in Fig. 4.2.

- *Feedback-free methods*, where the state and statistics of the measurement prediction are estimated by a (nonoptimal) estimator for all time instants and then the measurement prediction error and its statistics are used to estimate the noise CMs. In particular, this alternative approach decomposes an inherently nonlinear task into two coupled simpler tasks: i) non-optimal estimation of the state and computation of the measurement prediction error sequence (i.e. the innovation), ii) noise CM estimation by a statistical analysis of the innovation sequence. The estimate of the noise CMs is thus not needed for the state estimate, hence the name feedback-free methods. The architecture of these methods is depicted in Fig. 4.3. The feedback-free noise CM estimation methods cover the correlation methods and the MLMs. The correlation methods are based on an analysis of the innovation sequence properties of a suboptimal linear filter, see [34, 36, 38, 39, 49, 192, 66, 70, 195, 71, 76, 78, 80, 109, 113, 114, 204, 117, 132, 141, 144, 148, 157, 218, 174, 176]. Instead, the MLMs are based on optimization techniques from the system identification designed for input-output models, see for instance [179, 44, 46, 79, 197, 100, 152, 156, 28, 164, 167]. The feedback-free methods for linear models are based on the system observability assumption. The methods can typically estimate all elements of the CM $\boldsymbol{R}$ and no more than $n_x \cdot n_y$ elements of the CM $\boldsymbol{Q}$ if an LTI model is considered.

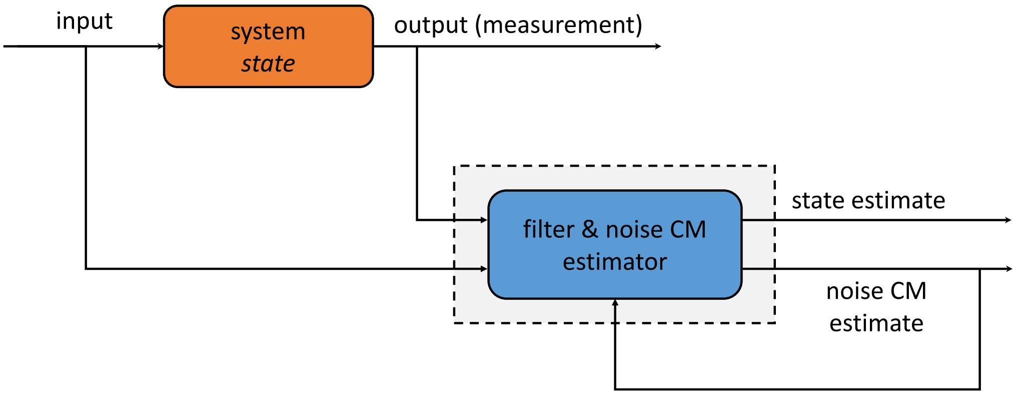


Figure 4.2: Illustration of the feedback architecture for noise CM estimation.

4.2.4 The correlation techniques for noise covariance matrices estimation

Estimation of the noise CMs was pioneered by Mehra in [123, 207] and is denoted as the *Indirect Correlation Method* following the proposition of the review [72]. The method is based on an analysis of the *Auto Correlation Function* (ACF) of the innovation of a linear predictor, which may not be optimal in the mean square error (MSE) sense. Correlation techniques are based on the idea that once the deterministic part of a plant is modelled accurately, the residuals from the deterministic part then carry information about the noises entering the plant. These residuals (or innovations) can then be correlated with each other to extract information about the covariance of the disturbances entering the plant. The introduced basic idea gave rise to a group of methods commonly denoted as the correlation methods. For the sake of simplicity, in this small review are covered only the basic ideas of the pioneering ICM method and later, of the newest research of correlation methods, the so-called *Direct Correlation Method* (DCM) presented by Odelson et al., see [136], which will be exploited for the implementation of a working estimation routine in Chapter 5. The DCM method, also known as *Auto Least-Square* (ALS) technique, offers significant advantages over other techniques in the literature, for instance the basic indirect approach. In particular, the ALS procedure solves a single least-squares problem while the other techniques estimate the covariances in two steps. Solving a single least-squares problem leads to smaller variance in the estimates as opposed to using two steps.

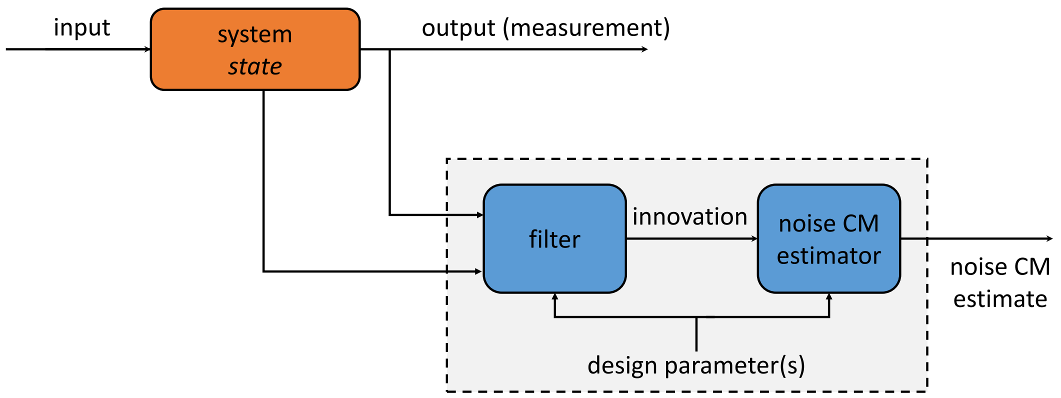


Figure 4.3: Illustration of the feedback-free architecture for noise CM estimation.

The Indirect Correlation Method

The ICM is designed for an LTI model and is based on the definition of an asymptotically stable linear predictor (or filter):

$$\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_k^- + \boldsymbol{L}(\boldsymbol{y}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_k^-) \tag{4.17a}$$

$$\hat{\boldsymbol{x}}_{k+1}^- = \boldsymbol{A}\hat{\boldsymbol{x}}_k + \boldsymbol{B}\boldsymbol{u}_k \tag{4.17b}$$

using the notation $\hat{\boldsymbol{x}}_k \equiv \hat{\boldsymbol{x}}_{k|k}$ and $\hat{\boldsymbol{x}}_k^- \equiv \hat{\boldsymbol{x}}_{k|k-1}$, and with an arbitrary initial condition $\hat{\boldsymbol{x}}_0^-$, where the predictor gain $\boldsymbol{L}$ is selected such that the matrix $\bar{\boldsymbol{A}} \equiv \boldsymbol{A} - \boldsymbol{A}\boldsymbol{L}\boldsymbol{C}$ is stable. Then, the state prediction error $\hat{\boldsymbol{\delta}}_k^- \equiv \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^-$ and the measurement prediction error (i.e. innovation) $\hat{\boldsymbol{\epsilon}}_k^- \equiv \boldsymbol{y}_k - \hat{\boldsymbol{y}}_k^-$ evolve according to the error-state model as follows:

$$\hat{\boldsymbol{\delta}}_{k+1}^- = \bar{\boldsymbol{A}}\hat{\boldsymbol{\delta}}_k^- + \boldsymbol{\Gamma}\bar{\boldsymbol{w}}_k \tag{4.18a}$$

$$\hat{\boldsymbol{\epsilon}}_k^- = \boldsymbol{C}\hat{\boldsymbol{\delta}}_k^- + \boldsymbol{v}_k \tag{4.18b}$$

where the measurement prediction is the usual $\hat{\boldsymbol{y}}_k^- = \boldsymbol{C}\hat{\boldsymbol{x}}_k^-$, and with $\boldsymbol{\Gamma} = [\boldsymbol{I}_{n_x}, -\boldsymbol{A}\boldsymbol{C}]$, $\bar{\boldsymbol{w}}_k = [\boldsymbol{w}_k^\top, \boldsymbol{v}_k^\top]^\top$, and $\boldsymbol{I}_{n_x}$ is the identity matrix of dimension n_x.

In the asymptotic case, the innovation is a zero-mean stochastic process whose *Auto Correlation Function* (ACF) is defined by:

$$\boldsymbol{P}_\epsilon(\tau) \equiv \mathbb{E}[\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_{k-\tau}^\top] = \begin{cases} \boldsymbol{C}\boldsymbol{P}\boldsymbol{C}^\top + \boldsymbol{R} & \text{if } \tau = 0 \\ \boldsymbol{C}\bar{\boldsymbol{A}}^{\tau-1}\boldsymbol{A}\big(\boldsymbol{P}\boldsymbol{C}^\top - \boldsymbol{L}\boldsymbol{P}_\epsilon(0)\big) & \text{if } \tau \neq 0 \end{cases} \tag{4.19}$$

where the CM of the steady-state innovation (measurement prediction error) $\boldsymbol{P} = \mathbb{E}[\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k^\top]$, when $k \to \infty$ is given by the solution of the Lyapunov equation of the form:

$$\boldsymbol{P} = \bar{\boldsymbol{A}}\boldsymbol{P}\bar{\boldsymbol{A}}^\top + \boldsymbol{A}\boldsymbol{L}\boldsymbol{R}\boldsymbol{L}^\top\boldsymbol{A}^\top + \boldsymbol{Q} \tag{4.20}$$

It is worth noting that ACF in Eq. (4.19) is a function of the known matrices $\boldsymbol{A}$, $\boldsymbol{C}$, and $\boldsymbol{L}$ and the unknown matrices $\boldsymbol{Q}$, $\boldsymbol{R}$, and $\boldsymbol{P}_{\epsilon}(\tau), \forall\tau$.

In Algorithm 4.1 it is presented the ICM algorithm for the noise CMs estimation. It is shown that the ICM method is based on several intermediate steps prior the computation of the estimates of the noise CMs, hence the name indirect.

Algorithm 4.1: The indirect correlation method algorithm for noise covariance matrices estimation.

```
Estimation of the ACF:
Design an asymptotically stable linear filter, see Eq. (4.17)
Compute the innovation sequence $\{\boldsymbol{\epsilon}_k\}_{k=0}^{i}$
Estimation of the ACF defined by Eq. (4.19) according to:
    $\hat{\boldsymbol{P}}_{\epsilon}(\tau) = \frac{1}{i-\tau}\sum_{k=\tau}^{i} \boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_{k-\tau}, \quad \tau = 0, 1, \dots, N-1 \qquad (4.21)$
  where $N$ is the number of the computed terms of the ACF in Eq. (4.21)
Noise CMs estimation:
Based on the system of $N$ linear matrix equations, see Eq. (4.19), the noise
  CMs are estimated using three subsequent steps:
Substitute $\hat{\boldsymbol{P}}_{\epsilon}(\tau)$ for $\boldsymbol{P}_{\epsilon}(\tau)$ in Eq. (4.19) for $\tau = 0, 1, \dots, N-1$, solve the
  system of equations with $\boldsymbol{P}_{\epsilon}(0) = \hat{\boldsymbol{P}}_{\epsilon}(0)$ and finally compute the Least
  Square (LS) estimate $\widehat{\boldsymbol{PC}^{\top}}$ of the term $\boldsymbol{PC}^{\top}$
Substitute $\widehat{\boldsymbol{PC}^{\top}}$ for $\boldsymbol{PC}^{\top}$ in Eq. (4.19) and compute the estimate $\hat{\boldsymbol{R}}$ of the
  measurement noise CM $\boldsymbol{R}$
Multiply and post-multiply both sides of Eq. (4.20) by $\boldsymbol{C}$ and $\boldsymbol{C}^{\top}$, respectively.
  Substitute $\widehat{\boldsymbol{PC}^{\top}}$ for $\boldsymbol{PC}^{\top}$. Finally compute the estimate $\hat{\boldsymbol{Q}}$ of the state noise
  CM $\boldsymbol{Q}$ using the LS method
```

The Direct Correlation Method

An upgrade of the ICM is proposed by the method from the work by Odelson [136], which reformulates the problem so that the noise CMs are estimated in a single step. Therefore, the method is referred to as the DCM. In the mentioned work, the DCM was first designed for the Gaussian LTI models.

Similarly to the previous correlation methods, the DCM computes and estimates the CMs of the innovation sequence of a stable linear predictor. The derived innovation sequence CMs are, in principle, the same as those used in the ICM, see Eq. (4.19). They are reformulated as follows:

$$\boldsymbol{P}_{\epsilon}(\tau) \equiv \mathbb{E}[\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_{k-\tau}^{\top}] = \begin{cases} \boldsymbol{CPC}^{\top} + \boldsymbol{R} & \text{if } \tau = 0 \\ \boldsymbol{C}\bar{\boldsymbol{A}}^{\tau}\boldsymbol{APC}^{\top} - \boldsymbol{C}\bar{\boldsymbol{A}}^{\tau-1}\boldsymbol{ALR} & \text{if } \tau \neq 0 \end{cases} \tag{4.22}$$

where $\tau = 0, 1, \dots, N-1$. Again, the steady-state CM of the state prediction error is given by the solution to the following extended Lyapunov equation:

$$\boldsymbol{P} = \bar{\boldsymbol{A}}\boldsymbol{P}\bar{\boldsymbol{A}}^{\mathsf{T}} + \boldsymbol{\Gamma}\begin{bmatrix}\boldsymbol{Q} & \boldsymbol{0}_{n_x \times n_y}\\ \boldsymbol{0}_{n_y \times n_x} & \boldsymbol{R}\end{bmatrix}\boldsymbol{\Gamma}^{\mathsf{T}} \tag{4.23}$$

Using the notation ⊗ for the Kronecker product and $\boldsymbol{P}_s$ for the column-wise stacking of matrix $\boldsymbol{P}$ into a vector [53], the solution of Eq. (4.23) can be explicitly written as follows:

$$\boldsymbol{P}_s = (\boldsymbol{I}_{n_x^2} - \bar{\boldsymbol{A}} \otimes \bar{\boldsymbol{A}})^{-1}(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\begin{bmatrix}\boldsymbol{Q} & \boldsymbol{0}_{n_x \times n_y}\\ \boldsymbol{0}_{n_y \times n_x} & \boldsymbol{R}\end{bmatrix}_s \tag{4.24}$$

Substituting Eq. (4.24) into Eq. (4.22) results in a system of linear equations for the elements of $\boldsymbol{Q}$ and $\boldsymbol{R}$. As already seen for the ICM, the DCM algorithm is presented in Algorithm 4.2.

Algorithm 4.2: The direct correlation method algorithm for noise covariance matrices estimation.

```
  Estimation of the ACF:
1 Design an asymptotically stable linear filter, see Eq. (4.17)
2 Compute the innovation sequence $\{\boldsymbol{\epsilon}_k\}_{k=0}^{i}$
3 Estimation of the ACF defined by Eq. (4.22) according to:

    $$\hat{\boldsymbol{P}}_{\epsilon}(\tau) = \frac{1}{i-\tau}\sum_{k=\tau}^{i}\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_{k-\tau}, \quad \tau = 0, 1, \ldots, N-1 \tag{4.25}$$

    where $N$ is the number of the computed terms of the ACF in Eq. (4.21)
  Noise CMs estimation:
  Based on the system of $N$ linear matrix equations, see Eq. (4.22), the noise
    CMs are estimated directly using a single step:
4 Compute the estimates $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{R}}$ of the state noise CMs using the LS method
    with respect to Eq. (4.24)
```

The DCM in Algorithm 4.2 provides asymptotically unbiased estimates. In the work by Rajamani and Rawlings[144] the estimation of the structural properties of the state noise within the concept of the DCM was discussed. Instead, in the work by Duník et al. [70] the impact of the user-defined parameters, namely, the gain matrix $\boldsymbol{L}$ and the number of equations N, was discussed and a method for their optimal setting was proposed as well. The method can be understood as a compensation for the effect of nonoptimal weighting matrix (typically, the identity matrix) used in the LSM solution of the DCM. The method was originally designed for the LTI models and later was extended for: (i) LTI models with correlated state and measurement noises [36, 117] providing asymptotically unbiased estimates, (ii) linear periodic models providing asymptotically unbiased estimates [157], (iii) linear models with time-correlated noises providing asymptotically unbiased estimates [195, 204], (iv) nonlinear models (based on a linearization of the model) at the cost of losing the property of the asymptotically unbiased estimates [34, 80, 113, 114], and (v) systems with LTI state equation and LTV or nonlinear measurement equation with a possibility to provide asymptotically unbiased estimates [71]. The computational efficiency of the method for high-dimensional models was discussed in the work by Zagrobelny and Rawlings [174]. Theoretical insights into the estimation of the unique elements of the noise CMs can be found in the work by Kost et al. [104]. For a general review and introduction see also the Ph.D. thesis of Rajamani [185].

4.2.5 The direct solution: a one-step approach

In opposition with the standard solution presented in Section 4.2.1, here is now explored the *direct solution* to the filtering design problem. This kind of paradigm, as the name suggests, relies on a direct approach which is based on a single step instead of two as in the standard solution. It also interesting to mention that the terminology is in analogy with the term *Direct Control* coined in the *Identification for Control* (I4C) research field, see [57, 181, 83, 88, 89, 201]. The development of the direct paradigm started when the practical problems of the standard solutions emerged in many applications, which are:

1. Only an approximated model $\hat{\mathcal{S}} \approx \mathcal{S}$ can be identified from the available dataset $\mathcal{D} = \{\boldsymbol{u}_k, \boldsymbol{y}_k, \boldsymbol{z}_k\}$ containing also samples of the variable of interest $\boldsymbol{z}_k$ to be filtered. For this reason, a filter which is optimal designed for the estimated model $\hat{\mathcal{S}}$ may, instead, display a large estimation error when applied to the real system $\mathcal{S}$.

2. In a non-linear system, designing a computationally tractable optimal filter is very difficult and often only approximate filters can be derived, whose stability is not even guaranteed.

Evaluating how these sources of approximation affect the filter estimation accuracy is still an open problem. For these reasons, the alternative approach of *Direct Filtering* (DF) is investigated in 2006, see [208], which is introduced to overcome the highlighted issues. Again, as the name suggests, this methodology makes use of the available dataset $\mathcal{D}$ for the direct design of the filter. As a consequence, the intermediate system identification step, which is a core step in the standard solution, is now skipped entirely. Following this rationale, the data are not used for the estimation of a model of the system but instead for the direct estimation of a model of the filter. The DF approach thus represents a paradigm shift in filter design, allowing the design of optimal filters and overcoming the mentioned problems.

2000s In the early 2000s pioneering contributions and ideas to *Direct Filtering* have been developed within a Set-Membership (SM) framework by Milanese et al. [125, 126, 129]. In these works the additive noises affecting the system are assumed unknown but bounded, and no parametric filter structure is used. Thus, they can be thought of as a *deterministic* solution to develop the DF paradigm. In the late 2000s, other contributions followed: i) a work by Novara for linear time-invariant (LTI) systems [133], ii) a work by Ruiz for linear parameter-varying (LPV) systems [149], and iii) a work by Milanese for non-linear systems [128]. Practical applications, mainly in the automotive field, can be also found in the literature developed in these years in [193, 209, 210].

In particular, it is now presented a reproduction of the simulated example of the work in [208], where the term "direct" was coined for the first time. Also, in the mentioned work it is also made available a founding comparison between the new direct filtering paradigm and the standard solution based on the two-step approach. In the example, it is considered the case of a direct filtering design for the analysis of a non-liner system. In particular, the deterministic non-linear system under analysis is the *Lorenz Attractor*, which consists of a three-dimensional dynamical system derived from the simplified equations of convection rolls arising in the dynamical equations of the atmosphere. For a certain set of parameters the system exhibits chaotic behavior and displays what is called a strange attractor, as displayed for instance by the parametric solution in phase space in Fig. 4.4. The main features of a chaotic system are high sensitivity to initial conditions and unstable solutions. In particular, the discretization in time of the differential equations of the Lorenz system results in the following discrete-time system:

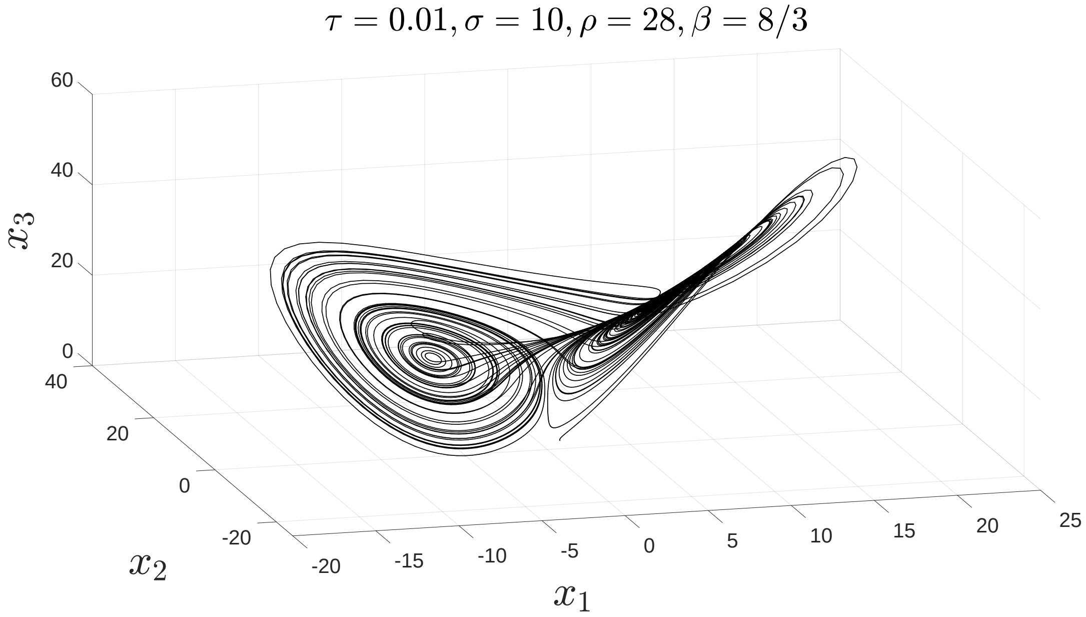


Figure 4.4: The graphical solution displays the chaotic behavior of the Lorenz attractor problem for the particular choice of parameters: $\tau = 0.01, \sigma = 10, \rho = 28, \beta = \frac{8}{3}$.

$$x_{k+1}^1 = (1 - \tau\sigma)x_k^1 + \tau\sigma x_k^2 \quad (4.26a)$$
$$x_{k+1}^2 = (1 - \tau)x_k^2 - \tau x_k^1 x_k^3 + \tau\rho x_k^1 \quad (4.26b)$$
$$x_{k+1}^3 = (1 - \tau\beta)x_k^3 + \tau x_k^1 x_k^2 \quad (4.26c)$$
$$y_k = x_k^1 + v_k \quad (4.26d)$$
$$z_k = x_k^2 x_k^3 + e_k \quad (4.26e)$$

where in this example the notation with the exponent x_k^i stands for the i-th component of the state variable at the k-th time instant, normally encoded in the compacted vector form as $\boldsymbol{x}_k$ in previous examples. The variables τ, σ, ρ, and β are positive parameters of the system. The noises v_k and e_k are i.i.d. Gaussian processes with the following statistics:

$$v_k \sim \mathcal{G}(0, 0.02^2) \tag{4.27}$$

$$e_k \sim \mathcal{G}(0, 20^2) \tag{4.28}$$

Following the reproduction of the example, the system in Eq. (4.26) has been sampled to collect a dataset consisting of 8000 samples. In particular, the key idea for the direct filtering paradigm is found in Eq. (4.26e), that shows that a dedicated equation is available to collect some samples of the desired variable, here labeled as z_k, to be estimated. The desired variable equation reflects the idea that for the designer of the filter, the variable can be measured for a limited amount of time. This is the case, for instance, of real world applications when industrial prototyping enable this possibility. On the other hand, after the filter is designed and ready to be used, the desired variable equation is no longer available, and the idea is that the filter must work without that knowledge. In other words, the information encoded in the desired variable equation is exploited ex-ante, only for the a-priori design of the filter. Instead, the designed direct filters works with just the knowledge of input-output data collected from the operation of the dynamical system.

To this end, returning to the example, the dataset simulated from the example has been partitioned in two sets:

$$\mathcal{D}_{\text{ID}} = \left\{ (u_k, y_k, z_k), k = 1, \dots, N_{\text{ID}} \right\} \tag{4.29}$$

$$\mathcal{D}_{\text{VL}} = \left\{ (u_k, y_k), k = N_{\text{ID}+1}, \dots, N_{\text{VL}} \right\} \tag{4.30}$$

with $N_{\text{ID}} = 6000$ and $N_{\text{VL}} = 8000$.

Then, in a parameterized stochastic setting, the dataset formed for the design of the filter, labeled as $\mathcal{D}_{\text{ID}}$, is used for the identification of the parameters of the following filter:

$$\hat{z}_k = f_{\text{DF}}(\boldsymbol{\theta}_{\text{DF}}, \boldsymbol{Y}_k, \boldsymbol{U}_k, \boldsymbol{Z}_k) \tag{4.31}$$

where $\boldsymbol{Y}_k$ and $\boldsymbol{Y}_k$ in general are formed from **Lem. 4.1** whose proof is also contained in [208]. The idea of this lemma is to have the existence of a desirable function that depends on available input and output data collected from the dynamical system. In other words it proves the existence of a function that can be used for the direct filtering paradigm.

Lemma 4.1. *Consider the general system in Eq. (4.1). If (f, h_y) is observable, then $\exists f_0$ and integers $n_1, n_2 \le n_x$ such that:*

$$z_k = f_0(\boldsymbol{Y}_k, \boldsymbol{U}_k) \tag{4.32}$$

$$\boldsymbol{Y}_k = [\boldsymbol{y}_k, \boldsymbol{y}_{k-1}, \cdots, \boldsymbol{y}_{k-n_1+1}] \tag{4.33}$$

$$\boldsymbol{U}_k = [\boldsymbol{u}_k, \boldsymbol{u}_{k-1}, \cdots, \boldsymbol{u}_{k-n_2+1}] \tag{4.34}$$

In particular, the regression function in Eq. (4.31) used for the filter structure is a neural network categorized as having a one hidden layer with r neurons (the simplest feedforward neural network):

$$f_{\mathrm{DF}}(\boldsymbol{\theta}_{\mathrm{DF}}, Y_k, U_k, Z_k) = \sum_{i=1}^{r} \alpha_i \sigma \left[\beta_i (Y_k, U_k, Z_k) - \lambda_i \right] + \zeta \tag{4.35}$$

where $\boldsymbol{\theta}_{\mathrm{DF}} = \{\alpha_i, \lambda_i, \zeta, \beta_i, i = 1, \cdots, r\}$ is the set of parameters and $\sigma(x)$ is the sigmoidal function used as the activation function of the neural network:

$$\sigma(x) = \frac{2}{1 + e^{-2x}} \tag{4.36}$$

To conclude the design of the filter, several neural networks of the form in Eq. (4.35) with different number of neurons (from $r = 2$ to $r = 15$) have been trained on the identification set $\mathcal{D}_{\mathrm{ID}}$. In the end, the neural network showing the lowest estimation error on the validation dataset $\mathcal{D}_{\mathrm{VL}}$ in a MSE (*Mean Square Error*) sense has been chosen. Results are displayed in Fig. 4.5 where it is possible to compare the filtered states using the available input and output data for $k = N_{\mathrm{ID}} + 1, \cdots, N_{\mathrm{VL}}$ with the actual values simulated from the system. In other words, after the identified filter has been applied to the validation set $\mathcal{D}_{\mathrm{VL}}$, the filter performance turned out to be quite satisfactory. Indeed, the resulting *Root Mean Square Estimation Error* (RMSEE) is 24, which is close to the standard deviation of the noise affecting the desirable variable z_k.

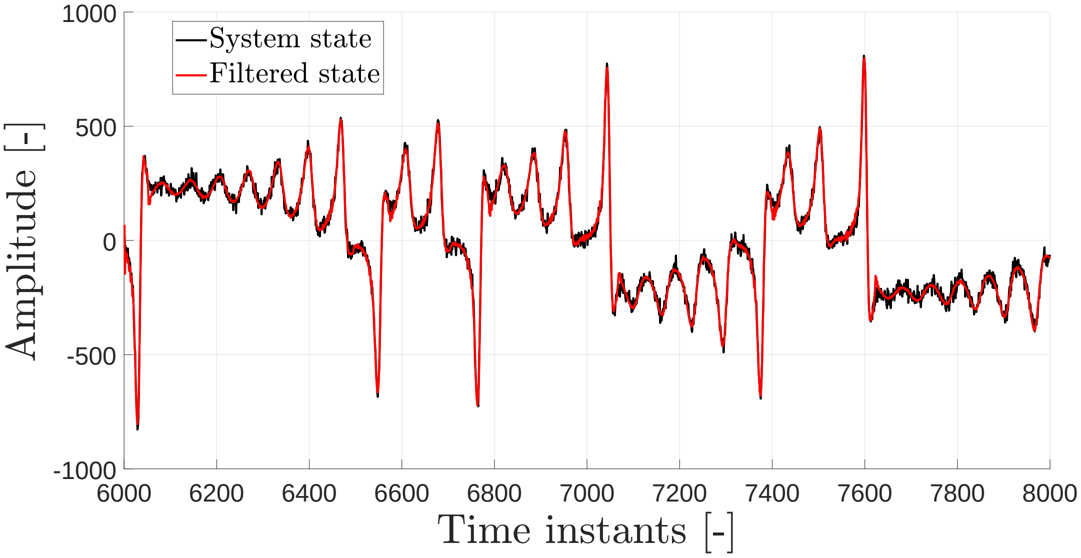


Figure 4.5: A comparison of the Lorenz attractor solution on the validation set $\mathcal{D}_{\mathrm{VL}}$: in black the true data z_k simulated from the system in Eq. (4.26) and in red the estimated data $\hat{z}_k$ after applying the direct filtering idea.

