Czech Technical University in Prague

Faculty of Transportation Sciences

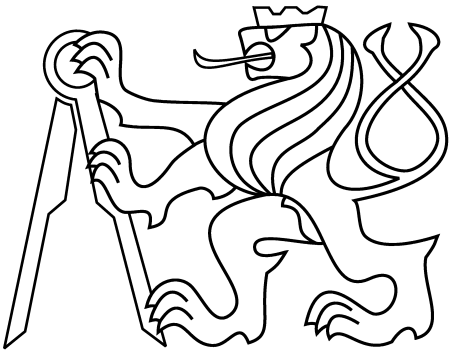
Department of Applied Mathematics

PhD THESIS

Prague, August 2010

Ing. Kamil Dedecius

Czech Technical University in Prague
Faculty of Transportation Sciences
Department of Applied Mathematics



Partial Forgetting in Bayesian Estimation

Thesis

Kamil Dedecius

PhD study program: Engineering Informatics
Branch of study: Engineering Informatics in Transportation and Telecommunications

Prague, August 2010

This thesis was written during the PhD study at the Department of Applied Mathematics, Faculty of Transportation Sciences, Czech Technical University in Prague.

Candidate:	Ing. Kamil Dedecius Department of Applied Mathematics Faculty of Transportation Sciences Czech Technical University in Prague Konviktská 20, 110 00 Praha 1
Supervisor:	Doc. Ing. Ivan Nagy, CSc. Department of Applied Mathematics Faculty of Transportation Sciences Czech Technical University in Prague Konviktská 20, 110 00 Praha 1

Date of state exam:	May 16, 2008
Date of thesis submission:	

Acknowledgement

First of all, I would like to express the deepest gratitude to an extraordinary man Ivan Nagy, my supervisor, for his kind patience and endurance in keeping me on track during the long journey to this thesis. My gratitude also goes to doctor Miroslav Kárný, who always patiently explained me the difficult aspects of the science.

I am very grateful to my friend and 'colleague in arms' Radek Hofman, an inexhaustible source of inspiration and outstanding young scientist, for our interesting discussions regarding every possible scientific and non-scientific topic.

Last, but not least, my deepest thanks belong to my wife Iva for her incredible support both in good times and in times when the real life was tough.

I would like to dedicate this work to the memory of my mother.

Prague, August 2010

Kamil Dedecius

This work was supported by grant GAČR 102/08/0567 Fully probabilistic design of dynamic decision strategies and by the Research center DAR, project of MŠMT 1M0572.

Contents

List of Figures

List of Tables

Notational conventions

Operators, intervals	
$\equiv$	equivalence by definition
$\propto$	proportionality, equivalence up to a constant factor
$\oplus$	direct sum
$\int_{x^*}$	integration over set x^*
$\sum_{i=a}^{b}$	summation over indices $i = a, \ldots, b$
$\circ$	composed mapping
$\lfloor a \rfloor$	floor of a
$\lvert a \rvert$	absolute value of $a \in \mathbb{R}^1$
$\langle a, b \rangle$	closed interval from a to b
(a, b)	open interval from a to b
$(a, b\rangle$, $\langle a, b)$	left-open and right-open intervals from a to b, respectively

Number systems	
$\mathbb{C}$	set of complex numbers
$\mathfrak{R}(\mathbb{C})$	real part of a complex number
$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}^+$	set of positive real numbers
$\mathbb{Z}$	set of integers
$\mathbb{Z}^+$	set of positive integers

Algebraic notation	
$A \in \mathbb{R}^{m \times n}$	matrix of dimension $m \times n$
$a \in \mathbb{R}^n$	column vector of length n
$I \in \mathbb{R}^{n \times n}$	identity matrix, i.e., square matrix with 1's on the main diagonal and zeros elsewhere
A'	transpose of matrix A
$V = L'DL$	factorization of positive definite matrix $V \in \mathbb{R}^{n \times n}$ into lower triangular matrix with unit diagonal $L \in \mathbb{R}^{n \times n}$ and diagonal matrix $D \in \mathbb{R}^{n \times n}$
$\det(A) = \lvert A \rvert$	determinant of matrix A
$\mathrm{diag}(A)$	diagonal of matrix A

$\mathrm{Tr}(A)$	trace of matrix A

Probability and modelling	
d	data (relevant inputs and outputs)
$\mathrm{D}\left(f\|\|g\right)$	Kullback-Leibler divergence of f on g
$\mathrm{E}\left[\cdot\right]$	expected value of argument $\cdot$
$f(x)$	probability density function determined by its argument x
$f(x\|y)$	probability density function determined by its argument x, conditioned by y
${}^{\tau}f$	true probability density function
$\mathcal{H}^*$	set of hypotheses (Chapter 5)
$\mathcal{L}(\cdot)$	likelihood
$\mathrm{P}(\cdot)$	probability of argument $\cdot$
$\hat{x}$	point estimate of x
$\tilde{x}$	approximate of x
$\mathrm{var}(X)$	variance of a random variable X
ψ	regression vector
$\Psi = [y, \psi']'$	extended regression vector
θ	vector of regression coefficients
Θ	model parameters
$\mathcal{D}i(\lambda, \alpha)$	Dirichlet distribution with parameters λ and α
$\mathcal{G}i\mathcal{W}(V, \nu)$	Gauss-inverse-Wishart distribution with information matrix V and degrees of freedom ν
$i\mathcal{W}(V, \nu)$	inverse-Wishart distribution with matrix V and degrees of freedom ν
$\mathcal{N}(\mu, \sigma^2)$	Gaussian (normal) distribution with mean value μ and variance σ^2
$\mathcal{T}_\nu$	Student distribution with ν degrees of freedom

Sets, set operators	
x^*	set of x values
$x_{i:j} = \{x_i, x_{i+1}, \ldots, x_j\}$	ordered set of variables $x_a, \quad a = i, \ldots, j$
$x(j) = \{x_1, x_2, \ldots, x_j\}$	ordered set of variables $x_a, \quad a = 1, \ldots, j$
X^C	complement of set X
$\mathrm{card}(A)$	cardinality of set A
$A \subset B$	A is subset of B

Functions, constants	
$\exp(x)$	exponential function of argument x, i.e. e^x
$B(x)$	beta function of argument x
γ	Euler or Euler-Mascheroni constant (Appendix A.6)
$\Gamma(x)$	gamma function of argument x (Appendix A.5)

$\Gamma_n(x)$	multivariate gamma function of argument x (Chapter 4)
$\psi_0(x)$	digamma function of argument x

List of Acronyms

AF	Alternative forgetting
AR	Autoregressive model
ARMA	Autoregressive model with moving average
ARX	Autoregressive model with exogenous inputs
ASCR	Academy of Sciences of the Czech Republic
CSAS	Czechoslovak Academy of Sciences
DF	Directional forgetting
EF	Exponential forgetting
EKF	Extended Kalman filter
GA	Genetic algorithm
IFAC	International Federation for Automatic Control
KF	Kalman filter
LF	Linear forgetting
LPV	Linear parameter-varying (system)
LS	Least squares
LTI	Linear time-invariant (system)
LTV	Linear time-varying (system)
MA	Moving average
MCMC	Markov chain Monte Carlo
OLS	Ordinary least squares
PFM	Partial forgetting method
RLS	Recursive Least Squares
SEF	Stabilized exponential forgetting
SF	Selective forgetting
SLF	Stabilized linear forgetting
TWLS	Time weighted least squares
UKF	Unscented Kalman filter
UTIA	Institute of Information Theory and Automation
cdf	cumulative distribution function
i.i.d.	independent, identically distributed (random variable)
pdf	probability density function

Chapter 1

Introduction

In order to describe the reality of the world around us and the influence of various phenomena, the human beings make models [63]. These models, in whatever form they take, are intended to be abstractions of real objects, carrying their important properties. Among the main advantages of the modelling is the ability to avoid the potentially dangerous *trial and error* experiments on real systems.

This thesis is focused on mathematical models, which employ mathematical principles to describe the systems with input and output signals and in which a transformation of these quantities is performed [75]. Such a topic is covered by the signal processing science. One of the main tasks of sciences with the mathematical modelling orientation is the extraction of desired quantitative properties of signals and systems. This field is called system identification [63, 51].

Often, the reality changes over the time and the system models have to reflect this unavoidable fact, otherwise they would hardly be viable. For example, the traffic intensity on a road varies during the day, week and year; the car dynamics given by speed, acceleration and direction vary during the ride, etc. If the models are intended to predict the future system properties (e.g., for control), the assumption of constant parameters must be released. In this case, the parameter identification methods capable of tracking the changes and self-adapting must be applied, which turns the ordinary parameter identification methods into parameter tracking, belonging to the group of adaptive filtration methods.

The tracking of varying model parameters was given a thorough attention during the last 50 years, however, the results are still not very satisfactory. The reason consists in the two difficulties – the trade-off between the tracking and noise rejection abilities and the common lack of the prior information needed for the solution of the trade-off problem.

There are multiple possible approaches to stochastic mathematical system modelling, the two most common are the Bayesian methodology [63, 40] and the frequentists' methodology. While the second one introduces mostly the algebraic solutions to the parameter tracking, e.g., the ordinary least squares (OLS), recursive least squares (RLS), the Kalman filter in its classical interpretation as the optimal linear least squares filter, the Bayesian methodology employs a fully probabilistic view expressed in terms of distributions and their properties (divergence, moments etc.).

In this thesis, a new method for recursive identification of slowly varying model parameters in a fully Bayesian framework is proposed. The developed partial forgetting method employs an alternative information to decide which parameters are about to be released and their variability in time admitted. It allows the user to construct hypotheses about the parameters, which under the knowledge of their potential behaviour leads to a significant improvement of the model reliability.

1.1 The state of the art

Most of the parameter estimation methods are based on the least squares (LS) method, originally introduced by French mathematician Legendre in 1805, however, it is mostly credited to Gauss who independently published the idea in 1809 in *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium* [25] and who claimed to have worked out the basics as soon as in 1795 at the age of eighteen. The claim started a long dispute between the two authors. In addition, the method was published in the United States in 1808 by Adrain. All of these men invented the method for the purposes of physics, the first two of them for astronomical measurements, Adrian for survey measurements. Later on, Gauss found a recursive version of the least squares algorithm (RLS), which avoided matrix inversion by using a matrix inversion lemma[1] (which was not known in his times) [26]. The recursive method was forgotten for about 120 years, when it was rediscovered by Plackett in 1950 [65] but it was first overlooked by practitioners, who preferred batch processing, but due to the development of the computers the recursive approach was finally fully recognized. The method was further generalized for dynamic systems where the noise may affect both the system model and the measurements by Kalman in 1960 in [38] and later in [37], however, the Danish mathematician and physician Thiele and American mathematician Swerling developed a similar algorithm earlier. The nowadays known Kalman filter (KF) was initially met with scepticism until its use in the Apollo program, which led to its recognition and widespread use. Various modifications of the filter appeared later, e.g., the sequential KF [11], information filtration replacing the covariance matrix by its inverse, square-root filtration, U-D filtration [7], steady state filtration, $\alpha - \beta$ and $\alpha - \beta - \gamma$ filtration and many others [74]. The Kalman filter has the important property of optimality for linear systems with both Gaussian and non-Gaussian noise, however, from the Bayesian viewpoint only the first one can be proved.

Apart from the estimation from the finite batch of data, also known as finite window estimation or limited memory filtration, the first notions of forgetting in recursive estimation algorithms appeared in the mid-1960's in the fashion of exponential data weighting [56, 63, 34]. This method, called exponential forgetting (EF), applied a positive weighting factor less than one to the parameter covariance matrix to discard old and potentially outdated information, which made the RLS algorithms adaptive. An the same time, the divergence properties of the Kalman filter were investigated and some fading-memory approaches and the KF with exponential data weighting [21] were proposed in order to improve the convergence properties. Simultaneously, the time-varying parameters changes were viewed as a random walk with a constant covari-

[1] Also known as Woodbury identity or Sherman-Morrison formula

ance matrix in state-space models, which established the basic principle of the linear forgetting (LF)[56].

However, both the linear and exponential forgetting encounter problems, when the signals are not persistently exciting and the forgetting is not compensated by the innovations. This issue leads to the phenomenon called covariance blow-up, when the covariance matrix grows without bounds, which causes extreme noise sensitivity of the estimation algorithm due to its rapidly increasing gain [56, 49]. This led many scientists to use the normalized weighted least squares (NLMS) algorithm instead, independently suggested by Naguma and Noda in [59] and Albert and Gardner in [1]. In spite of the computational efficiency of the NLMS, its worse parameter convergence led to further research in the field of 'classical' recursive least squares. There appeared a few solutions to the covariance blow-up problem by imposing the upper bound on the eigenvalues in the covariance matrix [8, 54]. Fortescue et al. in [24] proposed to use a variable forgetting factor, which was a partial solution working well in deterministic cases, but not eliminating the blow-up in stochastic systems, which was shown by Saelid et al. in [72]. This paper also introduced vector-variable forgetting factor for stabilization, however, as Kraus showed in [49] that this method could loose the adaptiveness as well as the Lozano and Goodwin's method of constant trace [53]. Goodwin et al. also proposed periodic resetting of the covariance matrix [28], which indeed avoided the above problems, but at the cost of serious non-optimality due to the information loss.

A promising way to stable estimation with state-space models represented the stabilized linear forgetting (SLF), which was given a particular interest in 1980–1990's, e.g., in [56, 57] etc. There also appeared stabilized exponential forgetting (SEF); its stochastic interpretation appears, e.g., in [46].

Another approach to time-varying parameters are characteristic with their orientation on individual parameter elements. This group of solutions is often called the directional forgetting (DF). The method was extensively developed by Kulhavý in [43] and Kulhavý and Kárný in [44], Kraus in [41], Hägglund in [29] and Cao and Schwartz in [13]. However, in [49] it is shown that the first algorithms lack the convergence.

Another class of algorithms represents the selective forgetting (SF). These are related to the SEF and SLF methods but their initial convergence for non-persistently exciting signals is much better. Their drawback is that they usually require performing numerically expensive eigenvalue decompositions. More on this topic can be found, e.g., in [62].

If we leave the idea of forgetting the old information, there exist yet another approaches to the issue of slowly varying parameters. Some of them are extended and unscented Kalman filtration (EKF and UKF, respectively), neural networks and Markov chain Monte Carlo methods (MCMC) [79, 58], smoothing-based estimation [78], identification based on orthonormal basis functions [77], a confidence region approach [50] and others [35, 36].

1.2 The goal, main contributions and organization

However the issue of the time-varying model parameters was given intensive research effort, the reached results are not always very satisfactory. The solutions suffer either from incapability to

track the parameters under certain conditions (e.g., non-informative data, different variability of parameters...), or from high computational demands, or both. This thesis brings an alternative approach to the problem and tries to fill the gap.
The main goals are:

- To contribute to the subject of tracking of time-varying model parameters in stochastic models and to present a new method, able to steadily track the parameters even in the case of their different variability. The chosen methodology is Bayesian. The method should be rather general and applicable to a wide variety of models.

- To specialize the method to the popular Gaussian models and demonstrate its applicability.

The main contributions of the thesis are as follows:

- The summarization of the existing forgetting techniques in the Bayesian framework is presented. Although there are many parameter tracking methods, only few of them exist in this framework, which is probably due to the more complicated description of the reality with distributions rather than with numbers. Their pros and cons are discussed.

- The method of the partial forgetting is described generally in the terms of any general distributions and their pdfs in particular, which opens the applicability to a wide variety of models – both regressive and non-regressive, Gaussian, Dirichlet... The hypothesis-oriented viewpoint on the parameter distribution is given. This allows the statisticians yet another approach to the parameters – not just as a general statement but a set of different statements, which are mixed together then.

- The partial forgetting method is derived for the very popular Gaussian model, which allows its immediate use in practical situations. The suitability of the derived algorithm is verified in the experimental part of the thesis.

- Proposals for tuning the forgetting are shortly given. The method is very complex, which lays extensive demands on the tuning techniques. Here, just a few ideas about the alternative information and a simplifying modification of the partial forgetting method are given.

The material of this thesis is organized as follows:

Chapter 2 introduces the basics of the Bayesian inference, necessary for understanding the further reading and derivation of the method. It starts with the processes and their probabilistic description, the mathematical system model and the regressive models in particular follow. The sections about the Bayesian learning describe the related theory of parameter estimation and output prediction, the role of the prior information and the model divergence issues. This chapter is built upon the generally known theory.

Chapter 3 brings the deeper insight into the current theory of estimation of slowly time-varying model parameters. A fundamental approach to the issue consists in discarding the old and potentially outdated information from the model. This methodology is called forgetting

and a selected important subclass of its methods is presented. The role of the covariance blow-up, occurring under specific conditions and invalidating the models, is discussed as well.

Chapter 4 deals with the Gaussian model and practically follows the structure of the previous chapter. At first, it introduces the Gauss-inverse-Wishart distribution, serving as the conjugate distribution suitable for Bayesian modelling. Its data update and time update (i.e. forgetting) are given. The methods, introduced in generality in Chapter 3, are specialized to the Gaussian model. The chapter ends with the theory of prediction.

Chapter 5 describes the theory of the partial forgetting method. It is the key part of the thesis, giving the own scientific research results of the author. After the discussion of the general principle, the derivation follows. The structure is the following:

- The role and the meaning of the hypotheses are described. This part is probably the most important, because it decides about the feasibility of the method.
- The information from the particular hypotheses is put together in the form of a finite mixture. This allows to put our knowledge together.
- To avoid the work with a mixture, which is rather complicated, we find its best approximation. This approximation is found as minimizer of the Kullback-Leibler divergence.
- The overall algorithm of the partial forgetting-based parameter estimation is described.
- The method needs an alternative information, three possible sources are given.
- Finally, the idea of the online optimization of weights of hypotheses is described.

Chapter 6 contains the specialization of the method for Gaussian model. As the need for approximation of the digamma function arises, it is given as well.

Chapter 7 checks experimentally the method. As the exponential forgetting is the most popular approach to the tracking of slowly varying parameters, the proposed method is compared to it. The first experiment checks the ability to avoid the covariance blow-up, it exploits a model from literature. The second experiment demonstrates the use of the method, employed with online optimization, on estimation of signal constructed with time-varying parameters. The last experiment uses real data – traffic intensities – to check the ability of the method to track and to predict them.

Chapter 8 discusses the reached results.

Appendix contains the basic and advanced mathematics used in the thesis.

Chapter 2

Bayesian modelling

2.1 Introductory definitions

The theory of stochastic processes is built upon the probability theory. To make the reading of the theory of the partial forgetting as clear as possible, it is necessary to introduce some important concepts of the mathematical nomenclature and related principles. In the further text, the basic knowledge of the measure-theoretic approach to the probability theory and statistics in the extent similar to the first chapter in [73] is supposed.

First, let us introduce the σ-algebra, defined as follows

Definition 1 (σ-algebra)**.** *A collection $\mathcal{F}$ of subsets of a set Ω is called σ-algebra in Ω if $\mathcal{F}$ has the following properties:*

i. $\Omega \in \mathcal{F}$

ii. If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$ where A^C denotes the complement of A relative to Ω.

iii. If $A = \bigcup_{n=1}^{\infty} A_n, A_n \in \mathcal{F}$ for $n = 1, 2, \ldots$, then also $A \in \mathcal{F}$.

An important example is the Borel σ-algebra, generated by the open sets (or equivalently by the closed sets). If $\mathcal{F}$ is a σ-algebra in Ω, then Ω is called a measurable space and the elements of $\mathcal{F}$ are measurable sets in Ω. The measurable space is a space which has a positive measure defined on the σ-algebra of its measurable sets.

Definition 2 (Positive measure)**.** *A positive measure is a function μ, defined on a σ-algebra $\mathcal{F}$, whose range is in $\langle 0, \infty \rangle$ and which is countably additive, i.e., if $\{A_i\}$ are disjoint countable subsets of $\mathcal{F}$, then*

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \tag{2.1}$$

Some important properties of the measure are given in [71].

An important example of a measurable space is the probability space, which is a space with a unit measure.

Definition 3 (Probability space)**.** *A probability space is a triple* $(\Omega, \mathcal{F}, P)$*, which consists of*

Ω *– a set called the sample space,*

$\mathcal{F}$ *– a* σ*-algebra defined on* Ω*,*

P *– a measure called probability,* $P(\Omega) = 1$

The definitions above introduce the probability as a normalized non-negative σ-additive function [27]. Let us now introduce the random variable.

Definition 4 (Random variable)**.** *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space and* $(\mathbb{R}, \mathcal{B})$ *a measurable observation space. The function*

$$X : \Omega \to \mathbb{R}$$

where $\mathcal{B}$ *is a Borel* σ*-algebra, denotes the random variable if it is measurable, i.e., the function satisfies* $X^{-1}(\mathcal{B}) \subset \mathcal{F}$*.*

With this definition, we finally can introduce the notion of random distribution, which will be important in the further reading. First, note that the mapping $P \circ X^{-1}$ is called the distribution of X. The measure-theoretic approach to the random distribution follows [23].

Definition 5 (Random distribution)**.** *Let* $(\mathbb{R}, \mathcal{B})$ *be a measurable space and denote* Λ *the set of probability measures on* $(\mathbb{R}, \mathcal{B})$*. The measurable space of probability measures on* $(\mathbb{R}, \mathcal{B})$ *is* $(\Lambda, \mathfrak{G})$*, where* $\mathfrak{G}$ *is the smallest* σ*-algebra such, that for each* $B \in \mathcal{B}$ *the function from* Λ *to* $\langle 0, 1\rangle$ *is* $\mathfrak{G}$*-measurable. A random distribution on* $(\mathbb{R}, \mathcal{B})$ *is then a measurable function from* $(\Omega, \mathcal{F}, P)$ *to* $(\Lambda, \mathfrak{G})$*.*

The idea of random distribution will be used for definition of the true pdf for derivation of the partial forgetting method. Next, we define the discrete stochastic process [60].

Definition 6 (Discrete stochastic process)**.** *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space. A stochastic process* $X(t)$ *is a collection* $\{X_t | t \in T\}$ *of random variables* X_t *defined on* $(\Omega, \mathcal{F}, P)$*.* T *is an index set of the stochastic process, in the case of discreetness* $T \subseteq \mathbb{Z}^+$*.*

An example of discrete stochastic process is, e.g., a discrete-time random walk. The time series have the property of a discrete stochastic process with a finite (or at most countable) ordered sequence of real numbers [66].
Now, let us introduce the probability density function.

Definition 7 (Probability density function)**.** *Let* $(\Omega, \mathcal{F}, P)$ *be a measure space. A probability density function on* Ω *is a function* $f : \Omega \to \mathbb{R}$ *such that*

- f *is* μ*-measurable,*
- f *is nonnegative* μ*-almost everywhere,*

- *f satisfies*

$$\int_{\Omega} f(x)\,\mathrm{d}\mu(x) = 1. \tag{2.2}$$

Remark 1. *In the further text, we will use the abbreviation pdf for probability density function.*

Remark 2. *The pdf of a random variable X in x is in the classical literature often denoted as $f_X(x)$ but as in the further reading the argument always clearly determines the related random variable, we will omit the subscript.*

In the theory of probability and statistical inference, the notion of expected value plays a significant role. Let us introduce it with the following definition.

Definition 8 (Expected value)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a random variable. The expected value, also known as mean value or expectation, is defined*

$$\mathrm{E}[X] = \int_{\Omega} X \mathrm{d}P \tag{2.3}$$

where the integral is understood as the Lebesgue integral with respect to the measure P.

Not all random variables have an expected value, since the integral may not exist. Some of the important properties of the expected value are:

- Expected value of a constant is equal to the constant itself, i.e.

$$\mathrm{E}[c] = c.$$

- Monotonicity – let X and Y be two random variables, then

$$\mathrm{E}[X] \leq \mathrm{E}[Y] \text{ if } X \leq Y \text{ almost surely}$$

- Linearity – let X and Y be two random variables and a, b real constants, then

$$\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$$

$$\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$$

We can work with either discrete or continuous random variables. While in the first case the appropriate measure is usually the counting measure, in the latter it is the Lebesgue one. As the above definition allows us to use both these measures, we can construct the respective pdfs for both types of the random variables. Moreover, this fact is justified by the Radon-Nikodým theorem (e.g. [73]). Using this fact, let us introduce the simplification of the notation of integrations and instead of $\mathrm{d}\mu(x)$ write shortly $\mathrm{d}x$.