An example related to the Lorenz attractor is presented to demonstrate the effectiveness of the presented approach.

To conclude, the results of the reproduction of the example in [208] shows that a direct filtering paradigm is indeed interesting and demonstrates its effectiveness. Moreover, the work contributes also with theoretical insights showing that the two-step procedure found in the classical solution is proved to perform, in the case of exact modeling, no better than the direct approach. In the presence of modeling errors, the directly identified filter is proved to be anyway the minimum variance estimator, among the selected approximating filter class. A similar result is not assured by the two-step design, whose performance deterioration due to modeling errors may be significantly larger. Another relevant point is that minimum variance filters for nonlinear systems are in general difficult to derive and/or to implement, and widely used approximate solutions quite often exhibit poor performance. On the contrary, the recent progresses in nonlinear identification methods may allow the direct filter identification.

Further work in a more structured *parametric-stochastic* setting has been introduced also in the early 2010s by Novara et al. in [135], and it is one of the focus of the present dissertation to be further investigated. To this end, Chapter 6 will be dedicated to formulate the parametric-stochastic setting for the direct filtering idea for steady-state filters derived from LTI systems. The aim of the contribution is to propose a direct solution to the filtering design problem where both an appropriate model structure and a model order complexity for the parametric estimation of the filter is derived. Then, a fair comparison of filtering performance between the direct and classical solutions will be analyzed in different experimental settings in Chapter 7. The results of the comparison aims to validate whether the direct solution gives better or no-worse filtering performance than the classical solution when critical practical problems are taken into account, i.e. estimation of noise covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$, as highlighted in the classical solution.

CHAPTER 5

The standard solution to the filtering design from data problem

As already seen in Chapter 4, the derivation of data-driven solutions, namely the standard one and the direct one, to the filtering design problem is carried out considering the following general description of a discrete-time LTI system $\mathcal{S}$:

$$\mathcal{S}: \quad \begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{G}\boldsymbol{w}_k && (5.1a)\\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{D}\boldsymbol{u}_k + \boldsymbol{v}_k && (5.1b)\\ \boldsymbol{z}_k &= \boldsymbol{C}_2\boldsymbol{x}_k + \boldsymbol{e}_k && (5.1c) \end{aligned}$$

where the new Eq. (5.1c) describes the possibility to measure the desired variable $\boldsymbol{z}_k$ to be filtered, for a limited amount of time, i.e. for $k = 1, \dots, N_{\text{ID}}$. In Eq. (5.1c), $\boldsymbol{e}_k$ denotes the measurement noise affecting the equation.

For the derivation of filtering solutions, the general framework in Eq. (5.1) is further simplified as follows:

$$\mathcal{S}: \quad \begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{w}_k && (5.2a)\\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k && (5.2b)\\ \boldsymbol{z}_k &= \boldsymbol{x}_k + \boldsymbol{e}_k && (5.2c) \end{aligned}$$

by considering only the stochastic component of the sub-system, i.e. $\boldsymbol{B} = \boldsymbol{D} = \boldsymbol{0}$, the process noise shaping matrix is the identity matrix for the sake of simplicity, i.e. $\boldsymbol{G} = \boldsymbol{I}$, and the measurement matrix of the desired variable is the identity matrix, i.e. $\boldsymbol{C}_2 = \boldsymbol{I}$.

Later, the assumption that $\boldsymbol{B} = \boldsymbol{0}$ is relaxed in the presence of a deterministic exogenous input, i.e. $\boldsymbol{B} \neq \boldsymbol{0}$, in Chapter 7 when performing experimental tests.

Considering the the standard solution to the filtering design from data problem, in general, assuming that $\boldsymbol{B} \neq \boldsymbol{0}$ and $\boldsymbol{C}_2 \neq \boldsymbol{I}$ requires to identify a model for those unknown matrices by exploiting the available measurements. For instance, considering the case of estimating the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$, a subspace system identification method could be applied. These case studies are not the main focus of the research to be presented in this dissertation and have therefore not been explored further.

However, it remains interesting to evaluate the potential of these studies in future research projects dealing with same filtering context.

Nonetheless, as seen in Section 4.2, the standard solution to the filtering design from data problem is derived considering a two-step approach:

1. First, a data-driven system identification step is performed.
2. Then, a filter design step is performed on the identified system.

Following this scheme, the chapter is organized as follow:

1. Section 5.1 introduces sequential steps that need to be performed in order to estimate the deterministic components of the system, matrice $\boldsymbol{A}$ and $\boldsymbol{C}$. In other words, the undertaken system identification routine is introduced.
2. Section 5.2 introduces sequential steps that need to be performed in order to estimate the stochastic components of the system, matrice $\boldsymbol{Q}$ and $\boldsymbol{R}$, required for the design of the filter. In other words, the undertaken filter design routine is introduced.

5.1 The system identification step

Considering the system identification step, instead of using classical techniques as described in the brief historical review in Section 4.2.2, in this section the new Eq. (5.1c) is exploited in order to estimate the deterministic component of the system $\mathcal{S}$. This idea is important since the new available equation encodes new knowledge that could not be used in the classical solutions but can be exploited now. This is also in accordance with the fact that the same knowledge will be used in the development of the direct solution methodology in Chapter 6, thus enabling a fair comparison of framework and (available and used) "ingredients".

In the next sections, it will be shown how the standard solution makes use of the available data of the system in order to estimate matrices $\boldsymbol{A}$ and $\boldsymbol{C}$.

5.1.1 Estimation of the state matrix

Since system $\mathcal{S}$ is LTI and regressor data $\boldsymbol{z}_{1:N_{\text{ID}}}$ are available, the state matrix $\boldsymbol{A}$ can be estimated considering Eq. (5.1c) substituted in (5.1a) giving the following regression model:

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{w}_k \tag{5.3a}$$

$$\boldsymbol{z}_{k+1} - \boldsymbol{e}_{k+1} = \boldsymbol{A}(\boldsymbol{z}_k - \boldsymbol{e}_k) + \boldsymbol{w}_k \tag{5.3b}$$

$$\boldsymbol{z}_{k+1} = \boldsymbol{A}\boldsymbol{z}_k + \boldsymbol{e}_{k+1} - \boldsymbol{A}\boldsymbol{e}_k + \boldsymbol{w}_k \tag{5.3c}$$

$$\boldsymbol{z}_{k+1} = \boldsymbol{A}\boldsymbol{z}_k + \boldsymbol{\eta}_k \tag{5.3d}$$

where $\boldsymbol{\eta}_k \equiv \boldsymbol{e}_{k+1} - \boldsymbol{A}\boldsymbol{e}_k + \boldsymbol{w}_k$.

Statistics of the noise $\boldsymbol{\eta}_k$ can be computed as follows:

$$\boldsymbol{\mu}_\eta \equiv \mathbb{E}[\boldsymbol{\eta}_k] = \mathbb{E}[\boldsymbol{e}_{k+1} - \boldsymbol{A}\boldsymbol{e}_k + \boldsymbol{w}_k] = \cancel{\mathbb{E}[\boldsymbol{e}_{k+1}]} - \boldsymbol{A}\cancel{\mathbb{E}[\boldsymbol{e}_k]} + \cancel{\mathbb{E}[\boldsymbol{w}_k]} = \boldsymbol{0}, \quad \forall k \tag{5.4}$$

$$
\begin{align}
\boldsymbol{\Sigma}_{\boldsymbol{\eta}} &\equiv \mathbb{E}[(\boldsymbol{\eta}_k - \boldsymbol{\mu}_{\boldsymbol{\eta}})(\boldsymbol{\eta}_k - \boldsymbol{\mu}_{\boldsymbol{\eta}})^\top] = \mathbb{E}[\boldsymbol{\eta}_k\boldsymbol{\eta}_k^\top] \tag{5.5a}\\
&= \mathbb{E}[(\boldsymbol{e}_{k+1} - \boldsymbol{A}\boldsymbol{e}_k + \boldsymbol{w}_k)(\boldsymbol{e}_{k+1} - \boldsymbol{A}\boldsymbol{e}_k + \boldsymbol{w}_k)^\top] \tag{5.5b}\\
&= \mathbb{E}[\boldsymbol{e}_{k+1}\boldsymbol{e}_{k+1}^\top] - \cancel{\mathbb{E}[\boldsymbol{e}_{k+1}\boldsymbol{e}_k^\top]}\boldsymbol{A}^\top + \cancel{\mathbb{E}[\boldsymbol{e}_{k+1}\boldsymbol{w}_k^\top]} - \boldsymbol{A}\cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_{k+1}^\top]} + \boldsymbol{A}\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_k^\top]\boldsymbol{A}^\top \nonumber\\
&\quad - \boldsymbol{A}\cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{w}_k^\top]} + \cancel{\mathbb{E}[\boldsymbol{w}_k\boldsymbol{e}_{k+1}^\top]} - \cancel{\mathbb{E}[\boldsymbol{w}_k\boldsymbol{e}_k^\top]}\boldsymbol{A}^\top + \mathbb{E}[\boldsymbol{w}_k\boldsymbol{w}_k^\top] \tag{5.5c}\\
&= \boldsymbol{\Sigma}_{\boldsymbol{e}} + \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{e}}\boldsymbol{A}^\top + \boldsymbol{Q}, \quad \forall k \tag{5.5d}
\end{align}
$$

By computing the cross-covariance function $\boldsymbol{\Sigma}_{\boldsymbol{z}\boldsymbol{\eta}}$ it turns out that the regressor $\boldsymbol{z}_k$ is correlated with the noise $\boldsymbol{\eta}_k$:

$$
\begin{align}
\boldsymbol{\Sigma}_{\boldsymbol{z}\boldsymbol{\eta}} &= \mathbb{E}[(\boldsymbol{z}_k - \boldsymbol{\mu}_{\boldsymbol{z}})(\boldsymbol{\eta}_k - \boldsymbol{\mu}_{\boldsymbol{\eta}})^\top] = \mathbb{E}[\boldsymbol{z}_k\boldsymbol{\eta}_k^\top] \tag{5.6a}\\
&= \mathbb{E}[(\boldsymbol{x}_k + \boldsymbol{e}_k)(\boldsymbol{e}_{k+1} - \boldsymbol{A}\boldsymbol{e}_k + \boldsymbol{w}_k)^\top] \tag{5.6b}\\
&= \mathbb{E}[\boldsymbol{x}_k\boldsymbol{e}_{k+1}^\top - \boldsymbol{x}_k\boldsymbol{e}_k^\top\boldsymbol{A}^\top + \boldsymbol{x}_k\boldsymbol{w}_k^\top + \boldsymbol{e}_k\boldsymbol{e}_{k+1}^\top - \boldsymbol{e}_k\boldsymbol{e}_k^\top\boldsymbol{A}^\top + \boldsymbol{e}_k\boldsymbol{w}_k^\top] \tag{5.6c}\\
&= \cancel{\mathbb{E}[\boldsymbol{x}_k\boldsymbol{e}_{k+1}^\top]} - \cancel{\mathbb{E}[\boldsymbol{x}_k\boldsymbol{e}_k^\top]}\boldsymbol{A}^\top + \cancel{\mathbb{E}[\boldsymbol{x}_k\boldsymbol{w}_k^\top]} + \cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_{k+1}^\top]} - \mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_k^\top]\boldsymbol{A}^\top + \cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{w}_k^\top]} \tag{5.6d}\\
&= -\boldsymbol{\Sigma}_{\boldsymbol{e}}\boldsymbol{A}^\top \tag{5.6e}
\end{align}
$$

$$
\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{z}} \equiv \boldsymbol{\Sigma}_{\boldsymbol{z}\boldsymbol{\eta}}^\top = -(\boldsymbol{A}^\top)^\top\boldsymbol{\Sigma}_{\boldsymbol{e}}^\top = -\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{e}} \tag{5.7}
$$

Finally, the system matrix estimate $\hat{\boldsymbol{A}}_{\text{LS}}$ is:

$$
\begin{align}
\hat{\boldsymbol{A}}_{\text{LS}} &= \boldsymbol{z}_{k+1}\boldsymbol{z}_k^\top(\boldsymbol{z}_k\boldsymbol{z}_k^\top)^{-1} \tag{5.8a}\\
&= (\boldsymbol{A}\boldsymbol{z}_k + \boldsymbol{\eta}_k)\boldsymbol{z}_k^\top(\boldsymbol{z}_k\boldsymbol{z}_k^\top)^{-1} \tag{5.8b}\\
&= \boldsymbol{A}\boldsymbol{z}_k\boldsymbol{z}_k^\top(\boldsymbol{z}_k\boldsymbol{z}_k^\top)^{-1} + \boldsymbol{\eta}_k\boldsymbol{z}_k^\top(\boldsymbol{z}_k\boldsymbol{z}_k^\top)^{-1} \tag{5.8c}\\
&= \boldsymbol{A} + \boldsymbol{\eta}_k\boldsymbol{z}_k^\top(\boldsymbol{z}_k\boldsymbol{z}_k^\top)^{-1} \tag{5.8d}
\end{align}
$$

where estimate $\hat{\boldsymbol{A}}_{\text{LS}}$ is biased due to regressor $\boldsymbol{z}_k$ being endogenous, i.e. $\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{z}} = -\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{e}} \neq \boldsymbol{0}$, see Eq. (5.7).

In order to compute an unbiased estimate, an instrumental variable approach might be considered. To this end, the following instrumental variable is considered:

$$
\boldsymbol{\phi}_k = \boldsymbol{z}_{k-1} \tag{5.9}
$$

By computing the cross-covariance function $\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\phi}}$ it turns out that the instrument $\boldsymbol{\phi}_k$ is uncorrelated with the noise $\boldsymbol{\eta}_k$:

$$
\begin{align}
\boldsymbol{\Sigma}_{\boldsymbol{\eta}\boldsymbol{\phi}} &= \mathbb{E}[(\boldsymbol{\eta}_k - \boldsymbol{\mu}_{\boldsymbol{\eta}})(\boldsymbol{\phi}_k - \boldsymbol{\mu}_{\boldsymbol{\phi}})^\top] = \mathbb{E}[\boldsymbol{\eta}_k\boldsymbol{\phi}_k^\top] \tag{5.10a}\\
&= \mathbb{E}[(\boldsymbol{e}_{k+1} - \boldsymbol{A}\boldsymbol{e}_k + \boldsymbol{w}_k)(\boldsymbol{x}_{k-1} + \boldsymbol{e}_{k-1})^\top] \tag{5.10b}\\
&= \mathbb{E}[\boldsymbol{e}_{k+1}\boldsymbol{x}_{k-1}^\top + \boldsymbol{e}_{k+1}\boldsymbol{e}_{k-1}^\top - \boldsymbol{A}\boldsymbol{e}_k\boldsymbol{x}_{k-1}^\top - \boldsymbol{A}\boldsymbol{e}_k\boldsymbol{e}_{k-1}^\top + \boldsymbol{w}_k\boldsymbol{x}_{k-1}^\top + \boldsymbol{w}_k\boldsymbol{e}_{k-1}^\top] \tag{5.10c}\\
&= \cancel{\mathbb{E}[\boldsymbol{e}_{k+1}\boldsymbol{x}_{k-1}^\top]} + \cancel{\mathbb{E}[\boldsymbol{e}_{k+1}\boldsymbol{e}_{k-1}^\top]} - \boldsymbol{A}\cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{x}_{k-1}^\top]} - \boldsymbol{A}\cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_{k-1}^\top]} \nonumber\\
&\quad + \cancel{\mathbb{E}[\boldsymbol{w}_k\boldsymbol{x}_{k-1}^\top]} + \cancel{\mathbb{E}[\boldsymbol{w}_k\boldsymbol{e}_{k-1}^\top]} \tag{5.10d}\\
&= \boldsymbol{0} \tag{5.10e}
\end{align}
$$

Then, the unbiased estimate $\hat{A}_{\mathrm{IV}}$ is given by:

$$\begin{aligned}
\hat{A}_{\mathrm{IV}} &= z_{k+1}\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1} \qquad (5.11a)\\
&= (\boldsymbol{A}z_k + \boldsymbol{\eta}_k)\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1} \qquad (5.11b)\\
&= \boldsymbol{A}z_k\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1} + \boldsymbol{\eta}_k\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1} \qquad (5.11c)\\
&= \boldsymbol{A} + \cancel{\boldsymbol{\eta}_k\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1}} \qquad (5.11d)
\end{aligned}$$

where now $\boldsymbol{\phi}_k$ is exogenous, i.e. $\Sigma_{\eta\phi} = \mathbf{0}$.

5.1.2 Estimation of the output matrix

Similar reasoning can be applied to the estimation of the output matrix $\boldsymbol{C}$, where the regression model is given by Eq. (5.1c) substituted in (5.1b):

$$\begin{aligned}
\boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k \qquad (5.12a)\\
\boldsymbol{y}_k &= \boldsymbol{C}(z_k - \boldsymbol{e}_k) + \boldsymbol{v}_k \qquad (5.12b)\\
\boldsymbol{y}_k &= \boldsymbol{C}z_k + \boldsymbol{\kappa}_k, \quad \boldsymbol{\kappa}_k \sim \mathcal{G}\left(\boldsymbol{\mu}_{\boldsymbol{\kappa}}, \Sigma_{\boldsymbol{\kappa}}\right) \qquad (5.12c)
\end{aligned}$$

where $\boldsymbol{\kappa}_k \equiv \boldsymbol{v}_k - \boldsymbol{C}\boldsymbol{e}_k$.

Statistics of the noise $\boldsymbol{\kappa}_k$ can be computed as follows:

$$\boldsymbol{\mu}_{\boldsymbol{\kappa}} \equiv \mathbb{E}[\boldsymbol{\kappa}_k] = \mathbb{E}[\boldsymbol{v}_k - \boldsymbol{C}\boldsymbol{e}_k] = \cancel{\mathbb{E}[\boldsymbol{v}_k]} - \boldsymbol{C}\cancel{\mathbb{E}[\boldsymbol{e}_k]} = \mathbf{0}, \quad \forall k \qquad (5.13)$$

$$\begin{aligned}
\Sigma_{\boldsymbol{\kappa}} &\equiv \mathbb{E}[(\boldsymbol{\kappa}_k - \boldsymbol{\mu}_{\boldsymbol{\kappa}})(\boldsymbol{\kappa}_k - \boldsymbol{\mu}_{\boldsymbol{\kappa}})^{\mathsf{T}}] = \mathbb{E}[\boldsymbol{\kappa}_k\boldsymbol{\kappa}_k^{\mathsf{T}}] \qquad (5.14a)\\
&= \mathbb{E}[(\boldsymbol{v}_k - \boldsymbol{C}\boldsymbol{e}_k)(\boldsymbol{v}_k - \boldsymbol{C}\boldsymbol{e}_k)^{\mathsf{T}}] \qquad (5.14b)\\
&= \mathbb{E}[\boldsymbol{v}_k\boldsymbol{v}_k^{\mathsf{T}}] - \boldsymbol{C}\cancel{\mathbb{E}[\boldsymbol{v}_k\boldsymbol{e}_k^{\mathsf{T}}]} - \boldsymbol{C}\cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{v}_k^{\mathsf{T}}]} + \boldsymbol{C}\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_k^{\mathsf{T}}]\boldsymbol{C}^{\mathsf{T}} \qquad (5.14c)\\
&= \boldsymbol{C}\Sigma_{\boldsymbol{e}}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{R}, \quad \forall k \qquad (5.14d)
\end{aligned}$$

By computing the cross-covariance function $\Sigma_{\boldsymbol{\kappa}z}$ it turns out that the regressor z_k is correlated with the noise $\boldsymbol{\kappa}_k$:

$$\begin{aligned}
\Sigma_{\boldsymbol{\kappa}z} &= \mathbb{E}\left[(\boldsymbol{\kappa}_k - \boldsymbol{\mu}_{\boldsymbol{\kappa}})(z_k - \boldsymbol{\mu}_z)^{\mathsf{T}}\right] = \mathbb{E}[\boldsymbol{\kappa}_k z_k^{\mathsf{T}}] \qquad (5.15a)\\
&= \mathbb{E}\left[(\boldsymbol{v}_k - \boldsymbol{C}\boldsymbol{e}_k)(\boldsymbol{x}_k + \boldsymbol{e}_k)^{\mathsf{T}}\right] \qquad (5.15b)\\
&= \mathbb{E}[\boldsymbol{v}_k\boldsymbol{x}_k^{\mathsf{T}} + \boldsymbol{v}_k\boldsymbol{e}_k^{\mathsf{T}} - \boldsymbol{C}\boldsymbol{e}_k\boldsymbol{x}_k^{\mathsf{T}} - \boldsymbol{C}\boldsymbol{e}_k\boldsymbol{e}_k^{\mathsf{T}}] \qquad (5.15c)\\
&= \cancel{\mathbb{E}[\boldsymbol{v}_k\boldsymbol{x}_k^{\mathsf{T}}]} + \cancel{\mathbb{E}[\boldsymbol{v}_k\boldsymbol{e}_k^{\mathsf{T}}]} - \boldsymbol{C}\cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{x}_k^{\mathsf{T}}]} - \boldsymbol{C}\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_k^{\mathsf{T}}] \qquad (5.15d)\\
&= -\boldsymbol{C}\Sigma_{\boldsymbol{e}}, \quad \forall k \qquad (5.15e)
\end{aligned}$$

Finally, the system matrix estimate $\hat{C}_{\mathrm{LS}}$ is:

$$\begin{aligned}
\hat{C}_{\mathrm{LS}} &= \boldsymbol{y}_k z_k^{\mathsf{T}}(z_k z_k^{\mathsf{T}})^{-1} \qquad (5.16a)\\
&= (\boldsymbol{C}z_k + \boldsymbol{\kappa}_k)z_k^{\mathsf{T}}(z_k z_k^{\mathsf{T}})^{-1} \qquad (5.16b)\\
&= \boldsymbol{C}z_k z_k^{\mathsf{T}}(z_k z_k^{\mathsf{T}})^{-1} + \boldsymbol{\kappa}_k z_k^{\mathsf{T}}(z_k z_k^{\mathsf{T}})^{-1} \qquad (5.16c)\\
&= \boldsymbol{C} + \boldsymbol{\kappa}_k z_k^{\mathsf{T}}(z_k z_k^{\mathsf{T}})^{-1} \qquad (5.16d)
\end{aligned}$$

where estimate $\hat{C}_{\mathrm{LS}}$ is biased due to regressor z_k being endogenous, i.e. $\varSigma_{\boldsymbol{\kappa} z} = -C\varSigma_{\boldsymbol{e}} \neq \boldsymbol{0}$, see Eq. (5.15a).

In order to compute an unbiased estimate, an instrumental variable approach might be considered. To this end, the following instrumental variable is considered:

$$\boldsymbol{\phi}_k = z_{k-1} \tag{5.17}$$

By computing the cross-covariance function $\varSigma_{\boldsymbol{\kappa\phi}}$ it turns out that the instrument $\boldsymbol{\phi}_k$ is uncorrelated with the noise $\boldsymbol{\kappa}_k$:

$$\begin{aligned}
\varSigma_{\boldsymbol{\kappa\phi}} &= \mathbb{E}[(\boldsymbol{\kappa}_k - \boldsymbol{\mu}_{\boldsymbol{\kappa}})(\boldsymbol{\phi}_k - \boldsymbol{\mu}_{\boldsymbol{\phi}})^{\mathsf{T}}] = \mathbb{E}[\boldsymbol{\kappa}_k\boldsymbol{\phi}_k^{\mathsf{T}}] && \text{(5.18a)}\\
&= \mathbb{E}\left[(\boldsymbol{v}_k - C\boldsymbol{e}_k)(\boldsymbol{x}_{k-1} + \boldsymbol{e}_{k-1})^{\mathsf{T}}\right] && \text{(5.18b)}\\
&= \mathbb{E}[\boldsymbol{v}_k\boldsymbol{x}_{k-1}^{\mathsf{T}} + \boldsymbol{v}_k\boldsymbol{e}_{k-1}^{\mathsf{T}} - C\boldsymbol{e}_k\boldsymbol{x}_{k-1}^{\mathsf{T}} - C\boldsymbol{e}_k\boldsymbol{e}_{k-1}^{\mathsf{T}}] && \text{(5.18c)}\\
&= \cancel{\mathbb{E}[\boldsymbol{v}_k\boldsymbol{x}_{k-1}^{\mathsf{T}}]} + \cancel{\mathbb{E}[\boldsymbol{v}_k\boldsymbol{e}_{k-1}^{\mathsf{T}}]} - C\cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{x}_{k-1}^{\mathsf{T}}]} - C\cancel{\mathbb{E}[\boldsymbol{e}_k\boldsymbol{e}_{k-1}^{\mathsf{T}}]} && \text{(5.18d)}\\
&= \boldsymbol{0}, \quad \forall k && \text{(5.18e)}
\end{aligned}$$

Then, the unbiased estimate $\hat{C}_{\mathrm{IV}}$ is given by:

$$\begin{aligned}
\hat{C}_{\mathrm{IV}} &= y_k\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1} && \text{(5.19a)}\\
&= (Cz_k + \boldsymbol{\kappa}_k)\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1} && \text{(5.19b)}\\
&= Cz_k\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1} + \boldsymbol{\kappa}_k\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1} && \text{(5.19c)}\\
&= C + \cancel{\boldsymbol{\kappa}_k\boldsymbol{\phi}_k^{\mathsf{T}}(z_k\boldsymbol{\phi}_k^{\mathsf{T}})^{-1}} && \text{(5.19d)}
\end{aligned}$$

where now $\boldsymbol{\phi}_k$ is exogenous, i.e. $\varSigma_{\boldsymbol{\kappa\phi}} = \boldsymbol{0}$.

5.1.3 Last notes about the system identification step

Note that, in the remaining chapter it is considered that:

$$A = \hat{A}_{\mathrm{IV}}, \quad C = \hat{C}_{\mathrm{IV}} \tag{5.20}$$

In the case of using other identification techniques, for instance subspace methods, the resulting estimates are used as if they were the real matrices of the system:

$$A = \hat{A}_{\mathrm{SS}}, \quad B = \hat{B}_{\mathrm{SS}}, \quad C = \hat{C}_{\mathrm{SS}}, \quad D = \hat{D}_{\mathrm{SS}} \tag{5.21}$$

5.2 The filter design step

Once the system is identified it is possible to design the optimal filter by using the KF theory. Note that in this section Eq. (5.1c) is not used since the desired variable can be measured only for a limited amount of time. The solution to the KF consists of an iterative prediction-correction process. In particular, in the prediction step, the time-update maps the state one-step ahead:

$$\hat{\boldsymbol{x}}_k^- = A\hat{\boldsymbol{x}}_{k-1} + B\boldsymbol{u}_{k-1} \tag{5.22}$$

$$P_k^- = AP_{k-1}A^{\mathsf{T}} + Q \tag{5.23}$$

Note that, in this step, it is also possible to predict the output from Eq. (5.1b) as $\hat{\boldsymbol{y}}_k^- = \boldsymbol{C}\hat{\boldsymbol{x}}_k^-$ and to compute the time-variant KF gain $\boldsymbol{K}_k$ as follows:

$$\boldsymbol{K}_k = \boldsymbol{P}_k^- \boldsymbol{C}^\mathsf{T}(\boldsymbol{C}\boldsymbol{P}_k^- \boldsymbol{C}^\mathsf{T} + \boldsymbol{R}) \tag{5.24}$$

Instead, in the correction step, the previous prediction is updated with the new available measurement $\boldsymbol{y}_k$:

$$\hat{\boldsymbol{x}}_k = \hat{\boldsymbol{x}}_k^- + \boldsymbol{K}_k \boldsymbol{\epsilon}_k^- \tag{5.25}$$

$$\boldsymbol{P}_k = (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{C})\boldsymbol{P}_k^- \tag{5.26}$$

where the (output) innovation is defined as the one-step ahead prediction error $\boldsymbol{\epsilon}_k^- \equiv \boldsymbol{y}_k - \hat{\boldsymbol{y}}_k^-$.