In the statistical inference and decision theory, we usually do not know all properties of the probability space $(\Omega, \mathcal{F}, P)$ and estimate them from samples from the space Ω. A sample is a random vector $X = [X_1, \ldots, X_n]'$, i.e., a vector of random variables, whose realizations (independently measured values) are $[x_1, \ldots, x_n]'$.

For n random variables from the same probability space (usually $\mathbb{R}^n$), it is possible to define a joint probability density function, which reflects them altogether.

Definition 9 (Joint pdf)**.** *The joint pdf of n random variables from the same probability space Ω is a function $f : \mathbb{R}^n \to \mathbb{R}^+$ such that for any measurable domain $D \subset \mathbb{R}^n$*

$$\mathrm{P}([x_1, x_2, \ldots, x_n] \in D) = \int_D f(u_1, u_2, \ldots, u_n)\mathrm{d}u_1\mathrm{d}u_2 \ldots \mathrm{d}u_n. \tag{2.4}$$

Just as in the case of the pdf of a single variable, the joint pdf also satisfies the following requirements

$$f(x_1, x_2, \ldots, x_n) \geq 0,$$

$$\int_{x^*} f(x_1, x_2, \ldots, x_n)\mathrm{d}x_1\mathrm{d}x_2 \ldots \mathrm{d}x_n = 1. \tag{2.5}$$

The relation (2.5) assigns probability one to the probability space. Evidently, one can eliminate any number of random variables from the joint pdf by integrating them out

$$\int_{x^*} f(x_1, x_2, \ldots, x_n)\mathrm{d}x_1 \ldots \mathrm{d}x_{i-1}\mathrm{d}x_{i+1} \ldots \mathrm{d}x_n = f(x_i).$$

Definition 10 (Conditional pdf)**.** *Let us have two random variables with joint pdf $f(a, b)$ and let the marginal pdf $f(b) > 0$. The conditional pdf of random variable A assuming B, denoted $f(a|b)$, is defined by the ratio*

$$f(a|b) = \frac{f(a, b)}{f(b)}. \tag{2.6}$$

Proposition 1 (Bayes' theorem)**.** *For two random variables with conditional pdf $f(b|a)$ and $f(b) > 0$, the Bayes' theorem states*

$$f(a|b) = \frac{f(b|a)f(a)}{f(b)} = \frac{f(b|a)f(a)}{\int f(b|a)f(a)\mathrm{d}a}. \tag{2.7}$$

If we suppose that the random variable A represents any outcome influenced by the random variable B corresponding to relevant piece of evidence or data, the following four pdfs can be defined [4].

Definition 11 (Prior, posterior and predictive pdf, likelihood)**.** *Let a, b be two random variables with joint pdf $f(a, b)$ with marginals $f(a), f(b) \neq 0$. Then, we define*

i. $f(a)$ *– the marginal pdf defining the prior pdf,*

ii. $f(b)$ *– the marginal pdf defining the predictive pdf,*

iii. $f(a|b)$ *– the posterior pdf of the random variable A,*

iv. $f(b|a)$ *– the likelihood of a given b.*

The likelihood $f(b|a)$, often denoted as $\mathcal{L}(a,b)$, is a function reflecting all what an experiment can say about the unknown variable A. It is a function of the second argument when the first argument remains fixed, and it is preserved for any positive proportionality constant $\alpha > 0$, $\mathcal{L}(a,b) = \alpha f(b|a)$. The likelihood is often useful for an equivalent Bayes' theorem expression

$$f(a|b) \propto \mathcal{L}(a,b) f(a),$$

which in verbal expression means that the posterior pdf is proportional to the likelihood and prior pdf. The term $\int f(b|a) f(a) \mathrm{d}b$ is referred to as the normalization integral or normalizing constant of $f(a|b)$, guaranteeing the unity of the the area under the probability density function. It often coincides with the marginal distribution of the data or the prior predictive distribution. In most models and application, it does not have an analytic closed form, which yields the analytic unclosedness of the right-hand pdf as well. This issue led to an enormous literature for computational methods for sampling from the pdf or from the constant [32]. If the true value of one random variable, say A, does not bring any additional information about another random variable B, we can talk about conditional independence. In this case, the following statements come true

$$f(a|b) = f(a), \qquad f(b|a) = f(b),$$

$$f(a,b) = f(a) f(b).$$

Besides the Bayes' theorem, the following additional operations given by the proposition form the very basics of the Bayesian inference.

Proposition 2 (Calculus with pdfs)**.** *Let $f(a,b,c)$ be the joint pdf of the variables A, B and C. The following operations are defined:*

1. *Normalization*

$$\int f(a,b|c) \mathrm{d}a \mathrm{d}b = \int f(a|b,c) \mathrm{d}a = f(b|a,c) \mathrm{d}b = 1, \tag{2.8}$$

2. *Chain rule*

$$f(a,b|c) = f(a|b,c) f(b|c) = f(b|a,c) f(a|c), \tag{2.9}$$

3. *Marginalization*

$$f(b|c) = \int f(a,b|c) \mathrm{d}a, \tag{2.10}$$

$$f(a|c) = \int f(a,b|c) \mathrm{d}b. \tag{2.11}$$

Sometimes the description of a random variable by an unimodal pdf is not sufficient. Therefore, the mixture concept was introduced by Karl Pearson in 1894. Although there is a complete theory around this topic, for the purpose of this thesis only the principle of finite mixture is needed.

Definition 12 (Finite mixture)**.** *Given a set of pdfs* $f_1(x), \ldots, f_n(x), n < \infty$ *and their weights (probabilities)* $w_1, \ldots, w_n$ *such that* $0 \leq w_i \leq 1$ *and* $\sum_{i=1}^{n} w_i = 1$*; the convex combination*

$$f(x) = \sum_{i=1}^{n} w_i f_i(x) \tag{2.12}$$

is a finite mixture density or shortly finite mixture or just mixture. The terms $f_i, i = 1, \ldots, n$ *are called components.*

The convex combination has the important property that it preserves both the non-negativity and the unit area under the mixture.

Remark 3. *All mixtures in the further reading are finite.*

2.2 System and its mathematical model

The system is understood as a part of the reality or the real world. From the observer's view, the following quantities can be distinguished [63]:

- u – inputs to the system, i.e., quantities expediently enforced on the system. If the system is inputless, it is called autonomous.
- y – outputs from the system, i.e., passively observable quantities, which can be directly driven through the preceding inputs only, if ever. They may be influenced by noise and, if the system is regressive, by previous outputs.

The system output is often determined by initial conditions. Both inputs and outputs are observed at discrete time instants t. The couples of inputs and outputs at the same time t compose the data $d_t = [u_t, y_t]'$; a vector of data from time instant 1 to time t describes the evolution of the system. It is denoted with $d(t) = [d_1, d_2, \ldots, d_t]'$. The quantity e denotes the noise, which, in our case, is discrete and white; it will be described in the next section in detail. A basic diagram of the input-output model is depicted in the Figure 2.1. In general, the system which possesses the

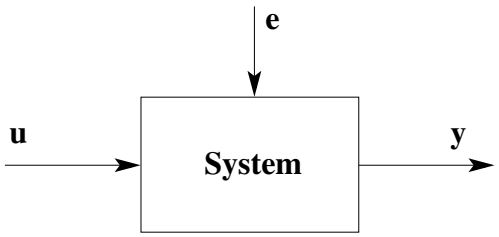


Figure 2.1: Input-output system model

property of superposition is called linear. This means that if the input to the system is a weighted sum of several signals, then the output is simply a weighted sum (superposition) of responses to each of those signals [61]. However, in stochastic systems, this assumption got a weaker form, mainly due to the noise.

2.2.1 Discrete white noise

Definition 13 (Discrete white noise)**.** *The stochastic process* $\{e_t, t \in T\}$ *where* $T \subset \mathbb{Z}^+$ *is said to be white noise if it satisfies*

$$\mathrm{E}\left[e_t\right] = 0 \tag{2.13}$$

$$\operatorname{cov} e_t = \mathrm{E}\left[e_t e_t'\right] = r = \text{const.} \tag{2.14}$$

$$\mathrm{E}\left[e_t e_{t-h}'\right] = \begin{cases} 0 & \text{if } h \neq 0 \\ r & \text{if } h = 0 \end{cases} \tag{2.15}$$

Remark 4. *In other words, the white noise is presented by a sequence of independent identically distributed (i.i.d.) variables with a zero mean value.*

The Definition 13 introduces the discrete white noise in a highly idealized form as a stochastic process with a constant variance $r \in \mathbb{R}^+$, which in the case of a multivariate model leads to a covariance matrix $R \in \mathbb{R}^{n\times n}$, however, this is out of the scope of this thesis.

In the system theory, the discrete white noise is used for modelling of the difference between the system output y_t and its estimate $\hat{y}_t$, conditioned on the past data and the input at the time instant t, i.e. [63]

$$e_t = y_t - \mathrm{E}\left[y_t | u_t, d(t-1)\right] \tag{2.16}$$

$$= y_t - \int y_t f(y_t | u_t, d(t-1)) \mathrm{d}y_t$$

$$= y_t - \hat{y}_t(u_t, d(t-1)). \tag{2.17}$$

With regard to Definition 13 and Equation (2.16), it may be concluded that the additional properties of the discrete white noise are

$$\mathrm{E}\left[e_t | u_t, d(t-1)\right] = 0, \tag{2.18}$$

$$\mathrm{E}\left[e_t u_{t-h}'\right] = 0 \quad \text{if } h = \{0, 1, \ldots, t-1\}, \tag{2.19}$$

$$\mathrm{E}\left[e_t y_{t-h}'\right] = 0 \quad \text{if } h = \{1, \ldots, t-1\}. \tag{2.20}$$

i.e., the zero covariances of the discrete white noise and the input and output indicate, that the noise is independent of the input-output process [63].

2.2.2 Linear regressive models

The linear regressive model expresses the relation between dependent variables (regressands) and the explanatory variables (regressors) in a noisy environment. Generally, it can be described by a conditional pdf in the form

$$f(y_t | u_t, d(t-1), \Theta) = f(y_t | \psi_t, \Theta), \tag{2.21}$$

where the term

$$\psi_t = [u_t, d(t-1)]' \tag{2.22}$$

denotes the column regression vector. Its impact is quantified by the regression coefficients in the column vector $\theta \in \Theta$, where Θ is a set of model parameters. The linear regressive model may be expressed by the relation

$$y_t = \psi_t'\theta + e_t. \tag{2.23}$$

Let the discrete white noise e_t introduced in the previous section have the pdf $f(e_t|u_t, d(t-1))$, then, if (2.19) and (2.20) are fulfilled, we can make the following simplification:

$$f(e_t|u_t, d(t-1)) = f(e_t).$$

Assuming such a pdf to be time-invariant (making the white noise process stationary) density function of a Gaussian distribution, $e_t \sim \mathcal{N}(0, r)$. Then, the model (2.21) is determined by its covariance r and the mean value $\hat{y}_t \equiv \psi_t'\theta$ (cf. Eq. (2.17)) and we talk about the Gaussian regressive model

$$f(y_t|\psi_t, \Theta) \sim \mathcal{N}(\psi_t'\theta, r). \tag{2.24}$$

2.3 Bayesian learning

Let us suppose, that the system model (2.21) is known up to a finite set of parameters Θ, i.e., the conditional probability distributions

$$f(y_t|u_t, d(t-1), \Theta), \quad t = 1, 2, \ldots \tag{2.25}$$

are given. If it is possible to observe the system up to the time instant t and hence keep the data $d(t)$ at disposal, the need of a prediction of the next output y_{t+1} based on the data and, if applied, the next system input u_{t+1} comes into question, e.g., for control purposes, regulation etc. In other words, the need of determination of the distribution $f(y_{t+1}|u_{t+1}, d(t))$ arises but the only known quantity is the history, i.e., data, while the parameter of the model remains unknown. To resolve this issue, we need to reflect the lack of knowledge of the parameter Θ somehow. If we apply the marginalization (2.10) to eliminate it from $f(y_t|u_t, d(t-1), \Theta)$ and use the chain rule (2.9), we obtain

$$f(y_t|u_t, d(t-1)) = \int f(y_t|u_t, d(t-1), \Theta) f(\Theta|u_t, d(t-1)) \mathrm{d}\Theta. \tag{2.26}$$

The first integrand in (2.26) is the model pdf (2.25) and it is defined by the model structure, while the second one, namely $f(\Theta|u_t, d(t-1))$ stands for the posterior pdf.
Apparently, the Bayesian modelling consists of two steps:

1. Choice of the model structure, defining the pdf $f(y_t|u_t, d(t-1), \Theta)$.

2. Estimation of model parameters, given by the posterior pdf $f(\Theta|u_t, d(t-1))$.

The statistical models are intended for modelling and prediction purposes and when the structure is known, their feasibility is mainly based just on the knowledge of the inputs and outputs, while the parameters stay fixed. Hence, there arises the need of explicit expression of the distribution $f(\Theta|u_t, d(t-1))$, which is evidently complicated by the presence of the input variable u_t. Fortunately, under very general conditions, the situation gets significantly simplified due to the effect called 'natural conditions of control'.

2.3.1 Natural conditions of control

If the control strategy does not incorporate more information about the unknown parameters than the past input and output data, then the distribution of the current input is given by

$$f(u_t|d(t-1), \Theta) = f(u_t|d(t-1)). \tag{2.27}$$

This identity is called the natural conditions of control [63]. It means, that the controller uses just the same data, as he would have as an observer and that no knowledge about the unknown model parameters is available to him. Let us consider the following relation for further reading

$$f(u_t|d(t-1), \Theta) f(\Theta|d(t-1)) = f(\Theta|u_t, d(t-1)) f(u_t|d(t-1)), \tag{2.28}$$

where the left-hand side was derived from the right one using the chain rule (2.9).

From identities (2.27) and (2.28) it follows another one, describing the conditional distribution of the model parameters. Indeed, it is conditioned just by past data

$$f(\Theta|u_t, d(t-1)) = f(\Theta|d(t-1)),$$

meaning that the information about the unknown parameters Θ could be extracted from the past data only and the knowledge of u_t does not bring any new information about them [63].

2.3.2 Bayesian parameter estimation and filtration

As it is evident from the previous section, under the natural conditions of control, the variables u_t and Θ are conditionally independent when the past data $d(t-1)$ is given. This fact will be very useful in the following derivation of the Bayesian parameter estimation. Recall the Bayes' theorem and the chain rule (2.9) to obtain the Bayes' rule in the following form

$$f(a|b, c) = \frac{f(b|a, c) f(a|c)}{\int f(b|a, c) f(a|c) \mathrm{d}a}$$

and set $a = \Theta$, $b = y_t$, $c = (u_t, d(t-1))$ and remind that $d_t = (y_t, u_t)$. Then the rule reads

$$f(\Theta|u_t, d(t)) = \frac{f(y_t|d(t-1), \Theta) f(\Theta|u_t, d(t-1))}{\int f(y_t|u_t, d(t-1), \Theta) f(\Theta|u_t, d(t-1)) \mathrm{d}\Theta}.$$

We already know, that under the natural conditions of control, the distribution of the parameter is conditionally independent of the current input u_t, which yields

$$f(\Theta|d(t)) = \frac{f(y_t|u_t, d(t-1), \Theta) f(\Theta|d(t-1))}{\int f(y_t|u_t, d(t-1), \Theta) f(\Theta|d(t-1)) \mathrm{d}\Theta}. \tag{2.29}$$

This estimation procedure is viewed as the data update, because, in fact, it incorporates the new information carried by data d_t into the parameter estimation. Omitting the denominator, the rule in proportional form reads

$$f(\Theta|d(t)) \propto f(y_t|u_t, d(t-1), \Theta) f(\Theta|d(t-1)). \tag{2.30}$$

It is useful to emphasize, that if the parameters are constant (as described above), each data update gradually corrects their distribution and the parameter posterior pdf $f(\Theta|d(t))$ concentrates on a small set or even a point [5].

Sometimes the assumption of parameters' invariance fails and it is necessary to reflect their time evolution $\Theta_t \to \Theta_{t+1}$. This step, called time update, is the key point of various forgetting techniques, when the posterior parameter pdf is manipulated so that it corresponds with the reality more precisely. It has the following form

$$f(\Theta_{t+1}|d(t)) = \int_{\Theta*} f(\Theta_{t+1}|d(t), \Theta_t) f(\Theta_t|d(t)) \mathrm{d}\Theta_t. \tag{2.31}$$

If the parameters were time-invariant ($\Theta_{t+1} = \Theta_t$), the time update would present a rather formal step. The first pdf in (2.31) represents an explicit model of parameters evolution, however, it is usually not available. Therefore, we usually employ other techniques of time updating by modifying the whole right-hand side of the equation. That approach will be discussed in the next chapter.

The combination of the data update (2.30) and the time update (2.31) is called Bayesian filtration of unknown parameters Θ_t. The observed data incorporated in the data update represents the only connection of the model to the reality.

Note, that in the case of constant parameters, the posterior pdf after the data update may be viewed as the prior pdf for the next estimation step. If the parameters vary and the time-update takes place in the estimation, then the posterior pdf after this turns into the prior pdf to the data update. From the above reading follows, that the (Gaussian) distribution of parameters during the modelling is preserved by multiplication of appropriate pdfs. Obviously, multiplication by zero or nearly zero value should not occur – such an event usually indicates a modelling or a measurement error.

2.3.3 Discussion of estimation and prediction

Recall the recursive parameter estimation relation (2.29) and let us reuse it to find a prediction of the output y_{t+1}, which is supposed to occur at the next time instant $t + 1$. Let us focus on the denominator of (2.29) to realize, that it can be viewed as a Bayesian one-step-ahead predictor. Indeed, (2.29) can be rewritten into two parts:

- the predictor

$$f(y_t|d(t-1)) = \int f(y_t|u_t, d(t-1), \Theta) f(\Theta|d(t-1))\mathrm{d}\Theta, \tag{2.32}$$

- the parameter estimator (which employs the predictor)

$$f(\Theta|d(t)) = \frac{f(y_t|u_t, d(t-1), \Theta)}{f(y_t|u_t, d(t-1))} f(\Theta|d(t-1)). \tag{2.33}$$

The predictor uses the parameter estimate computed from the past data $d(t-1)$ to seek the new distribution of y_t, using the currently determined input u_t. The obtained output y_t substituted with the input into the distribution $f(y_t|u_t, d(t-1))$ leads to a possibility to calculate the fraction in (2.33), i.e.

$$g(\Theta) = \frac{f(y_t|u_t, d(t-1), \Theta)}{f(y_t|u_t, d(t-1))},$$

where in the numerator we can use all possible values of the parameter Θ on condition of fixed data. Calculation of the parameter estimate should theoretically be a simple task then, however, it is not. The first trouble lies in the possible high multidimensionality of the unknown parameter Θ, the second one in the need of recomputing the factor $g(\Theta)$ in each time step t. Both these facts may lead to an extensive computational burden, in particular if this approach is used in embedded devices, where the memory and the performance is a key. A relatively simple solution represents the use of conjugate distribution families and sufficient statistics, which ensure the reproduction of the distribution and reduction of the problem dimensionality. This concept will be described for a selected class of models further in this chapter.

If the purpose of the system is not the parameter estimation but the output prediction, there arises the question whether and how it is possible to prepare such a theoretical concept, that omits the parameter estimation and allows to directly update the conditional pdf for the next output. Let us rewrite the relation (2.30) into the form with likelihoods

$$f(\Theta|d(t)) = \frac{\mathcal{L}_t(\Theta, d(t)) f(\Theta|d_0)}{\int \mathcal{L}_t(\Theta, d(t)) f(\Theta|d_0)\mathrm{d}\Theta}, \tag{2.34}$$

where

$$\mathcal{L}_t(\Theta, d(t)) = \prod_{\tau=t_0+1}^{t} f(y_\tau|u_\tau, d(\tau-1), \Theta) \tag{2.35}$$

is the likelihood combining the model likelihoods from time $t_0 + 1$. The instant t_0 does not need to be the real beginning of the stochastic process, the Bayesian approach allows to evaluate the likelihood from data between the time t_0 and t; $t > t_0$ [63]. The predictor (2.32) with (2.34) substituted for $f(\Theta|d(t))$ is for the future time instant $t + 1$

$$f(y_{t+1}|u_{t+1}, d(t)) = \frac{\int f(y_{t+1}|u_{t+1}, d(t), \Theta)\, \mathcal{L}_t(\Theta, d(t))\, f(\Theta|d(t))\mathrm{d}\Theta}{\int \mathcal{L}_t(\Theta, d(t))\, f(\Theta|d(t_0))\mathrm{d}\Theta}.$$

From the conditional likelihood (2.35) it can be seen, that the first two terms in the nominator integral produce a new likelihood

$$f(y_{t+1}|u_{t+1}, d(t), \Theta)\, \mathcal{L}_t(\Theta, d(t)) = f(y_{t+1}|u_{t+1}, d(t), \Theta) \times \prod_{\tau=t_0+1}^{t} f(y_\tau|u_\tau, d(\tau-1), \Theta)$$

$$= \mathcal{L}_{t+1}(\Theta, d(t+1)),$$

hence

$$f(y_{t+1}|u_{t+1}, d(t)) = \frac{\int \mathcal{L}_{t+1}(\Theta, d(t+1)) f(\Theta|d(t))\mathrm{d}\Theta}{\int \mathcal{L}_t(\Theta, d(t))\, f(\Theta|d(t_0))\mathrm{d}\Theta} \tag{2.36}$$

$$= \frac{I_{t+1}(d(t+1))}{I_t(d(t))}, \tag{2.37}$$

where

$$I_t(d(t)) = \int \mathcal{L}_t(\Theta, d(t))\, f(\Theta|d(t_0))\mathrm{d}\Theta.$$

This shows, that under certain conditions, it should be possible to find a posterior distribution for the next (predicted) output. However, this supposes that we are able to express the prediction integrals in (2.37) as a function of the extended regression vector $\Psi_t = [y_t, \psi_t']'$. Fortunately, for a certain class of models, this is possible. As this thesis deals with the Gaussian models, it will be shown that the predictive pdf for such a class of models is the Student's one.

2.3.4 Prior pdf

The prior pdf $f(\Theta)$ quantifies prior (expert) knowledge or belief about an unknown random variable a priori, before the new data was observed and incorporated into the variable distribution. The prior pdf enters the Bayes' theorem on the right-hand side and its multiplication with the likelihood produces the poster pdf. In general, the prior pdf is often more flat than the likelihood, which sharpens by the Bayes' theorem. In the Bayesian framework, the parameter pdf is characterized by the hyperparameter, to avoid confusion of terms.

The role of the prior information becomes important in the very beginning of the modelling, when either no data was produced yet or when the modelling started later than the data generating process and the initial data was not registered. Let us denote the time when the modelling started as t_0 and suppose, that the sampling is running for $t \gg t_0$. Then the data $d(t_0)$ can be simply neglected, as the information in $d(t)$ dominates and the approximation

$$f(\Theta|d(t_0)) \approx f(\Theta)$$

is well acceptable if the data is informative [63]. And we can proceed further with this topic yet, introducing the concept of non-informative prior. Later, in the following chapters, we will use the convenient properties of the prior pdf for forgetting.

2.3.5 Non-informative prior and stable estimation

The principle of non-informative prior information rises from the attitude 'Let the data speak for themselves' [63, 4]. The formal idea suggests, that there is only a little known a priori and the prior distribution should reflect this state, instead of introducing any potentially bad belief into the modelling process. During the runtime, as the new informative data is observed and incorporated into the distribution, it will get its proper shape and peak(s).

The principle of stable estimation significantly eliminates the problems with an improper or non-informative prior distribution, because it leads to its transformation by the data bringing information into the model. From this point of view, it may seem evident that the most convenient prior distribution is the uniform distribution, assigning all the values on the support the same probability measure. However, this approach may lead to problems, especially if the parameter is real, when the normalization may not be possible [70]. This problem can be circumvented by using a very flat Gaussian prior with zero mean, i.e., a Gaussian distribution with a very large variance. In [9] it is shown, that the uniform prior distribution leads to identically the same parameter estimates as the Gaussian prior in the limit as the hyperparameter (variance) approaches infinity.

2.3.6 Sufficient statistics and conjugate prior

The notion of sufficient statistics is very handy or even necessary for processing large amount of data as it gets impractical or impossible to work with it directly. Therefore, the data is 'transformed' into a set of quantities of a smaller non-increasing dimension [63], which are functions of the data and which carry the same information for estimation as the data themselves. Let $V_t = V_t(d(t))$ be the sufficient statistic for the data. Then it must be true that

$$f(x|d(t)) = f(x|V_t). \tag{2.38}$$

where x is a realization of some random variable. If there exists a sufficient statistic, then the relation (2.38) keeps the same dimension regardless on the time index t.