Steady-state considerations and the offline design of the filter

If steady-state is considered, the KF gain $\boldsymbol{K}_\infty$ can be computed off-line [29] from Eq. (5.24) by solving the DARE:

$$\boldsymbol{P}_\infty = \boldsymbol{A}\boldsymbol{P}_\infty \boldsymbol{A}^\mathsf{T} + \boldsymbol{Q} - \boldsymbol{A}\boldsymbol{P}_\infty \boldsymbol{C}^\mathsf{T}(\boldsymbol{C}\boldsymbol{P}_\infty \boldsymbol{C}^\mathsf{T} + \boldsymbol{R})^{-1}\boldsymbol{C}\boldsymbol{P}_\infty \boldsymbol{A}^\mathsf{T} \tag{5.27}$$

5.2.1 Estimation of the noise covariance matrices

Note that the derivation of previous KF equations require also complete knowledge of statistics of the noises affecting the system, i.e. the noise *Covariance Matrices* (CMs) $\boldsymbol{Q}$ and $\boldsymbol{R}$. The assumption is, however, questionable in many cases and an incorrect description can cause worsening of estimation quality. Therefore, since the 1970s research interest has been focused on identifying noise CMs [47, 123, 207] through different identification methods, namely Bayesian methods, covariance matching methods, maximum likelihood methods, and correlation methods [137]. Among them, the class of the correlation methods is the most studied as it can be derived analytically with minimal assumptions on the model and provides consistent and unbiased estimates, as reviewed in [72]. The *Autocovariance Least-Square* (ALS) method adhere summarized is taken from [70] and is just one of the existing formulations [36, 137, 144]. All of these formulations refer to the *Direct Correlation Method* class, as seen previously in Section 4.2.4. In particular, these methods are based on an analysis of the second-order statistics of the state (one-step ahead) prediction error $\boldsymbol{\delta}_k \equiv \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k^-$ produced by the linear state predictor derived from Eq. (5.25) substituted in Eq. (5.22):

$$\hat{\boldsymbol{x}}_{k+1}^- = \boldsymbol{A}(\hat{\boldsymbol{x}}_k^- + \boldsymbol{L}\boldsymbol{\epsilon}_k^-) \tag{5.28}$$

where the user-defined parameter $\boldsymbol{L} \neq \boldsymbol{K}_\infty$ is a steady-state, stabilizing, and not optimal filter gain and, without loss of generality, the deterministic sub-system is not considered, i.e. $\boldsymbol{B} = \boldsymbol{0}$.

The evolution $\boldsymbol{\delta}_{k+1}$ is written by subtracting recursive Eq. (5.28) from Eq. (5.1a) and substituting Eq. (5.1b):

$$\boldsymbol{\delta}_{k+1} = (\boldsymbol{A} - \boldsymbol{A}\boldsymbol{L}\boldsymbol{C})\boldsymbol{\delta}_k + [\boldsymbol{I}_{n_x}, -\boldsymbol{A}\boldsymbol{C}][\boldsymbol{w}_k^\mathsf{T}, \boldsymbol{v}_k^\mathsf{T}]^\mathsf{T} \tag{5.29}$$

$$= \bar{\boldsymbol{A}}\boldsymbol{\delta}_k + \bar{\boldsymbol{G}}\bar{\boldsymbol{v}}_k \tag{5.30}$$

where matrices $\bar{A}$ and $\bar{G}$ are defined as follows:

$$\bar{\boldsymbol{A}} \equiv \boldsymbol{A} - \boldsymbol{ALC} \tag{5.31}$$

$$\bar{\boldsymbol{G}} \equiv [\boldsymbol{I}_{n_x}, -\boldsymbol{AC}] \tag{5.32}$$

Considering the steady-state covariance matrix of the state innovation described by the Lyapunov equation:

$$\boldsymbol{P}_{\delta} = \bar{\boldsymbol{A}}\boldsymbol{P}_{\delta}\bar{\boldsymbol{A}}^{\top} + \bar{\boldsymbol{G}} \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix} \bar{\boldsymbol{G}}^{\top} \tag{5.33}$$

The covariance of the output innovation sequence is described by:

$$\boldsymbol{P}_{\epsilon}(0) = \boldsymbol{C}\boldsymbol{P}_{\delta}\boldsymbol{C}^{\top} + \boldsymbol{R} \tag{5.34}$$

$$\boldsymbol{P}_{\epsilon}(\tau) = \boldsymbol{C}\bar{\boldsymbol{A}}^{\tau}\boldsymbol{P}_{\delta}\boldsymbol{C}^{\top} - \boldsymbol{C}\bar{\boldsymbol{A}}^{\tau-1}\boldsymbol{ALR} \tag{5.35}$$

where $\tau = 1, \dots, m-1$ is the time delay and the user-defined parameter m is the maximum lag.

Solution to Eq. (5.33) and its substitution into Eq. (5.34) and Eq. (5.35) gives the linear system $\mathcal{A} = \boldsymbol{\vartheta b}$, where the design matrix $\mathcal{A}$ is defined as:

$$\mathcal{A} = \left[\mathcal{D}, \mathcal{D}(\boldsymbol{AL} \otimes \boldsymbol{AL}) + (\boldsymbol{I}_{n_y} \otimes \boldsymbol{\Gamma})\right] \tag{5.36}$$

$$\mathcal{D} = (\boldsymbol{C} \otimes \mathcal{O})(\boldsymbol{I}_{n_x^2} - \bar{\boldsymbol{A}} \otimes \bar{\boldsymbol{A}})^{-1} \tag{5.37}$$

$$\mathcal{O} = \left[\boldsymbol{C}^{\top}, (\boldsymbol{C}\bar{\boldsymbol{A}})^{\top}, \dots, (\boldsymbol{C}\bar{\boldsymbol{A}}^{m-1})^{\top}\right]^{\top} \tag{5.38}$$

$$\boldsymbol{\Gamma} = \left[\boldsymbol{I}_{n_y}, -(\boldsymbol{CAL})^{\top}, \dots, -(\boldsymbol{C}\bar{\boldsymbol{A}}^{m-2}\boldsymbol{AL})^{\top}\right]^{\top} \tag{5.39}$$

The unknown parameter vector is composed by the noise CMs $\boldsymbol{\vartheta} = [\boldsymbol{Q}_s^{\top}, \boldsymbol{R}_s^{\top}]^{\top}$. Instead, the known variable is given by the innovation covariance matrix as:

$$\boldsymbol{b} = \mathcal{C}_s(m) \tag{5.40}$$

$$\mathcal{C}(m) = [\boldsymbol{P}_{\epsilon}(0)^{\top}, \boldsymbol{P}_{\epsilon}(1)^{\top}, \dots, \boldsymbol{P}_{\epsilon}(m-1)^{\top}]^{\top} \tag{5.41}$$

where the symbol $\otimes$ stands for the Kronecker product and the notation $\boldsymbol{A}_s$ means the column-wise stacking of the matrix $\boldsymbol{A}$ into a vector [53, 27].

The estimate $\hat{\boldsymbol{\vartheta}}$ in the least-squares sense is given by:

$$\hat{\boldsymbol{\vartheta}} = (\mathcal{A}^{\top}\mathcal{A})^{-1}\mathcal{A}^{\top}\hat{\boldsymbol{b}} = \mathcal{A}^{\dagger}\hat{\boldsymbol{b}} \tag{5.42}$$

where, due to the ergodicity process assumption, the estimate $\hat{\boldsymbol{b}} = \hat{\mathcal{C}}_s(m)$ is computed from Eq. (5.41) with:

$$\hat{\boldsymbol{P}}_{\epsilon}(\tau) = \frac{1}{N-\tau}\sum_{i=1}^{N-\tau} \boldsymbol{\epsilon}_{i+\tau}\boldsymbol{\epsilon}_i^{\top} \tag{5.43}$$

The ALS estimator is proven to be unbiased and consistent [137]. The efficiency of the estimator is studied in [144] by a weighted LS, although the optimal weighting is impractical using current computational techniques. Also, the case of a given process noise shaping matrix $\boldsymbol{G} \neq \boldsymbol{I}$ is studied in [144]. Identifiability conditions of the unique elements of the CMs are recently studied in [104].

In some cases, the noise CMs formed by the ALS estimate may not be positive semidefinite, i.e. they are physically meaningless. Few contributions in literature do address this issue, see [50, 137, 144]. In these cases, a practical solution is to use *Semi-Definite Programming* (SDP) to enforce the constraints:

$$\hat{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta}} \|\mathcal{A}\boldsymbol{\vartheta} - \hat{\boldsymbol{b}}\|_2^2 \tag{5.44a}$$

$$\text{s.t. } \boldsymbol{Q} \geq \boldsymbol{0}, \boldsymbol{R} \geq \boldsymbol{0} \tag{5.44b}$$

CHAPTER 6

The direct solution to the filtering design from data problem

As already seen in Chapter 4, the derivation of data-driven solutions, namely the standard one and the direct one, to the filtering design problem is carried out considering the following description of a discrete-time LTI system $\mathcal{S}$:

$$\mathcal{S}: \quad \begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{G}\boldsymbol{w}_k && (6.1a)\\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{D}\boldsymbol{u}_k + \boldsymbol{v}_k && (6.1b)\\ \boldsymbol{z}_k &= \boldsymbol{C}_2\boldsymbol{x}_k + \boldsymbol{e}_k && (6.1c) \end{aligned}$$

where the new Eq. (6.1c) describes the possibility to measure the desired variable $\boldsymbol{z}_k$ to be filtered, for a limited amount of time, i.e. for $k = 1, \dots, N_{\text{ID}}$. In Eq. (6.1c), $\boldsymbol{e}_k$ denotes the measurement noise affecting the equation, which by construction, is a white noise process uncorrelated with the other noise process affecting the system.

For the derivation of filtering solutions, the general framework in Eq. (5.1) is further simplified as follows:

$$\mathcal{S}: \quad \begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{w}_k && (6.2a)\\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k && (6.2b)\\ \boldsymbol{z}_k &= \boldsymbol{x}_k + \boldsymbol{e}_k && (6.2c) \end{aligned}$$

by considering only the stochastic component of the sub-system, i.e. $\boldsymbol{B} = \boldsymbol{D} = \boldsymbol{0}$, the process noise shaping matrix is the identity matrix for the sake of simplicity, i.e. $\boldsymbol{G} = \boldsymbol{I}$, and the measurement matrix of the desired variable is the identity matrix, i.e. $\boldsymbol{C}_2 = \boldsymbol{I}$.

Later, the assumption that $\boldsymbol{B} = \boldsymbol{0}$ is relaxed in the presence of a deterministic exogenous input, i.e. $\boldsymbol{B} \neq \boldsymbol{0}$, in Chapter 7 when performing experimental tests. For this reason in this chapter, the derivation of the data-driven direct solution to the filtering design problem considers the simplified Eq. (6.2) with $\boldsymbol{B} \neq \boldsymbol{0}$, so that the derivation is generalized when needed and be easily reduced to the case $\boldsymbol{B} = \boldsymbol{0}$.

Under this premise, this chapter reviews the contribution in the development of the data-driven direct solution to the filtering design problem. In particular, this chapter is organized as follow:

1. Section 6.1 introduces the mathematical framework and a set of assumptions that must be undertaken in order to define exactly what is the problem statement of the filtering design problem in the direct paradigm based on a single-step approach.
2. Section 6.2 is dedicated to briefly formulate the methodology for the design of the filter, which consists in a optimization routine to estimate the parametric-stochastic structure of the filter.

6.1 Assumptions and problem statement

Next, in order to develop the data-driven direct solution for the filtering design problem for LTI systems, let suppose the following basic assumptions:

Assumption 6.1. The assumptions are:

- The system functions (matrices) $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ defining the system $\mathcal{S}$ to be filtered are unknown.
- The couple $(\boldsymbol{A}, \boldsymbol{C})$ is observable.
- A set of data (the identification dataset) is available for the design of the filter:

$$\mathcal{D}_{\mathrm{ID}} = \left\{(u_k, y_k, z_k), k = 1, \dots, N_{\mathrm{ID}}\right\} \tag{6.3}$$

- The noises $\boldsymbol{w}_k$, $\boldsymbol{v}_k$, and $\boldsymbol{e}_k$ are unmeasured stochastic variables.

Then, the filter design problem is defined as follows:

The filtering design problem

Design a causal filter using the identification dataset $\mathcal{D}_{\mathrm{ID}}$ that, operating on the input-output data $\left\{(u_k, y_k), k = 1, \dots, N_{\mathrm{ID}}\right\}$ gives an estimate $\hat{z}_k$ of the desired variable z_k, having the minimum estimation error variance property $\mathbb{E}[z_k - \hat{z}_k]$ for any k.

From the observability assumption in **Asm. 6.1** due to **Lem. 4.1** there exists a function f_0 that operating on the input-output data as required by the filtering design problem statement can be used for the estimation of the desired variable, let this function be termed as filter:

$$\hat{z}_{k|k} = f_0(\boldsymbol{u}_k, \boldsymbol{u}_{k-1}, \cdots, \boldsymbol{u}_{k-n_2+1}, \boldsymbol{y}_k, \boldsymbol{y}_{k-1}, \cdots, \boldsymbol{y}_{k-n_1+1}, \hat{z}_{0|0}) \tag{6.4}$$

Moreover, note that since the system $\mathcal{S}$ under analysis is linear, also function f_0 is expected to be linear.

Remark 6.1. Note also that, differently from the classical solution where the filter structure is not chosen in the two-step procedure and it just depends on the structure of the identified model, now, defining the methodology for the direct solution there is the need to select a parametric structure for the filter to be designed. In other words, compared to the *Prediction Error Minimization* (PEM) methodology [20, 30] in the *System Identification theory*, the problem of choosing one particular model structure with a suitable order complexity is shifted from the estimation of a model of the system to the estimation of a model of the filter.

6.2 The direct filter design methodology

Once the mathematical framework and the assumptions are defined, it is time to define the methodology to develop the direct data-driven solution to the filtering design problem. Recalling from Chapter 4, where the one-step approach was introduced, there is the interest to develop a methodology for a parametric-stochastic framework. By means of the terminology developed and involved with the *prediction-error* (PE) framework, let then the true filter notation to be $\mathcal{F}$.

It is now possible to define the filter by the general parametric model structure:

$$\mathcal{M}(\boldsymbol{\theta}_{\mathrm{DF}}) \tag{6.5}$$

where $\boldsymbol{\theta}_{\mathrm{DF}}$ is the unknown parameter to be estimated.

As usual, the model structure then defines the hypothesis set $\mathcal{H}$ containing all the feasible models that can result from the estimation considering its possible constraints:

$$\mathcal{H} = \left\{ \mathcal{M}(\boldsymbol{\theta}_{\mathrm{DF}}) \mid \boldsymbol{\theta}_{\mathrm{DF}} \in \boldsymbol{\Theta}_{\mathrm{DF}} \subset \mathbb{R}^d \right\} \tag{6.6}$$

Then, under the assumption that the filter is contained in the hypothesis set, i.e. $\mathcal{F} \in \mathcal{H}$, the best approximating (thus feasible) filter model $\hat{\mathcal{F}} = \mathcal{M}(\hat{\boldsymbol{\theta}}_{\mathrm{DF}})$ is searched for in $\mathcal{H}$ by means of solving the following optimization problem:

$$\hat{\boldsymbol{\theta}}_{\mathrm{DF}} = \underset{\boldsymbol{\theta}_{\mathrm{DF}} \in \boldsymbol{\Theta}_{\mathrm{DF}}}{\arg\min}\, J_{N_{\mathrm{ID}}}(\boldsymbol{\theta}_{\mathrm{DF}}) \tag{6.7}$$

$$\text{with } J_{N_{\mathrm{ID}}}(\boldsymbol{\theta}_{\mathrm{DF}}) = \frac{1}{N_{\mathrm{ID}}} \sum_{i=1}^{N_{\mathrm{ID}}} \left\| \boldsymbol{\varepsilon}_i(\boldsymbol{\theta}_{\mathrm{DF}}) \right\|_2^2$$

where $\left\{ \boldsymbol{\varepsilon}_i(\boldsymbol{\theta}_{\mathrm{DF}}) \equiv z_i - \hat{z}_i(\boldsymbol{\theta}_{\mathrm{DF}}) \right\}_{i=1}^{N}$ is the estimation error sequence of the model $M(\boldsymbol{\theta}_{\mathrm{DF}})$, and $\|\cdot\|$ is a scalar-valued norm function, for instance the l_2 norm, see [20].

It is worth mentioning that the estimation error sequence $\left\{ \boldsymbol{\varepsilon}_i(\boldsymbol{\theta}_{\mathrm{DF}}) \equiv z_i - \hat{z}_i(\boldsymbol{\theta}_{\mathrm{DF}}) \right\}_{i=1}^{N}$ relies on the estimate $\hat{z}_k$ which could potentially both the predicted or the filtered variable. As for now, when considering that the direct data-driven paradigm theory is not mature enough, this is a open question still not resolved. However some considerations can be made:

- By considering the prediction as the estimate, the designer could potentially exploit some theoretical insights derived from the well-known *Prediction Error Minimization* methodology.

- Instead, if considering filtering performance and goodness of estimation (i.e. closeness of the estimated parameter with the best optimal one, the true parameter), it is also interesting to know whether there is a difference developing the direct data-driven methodology by considering the filtered variable as the estimate.

In the following methodology the second way was chosen in the direct data-driven strategy.

6.3 Selection of the filter model structure and filter model complexity

Returning to the methodology, when considering the interest in developing a parametric-stochastic framework for the data-driven direct filter design problem, recall the difficulty highlighted in Rem. 6.1.

The natural question that arises, that is also very interesting to be answered, is: **whether there is some benefit to pick a particular model structure with a defined model complexity directly in the filtering design process or if it possible at all to have some insights**.

It is known that, when considering the classical solution based on the two-step approach, the filter model structure is defined automatically by means of applying the optimal filtering theory on the system. In other words the system structure naturally defines the filter model structure and model complexity. Instead, in the direct paradigm theory there could be a benefit in choosing a particular structure for the identification of a good filter. To this end, considering only the type of system under analysis, see again Eq. (6.1), i.e. a dynamical discrete-time LTI system, there are some potential insights on the optimal model structure and model complexity by exploiting the BLUE solution derived in Chapter 3 for this type of systems. In particular, the idea is to derive a filtering solution to the LTI case that depends only on the available dataset, that is, in the direct paradigm, on the identification dataset $\mathcal{D}_{\mathrm{ID}}$. To this end, recap from Chapter 3 the correction and prediction equations derived from the Bayesian approach. Note that the equations are modified to take care of the *time-invariant* property of the system matrices. In other words, let the general time-variant system matrices to be fixed with respect to time:

$$\boldsymbol{A}_k = \boldsymbol{A}, \quad \forall k \tag{6.8a}$$
$$\boldsymbol{B}_k = \boldsymbol{B}, \quad \forall k \tag{6.8b}$$
$$\boldsymbol{C}_k = \boldsymbol{C}, \quad \forall k \tag{6.8c}$$
$$\boldsymbol{D}_k = \boldsymbol{D}, \quad \forall k \tag{6.8d}$$
$$\boldsymbol{G}_k = \boldsymbol{G}, \quad \forall k \tag{6.8e}$$
$$\boldsymbol{Q}_k = \boldsymbol{Q}, \quad \forall k \tag{6.8f}$$
$$\boldsymbol{R}_k = \boldsymbol{R}, \quad \forall k \tag{6.8g}$$

Then, the statistics values for the mean and covariance for the correction step are as follows:

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_{k|k-1} - \boldsymbol{D}\boldsymbol{u}_k) \tag{6.9a}$$

$$\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k \boldsymbol{S}_k \boldsymbol{K}_k^{\mathsf{T}} \tag{6.9b}$$

$$= \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k \boldsymbol{S}_k \boldsymbol{K}_k^{\mathsf{T}} \tag{6.9c}$$

where $\boldsymbol{K}_k$ is known as the *Kalman gain*:

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\boldsymbol{C}^{\mathsf{T}}(\boldsymbol{C}\boldsymbol{P}_{k|k-1}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{R}) \tag{6.10a}$$

$$= \boldsymbol{P}_{k|k-1}\boldsymbol{C}^{\mathsf{T}}\boldsymbol{S}_k \tag{6.10b}$$

with:

$$\boldsymbol{S}_k = \boldsymbol{C}\boldsymbol{P}_{k|k-1}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{R} \tag{6.11}$$

Instead, the statistics values for the mean and covariance for the prediction step are as follows:

$$\hat{\boldsymbol{x}}_{k+1|k} = \boldsymbol{A}\hat{\boldsymbol{x}}_{k|k} + \boldsymbol{B}\boldsymbol{u}_k \tag{6.12a}$$

$$\boldsymbol{P}_{k+1|k} = \boldsymbol{A}\boldsymbol{P}_{k|k}\boldsymbol{A}^{\mathsf{T}} + \boldsymbol{G}\boldsymbol{Q}\boldsymbol{G}^{\mathsf{T}} \tag{6.12b}$$

Next, consider the prediction equation for the mean statistics in Eq. (6.12a) and shift it one time instant in the past as follows:

$$\hat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{A}\hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{B}\boldsymbol{u}_{k-1} \tag{6.13}$$

Then, it possible to substitute Eq. (6.13) into Eq. (6.9a), yielding:

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{C}\hat{\boldsymbol{x}}_{k|k-1} - \boldsymbol{D}\boldsymbol{u}_k) \tag{6.14}$$

$$\hat{\boldsymbol{x}}_{k|k} \overset{(6.13)}{=} \boldsymbol{A}\hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{B}\boldsymbol{u}_{k-1} + \boldsymbol{K}_k\big(\boldsymbol{y}_k - \boldsymbol{C}(\boldsymbol{A}\hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{B}\boldsymbol{u}_{k-1}) - \boldsymbol{D}\boldsymbol{u}_k\big) \tag{6.15}$$

$$\hat{\boldsymbol{x}}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{C})\boldsymbol{A}\hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{K}_k\boldsymbol{y}_k + (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{C})\boldsymbol{B}\boldsymbol{u}_{k-1} - \boldsymbol{K}_k\boldsymbol{D}\boldsymbol{u}_k \tag{6.16}$$

Note that Eq. (6.16) is composed by the required ingredients, namely the left side of the equation has the filtered estimate explicated, while the right side of the equation is based on past and present available data. Anyway, in order to reduce the complexity of the methodology, two more assumptions are made:

- First, Eq. (6.16) suffers from being based on time-variant parameters. This can be observed when analyzing the Kalman gain $\boldsymbol{K}_k$. In other words, even if the system $\mathcal{S}$ is time-invariant, the derived filter may as well be time-variant. In order to solve this difficulty, **the steady-state behavior of the filter is assumed from now on**. To this end, when considering the steady-state behavior, it is well-known that the KF gain $\boldsymbol{K}_\infty$ can be computed off-line, see for instance [29], from Eq. (5.24) by solving the DARE (*Discrete Arithmetic Riccati Equation*), that is:

$$\boldsymbol{P}_\infty = \boldsymbol{A}\boldsymbol{P}_\infty\boldsymbol{A}^{\mathsf{T}} + \boldsymbol{Q} - \boldsymbol{A}\boldsymbol{P}_\infty\boldsymbol{C}^{\mathsf{T}}(\boldsymbol{C}\boldsymbol{P}_\infty\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{R})^{-1}\boldsymbol{C}\boldsymbol{P}_\infty\boldsymbol{A}^{\mathsf{T}} \tag{6.17}$$

  using the assumption (made in the beginning) that $\boldsymbol{G} = \boldsymbol{I}$.

- Second, often the feedthrough matrix is assumed null, i.e. $\boldsymbol{D} = \boldsymbol{0}$. This is the case when there is no static input-output component from the dynamical system, as it often already filtered out when dealing with dynamical models. Following this rationale, for the sake of simplicity, from now on it also assumed that the feedthrough matrix is $\boldsymbol{D} = \boldsymbol{0}$.

By applying the mentioned assumptions, Eq. (6.16) can be reduced even more to:

$$\hat{\boldsymbol{x}}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_\infty \boldsymbol{C})\boldsymbol{A}\hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{K}_\infty \boldsymbol{y}_k + (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{C})\boldsymbol{B}\boldsymbol{u}_{k-1} \tag{6.18}$$

where Eq. (6.18) is considered as the *reference equation* to derive an appropriate model structure for the design of the filter.

To this end, using matrix polynomials, Eq. (6.18) turns in:

$$\boldsymbol{A}(z^{-1})\hat{\boldsymbol{x}}_{k|k} = \boldsymbol{B}_y(z^{-1})\boldsymbol{y}_k + \boldsymbol{B}_{\boldsymbol{u}}(z^{-1})\boldsymbol{u}_k \tag{6.19}$$

where the matrix polynomials are defined as follows:

$$\boldsymbol{A}(z^{-1}) = \boldsymbol{A}^{[0]} + \boldsymbol{A}^{[1]}z^{-1} \tag{6.20a}$$
$$\boldsymbol{B}_y(z^{-1}) = \boldsymbol{B}_y^{[0]} \tag{6.20b}$$
$$\boldsymbol{B}_{\boldsymbol{u}}(z^{-1}) = \boldsymbol{B}_{\boldsymbol{u}}^{[0]} + \boldsymbol{B}_{\boldsymbol{u}}^{[1]}z^{-1} \tag{6.20c}$$

with the following values:

$$\boldsymbol{A}^{[0]} = \boldsymbol{I} \tag{6.21a}$$
$$\boldsymbol{A}^{[1]} = (\boldsymbol{I} - \boldsymbol{K}_\infty \boldsymbol{C})\boldsymbol{A} \tag{6.21b}$$
$$\boldsymbol{B}_y^{[0]} = \boldsymbol{K}_\infty \tag{6.21c}$$
$$\boldsymbol{B}_{\boldsymbol{u}}^{[0]} = \boldsymbol{0} \tag{6.21d}$$
$$\boldsymbol{B}_{\boldsymbol{u}}^{[1]} = (\boldsymbol{I} - \boldsymbol{K}_\infty \boldsymbol{C})\boldsymbol{B} \tag{6.21e}$$

Using matrix fraction descriptions (MFD) [15], Eq. (6.19) can be seen as a model of the form:

$$\hat{\boldsymbol{x}}_{k|k} = \frac{\boldsymbol{B}_y(z^{-1})}{\boldsymbol{A}(z^{-1})}\boldsymbol{y}_k + \frac{\boldsymbol{B}_{\boldsymbol{u}}(z^{-1})}{\boldsymbol{A}(z^{-1})}\boldsymbol{u}_k \tag{6.22}$$

Then, by defining the true filter model $G_0(z^{-1};\boldsymbol{\theta}_0)$ as:

$$G_0(z^{-1};\boldsymbol{\theta}_0) = \left[\frac{\boldsymbol{B}_y(z^{-1})}{\boldsymbol{A}(z^{-1})}, \frac{\boldsymbol{B}_{\boldsymbol{u}}(z^{-1})}{\boldsymbol{A}(z^{-1})}\right] \tag{6.23}$$

it is possible to rewrite Eq. (6.22) into a compact form as follows:

$$\hat{\boldsymbol{x}}_{k|k} = G_0(z^{-1};\boldsymbol{\theta}_0)[\boldsymbol{y}_k^\top, \boldsymbol{u}_k^\top]^\top \tag{6.24}$$

with the (free) true parameter $\boldsymbol{\theta}_0$ defined by the free known values:

$$\boldsymbol{\theta}_0 = \left[\boldsymbol{A}^{[1]}, \boldsymbol{B}_y^{[0]}, \boldsymbol{B}_{\boldsymbol{u}}^{[1]}\right]^\top \tag{6.25}$$

with the number of elements of the parameter calculated as:

$$\begin{aligned} n_{\theta_0} &= \overbrace{n_x \times n_x}^{A^{[1]}} + \overbrace{n_x \times n_y}^{B_y^{[0]}} + \overbrace{n_x \times n_u}^{B_u^{[1]}} && (6.26a) \\ &= n_x \times (n_x + n_y + n_u) && (6.26b) \end{aligned}$$

Note that the fixed matrices $\boldsymbol{A}^{[0]} = \boldsymbol{I}$ and $\boldsymbol{B}_{\boldsymbol{u}}^{[0]} = \boldsymbol{0}$ were left out. Otherwise, the (full) true parameter would be defined as:

$$\boldsymbol{\theta}_0 = \left[\boldsymbol{A}^{[0]}, \boldsymbol{A}^{[1]}, \boldsymbol{B}_y^{[0]}, \boldsymbol{B}_{\boldsymbol{u}}^{[0]}, \boldsymbol{B}_{\boldsymbol{u}}^{[1]}\right]^\mathsf{T} \tag{6.27}$$

with the number of elements of the parameter calculated as:

$$\begin{aligned} n_{\theta_0} &= \overbrace{n_x \times n_x}^{A^{[0]}} + \overbrace{n_x \times n_x}^{A^{[1]}} + \overbrace{n_x \times n_y}^{B_y^{[0]}} + \overbrace{n_x \times n_u}^{B_u^{[0]}} + \overbrace{n_x \times n_u}^{B_u^{[1]}} && (6.28a) \\ &= n_x \times (2 \cdot n_x + n_y + 2 \cdot n_u) && (6.28b) \end{aligned}$$

In order to rewrite Eq. (6.24) in terms of the desired variable z_k, consider the availability in the design step, i.e. $k \leq N_{\mathrm{ID}}$, of Eq. (6.1c) and the state filtering error equation related to the KF theory:

$$\boldsymbol{\delta}_k = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k} \tag{6.29}$$

then, by substitution, Eq. (6.22) can be reformed as:

$$z_k = G_0(z^{-1}; \boldsymbol{\theta}_0)[\boldsymbol{y}_k^\mathsf{T}, \boldsymbol{u}_k^\mathsf{T}]^\mathsf{T} + \boldsymbol{\rho}_k \tag{6.30}$$

where the new stochastic noise $\boldsymbol{\rho}_k$ is defined as follows:

$$\boldsymbol{\rho}_k \equiv \boldsymbol{\delta}_k + \boldsymbol{e}_k \tag{6.31}$$

Note that Eq. (6.30) can be used as the reference model for the direct filtering estimation since the input of the reference model are the input/output data of the system, and the output of the reference model is the desired variable to be estimated, for which some samples are available during the design step. Following this rationale, the signal model structure of the direct filter is given by Eq. (6.23) whereas the filter model complexity is given, respectively, by the polynomials and the known quantities in Eqs. (6.20) and (6.21).