In modelling, we are generally interested not only in the unknown parameters Θ but in the prediction of the future outputs y_{t+1} as well. Each of these variables can possess its own sufficient statistics, however, existence of the former does not necessarily induce the existence of the latter but fortunately for a certain class of system models these statistics exist [63]. Such models have the special property, that during the runtime the distribution of the parameters and the outputs is reproduced as they are recursively updated. Recall again from (2.30) the data update

$$f(\Theta|d(t)) \propto f(y_t|u_t, d(t-1), \Theta) f(\Theta|d(t-1)).$$

To make the problem tractable, we need the prior pdf $f(\Theta|d(t-1))$ and the posterior pdf $f(\Theta|d(t))$ belong to the same functional form. Such a distribution is called self-reproducing [63] or more frequently conjugate [5].

The topic of conjugate priors was discussed, e.g., in [69]. As it already has been written above, the existence of the conjugate prior distribution fully depends on the existence of the

sufficient statistic. It is possible to prove, that the existence of a set of fixed-dimensional sufficient statistics implies the existence of the conjugate prior family [69, 19]. Once the family is known, it remains to determine which one of its members represents the prior information best. To this point, one needs to find the hyperparameters of the conjugate distribution [22]. It may be shown, that if the data generating process belongs to the exponential family [10, 22] and meets certain conditions, then there exists a set of sufficient statistics.

2.4 Information divergence

For the purpose of this thesis, the Kullback-Leibler divergence serves as a measure of the information divergence, or, in other words, a dissimilarity of two distributions characterized by their probability density function. Let us formulate its definition:

Definition 14 (Kullback-Leibler divergence)**.** *Let f and g be two pdfs of a random variable X, acting on a common set x^*. The Kullback-Leibler divergence, also known as relative entropy, is defined as*

$$\mathrm{D}\left(f||g\right) = \int_{x^*} f(x) \ln \frac{f(x)}{g(x)} \mathrm{d}x. \tag{2.39}$$

In our application it measures the divergence of a pair of pdfs f and g, acting on a set x^*. However, it cannot be considered as a distance measure, since it does not satisfy neither the symmetry, $\mathrm{D}\left(f||g\right) \neq \mathrm{D}\left(g||f\right)$, nor the triangle inequality. Some interesting properties of the Kullback-Leibler divergence are stated in the following corollary.

Corollary 1 (Information inequality)**.** *The Kullback-Leibler divergence has the following important properties*

- $\mathrm{D}\left(f||g\right) \geq 0$.
- $\mathrm{D}\left(f||g\right) = 0$ *if and only if* $f(x) = g(x)$ *almost everywhere on* x^*.
- $\mathrm{D}\left(f||g\right) = \infty$ *if and only if* $g(x) = 0$ *and* $f(x) > 0$ *on a set of positive Lebesgue measure.*

The proof of these statements can be found, e.g., in [47].

The Kullback-Leibler divergence can be used in various scenarios, for instance, as the information gain in Bayesian inference, measuring the innovation when moving from the prior pdf to the posterior (the Bayesian d-optimal design) or, as later in our case, for approximation by its minimization. The Kullback-Leibler divergence has a strong relation to information and entropy measures, e.g., the self-information or the Shannon entropy [18].

Definition 15 (Kerridge inaccuracy)**.** *Let f and g be two pdfs of a random variable X. The Kerridge inaccuracy is defined as*

$$\mathrm{K}\left(f, g\right) = \int f(x) \ln \frac{1}{g(x)} \mathrm{d}x. \tag{2.40}$$

The Kerridge inaccuracy is particularly useful for finding such a pdf g, that is the closest to f in the Kullback-Leibler sense. The proof is trivial and therefore omitted.

2.5 Estimation divergence issues

Even though the estimation theory is correct, the filter (2.30) and (2.31) may not work properly. There are two possible causes for the failure [74]:

- finite precision arithmetic
- modelling errors

The finite precision arithmetic in computers leads to roundoff errors, when the digital processor approximates the infinite set of real numbers with a finite set [55]. Since the covariance matrices are often very badly conditioned, the risk of producing inappropriate results increases, when they are directly manipulated. Even a small perturbation connected with the rounding may lead to a very large change of the results. To solve this issue, multiple methods for handling with the ill-conditioned matrices are employed. They are mainly based on increasing of the precision (from float to double or even quadruple), or the matrix factorization, e.g., the L'DL and LDL' factorizations, Cholesky factorization, UD factorization, Gram-Schmidt orthogonalization procedure, Givens and Householder transformations and others [74, 7, 12].

The other cause, the modelling errors, arises for many different reasons. The model is often treated as known, i.e., its order, noise properties, covariances etc. were successfully determined before the modelling. If any of these assumptions are violated, the filter may start to diverge.

Besides the numerical improvements, we can use several strategies to face the specified problems, e.g. [74]:

- use a fading-memory filter, which is a filter admitting the change of the reality (parameters) and therefore placing more emphasis on more recent measurement by forgetting the measurements from the distant past. This theoretically results in a non-optimal filter but it may restore its convergence and stability. Indeed, it is better to have a convergent and stable filter, which is sub-optimal but works, than to have an optimal yet nonfunctional one. There are many methods focused on this problematics and this thesis introduces one of them.
- use fictious process noise – it is very easy to implement addition of fictious noise. In fact, it expresses the statistician's lack of confidence in the filter. Such a filter places more emphases on the measurements than a usual filter [34].

If we deal with forgetting, the mismodelling issue is often tight with the covariance blow-up phenomenon, when the covariance matrix eigenvalues quickly increase without bounds (see Section 3.1). This situation occurs, when the old information is discarded by the forgetting without incorporating any new information. We will widely discuss this topic in the following chapter.

Chapter 3

Estimation of time-varying parameters

In a real situation, the assumption that a set of model parameters is constant in time is fulfilled only approximately and/or temporarily. Moreover, as the mathematical model is only an approximative description of the reality, it may happen that for chosen model structure, more or less time-varying parameters would be appropriate during the time run. The reflection of the parameter changes turns the parameter estimation into the parameter tracking [63].

The time variability of the parameters can be of three types:

a) Constant parameters

b) Slowly time-varying parameters

c) Fast time-varying parameters

The case of constant parameters is the simplest, as the relation $\Theta_t = \Theta_{t-1}$ holds for each time instant $t \in t^*$. The estimation of constant parameters was thoroughly presented in the previous chapter.

All the cases of parameters variability, but the varying parameters in particular, may be generally modelled by the conditional probability distribution for Θ_t on condition on Θ_{t-1} and (possibly) previous data, that occurs in the predictive part of the Bayes' rule

$$f(\Theta_t|d(t-1)) = \int f(\Theta_t|\Theta_{t-1}, d(t-1)) f(\Theta_{t-1}|d(t-1)) \mathrm{d}\Theta_{t-1}, \tag{3.1}$$

which results in a flattening of the pdf $f(\Theta_{t-1}|d(t-1))$ to obtain the new pdf $f(\Theta_t|d(t-1))$ [63, 40].

The rather vague term 'slowly varying parameters' describes the case when the true parameter value Θ_t does not lay far from Θ_{t-1}, thus $\Theta_t \approx \Theta_{t-1}$. In this case, instead of a direct use of the formula (3.1) and trying to find the explicit time-evolution model, which would be often too complicated or even impossible, it is more convenient to increase the uncertainty of belief in the old estimate by manipulating its pdf $f(\Theta_{t-1}|d(t-1))$. Such an approach is called forgetting and it substitutes the time-update step (2.31) in the case of time-variant parameters.

3.1 Covariance blow-up in general and ARX application

The most basic forgetting methods are accompanied with the phenomenon of the covariance matrix blow-up. It occurs, if the observation and forgetting updates are not well balanced and the increase of the trace of the parameter covariance matrix due to forgetting is not compensated by the data updates. The covariance grows without bounds for less exciting or non-exciting signals, which in fact causes unbounded grow of the estimation error variance and therefore decrease in the reliability of the estimates. From another viewpoint, it can be presented as a loss of information stored in the covariance matrix. In addition, besides the unreliable estimation, the blow-up conditions lead to numerical difficulties like overflows and roundoff errors [56].

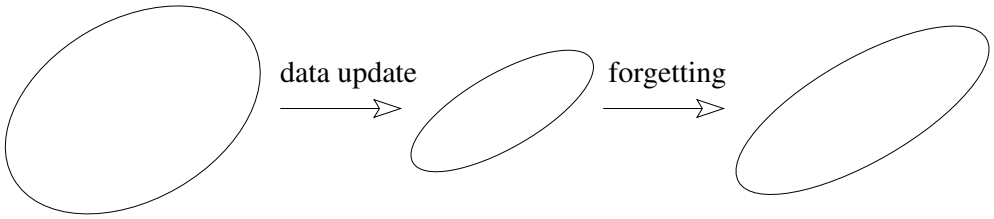


Figure 3.1: Evolution of the covariance matrix due data update and time update (exponential forgetting). The ellipse represents an equiprobability curve. With incoming informative data, the ellipse gets more concentrated, while the forgetting has the opposite effect. If non-informative data is coming, then forgetting causes a potentially enormous increase of the covariance matrix, because it is not compensated by the data updates.

Apart from the Bayesian estimation based on the batch of recent data, several techniques reflecting the slow change in the model parameters values exist. This chapter briefly describes four forgetting methods – the exponential forgetting [63, 34], which is the most basic yet popular, alternative forgetting based on it [45, 40], linear forgetting [45, 56] typical rather for state-space models and directional forgetting [43, 44]. All these methods are described in the Bayesian framework, however, there exist several non-Bayesian approaches to the issue, e.g. [13, 35]. These will be not discussed, as they are out of our focus. The rest of the chapter summarizes the requirements imposed on the partial forgetting.

The majority of the methods developed to suppress or completely eliminate the blow-up conditions are ARX-related. Some of them were introduced in the introductory part of this thesis. While most of them deal with the state-space models (e.g., RLS or Kalman filter), the Bayesian input-output models did not attain such an intensive focus. However, the Bayesian view is much more general in terms of its applicability to any parametric model $f(y_t|u_t, d(t-1), \Theta_{t-1})$ and any $f(\Theta_{t-1}|d(t-1))$. This universality is one of the main advantages of the partial forgetting method.

3.1.1 Exponential forgetting

The exponential forgetting (EF), also known for Gaussian ARX models as time-weighted least squares (TWLS) [34], or flattening the posterior pdf [63], dominates the methods of solution of slowly time-variant-parameters issue. This approach introduces a new 'tuning knob' $\lambda \in \langle 0, 1\rangle$ called forgetting factor[1]. This factor causes the flattening of the pdf by its exponentiation

$$f(\Theta_t|d(t-1)) \propto [f(\Theta_{t-1}|d(t-1))]^{\lambda}. \tag{3.2}$$

For the exponential forgetting it is typical, that it does not reflect the character of measurements obtained during the time run, and causes uniform forgetting of all the old information in the parameter pdf. Although this method is the most popular one, it has the highest liability to the covariance blow-up, occurring when the forgetting and learning rates are not well balanced, and the useful information is continually suppressed.

There appeared many other methods, based on the exponential forgetting, which tried to solve its limitations, e.g., stabilized exponential forgetting, restricted exponential forgetting [42] etc.

3.1.2 Alternative forgetting

Sometimes the exponential forgetting is viewed as an optimization problem of 'balancing' two parameter pdfs f_1 and f_2. Such a method is called alternative or stabilized exponential forgetting. It introduces an alternative distribution of parameters, which has the pdf f_2. This distribution expresses our knowledge of the possible alternative behaviour of parameters. Thus, by regarding this alternative information, the method allows to track them with higher stability than the basic exponential forgetting. The estimate of the true pdf $\hat{f}$, describing the distribution of model parameters, then equals to the weighted geometric mean of f_1 and f_2 according to the following proposition [45, 40].

Proposition 3. *Let an unknown true pdf f, describing the distribution of unknown parameters, be equal to pdf f_1 with probability $\lambda \in \langle 0, 1\rangle$ and to pdf f_2 with probability $1-\lambda$. Let the pdfs f_1 and f_2 be mutually non-orthogonal. Then, the relation*

$$\hat{f}(\Theta) \propto [f_1(\Theta)]^{\lambda}[f_2(\Theta)]^{1-\lambda} \tag{3.3}$$

describes the estimate of f, obtained as a solution of the optimization problem

$$\min_{f}[\lambda \mathrm{D}\left(f||f_1\right) + (1-\lambda)\mathrm{D}\left(f||f_2\right)]. \tag{3.4}$$

[1] Usually, the forgetting factor of the exponential forgetting λ is not lower than 0.95

Proof. Let us denote $\lambda_1 = \lambda$ and $\lambda_2 = 1 - \lambda$ and rewrite the minimized functional in (3.4) as follows

$$\sum_{i=1}^{2} \lambda_i \left(\int f(\Theta) \ln \frac{f(\Theta)}{f_i(\Theta)} \mathrm{d}\Theta \right) = \int f(\Theta) \ln \frac{f(\Theta)}{\frac{\prod_{i=1}^{2} \lambda_i f_i(\Theta)}{\int \prod_{i=1}^{2} \lambda_i f_i(\Theta) \mathrm{d}\Theta}} \mathrm{d}\Theta$$

$$- \ln \int \prod_{i=1}^{2} \lambda_i f_i(\Theta) \mathrm{d}\Theta = \mathrm{D}\left(f \middle|\middle| \hat{f} \right) + \text{constant}$$

where the constant is independent of the optimized f. The assumption of non-orthogonality of the two pdfs f_1, f_2 ensures that the integral $\ln \int \prod_{i=1}^{2} \lambda_i f_i(\Theta) \mathrm{d}\Theta > 0$. The properties of the Kullback-Leibler divergence given in Corrolary 1 ensure that the optimization has only one solution $f = \hat{f}$. □

In practical use, the pdf f_1 is usually obtained after the data update (2.30), while the other one (f_2) is any appropriate (preferably flat, e.g., the prior) pdf. The relation (3.3) defines the time update step. Here, the issue of EF arises – if the pdf f_2 is proportional to constant, then the basic exponential forgetting is obtained. This explains its sensitivity – we admit arbitrary position of Θ_t with probability λ_2.

The exponential forgetting method is the most basic approach to time-variant parameters, however, it lacks some useful properties like adaptability and ability to track multiple parameters which vary each with different rates. Another problem arises from the so-called estimator windup [13] which occurs if the system input is not persistently excited [2]. To solve the drawbacks, some modification of it were developed.

3.1.3 Linear forgetting

The linear forgetting represents another approach to time-varying parameters. Like the exponential forgetting method, this one is also achieved as an optimization problem, which is even dual to the exponential forgetting [45]. The drawback of this method is that it usually does not preserve the distribution family. On the other side, its advantage consists in the fact, that it uses the proper form of Kullback-Leibler divergence for the approximation [4, 39].

Proposition 4. *Let an unknown true pdf f, describing the distribution of unknown parameters, be equal to pdf f_1 with probability $\lambda \in \langle 0, 1 \rangle$ and to pdf f_2 with probability $1 - \lambda$. Let the pdfs f_1 and f_2 be mutually non-orthogonal and have the same support Θ^*. Then, the relation*

$$\hat{f}(\Theta) \propto \lambda f_1(\Theta) + (1 - \lambda) f_2(\Theta) \tag{3.5}$$

describes the estimate $\hat{f}$ of f, obtained as a solution of the optimization problem

$$\min_{f} [\lambda \mathrm{D}\left(f_1 || f\right) + (1 - \lambda) \mathrm{D}\left(f_2 || f\right)]. \tag{3.6}$$

Proof. Let us denote $\lambda_0 = \lambda$ and $\lambda_1 = 1-\lambda$ and rewrite the optimized functional (3.6) as follows

$$\sum_{i=1}^{2} \lambda_i \left(\int f_i(\Theta) \ln \frac{f_i(\Theta)}{f(\Theta)} \mathrm{d}\Theta \right) = \int \left(\sum_{i=1}^{2} \lambda_i f_i(\Theta) \right) \ln \frac{\sum_{i=1}^{2} \lambda_i f_i(\Theta)}{f(\Theta)} \mathrm{d}\Theta$$

$$- \int \left(\sum_{i=1}^{2} \lambda_i f_i(\Theta) \right) \ln \left(\sum_{i=1}^{2} \lambda_i f_i(\Theta) \right) \mathrm{d}\Theta + \sum_{i=1}^{2} \lambda_i \left(\int f_i(\Theta) \ln f_i(\Theta) \mathrm{d}\Theta \right)$$

$$= \mathrm{D}\left(\hat{f} \middle|\middle| f \right) + \text{constant}$$

where the constant is independent of the optimized f. The properties of the Kullback-Leibler divergence given in Corrolary 1 imply that $f = \hat{f}$ is the only solution of the optimization problem. □

Remark 5. *If we search for the estimate of the optimization* (3.5) *in a specific form, we talk about the restricted linear forgetting. Such a case occurs, e.g., if f_1 and f_2 belong to the class of pdfs conjugated to the exponential family, while the mixture* (3.5) *does not [45].*

3.1.4 Directional forgetting

The directional forgetting (DF) was first introduced in the research report [43], followed by paper [44]. The principle of the method consists in a transformation

$$\phi = \{\phi_1, \phi_2\} = F_t(\Theta), \qquad \phi_1 \in S_1; \phi_2 \in S_2 \tag{3.7}$$

where $F_t(\Theta)$ is a (time-dependent) function of the parameters Θ, which maps the multivariate parameter into two complementary subspaces S_1 and S_2. On the first subspace the probability distribution remains unchanged by the data update, i.e., it is spanned on the non-excited directions. The other subspace S_2 serves for the rest of parameters, for that the exponential forgetting is applied. The two subspaces should be stochastically independent, because only in that case a flattening of one pdf does not influence the second one [44]. Recall (3.7) and write

$$f_t(\phi) = f_t(\phi_1) f_t(\phi_2) \tag{3.8}$$

where f_t is a pdf of the argument (transformed parameter) at time t. Because the new data do not correct the pdf of ϕ_1

$$f_t(\phi_1) = f_{t-1}(\phi_1).$$

When the largest parameters subspace uninfluenced by data is found, then the directional time-updating rule reads

$$f_t(\phi) = f_t(\phi_1) f_t(\phi_2)$$

$$= f_{t-1}(\phi_1) [f_{t-1}(\phi_2)]^{\lambda},$$

where $\lambda \in \langle 0, 1\rangle$ stands for the exponential forgetting factor (c.f. exponential forgetting, Section 3.1.1).

After performing the directional forgetting step, the original parameter space is to be reached by the inverse mapping

$$\Theta_t = F_t^{-1}(\phi).$$

Remark 6. *A similar idea was introduced by Häggelund in [29], however for ARX models only, hence it lacks the generality of that one given above [43].*

3.2 Summary of desired properties of the forgetting

If we pick up the advantages and disadvantages of the selected methods presented above from the related literature, we can write down a summary of reasonable requirements imposed on forgetting:

- covariance matrix blow-up elimination – the proposed algorithm should not suffer from the covariance blow-up phenomenon. There should be a mechanism to eliminate it.
- numerical stability – the algorithm should be stable and lead to well-conditioned covariance matrices. It should not suffer from roundoff errors and the matrix inversions should occur rather rarely.
- configurability and adaptability – the method must allow tuning to reflect the modelled reality.
- universality – a method, which fulfills this requirement, is applicable in a wider class of real problems, than any specialized method. This point can be reached using the Bayesian framework.
- Bayesian framework – the Bayesian methods fix the data, not the parameters as the frequentists methods. This allows to infer even from a small batches of data.

The new forgetting method should fulfill these requirements.

Chapter 4

Parameter tracking for ARX model

The Gaussian (normal) distribution is central to statistical inference and modelling [17]. It is fully characterized by its two parameters – the mean value as the measure of location and the variance, describing the scatter around the mean. The central limit theorem from both the frequentists' and Bayesian viewpoint [6] justifies the use of this distribution as a suitable approximation for the posterior probability distribution even in the case of non-normality. In [17] one can find useful references to the literature handling with multimodal, skewed and another cases, in which the Gaussian distribution may be (limitly) used.

The parameterized Gaussian pdf for normal model is defined as follows.

Definition 16 (Gaussian pdf)**.** *The pdf of the Gaussian model has the following form*

$$\mathcal{N}(\theta'\psi, r) \equiv \frac{1}{\sqrt{2\pi r}} \exp\left\{-\frac{1}{2r}(y-\theta'\psi)^2\right\} \tag{4.1}$$

$$\equiv \frac{1}{\sqrt{2\pi r}} \exp\left\{-\frac{1}{2r}\operatorname{Tr}(\Psi\Psi'[-1,\theta']'[-1,\theta'])\right\}, \tag{4.2}$$

where

y *is the scalar data,*

ψ *is the regression vector,*

θ *is the vector of regression coefficients,*

r *is the variance (also denoted by* σ^2*),*

$\Psi = [y, \psi']'$ *is the extended regression vector*

The Gaussian model parameters are $\Theta = \{\theta, r\}$. If we employ it in the Bayesian modelling, there a need for a convenient conjugate prior distribution arises (cf. Section 2.3.6). We use the inverse-Wishart distribution, whose pdf is defined as follows.

Definition 17 (inverse-Wishart pdf)**.** *Let X be an $m \times n$ matrix whose rows are independently normally distributed with zero mean. Then, the distribution of $\Xi = (X'X)^{-1}$ is inverse-Wishart with matrix V and degrees of freedom ν, thus $\Xi \sim i\mathcal{W}(V, \nu)$ with pdf*

$$iW(V,\nu) \equiv \frac{|V|^{\frac{\nu}{2}}|\Xi|^{-\frac{\nu+n+2}{2}} \exp\left[-\frac{\mathrm{Tr}(V\Xi^{-1})}{2}\right]}{2^{\frac{\nu n}{2}}\Gamma_n\left(\frac{\nu}{2}\right)} \tag{4.3}$$

where

$$\Gamma_n(a) \equiv \pi^{\frac{n(n-1)}{4}} \prod_{j=1}^{n} \Gamma\left(a + \frac{1-j}{2}\right) \tag{4.4}$$

is the multivariate gamma function [33].

Usually, the rows of the matrix X represent the independent n-variate normally distributed variables $[x_1, \ldots, x_m]'$. The inverse-Wishart distribution is a multivariate generalization of the inverse-gamma distribution, which, in turn, is a generalization of the inverse χ^2 distribution for non-integer degrees of freedom.

4.1 Gauss-inverse-Wishart distribution

If we assume normality of the model (2.21), we can model the parameters by Gauss-inverse-Wishart ($\mathcal{G}i\mathcal{W}$) distribution defined as follows [40]:

Definition 18 (Gauss-inverse-Wishart pdf)**.** *The pdf of the Gauss-inverse-Wishart distribution has the form*

$$\mathcal{G}i\mathcal{W}(V,\nu) \equiv \frac{r^{-0.5(\nu+n+2)}}{\mathcal{I}(V,\nu)} \exp\left\{\frac{-1}{2r}\begin{bmatrix}-1\\ \theta\end{bmatrix}' V \begin{bmatrix}-1\\ \theta\end{bmatrix}\right\} \tag{4.5}$$

or

$$\mathcal{G}i\mathcal{W}(L,D,\nu) \equiv \frac{r^{-0.5(\nu+n+2)}}{\mathcal{I}(L,D,\nu)} \exp\left\{\frac{-1}{2r}\left[(\theta-\hat{\theta})'C^{-1}(\theta-\hat{\theta}) + D_{LSR}\right]\right\} \tag{4.6}$$

where

$$V = L'DL = \begin{bmatrix}1 & 0\\ L_{21} & L_{22}\end{bmatrix}' \begin{bmatrix}D_{11} & 0\\ 0 & D_{22}\end{bmatrix} \begin{bmatrix}1 & 0\\ L_{21} & L_{22}\end{bmatrix} \tag{4.7}$$

and where the individual terms have the following meaning:

V *is the extended information matrix, i.e., symmetric square $n \times n$ dimensional positive definite matrix, which carries the information about the past data important for estimation. By its $L'DL$ factorization, the terms L and D are obtained.*

ν *stands for the degrees of freedom (counter),*

n *denotes the length of the extended regression vector Ψ,*

r *is the model noise variance,*

$\hat{\theta} \equiv L_{22}^{-1} L_{21}$ *is the least-squares (LS) estimate of* θ*,*

$C \equiv L_{22}^{-1} D_{22}^{-1} (L_{22}^{-1})'$ *is the covariance of estimate of* θ*,*

$D_{LSR} \equiv D_{11}$ *is the LS remainder*

$\mathcal{I}$ *stands for the normalization integral*

$$\mathcal{I}(L, D, \nu) = \Gamma(0.5\nu) \sqrt{\frac{2^{\nu} (2\pi)^n}{D_{LSR}^{\nu} \det(D_{22})}} \,. \tag{4.8}$$

The variance of the $\mathcal{G}i\mathcal{W}$ distribution comes from the inverse-Wishart distribution with the least-squares remainder and the number of degrees as parameters, $i\mathcal{W}(D_{LSR}, \nu)$. Its moments are

$$\mathrm{E}\left[r|L, D, \nu\right] = \frac{D_{LSR}}{\nu - 2} \equiv \hat{r} \tag{4.9}$$

$$\mathrm{var}[r|L, D, \nu] = \frac{2\hat{r}^2}{\nu - 4} \tag{4.10}$$

$$\mathrm{E}\left[r^{-1}|L, D, \nu\right] = \frac{\nu}{D_{LSR}} \tag{4.11}$$

The $L'DL$ factorization of the extended information matrix V allows to increase the numerical stability of related computations and therefore it plays an important role. Its definition is given in Appendix A.3.

4.2 Parameter estimation in Gaussian model

Now, we have to reflect the data update (2.30), when new data is introduced into the probability density function $f(\theta|d(t))$. In the case of model normality it means, that we need to update the Gauss-inverse-Wishart pdf in two steps:

1. update the extended information matrix V, and
2. update the counter (degrees of freedom)

If the update does not use any weighting of the new data, which is equivalent to no forgetting, the procedure is simply straightforward as shows the following proposition.

Proposition 5 (Data-update of $\mathcal{G}i\mathcal{W}$ pdf)**.** *Let* $f(\theta|d(t)) \sim \mathcal{G}i\mathcal{W}(V, \nu)$ *where* V *is the extended data matrix and* ν *are the degrees of freedom. Let* $\Psi_t = [y_t, \psi_t']'$ *where* y_t *is the last output and* ψ_t *is the corresponding regression vector. Let us use the time indices to denote the transition from*

the time instant $t-1$ to t. Then, the data update of the $\mathcal{G}i\mathcal{W}$ pdf is equivalent to the rank-one update and has the form

$$V_t = V_{t-1} + \Psi_t \Psi_t' \tag{4.12}$$

$$\nu_t = \nu_{t-1} + 1 \tag{4.13}$$

Recall from the definition of the Gauss-inverse-Wishart pdf that the least-squares parameter estimates are obtained from the identity

$$\hat{\theta} \equiv L_{22}^{-1} L_{21} \tag{4.14}$$

and their covariance is

$$C \equiv L_{22}^{-1} D_{22}^{-1} (L_{22}^{-1})'. \tag{4.15}$$

4.3 Forgetting in Gaussian model

This section describes the application of forgetting techniques from the previous chapter on Gaussian model. In this case, we are focused mainly on tracking of varying vector of regression coefficients θ.

4.3.1 Exponential forgetting

Recall from Section (3.1.1) that this most basic type of forgetting does nothing special then a simple exponentiation of the posterior pdf.

$$f(\theta_t | d(t-1)) = \left[f(\theta_{t-1} | d(t-1)) \right]^{\lambda}. \tag{4.16}$$

Let us suppose the normality of the model (2.21). We already know, that the parameters can be modelled by a Gauss-inverse-Wishart distribution, which is conjugate and therefore it is reproduced at each instant during the estimation. The exponentiation of the posterior pdf can be seen as a multiplication of existing exponents of the pdf (4.5). The $\mathcal{G}i\mathcal{W}$ distribution has two parameters, both of them are sufficient statistics – the extended information matrix V saving the data evolution and the degrees of freedom ν. The exponential forgetting leads to the following modification of them

$$V_t = \lambda V_{t-1} + \Psi_t \Psi_t' \tag{4.17}$$

$$\nu_t = \lambda \nu_{t-1} + 1 \tag{4.18}$$

If we employ the L'DL-decomposed information matrix, then the forgetting in (4.17) related to V_{t-1} manifests itself as a multiplication of the diagonal matrix D by the forgetting factor λ

$$\lambda V_{t-1} = \lambda (L' D L)_{t-1} = L_{t-1}' \lambda D_{t-1} L_{t-1}. \tag{4.19}$$

4.3.2 Alternative forgetting

The concept of the alternative (or stabilized exponential) forgetting was introduced in Section 3.1.2. The approach is similar to the exponential forgetting, the key difference lies in the usage of an alternative distribution for model parameters, which preserves the $\mathcal{G}i\mathcal{W}$ distribution. Let us consistently with the previous chapter denote f_1 the data-updated pdf (2.30), f_2 the alternative one and λ the weight (probability) of f_1. Let V_1 and V_2 be the extended information matrices of the first or the second pdf, respectively. Also let ν_1 and ν_2 be their degrees of freedom. Then the alternative forgetting recomputes the statistics as follows

$$V_t = [\lambda V_{1;t-1} + (1-\lambda)V_{2,t-1}] + \Psi_t \Psi_t' \tag{4.20}$$

$$\nu_t = [\lambda \nu_{1;t-1} + (1-\lambda)\nu_{2;t-1}] + 1 \tag{4.21}$$

4.3.3 Linear forgetting

The linear forgetting was introduced in Section 3.1.3. Reconsider the alternative forgetting once again. While in that case the forgetting was performed as a product of two pdfs, one data-updated and the other some suitable alternative, here the multiplication is replaced by a weighted summation. As the weights are complementary to unity, the summation is a special case – a convex combination. In fact, here is the source of the main problem of this forgetting method – the result is usually neither a single density nor a density of a conjugate family, but a mixture of densities. This means, that after the time update, i.e., forgetting, the result does not belong to the same family as the prior pdf.

4.3.4 Directional forgetting

The directional forgetting (DF) introduced in Section 3.1.4 employs a modification of the Gaussian pdf described by Definition 16. This technique was derived for the Gaussian model in [43] and [44]. In this reading, it is modified for scalar output, $y_t \in \mathbb{R}$, which leads to scalar prediction error $\hat{e}_t \in \mathbb{R}$.

The forgetting itself runs in two branches, between which the switching is decided upon the variable

$$\zeta_{t-1} = \psi_{t-1}' C_{t-1} \psi_{t-1}. \tag{4.22}$$

where, using notation introduced in Definitions 16 and 18, C denotes the covariance of the estimate of θ and ψ is the regression vector. There exist two relevant cases:

1. $\zeta_t \neq 0$ – the regular case

$$\nu_t = \lambda(\nu_{t-1} + 1) \tag{4.23}$$

$$D_{LSR;t} = \lambda \left(D_{LSR;t-1} + \frac{\hat{e}_{t-1}^2}{1 + \zeta_{t-1}} \right) \tag{4.24}$$

$$\hat{\theta}_t = \hat{\theta}_{t-1} + \frac{C_{t-1}\psi_{t-1}\hat{e}_{t-1}}{1 + \zeta_{t-1}} \tag{4.25}$$

$$C_t = C_{t-1} - \frac{C_{t-1}\psi_{t-1}\psi'_{t-1}C_{t-1}}{\epsilon_{t-1}^{-1} + \zeta_{t-1}} \tag{4.26}$$

where the weighting factor

$$\epsilon_{t-1} = \lambda - \frac{1 - \lambda}{\zeta_{t-1}} \tag{4.27}$$

2. $\zeta_{t-1} = 0$ – the singular case

$$\nu_t = \lambda(\nu_{t-1} - 1) + 2 \tag{4.28}$$

$$D_{LSR;t} = \lambda(D_{LSR;t-1} + \hat{e}_{t-1}^2) \tag{4.29}$$

$$\hat{\theta}_t = \hat{\theta}_{t-1} \tag{4.30}$$

$$C_t = C_{t-1} \tag{4.31}$$

The individual terms have the meaning as introduced in Definition 18. The factor $\lambda \in \langle 0, 1\rangle$ is identical to that one in exponential forgetting.

Remark 7. *The major innovation brought by the directional forgetting lies in the modified evaluation of the parameter covariance matrix C – see expression* (4.26)*. It might be shown, that the exponential forgetting evolves it according to [44]*

$$C_t = \frac{1}{\lambda} \left[C_{t-1} - \frac{C_{t-1}\psi_{t-1}\psi'_{t-1}C_{t-1}}{1 + \zeta_{t-1}}. \right] \tag{4.32}$$

It is easy to show, that (4.32) *is a rank-one update of the covariance matrix, also known as the Sherman-Morrison formula, in this case weighted by factor $\lambda \in \langle 0, 1\rangle$.*

4.4 Prediction with Gaussian model

The basic principles of prediction of the model output were discussed in Section 2.3.3. Now, consider that the model is Gaussian, thus it is built on the principles from the last two chapters. The predictive pdf for this case is the Student pdf as described in the following proposition.

Proposition 6 (Predictive pdf)**.** *Let V be the extended information matrix, ν the degrees of freedom and $\Psi = [y, \psi']'$ the extended regression vector. The predictive pdf for Gaussian model belongs to the Student's $\mathcal{T}_\nu$ distribution and has the form*

$$f(y|\psi, V, \nu) = \frac{\mathcal{I}(V + \Psi\Psi', \nu + 1)}{\sqrt{2\pi\mathcal{I}(V, \nu)}} \tag{4.33}$$

or

$$f(y|\psi, L, D, \nu) = \frac{\Gamma\left(0.5(\nu+1)\right)\left[D_{LSR}(1+\zeta)\right]^{-\frac{1}{2}}}{\sqrt{\pi}\,\Gamma(0.5\nu)\left(1+\frac{\hat{e}^2}{D_{LSR}(1+\zeta)}\right)^{\frac{1}{2}(\nu+1)}}, \tag{4.34}$$

where $\hat{e}$ is the prediction error $y - \theta'\psi$ and

$$\zeta = \psi' C \psi. \tag{4.35}$$

The terms C, L, D are described in Definition 18.

The proof can be found, e.g., in [40].

Chapter 5

Partial forgetting method (PFM)

The content of this chapter represents the new contribution of the author to the theory of tracking of time-varying parameters. The partial forgetting is a currently developed method to estimate the slowly varying parameters of input-output models even in the case, when they change with different rates. Suppose, that there is a model with a parameter vector of length n,

$$\Theta = [\Theta_1, \Theta_2, \ldots, \Theta_n]'$$

and the goal is to track the elements. In fact, the partial forgetting is an extension of the 'standard' forgetting in the sense that it allows to discard only that information, which is related to the varying parameters.

5.1 Principle of the method

The essential idea of the partial forgetting method lies in the notion that the parameters must have some distribution, which characterizes them. Such a distribution exists during the modelling process, however, it is unknown and it changes as the parameters vary. To resolve such a situation in the Bayesian framework, two ways to obtain this distribution are possible:

i. consider a random true pdf ${}^{T}f(\Theta|d(t))$ as a 'hyperdistribution' (c.f. Def. 5), describing the particular distributions during the modelling,

ii. find some convenient approximation of the parameter distribution either at each time instant or during the whole time run by modelling the projections of ${}^{T}f(\Theta|d(t))$ according to hypotheses in the form $\mathrm{E}\left[{}^{T}f(\Theta|d(t))|\Theta, d(t), H_i\right] = f_{H_i}(\Theta|d(t))$ where $i = 0, 1, \ldots$ and each hypothesis H_i is a element of a set $\mathcal{H}^*$.

The first point would probably be more rigorous from the Bayesian viewpoint, however, the complexity of the problem would enormously increase and the deployment of a method based on it would be rather questionable. The computational burden would probably hardly (if ever) balance the effectiveness and feasibility of the method.

The latter solution was made the keystone of the partial forgetting, because it is relatively simple, at least in comparison to the first one.

Let us denote the true parameter pdf ${}^{T}f(\Theta|d(t))$ and stay aware that it is unknown. It ideally describes the actual true distribution of the model parameters at each time instant t. Our aim is to find its best approximation, based on our knowledge of the reality and expert information. For this purpose we can formulate various hypotheses (Sec. 5.2) about the parameters distribution, namely about ${}^{T}f(\Theta|d(t))$. Each of these hypotheses represents a statement about the variability of individual parameter elements, i.e., whether and which configuration of parameters changes and it introduces a pdf $f_{H_i}(\Theta|d(t))$, that should be used on condition of its validity.

The particular hypotheses are supposed to be valid with some probabilities (weights). We denote them $\lambda_i, i = 0, 1, \ldots$

In this manner, we enumerate several statements about the parameters' behaviour, regardless of the knowledge, which one is true at the particular moment. The hypothetic pdfs weighted by their probabilities then form a mixture of densities (Sec. 5.3), that approximatively describes the true density ${}^{T}f(\Theta|d(t))$. An approximation of the mixture by a single density (Sec. 5.4) and its use in modelling instead of ${}^{T}f(\Theta|d(t))$ is possible then.

The approximate pdf is constructed so that it would minimize the expectation of a distance between the mixture and itself. As the distance (or more correctly divergence) measure, we use the Kullback-Leibler divergence [47, 48, 4].

5.2 Hypotheses

As it has been mentioned above, we will follow the approximation approach to ${}^{T}f(\Theta|d(t))$. Though such a distribution is random (see Def. 5) and unknown, we can construct its projections according to several hypotheses about the parameter variability, $\mathrm{E}\left[{}^{T}f(\Theta|d(t))|\Theta, d(t), H_i\right] = f_{H_i}(\Theta|d(t))$ where $i = 0, 1, \ldots$ These projections enumerate meaningful cases of parameters variability. To this end we need:

1. Hypotheses $H_i \in \mathcal{H}^*, i = 1, 2, \ldots$ about the true parameter pdf ${}^{T}f(\Theta|d(t))$. $\mathcal{H}^*$ is a set of all hypotheses being considered, enumerated further in (5.6). These hypotheses are based on our knowledge of the current situation, the expected future development, the past data or any other useful information that expresses the further development of the system. Each of these hypotheses can be viewed as a point estimate of the pdf ${}^{T}f(\Theta|d(t))$, i.e.

$$H_i : \mathrm{E}\left[{}^{T}f(\Theta|d(t))|\Theta, d(t), H_i\right] = f_{H_i}(\Theta|d(t)). \tag{5.1}$$

   where $f_{H_i}(\Theta|d(t))$ is a pdf expressing our expectation of the present parameter distribution with respect to the potential parameters variability.

   The basic partial forgetting method generates all possible hypotheses about all combinations of the parameters in the form of various expectations of the true pdf. If the model has n regression coefficients, then the cardinality of the set of hypotheses $\mathcal{H}^*$ is 2^n. We denote them $H_0, \ldots, H_{2^n-1}$. The cardinality grows exponentially with n, but a portion

of the hypotheses is likely to vanish during the modelling process due to their probability approaching zero, which can potentially be regarded by the forgetting algorithm.

It is worth to emphasize that in the hypotheses definition, the random element is the whole true pdf ${}^{T}f(\Theta|d(t))$. All other variables like parameters and data occur in the condition and thus they are treated as known. Hence the expectation is taken over all possible forms of ${}^{T}f(\Theta|d(t))$.

2. Probabilities (weights) λ_i of realization of the particular hypotheses at the following time instant(s). As each hypothesis stands for an atomic random event and they altogether form entire observation (sample) space, it must hold

$$\sum_i \lambda_i = 1, \qquad \lambda_i \in \langle 0, 1 \rangle.$$

Recall, that the multivariate parameter pdf $f(\Theta|d(t)), \Theta \in \mathbb{R}^n$ can be decomposed by the chain rule (2.9), allowing us to access the marginal pdfs of parameters. This is the key of the method – instead of forgetting applied on the whole parameter pdf, we can apply it only on related marginals, which will be demonstrated later.

Now, let us construct a set of hypotheses $\mathcal{H}^*$ about ${}^{T}f(\Theta|d(t))$. Obviously, we can expect that no change in parameters values will occur, thus the posterior pdf (2.30) fits for the true pdf. Denote it the null hypothesis H_0

$$H_0 : \ \mathrm{E}\left[{}^{T}f(\Theta|d(t))|\Theta, d(t), H_0\right] = f(\Theta|d(t)). \tag{5.2}$$

This is one of two extreme cases. The other one says that all the parameters vary and thus have some alternative behaviour. As this case comprises all parameters at once, let us denote it, for the sake of further readability, as the last hypothesis H_{2^n-1}. We use some suitable alternative, e.g., a flat pdf $f_A(\Theta)$ from a priori gathered information

$$H_{2^n-1} : \ \mathrm{E}\left[{}^{T}f(\Theta|d(t))|\Theta, d(t), H_{2^n-1}\right] = f_A(\Theta). \tag{5.3}$$

Remark 8. *There are many possible sources of the alternative information. One of them is the prior pdf, either obtained from the first few data or from an expert. Another one is the flattened posterior pdf. In general, the choice of the prior or the alternative information is the common task in the Bayesian methods which the partial forgetting can take advantage of. Later in this theses, some sources of the alternative will be given.*

Now, we have to construct the $2^n - 2$ remaining intermediate hypotheses $H_1, \ldots, H_{2^n-2}$ about all possible configurations of potentially slowly varying parameters. As the previously introduced pdf $f_A(\Theta)$ carries the information about all the parameters, the most straightforward way is to extract just the necessary parts, related to the parameters specified as varying by the particular hypothesis. To this end, let us make use of the chain rule (2.9) and rewrite it in the form

$$f(\Theta) = f(\Theta_1, \ldots, \Theta_n) = f(\Theta_1) \prod_{i=2}^{n} f(\Theta_i|\Theta_{i-1}, \ldots, \Theta_1). \tag{5.4}$$

Indeed, we are able to decompose both the data-updated pdf $f(\Theta|d(t))$ from (5.2) and the alternative one $f_A(\Theta)$ from (5.2) and, in the data-updated pdf, replace the affected parts, i.e., its marginals of potentially varying parameters. If we denote Θ_α any subset of the Θ elements and Θ_β its complement, we can formulate the additional hypotheses

$$H_j:\ \mathrm{E}\left[{}^{T}f(\Theta|d(t))|\Theta, d(t), H_j\right] = f(\Theta_\alpha|\Theta_\beta, d(t))f_A(\Theta_\beta). \tag{5.5}$$

Following the guidelines above leads to the complete list of all possible hypotheses about ${}^{T}f(\Theta|d(t))$. Here is the enumeration:

$$\begin{aligned}
&H_0:\ \mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\Theta, d(t), H_0\right] = f(\Theta|d(t))\\
&H_1:\ \mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\Theta, d(t), H_1\right] = f(\Theta_2,\ldots,\Theta_n|\Theta_1, d(t))f_A(\Theta_1)\\
&H_2:\ \mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\Theta, d(t), H_2\right] = f(\Theta_1,\Theta_3,\ldots,\Theta_n|\Theta_2, d(t))f_A(\Theta_2)\\
&\ldots\\
&H_n:\ \mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\Theta, d(t), H_n\right] = f(\Theta_1,\ldots,\Theta_{n-1}|\Theta_n, d(t))f_A(\Theta_n)\\
&H_{n+1}:\ \mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\Theta, d(t), H_{n+1}\right] = f(\Theta_3,\ldots,\Theta_n|\Theta_1,\Theta_2, d(t))f_A(\Theta_1,\Theta_2)\\
&H_{n+2}:\ \mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\Theta, d(t), H_{n+2}\right] = f(\Theta_2,\Theta_4\ldots,\Theta_n|\Theta_1,\Theta_3, d(t))f_A(\Theta_1,\Theta_3)\\
&\ldots\\
&H_{2^n-2}:\ \mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\Theta, d(t), H_{2^n-2}\right] = f(\Theta_n|\Theta_1,\ldots,\Theta_{n-1}, d(t))f_A(\Theta_1,\ldots,\Theta_{n-1})\\
&H_{2^n-1}:\ \mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|d(t), H_{2^n-1}\right] = f_A(\Theta),
\end{aligned} \tag{5.6}$$

where f_A is an alternative pdf, expressing uncertainty arising from parameter changes.

The verbal expression of the given hypotheses is the following: H_0 assumes that no parameter varies, hence the data-updated pdf is used in (2.30) directly. The hypotheses $H_1 - H_n$ represent the cases when only one parameter varies and its marginal pdf is replaced with an alternative pdf. The remaining hypotheses enumerate cases when a specific subset of parameters vary. The last hypothesis H_{2^n-1} expresses the case when all parameters vary. Here, the whole data-updated pdf is substituted by an alternative.

The next step is to assign probabilities $\lambda_0,\ldots,\lambda_{2^n-1}$ to the hypotheses. As the parameters are supposed to vary slowly, we can usually expect that the null hypothesis H_0 will have the probability rather close to one, while the other hypotheses (the last one in particular) close to zero. Indeed, it must hold, that $\sum_i \lambda_i = 1$. The search for suitable probabilities is discussed in Section 5.7.

It is evident, that with the increasing number of model parameters, the number of hypotheses grows exponentially, which imposes demands on effectiveness of the computational environment. The preferred approach is to choose from the set of hypotheses $\mathcal{H}^*$ only those, which can become significant during the time development, i.e., select only a subset of the hypotheses set $\mathcal{H}^*$ and consider the remaining possible hypotheses to have weights equal to zero. However, to be able to do this, we need some expert information.

5.3 Mixture

Each hypothesis $H_i \in \mathcal{H}^*, i = 0, \ldots, 2^n - 1$ represents an atomic event with a probability λ_i that it is true. Hence weighted combination of all the hypothetical situations (i.e., pdfs) should describe the whole reality at once. For the sake of convenience, let us denote

$$\mathcal{C} = \{\Theta, d(t)\}, \tag{5.7}$$

and see, that the convex combination of the probability density functions, according to individual hypotheses, produces the expectation of the true parameter pdf in the form of a finite mixture (see Def. 12)

$$\begin{aligned} &\mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\mathcal{C}\right] = \mathrm{E}\left[\mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\mathcal{C}, H_i\right]\middle|\mathcal{C}\right] = \\ &= \sum_{i=0}^{2^n-1} \lambda_i \mathrm{E}\left[{}^{T}f(\Theta|d(t))\middle|\mathcal{C}, H_i\right] = \sum_{i=0}^{2^n-1} \lambda_i f_{H_i}(\Theta|d(t)), \end{aligned} \tag{5.8}$$

where $f_{H_i}(\Theta|d(t))$ are pdfs introduced by related hypotheses (5.1).

5.4 Approximation

A direct use of the mixture (5.8) is impractical, because it prevents on-line evaluation of the method and operating on mixtures is generally more complicated. To circumvent this issue, we will approximate it by a single density.

We search for an approximate pdf $\tilde{f}(\Theta|d(t))$ of the mixture (5.8) that belongs to the same family of distributions as the mixture components. According to the previous reading (Section 2.4), under general conditions, it is apt to use the Kullback-Leibler divergence as a 'measure' of dissimilarity between two distributions [4]. The smaller this divergence is, the more similar are two distributions are. Hence the approximative pdf should be selected as that one which minimizes the expected divergence between the mixture and itself

$$\begin{aligned} \tilde{f}(\Theta|d(t)) &= \underset{\tilde{f}\in\tilde{f}^*(\Theta|d(t))}{\arg\min}\ \mathrm{E}\left[\mathrm{D}\left({}^{T}f\middle\|\tilde{f}\right)\middle|\mathcal{C}\right] = \\ &= \underset{\tilde{f}\in\tilde{f}^*(\Theta|d(t))}{\arg\min}\ \mathrm{E}\left[\int_{\Theta^*} {}^{T}f(\Theta|d(t)) \ln\frac{{}^{T}f(\Theta|d(t))}{\tilde{f}(\Theta|d(t))}\,\mathrm{d}\Theta\middle|\mathcal{C}\right]. \end{aligned} \tag{5.9}$$

In Section 2.4 we introduced the Kerridge inaccuracy in Definition 15. It has a nice property, that the minimum of the Kullback-Leibler divergence and this inaccuracy measure are identical,

which allows to simplify the derivations. This allows us to express (5.9) as follows

$$\begin{aligned}
\tilde{f}(\Theta|d(t)) &= \underset{f^*(\Theta|d(t)}{\arg\min}\, \mathrm{E}\left[\mathrm{K}\left({}^{T}f, \tilde{f}\right)\Big|\mathcal{C}\right] = \\
&= \underset{\tilde{f}\in\tilde{f}^*(\Theta|d(t))}{\arg\min}\, \mathrm{E}\left[\int_{\Theta^*} {}^{T}f(\Theta|d(t)) \ln\frac{1}{\tilde{f}(\Theta|d(t))}\,\mathrm{d}\Theta\Big|\mathcal{C}\right] = \\
&= \Big|\ \text{substituting expectation for unknown } {}^{T}f(\Theta|d(t))\ \Big| = \\
&= \underset{\tilde{f}\in\tilde{f}^*(\Theta|d(t))}{\arg\min} \int_{\Theta^*} \mathrm{E}\left[{}^{T}f(\Theta|d(t))\Big|\mathcal{C}, H_i\right] \ln\frac{1}{\tilde{f}(\Theta|d(t))}\,\mathrm{d}\Theta = \\
&= \Big|\ \text{using relation (5.8)}\ \Big| = \\
&= \underset{\tilde{f}\in\tilde{f}^*(\Theta|d(t))}{\arg\min} \int_{\Theta^*} \sum_{i=0}^{2^n-1} \lambda_i f_{H_i}(\Theta|d(t)) \ln\frac{1}{\tilde{f}(\Theta|d(t))}\,\mathrm{d}\Theta. \qquad (5.10)
\end{aligned}$$

Using the relation (5.10), we have found the best approximation of the true parameter probability density function $\tilde{f}(\Theta|d(t))$. This pdf ideally approximates the probabilistic description of the real behaviour of model parameters, represented by the true but unknown pdf ${}^{T}f(\Theta|d(t))$. Hence, it is possible to use it as a time-updated pdf for parameter estimation, which (among others) implies its convenience for prediction purposes. The next section summarizes the discussed steps in the algorithmic form.