In particular, observe that the model of the DF is:

$$\mathcal{M}(z^{-1}; \boldsymbol{\theta}_{\mathrm{DF}}) = G_0(z^{-1}; \boldsymbol{\theta}_0) \tag{6.32}$$

where the model structure of the direct filter can be decomposed into the well-known signal model $G(z^{-1}; \boldsymbol{\theta}_{\mathrm{DF}})$ and noise model $H(z^{-1}; \boldsymbol{\theta}_{\mathrm{DF}})$ from the system identification theory [20]:

$$\mathcal{M}(z^{-1}; \boldsymbol{\theta}_{\mathrm{DF}}) = \left\{ G(z^{-1}; \boldsymbol{\theta}_{\mathrm{DF}}),\ H(z^{-1}; \boldsymbol{\theta}_{\mathrm{DF}}) \mid \boldsymbol{\theta}_{\mathrm{DF}} \in \boldsymbol{\Theta}_{\mathrm{DF}} \right\} \tag{6.33}$$

where it is now clear that the aim of the direct filtering design step following the estimation methodology in Section 6.2 is to estimate the direct signal model as close as possible to the one derived from the optimal KF theory. In other words, it is desirable that:

$$G(z^{-1};\boldsymbol{\theta}_{\mathrm{DF}}) \rightarrow G_0(z^{-1};\boldsymbol{\theta}_0) \tag{6.34a}$$

$$\boldsymbol{\theta}_{\mathrm{DF}} \rightarrow \boldsymbol{\theta}_0 \tag{6.34b}$$

It is worth considering the general structure of the direct filtering by explicating also the filter noise model in Eq. (6.30), yielding:

$$\boldsymbol{z}_k = \boldsymbol{G}(z^{-1};\boldsymbol{\theta}_{\mathrm{DF}})[\boldsymbol{y}_k^{\mathrm{T}}, \boldsymbol{u}_k^{\mathrm{T}}]^{\mathrm{T}} + \boldsymbol{H}(z^{-1};\boldsymbol{\theta}_{\mathrm{DF}})\boldsymbol{\rho}_k \tag{6.35}$$

where the problem of selecting the model structure and model order of the noise model $\boldsymbol{H}(z^{-1};\boldsymbol{\theta}_{\mathrm{DF}})$ can be tackle down by considering the innovations point of view [94, 96]:

> “one criterium to justify the optimality of the solution to a filtering problem is to check how white the pseudo-innovations are, the whiter the more optimal, see [59]”.

Following this rationale, under the belief that the KF assumptions hold true in the derivation of the reference model in the steady-state case, it is expected that the state filtering error $\boldsymbol{\delta}_k$ in Eq. (6.29) is a white process, otherwise losing the optimality condition of the theory. Moreover, also the measurement noise of the desire variable $\boldsymbol{e}_k$ is, by construction, a white process. Then in general, the linear combination of these noises in Eq. (6.31) forms the colored noise process $\boldsymbol{\rho}_k$. Moreover, the colored noise $\boldsymbol{\rho}_k$ can be modeled by a dynamical system, here the noise model, with a white noise process as input, here termed again for simplicity as $\boldsymbol{\rho}_k$. Using post-estimation model validation techniques such as residual analysis, a proper structure for the noise model can be validated.

If it is expected that the noise model for the filter has no dynamic component, then the model structure can be fixed to the the unitary matrix as in the *Output-Error* (OE) model, i.e.

$$\boldsymbol{H}(z^{-1};\boldsymbol{\theta}_{\mathrm{DF}}) = \boldsymbol{I} \tag{6.36}$$

If that is not the case, other linear model structures could be used, such as the *AutoRegressive with eXogenous inputs* (ARX) model structure. In the end, once the filter design step is finished, the resulting parameter estimate $\hat{\boldsymbol{\theta}}_{\mathrm{DF}}$ can be used to operate with the identified filter as in the *Linear In Parameter* (LIP) regression in Eq. (6.16) to obtain filtered estimates of the desired variable as follows:

$$\hat{\boldsymbol{z}}_{k|k} = \hat{\boldsymbol{\theta}}_{\mathrm{DF}} \cdot \boldsymbol{\Phi}_k \tag{6.37}$$

with a user-selected starting condition parameter $\hat{\boldsymbol{z}}_{0|0}$.

In particular, the regressor variable $\boldsymbol{\Phi}_k$ is defined as follows:

$$\boldsymbol{\Phi}_k = [\underbrace{\hat{\boldsymbol{z}}_{k-1|k-1}^{\mathrm{T}}}_{1\times n_x}, \underbrace{\boldsymbol{y}_k^{\mathrm{T}}}_{1\times n_y}, \underbrace{\boldsymbol{u}_{k-1}^{\mathrm{T}}}_{1\times n_u}]^{\mathrm{T}} \in \mathbb{R}^{(n_x+n_y+n_u)\times 1} \tag{6.38}$$

and the estimated parameter $\hat{\boldsymbol{\theta}}_{\mathrm{DF}}$ has the following structure:

$$\hat{\boldsymbol{\theta}}_{\mathrm{DF}} = \left[\hat{\boldsymbol{A}}^{[1]}, \hat{\boldsymbol{B}}_{\boldsymbol{y}}^{[0]}, \hat{\boldsymbol{B}}_{\boldsymbol{u}}^{[1]}\right]^{\mathsf{T}} \in \mathbb{R}^{n_x \times (n_x + n_y + n_u)} \tag{6.39a}$$

$$= \begin{bmatrix} \hat{\theta}_{11} & \cdots & \hat{\theta}_{1\square} \\ \vdots & \ddots & \vdots \\ \hat{\theta}_{\triangle 1} & \cdots & \hat{\theta}_{\triangle\square} \end{bmatrix} \in \mathbb{R}^{\triangle \times \square} \tag{6.39b}$$

where the symbols $\triangle$ and $\square$ denotes, respectively, the quantities $n_{\boldsymbol{x}}$ and $n_{\boldsymbol{x}} + n_{\boldsymbol{y}} + n_{\boldsymbol{u}}$.

CHAPTER 7

Comparison between the standard and direct solutions

This chapter reviews the experimental results about the developed data-driven solutions to the filtering design problem. Two different paradigm, as seen in Chapter 4, are compared through the derived solutions, respectively the standard solution based on a two-step approach derived in Chapter 5 and the direct solution based on a one-step approach derived in Chapter 6. In particular, this chapter is organized as follow:

1. Section 7.1 introduces the results for a simplified univariate example, with no exogenous inputs. In this example, different experimental settings are simulated and both the parameter estimates of the filter model as well as its filtering performance are analyzed.
2. Section 7.2 introduces briefly a discussion about the difficulties to expand the developed filtering solutions and the practical experimental comparison to a multivariate system with exogenous input. In particular, the system under analysis is taken from an example of an aerospace industrial application about flight control.

As already seen in Chapter 5 and Chapter 6, the derived data-driven solutions to the filtering design problem are carried out considering the following general description of a discrete-time LTI system $\mathcal{S}$:

$$\mathcal{S}: \quad \begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k && (7.1a) \\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{D}\boldsymbol{u}_k + \boldsymbol{v}_k && (7.1b) \\ \boldsymbol{z}_k &= \boldsymbol{x}_k + \boldsymbol{e}_k && (7.1c) \end{aligned}$$

7.1 A simple univariate academic example

First, for the sake of easiness in the comparison of the derived data-driven solutions, the discrete-time LTI system $\mathcal{S}$ in Eq. (7.1) is simplified further by considering an univariate SISO (*Single Input, Single Output*) with no exogenous inputs. In particular,

Table 7.1: Different experimental settings regarding the SNR and the number of available samples for identification purposes. Each combination of settings define an experiment $E_i, i = 1, \dots, 6$ for the case of the univariate with no exogenous inputs example.

Experiment	E_1	E_2	E_3	E_4	E_5	E_6
SNR	100 $\boldsymbol{Q} = 1.000 \cdot 10^{-1}$ $\boldsymbol{R} = 4.999 \cdot 10^{-2}$ $\boldsymbol{\Sigma_e} = 1.666 \cdot 10^{-2}$		10 $\boldsymbol{Q} = 1.000 \cdot 10^{-1}$ $\boldsymbol{R} = 1.580 \cdot 10^{-1}$ $\boldsymbol{\Sigma_e} = 5.265 \cdot 10^{-2}$		3 $\boldsymbol{Q} = 1.000 \cdot 10^{-1}$ $\boldsymbol{R} = 2.883 \cdot 10^{-1}$ $\boldsymbol{\Sigma_e} = 9.610 \cdot 10^{-2}$	
$N_{\mathbf{VL}}$	4000					
$N_{\mathbf{ID}}$	6000	200	6000	200	6000	200

the highlighted example is one of the simplest and thus, can be considered a tutorial and academic example adhere. Finally, the system matrices take the following values:

$$\boldsymbol{A} = 0.8 \tag{7.2a}$$
$$\boldsymbol{B} = 0 \tag{7.2b}$$
$$\boldsymbol{C} = 3 \tag{7.2c}$$
$$\boldsymbol{D} = 0 \tag{7.2d}$$

and the considered system $\mathcal{S}$ is reduced as follows:

$$\mathcal{S}: \quad \begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{w}_k && (7.3a) \\ \boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k && (7.3b) \\ \boldsymbol{z}_k &= \boldsymbol{x}_k + \boldsymbol{e}_k && (7.3c) \end{aligned}$$

7.1.1 Experimental settings

Considering the stochastic part of the system, different noise CMs $\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{\Sigma_e}$ related, respectively, to the noise processes $\boldsymbol{w}_k$, $\boldsymbol{v}_k$, and $\boldsymbol{e}_k$ are tested in experimental settings summarized in Tab. 7.1. In particular, the experimental settings differ in terms of the SNR (*Single to Noise Ration*) of the noise processes affecting the system, defined, for instance, as follows:

$$SNR[\boldsymbol{y}] = \frac{\mathbb{E}[(\boldsymbol{C}\boldsymbol{x}_k)^2]}{\mathbb{E}[\boldsymbol{v}_k^2]} = \frac{\sum_{k=1}^{N_{\mathrm{VL}}} (\boldsymbol{C}\boldsymbol{x}_i)^2}{\cancel{N_{\mathrm{VL}}}} \frac{\cancel{N_{\mathrm{VL}}}}{\sum_{k=1}^{N_{\mathrm{VL}}} \boldsymbol{v}_i^2} \tag{7.4a}$$

$$SNR[\boldsymbol{z}] = \frac{\mathbb{E}[\boldsymbol{x}_k^2]}{\mathbb{E}[\boldsymbol{e}_k^2]} = \frac{\sum_{k=1}^{N_{\mathrm{VL}}} \boldsymbol{x}_i^2}{\cancel{N_{\mathrm{VL}}}} \frac{\cancel{N_{\mathrm{VL}}}}{\sum_{k=1}^{N_{\mathrm{VL}}} \boldsymbol{e}_i^2} \tag{7.4b}$$

where the computation can be carried out thanks to the fact that experiments are performed in simulation, thus noise processes data and their statistics (the mean of its squares) are available to the designer.

It is worth mentioning that only the measured variables were considered in this sense, namely the output of the system $\boldsymbol{y}_k$ and the desired variable to be estimated $\boldsymbol{z}_k$, which is considered to be measured for a limited amount of time. On the other hand, the process noise variable $\boldsymbol{w}_k$ and thus the system state $\boldsymbol{x}_k$ were neglected, as the process noise reflects the idea of errors in the modeling part of the system, which in practical cases cannot be considered to be measurable, else the better modeling. In the end, the following values of SNR were evaluated:

> **Experimental settings: the SNR of the measurable variables**
>
> - High $SNR = 100$: in this case the signal, even if affected by disturbance noise, is acquired without losing much information;
> - Medium $SNR = 10$;
> - Low $SNR = 3$: in this case the noise affecting the signal disturbs the acquisition resulting in inaccurate measurements;
>
> Note that the values are in a decimal scale and not in a logarithmic scale.

The experimental settings also differ in the number of data samples actually used for identification purpose, i.e. data used for the estimation of the system and/or filter models. In particular, the following values were considered:

> **Experimental settings: the number of samples for the identification**
>
> The number of samples used for the identification dataset in the experiments is as follows:
> - $N_{\mathrm{ID}} = 200$ samples (considered as a low number);
> - $N_{\mathrm{ID}} = 6000$ samples (considered a high number);

> **Experimental settings: the number of samples for the validation**
>
> The number of samples used for the validation dataset in the experiments is as follows:
> - $N_{\mathrm{VD}} = 4000$ samples (always the same);

The number of samples jointly with the SNR settings define a combinational discrete grid of six experimental settings, each one termed as *Experiment* and labeled accordingly as $E_i, i = 1, \dots, 6$, see again Table 7.1.

For instance, for the sake of completeness, Experiment E_1 is implemented as follows:

$$\mathcal{S}: \begin{cases} \boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{w}_k, & k = 1, \dots, N = 10000 \quad (7.5a) \\ \boldsymbol{y}_k = \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k, & k = 1, \dots, N = 10000 \quad (7.5b) \\ \boldsymbol{z}_k = \boldsymbol{x}_k + \boldsymbol{e}_k, & k = 1, \dots, N_{\mathrm{ID}} = 6000 \quad (7.5c) \end{cases}$$

with $\boldsymbol{w}_k \sim \mathcal{G}(0, 1.0 \cdot 10^{-1})$, $\boldsymbol{v}_k \sim \mathcal{G}(0, 4.999 \cdot 10^{-2})$, and $\boldsymbol{e}_k \sim \mathcal{G}(0, 1.666 \cdot 10^{-2})$.

Instead, for instance, Experiment E_4 is implemented as follows:

$$\mathcal{S}: \begin{cases} x_{k+1} = Ax_k + w_k, & k = 1, \dots, N = 4200 & (7.6a) \\ y_k = Cx_k + v_k, & k = 1, \dots, N = 4200 & (7.6b) \\ z_k = x_k + e_k, & k = 1, \dots, N_{\text{ID}} = 200 & (7.6c) \end{cases}$$

with $w_k \sim \mathcal{G}(0, 1.0 \cdot 10^{-1})$, $v_k \sim \mathcal{G}(0, 1.580 \cdot 10^{-1})$, and $e_k \sim \mathcal{G}(0, 5.265 \cdot 10^{-2})$.

7.1.2 MC simulations and collected datasets

Moreover, each experiment is made of $N_{\text{MC}} = 1000$ Monte-Carlo simulations in order not to consider potential outline cases. Following this rationale, the results were observed and computed on the empirical distribution of the MC simulation, or summarized by some of its statistics such as the mean value.

In the end, for each MC simulation, the realization collects $k = 1, \dots, N$ samples, where $N = N_{\text{ID}} + N_{\text{VL}}$, divided into two dataset, one for identification purpose and one for validation purpose as already partly mentioned in **Asm. 6.1**. The following datasets are thus available:

- $\mathcal{D}_{\text{ID}} = (y_{1:N_{\text{ID}}}, z_{1:N_{\text{ID}}})$ for the filter identification;
- $\mathcal{D}_{\text{VL}} = (y_{N_{\text{ID}}+1:N})$ for the validation and evaluation of filtering performance;

7.1.3 The filtering solutions and related parameter estimates

In the following section are presented in details the different kind of solutions that will be compared specifically on the mentioned univariate framework in Section 7.1. The analyzed solutions will be also used in the comparison of the multivariate example in Section 7.2 with an appropriate extension of the highlighted ingredients. These filtering solutions are as follows:

1. **Baseline** refers to the filtering solution based on the KF theory in the classical model-based approach. In other words, the deterministic and stochastic components of the system are known, i.e. the system matrices A, C, Q, R are available. Then, the filter design is derived by applying the KF theory on the system, see Chapter 3. In particular, the filter estimator that will be considered is the recursive equation given by Eq. (6.24). Since all quantities of interested are known and the filtering theory is optimal, this solution is termed as *baseline* as it will serve as the grounding example to compare other solutions against it.

2. **Standard $\hat{Q}\hat{R}$** refers to the filtering solution based on the standard data-driven two-step approach. Firstly, the available identification dataset is used to estimate the determinist components of the systems using the ad-hoc instrumental variable least squares solution implemented in Chapter 5. Then, the state-of-the-art correlation method for noise covariance matrices estimation is applied to identify also the stochastic components of the systems, namely the unknown noise covariance matrices Q and R, again see Chapter 5. The filter estimator for this solution is given again by Eq. (6.24), substituting the unknown matrices A, C, Q, R with their appropriate estimates, respectively $\hat{A}, \hat{C}, \hat{Q}, \hat{R}$. The resulting unknown parameter vector is termed $\hat{\theta}_{\text{KF}}$.

3. **Standard QR** refers to the same filtering solution based on the standard data-driven two-step approach as in the **Standard $\hat{Q}\hat{R}$**. The only difference is that in this case, the noise covariance matrices are not estimated but are given as if they were known. This example is very interesting since it serves for a couple of reasons:

   (a) First, the ad-hoc instrumental variable least-square method used for the identification of the deterministic component of the system can be checked whether its use deteriorate the filtering performance. In other words, **Standard $\hat{Q}\hat{R}$** suffers from having sequential estimation routines in other to estimate all the required ingredients for the KF to be applied. By getting rid of the noise covariance matrices estimation the designer can see the effects of the system identification techniques.

   (b) Second, it shows how the practical problem of noise covariance matrices estimation left to the filter designer is not trivial at all. As a consequence, when the stochastic properties are not fine-tuned the performance of the filter worsen. Moreover, theoretical insights derived in the model-based paradigm may fail to be applied when the practical situation enforces different assumptions. This motivates further the search for a new paradigm found in the direct solution. The filter estimator for this solution is given again by Eq. (6.24), substituting the unknown matrices $\boldsymbol{A}, \boldsymbol{C}$ with their appropriate estimates, respectively $\hat{\boldsymbol{A}}, \hat{\boldsymbol{C}}$, and exploiting the known matrices $\boldsymbol{Q}, \boldsymbol{R}$. The resulting unknown parameter vector is termed $\hat{\boldsymbol{\theta}}_{\mathrm{KF}}$.

4. **Direct** refers to the filtering solution based on the new direct data-driven one-step approach developed in this thesis. The filter is estimated by Eq. (6.7) following the developed methodology in Chapter 6 using a linear model structure defined by Eq. (6.35), where both an OE and an ARX model structure were tested.

Once the filter design step is finished, whatever the filtering solution is considered, since the common unifying framework is linear and its built on being statistical-parametric, the identified model of the filter can be represented by the estimated parameter $\hat{\boldsymbol{\theta}}$. Thus, let the identified filter model be defined as:

$$\hat{\mathcal{M}} = \mathcal{M}(\hat{\boldsymbol{\theta}}) \tag{7.7}$$

For instance, for the baseline solution, the $\hat{\boldsymbol{\theta}}$ is given by the true value in Eq. (6.25):

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \boldsymbol{\theta}_0 && \text{(7.8a)}\\
&= \left[\boldsymbol{A}^{[1]}, \boldsymbol{B}_y^{[0]}, \boldsymbol{B}_u^{[1]}\right]^{\mathsf{T}} && \text{(7.8b)}\\
&= \left[(\boldsymbol{I} - \boldsymbol{K}_\infty \boldsymbol{C})\boldsymbol{A}, \boldsymbol{K}_\infty\right]^{\mathsf{T}} && \text{(7.8c)}
\end{aligned}$$

where:

- from Eq. (7.8b) to Eq. (7.8c) the element $\boldsymbol{B}_u^{[1]}$ was not considered since the univariate example does not have an exogenous input, i.e.

$$n_u = 0 \tag{7.9}$$

- in Eq. (7.8c) the (remaining) first two elements reduce to scalar values, again due to the univariate examples, i.e.

$$n_x = n_y = 1 \tag{7.10}$$

Instead, for instance, for the standard $\hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ solution, the $\hat{\boldsymbol{\theta}}$ is given by the estimated values from the algorithm in Chapter 5:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \boldsymbol{\theta}_{\text{KF}} & (7.11\text{a}) \\ &= \left[\hat{\boldsymbol{A}}^{[1]}, \hat{\boldsymbol{B}}_{\boldsymbol{y}}^{[0]}, \hat{\boldsymbol{B}}_{\boldsymbol{u}}^{[1]}\right]^{\mathsf{T}} & (7.11\text{b}) \\ &= \left[(\boldsymbol{I} - \hat{\boldsymbol{K}}_{\infty}\hat{\boldsymbol{C}})\hat{\boldsymbol{A}}, \hat{\boldsymbol{K}}_{\infty}\right]^{\mathsf{T}} & (7.11\text{c}) \end{aligned}$$

where:

- from Eq. (7.11b) to Eq. (7.11c) the element $\boldsymbol{B}_{\boldsymbol{u}}^{[1]}$ was not considered since the univariate example does not have an exogenous input, i.e.

$$n_u = 0 \tag{7.12}$$

- in Eq. (7.11c) the (remaining) first two elements reduce to scalar values, again due to the univariate examples, i.e.

$$n_x = n_y = 1 \tag{7.13}$$

- in Eq. (7.11c) the estimates $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{C}}$ are given by the ad-hoc instrumental variable least square algorithm given by, respectively, Eq. (5.11) and Eq. (5.19). The other estimate $\hat{\boldsymbol{K}}_{\infty}$ is given by estimates $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{R}}$ identified from Eq. (5.44) required for the calculation of the DARE in Eq. (6.17).

Instead, for instance, for the direct solution, the $\hat{\boldsymbol{\theta}}$ is given by the optimization problem defined in the methodology of the direct solution in Eq. (6.7):

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \boldsymbol{\theta}_{\text{DF}} & (7.14\text{a}) \\ &= \left[\hat{\theta}_1, \hat{\theta}_2\right]^{\mathsf{T}} & (7.14\text{b}) \end{aligned}$$

where:

- in Eq. (7.14b) the two elements are scalar values, due to the univariate examples, i.e.

$$n_u = 0 \tag{7.15}$$

$$n_x = n_y = 1 \tag{7.16}$$

7.1.4 System identification and noise covariance matrices estimates in the standard solution

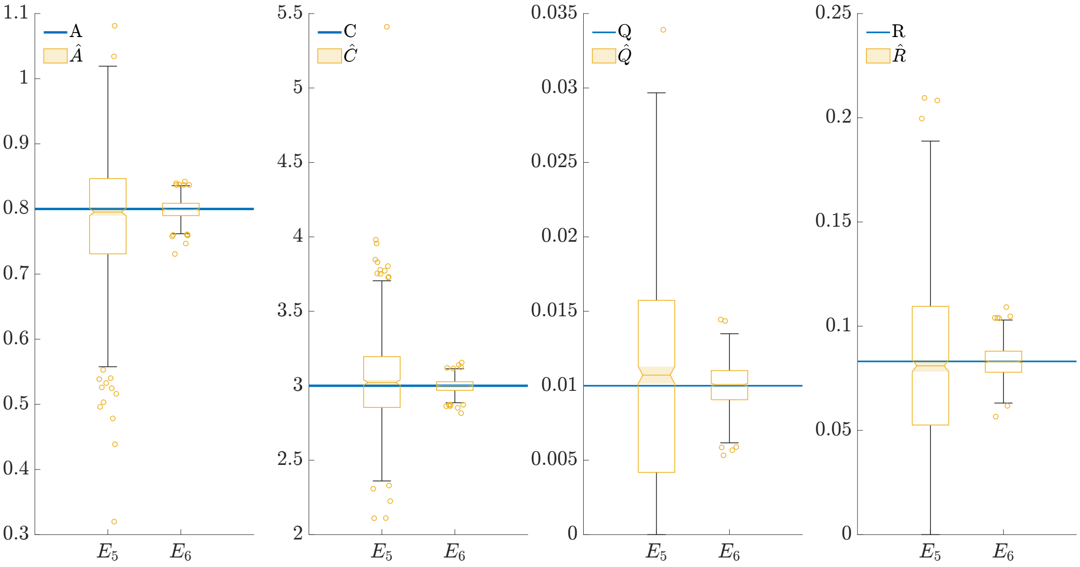


Figure 7.1: The deterministic and stochastic components of the identified system $\hat{A}, \hat{C}, \hat{Q}, \hat{R}$ in the univariate case for worst-case scenario experiments E_5 and E_6. In blu the **Baseline** solution and in yellow the **Standard $\hat{Q}\hat{R}$** solution.

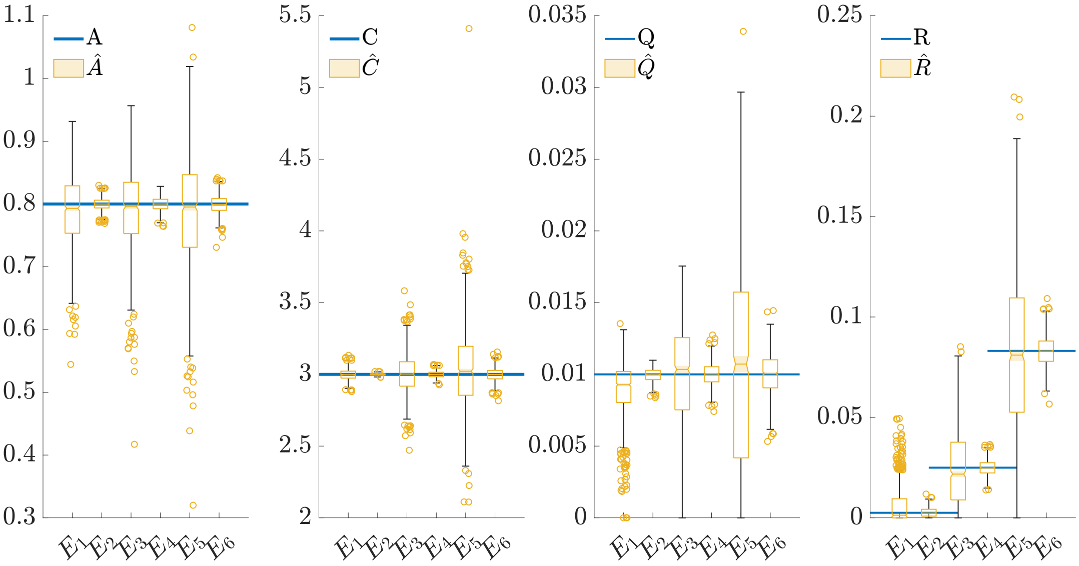


Figure 7.2: The deterministic and stochastic components of the identified system $\hat{A}, \hat{C}, \hat{Q}, \hat{R}$ in the univariate case for all experiments. In blu the **Baseline** solution and in yellow the **Standard $\hat{Q}\hat{R}$** solution.

For the sake of completeness, the two-step approach of the **Standard $\hat{Q}\hat{R}$** solution is explored in more details. In particular, the results of the system identification estimates, namely matrices $\hat{A}$ and $\hat{C}$ identified through the ad-hoc instrumental variable least square method of Section 5.1.1 and the noise covariance matrices $\hat{Q}$ and $\hat{R}$ identified through the *Direct Correlation Method* (DCM) of Section 5.2.1 are shown by means of their empirical distributions through boxplots in Fig. 7.1 for the univariate example for the worst-case scenario, i.e. experiments E_5 and E_6. Instead, the same figure containing all the experiments is available in Fig. 7.2.