5.5 Algorithm of the partial forgetting

The algorithm of the partial forgetting method can be described as follows:

Initial mode, for $t = 0$

- Specify the appropriate hypotheses $H_i \in \mathcal{H}^*$, $i = 0, \ldots, 2^n - 1$ about expectation of the true parameter pdf ${}^{T}f(\Theta|d(t))$ – (5.6) – or its subset.
- Select the (initial) probabilities $\lambda_i \in \langle 0, 1\rangle$, $i = 0, \ldots, 2^n - 1$ of relevance of individual hypothesis $H_i \in \mathcal{H}^*$.
- Specify proper alternative pdfs for subset of hypotheses $\mathcal{H}^* \backslash H_0$ (e.g., using the prior pdf).

On-line mode, for $t > 0$

1. Collect the newest data d_t

2. Perform the data update (2.30)

3. Construct corresponding pdf for each hypothesis from $\mathcal{H}^*$ (5.6) with a proper alternative parameter behaviour, i.e.:

   (a) Compute the pdf after the data update – Prop. 5;

   (b) Change the pdf describing the related parameter(s) with its alternative (5.5);

4. Compute the minimally divergent pdf (5.10)

5. If $t \leq t_{end}$ (t_{end} is the ending time of the estimation), go to the step 1.

The pdf from the step 4 forms the optimal estimate of the true parameter probability density function.

5.6 Sources of alternative information

5.6.1 Prior information

The prior information represents the most basic approach to the problem of an alternative information. Let us suppose, that there is a pdf, which is 'learnt' in the first few initial time steps so, that

i. it is flat enough not to introduce too invalid information into the estimation process,

ii. it is supposed to carry perhaps weak but valid information about the parameters behaviour.

Let use abstract from non-informative prior and consider just the case of an informative but flat prior pdf. Then, this pdf can contain certain useful information, however, even in the case that it is not very accurate, it will not lead to deterioration of the parameter estimation process.

A special case of the prior information technique is the prior renewal on a finite time window. If we expect that the reality significantly changes during the time run, especially on long horizons, then the 'old' prior may get very inaccurate and its use in the modelling can lead to considerable decrease of the estimation quality. Therefore, it is reasonable to renew the alternative information in certain time intervals, and use it in the following time window.

5.6.2 Expert information

This approach is preferred when an expert can provide us with a useful piece of information about a parameter behaviour in the form of a probability density function. It can be, e.g., the case when the parameter can take a value from a predefined small set of possible values. However, this case is not very likely in most applications, often because it is complicated to express an information in the form of a pdf.

5.6.3 Flattened posterior pdf

This variant is inspired by the exponential forgetting of the posterior probability density function. As a source of the alternative information serves the posterior pdf from the previous time step. It is flattened by exponentiation by some factor $\lambda \in (0, 1)$. The alternative pdf for the partial forgetting has often to be very flat, making the factor rather low, which is different from the exponential forgetting.

Let us suppose, that at the time instant t, the posterior pdf is $f(\Theta|d(t))$. We construct the alternative pdf for the next step as follows:

$$f_{A,t+1}(\Theta) = [f(\Theta|d(t))]^{\lambda}, \qquad \lambda \in (0, 1).$$

It is also possible to employ the alternative forgetting (see Section 3.1.2), based on the exponential one, and construct the alternative pdf as

$$f_{A,t+1}(\Theta) = [f(\Theta|d(t))]^{\lambda}[f_E(\Theta|d(t))]^{1-\lambda}, \qquad \lambda \in (0, 1),$$

where $f_E(\Theta|d(t))$ is a pdf from an expert.

5.7 Determination of weights

The determination of the hypotheses weights λ_i is not easy. The main reasons are the nonlinearity of the problem and its potentially high dimensionality, related to the number of hypotheses. Two concepts were successfully tested – offline determination and online determination.

5.7.1 Offline determination of weights

This approach is based on a search of (sub)optimal weights λ_i from a batch of observed data. Several optimization methods are suitable for this purpose, e.g., from the group of the evolutionary algorithms like genetic algorithms [31, 16], methods based on Bayes factors [68] etc. This approach supposes the use of found weights for the future data batch. Obviously, this works well if the character of the parameters variability catched in the previous batch is valid for the future data, but it can easily fail if it is not the case. The computational burden is very high in the optimization step but the resources are saved during the successive estimation process.

5.7.2 Online determination of weights

The online determination is based on a search of (sub)optimal hypotheses weights λ_i during the run of the estimation. The advantage is obvious – the weights are tuned according to the latest reality, which allows to take their changes immediately in regard, however, at the cost of increased computational burden.

The online determination is performed each time step after enumeration of the hypotheses of the partial forgetting. It has just one off-line step realized a priori – setting the initial weights in a form of statistic of the Dirichlet distribution. The Dirichlet distribution pdf is defined as follows:

Definition 19 (Dirichlet distribution)**.** *The pdf of Dirichlet distribution has the form*

$$\mathcal{D}i(\lambda_1,\ldots,\lambda_{K-1};\alpha_1,\ldots,\alpha_K) \equiv \frac{1}{B(\alpha)}\prod_{i=1}^{K}\lambda_i^{\alpha_i-1} \tag{5.11}$$

for all $\lambda_1,\ldots,\lambda_{K-1},\lambda_K > 0$ *so that* $\lambda_K = 1-\sum_{i=1}^{K-1}\lambda_i$*. The normalizing constant is the multivariate beta function*

$$B(\alpha) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}, \qquad \alpha = (\alpha_1,\ldots,\alpha_K). \tag{5.12}$$

We model the weights with the Dirichlet pdf as follows – first, we set the prior statistics α_i, related to the hypotheses weights λ_i of the partial forgetting method. We can do so either on base of an expert knowledge, or simply set the weights as uniformly distributed. Then, we enhance the classical estimation algorithm with partial forgetting so, that it includes the tuning of weights.

Algorithm of the partial forgetting with online weights tuning

Initial mode, for $t = 0$

- Specify the appropriate hypotheses $H_i \in \mathcal{H}^*$, $i = 0,\ldots,2^n-1$ about expectation of the true parameter pdf ${}^{T}f(\Theta|d(t))$ – (5.6) – or its subset.
- Select the (initial) probabilities $\lambda_i \in \langle 0,1\rangle$, $i = 0,\ldots,2^n-1$ of relevance of individual hypothesis $H_i \in \mathcal{H}^*$.
- Specify proper alternative pdfs for subset of hypotheses $\mathcal{H}^*\backslash H_0$ (e.g., using the prior pdf).
- Set initial statistics α_i of the Dirichlet distribution, related to weights λ_i.

On-line mode, for $t > 0$

1. Collect the newest data d_t.
2. Perform the data update (2.30).
3. Construct corresponding pdf for each hypothesis from $\mathcal{H}^*$ (5.6) with a proper alternative parameter behaviour, i.e.:
   (a) Compute the pdf after the data update – Prop. 5;
   (b) Change the pdf describing the related parameter(s) with its alternative (5.5);
   (c) Perform the data update of the hypothetic pdfs with the next available data.
   (d) Calculate the likelihoods of the particular hypothetic pdfs.
   (e) Tune statistics α_i in (5.11) proportionally to the likelihoods of hypotheses.

(f) Express the weights of particular hypotheses by expectation of the Dirichlet distribution

$$\lambda_i = \alpha_i \left(\sum_{j=1}^{2^n-1} \alpha_j \right)^{-1}, \qquad i = 1, \ldots, 2^n - 1.$$

4. Compute the minimally divergent pdf (5.10).

5. If $t \leq t_{end}$ (t_{end} is the ending time of the estimation), go to the step 1.

The pdf from step 4 forms the optimal estimate of the true parameter probability density function. The update of the statistics – step 3(e) – may be done in several different ways, e.g., by increasing the statistics by one or proportionally to a value of the likelihood.

Chapter 6

Application to Gaussian model

In this chapter, the partial forgetting method will be derived for a Gaussian model, whose conjugated prior has the Gauss-inverse-Wishart distribution (see Chapter 4). The choice of this type of distribution was based on practical needs for modelling of multivariate normally distributed data in the Bayesian framework.

In this derivation, we will follow steps given in Chapter 5 and specify them for the Gaussian model parameters, namely regression coefficients. Let us remind that we will be confronted with the following needs:

i. A set of hypotheses $\mathcal{H}^*$ has to be generated according to (5.6). Therefore we need to find a way how to decompose the Gauss-inverse-Wishart pdf with the chain rule (5.4). This will allow us to extract useful information from the data-updated pdf $f(\Theta|d(t))$ and from the alternative pdf $f_A(\Theta)$ and upon these to construct a new hypothetic pdf.

ii. The pdfs obtained from the previous step will be assigned probabilities (weights) $\lambda_i \in \langle 0, 1\rangle$, $i = 0, \ldots, \mathrm{card}(\mathcal{H}^*) - 1$ and a finite mixture (5.8) will be built. This step is a counterpart of Equation (5.8) and there is nothing special to be defined.

iii. The mixture (5.8) will be approximated by a single Gauss-inverse-Wishart density according to (5.10), which is a bit complicated task. We will use the Kullback-Leibler divergence of two pdfs (Def. 14):

   - $\tilde{f}(\Theta|d(t))$ – the approximate $\mathcal{G}i\mathcal{W}$ pdf to be found
   - $\mathrm{E}\left[{}^{T}f(\Theta|d(t))|\Theta, d(t)\right]$ – the expectation of the true $\mathcal{G}i\mathcal{W}$ pdf represented by the mixture (5.8)

   The expression will be minimized by differentiating it with respect to characteristics of $\tilde{f}(\Theta|d(t))$ and laying equal to zero. Surprisingly, as it will be shown in the further reading, the majority of the work can be done analytically.

6.1 Construction of hypotheses

The Gaussian model employs the $\mathcal{G}i\mathcal{W}$ pdf as the conjugated prior pdf. Suppose, that the data-update step according to Proposition 5 has been already evaluated and now it is time to decide of the forgetting of regression coefficients θ. At first, we have to specify appropriate hypotheses about the individual model parameters' behaviour as shown in Equations (5.6). The parameter connected with the $\mathcal{G}i\mathcal{W}$ distribution has the form

$$\Theta = \{\theta', r\},$$

thus the vector of coefficients is augmented by the variance r. Therefore, the hypotheses about the true parameter pdf ${}^{T}f(\Theta|d(t))$ will have the following general form

$$H_i : \mathrm{E}\left[{}^{T}f(\theta, r|d(t))|\theta, r, d(t), H_i\right] = f_{H_i}(\Theta|d(t)), \qquad H_i \in \mathcal{H}^*$$
$$i = 0, \ldots, \mathrm{card}(\mathcal{H}^*).$$

The hypotheses' construction presumes a knowledge of the $\mathcal{G}i\mathcal{W}$ pdf decomposition according to the chain rule (2.9) resp. (5.4). The decomposed parts then allow to do exchanges of the information related to the varying parameters. The next proposition defines how the decomposition may be done.

Proposition 7 (Marginal and conditional pdfs of $\mathcal{G}i\mathcal{W}$ pdf)**.** *Given a distribution $\mathcal{G}i\mathcal{W}(V, \nu)$ of $\theta = [\theta'_\alpha, \theta'_\beta]'$. Let $L'DL$ be the factorization of the extended information matrix V of its pdf (with corresponding dimensions) as follows:*

$$L \equiv \begin{bmatrix} 1 & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix}, D \equiv \begin{bmatrix} D_{11} & 0 & 0 \\ 0 & D_{22} & 0 \\ 0 & 0 & D_{33} \end{bmatrix}$$

Then, the $\mathcal{G}i\mathcal{W}$ pdf may be decomposed to the low-dimensional marginal pdf

$$f(\theta_\alpha, r) \sim \mathcal{G}i\mathcal{W}_{\theta_\alpha, r}\left(\begin{bmatrix} 1 & 0 \\ L_{21} & L_{22} \end{bmatrix}, \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \end{bmatrix}, \nu\right)$$
$$\propto r^{-\frac{\nu+n_\alpha+2}{2}} \exp\left\{-\frac{1}{2r}\begin{bmatrix}-1 \\ \theta_\alpha\end{bmatrix}'\begin{bmatrix} 1 & 0 \\ L_{21} & L_{22}\end{bmatrix}'\begin{bmatrix} D_{11} & 0 \\ 0 & D_{22}\end{bmatrix}\begin{bmatrix} 1 & 0 \\ L_{21} & L_{22}\end{bmatrix}\begin{bmatrix}-1 \\ \theta_\alpha\end{bmatrix}\right\}$$

and the low-dimensional conditional pdf

$$f(\theta_\beta|\theta_\alpha, r) \sim \mathcal{N}_{\theta_\beta}\left(L_{33}^{-1}\left(L_{31} - L_{32}\theta_\alpha\right), r\left(L'_{33}D_{33}L_{33}\right)^{-1}\right)$$
$$\propto r^{-\frac{n_\beta}{2}} \exp\left\{-\frac{1}{2r}\begin{bmatrix}-1 \\ \theta_\alpha \\ \theta_\beta\end{bmatrix}'\left[L_{31}L_{32}L_{33}\right]' D_{33}\left[L_{31}L_{32}L_{33}\right]\begin{bmatrix}-1 \\ \theta_\alpha \\ \theta_\beta\end{bmatrix}\right\}$$

where n_α and n_β denote the lengths of related parameter vector (cf. Definition 18). The proof can be found in [40]

This proposition provides us with the ability to construct the pdfs for each of the hypotheses given in (5.6) in a straightforward way. Except for the zero and the last hypothetical pdfs, which are obtained from filtration and alternative information, respectively, they are built from relevant parts of the decomposed data-updated and alternative pdfs. In general, this action is equivalent to the replacement of proper rows in the $L'DL$-factorized information matrix.

To change the marginal pdf inherent to parameter θ_β in Proposition 7, it is necessary to perform such a permutation of the proper rows of the information matrix, which preserves the structure of vectors in $[-1, \theta']\, V\, [-1, \theta']'$. The permutation is given by the following proposition:

Proposition 8. *Let $V = L'DL$ be the decomposition of the extended information matrix inherent to the Gaussian model with the regression vector $\psi = [\psi_1', \psi_2, \psi_3, \psi_4']'$ where $\psi_1 \in \mathbb{R}^m, \psi_4 \in \mathbb{R}^n$ are column vectors and ψ_2, ψ_3 are real scalars. Let $\theta = [\theta_1', \theta_2, \theta_3, \theta_4']'$ be the vector of real regression coefficients which entries correspond to the entries in ψ. Let the L, D matrices take the following general form:*

$$L = \begin{bmatrix} L_{11} & 0 & 0 & 0 \\ L_{21} & 1 & 0 & 0 \\ L_{31} & L_{32} & 1 & 0 \\ L_{41} & L_{42} & L_{43} & L_{44} \end{bmatrix} \qquad D = \begin{bmatrix} D_{11} & 0 & 0 & 0 \\ 0 & D_{22} & 0 & 0 \\ 0 & 0 & D_{33} & 0 \\ 0 & 0 & 0 & D_{44} \end{bmatrix}$$

where $L_{11} \in \mathbb{R}^{m\times m}$ and $L_{44} \in \mathbb{R}^{n\times n}$ are the unit lower triangular matrices and $D_{11} \in \mathbb{R}^{m\times m}$ and $D_{44} \in \mathbb{R}^{n\times n}$ are positive definite diagonal matrices. Then, the swapping of the second and third entries of the vectors, under the condition of preserving the quadratic form $[-1, \theta']V[-1, \theta']'$, leads to the permutation of corresponding rows of the L and D matrices in the form:

$$\tilde{D} = \begin{bmatrix} D_{11} & 0 & 0 & 0 \\ 0 & \frac{D_{22}D_{33}}{D_{22}+L_{32}^2 D_{33}} & 0 & 0 \\ 0 & 0 & D_{22} + L_{32}^2 D_{33} & 0 \\ 0 & 0 & 0 & D_{44} \end{bmatrix}$$

$$= D_{11} \oplus \frac{D_{22}D_{33}}{D_{22} + L_{32}^2 D_{33}} \oplus D_{22} + L_{32}^2 D_{33} \oplus D_{44}$$

$$\tilde{L} = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & -L_{32} & 1 & 0 \\ 0 & \frac{D_{22}}{D_{22}+L_{32}^2 D_{33}} & \frac{L_{32}D_{33}}{D_{22}+L_{32}^2 D_{33}} & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \times L \times \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix}$$

where I are identity matrices of the same dimensions as the corresponding blocks in the $\tilde{D}$ matrix.

The proof and some additional properties are given in Appendix A.4.

6.2 Mixture of $\mathcal{G}i\mathcal{W}$ pdfs and approximation

As given in Section 5.4, the convex combination of the hypothetic pdfs with weights λ_i leads to the mixture of densities approximating the expectation of the true parameter probability density function ${}^{T}f(\Theta|d(t))$. To approximate this mixture with a single $\mathcal{G}i\mathcal{W}$ density, we are searching for the minimally divergent (in the Kullback-Leibler divergence sense) pdf as given in (5.10). The Kullback-Leibler divergence introduced by (2.39) of two $\mathcal{G}i\mathcal{W}$ distributions is given by the following proposition [40]:

Proposition 9 (KL divergence of two $\mathcal{G}i\mathcal{W}$ pdfs)**.** *Given two Gauss-inverse-Wishart distributions with probability density functions f and $\tilde{f}$. The Kullback-Leibler divergence of these two functions has the following form*

$$\begin{aligned}\mathrm{D}\left(f\middle\|\tilde{f}\right) = \ln\frac{\Gamma(0.5\tilde{\nu})}{\Gamma(0.5\nu)} - 0.5\ln|C\tilde{C}^{-1}| + 0.5\tilde{\nu}\ln\frac{D_{LSR}}{\tilde{D}_{LSR}} \\ + 0.5(\nu-\tilde{\nu})\psi_0(0.5\nu) - 0.5n - 0.5\nu + 0.5\mathrm{Tr}\left(C\tilde{C}^{-1}\right) \\ + 0.5\frac{\nu}{D_{LSR}}\left[\left(\hat{\theta}-\hat{\tilde{\theta}}\right)'\tilde{C}^{-1}\left(\hat{\theta}-\hat{\tilde{\theta}}\right)+\tilde{D}_{LSR}\right]\end{aligned}$$

where $\psi_0(\cdot)$ denotes the digamma function, i.e., the first logarithmic derivative of the gamma function $\Gamma(\cdot)$.

The proof is not trivial and is given in [40].

In our application, the term f stands for ${}^{T}f(\theta, r|d(t))$, which is represented by the mixture of $\mathcal{G}i\mathcal{W}$ components generated by the hypotheses H_i. To find the best approximation of this mixture we need to find the minimum of the Kullback-Leibler divergence (Proposition 9) by taking derivatives with respect to $\hat{\tilde{\theta}}, \tilde{C}, \tilde{D}_{LSR}$ and $\tilde{\nu}$. Useful identities are $\frac{\partial}{\partial X}a'Xb = ab'$, $\frac{\partial}{\partial X}\ln|AXB| = (X^{-1})'$ and $\frac{\partial}{\partial X}\mathrm{Tr}(AX) = A'$).

Proposition 10. *Given a convex combination (mixture) of n Gauss-inverse-Wishart pdfs, its best approximation in the sense of a minimization of the Kullback-Leibler divergence in the form of the $\mathcal{G}i\mathcal{W}$ distribution, is given by the following parameters (statistics)*

- *$\tilde{\theta}$ – the regression coefficients*

$$\hat{\tilde{\theta}} = \left(\sum_{i=0}^{2^n-1}\lambda_i\frac{\nu_i}{D_{LSR,i}}\right)^{-1}\left(\sum_{i=0}^{2^n-1}\lambda_i\frac{\nu_i}{D_{LSR,i}}\hat{\theta}_i\right) \tag{6.1}$$

- *$\tilde{D}_{LSR}$ – the least-squares reminder*

$$\tilde{D}_{LSR} = \tilde{\nu}\left(\sum_{i=0}^{2^n-1}\lambda_i\frac{\nu_i}{D_{LSR,i}}\right)^{-1} \tag{6.2}$$

- $\tilde{C}$ *– the least-square covariance matrix*

$$\tilde{C} = \sum_{i=0}^{2^n-1} \lambda_i C_i + \sum_{i=0}^{2^n-1} \lambda_i \frac{\nu_i}{D_{LSR;i}} \left[\left(\hat{\theta}_i - \hat{\tilde{\theta}} \right) \left(\hat{\theta}_i - \hat{\tilde{\theta}} \right)' \right] \tag{6.3}$$

- *and the counter (degrees of freedom)*

$$\tilde{\nu} = \frac{1 + \sqrt{1 + \frac{4}{3}(A - \ln 2)}}{2(A - \ln 2)} \tag{6.4}$$

where

$$A = \ln \left(\sum_{i=0}^{2^n-1} \lambda_i \frac{\nu_i}{D_{LSR,i}} \right) + \sum_{i=0}^{2^n-1} \lambda_i \ln D_{LSR,i} - \sum_{i=0}^{2^n-1} \lambda_i \ \psi_0(0.5\nu_i)$$

Proof. Most results are obtained directly using the derivative rules given above. The only exception is the counter, which required approximation of the digamma function $\psi_0(\tilde{\nu})$ (see Appendix A.6). The approximation was done on base of the Bernoulli numbers, however, other methods can be used as well (see, e.g., [3, 76, 15]).

Differentiation of the Kullback-Leibler divergence with f exchanged with the mixture of $\mathcal{G}i\mathcal{W}$ densities with respect to $\tilde{\nu}$ yields

$$\frac{\partial \mathrm{D}\left(f \middle|\middle| \tilde{f}\right)}{\partial \tilde{\nu}} = \psi_0(0.5\tilde{\nu}) - \ln \tilde{\nu} + \ln \left(\sum_{i=0}^{2^n-1} \lambda_i \frac{\nu_i}{D_{LSR,i}} \right) + \sum_{i=0}^{2^n-1} \lambda_i \ln D_{LSR,i} - \sum_{i=0}^{2^n-1} \lambda_i \ \psi_0(0.5\nu_i).$$

To find the minimum, we need to lay it equal to zero. For simplification purpose, let us substitute the independent part

$$A = \ln \left(\sum_{i=0}^{2^n-1} \lambda_i \frac{\nu_i}{D_{LSR,i}} \right) + \sum_{i=0}^{2^n-1} \lambda_i \ln D_{LSR,i} - \sum_{i=0}^{2^n-1} \lambda_i \ \psi_0(0.5\nu_i),$$

which leads to the expression

$$\psi_0(0.5\tilde{\nu}) - \ln \tilde{\nu} + A = 0. \tag{6.5}$$

The non-trivial digamma function $\psi_0(0.5\tilde{\nu})$ has to be approximated. We introduce the approximation based on the Bernoulli numbers (see Appendix A.6) and employ the first three terms which are

$$\psi_0(x) = \ln x - \frac{1}{2x} - \frac{1}{12x^2} + \underbrace{O\left(\frac{1}{x^4}\right)}_{\to 0},$$

hence for $x = 0.5\tilde{\nu}$ and substitution with A

$$\ln 0.5\tilde{\nu} - \frac{1}{\tilde{\nu}} - \frac{1}{3\tilde{\nu}} - \ln \tilde{\nu} + A = 0$$

$$\ln \frac{1}{2} - \frac{1}{\tilde{\nu}} - \frac{1}{3\tilde{\nu}^2} = 0. \tag{6.6}$$

To have a single single solution for $\tilde{\nu}$, the monotonicity should be reached. To check, the first derivative is

$$\frac{\partial}{\partial \tilde{\nu}} \left(\ln \frac{1}{2} - \frac{1}{\tilde{\nu}} - \frac{1}{3\tilde{\nu}^2} \right) = \frac{1}{\tilde{\nu}^2} + \frac{2}{3\tilde{\nu}^3},$$

which leads to a local minimum $\tilde{\nu} = -\frac{2}{3}$, which is in contradiction with the rule $\tilde{\nu} > 0$. Thus (6.5) is monotone on its domain. Also, the first derivative on it is positive, hence the monotone increase has been proved and there must be only one solution of (6.6). After a few simple rearrangements it is found to be

$$\tilde{\nu} = \frac{1 + \sqrt{1 + \frac{4}{3}(A - \ln 2)}}{2(A - \ln 2)},$$

Which was to be proved. □

A Gauss-inverse-Wishart probability density function (4.6) constructed with found terms (6.1), (6.2), (6.3) and (6.4) may be used as the best approximation of the parameters distribution.

Chapter 7

Experiments

This chapter deals with an experimental verification of the partial forgetting method's workability and reliability. The experiments are designed to check whether the method meets the needs laid out in the former part of this thesis, namely in Introduction and Section 3.2, i.e.

- The ability to track systems with time-varying parameters with different rates of changes.
- The ability to avoid the covariance blow-up.

Some of the experiments were inspired by related literature, e.g., the covariance blow-up suppression (Section 7.1).

The input-output models, built within the Bayesian framework, are compared in these experiments. It means, that we fully abstract from the state-space models like the recursive least-squares (RLS) or the Kalman filter (KF), which represent a different approach to modelling of time variations. Several forgetting methods in the domain of our focus exist, however, the real world mostly prefers simple solutions to the sophisticated ones. While the most basic approach to slowly varying parameters – the exponential forgetting – has become the most popular in spite of its drawbacks, other methods have not attained such success, no matter how sophisticated they are. That is the reason why the partial forgetting-based estimation is compared to it.

As the measure of prediction quality, the relative prediction error (RPE) or root mean squared error (RMSE) were used. The RMSE for prediction is expressed by the following relation

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}, \tag{7.1}$$

while the RPE is

$$\mathrm{RPE} = \frac{1}{s}\sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}} = \frac{RMSE}{s}. \tag{7.2}$$

Here y_t denotes the real system output, $\hat{y}_t$ is the predicted output and s is the sample standard deviation of data on a horizon T.

Under a knowledge of true parameters, we use the mean squared error (MSE) as the measure of estimation quality

$$\mathrm{MSE}(\hat{\theta}) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_t - \theta_t)^2, \qquad \theta, \hat{\theta} \in \mathbb{R}^n. \tag{7.3}$$

It is the second moment of the estimation error, and thus incorporates both the variance of the estimator and its bias.

7.1 Covariance blow-up prevention

7.1.1 Experiment design

This example is inspired by a research report [43]. There, it is used to demonstrate a similar property of the directional forgetting method. It introduces a system given by the regressive model

$$y_t = u_{1,t} - u_{2,t} + k_t + e_t,$$

where, for $t \leq 100$, the input signals $u_{1,t}$ and $u_{2,t}$ were generated in an open loop as two series of independent identically distributed normal random variables with zero mean value and unit variance. For the time instants $t > 100$, the inputs are fixed at the last randomly generated values $u_{1,100}$ and $u_{2,100}$, i.e.

- for $t = 1, \ldots, 100$

$$u_{1,t}, u_{2,t} \sim N(0, 1)$$

$$\mathrm{cov}(u_{1,t}, u_2) = 0$$

- for $t > 100$

$$u_{1,t} = u_{1,100}$$

$$u_{2,t} = u_{2,100}$$

and the absolute term k_t was the periodical function of the time

$$k_t = \sin\left(\frac{2\pi t}{200}\right).$$

The term e_t is normally distributed white noise with zero mean and variance 2.5×10^{-3}. The time horizon for the simulation was 1000 data samples. It is worth to notice that due to the computers improvement and decimal data representation, the simulation does not lead to the same results as in the referred paper. In addition, its initial seed is also unknown.

To simulate the conditions in embedded systems, the extended information matrix was represented within single precision.

The vector of regression coefficients θ_t and the regression vector ψ_t (2.22) read

$$\theta_t = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ \sin\left(\frac{2\pi t}{200}\right) \end{bmatrix} \qquad \psi_t = \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ 1 \end{bmatrix}.$$

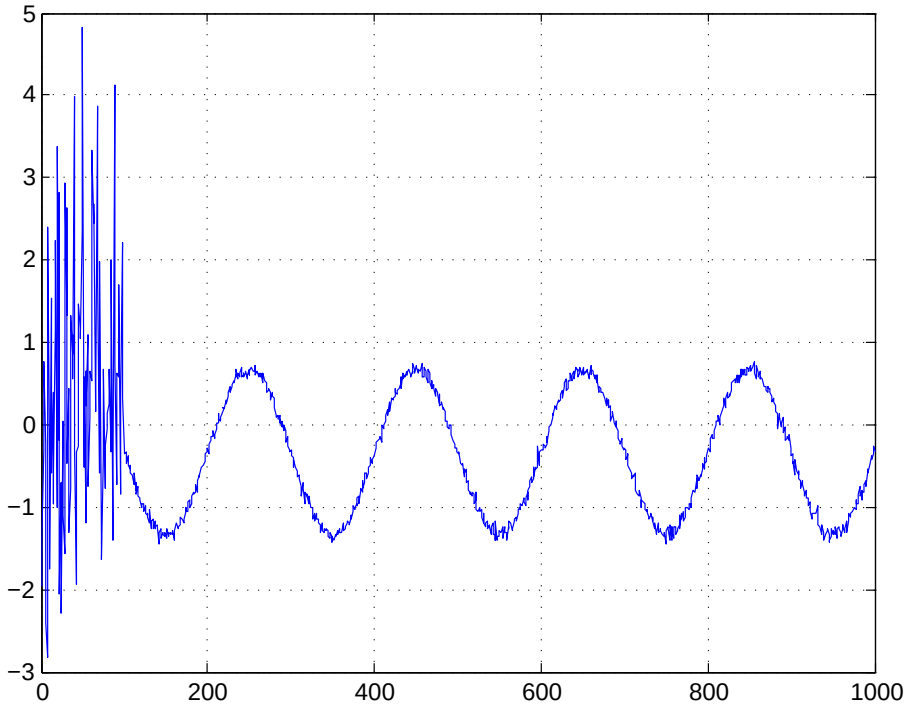


Figure 7.1: Output of the system $y_t = u_{1,t} - u_{2,t} + k_t + e_t$

The simulation started with a prior information acquired from the first 10 data vectors and this prior was used as the source for alternative pdf for the partial forgetting method, see Section 5.6.1.

7.1.2 Results

Until the signal was exciting, the both forgetting methods led to reasonable results and the covariance matrix stayed bounded. However, from the time instant $t = 100$, its non-exciting properties caused problems if forgetting was present. This is depicted in the Figure 7.2 for partial forgetting and in Fig. 7.3 for exponential forgetting. The only exciting component of the signal – the noise – was not sufficient to preserve the covariance bounded, which in the former case led to rapid exponential growth of the eigenvalues of the covariance matrix C. The latter figure shows, that the partial forgetting can eliminate such phenomenon by reintroducing the information about the parameters back into the estimation process. Generally speaking, if the model

missed any information, it used the best it had at disposal from the past. This was done by setting just a very small value $\lambda_4 = 0.0002$ to the hypothesis $H_4 : \mathrm{E}\left[{}^{T}f(\Theta|d(t))|\Theta, d(t), H_4\right] = f(\theta_3|\theta_1, \theta_2, r) f_A(\theta_1, \theta_2, r)$ and weight $1 - \lambda_4$ to the null hypothesis representing the unknown true pdf by the data-updated one. The remaining hypotheses had zero weights.

The divergent property of the exponential forgetting is further demonstrated by Figure 7.4, showing the same situation as above, but here for $\lambda = 0.98$. While $\lambda = 0.99$ led on 1000 steps to the maximum eigenvalue between 60 and 70, one per cent change of the forgetting factor increased it nearly to $9 \cdot 10^5$.

Characteristics	EF	PFM
RPE	0.7616	0.7469
RMSE	0.6559	0.6434
maximum prediction error	5.31	5.26
mean prediction error	-0.192	-0.2897
median of prediction error	-0.164	-0.248
prediction error – standard deviation	0.628	0.5752

Table 7.1: Covariance blow-up prevention – results of the 1 step-ahead prediction.

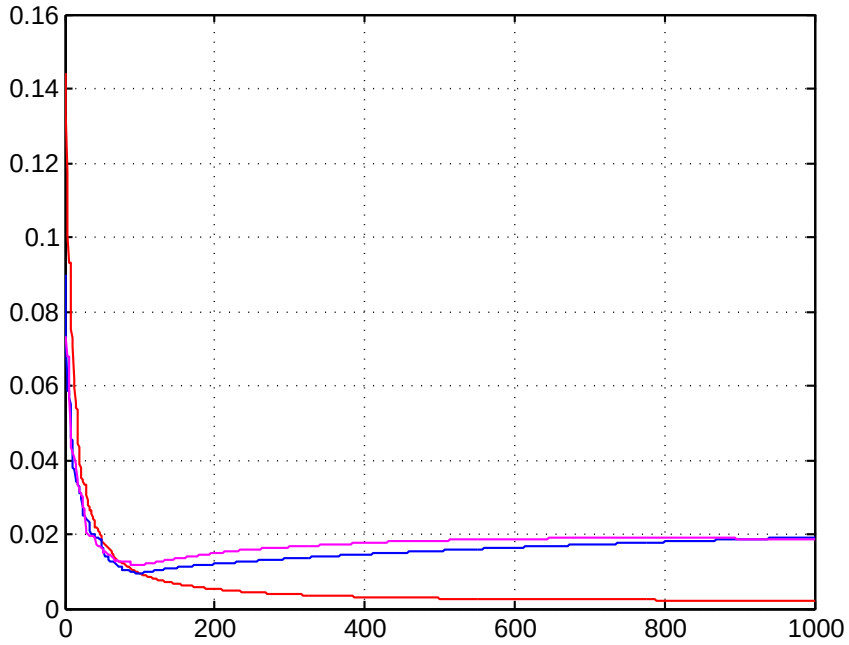


Figure 7.2: Partial forgetting – evolution of eigenvalues of the parameter covariance matrix.

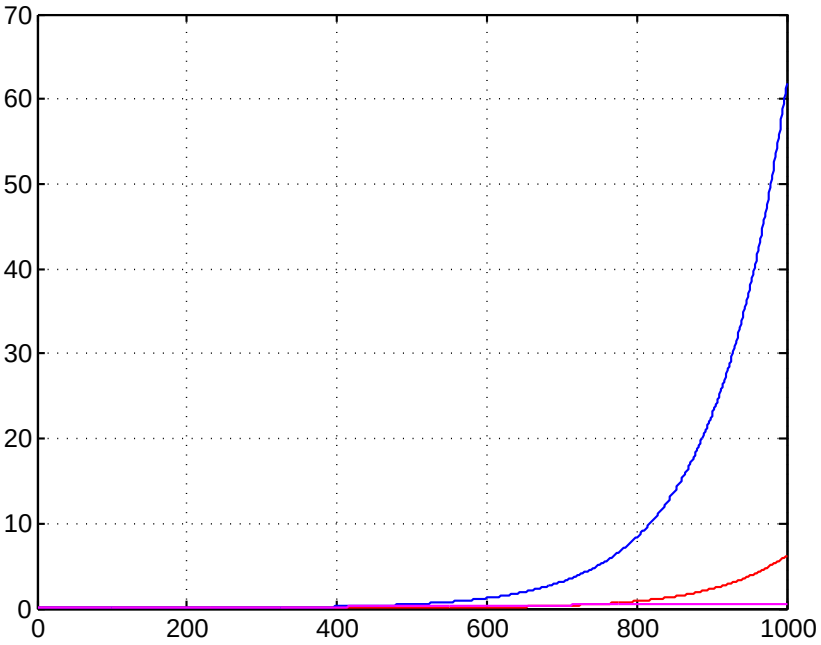


Figure 7.3: Exponential forgetting – evolution of eigenvalues of the parameter covariance matrix for $\lambda = 0.99$.

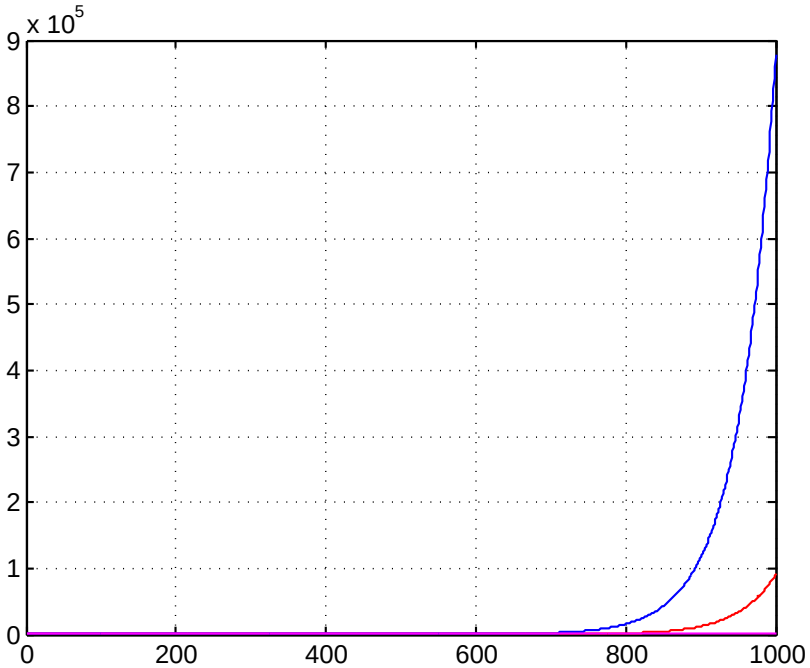


Figure 7.4: Exponential forgetting – evolution of eigenvalues of the parameter covariance matrix for $\lambda = 0.98$. A 0.01 change in the forgetting factor led to more than 10^4 times larger eigenvalue than in the previous case.

7.1.3 Discussion

Since the purpose of this example was the demonstration of the covariance blow-up suppression, the prediction ability is rather marginal in this reading. However, for the sake of completeness, the Tab. 7.1 summarizes the results of one-step-ahead prediction. The RPE (7.2) and RMSE (7.1) and the consistency are better in the case of the partial forgetting, yet its predictions are slightly biased. The prediction quality could be improved by fine-tuning of the alternative information and hypotheses weights yet, rather than by setting them just by hand within a few try-fail steps. The important conclusion is that the partial forgetting-based estimation helped to stay stable and produce reasonable results, which is not the case of the second method.

7.2 Time-varying parameters

7.2.1 Experiment design

This experiment justifies the use of the partial forgetting method for tracking time-varying parameters. In this example, we use the first-order autoregressive model with exogenous input (ARX) in the form

$$y_t = \theta_1 y_{t-1} + \theta_2 u_t + \theta_3 + e_t, \qquad t = 1, \ldots, 300,$$

where y_{t-1} denotes the previous output, u_t is the current input (a discrete white noise sequence with zero mean and unit variance) and e_t is the discrete white noise with zero mean and variance 0.5. The parameters θ_i have the following values:

- $\theta_1 = 0.5$ characterizes the model dynamics.
- $\theta_2 = 1$ is the gain of the input.
- θ_3 is a (relatively fast) oscillating absolute term:

$$\theta_3 = \begin{cases} 0.5 & \text{for } \lfloor \frac{t-1}{50} \rfloor \text{ odd} \\ -0.5 & \text{elsewhere} \end{cases}$$

The weight for the exponential forgetting (Sections 3.1.1 and 4.3.1) was determined a priori and then used during the whole runtime. The partial forgetting (Chapters 5 and 6) was run in two modes – the online mode employing tuning of the weights during the parameter identification process with prior Dirichlet statistics $[100, 10, 10, \ldots, 10]$ (Sec. 5.7.2) and the offline mode, when the final weights from the former method were used directly from the start (Sec. 5.7.1). The first mode represents the case when nothing is known at the beginning. The second mode stands for the case when we have knowledge from the past or from a (rather computer intensive) prior optimization. As a source of alternative information we used the posterior pdf, flattened by the factor 0.1.

The course of the output is depicted in the Figure 7.5. Obviously, it is significantly corrupted by additive white noise with zero mean and unit variance, hence problems with parameter estimation are expected.

7.2.2 Results

For the sake of convenience, the results of parameter tracking are discussed directly under individual figures. The quality of estimation summarizes Table 7.2.

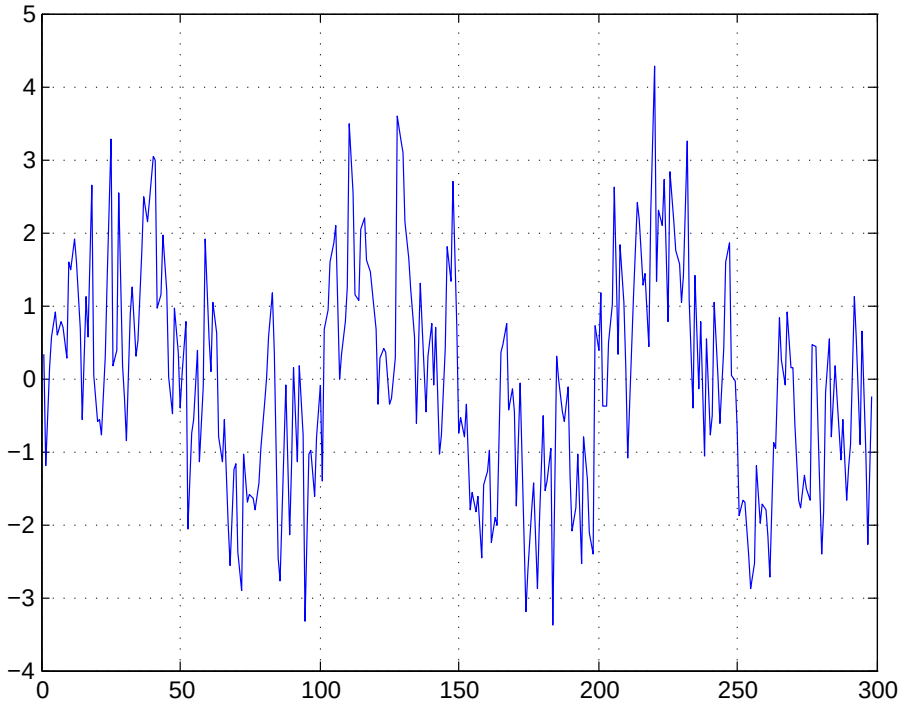


Figure 7.5: Simulated data course of ARX system $y_t = \theta_1 y_{t-1} + \theta_2 u_t + \theta_3 + e_t$

Method	**Weights**	**MSE**(θ_1)	**MSE**(θ_2)	**MSE**(θ_3)
Exponential	0.95	$2.46 \cdot 10^{-2}$	$0.51 \cdot 10^{-2}$	$15.72 \cdot 10^{-2}$
Partial (online)	[74.2, 2.13, ..., 2.13, 13.01]$\cdot 10^{-2}$	$1.46 \cdot 10^{-2}$	$1.35 \cdot 10^{-2}$	$5.61 \cdot 10^{-2}$
Partial (offline)	[74.2, 2.13, ..., 2.13, 13.01]$\cdot 10^{-2}$	$1.29 \cdot 10^{-2}$	$1.16 \cdot 10^{-2}$	$4.89 \cdot 10^{-2}$

Table 7.2: Mean squared errors of parameter estimation with exponential and partial forgetting with both online weights tuning and offline weights. The estimation with partial forgetting evidently led to better estimation of parameters θ_1 and mainly θ_3, which represents the oscillatory absolute term. The only exception is for parameter θ_2. If we knew the weights a priori to avoid the online tuning, the estimation would be slightly better yet.

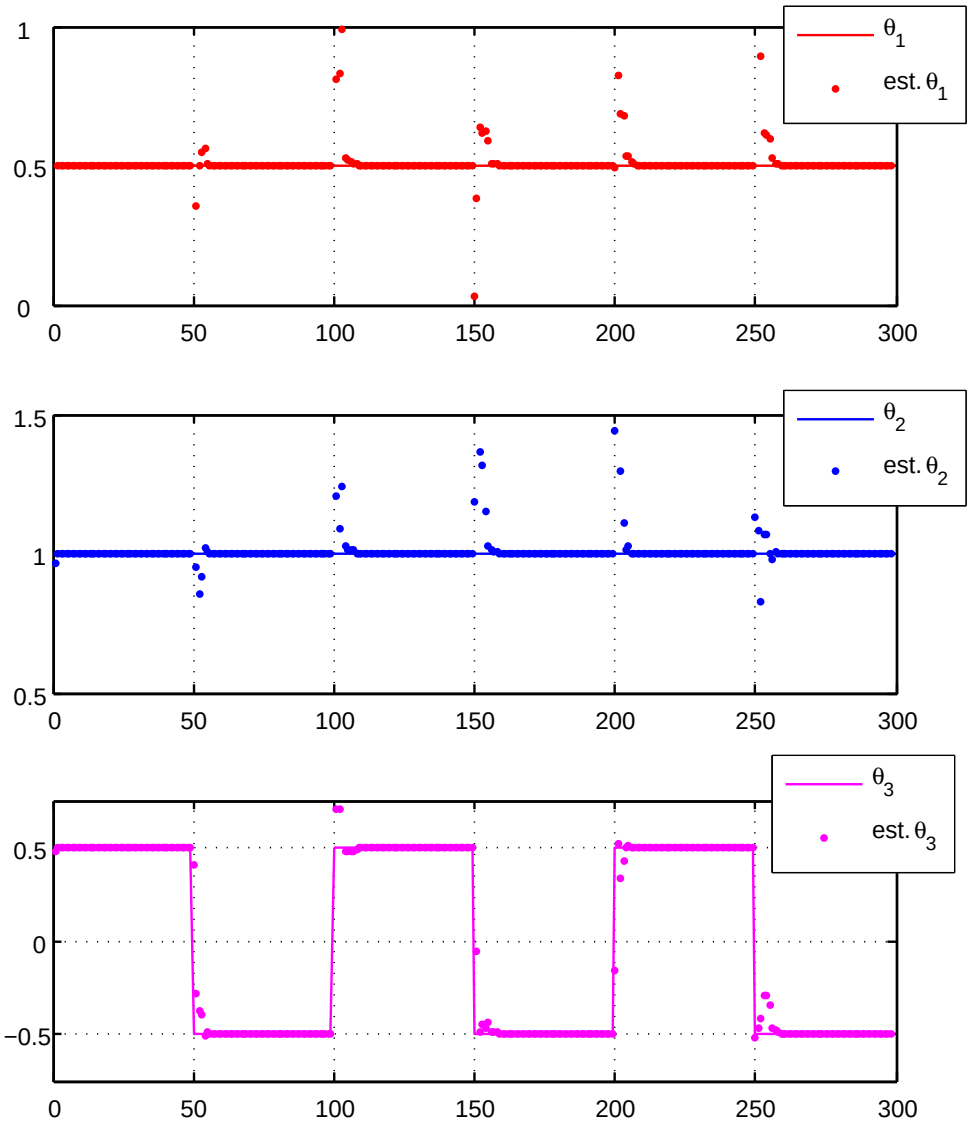


Figure 7.6: Partial forgetting with online weights tuning – parameter estimates. The estimation worked quite fine, the only difficulties were observed when the sign of the absolute term changed. It is worth to stress the estimation of θ_3, in which case the reactions to the changes are very fast and accurate.