From Fig. 7.2 is possible to observe that inter-experiments:

- with a decreasing SNR setting from experiments E_1 to experiments E_6, see Table 7.1, the variance of the estimates increases as expected from a standard bias-variance tradeoff discussion regarding estimation techniques.

On the other hand, from Fig. 7.1 is possible to observe that intra-experiments:

- the worst-case scenario experiments E_5 with a limited number of available identification samples fails to return unbiased estimates. The reason is found from the standard instrumental variable least square theory used to estimate the deterministic components of the system, namely matrice $\hat{\boldsymbol{A}}, \hat{\boldsymbol{C}}$, but also from the DCM method used to estimate the stochastic components of the system, namely matrices $\hat{\boldsymbol{Q}}, \hat{\boldsymbol{R}}$. Indeed both methods are founded on the asymptotic case theory, when the number of samples available for the estimation tend to infinity. Since the setting under analysis is to have a limited number of available data, in exact contrast with the required assumption, the result of having biased estimates could be expected.

7.1.5 Parameter estimates of the identified filter solutions

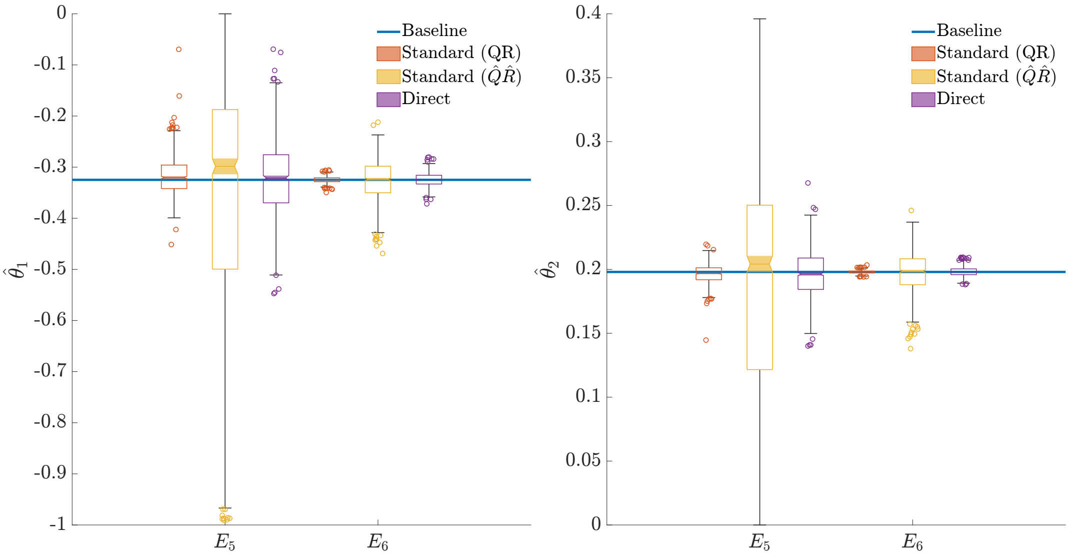


Figure 7.3: The empirical distribution of the identified filter parameter $\hat{\boldsymbol{\theta}}$ in the univariate case for the worst-case scenario of experiments E_5 and E_6. Different filtering solutions are compared: in blu the **Baseline** solution, in red the **Standard** $\boldsymbol{QR}$ solution, in yellow the **Standard** $\hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ solution, and finally in purple the **Direct** solution solution.

For the sake of completeness, results about the identified parameter $\hat{\boldsymbol{\theta}}$ of the discussed filtering solutions are displayed by means of their empirical distribution through box-plots in Fig. 7.3, where worst-case scenario settings are emphasized by experiments E_5 and E_6, and in Fig. 7.4 instead for a general overview of all experiments.

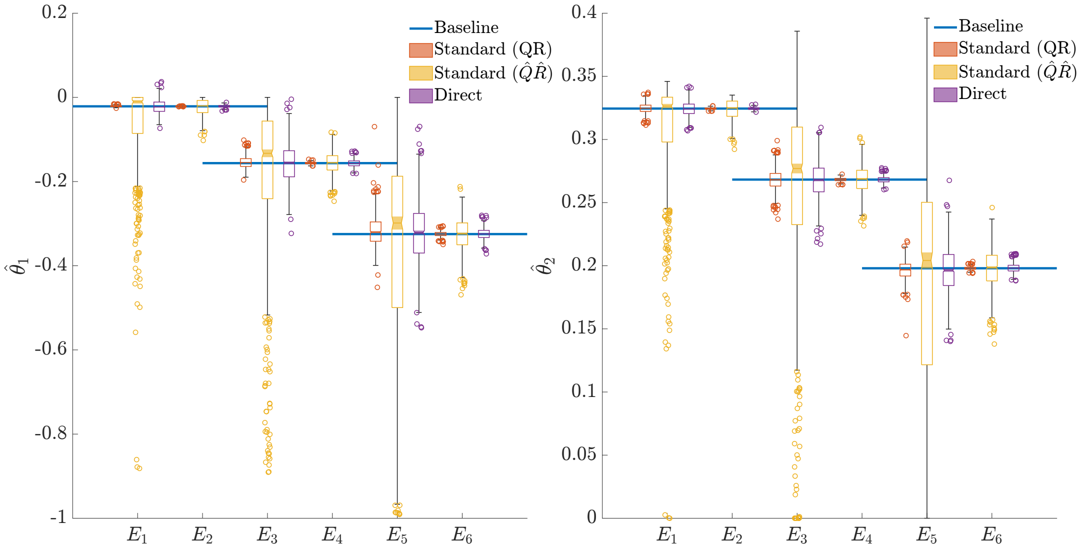


Figure 7.4: The empirical distribution of the identified filter parameter $\hat{\boldsymbol{\theta}}$ in the univariate case for all the experiments. Different filtering solutions are compared: in blu the **Baseline** solution, in red the **Standard** QR solution, in yellow the **Standard** $\hat{Q}\hat{R}$ solution, and finally in purple the **Direct** solution solution.

In particular, considering the univariate example under analysis, whatever the filtering solution is considered, from the unified framework the signal model of the filter is expected to be as follows:

$$\hat{G} = G(\hat{\boldsymbol{\theta}}) = \frac{\hat{\theta}_2}{1 + \hat{\theta}_1 z^{-1}} \tag{7.17}$$

thus, the general form of the estimated parameter $\hat{\boldsymbol{\theta}}$ consists of two scalar elements as follows:

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} \tag{7.18}$$

By observing Figs. 7.3 and 7.4 it is interesting to note that:

- In general, the empirical distributions of the identified parameter for all solutions seem to have the property of unbiasness, thus estimating accurately the expected filter parameters. The only exception is for experiments E_1, E_3, and E_5 for which the **Standard** $\hat{Q}\hat{R}$ solution suffers from having the property of unbiasness estimates. This problematic may be explained by noting that all those experiments share the same settings of having a limited number of available identification samples, see Table 7.1. Following this rationale, the consideration is that **Standard** $\hat{Q}\hat{R}$ solution is built on sequential estimation routines, see Chapter 5, that are based on the asymptotic case scenario. For this reason, since the setting under analysis is to have a limited number of available data, in exact contrast with the required assumption, the result of having biased estimates could be expected.
- It is also worth noting that the goal of a filtering design process is to design a filter which ultimately is successful in its reconstruction of the desired estimate.

Following this rationale, the focus should be posed on the filtering performance and not on the identified filter model. Then, the issues found from observing the empirical distribution of the identified parameter $\hat{\boldsymbol{\theta}}$ may be overlooked.

Remark 7.1. Note that in Figs. 7.3 and 7.4 the values of the identified parameter $\hat{\boldsymbol{\theta}}$ for the **Baseline** solution is a straight line since no estimation occurs. Indeed, in **Baseline** solution, as already seen in Section 7.1.3, the identified parameter $\hat{\boldsymbol{\theta}}$ is given by the true value $\boldsymbol{\theta}_0$ that can be computed by substitution of the known values of the system deterministic and stochastic matrices.

7.1.6 Model validation and model structure determination for the direct solution

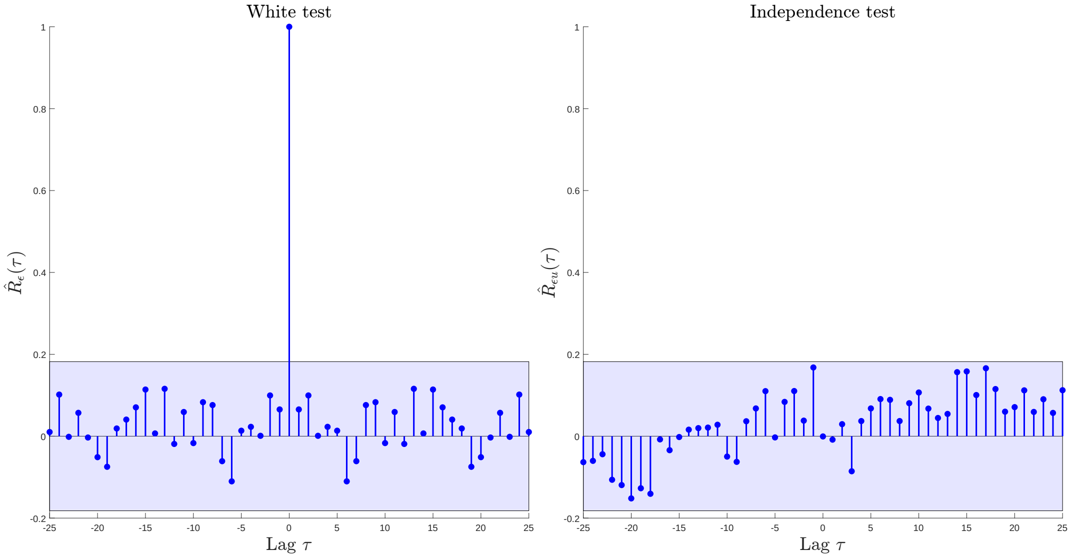


Figure 7.5: Results for the residual analysis test to validate the OE model structure of the direct filtering solution in the univariate example. On the left the white test showing the residual auto-covariance $\hat{R}_\epsilon(\tau)$, on the right the independence test showing the (filter) input-residual cross-covariance $\hat{R}_{\epsilon u}(\tau)$.

Limited to the worst-case scenario of experiment E_5, considering the **Direct** solution to the filtering design problem, the difficulty of selecting a proper noise model structure for the filter is still a open question not answered from the developed methodology. For these reasons, there is still room for more theoretical research. Anyway, in general, when searching for the correct model order different questions are raised, such as:

- Is a given model flexible enough?
- Is a given model too complex? (relevant also to model reduction)
- Which model structure between different candidates should be chosen?

To this end, statistical tests based on the empirical estimation error $\epsilon_k(\hat{\boldsymbol{\theta}})$ can be used to give some insights. In particular they are based on some assumptions [30]:

- Assumption 1: $\epsilon_k(\hat{\boldsymbol{\theta}})$ is a zero-mean white noise;

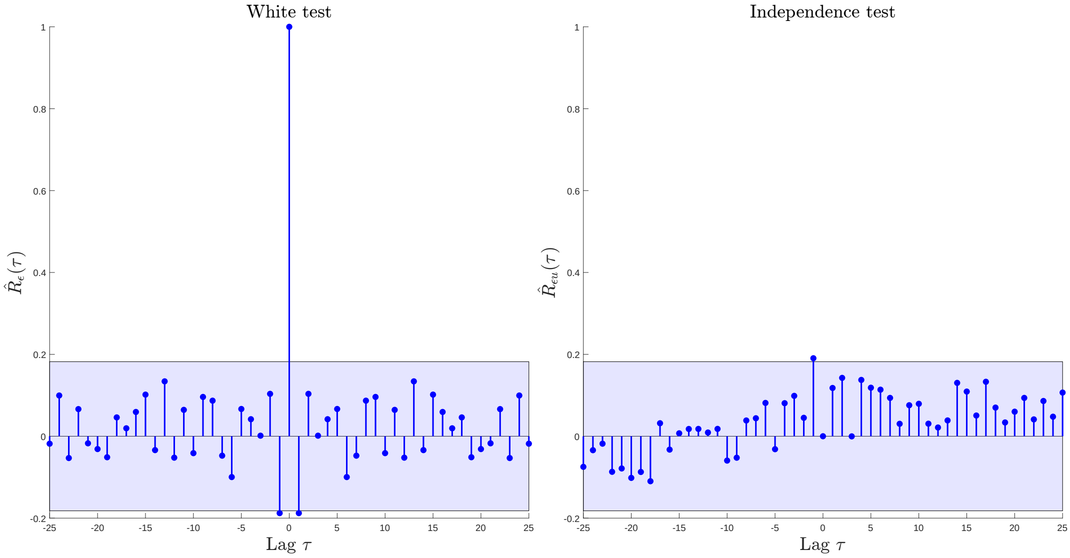


Figure 7.6: Results for the residual analysis test to validate the ARX model structure of the direct filtering solution in the univariate example. On the left the white test showing the residual auto-covariance $\hat{R}_\epsilon(\tau)$, on the right the independence test showing the (filter) input-residual cross-covariance $\hat{R}_{\epsilon u}(\tau)$.

- Assumption 2: $\epsilon_k(\hat{\boldsymbol{\theta}})$ has a symmetric distribution;
- Assumption 3: $\epsilon_k(\hat{\boldsymbol{\theta}})$ is independent of past inputs;
- Assumption 4: $\epsilon_k(\hat{\boldsymbol{\theta}})$ is independent of all inputs (applicable when the system is operating in open-loop);

Among these statistical tests that check for the mentioned assumptions, one of the most used thanks to its spread use in the system identification field [20, 30], is the *Residual Analysis* test which was performed on two particular model structure for the direct filter:

- An *Output-Error* (OE) model structure for the filter. The residual analysis after the estimation of the filtering parameters is shown in Fig. 7.5.
- An *AutoRegressive with eXogenous input* (ARX) model structure for the filter. The residual analysis after the estimation of the filtering parameters is shown in Fig. 7.6.

For the univariate case, the results suggest that the OE model structure is superior to the ARX model structure since:

- For the white test which tests the whiteness of the residuals, the sample auto-covariance $\hat{R}_\epsilon(\tau)$ of the residuals is bounded properly for every lags except zero, i.e. $\forall\tau \neq 0$, for the OE model structure. Instead, for the model structure ARX there is one violation at lags $|\tau| = 1$, suggesting an inaccurate model structure.
- For the independence test which tests the uncorrelated assumption between inputs and residuals, the sample cross-covariance $\hat{R}_{\epsilon u}(\tau)$ of the residuals is bounded properly for lags $\forall\tau \neq 0$, for the OE model structure.

In particular, this is the desirable output of the test for OE models, since the focus is on the independence of residuals and inputs and, not on results of the whiteness of the residuals due to the fact that the modeling focus is on the dynamics of the signal model $G(\hat{\boldsymbol{\theta}}_{\mathrm{DF}})$ and not on the disturbance properties of the noise model $H(\hat{\boldsymbol{\theta}}_{\mathrm{DF}})$. Instead, for the model structure ARX there is one violation for the negative lag $\tau = -1$, which is not particularly significant since cross-correlation between residuals and inputs for negative lags is not necessarily an indication of an inaccurate model. In particular, in this type of test the focus is concentrated on the positive lags in the cross-correlation plot during model validation (when current residuals affect future input values) which may suggest the presence of feedback in the system under analysis.

7.1.7 Operational use and filtering performance

Once the filter design step is finished, the unifying statistical-parametric framework returns the estimate $\hat{\boldsymbol{\theta}}$ as seen in previous section. Then, from the recursive filtering equation in Eq. (6.16) shared by all the solutions, substituting its appropriate estimate $\hat{\boldsymbol{\theta}}$, the filtered variable is returned. In other words, the desired filtered variable is given by means of following linear combination:

$$\hat{\boldsymbol{x}}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k \boldsymbol{C})\boldsymbol{A}\hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{K}_k \boldsymbol{y}_k \tag{7.19a}$$

$$\hat{z}_{k|k} = \hat{\boldsymbol{\theta}}\boldsymbol{\Phi}_k, \qquad \forall k = N_{\mathrm{ID}}, \cdots, N \tag{7.19b}$$

where:

- $\hat{\boldsymbol{x}}_{k|k}$ is considered as $\hat{z}_{k|k}$ since $\hat{z}_{k|k} = \boldsymbol{C}_2\hat{\boldsymbol{x}}_{k|k}$ considering the assumption $\boldsymbol{C}_2 = \boldsymbol{I}$.
- Regression data $\boldsymbol{\phi}_k = [\hat{z}^{\mathrm{T}}_{k-1|k-1}, \boldsymbol{y}^{\mathrm{T}}_k]^{\mathrm{T}}$ are taken from the validation dataset $\mathcal{D}_{\mathrm{VL}}$ ad denoted in Eq. (7.19b).
- Equation (6.16) is reduced to Eq. (7.19a) considering the assumption that there is no exogenous input: $\boldsymbol{B} = \boldsymbol{D} = \boldsymbol{0}$ and $n_{\boldsymbol{u}} = 0$.

In order to evaluate filtering performance using the validation dataset a performance fitness criterium is needed. To this end, the *Normalized Root Mean Square Error* fitness indicator, expressed in percentage, was selected as follows:

$$\left(1 - \frac{\|z_k - \hat{z}_{k|k}\|_2}{\|z_k - \bar{z}\|_2}\right) \cdot 100 \tag{7.20}$$

where:

- z_k denotes the true desired variable, in other words the ground truth. Note that in practice this knowledge is not available, but considering that the univariate example is performed in simulation this knowledge is easily accessible.
- $\bar{z}$ denotes the sample mean of the true state z_k.

In particular, for the sake of compactness, the discussion about filtering performance is focused just on experiments E_5 and E_6 since comparison differences are more emphasized in these scenarios. In details, experiments E_5 and E_6 deal with the worst-case scenario when the SNR of the measured data is the lowest among the experiments, see again Table 7.1. In particular, experiment E_5 has also the lowest number of available identification samples setting. To the end of comparison, results are depicted graphically in Fig. 7.7 where the filtering performance is expressed through the mentioned NRMSR fitness criterium. From Fig. 7.7 the following considerations can be observed when considering *intra-experiments*, namely when

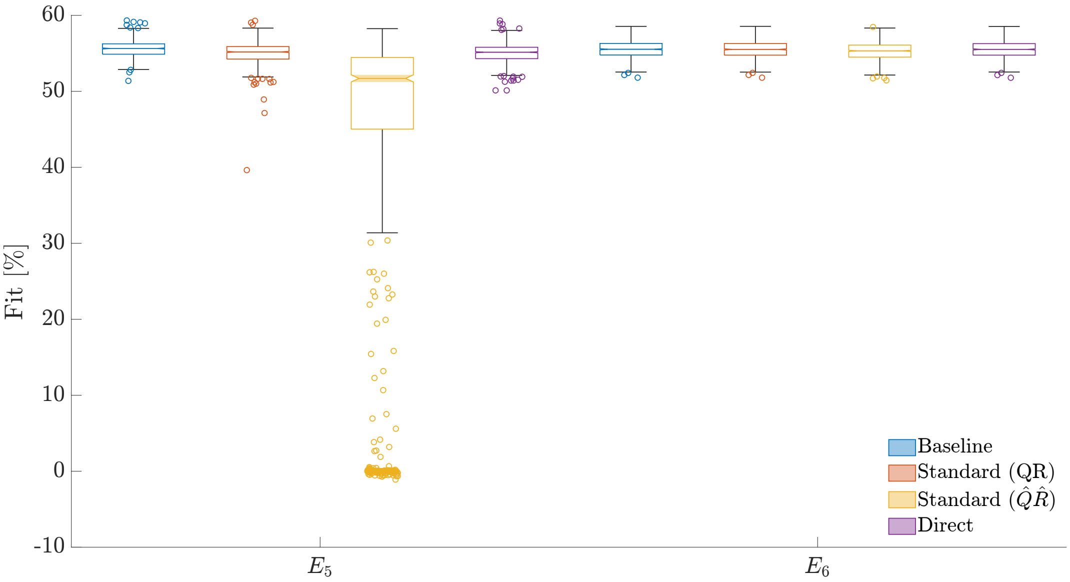


Figure 7.7: Filtering performance in the univariate case for worst-case scenario experiments E_5 and E_6. Different solutions are compared by means of the NRMSE fitness criterium: in blu the **Baseline** solution, in red the **Standard QR** solution, in yellow the **Standard $\hat{Q}\hat{R}$** solution, and finally in purple the **Direct** solution solution.

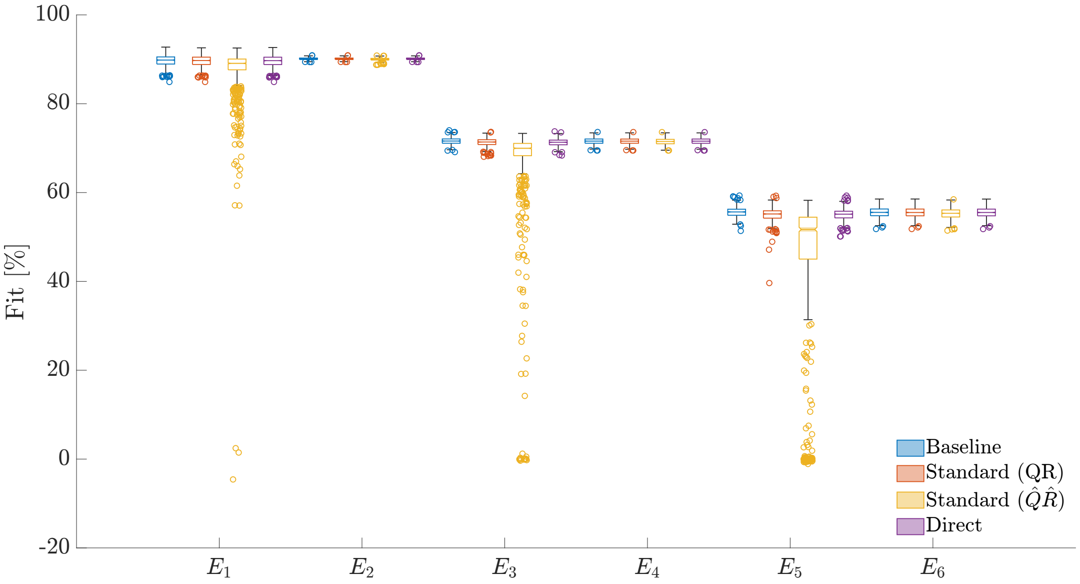


Figure 7.8: Filtering performance in the univariate case for all experiments. Different solutions are compared by means of the NRMSE fitness criterium: in blu the **Baseline** solution, in red the **Standard QR** solution, in yellow the **Standard $\hat{Q}\hat{R}$** solution, and finally in purple the **Direct** solution solution.

comparing the different solutions in the same experiment. In this case, the focus is on the worst-case scenario of experiment E_5:

- The **Baseline** solution shows the best performance, having always the best empirical distribution, as shown from the boxplot with the highest median and the lowest variance and few outliers. Since it was defined as the baseline solution due to the fact that the optimal filtering theory is applied when all the required ingredients are known, this result is as expected.
- The **Standard $\boldsymbol{QR}$** solution and the **Direct** solution perform visually the same, having similar boxplots statistics. This result suggests multiple points:
  - If the practical problem of having to estimate somehow the unknown stochastic components of the system, namely the noise covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$, was not realistic, then the **Standard $\boldsymbol{QR}$** solution could be - de facto - the standard viable solution, showing not problem whatsoever.
  - In the worst case scenario for the **Direct** solution, i.e. when there is no need to estimate the noise covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ that is one of the reasons that fueled its development, the solution performs the same as the standard solution, with respect to filtering performance.
- The **Standard $\hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$** solution, instead, shows the worst filtering performance just in experiment E_5, that is the one with the smallest number of available identification samples. In particular, the median is graphically few percentage points lower than the other solutions and both the lower inter-quantile value is much deeper than the other solutions. Moreover it shows many outlier realizations with poor filtering performance. This could be expected if observing that the **Standard $\hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$** solution is implemented by means of multiple estimation routines needed to estimate the required ingredients. Following this rationale it is expected that the overall filtering performance is deteriorated from the uncertainty imparted from the various sequential estimation routines, even when state-of-art solution are employed (see for instance the one-step DCM technique used to estimate the noise CMs in Chapter 5).

Instead, from Fig. 7.7 the following considerations can be observed when considering *inter-experiments*, namely when comparing the solutions among different experiments. In this case the focus is on experiments E_5 and E_6:

- The different filtering solutions show comparable filtering performance as denoted graphically in experiment E_6. The result suggests that the mentioned inter-experiment considerations are emphasized only when worst-case scenario settings are considered. In particular, in this case, the difference between experiments E_5 and E_6 is related to the number of available identification samples that can be used in the filter design step of the different solutions. This suggests that:
  - The **Direct** solution may prove to be more viable and effective in terms of filtering performance in the case of critical few available samples.
  - The **Standard $\hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$** performs poorly as could be expected since its solution suffers from exploiting estimation routines which are derived from

the asymptotic case scenario, that is in exact contrast when considering limited data in practice. The questioning of how much estimation variance of this solution based on the two-step approach is carried over from the sequential steps is still not answered. Anyway, insights in the inter-experiments considerations suggest that the noise CMs estimation routines are the main actor in this sense when noting the difference between the **Standard** $\hat{Q}\hat{R}$ solution and the **Standard** $\hat{\boldsymbol{Q}}\hat{\boldsymbol{R}}$ solution.

7.2 A multivariate example with exogenous input

The aim of the following section is to extend the experimental framework and the comparison of the filtering solutions from the univariate example in Section 7.1 to more complex example. In particular, the focus is on:

1. Incorporate a multivariate example, i.e. $n_x > 1$
2. Incorporate an example with exogenous input, i.e. $n_u \geq 1, \boldsymbol{B} \neq \boldsymbol{0}$

To this end, a second multivariate case with exogenous input, taken from an aerospace example of an identified system related to the longitudinal flight control, is considered as follows:

$$\begin{aligned}
\boldsymbol{A} &= \begin{bmatrix} 0.9944 & -0.1203 & -0.4302 \\ 0.0017 & 0.9902 & -0.0747 \\ 0 & 0.8187 & 0 \end{bmatrix}, & \boldsymbol{B} &= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\
\boldsymbol{C} &= \begin{bmatrix} 0 & 3 & 0 \end{bmatrix}, & \boldsymbol{D} &= \boldsymbol{0} \\
\boldsymbol{Q} &= \begin{bmatrix} 0.01 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 0.01 \end{bmatrix}, & \boldsymbol{R} &= 0.01
\end{aligned} \tag{7.21}$$

Using the same reasoning of the univariate example in Section 7.1, the considered system $\mathcal{S}$ is reduced as follows:

$$\mathcal{S}: \quad \begin{aligned}
\boldsymbol{x}_{k+1} &= \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{w}_k & \text{(7.22a)} \\
\boldsymbol{y}_k &= \boldsymbol{C}\boldsymbol{x}_k + \boldsymbol{v}_k & \text{(7.22b)} \\
\boldsymbol{z}_k &= \boldsymbol{x}_k + \boldsymbol{e}_k & \text{(7.22c)}
\end{aligned}$$

where the measurement noise $\boldsymbol{e}_k$ affecting Eq. (7.22c) is a zero-mean white process with noise covariance matrix $\boldsymbol{\Sigma_e}$ as follows:

$$\boldsymbol{e}_k \sim \mathcal{G}\left(\boldsymbol{\mu_e} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma_e} = \begin{bmatrix} 0.04 & 0 & 0 \\ 0 & 0.04 & 0 \\ 0 & 0 & 0.04 \end{bmatrix}\right) \tag{7.23}$$

Note that, for the sake of simplicity, experiments settings defined for the univariate example in Table 7.1 are no more considered. Following this rationale, the tested noise covariance matrices values are the ones in Eqs. (7.21) and (7.23). Regarding the identification dataset, just the worst-case scenario is considered, defined as follows:

- $N_{\text{ID}} = 200$ samples (considered as a low number);
- $N_{\text{VL}} = 4000$ samples;
- $N = N_{\text{ID}} + N_{\text{ID}} = 4200$ samples;

These samples form the identification dataset and the validation dataset defined in the same manner as in the univariate example in Section 7.1.2.

Moreover the control input $\boldsymbol{u}_k$ was selected to be a *multisine signal*[1] for its property of being exciting in a wide range of frequencies, thus allowing a nice-to-have input excitation for the estimation routines [116][30]. In particular the selected multisine signal has:

- A period equals to its length, i.e. $N = 4200$ samples
- A range limited by $[-0.1, 0.1]$, i.e. $-0.1 \le \boldsymbol{u}_k \le 0.1, \qquad \forall k$

7.2.1 System identification and noise covariance matrices estimates in the standard solution

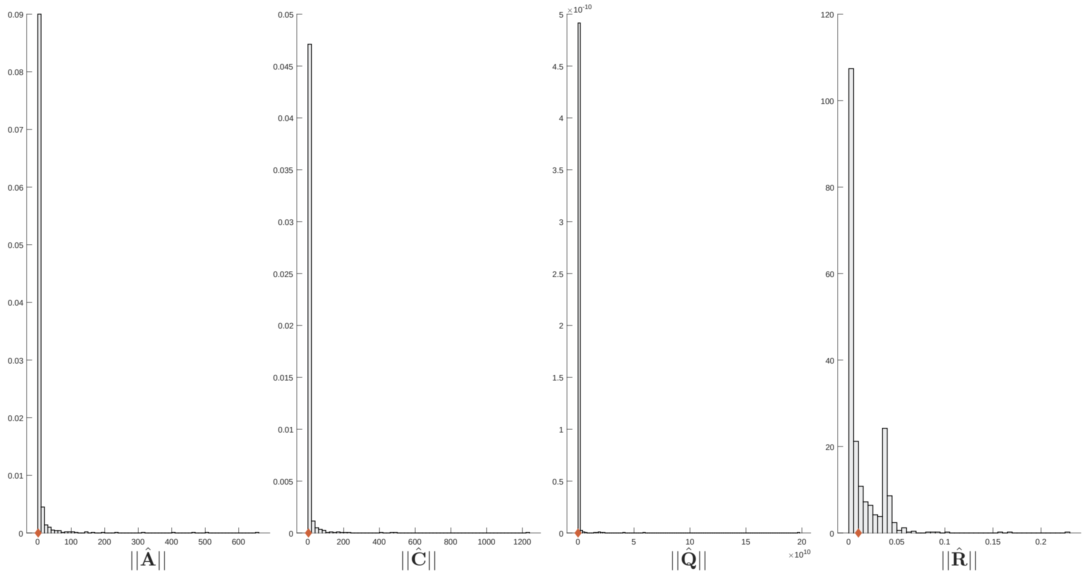


Figure 7.9: The empirical distribution of the identified system matrices $\hat{A}$, $\hat{C}$, $\hat{Q}$, and $\hat{R}$ expressed by scalar values by means of the spectral norm (2-norm) $\|\cdot\|_2$, in the multivariate case with exogenous input for the **Standard $\hat{Q}\hat{R}$** and **Standard QR** solutions.

For the sake of completeness, the two-step approach of the **Standard $\hat{Q}\hat{R}$** solution is explored in more details. In particular, the results of the system identification estimates, namely matrices $\hat{A}$ and $\hat{C}$ identified through the ad-hoc instrumental variable least square method of Section 5.1.1 and the noise covariance matrices $\hat{Q}$

[1] A multi-sine signal is a sum of a number of harmonically related sinusoids with freely adjustable amplitudes and phases. This control input signal is often used in *System Identification*.

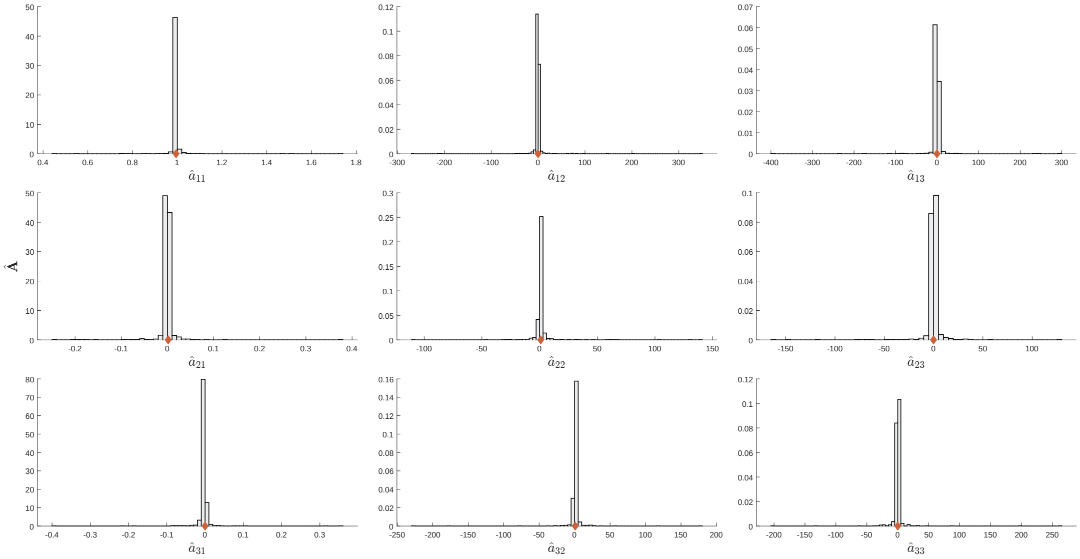


Figure 7.10: The empirical distribution of all the elements forming the identified system matrix $\hat{\boldsymbol{A}}$, with $n_x = 3$, in the multivariate case with exogenous input for the **Standard $\hat{Q}\hat{R}$** and **Standard QR** solutions.

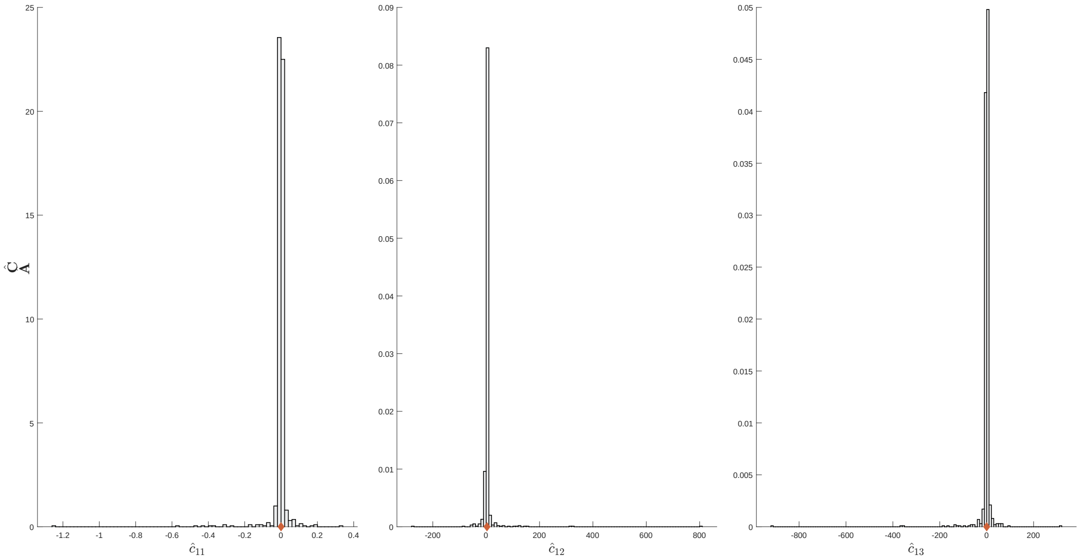


Figure 7.11: The empirical distribution of all the elements forming the identified system matrix $\hat{\boldsymbol{C}}$, with $n_y = 1$, in the multivariate case with exogenous input for the **Standard $\hat{Q}\hat{R}$** and **Standard QR** solutions.

and $\hat{\boldsymbol{R}}$ identified through the *Direct Correlation Method* (DCM) of Section 5.2.1 are shown by means of their empirical distributions through histograms in Fig. 7.9 for the multivariate example using the spectral norm to reduce matrices to scalar values. Instead, the empirical histograms of every elements of identified matrices $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{C}}$, $\hat{\boldsymbol{Q}}$, and $\hat{\boldsymbol{R}}$ is available, respectively, in Figs. 7.10 to 7.13.

In particular, few difficulties arise when dealing with the multivariate case:

- The deterministic component brought by the extension with the exogenous input, namely matrix $\boldsymbol{B}$ was no estimated in the multivariate case but was

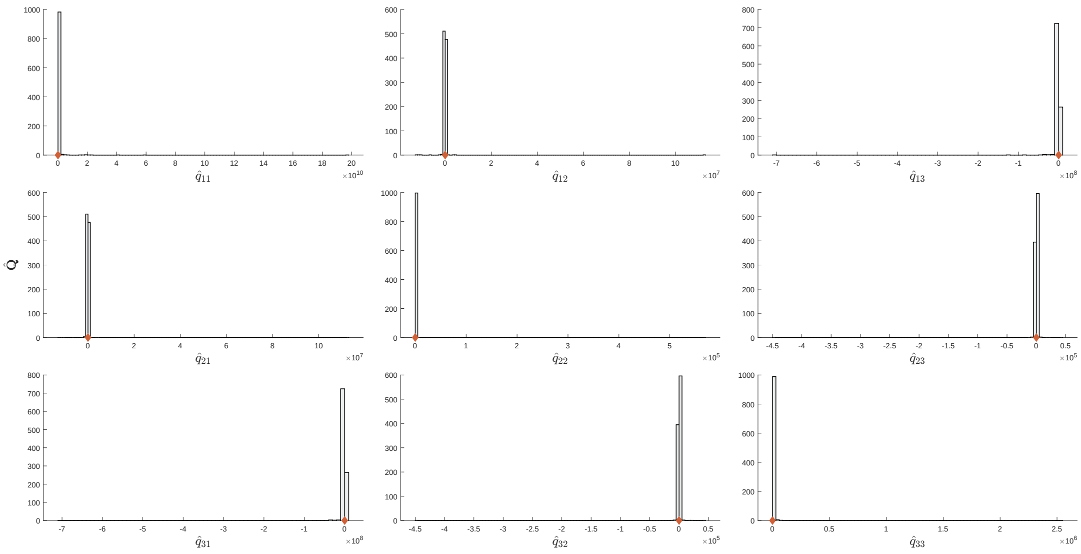


Figure 7.12: The empirical distribution of all the elements forming the identified system matrix $\hat{Q}$, with $n_x = 3$, in the multivariate case with exogenous input for the **Standard $\hat{Q}\hat{R}$** solution.

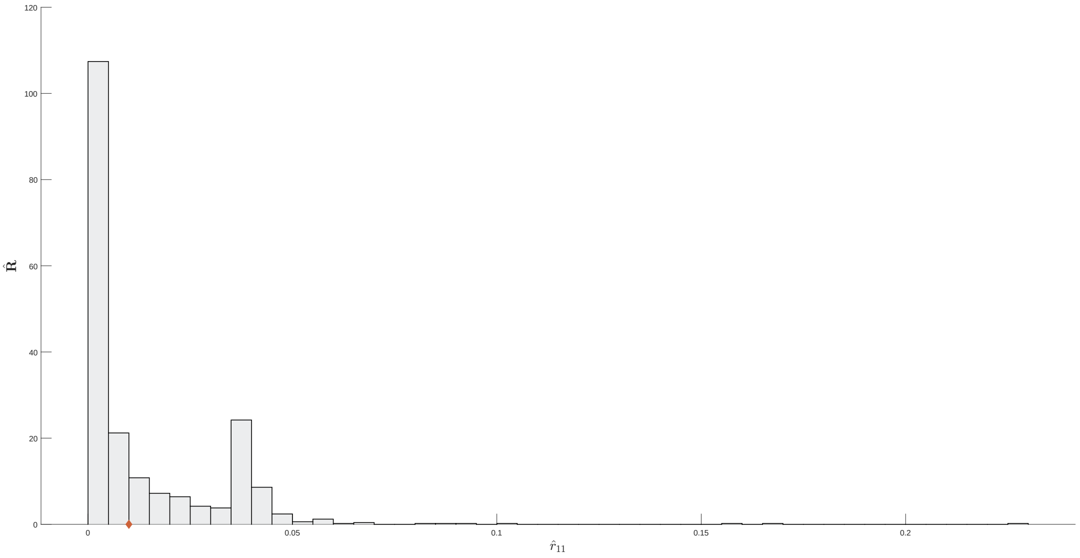


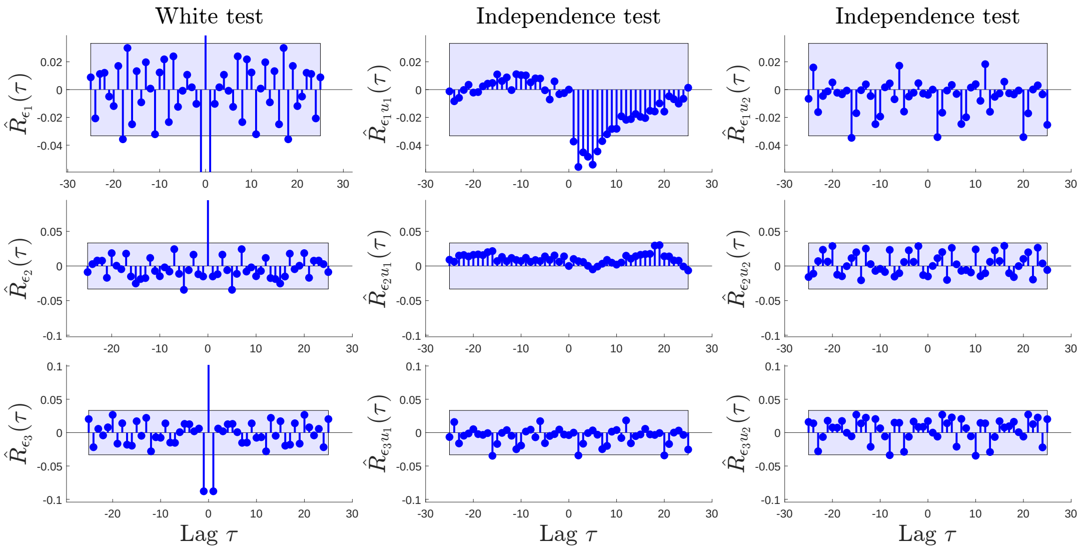
Figure 7.13: The empirical distribution of all the elements forming the identified system matrix $\hat{R}$, with $n_y = 1$, in the multivariate case with exogenous input for the **Standard $\hat{Q}\hat{R}$** solution.

used as if its value was known. The rationale is that in general estimation framework are focused on the identification of the stochastic component of the system. On the other hand, the developed ad-hoc instrumental variable least square routine used to estimate matrices $\boldsymbol{A}$ and $\boldsymbol{C}$, could not be extended easily. In this sense, a fertile field of research would be to incorporate other methods to perform system identification, with the suggestion to incorporate also sub-space method for their advantage with estimating all the system matrices in multivariate case examples.

- In general, the histogram are visualized as peaks due to the fact that there are many bad realization with outliner values, thus distorting the figure scale limits. The effect can be appreciate more on the estimates for the noise covariance matrices $\hat{Q}$, and $\hat{R}$ suggesting that the (DCM) method used for their estimation suffers in the multivariate case. The insights was researched briskly in the literature, leading to theoretical considerations about the identifiability property for the DCM method and the number of unique elements of the noise covariance matrices that can be estimated properly. For a full reference, please refer to the recent dissertation in [104]. In the considered example, the effect was not studied further, leaving fertile room for improvement both in theory and practice.

- It also possible to observe another effect more emphasized on the estimates for the noise covariance matrices $\hat{Q}$, and $\hat{R}$. In particular, the effect consists of having their empirical distribution more skewed toward the zero, as if there was a mode value around it. The insights lead to implementation detail, as discussed in Chapter 5, about the SDP constraints added to the optimization routine in order to enforce the positive semi-definite property on the covariance estimates. Following this rationale, it is believed that in the multivariate case the bound is hit more often due to the burden of having to deal with higher dimensionalities.

7.2.2 Model validation and model structure determination for the direct solution



Results of the residual analysis test for the OE model structure of the direct filtering solution in the multivariate with exogenous input example. On the left column the white test showing the residual auto-covariance $\hat{R}_{\epsilon}(\tau)$. On the middle column the independence test showing the first (output $\boldsymbol{y}_k$) input-residual cross-covariance $\hat{R}_{\epsilon u}(\tau)$. On the right column the independence test showing the second (input $\boldsymbol{u}_k$) input-residual cross-covariance $\hat{R}_{\epsilon u}(\tau)$.

Considering the multivariate case, for the sake of simplicity, only the following model-structure was tested in the **Direct** solution to the filtering design problem:

- An *Output-Error* (OE) model structure for the filter. The residual analysis after the estimation of the filtering parameters is shown in Section 7.2.2.

The choice of this structure is simple given by the fact that in the univariate example it gave better validation result when doing the residual analysis. In particular, in Section 7.2.2 are shown three columns:

- For the white test which tests the whiteness of the residuals, the sample auto-covariance $\hat{R}_\epsilon(\tau)$ of the residuals is shown on the first column. For the model structure OE there are some violations at lags $|\tau| = 1$ with respect to the first and third component of the residual vector. Since the desirable output of the test for OE models is to have independence of residuals and inputs, the observed results on the whiteness of the residuals can be overlooked due to the fact that the modeling focus is on the dynamics of the signal model $G(\hat{\boldsymbol{\theta}}_{\mathrm{DF}})$ and not on the disturbance properties of the noise model $H(\hat{\boldsymbol{\theta}}_{\mathrm{DF}})$.
- For the independence test which tests the uncorrelated assumption between inputs and residuals, the sample cross-covariance $\hat{R}_{\epsilon u}(\tau)$ of the residuals is shown in the second column, when testing the first input u_1 of the filter, that is the output of the system $\boldsymbol{y}_k$, and in the third column, when testing the second input u_2 of the filter, that is the exogenous input of the system $\boldsymbol{u}_k$. More concerning is the observation of positive lags in the cross-correlation plot during model validation (when current residuals affect future input values) which may suggest the presence of feedback in the system under analysis.

7.2.3 Operational use and filtering performance

Once the filter design step is finished, the unifying statistical-parametric framework returns the estimate $\hat{\boldsymbol{\theta}}$ as seen in previous section. Then, from the recursive filtering equation in Eq. (6.16) shared by all the solutions, substituting its appropriate estimate $\hat{\boldsymbol{\theta}}$, the filtered variable is returned. In other words, the desired filtered variable is given by means of following linear combination:

$$\hat{\boldsymbol{x}}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{C})\boldsymbol{A}\hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{K}_k\boldsymbol{y}_k + (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{C})\boldsymbol{B}\boldsymbol{u}_{k-1} \tag{7.24a}$$

$$\hat{\boldsymbol{z}}_{k|k} = \hat{\boldsymbol{\theta}}\boldsymbol{\Phi}_k, \qquad \forall k = N_{\mathrm{ID}}, \cdots, N \tag{7.24b}$$

where:

- $\hat{\boldsymbol{x}}_{k|k}$ is considered as $\hat{\boldsymbol{z}}_{k|k}$ since $\hat{\boldsymbol{z}}_{k|k} = \boldsymbol{C}_2\hat{\boldsymbol{x}}_{k|k}$ considering the assumption $\boldsymbol{C}_2 = \boldsymbol{I}$.
- Regression data $\boldsymbol{\phi}_k = [\hat{\boldsymbol{z}}_{k-1|k-1}^\top, \boldsymbol{y}_k^\top, \boldsymbol{y}_{k-1}^\top]^\top$ are taken from the validation dataset $\mathcal{D}_{\mathrm{VL}}$ ad denoted in Eq. (7.24b).
- Equation (6.16) is reduced to Eq. (7.24a) considering the assumption that there is a exogenous input: $\boldsymbol{B} \neq \boldsymbol{0}$ and $n_{\boldsymbol{u}} = 1$, also $\boldsymbol{D} = \boldsymbol{0}$.

In order to evaluate filtering performance using the validation dataset a performance fitness criterium is needed. To this end, again the *Normalized Root Mean Square Error* fitness indicator, expressed in percentage, was selected as follows:

$$\left(1 - \frac{\|z_k - \hat{z}_{k|k}\|_2}{\|z_k - \bar{z}\|_2}\right) \cdot 100 \tag{7.25}$$

where:

- z_k denotes the true desired variable, in other words the ground truth. Note that in practice this knowledge is not available, but considering that the multivariate example is performed in simulation this knowledge is easily accessible.
- $\bar{z}$ denotes the sample mean of the true state z_k.

In order to compare filtering performance, results are depicted graphically in Fig. 7.14 through the mentioned NRMSR fitness criterium. From Fig. 7.14 the following

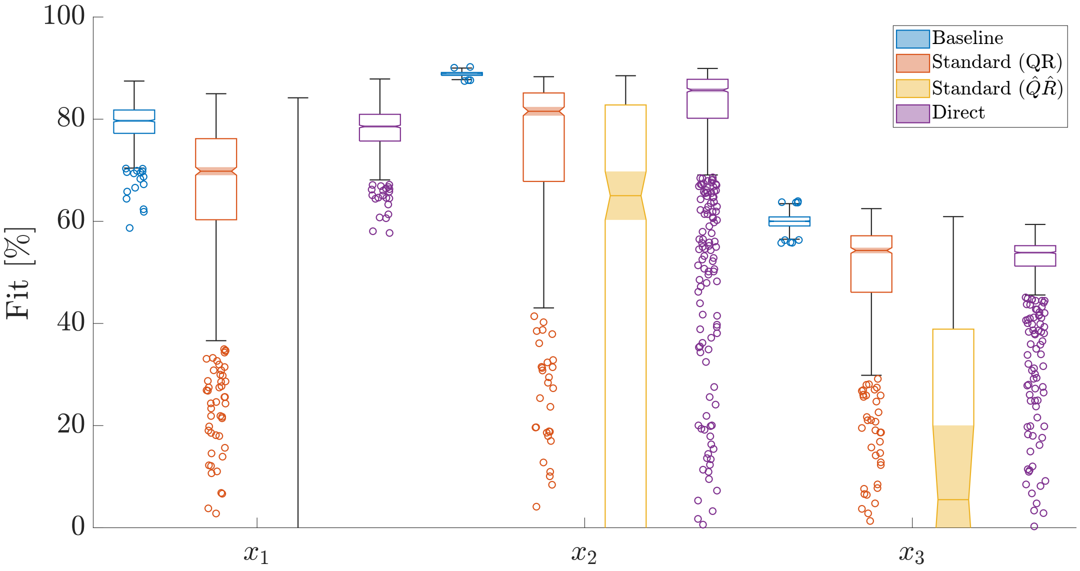


Figure 7.14: Filtering performance in the multivariate case with exogenous input. Different solutions are compared by means of the NRMSE fitness criterium: in blu the **Baseline** solution, in red the **Standard** $\boldsymbol{QR}$ solution, in yellow the **Standard** $\boldsymbol{\hat{Q}\hat{R}}$ solution, and finally in purple the **Direct** solution solution.

considerations can be observed when considering *inter-experiments*, namely when comparing the different solutions in the same experiment. In this case, the focus is on the worst-case scenario of experiment E_5:

- In general in the multivariate case the fitting performance worsen in all the filtering solutions. Anyway, with respect to the **Direct** solution, the **Standard** $\boldsymbol{\hat{Q}\hat{R}}$ has a greater decrease in performance, emphasized by the boxplots out of the scale. Moreover, also the **Standard** $\boldsymbol{QR}$, which in univariate case was comparable with the **Direct** solution in terms of performance, has now worse performance. This is interesting since it suggests that in the multivariate case the simplification brought by the **Direct** solution pays off more and has a greater advantage than in the univariate case.
- Another consideration is that the **Direct** solution, the **Standard** $\boldsymbol{\hat{Q}\hat{R}}$, and the **Standard** $\boldsymbol{QR}$ suffer all from scaling with the dimensionality. In particular, both **Standard** $\boldsymbol{\hat{Q}\hat{R}}$ and the **Standard** $\boldsymbol{QR}$ solutions have problems, as already mentioned in Section 7.2.1, in the system identification step when system matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ are estimated but also when applying the DCM method in order to estimate the stochastic components $\boldsymbol{Q}$ and $\boldsymbol{R}$ due to the numerical

problems of the routine and mentioned theoretical identifiability property difficulties. Instead, the **Direct** solution suffers when the estimation routines used to identify the chosen OE model structure (the `oe` and `tfest` MATLAB functions) has no ability to enforce a common matrix polynomial denominator in the optimization problem. To solve this problematic, the empirical tests were carried out by a different implementation of the direct filter estimation. In particular, the optimization problem in Eq. (6.7) was solved by means of the powerful `fmincon` routine which permits the specification of the required constraints.

Conclusion

This thesis proposed four different contributions to the research field of data-driven filtering design problem.

To begin with, an introduction of the standard solution based on the two-step approach is given in Chapter 4. Of this solution, the practical difficulties in estimating the noise covariance matrices, needed to properly apply the optimal Kalman filter theory, are highlighted during the system identification stage. Next, a first contribution concerns the reformulation of the solution idea into one that is developed directly, based on one-step approach, namely, the methodology of direct identification of the desired filter. Finally, the mix of these two initial works results in the introduction of a new unifying working framework limited, here, to LTI systems and their steady-state LTI filters. The special feature of the new working framework is the introduction of a new measurement equation, expressive of the fact that it is assumed possible to sample the variable of interest to be filtered for a given limited amount of time. Following this rationale, the hardships of defining a common set of settings in the unifying framework are paid off by being able to develop the mentioned direct methodology, and to compare the researched data-driven filtering solutions in a fair and common manner through experimental tests.

It is precisely in this common framework, that a second practical contribution is developed in Chapter 5. Here, the sequential steps with their requirements needed to implement the standard solution to the filtering design problem are explored in details. First, to the end of identifying the deterministic components of the system, ad-hoc instrumental variable least square routines are derived from scratch by exploiting the knowledge of the innovative measurement equation of the desired variable. The outputs of this identification task are the unbiased estimates of the deterministic components of the system, namely the matrices $\boldsymbol{A}$ and $\boldsymbol{C}$. Second, regarding the stochastic components of the system, the state-of-the-art *Direct Correlation Method* (DCM) is implemented with a slightly modification of the original implementation in order to enforce the SDP constraints on the identified covariance matrices. The results of this second estimation routine are the unbiased estimates of the stochastic components of the system, namely the matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$. The combination of the above algorithms allows the derivation of the standard solution to the filtering design problem based on a two-step approach: first the identification of a model of the system, then the optimal Kalman theory can be applied on the identified system to derive a filter for the variable of interest.

Following the development of the second contribution, a third practical contribution is developed in Chapter 6. Here, the sequential steps and the required ingredients to implement the direct solution to the filtering design problem are explored in details. First, a recursive filter equation, termed as reference equation, which makes use of input and output data samples up to the current time, is derived from the steady-state LTI Kalman filter theory. This equation is then manipulated in order to define a proper signal and noise model by exploiting the new measurement equation of the desired variable, thus exploiting the available information for a limited amount of time in the design of the filter. The design is carried out by casting a statistical-parametric optimization problem, resulting in the estimation of a linear model structure of the filter as close as possible to the derived reference equation, similar to how data-driven transfer functions are identified do date. Here the correct choice of the noise model remains an open theoretical question which is tackled down in practice by exploiting the residual analysis theory of system identification in the experimental comparison tests.

The final contribution in Chapter 7 is the natural ending of the previous commitments, that is the development of an experimental comparison to the derived data-driven filtering solutions, the standard and the direct ones, in a common LTI framework. Here, first a simple univariate example with no exogenous input is explored in details. The purpose of this example is to present a use-case where experimental insights can be observed in a simple way, thus serving as an academic tutorial or guideline. Later on, the example is extended to a more general multivariate with exogenous input case. Regarding the implementation of the routines, difficulties related to the scaling of dimensionality and to numerical problems are highlighted in various scenarios. Nonetheless, estimates of the desired variable are inferred from the identified filters and performance are evaluated by means of the NRMSE fitness criterium. In the end, results indicate that the direct data-driven solution to the filter design problem is viable and, in some cases, its filtering performance can be superior to the standard solution.

Future developments After presenting the results of each contribution, it is worth exploring also future developments in more details:

1. The first future development concerns the estimation of matrix $\boldsymbol{B}$ in the case of an exogenous input as in the mentioned multivariate example. Do date, the matrix $\boldsymbol{B}$ is not estimated but is considering known due to lack of time in investigating system identification routines that could output a proper stochastic realization of the system or extend the ad-hoc instrumental variable least-squares routines. To this end, it is suggested to investigate methods from the sub-space identification theory to check if they are a proper match, see [21, 142], and how the new measurement equation could be exploited in those frameworks.

2. More theoretical efforts and, in general, a greater theoretical background are needed in the formulation of the new direct paradigm in order to:

   (a) Refine the to-be-taken assumptions in the unifying framework and support their importance with theoretical insights.