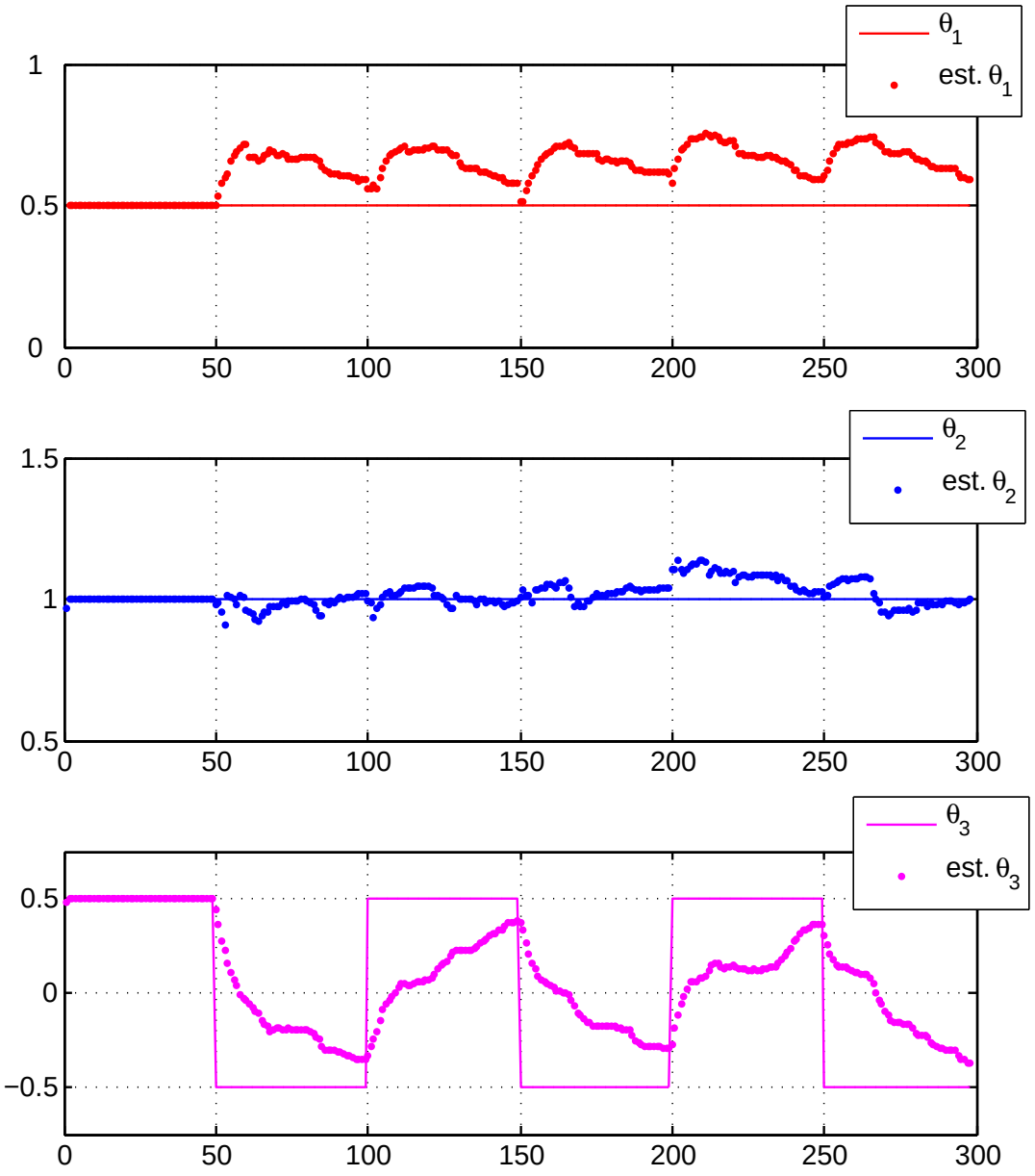


Figure 7.7: Exponential forgetting – parameter estimates. Here the situation is much worse. Only when the phase of the absolute term changes, the amplitude of the errors is smaller in comparison to the partial forgetting method. Still, the estimations of θ_1 and θ_2 are much more biased. The method completely failed with respect to the absolute term, when it was not able to react on the value changes.

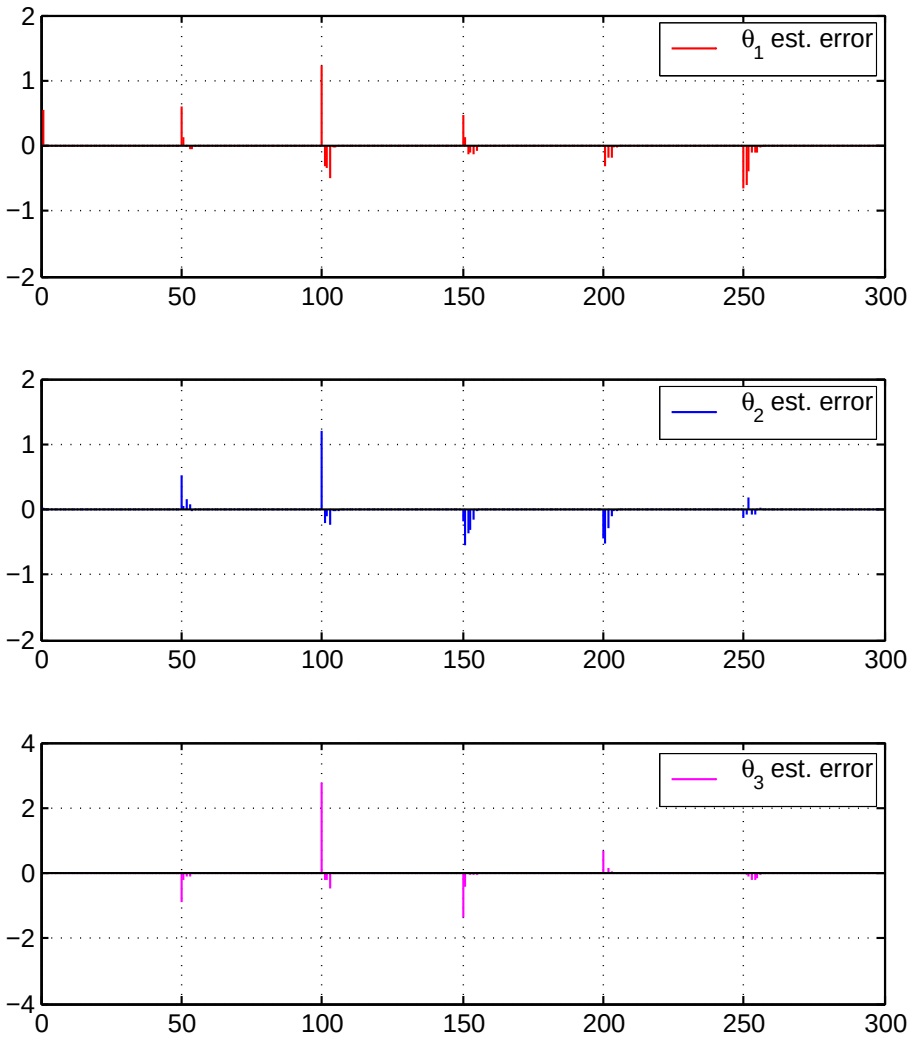


Figure 7.8: Partial forgetting with online weights tuning – estimation errors. The figure shows what was written above – the only problematic points occur when the absolute term changes its sign. This change has significant impact on all the parameters estimates, however, the estimation stabilizes quickly.

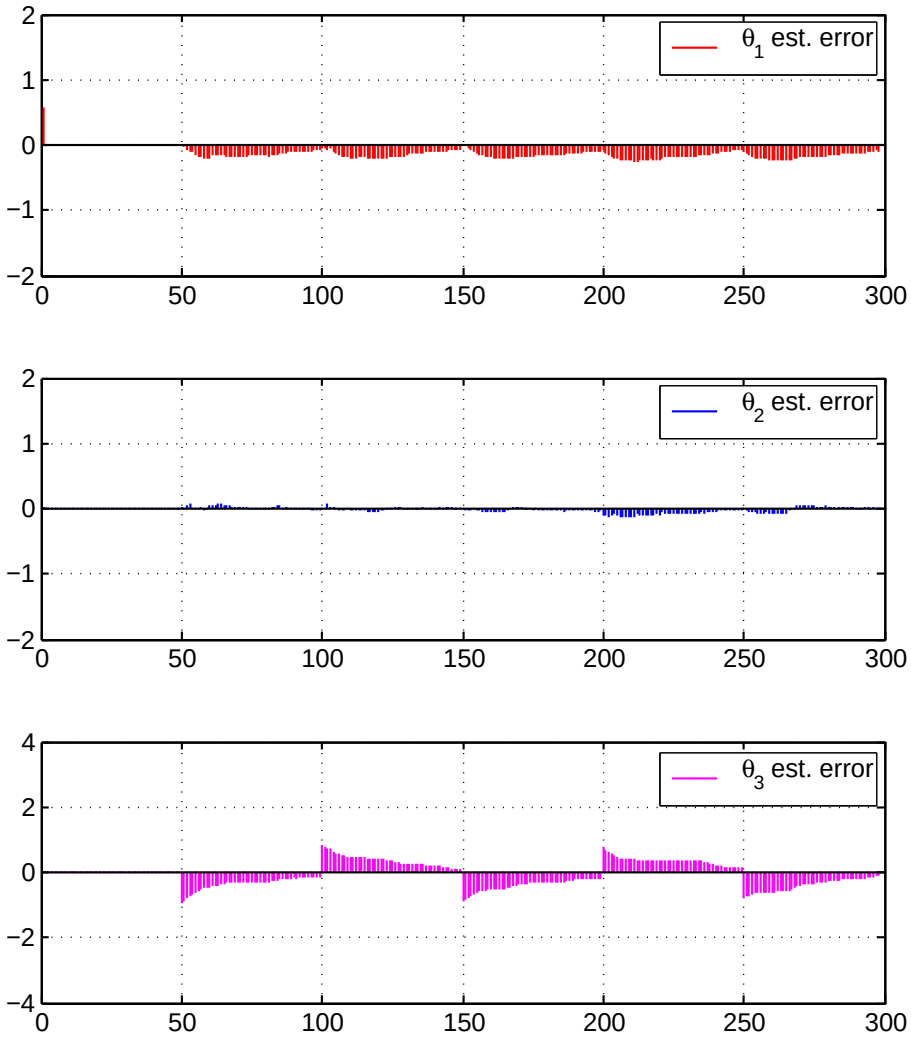


Figure 7.9: Exponential forgetting – estimation errors. The biasedness of the parameter estimation is evident. Each change of the absolute term θ_3 causes a significant decrease of the estimation quality. The process stabilizes later but with every new change the issue reappears.

7.2.3 Discussion

The example demonstrated the abilities of the partial forgetting method to track time-varying parameters of an ARX model. Two optimization methods were chosen – the online optimization, when the tracking was started with initial weights preferring the null hypothesis (i.e., the not varying parameters), and the offline optimization with the posterior weights obtained by the first method. Logically, the online optimization led to slightly worse results, but still much better than the exponential forgetting-based estimation.

As a source of an alternative pdf, the posterior pdf flattened by factor 0.1 was chosen. This choice allowed the tracking to response rapidly to the change in parameters, which, in fact, was rather large. However, to show that the method does not depend on complicated setting of the initial 'tuning knobs', this factor was chosen by the user without any optimization, just like the initial weights for online optimization were.

7.3 Real-data prediction

7.3.1 Experiment design

The autoregressive model (AR) was used for a prediction of real traffic data. The dataset consisted of traffic intensities measured in Prague, Czech Republic, with a sampling period equal to five minutes. For testing purposes, the window of 300 samples was used. This window contained one traffic peak and a short period before it. The course of the data is depicted in the Figure 7.10. We could anticipate problems with prediction of the future intensity values. At first, the course was rather flat at the beginning, then it rapidly grew and after some period it decreased. Obviously, the data had a very noisy character.

We tried to run 3 steps-ahead prediction with exponential and partial forgetting with the first 5 data vectors as the source of the prior information. The model used for this purpose was a second-order autoregressive model AR(2) in the form

$$y_t = \theta_1 y_{t-2} + \theta_2 y_{t-1} + \theta_3 + e_t, \qquad t = 1, \ldots, 300.$$

The alternative information for the partial forgetting was made from the posterior pdf by flattening it with a factor 0.9. This value was chosen by the user, according to his presumption that the parameters will vary just very slowly, to avoid extremely low weights of the hypotheses.

7.3.2 Results

Table 7.3 summarizes the results of the modelling process, i.e., the relative prediction error RPE (7.2), root mean squared error RMSE (7.1) of the prediction and statistics of the prediction errors – maximum absolute prediction error, mean and standard deviation.
As the figures look very similar, only those ones related to the partial forgetting are presented.

Apparently, the parameter estimation with partial forgetting led to a very slight improvement in the prediction quality. The advantage of the method in comparison to the exponential forgetting consisted in the ability to mix multiple information together at once. While the optimal

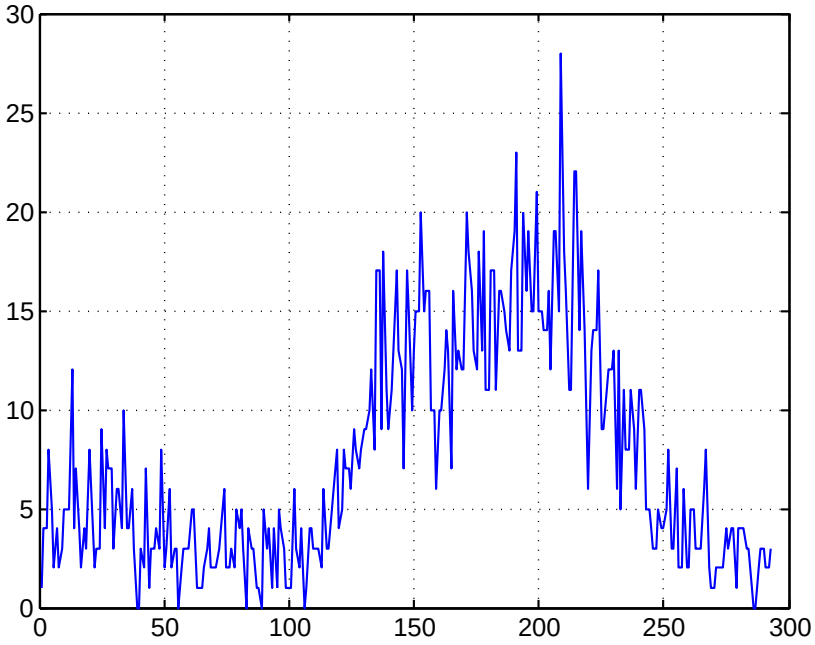


Figure 7.10: Traffic data

Characteristics	EF	PFM
RPE	0.5785	0.5640
RMSE	3.3052	3.2224
maximum absolute prediction error	12.67	15.1
mean prediction error	-0.0675	-0.0077
prediction error – standard deviation	3.3100	3.2224

Table 7.3: Traffic data – results of the 3 steps-ahead prediction

weight of the exponential forgetting was 0.95, the weights of hypotheses of the partial forgetting, determined by the offline optimization, were [17.73, 0, 0, 0, 0.35, 0.35, 81.56]$\times 10^{-2}$. The combination of the fully alternative information (the flattened posterior pdf) with the data-updated pdf was preferred.

7.3.3 Discussion

The purpose of this experiment was to demonstrate the use of the partial forgetting in the prediction of the true data. The method can improve the prediction quality even in complicated tasks like the traffic intensities prediction. Although the difference between the two methods is very small, it showed a way for the further research. First, a better alternative information could significantly improve the modelling. Second, the author was proposed with the use of log-normal

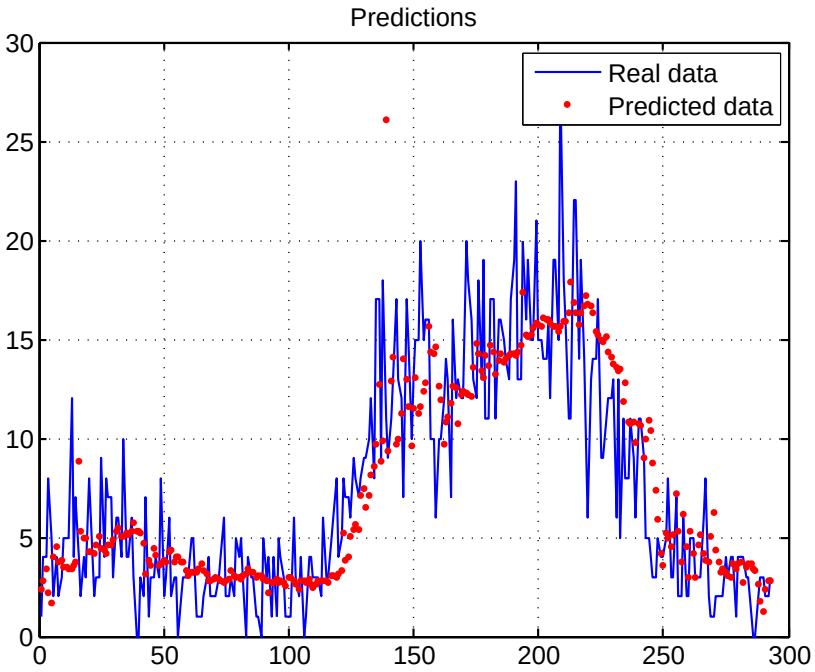


Figure 7.11: Traffic data – partial forgetting, true data and predictions

model instead of the Gaussian one, as the relation between its mean value and variance could be advantage. This will will be solved as a part of the further research .

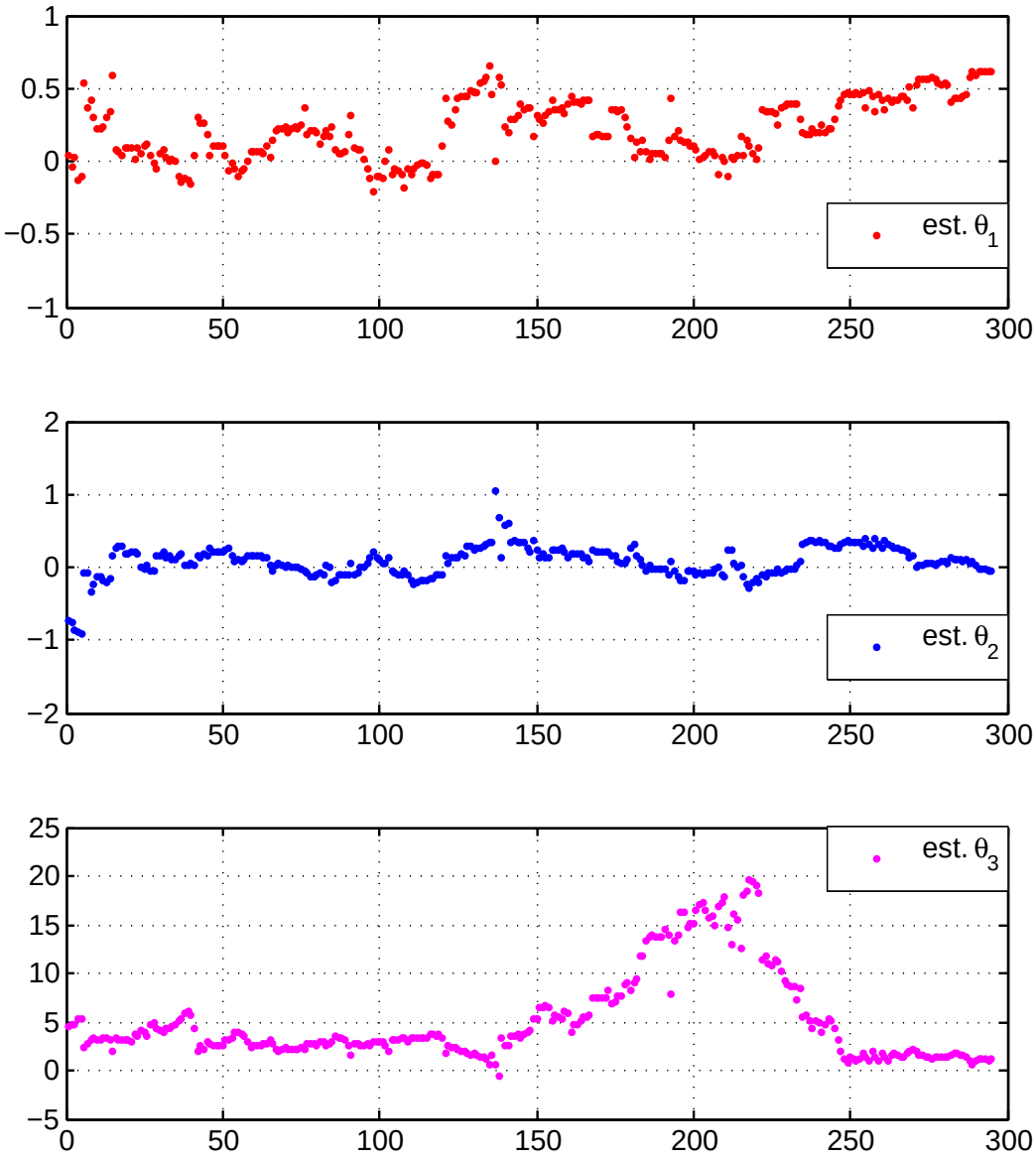


Figure 7.12: Traffic data – partial forgetting, estimated parameters

Chapter 8

Conclusions

8.1 Thesis summary

Although an intensive research effort was given to the issue of slowly varying parameters of stochastic models, it still represents a challenging problem. This thesis presents a new approach how to solve it. The method is, due to the Bayesian methodology, suitable for a wide class of parametric models.

In this work, we developed a new method suitable for tracking slowly varying parameters of stochastic models even in cases when they change with different rates. The partial forgetting method is based on the formulation of various hypotheses about the true but unknown parameter pdf, which would ideally describe the distribution of the model parameters and would be the only perfect information but which is unknown. The generality of the approach is obvious – there is no explicit need for a form of the parametric model. It can be either regressive or not; Gaussian, Poisson, log-normal, multinomial etc. In the thesis, we developed it for the Gaussian one.

The hypotheses express our expectations of the true pdf. Each of them introduces one pdf, which should be used on condition of its validity. However, we do not know which one is valid (or the nearest to the reality) at the moment – otherwise we would use it directly – but we can assign all of them with probabilities, i.e., weights, with which they are (supposed to be) valid. This step leads to a mixture of pdfs weighted by their probabilities and our goal is to approximate it by a single pdf, which we call the best available approximation of the true pdf. As a criterion for the approximation, we use the Kullback-Leibler divergence. Here again a theoretical applicability to various models arises, as the divergence can be well evaluated for many distributions.

There exist several sources of the alternative information for the method. This thesis introduces a few of them, however, the expert information comprises a wide class of other sources, which an expert can take advantage of. Anyway, if there is a lack for the expert knowledge, the issue can be easily solved, e.g., by flattening of the posterior pdf.

The hypotheses weights can be tuned either offline or online. While the offline tuning decreases the need for computational resources but at the cost of worse adaptivity, the online tuning allows to react smoothly on the changes of the modelled reality. In addition, the offline tuning

has the advantage of the independence of the optimization algorithm, hence the user can employ any existing and suitable optimization technique that he prefers.

Like everything in the world, this method has its pros and cons. The main advantage lies in its ability to track the parameters of a wide range of models and make these models stable, even in the case of different parameters variability. It helps to suppress the phenomenon known as covariance blow-up, when the eigenvalues of the covariance matrix grow without bounds in any or all directions determined by the eigenvectors. Such a situation was simulated in an experiment. The main drawback of the method consists in its dimensionality – it is a bit tricky to express a rule for the hypotheses formulation and their weights assignment. It remains as a task for the further research.

8.2 Future research directions

The partial forgetting method opened a new way to deal with time-varying parameters of stochastic models. As the method was developed in the Bayesian paradigm, it is very universal and may be used in a wide class of problems. Among the possible future research topics belong, e.g.:

- A methodology for enumeration of the significant hypotheses from the set $\mathcal{H}^*$. After resolving this issue, the computational burden could be significantly lowered. Besides the general mathematical solution, different science domains could benefit from their experts' information. This can be the case of traffic modelling, biological modelling, automation and engineering etc.

- The applications for many various models can be carried out, e.g., for the log-normal model, mentioned in Experiment 7.3 and other parametric models. The only potential difficulty lies in the solution of the Kullback-Leibler divergence for particular distributions.

Appendix A

Mathematics

This appendix summarizes the necessary mathematical background. The notation is consistent with the previous chapters. The only exception is the representation of sparse matrices where, for simplified reading, the empty positions are 0's.