   (b) Evaluate whether the new methodology of the direct solution can be extended by relaxing some assumptions. For instance, assuming one of the following: $\boldsymbol{C}_2 \neq \boldsymbol{I}$, $\boldsymbol{G} \neq \boldsymbol{I}$, and $\boldsymbol{D} \neq \boldsymbol{0}$.

(c) Evaluate whether the new methodology of the direct solution can be extended by considering generic filters beyond steady-state filters.

(d) Evaluate whether the new methodology of the direct solution can be extended to dynamical systems beyond LTI systems.

(e) Derive more fruitful considerations for a potential guided choice of the linear model structure and model complexity of the direct filter with respect to the noise model.

3. Practical and theoretical contributions to the direct solution could be investigated in order to evaluate the possibility of developing an ad-hoc dedicated estimation routine when considering the multivariate case. These efforts are needed to:

   (a) tackle down the implementation burden of enforcing the constraints of the common denominator in the model structure of the filter (to date, MATLAB routines `oe` and `tfest` do not permit to cast these constraints)

   (b) carry out the estimation routine for the design of the filter efficiently from a numerical point of view. To date, the dedicated routine to estimate the direct filter in the multivariate case uses the `fmincon` MATLAB function without fine-tuning its configurable parameters. A scouting of other available optimization routines is recommended.

4. Moreover, it is clear that the standard solution is a grouping of different kind of routines developed from different theoretical frameworks. In other words, the work of implementing the standard solution and, at the same time, solve for its many practical problems is far from being answered. Thus, a fertile field of study is found in the state-of-the-art noise CMs estimation routines where both practical and theoretical results should be researched more. To this end, the author suggest the reference in [104] and the review in [72] as starting points. In particular, in [104] theoretical insights are explored: the identifiability property and the difficulty of estimating all unique elements of the covariance matrices in multivariate example. On the other hand, practical efforts on enforcing a PSD programming to the DCM method is analyzed in [50, 137, 144]

5. Finally, the experimental comparison could be enhanced by validating the empirical results and insights on different set of examples. Also, in hindsight, a more careful choice of experimental settings should be wanted, in order to simplify the computation and focus on the interesting. To this end, the author suggests to drop the discrete grid of SNR and number of available samples for the design of the filter and work only with a mono-dimensional continuous grid related to the number of available samples.

Bibliography

Book references

[1] B. D. O. Anderson and J. B. Moore. *Optimal filtering.* Republication of the work originally published by Prentice-Hall, Inc., Englewood Cliffs, New Jersey, in 1979. Mineola, N.Y: Dover Publications, 2005. 368 pp. ISBN: 9780486439389.

[2] B. D. O. Anderson and S. Vongpanitlerd. *Network Analysis and Synthesis: A Modern Systems Theory Approach.* 1973. 548 pp. ISBN: 9780486453576.

[3] A. C. Antoulas, ed. *Mathematical System Theory.* Springer Berlin Heidelberg, 1991. DOI. URL.

[4] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models.* Springer Series in Statistics, Springer New York, Jan. 26, 2007. 653 pp. ISBN: 978-0-387-40264-2. URL.

[5] W. Cauer. *Siebschaltungen: Hrsg. mit unterstützung des Elektrotechnischen vereins ev, Berlin.* VDI-verlag gmbh, 1931.

[6] C. K. Chui and G. Chen. *Kalman Filtering.* Springer International Publishing, 2017. DOI. URL.

[7] A. I. Dale. *A History of Inverse Probability.* Springer New York, 1999. DOI. URL.

[8] P. Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* SPRINGER NATURE, Mar. 15, 2004. 556 pp. ISBN: 0387202684. DOI. URL.

[9] P. Del Moral. *Mean Field Simulation for Monte Carlo Integration.* Chapman and Hall/CRC, May 2013. ISBN: 9780429086960. DOI. URL.

[10] G. Evensen. *Data Assimilation.* Springer Berlin Heidelberg, Aug. 27, 2009. 332 pp. ISBN: 3642037100. URL.

[11] W. Gersch and G. Kitagawa. *Smoothness Priors Analysis of Time Series.* Springer New York, Aug. 9, 1996. 276 pp. ISBN: 0387948198. DOI. URL.

[12] B. Gold and C. M. Rader. *Digital processing of signals.* [Reprint. Originally published in New York by McGraw-Hill, 1969]. Malabar, Fla: Krieger, 1983. 269 pp. ISBN: 9780898745481.

[13] A. Graham. *Kronecker products and matrix calculus with applications.* [Reprint. Originally published in Chichester by Ellis Horwood Ltd, 1981]. Mineola, New York: Dover Publications, 2018. ISBN: 9780486824178.

[14] A. H. Jazwinski. *Stochastic Processes and Filtering Theory.* [Reprint of the Academic Press, New York, 1970 edition.] Dover Publications, Nov. 1, 2007. 376 pp. ISBN: 0486462749.

[15] T. Kailath. *Linear Systems.* Information and System Sciences Series. Prentice-Hall, 1980. ISBN: 9780135369616.

[16] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear estimation.* Upper Saddle River, N.J: Prentice Hall, 2000. ISBN: 9780130224644. URL.

[17] G. Kallianpur. *Stochastic Filtering Theory.* Springer New York, Sept. 1, 1980. 318 pp. ISBN: 038790445X. DOI. URL.

[18] R. Labbe. *Kalman and Bayesian Filters in Python.* 2018. URL.

[19] R. C. K. Lee. *Optimal Estimation, Identification, and Control.* The MIT Press, 1964. 152 pp. ISBN: 9780262120128. URL.

[20] L. Ljung. *System Identification: Theory for the User.* PRENTICE HALL, Dec. 31, 1998. 640 pp. ISBN: 0136566952.

[21] P. V. Overschee and B. D. Moor. *Subspace Identification for Linear Systems.* Springer US, Oct. 8, 2011. 272 pp. ISBN: 978-0-7923-9717-5. DOI. URL.

[22] L. R. Rabiner and B. Gold. *Theory and application of digital signal processing.* Englewood Cliffs, N.J: Prentice-Hall, 1975. ISBN: 9780139141010.

[23] J. B. Rawlings, D. Q. Mayne, and M. M. Diehl. *Model predictive control : theory, computation, and design.* Santa Barbara, California: Nob Hill Publishing, 2020. ISBN: 9780975937754.

[24] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications.* ARTECH HOUSE INC, Jan. 1, 2004. 299 pp. ISBN: 9781580536318. URL.

[25] S. Särkkä. *Bayesian Filtering and Smoothing.* Cambridge University Press, Nov. 1, 2017. 256 pp. ISBN: 110703065X. URL.

[26] S. M. Savaresi et al. *Semi-Active Suspension Control Design for Vehicles.* Elsevier Science & Techn., Aug. 13, 2010. 240 pp.

[27] S. R. Searle and A. I. Khuri. *Matrix Algebra Useful for Statistics.* WILEY, May 1, 2017. 512 pp. ISBN: 1118935144. URL.

[28] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications.* Springer International Publishing, 2017. ISBN: 978-3-319-52451-1. DOI. URL.

[29] D. Simon. *Optimal State Estimation.* John Wiley & Sons, June 12, 2006. 554 pp. ISBN: 0471708585. URL.

[30] T. Söderström and P. Stoica. *System identification.* [Reprint. Originally published in New Yorkk by Prentice Hall, 1989]. Uppsala University, Sweden: Prentice Hall, 2001. ISBN: 0138812365. URL.

[31] U. Spagnolini. *Statistical Signal Processing in Engineering.* John Wiley & Sons, Ltd, Jan. 2018. ISBN: 9781119293972. DOI.

[32] P. Swerling. *A proposed stagewise differential correction procedure for satellite tracking and prediciton.* Rand Corporation, 1958.

[33] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, With Engineering Applications.* Originally appears in 1942 as a classified National Defense Research Council Report and first published in 1949. Cambridge: The MIT Press, 1975. ISBN: 9780262230025. URL.

Article references

[34] M. F. Abdel-Hafez. "On the GPS/IMU sensors' noise estimation for enhanced navigation integrity". In: *Mathematics and Computers in Simulation* 86 (Dec. 2012), pp. 101–117. DOI. URL.

[35] H. Akaike. "Stochastic theory of minimal realization". In: *IEEE Transactions on Automatic Control* 19.6 (Dec. 1974), pp. 667–674. DOI. URL.

[36] B. M. Åkesson et al. "A generalized autocovariance least-squares method for Kalman filter tuning". In: *Journal of Process Control* 18.7-8 (Aug. 2008), pp. 769–779. DOI.

[37] D. F. Allinger and S. K. Mitter. "New results on the innovations problem for non-linear filtering". In: *Stochastics* 4.4 (Jan. 1981), pp. 339–348. DOI.

[38] D. Alspach. "Comments on "On the identification of variances and adaptive Kalman filtering"". In: *IEEE Transactions on Automatic Control* 17.6 (Dec. 1972), pp. 843–845. DOI. URL.

[39] B. N. Alsuwaidan, J. L. Crassidis, and Y. Cheng. "Generalized Multiple-Model Adaptive Estimation using an Autocorrelation Approach". In: *IEEE Transactions on Aerospace and Electronic Systems* 47.3 (July 2011), pp. 2138–2152. DOI. URL.

[40] E. N. Anagnostou and W. F. Krajewski. "Real-Time Radar Rainfall Estimation. Part I: Algorithm Formulation". In: *Journal of Atmospheric and Oceanic Technology* 16.2 (Feb. 1999), pp. 189–197. DOI. URL.

[41] B. D. O. Anderson and J. B. Moore. "The Kalman-Bucy Filter as a True Time-Varying Wiener Filter". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-1.2 (Apr. 1971), pp. 119–128. DOI. URL.

[42] T. J. Arnold and J. B. Rawlings. "Uniqueness Conditions for ALS Problems". In: *IFAC-PapersOnLine* 51.20 (2018), pp. 469–474. DOI. URL.

[43] K.-J. Åström and B. Torsten. “Numerical Identification of Linear Dynamic Systems from Normal Operating Records”. In: *IFAC Proceedings Volumes* 2.2 (Sept. 1965), pp. 96–111. DOI.

[44] P. Axelsson et al. “ML Estimation of Process Noise Variance in Dynamic Systems”. In: *IFAC Proceedings Volumes* 44.1 (Jan. 2011), pp. 5609–5614. DOI. URL.

[45] G. A. Barnard and T. Bayes. “Studies in the History of Probability and Statistics: IX. Thomas Bayes’s Essay Towards Solving a Problem in the Doctrine of Chances”. In: *Biometrika* 45.3/4 (Dec. 1958), p. 293. DOI. URL.

[46] V. A. Bavdekar, A. P. Deshpande, and S. C. Patwardhan. “Identification of process and measurement noise covariance for state and parameter estimation using extended Kalman filter”. In: *Journal of Process Control* 21.4 (Apr. 2011), pp. 585–601. DOI. URL.

[47] P. R. Bélanger. “Estimation of noise covariance matrices for a linear timevarying stochastic process”. In: *Automatica* 10.3 (May 1974), pp. 267–275. DOI. URL.

[48] V. E. Beneš. “Exact finite-dimensional filters for certain diffusions with nonlinear drift”. In: *Stochastics* 5.1-2 (June 1981), pp. 65–92. DOI.

[49] T. Berry and T. Sauer. “Adaptive ensemble Kalman filtering of non-linear systems”. In: *Tellus A: Dynamic Meteorology and Oceanography* 65.1 (July 2013), p. 20331. DOI. URL.

[50] F. Bianchi, S. Formentin, and L. Piroddi. “Process noise covariance estimation via stochastic approximation”. In: *International Journal of Adaptive Control and Signal Processing* 34.1 (Nov. 2019), pp. 63–76. DOI.

[51] J. Blanchard. “The History of Electrical Resonance”. In: *Bell System Technical Journal* 20.4 (Oct. 1941), pp. 415–433. DOI.

[52] H. W. Bode and C. E. Shannon. “A Simplified Derivation of Linear Least Square Smoothing and Prediction Theory”. In: *Proceedings of the IRE* 38.4 (Apr. 1950), pp. 417–425. DOI.

[53] J. Brewer. “Kronecker products and matrix calculus in system theory”. In: *IEEE Transactions on Circuits and Systems* 25.9 (Sept. 1978), pp. 772–781. DOI. URL.

[54] R. S. Bucy and K. D. Senne. “Digital synthesis of non-linear filters”. In: *Automatica* 7.3 (May 1971), pp. 287–298. DOI.

[55] S. Butterworth. “On the theory of filter amplifiers”. In: *Wireless Engineer* 7.6 (1930), pp. 536–541.

[56] G. A. Campbell. “Physical Theory of the Electric Wave-Filter”. In: *Bell System Technical Journal* 1.2 (Nov. 1922), pp. 1–32. DOI.

[57] M. C. Campi, A. Lecchini, and S. M. Savaresi. “Virtual reference feedback tuning: a direct method for the design of feedback controllers”. In: *Automatica* 38.8 (Aug. 2002), pp. 1337–1346. DOI.

[58] M. Canale, M. Milanese, and C. Novara. “Semi-Active Suspension Control Using “Fast” Model-Predictive Techniques”. In: *IEEE Transactions on Control Systems Technology* 14.6 (Nov. 2006), pp. 1034–1046. DOI.

[59] Z. Chen. “Bayesian filtering: From Kalman filters to particle filters, and beyond”. In: *Statistics: A Journal of Theoretical and Applied Statistics* 182.1 (2003), pp. 1–69. URL.

[60] Y. Chien and K. Fu. “On Bayesian Learning and Stochastic Approximation”. In: *IEEE Transactions on Systems Science and Cybernetics* 3.1 (1967), pp. 28–38. DOI. URL.

[61] A. Chiuso and G. Picci. “Consistency analysis of some closed-loop subspace identification methods”. In: *Automatica* 41.3 (Mar. 2005), pp. 377–391. DOI. URL.

[62] N. L. C. Chui and J. M. Maciejowski. “Realization of stable models with subspace methods”. In: *Automatica* 32.11 (Nov. 1996), pp. 1587–1595. DOI.

[63] R. Daley. “Estimating Model-Error Covariances for Application to Atmospheric Data Assimilation”. In: *Monthly Weather Review* 120.8 (Aug. 1992), pp. 1735–1746. DOI. URL.

[64] S. Darlington. “Synthesis of Reactance 4-Poles Which Produce Prescribed Insertion Loss Characteristics: Including Special Applications To Filter Design”. In: *Journal of Mathematics and Physics* 18.1-4 (Apr. 1939), pp. 257–353. DOI.

[65] F. Daum. “Exact finite-dimensional nonlinear filters”. In: *IEEE Transactions on Automatic Control* 31.7 (July 1986), pp. 616–622. DOI.

[66] D. Dee et al. “An efficient algorithm for estimating noise covariances in distributed systems”. In: *IEEE Transactions on Automatic Control* 30.11 (Nov. 1985), pp. 1057–1065. DOI. URL.

[67] D. P. Dee. “On-line Estimation of Error Covariance Parameters for Atmospheric Data Assimilation”. In: *Monthly Weather Review* 123.4 (Apr. 1995), pp. 1128–1145. DOI. URL.

[68] P. Del Moral. “Nonlinear filtering: Interacting particle resolution”. In: *Comptes Rendus de l’Académie des Sciences - Series I - Mathematics* 325.6 (Sept. 1997), pp. 653–658. DOI.

[69] A. Doucet and A. M. Johansen. “A Tutorial on Particle Filtering and Smoothing: Fifteen years later”. Version 1.1. In: *Oxford Handbook of Nonlinear Filtering, Oxford University Press* (2011).

[70] J. Dunik, O. Straka, and M. Simandl. “On Autocovariance Least-Squares Method for Noise Covariance Matrices Estimation”. In: *IEEE Transactions on Automatic Control* 62.2 (Feb. 2017), pp. 967–972. DOI. URL.

[71] J. Duník, O. Straka, and M. Šimandl. “Estimation of Noise Covariance Matrices for Linear Systems with Nonlinear Measurements”. In: *IFAC-PapersOnLine* 48.28 (2015), pp. 1130–1135. DOI. URL.

[72] J. Duník et al. “Noise covariance matrices in state-space models: A survey and comparison of estimation methods-Part I”. In: *International Journal of Adaptive Control and Signal Processing* 31.11 (May 2017), pp. 1505–1543. DOI. URL.

[73] Z. M. Durovic and B. D. Kovacevic. “Robust estimation with unknown noise statistics”. In: *IEEE Transactions on Automatic Control* 44.6 (June 1999), pp. 1292–1296. DOI. URL.

[74] G. Evensen. “Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics”. In: *Journal of Geophysical Research* 99.C5 (1994), p. 10143. DOI.

[75] S. Fekri, M. Athans, and A. Pascoal. “Issues, progress and new results in robust adaptive control”. In: *International Journal of Adaptive Control and Signal Processing* 20.10 (2006), pp. 519–579. DOI. URL.

[76] B. Feng et al. “Kalman Filter With Recursive Covariance Estimation - Sequentially Estimating Process Noise Covariance”. In: *IEEE Transactions on Industrial Electronics* 61.11 (Nov. 2014), pp. 6253–6263. DOI. URL.

[77] M. Frei and H. R. Künsch. “Sequential State and Observation Noise Covariance Estimation Using Combined Ensemble Kalman and Particle Filters”. In: *Monthly Weather Review* 140.5 (May 2012), pp. 1476–1495. DOI. URL.

[78] B. Friedland. “Estimating Noise Variances by Using Multiple Observers”. In: *IEEE Transactions on Aerospace and Electronic Systems* AES-18.4 (July 1982), pp. 442–448. DOI. URL.

[79] S. Gao et al. “Intelligent State Estimation for Continuous Fermenters Using Variational Bayesian Learning”. In: *IEEE Transactions on Industrial Informatics* 17.12 (Dec. 2021), pp. 8429–8437. DOI. URL.

[80] M. Ge and E. C. Kerrigan. “Noise Covariance Estimation for Time-varying and Nonlinear Systems”. In: *IFAC Proceedings Volumes* 47.3 (2014), pp. 9545–9550. DOI. URL.

[81] M. Gevers and T. Kailath. “An innovations approach to least-squares estimation–Part VI: Discrete-time innovations representations and recursive estimation”. In: *IEEE Transactions on Automatic Control* 18.6 (Dec. 1973), pp. 588–600. DOI. URL.

[82] M. Gevers. “A personal view of the development of system identification: A 30-year journey through an exciting field”. In: *IEEE Control Systems* 26.6 (Dec. 2006), pp. 93–105. DOI. URL.

[83] M. Gevers. “Identification for Control: From the Early Achievements to the Revival of Experiment Design”. In: *European Journal of Control* 11.4-5 (Jan. 2005), pp. 335–352. DOI. URL.

[84] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. “Novel approach to nonlinear / non-Gaussian Bayesian state estimation”. In: *IEE Proceedings F Radar and Signal Processing* 140.2 (1993), p. 107. DOI.

[85] P.-O. Gutman and M. Velger. “Tracking targets using adaptive Kalman filtering”. In: *IEEE Transactions on Aerospace and Electronic Systems* 26.5 (1990), pp. 691–699. DOI. URL.

[86] J. E. Handschin and D. Q. Mayne. “Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering”. In: *International Journal of Control* 9.5 (May 1969), pp. 547–559. DOI.

[87] C. Hilborn and D. Lainiotis. “Optimal Estimation in the Presence of Unknown Parameters”. In: *IEEE Transactions on Systems Science and Cybernetics* 5.1 (1969), pp. 38–43. DOI. URL.

[88] R. Hildebrand et al. “Prefiltering in Iterative Feedback Tuning: Optimization of the Prefilter for Accuracy”. In: *IEEE Transactions on Automatic Control* 49.10 (Oct. 2004), pp. 1801–1805. DOI. URL.

[89] H. Hjalmarsson. “Iterative feedback tuning—an overview”. In: *International Journal of Adaptive Control and Signal Processing* 16.5 (2002), pp. 373–395. DOI. URL.

[90] B. L. Ho and R. E. Kalman. “Editorial: Effective construction of linear state-variable models from input/output functions”. In: *Automatisierungstechnik* 14.1-12 (Dec. 1966), pp. 545–548. DOI. URL.

[91] Y. Ho and R. Lee. “A Bayesian approach to problems in stochastic estimation and control”. In: *IEEE Transactions on Automatic Control* 9.4 (Oct. 1964), pp. 333–339. DOI. URL.

[92] A. Jazwinski. “Filtering for nonlinear dynamical systems”. In: *IEEE Transactions on Automatic Control* 11.4 (Oct. 1966), pp. 765–766. DOI. URL.

[93] S. Julier, J. Uhlmann, and H. F. Durrant-Whyte. “A new method for the nonlinear transformation of means and covariances in filters and estimators”. In: *IEEE Transactions on Automatic Control* 45.3 (Mar. 2000), pp. 477–482. DOI.

[94] T. Kailath. “A view of three decades of linear filtering theory”. In: *IEEE Transactions on Information Theory* 20.2 (Mar. 1974), pp. 146–181. DOI. URL.

[95] T. Kailath. “An innovations approach to least-squares estimation–Part I: Linear filtering in additive white noise”. In: *IEEE Transactions on Automatic Control* 13.6 (Dec. 1968), pp. 646–655. DOI. URL.

[96] T. Kailath. “The innovations approach to detection and estimation theory”. In: *Proceedings of the IEEE* 58.5 (1970), pp. 680–695. DOI. URL.

[97] T. Kailath and P. Frost. “An innovations approach to least-squares estimation–Part II: Linear smoothing in additive white noise”. In: *IEEE Transactions on Automatic Control* 13.6 (Dec. 1968), pp. 655–660. DOI. URL.

[98] R. E. Kálmán. “A New Approach to Linear Filtering and Prediction Problems”. In: *Journal of Basic Engineering* 82.1 (Mar. 1960), pp. 35–45. DOI. URL.

[99] R. E. Kálmán and R. S. Bucy. “New Results in Linear Filtering and Prediction Theory”. In: *Journal of Basic Engineering* 83.1 (Mar. 1961), pp. 95–108. DOI. URL.

[100] R. Kashyap. “Maximum likelihood identification of stochastic linear systems”. In: *IEEE Transactions on Automatic Control* 15.1 (Feb. 1970), pp. 25–34. DOI. URL.

[101] T. Katayama and G. Picci. “Realization of stochastic systems with exogenous inputs and subspace identification methods”. In: *Automatica* 35.10 (Oct. 1999), pp. 1635–1652. DOI. URL.

[102] G. Kitagawa. “Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models”. In: *Journal of Computational and Graphical Statistics* 5.1 (Mar. 1996), pp. 1–25. DOI.

[103] A. N. Kolmogorov. “Stationary Sequences in Hilbert Space”. In: *Moscow University Mathematics Bulletin* 2 (6 1941), pp. 1–40.

[104] O. Kost, J. Duník, and O. Straka. “Identifiability of Unique Elements of Noise Covariances in State-Space Models”. In: *IFAC-PapersOnLine* 54.7 (2021), pp. 316–321. DOI. URL.

[105] H. J. Kushner. “Dynamical equations for optimal nonlinear filtering”. In: *Journal of Differential Equations* 3.2 (Apr. 1967), pp. 179–190. DOI.

[106] H. J. Kushner. “On the Differential Equations Satisfied by Conditional Probablitity Densities of Markov Processes, with Applications”. In: *Journal of the Society for Industrial and Applied Mathematics Series A Control* 2.1 (Jan. 1964), pp. 106–119. DOI.

[107] D. Lainiotis. “Optimal adaptive estimation: Structure and parameter adaption”. In: *IEEE Transactions on Automatic Control* 16.2 (Apr. 1971), pp. 160–170. DOI. URL.

[108] J. Leathrum. “On sequential estimation of state noise variances”. In: *IEEE Transactions on Automatic Control* 26.3 (June 1981), pp. 745–746. DOI. URL.

[109] T. Lee. “A direct approach to identify the noise covariances of Kalman filtering”. In: *IEEE Transactions on Automatic Control* 25.4 (Aug. 1980), pp. 841–842. DOI. URL.

[110] C. Leondes, J. Peller, and E. Stear. “Nonlinear Smoothing Theory”. In: *IEEE Transactions on Systems Science and Cybernetics* 6.1 (1970), pp. 63–71. DOI. URL.

[111] N. Levinson. “The Wiener (Root Mean Square) Error Criterion in Filter Design and Prediction”. In: *Journal of Mathematics and Physics* 25.1-4 (Apr. 1946), pp. 261–278. DOI.

[112] X. R. Li and Y. Bar-Shalom. “A recursive multiple model approach to noise identification”. In: *IEEE Transactions on Aerospace and Electronic Systems* 30.3 (July 1994), pp. 671–684. DOI. URL.

[113] F. V. Lima and J. B. Rawlings. “Nonlinear stochastic modeling to improve state estimation in process monitoring and control”. In: *AIChE Journal* 57.4 (May 2010), pp. 996–1007. DOI. URL.

[114] F. V. Lima et al. “Covariance and State Estimation of Weakly Observable Systems: Application to Polymerization Processes”. In: *IEEE Transactions on Control Systems Technology* 21.4 (July 2013), pp. 1249–1257. DOI. URL.

[115] T.-t. Lin and S. Yau. “Bayesian Approach to the Optimization of Adaptive Systems”. In: *IEEE Transactions on Systems Science and Cybernetics* 3.2 (1967), pp. 77–85. DOI. URL.

[116] L. Ljung. “Convergence analysis of parametric identification methods”. In: *IEEE Transactions on Automatic Control* 23.5 (Oct. 1978), pp. 770–783. DOI. URL.

[117] T. Ma et al. “Estimation of time series noise covariance using correlation technology”. In: *Journal of Control Theory and Applications* 9.2 (May 2011), pp. 165–170. DOI. URL.

[118] C. Magnant et al. “Bayesian non-parametric methods for dynamic state-noise covariance matrix estimation: Application to target tracking”. In: *Signal Processing* 127 (Oct. 2016), pp. 135–150. DOI. URL.

[119] H. Martz and G. H. Born. “Empirical Bayes Estimation of Observation Error Variances in Linear Systems”. In: *AIAA Journal* 9.6 (June 1971), pp. 1183–1187. DOI. URL.

[120] P. Matisko and V. Havlena. “Noise covariance estimation for Kalman filter tuning using Bayesian approach and Monte Carlo”. In: *International Journal of Adaptive Control and Signal Processing* 27.11 (Dec. 2012), pp. 957–973. DOI. URL.

[121] P. Matisko and V. Havlena. “Noise covariances estimation for Kalman filter tuning”. In: *IFAC Proceedings Volumes* 43.10 (2010), pp. 31–36. DOI. URL.

[122] R. J. McAulay and E. Denlinger. “A Decision - Directed Adaptive Tracker”. In: *IEEE Transactions on Aerospace and Electronic Systems* AES-9.2 (Mar. 1973), pp. 229–236. DOI. URL.

[123] R. Mehra. “On the identification of variances and adaptive Kalman filtering”. In: *IEEE Transactions on Automatic Control* 15.2 (Apr. 1970), pp. 175–184. DOI.

[124] N. Metropolis and S. Ulam. “The Monte Carlo Method”. In: *Journal of the American Statistical Association* 44.247 (Sept. 1949), pp. 335–341. DOI.

[125] M. Milanese and C. Novara. “Set Membership methods in identification, prediction and filtering of nonlinear systems”. In: *IFAC Proceedings Volumes* 42.10 (2009), pp. 263–272. DOI.

[126] M. Milanese and C. Novara. “Unified Set Membership theory for identification, prediction and filtering of nonlinear systems”. In: *Automatica* 47.10 (Oct. 2011), pp. 2141–2151. DOI.

[127] M. Milanese et al. “NONLINEAR VIRTUAL SENSORS DESIGN FROM DATA”. In: *IFAC Proceedings Volumes* 39.1 (2006), pp. 576–581. DOI.

[128] M. Milanese et al. “The filter design from data (FD2) problem: Nonlinear Set Membership approach”. In: *Automatica* 45.10 (Oct. 2009), pp. 2350–2357. DOI.

[129] M. Milanese and C. Novara. “Set Membership identification of nonlinear systems”. In: *Automatica* 40.6 (June 2004), pp. 957–975. DOI.

[130] A. Moghaddamjoo and R. L. Kirlin. “Robust adaptive Kalman filtering with unknown inputs”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.8 (1989), pp. 1166–1175. DOI. URL.

[131] K. Myers and B. Tapley. “Adaptive sequential estimation with unknown noise statistics”. In: *IEEE Transactions on Automatic Control* 21.4 (Aug. 1976), pp. 520–523. DOI. URL.

[132] G. Noriega and S. Pasupathy. “Adaptive estimation of noise covariance matrices in real-time preprocessing of geophysical data”. In: *IEEE Transactions on Geoscience and Remote Sensing* 35.5 (1997), pp. 1146–1159. DOI. URL.

[133] C. Novara, F. Ruiz, and M. Milanese. “Direct design of optimal filters from data”. In: *IFAC Proceedings Volumes* 41.2 (2008), pp. 462–467. DOI.

[134] C. Novara, F. Ruiz, and M. Milanese. “Direct Filtering: A New Approach to Optimal Filter Design for Nonlinear Systems”. In: *IEEE Transactions on Automatic Control* 58.1 (Jan. 2013), pp. 86–99. DOI. URL.

[135] C. Novara et al. “The filter design from data (FD2) problem: parametric-statistical approach”. In: *International Journal of Robust and Nonlinear Control* 22.16 (Sept. 2011), pp. 1853–1872. DOI. URL.

[136] B. J. Odelson, A. Lutz, and J. B. Rawlings. “The autocovariance least-squares method for estimating covariances: application to model-based control of chemical reactors”. In: *IEEE Transactions on Control Systems Technology* 14.3 (May 2006), pp. 532–540. DOI. URL.

[137] B. J. Odelson, M. R. Rajamani, and J. B. Rawlings. “A new autocovariance least-squares method for estimating noise covariances”. In: *Automatica* 42.2 (Feb. 2006), pp. 303–308. DOI.

[138] P. V. Overschee and B. D. Moor. “A unifying theorem for three subspace system identification algorithms”. In: *Automatica* 31.12 (Dec. 1995), pp. 1853–1864. DOI.

[139] P. V. Overschee and B. D. Moor. “N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems”. In: *Automatica* 30.1 (Jan. 1994), pp. 75–93. DOI.

[140] E. Özkan et al. “Marginalized adaptive particle filtering for nonlinear models with unknown time-varying noise parameters”. In: *Automatica* 49.6 (June 2013), pp. 1566–1575. DOI. URL.

[141] S. Park et al. “Measurement Noise Recommendation for Efficient Kalman Filtering over a Large Amount of Sensor Data”. In: *Sensors* 19.5 (Mar. 2019), p. 1168. DOI. URL.

[142] S. J. Qin. “An overview of subspace identification”. In: *Computers Chemical Engineering* 30.10-12 (Sept. 2006), pp. 1502–1513. DOI.

[143] S. J. Qin, W. Lin, and L. Ljung. “A novel subspace identification approach with enforced causal models”. In: *Automatica* 41.12 (Dec. 2005), pp. 2043–2053. DOI. URL.

[144] M. R. Rajamani and J. B. Rawlings. “Estimation of the disturbance structure from data using semidefinite programming and optimal weighting”. In: *Automatica* 45.1 (Jan. 2009), pp. 142–148. DOI.

[145] H. E. Rauch. “Solutions to the linear smoothing problem”. In: *IEEE Transactions on Automatic Control* 8.4 (Oct. 1963), pp. 371–372. DOI. URL.

[146] H. E. Rauch, F. Tung, and C. T. Striebel. “Maximum likelihood estimates of linear dynamic systems”. In: *AIAA Journal* 3.8 (Aug. 1965), pp. 1445–1450. DOI. URL.

[147] R. H. Reichle, W. T. Crow, and C. L. Keppenne. “An adaptive ensemble Kalman filter for soil moisture data assimilation”. In: *Water Resources Research* 44.3 (Mar. 2008). DOI. URL.

[148] R. G. Reynolds. “Robust estimation of covariance matrices”. In: *IEEE Transactions on Automatic Control* 35.9 (1990), pp. 1047–1051. DOI. URL.

[149] F. Ruiz, C. Novara, and M. Milanese. “Direct design from data of optimal filters for LPV systems”. In: *Systems & Control Letters* 59.1 (Jan. 2010), pp. 1–8. DOI.

[150] S. Sarkka and A. Nummenmaa. “Recursive Noise Adaptive Kalman Filtering by Variational Bayesian Approximations”. In: *IEEE Transactions on Automatic Control* 54.3 (Mar. 2009), pp. 596–600. DOI. URL.

[151] A. H. Sayed and T. Kailath. “A survey of spectral factorization methods”. In: *Numerical Linear Algebra with Applications* 8.6-7 (2001), pp. 467–496. DOI. URL.

[152] T. B. Schön, A. Wills, and B. Ninness. “System identification of nonlinear state-space models”. In: *Automatica* 47.1 (Jan. 2011), pp. 39–49. DOI. URL.

[153] B. D. Schutter. “Minimal state-space realization in linear system theory: an overview”. In: *Journal of Computational and Applied Mathematics* 121.1-2 (Sept. 2000), pp. 331–354. DOI.

[154] C. E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. DOI. URL.

[155] C. E. Shannon. “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.4 (Oct. 1948), pp. 623–656. DOI. URL.

[156] R. H. Shumway and D. S. Stoffer. “AN APPROACH TO TIME SERIES SMOOTHING AND FORECASTING USING THE EM ALGORITHM”. In: *Journal of Time Series Analysis* 3.4 (July 1982), pp. 253–264. DOI. URL.

[157] M. Šimandl and J. Duník. “Estimation of noise covariance matrices for periodic systems”. In: *International Journal of Adaptive Control and Signal Processing* 25.10 (July 2011), pp. 928–942. DOI. URL.

[158] F. Sims, D. Lainiotis, and D. Magill. “Recursive algorithm for the calculation of the adaptive Kalman filter weighting coefficients”. In: *IEEE Transactions on Automatic Control* 14.2 (Apr. 1969), pp. 215–218. DOI. URL.

[159] G. Smith. “Sequential estimation of observation error variances in a trajectoryestimation problem.” In: *AIAA Journal* 5.11 (Nov. 1967), pp. 1964–1970. DOI. URL.

[160] A. Solonen et al. “Estimating model error covariance matrix parameters in extended Kalman filtering”. In: *Nonlinear Processes in Geophysics* 21.5 (Sept. 2014), pp. 919–927. DOI. URL.

[161] J. Spragins. “A note on the iterative application of Bayes’ rule”. In: *IEEE Transactions on Information Theory* 11.4 (Oct. 1965), pp. 544–549. DOI. URL.

[162] R. L. Stratonovich. “Conditional Markov Processes”. In: *Theory of Probability & Its Applications* 5.2 (Jan. 1960), pp. 156–178. DOI.

[163] J. R. Stroud and T. Bengtsson. “Sequential State and Variance Estimation within the Ensemble Kalman Filter”. In: *Monthly Weather Review* 135.9 (Sept. 2007), pp. 3194–3208. DOI. URL.

[164] W. Tsang, J. Glover, and R. Bach. “Identifiability of unknown noise covariance matrices for some special cases of a linear, time-invariant, discrete-time dynamic system”. In: *IEEE Transactions on Automatic Control* 26.4 (Aug. 1981), pp. 970–974. DOI. URL.

[165] M. Verhaegen. “Identification of the deterministic part of MIMO state space models given in innovations form from input-output data”. In: *Automatica* 30.1 (Jan. 1994), pp. 61–74. DOI.

[166] M. Verhaegen and P. Dewilde. “Subspace model identification Part 1. The output-error state-space model identification class of algorithms”. In: *International Journal of Control* 56.5 (Nov. 1992), pp. 1187–1210. DOI. URL.

[167] M. Waller and H. Saxén. “Estimating the degree of time variance in a parametric model”. In: *Automatica* 36.4 (Apr. 2000), pp. 619–625. DOI. URL.

[168] X. Wang et al. “A novel approach of noise statistics estimate using H ∞ filter in target tracking”. In: *Frontiers of Information Technology & Electronic Engineering* 17.5 (May 2016), pp. 449–457. DOI. URL.

[169] D. M. Wiberg and D. G. DeWolf. “A convergent approximation of the continuous-time optimal parameter estimator”. In: *IEEE Transactions on Automatic Control* 38.4 (Apr. 1993), pp. 529–545. DOI. URL.

[170] D. M. Wiberg, T. D. Powell, and D. Ljungquist. “An online parameter estimator for quick convergence and time-varying linear systems”. In: *IEEE Transactions on Automatic Control* 45.10 (Oct. 2000), pp. 1854–1863. DOI. URL.

[171] D. M. Wiberg et al. “A Fix-Up for the EKF Parameter Estimator”. In: *IFAC Proceedings Volumes* 41.2 (2008), pp. 6502–6507. DOI. URL.

[172] L. A. Zadeh. “Optimum Nonlinear Filters”. In: *Journal of Applied Physics* 24.4 (Apr. 1953), pp. 396–404. DOI.

[173] L. A. Zadeh and J. R. Ragazzini. “An Extension of Wiener’s Theory of Prediction”. In: *Journal of Applied Physics* 21.7 (July 1950), pp. 645–655. DOI.

[174] M. A. Zagrobelny and J. B. Rawlings. “Practical improvements to autocovariance least-squares”. In: *AIChE Journal* 61.6 (Mar. 2015), pp. 1840–1855. DOI. URL.

[175] M. Zakai. “On the optimal filtering of diffusion processes”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 11.3 (1969), pp. 230–243. DOI.

[176] Y. Zhen and J. Harlim. “Adaptive error covariances estimation methods for ensemble Kalman filters”. In: *Journal of Computational Physics* 294 (Aug. 2015), pp. 619–638. DOI.

[177] J. Zhou and R. H. Luecke. “Estimation of the co variances of the process noise and measurement noise for a linear discrete dynamic system”. In: *Computers & Chemical Engineering* 19.2 (Feb. 1995), pp. 187–195. DOI. URL.

[178] O. J. Zobel. “Theory and Design of Uniform and Composite Electric Wavefilters”. In: *Bell System Technical Journal* 2.1 (Jan. 1923), pp. 1–46. DOI.

Ph.D. Theses

[179] P. D. Abramson. “Simultaneous estimation of the stateand noise statistics in linear dynamical systems”. PhD thesis. Institute of Technology, Massachusetts, May 10, 1968. URL.

[180] A. Doucet. “Monte Carlo algorithms for Bayesian estimation of hidden Markov models. Application to processing of radiation signals”. Ph.D. Thesis. Universite de Paris XI, 1997.

[181] S. Formentin. “Direct data-driven control system design : theory and applications”. PhD thesis. Politecnico di Milano, Feb. 2, 2012.

[182] G. W. Husa and A. P. Sage. “ADAPTIVE BAYES FILTERING WITHUNKNOWN PRIOR STATISTICS”. PhD thesis. Institute of Technology, Southern Methodist University, 1969. URL.

[183] M. Lovera. “Subspace identification methods: theory and applications”. PhD thesis. Politecnico di Milano, Dipartimento di Elettronica e Informazione, 1997.

[184] M. Mazzoleni. “Learning meets control: data analytics for dynamical systems”. PhD thesis. University of Bergamo, Department of Engineering and Applied Sciences, 2018. URL.

[185] M. R. Rajamani. “Data-based Techniques to Improve State Estimation in Model Predictive Control”. PhD thesis. University of Wisconsin–Madison, 2007.

[186] M. Scandella. “Regularized kernel-based learning for system identication”. PhD thesis. University of Bergamo, Department of Engineering and Applied Sciences, 2020. URL.

[187] J. C. Shellenbarger. “ESTIMATION OF COVARIANCE PARAMETERS FOR AN ADAPTIVE KALMAN FILTER”. PhD thesis. Iowa State University of Science and Technology, 1966. URL.

[188] L. M. Silverman. “Structural properties of time-variable linear systems”. PhD thesis. Department of Electrical Engineering, Columbia University, 1966.

Other references

[179] P. D. Abramson. “Simultaneous estimation of the stateand noise statistics in linear dynamical systems”. PhD thesis. Institute of Technology, Massachusetts, May 10, 1968. URL.

[189] B. D. O. Anderson and J. B. Moore. “Kalman Filtering: Whence, What and Whither?” In: *Mathematical System Theory*. Springer Berlin Heidelberg, 1991, pp. 41–54. DOI.

[190] S. Bittanti. “History and prehistory of the Riccati equation”. In: *Proceedings of 35th IEEE Conference on Decision and Control*. Vol. 2. IEEE, 1996, pp. 1599–1604. DOI. URL.

[191] S. Bittanti. “Count Riccati and the Early Days of the Riccati Equation”. In: *The Riccati Equation*. Springer Berlin Heidelberg, 1991, pp. 1–10. DOI. URL.

[192] W. T. Bundick. *Development of a Technique for Estimating Noise Covariances Using Multiple Observers*. Tech. rep. NASA Technical Memorandum 4020, 1988. URL.

[193] M. Canale et al. “A study on the use of virtual sensors in vehicle control”. In: *2008 47th IEEE Conference on Decision and Control*. IEEE, 2008. DOI.

[194] A. Chiuso and G. Picci. “Geometry of Oblique Splitting Subspaces, Minimality and Hankel Operators”. In: *Directions in Mathematical Systems Theory and Optimization*. Springer Berlin Heidelberg, 2002, pp. 85–126. DOI.

[180] A. Doucet. “Monte Carlo algorithms for Bayesian estimation of hidden Markov models. Application to processing of radiation signals”. Ph.D. Thesis. Universite de Paris XI, 1997.

[195] J. Duník et al. “Estimation of state and measurement noise characteristics”. In: *2015 18th International Conference on Information Fusion (Fusion)*. IEEE. 2015, pp. 1817–1824.

[196] P. L. Faurre. “Stochastic Realization Algorithms”. In: *Mathematics in Science and Engineering*. Elsevier, 1976, pp. 1–25. DOI. URL.

[181] S. Formentin. “Direct data-driven control system design : theory and applications”. PhD thesis. Politecnico di Milano, Feb. 2, 2012.

[197] G. C. Goodwin and J. C. Aguero. “Approximate EM Algorithms for Parameter and State Estimation in Nonlinear Stochastic Models”. In: *Proceedings of the 44th IEEE Conference on Decision and Control.* IEEE, Dec. 15, 2005. DOI. URL.

[182] G. W. Husa and A. P. Sage. “ADAPTIVE BAYES FILTERING WITHUNKNOWN PRIOR STATISTICS”. PhD thesis. Institute of Technology, Southern Methodist University, 1969. URL.

[198] S. J. Julier and J. K. Uhlmann. “New extension of the Kalman filter to nonlinear systems”. In: *SPIE Proceedings.* Ed. by I. Kadar. SPIE, July 1997. DOI.

[199] T. Kailath. “Lectures on Wiener and Kalman Filtering”. In: *Lectures on Wiener and Kalman Filtering.* Springer Vienna, 1981, pp. 1–143. DOI. URL.

[200] T. Kailath. “Norbert Wiener and the Development of Mathematical Engineering”. In: *Communications, Computation, Control, and Signal Processing.* Springer US, 1997, pp. 35–64. DOI. URL.

[201] A. Karimi, K. van Heusden, and D. Bonvin. “Non-iterative data-driven controller tuning using the correlation approach”. In: *2007 European Control Conference (ECC).* IEEE, July 2007. DOI. URL.

[202] W. E. Larimore. “Canonical variate analysis in identification, filtering, and adaptive control”. In: *29th IEEE Conference on Decision and Control.* IEEE, 1990. DOI.

[203] X. R. Li and Y. Bar-Shalom. “A recursive hybrid system approach to noise identification”. In: *The First IEEE Conference on Control Applications.* IEEE, 1992. DOI. URL.

[204] Z. Liqiang et al. “Colored noise estimation algorithm based on autocovariance least-squares method”. In: *2015 12th IEEE International Conference on Electronic Measurement & Instruments (ICEMI).* IEEE, July 2015. DOI. URL.

[205] J. Liu and M. West. “Combined Parameter and State Estimation in Simulation-Based Filtering”. In: *Sequential Monte Carlo Methods in Practice.* Springer New York, 2001, pp. 197–223. DOI. URL.

[206] L. Ljung. *System Identification Toolbox.* Ed. by MATLAB. 1987.

[183] M. Lovera. “Subspace identification methods: theory and applications”. PhD thesis. Politecnico di Milano, Dipartimento di Elettronica e Informazione, 1997.

[184] M. Mazzoleni. “Learning meets control: data analytics for dynamical systems”. PhD thesis. University of Bergamo, Department of Engineering and Applied Sciences, 2018. URL.

[207] R. Mehra. “Approaches to adaptive filtering”. In: *1970 IEEE Symposium on Adaptive Processes (9th) Decision and Control.* IEEE, Dec. 1970. DOI.

[208] M. Milanese et al. “Filter design from data: direct vs. two-step approaches”. In: *2006 American Control Conference.* IEEE, 2006. DOI.

[209] M. Milanese, D. Regruto, and A. Fortina. "Direct Virtual Sensor (DVS) design in vehicle sideslip angle estimation". In: *2007 American Control Conference*. IEEE, July 2007. DOI.

[210] M. Milanese, F. Ruiz, and M. Taragna. "Linear virtual sensors for vertical dynamics of vehicles with controlled suspensions". In: *2007 European Control Conference (ECC)*. IEEE, July 2007. DOI.

[185] M. R. Rajamani. "Data-based Techniques to Improve State Estimation in Model Predictive Control". PhD thesis. University of Wisconsin–Madison, 2007.

[186] M. Scandella. "Regularized kernel-based learning for system identication". PhD thesis. University of Bergamo, Department of Engineering and Applied Sciences, 2020. URL.

[211] T. B. Schon and F. Lindsten. "Learning of dynamical systems: Particle filters and Markov chain methods". draft manuscript. Aug. 23, 2017.

[187] J. C. Shellenbarger. "ESTIMATION OF COVARIANCE PARAMETERS FOR AN ADAPTIVE KALMAN FILTER". PhD thesis. Iowa State University of Science and Technology, 1966. URL.

[212] A. N. Shiryayev. "Interpolation and Extrapolation of Stationary Random Sequences". In: *Selected Works of A. N. Kolmogorov*. Springer Netherlands, 1992, pp. 272–280. DOI.

[188] L. M. Silverman. "Structural properties of time-variable linear systems". PhD thesis. Department of Electrical Engineering, Columbia University, 1966.

[213] R. L. Stratonovich. "Conditional Markov Processes and Their Application to the Theory of Optimal Control". In: *Non-Linear Transformations of Stochastic Processes*. Elsevier, 1965, pp. 427–453. DOI.

[214] M. Taragna, F. Ruiz, and M. Milanese. "Virtual sensors for linear dynamic systems: structure and identification". In: *3rd international IEEE scientific conference on physics and control (PhysCon 2007)*. 2007, pp. 1–7. URL.

[215] H. F. Wahab and R. Katebi. "Robust adaptive estimators for nonlinear systems". In: *2013 Conference on Control and Fault-Tolerant Systems (SysTol)*. IEEE, Oct. 2013. DOI. URL.

[216] G. Welch, G. Bishop, et al. *An introduction to the Kalman filter*. 1995. URL.

[217] N. Wiener and E. Hopf. "On a class of singular integral equations". In: *Proc. Prussian Acad. Math-Phys. Ser.* Vol. 636. 1931.

[218] P. J. Wojcik. "On-line Estimation of Signal and Noise Parameters and the Adaptive Kalman Filtering". In: *Approximate Kalman Filtering*. WORLD SCIENTIFIC, Aug. 1993, pp. 87–111. DOI. URL.

APPENDIX A

Random variables

The following appendix summarized important results on the study of random variables. The herein discussion is inspired from the works in [211, 31]. Please refer to them for a full reference. In particular, the chapter is organized as follow:

- Appendix A.1 introduces in general random variables sampled from a Gaussian distribution in a real-valued domain.
- Appendix A.2 introduces specifically results for the conditioning and marginalization of the partitioning of two or more random variables sampled from a Gaussian distribution in a real-valued domain.
- Appendix A.3 introduces specifically results for the conditioning and marginalization of affine transformations of random variables sampled from a Gaussian distribution in a real-valued domain.

A.1 Real-valued gaussian random variables

The following section introduces in general the concept of random variables with the properties of being sampled from a Gaussian distribution and having a real-valued domain. Specifically, results will be given directly working with multivariate random variables.

A.1.1 Multivariate gaussian

The multivariate gaussian distribution is the joint pdf for a set of RVs $x_1, x_2, \dots, x_{n_x}$ with gaussian pdfs, and arbitrary correlation. Given $\boldsymbol{x} = [x_1, x_2, \dots, x_{n_x}]^\top$ a vector of n_x real Gaussian RVs, the joint pdf is:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma_{\boldsymbol{xx}}|} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^\top \Sigma_{\boldsymbol{xx}}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})\right) \tag{A.1}$$

with mean $\bar{\boldsymbol{x}}$:

$$\bar{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}] = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{n_x}]^\top \tag{A.2}$$

and covariance Σ_{xx}:

$$\Sigma_{xx} = \mathbb{E}\left[(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^\mathsf{T}\right] \tag{A.3}$$

In a compact form it is customary to use the following notation to denote that the RV $\boldsymbol{x}$ is multivariate gaussian with mean value $\bar{\boldsymbol{x}}$ and covariance Σ_{xx}:

$$\boldsymbol{x} \sim \mathcal{G}(\bar{\boldsymbol{x}}, \Sigma_{xx}) \tag{A.4}$$

while the joint pdf in Eq. (A.1) can be compactly described using the notation:

$$\mathcal{N}(\boldsymbol{x}; \bar{\boldsymbol{x}}, \Sigma_{xx}) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma_{xx}|} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}})^\mathsf{T}\Sigma_{xx}^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}})\right) \tag{A.5}$$

A.1.2 Linear transformation of a multivariate gaussian

A multivariate gaussian RV $\boldsymbol{x}$ can be linearly transformed as:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \tag{A.6}$$

the resulting multivariate RV $\boldsymbol{y}$ is still gaussian:

$$\boldsymbol{y} \sim \mathcal{G}(\bar{\boldsymbol{y}}, \Sigma_{yy}) \tag{A.7}$$

with mean $\bar{\boldsymbol{y}}$:

$$\begin{aligned} \bar{\boldsymbol{y}} &= \mathbb{E}[\boldsymbol{y}] && \text{(A.8a)} \\ &= \mathbb{E}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] + \boldsymbol{b} && \text{(A.8b)} \\ &= \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{b} && \text{(A.8c)} \end{aligned}$$

and covariance Σ_{yy}:

$$\begin{aligned} \Sigma_{yy} &= \mathbb{E}\left[(\boldsymbol{y} - \bar{\boldsymbol{y}})(\boldsymbol{y} - \bar{\boldsymbol{y}})^\mathsf{T}\right] && \text{(A.9a)} \\ &= \mathbb{E}\left[(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b})(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} - \boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b})^\mathsf{T}\right] && \text{(A.9b)} \\ &= \boldsymbol{A}\mathbb{E}\left[(\boldsymbol{x} - \bar{\boldsymbol{x}})(\boldsymbol{x} - \bar{\boldsymbol{x}})^\mathsf{T}\right]\boldsymbol{A}^\mathsf{T} && \text{(A.9c)} \\ &= \boldsymbol{A}\Sigma_{xx}\boldsymbol{A}^\mathsf{T} && \text{(A.9d)} \end{aligned}$$

A.2 Partitioning of a multivariate gaussian pdf

Given two multivariate gaussian RVs, $\boldsymbol{x} = [x_1, \dots, x_{n_x}]^\mathsf{T}$ and $\boldsymbol{y} = [y_1, \dots, y_{n_y}]^\mathsf{T}$:

$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{G}(\bar{\boldsymbol{x}}, \Sigma_{xx}) && \text{(A.10)} \\ \boldsymbol{y} &\sim \mathcal{G}(\bar{\boldsymbol{y}}, \Sigma_{yy}) && \text{(A.11)} \end{aligned}$$

Let the vector $\boldsymbol{z}$ be obtained through the concatenation of $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \in \mathbb{R}^{n_x + n_y} \tag{A.12}$$

It is jointly Gaussian:

$$z \sim \mathcal{N}(z; \bar{z}, \Sigma_{zz}) \tag{A.13}$$

where mean and covariances are block-partitioned matrices collecting the two multivariate RVs $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively:

$$\bar{z} = \mathbb{E}[z] = \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \tag{A.14}$$

and

$$\Sigma_{zz} = \mathbb{E}[(z - \bar{z})(z - \bar{z})^{\mathsf{T}}] = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \tag{A.15}$$

Marginalization and conditioning of partitioned Gaussian pdfs are now stated.

A.2.1 Marginalization of gaussian pdfs

Let the random vector $\boldsymbol{x}$ be Gaussian distributed according to Eq. (A.5) and let it be partitioned according to Eq. (A.13), then the marginal pdf $p(\boldsymbol{x})$ is given by:

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \bar{\boldsymbol{x}}, \Sigma_{xx}) \tag{A.16}$$

A.2.2 Conditioning of gaussian pdfs

The conditional pdf of the Gaussian distribution is Gaussian, and conditioning (i.e. slicing the pdf) might change the mean and covariance of the resulting pdf. Derivation of conditional pdfs is particularly important in some estimation methods. The conditional pdf $p(\boldsymbol{x}|\boldsymbol{y})$ can be evaluated from the joint pdf $p(\boldsymbol{x}, \boldsymbol{y})$. In particular, let the random vector $\boldsymbol{x}$ be Gaussian distributed according to Eq. (A.5) and let it be partitioned according to Eq. (A.13), then the conditional pdf $p(\boldsymbol{x}|\boldsymbol{y})$ is given by the Bayes' rule:

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &= \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(z)}{p(\boldsymbol{y})} && \text{(A.17a)} \\ &= \frac{\mathcal{N}(z; \bar{z}, \Sigma_{zz})}{\mathcal{N}(\boldsymbol{y}; \bar{\boldsymbol{y}}, \Sigma_{yy})} && \text{(A.17b)} \\ &= \mathcal{N}(\boldsymbol{x}; \bar{\boldsymbol{x}} + \Sigma_{xy}\Sigma_{yy}^{-1}(\boldsymbol{y} - \bar{\boldsymbol{y}}), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}) && \text{(A.17c)} \end{aligned}$$

with mean:

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = \bar{\boldsymbol{x}} + \Sigma_{xy}\Sigma_{yy}^{-1}(\boldsymbol{y} - \bar{\boldsymbol{y}}) \tag{A.18}$$

and covariance

$$\mathbb{C}[\boldsymbol{x}|\boldsymbol{y}] = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \tag{A.19}$$

A.3 Affine transformation of gaussian pdfs

In the previous section the expressions for the marginal and conditional pdfs expressed in terms of the parameters of the joint pdf are derived in the case of a partitioned Gaussian pdf. A different starting point is now considered, namely that the marginal density $p(\boldsymbol{x})$ and the conditional density $p(\boldsymbol{y}|\boldsymbol{x})$ are given, and expressions for the joint density $p(\boldsymbol{x}, \boldsymbol{y})$, the marginal density $p(\boldsymbol{y})$, and the conditional density $p(\boldsymbol{x}|\boldsymbol{y})$ are derived.

Assume that the random vector $\boldsymbol{x}$, as well as $\boldsymbol{y}$ conditioned on $\boldsymbol{x}$, i.e. $\boldsymbol{y}|\boldsymbol{x}$, be Gaussian distributed:

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \bar{\boldsymbol{x}}, \Sigma_{xx}) \tag{A.20a}$$

$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \Sigma_{y|x}) \tag{A.20b}$$

where $\boldsymbol{A}$ is a matrix (of appropriate dimension) and $\boldsymbol{b}$ is a constant vector. The joint distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$ is then given by:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}; \begin{bmatrix}\bar{\boldsymbol{x}}\\ \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{b}\end{bmatrix}, \begin{bmatrix}\Sigma_{xx} & \Sigma_{xx}\boldsymbol{A}^{\mathsf{T}}\\ \boldsymbol{A}\Sigma_{xx} & \boldsymbol{A}\Sigma_{xx}\boldsymbol{A}^{\mathsf{T}} + \Sigma_{y|x}\end{bmatrix}\right) \tag{A.21}$$

Combining the results of Appendices A.2.1 and A.2.2 the marginal and conditional pdf are derived as well. In particular, the marginal $p(\boldsymbol{y})$ is given by:

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}; \bar{\boldsymbol{y}}, \Sigma_{yy}) \tag{A.22}$$

with mean

$$\bar{\boldsymbol{y}} = \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{b} \tag{A.23}$$

and covariance

$$\Sigma_{yy} = \boldsymbol{A}\Sigma_{xx}\boldsymbol{A}^{\mathsf{T}} + \Sigma_{y|x} \tag{A.24}$$

Instead, the conditional pdf $p(\boldsymbol{x}|\boldsymbol{y})$ is given by:

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \bar{\boldsymbol{x}} + \Sigma_{xx}\boldsymbol{A}^{\mathsf{T}}\Sigma_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{A}\bar{\boldsymbol{x}} - \boldsymbol{b}), \Sigma_{xx} - \Sigma_{xx}\boldsymbol{A}^{\mathsf{T}}\Sigma_{yy}^{-1}\Sigma_{xx}\boldsymbol{A}) \tag{A.25}$$