A.1 Matrix algebra

The following useful rules are used in the thesis (without proofs) [64]

$$(AB)' = B'A' \tag{A.1}$$

$$(AB)^{-1} = B^{-1}A^{-1} \tag{A.2}$$

$$(A')^{-1} = (A^{-1})' \tag{A.3}$$

$$\mathrm{Tr}(A) = \sum_i \mathrm{diag}(A) = \sum_i a_{ii} \tag{A.4}$$

$$\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) = \mathrm{Tr}(CAB) \tag{A.5}$$

A.2 Matrix calculus

The following rules are useful [64]. They are listed without proofs. Let the matrices A, B, X, Y and vectors a, b be real with compatible dimensions, and the matrices, where needed, square. Then:

$$\frac{\partial A}{\partial X} = 0 \qquad \text{A is constant} \tag{A.6}$$

$$\frac{\partial \alpha X}{\partial X} = \alpha \cdot \frac{\partial X}{\partial X} \qquad \alpha \text{ is a real constant} \tag{A.7}$$

$$\frac{\partial(X+Y)}{\partial X} = \frac{\partial X}{\partial X} + \frac{\partial Y}{\partial X} \tag{A.8}$$

$$\frac{\partial \det(X)}{\partial X} = \det(X)(X^{-1})' \tag{A.9}$$

$$\frac{\partial \det(AXB)}{\partial X} = \det(AXB)(X^{-1})' \tag{A.10}$$

$$\frac{\partial \ln|\det(X)|}{\partial X} = (X^{-1})' \tag{A.11}$$

$$\frac{\partial \operatorname{Tr}(AX^{-1}B)}{\partial X} = -(X^{-1}BAX^{-1})' \tag{A.12}$$

$$\frac{\partial x'a}{\partial x} = \frac{\partial a'x}{\partial x} = a \tag{A.13}$$

$$\frac{\partial a'Xb}{\partial X} = ab' \tag{A.14}$$

A.3 Useful matrix factorizations

Before introducing the L'DL factorization, we need to define the Cholesky factorization.

Definition 20 (Cholesky factorization)**.** *Given a symmetric positive definite matrix A. The Cholesky factorization of such a matrix has the form*

$$A = LL' \tag{A.15}$$

where L is a unique lower triangular matrix with positive diagonal and L' its transposition.

The L'DL factorization is a special case of the Cholesky factorization. It is commonly used in Bayesian inference, especially in the learning process, when the extended information matrix is being updated with new data. It is the only safe way how to decompose the often poorly conditioned information matrix [40] and work with it while avoiding the computation difficulties (unlike, e.g., the LDL' and other types of factorizations). The condition for existence of this type of factorization is the symmetry and positive definiteness of the decomposed matrix.

Definition 21 (L'DL factorization)**.** *Given a symmetric definite matrix A. The L'DL factorization of such a matrix has the form*

$$A = L'DL \tag{A.16}$$

where L is a unique unit lower triangular matrix, L' is its transposition and D is a diagonal matrix. Moreover, if the matrix A is positive definite, the diagonal entries are positive as well.

The decomposed parts are of two types

- **L** – unique unit lower triangular matrix, i.e., $\mathrm{diag}(L) = 1$
- **D** – unique diagonal matrix with zeros outside the main diagonal

For a symmetric positive definite matrix of rank equal to three, which is used in this thesis, the appropriate L'DL factorization looks like follows.

$$V = L'DL = \begin{bmatrix} 1 & & \\ L_1 & 1 & \\ L_2 & L_3 & 1 \end{bmatrix}' \begin{bmatrix} D_1 & & \\ & D_2 & \\ & & D_3 \end{bmatrix} \begin{bmatrix} 1 & & \\ L_1 & 1 & \\ L_2 & L_3 & 1 \end{bmatrix} \tag{A.17}$$

In the software representation, it is more useful to work with a LD matrix as it contains both the lower diagonal matrix L and the diagonal matrix D while all the entries outside the main diagonal in the L matrix stay unchanged as well as those of the D matrix.

$$LD = \begin{bmatrix} 1 & & \\ L_1 & 1 & \\ L_2 & L_3 & 1 \end{bmatrix} + \begin{bmatrix} D_1 & & \\ & D_2 & \\ & & D_3 \end{bmatrix} - \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} = \begin{bmatrix} D_1 & & \\ L_1 & D_2 & \\ L_2 & L_3 & D_3 \end{bmatrix} \tag{A.18}$$

A.4 Permutation of adjacent rows in L'DL factorized information matrix

The following proposition describes the permutation of two adjacent rows in L'DL factorized information matrix.

Proposition 11. *Let $V = L'DL$ be the decomposition of the extended information matrix inherent to the normal model with the regression vector $\Psi = [\psi_1', \psi_2, \psi_3, \psi_4']'$ where $\psi_1 \in \mathbb{R}^m, \psi_4 \in \mathbb{R}^n$ are column vectors and ψ_2, ψ_3 are real scalars. Let $\theta = [\theta_1', \theta_2, \theta_3, \theta_4']'$ be the vector of real regression coefficients which entries correspond to the entries in ψ. Let the L, D matrices take the following general form:*

$$L = \begin{bmatrix} L_{11} & & & \\ L_{21} & 1 & & \\ L_{31} & L_{32} & 1 & \\ L_{41} & L_{42} & L_{43} & L_{44} \end{bmatrix}, \qquad D = \begin{bmatrix} D_{11} & & & \\ & D_{22} & & \\ & & D_{33} & \\ & & & D_{44} \end{bmatrix}, \tag{A.19}$$

where $L_{11} \in \mathbb{R}^{m \times m}$ and $L_{44} \in \mathbb{R}^{n \times n}$ are the unit lower triangular matrices and $D_{11} \in \mathbb{R}^{m \times m}$ and $D_{44} \in \mathbb{R}^{n \times n}$ are positive definite diagonal matrices. Then, the swapping of the second and third entries of the vectors, under the condition of preserving the quadratic form

$[-1, \theta']V[-1, \theta']'$, *leads to the permutation of corresponding rows of the* L *and* D *matrices, which leads to the following new matrices:*

$$\tilde{D} = \begin{bmatrix} D_{11} & & & \\ & \frac{D_{22}D_{33}}{D_{22}+L_{32}^2 D_{33}} & & \\ & & D_{22} + L_{32}^2 D_{33} & \\ & & & D_{44} \end{bmatrix} \tag{A.20}$$

$$= D_{11} \oplus \frac{D_{22}D_{33}}{D_{22} + L_{32}^2 D_{33}} \oplus D_{22} + L_{32}^2 D_{33} \oplus D_{44} \tag{A.21}$$

$$\tilde{L} = \begin{bmatrix} I & & & \\ & -L_{32} & 1 & \\ & \frac{D_{22}}{D_{22}+L_{32}^2 D_{33}} & \frac{L_{32}D_{33}}{D_{22}+L_{32}^2 D_{33}} & \\ & & & I \end{bmatrix} \times L \times \begin{bmatrix} I & & & \\ & 0 & 1 & \\ & 1 & 0 & \\ & & & I \end{bmatrix} \tag{A.22}$$

$$= E_{2/3} L E_1 \tag{A.23}$$

where I *are identity matrices of the same dimensions as the corresponding blocks in the* $\tilde{D}$ *matrix and* $\oplus$ *denotes direct product.*

Proof. The proposition is proved in the following way:

1. Let $\tilde{E}_{2/3}$ denote the left-hand matrix which elements were created from the once permuted matrices $\tilde{L}$ and $\tilde{D}$. If the Equation (A.23) holds, then the following relation must be true:

$$\tilde{E}_{2/3} E_{2/3} L E_1 E_1 = \tilde{E}_{2/3} \tilde{L} E_1 = L \tag{A.24}$$

2. We must obtain the originating D by elementary manipulations with $\tilde{D}$ using the entry from $\tilde{L}$ in (A.21).

Both the points lead to just elementary algebraic operations validating the proposition. □

Remark 9. *The matrices in the Equation* (A.23) *have the following special properties [55]:*

1. *The matrix* $E_{2/3}$ *is a combination of Type II and Type III elementary matrices, i.e., multiplying and adding of rows if used as a left-hand multiplier;*
2. *The matrix* E_1 *is the Type I elementary matrix, i.e., matrix interchanging columns if used as a right-hand multiplier.*

A.5 Gamma function

The gamma function was used in this thesis several times. Its definition is as follows:

Definition 22 (Gamma function)**.** *Let $z \in \mathbb{C}$ with $\Re(z) > 0$. Then, the gamma function is defined by the integral*

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} \mathrm{d}t \tag{A.25}$$

Another definitions are

$$\Gamma(z) = \lim_{n\to\infty} \frac{n^z n!}{\prod_{k=0}^{n}(z+k)} \tag{A.26}$$

$$= e^{-\gamma z} \prod_{n=1}^{\infty} \left(1 + \frac{z}{n}\right) e^{\frac{z}{n}} \tag{A.27}$$

where the definition (A.26) *is the Euler's original one and*

$$\gamma = \int_0^\infty \left(\frac{-1}{1+t} - \frac{1}{e^t}\right) \frac{1}{t} \mathrm{d}t \tag{A.28}$$

$$= \lim_{n\to\infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots + \frac{1}{n} - \ln n\right) \tag{A.29}$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n} \left[\frac{1}{i} - \ln\left(1 + \frac{1}{i}\right)\right] \tag{A.30}$$

$$= -\Gamma'(1) \tag{A.31}$$

$$\doteq 0.577215\ldots \tag{A.32}$$

is the Euler (or Euler-Mascheroni) constant in various expressions, see e.g., [20].

A.6 Digamma function

The digamma function is the first logarithmic derivative of the gamma function.

Definition 23. *Let $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$. Then, the digamma function is defined*

$$\psi_0(x) = \frac{\partial}{\partial x} \ln \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}. \tag{A.33}$$

The digamma function is the first member of the family of the polygamma functions defined as an n-th logarithmic derivative of the gamma function (hence the "0" subscript).
There are several usable approximations of the digamma function, two popular of them are

$$x \to 0^+ \qquad \psi_0(x) \approx -\frac{1}{x} - \gamma, \tag{A.34}$$

$$x \to \infty \qquad \psi_0(x) \approx \ln x - \frac{1}{2x} - \sum_{i=1}^{n} \frac{B_{2j}}{2j} \frac{1}{2x^{2j}} + O\left(\frac{1}{x^{2m}}\right). \tag{A.35}$$

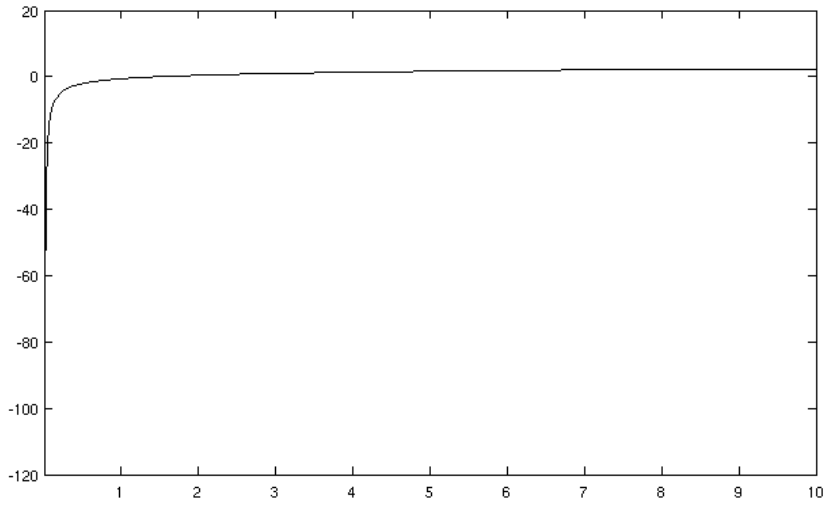


Figure A.1: Digamma function – positive half-plane

The first approximation A.34 is usable when x approaches zero from the right and does not seem to be very useful for the purpose of this thesis. The second approximation (A.35) is used when the function argument is big enough (and may even go to infinity). The terms B_{2j} stand for the Bernoulli numbers

$$B_{2j} = (-1)^{n+1} \frac{2(2n)!}{(2\pi)^{2n}} \left[1 + \frac{1}{2^{2n}} + \frac{1}{3^{2n}} + \frac{1}{4^{2n}} + \ldots \right].$$

For practical use, the first few Bernoulli numbers are tabulated.

$B_0 = 1$	$B_{12} = -\frac{691}{2730}$
$B_1 = -\frac{1}{2}$	$B_{14} = \frac{7}{6}$
$B_2 = \frac{1}{6}$	$B_{16} = -\frac{3617}{510}$
$B_4 = -\frac{1}{30}$	$B_{18} = \frac{43867}{798}$
$B_6 = \frac{1}{42}$	$B_{20} = -\frac{174611}{330}$
$B_8 = -\frac{1}{30}$	$B_{22} = \frac{854513}{138}$
$B_{10} = \frac{5}{66}$	$B_{24} = -\frac{236364091}{2730}$

Table A.1: First Bernoulli numbers

Bibliography

[1] Albert, A.E., Gardner, L.A., *Stochastic approximation and nonlinear regression*, Massechussets: IT Press Cambridge, 1967.

[2] Åström, K.J and Wittenmark, B., *Adaptive Control*, second edition, Addison-Wesley, 1995

[3] Bernardo, J.M. (1976). *Algorithm AS 103: Psi (digamma) function*, Applied Statistics, Vol. 25, No. 3 (1976), pp. 315–317.

[4] Bernardo, J.M. (1979). *Expected information as expected utility*. The Annals of Statistics, Vol. 7, No. 3, pp. 686–690.

[5] Bernardo, J.M., Smith, A.F.M., *Bayesian theory*. Wiley, Chichester, 1994.

[6] Berger, J., *Statistical Decision Theory and Bayesian Analysis (2nd edn)*. Springer-Verlag: New York, 1985.

[7] Bierman, G.J., *Measurement updating using the U-D factorization*, 1975 IEEE Conference on Decision and Control including the 14th Symposium on Adaptive Processes, vol. 14, 1975.

[8] Bittanti, S. & Campi, M. (1994). *Bounded Error Identification of Time-Varying Parameters by RLS Techniques*. IEEE Transactions on Automatic Control, Vol. 39, No. 5, pp. 1106–1110.

[9] Bretthorst, G. L., *Bayesian spectrum analysis and parameter estimation*. Springer-Verlag, 1989.

[10] Brown, L.D., *Fundamentals of statistical exponential families: with applications in statistical decision theory*. IMS, 1986. ISBN 0940600102.

[11] Bucy, R., Joseph, P., *Filtering for Stochastic Processes with Applications to Guidance*, John Wiley & Sons, New York, 1968.

[12] Candy, J.V., *Model-based signal processing*. Wiley, 2006. ISBN 9780471236320.

[13] Cao, L. & Schwartz, H. (2000). *Directional forgetting algorithm based on the decomposition of the information matrix*, Automatica, vol. 36, no. 11, pp. 1725–1731.

[14] Chaloner, K., Verdinelli, I., *Bayesian experimental design: A review*. Statistical Science, vol. 10, no. 3, pp. 273–304. 1995.

[15] Cody, W.J., Strecok, A.J. & Thacher, H.C. (1973). *Chebyshev Approximations for the Psi Function*, Mathematics of Computation, Vol. 27, No. 121 (1973), pp. 123–127.

[16] Coley, D.A., *An Introduction to Genetic Algorithms for Scientists and Engineers*. World Scientific, 1999.

[17] Congdon, P., *Bayesian statistical modelling*. Measurement Science and Technology, Vol. 13, Institute of Physics Publishing. 2002.

[18] Cover, T.M., Thomas, J.A., *Elements of information theory*. Wiley, 2006.

[19] DeGroot, M.H., *Optimal Statistical Decisions*, McGraw-Hill Book Company, New York, 1970.

[20] Dirichlet, G. L. (1836). Sur les intégrales eulériennes. J. reine angew. Math., 15, 258-263 (Werke Vol. I, pp. 273–282, G.Reimer, Berlin 1889

[21] Fagin S.L, *Recursive linear regression theory: Optimal filter theory and error analysis*, IEEE, Int. Conv. Rec., 12, pp. 216–240, 1964.

[22] Fink, D., *A Compendium of Conjugate Priors*, Cornell University, Tech. Rep., 1995

[23] Fristedt, B., Gray, L.F., *A modern approach to probability theory*. Birkhauser, 1997.

[24] Fortescue, T.R., Kershenbaum, L.S., Ydstie, B.E., *Implementation of Self-Tuning Regulators with Variable Forgetting Factors*. Automatica, Vol. 17, No. 6, pp. 831–835, 1981.

[25] Gauss, C.F., *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, 1809.

[26] Gauss, C.F., *Theoria Combinationis Observationum Erroribus Minimus Obnaxiae. Part 1, 1821; Part 2, 1823; Suppl., 1826*

[27] Gnedenko, BV and Kolmogorov, AN, *Limit distributions for sums of independent random variables*. Cambridge Mass, 1954.

[28] Goodwin, G.C., Teoh, E.K., Elliott, H., *Deterministic convergence of a self-tuning regulator with covariance resetting*, IEE Proceedings-Control Theory and Applications, vol. 130, no. 1, pp. 6–8, 1983.

[29] Hägglund, T., *Recursive Estimation of Slowly Time-Varying Parameters*, IFAC Symp. on Identification and System Parameter Estimation, York, pp. 1137–1142, 1985.

[30] Haykin, S., *Adaptive filter theory*. Prentice Hall, 2002.

[31] Houck, C.R. and Joines, J. and Kay, M., *A Genetic Algorithm for Function Optimization: A Matlab Implementation*. NCSU-IE TR, vol. 95, no. 09. 1995.

[32] Ibrahim, J.G., Chen, M.H., Sinha, D., *Bayesian survival analysis*. Springer-Verlag, 2001.

[33] James, A.T., *Distributions of Matrix Variates and Latent Roots Derived from Normal Samples*. Annals of Mathematical Statistics, Vol. 35, No. 2, pp. 475–501. 1964

[34] Jazwinski, A.H. (1970). *Stochastic processes and filtering theory*. New York: Academic Press.

[35] Jiang, J. and Zhang, Y., *A novel variable-length sliding window blockwise least-squares algorithm for on-line estimation of time-varying parameters*. International Journal of Adaptive Control and Signal Processing, Vol. 18, no. 6, pp. 505–521. UK, 2004.

[36] Kailath, T., *The innovations approach to detection and estimation theory*. Proceedings of the IEEE vol 58, no. 5, pp. 680–695. 1970.

[37] Kalman, R.E. & Bucy, R.S. (1961). *New Results in Linear Filtering and Prediction Theory*.

[38] Kalman, R.E. (1960). *A new approach to linear filtering and prediction problems*. Journal of Basic Engineering 82 (1), pp. 35–45.

[39] Kárný, M., Andrýsek, J., *Use of Kullback-Leibler divergence for forgetting*. International Journal of Adaptive Control and Signal Processing, vol. 23, no. 1, pp. 1–15. 2009.

[40] Kárný, M. et al. (2005). *Optimized Bayesian Dynamic Advising*, Springer.

[41] Kraus, F.J., *Parameter-Schätzverfahren*. Interen Bericht für Landis & Gyr Zug AG. 1983.

[42] Kulhavý, R. (1987) *Restricted exponential forgetting in real-time identification*, Automatica, vol. 23, no. 5, pp. 589–600.

[43] Kulhavý, R. *Směrové zapomínání a průběžná identifikace systémů s pomalu se měnícími parametry*. Research report no. 1170, CSAS, 1983.

[44] Kulhavý, R. & Kárný, M. (1984). *Tracking of slowly varying parameters by directional forgetting*, In Preprints of the 9th IFAC World Congress, Budapest, Vol. X, pp. 78–83.

[45] Kulhavý R. & Kraus, F.J. (1996). *On duality of regularized exponential and linear forgetting*, Automatica, vol. 32/10, pp. 1403–1415.

[46] Kulhavý R., Zarrop, M.B., *On a general concept of forgetting*, International Jurnal of Control, vol. 58, pp. 905–924, 1993.

[47] Kullback, S. *Information Theory and Statistics*. New York: Dover, 1968, 2nd ed. 1968.

[48] Kullback, S. and Leibler, R.A., *On information and sufficiency*. Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79–86, 1951.

[49] Kraus, F.J., *Das Vergessen in Rekursiven ParameterSchätzverfahren*. Zurich: ETH, 1986.

[50] Layton, K.J., Weyer, E., Campi, M. *Online Algorithms for the Construction of Guaranteed Confidence Sets for the Parameters of Time-Varying Systems*. Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009), pp. 426–431 [USB Disk]. Saint-Malo, France.

[51] Ljung, L. (1987). *System Identification: Theory for the User*. Prentince-Hall, Englewood Cliffs, N.J.

[52] Ljung, L. and Söderström, T., *Theory and practice of recursive identification*. MIT press Cambridge, Mass. 1983

[53] Lozano, R., Goodwin, G.C., *A globally convergent adaptive pole placement algorithm wihtout a persistency of excitation requirement*, IEEE transactions on automatic control, vol. 30, no. 8, pp. 795–798, 1985.

[54] Magnus, N. et al., *Neural Networks for Modelling and Control of Dynamic Systems: A Practitioner's Handbook*. New York: Springer-Verlag, 2000.

[55] Meyer, C.D., *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, 2000.

[56] Milek, J.J., *Stabilized adaptive forgetting in recursive parameter estimation*. Zurich: vdf Hochschulverlag AG, 1995.

[57] Milek, J.J., Kraus, F.J., *Stabilized least squares estimators: convergence and errorpropagation properties*, Proceedings of the 30th IEEE Conference on Decision and Control, pp. 3086–3087, 1991.

[58] Min, C. (1998). *A Gibbs Sampling Approach to Estimation and Prediction of Time-Varying-Parameter Models*. Computational Statistics and Data Analysis, Vol. 27, No. 2, pp. 171–194.

[59] Nagumo, J., Noda, A., *A learning method for system identification*, IEEE Transactions on Automatic Control, vol. 12, no. 3, pp. 282–287, 1967.

[60] Najim, K., Ikonen, E., Daoud, AK. (2004) *Stochastic processes: estimation, optimization and analysis*. Butterworth-Heinemann.

[61] Oppenheim, A.V., Schafer, R.W., *Discrete-time signal processing*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1989.

[62] Parkum,J.E., Poulsen N.K., Holst, J., *Recursive Forgetting Algorithms*, International Journal of Control, vol. 55, no. 1, pp. 109–128. 1992.

[63] Peterka, V. (1981). *Bayesian Approach to System Identification*, in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford.

[64] Petersen, K.B., Pedersen, M.S. (2008) *The Matrix Cookbook*. Available online on `http://matrixcookbook`.

[65] Plackett, r.L., *Some theorems in least squares*. Biometrika vol. 37, pp. 149–157, 1950.

[66] *Discrete stochastic process*. PlanetMath.org [cit. 23.1.2009]. Available online on `http://planetmath.org/encyclopedia/State2.html`.

[67] Pollock, DSG, *A Handbook of Time-series Analysis, Signal Processing and Dynamics*. Academic Press, 1999.

[68] Kass, R.E., Raftery, A.E., *Bayes factors*. Journal of the American Statistical Association, vol. 90, no. 430, 1995.

[69] Raiffa, H., Schlaifer, R., *Applied Statistical Decision Theory*, Boston: Harvard University, 1961.

[70] Ruanaidh, J. J. K., W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.

[71] Rudin, W., *Real and Complex Analysis*, McGraw-Hill, New York, 1987.

[72] Saelid, S., Foss, B., *Adaptive Controllers with a Vector Variable Forgetting Factor*. Decision and Control, 1983. The 22nd IEEE Conference on. Vol. 22, 1983.

[73] Shao, J., *Mathematical Statistics*. Berlin: Springer-Verlag, 1999. ISBN 0-387-98473-9.

[74] Simon, D. (2006). *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Wiley-Interscience.

[75] Söderström, T., Stoica, P., *System identification*. Prentice Hall Englewood Cliffs, NJ, 1989.

[76] Spouge, J.L. (1994). *Computation of the gamma, digamma, and trigamma functions*, SIAM Journal on Numerical Analysis, Vol. 31, No. 3 (1994), pp. 931–944.

[77] Tóth, R., Heuberger, P.S.C., Van den Hof, M.J. (2009). *An LPV Identification Framework Based on Orthonormal Basis Functions*. Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009), pp. 1328–1333 [USB Disk]. Saint-Malo, France.

[78] Young, P. (2009). *Time Variable Parameter Estimation*. Proceedings of the 15th IFAC Symposium on System Identification (SYSID 2009), pp. 432–437 [USB Disk]. Saint-Malo, France.

[79] Yingwei, L., Sundararajan, N. & Saratchandran, P. (1997). *Identification of Time-Varying Nonlinear Systems Using Minimalradial Basis Function Neural Networks*. IEE Proceedings-Control Theory and Applications, Vol. 144, No. 2, pp. 202–208.

Index