KUOPION YLIOPISTON JULKAISUJA C. LUONNONTIETEET JA YMPÄRISTÖTIETEET 213
KUOPIO UNIVERSITY PUBLICATIONS C. NATURAL AND ENVIRONMENTAL SCIENCES 213

STEFANOS GEORGIADIS

State-Space Modeling and Bayesian Methods for Evoked Potential Estimation

Doctoral dissertation

To be presented by permission of the Faculty of Natural and Environmental Sciences
of the University of Kuopio for public examination in Auditorium L21,
Snellmania building, University of Kuopio,
on Friday 25th May 2007, at 12 noon

Department of Physics
University of Kuopio

KUOPION YLIOPISTO

KUOPIO 2007

Distributor: Kuopio University Library
P.O. Box 1627
FI-70211 KUOPIO
FINLAND
Tel. +358 17 163 430
Fax +358 17 163 410
http://www.uku.fi/kirjasto/julkaisutoiminta/julkmyyn.html

Series Editors: Professor Pertti Pasanen, Ph.D.
Department of Environmental Sciences

Professor Jari Kaipio, Ph.D.
Department of Physics

Author's address: Department of Physics
University of Kuopio
P.O. Box 1627
FI-70211 KUOPIO
FINLAND
Tel. +358 17 162 362
Fax +358 17 162 373
E-mail: Stefanos.Georgiadis@uku.fi

Supervisors: Professor Pasi Karjalainen, Ph.D.
Department of Physics
University of Kuopio

Mika Tarvainen, Ph.D.
Department of Physics
University of Kuopio

Reviewers: Professor Tapio Seppänen, Dr.Tech.
Computer Engineering Laboratory
Department of Electrical and Information Engineering
University of Oulu
Oulu, Finland

Professor Tarmo Lipping, Dr.Tech.
University Consortium of Pori
Tampere University of Technology
Pori, Finland

Opponent: Docent Mark van Gils, Ph.D.
VTT Technical Research Center of Finland
Tampere, Finland

ISBN 978-951-27-0691-4
ISBN 978-951-27-0786-7 (PDF)
ISSN 1235-0486

Kopijyvä
Kuopio 2007
Finland

Georgiadis, Stefanos. State-Space Modeling and Bayesian Methods for Evoked Potential Estimation. Kuopio University Publications C. Natural and Environmental Sciences 213. 2007. 179 p.
ISBN 978-951-27-0691-4
ISBN 978-951-27-0786-7 (PDF)
ISSN 1235-0486

ABSTRACT

Electroencephalogram (EEG) provides a high-temporal resolution imaging modality for relating brain activity to cognitive function. However, individual EEG channels measure superimposed activity generated simultaneously from various brain and extra-brain sources. Therefore, signal processing methods are required in order to enhance and categorize brain related activity.

Evoked potentials (EPs) reflect changes in the brain's electrical activity due to stimulation. A significant advantage of EP research is that cortical reactivity and function can be assessed with high-temporal resolution. Therefore, EPs are used to observe changes of brain function, and to explain cognitive processes. Evoked potentials are traditionally separated from ongoing brain activity and noise by forming averages of time-locked EEG epochs. This signal enhancement leads to significant loss off information about the physiological mechanism.

Single-trial estimation methods can be used to provide information about trial-to-trial phenomena. Of special interest is the case where some parameters of the potentials change dynamically from stimulus to stimulus. This could be a trend-like change in amplitude or latency of some component of the EPs.

In this thesis, novel methods for EP denoising and enhancement are presented. The proposed methods involve state-space modeling and identification techniques. These are developed within the framework of Bayesian mean square estimation and regularization theory. Estimates of the EPs are obtained with Kalman filter and smoother algorithms. The methods are able to track dynamic variability from trial-to-trial. This is demonstrated with simulated and real EP measurements. The estimates could, for example, be used to detect changes in cognitive state such as habituation effects, or to monitor cerebral activity during anesthesia.

AMS (MOS) Classification: 92C55, 60G35, 93E11, 93E12, 93E14, 93E24, 62H25, 15A18
National Library of Medicine Classification: QT 36, WL 26.5, WL 150, WL 102, WV 270
INSPEC Thesaurus: medical signal processing; signal denoising; state-space methods; Bayes methods; Kalman filters; smoothing methods; state estimation; singular value decomposition; principal component analysis; independent component analysis; blind source separation; electroencephalography; bioelectric potentials; auditory evoked potentials

Acknowledgments

This study was carried out at the Department of Physics, University of Kuopio during the years 2001–2007. The study was supported by CIMO (Center of International Mobility, Finnish Ministry of Education), Magnus Ehrnrooth Foundation, Instrumentarium Foundation for Science, University of Kuopio and Graduate School of Functional Research in Medicine.

I would like to express my gratitude to my main supervisor Professor Pasi Karjalainen, Ph.D., for his guidance, support, and confidence in me during all these years. I would like to thank my supervisor and colleague Mika Tarvainen, Ph.D., for his support and useful suggestions for improving this work. In addition, I would like to thank my colleague Perttu Ranta-aho, M.Sc., for his support and for fruitful collaboration.

I wish to express my appreciation to the official reviewers Professor Tapio Seppänen, Dr.Tech., and Professor Tarmo Lipping, Dr.Tech., for their expert comments and constructive criticism. I would also like to thank Professor Jari Kaipio, Ph.D., for his encouragement and support.

I wish to thank the whole staff of the Department of Physics for creating a pleasant working atmosphere. Many thanks to all the members of Biosignal Analysis and Medical Imaging Group. Special thanks to Mikko Kervinen, M.Sc., and Juha-Pekka Niskanen, M.Sc., for many scientific and non-scientific discussions. I also want to thank my friend Darin Peterson for improving the language in several parts of this thesis.

Finally, I would like to thank my parents Dimitris and Jenny for their constant support to my studies, and my sister Stevy for her encouragement. I am also grateful to my other relatives in Greece and in Finland, and to all my friends for their sincere interest towards me and my work. Especially, I want to thank my dear wife Leena for her endless love and support.

Kuopio, May 2007

Stefanos Georgiadis

Abbreviations

AEP	Auditory evoked potential
BSS	Blind source separation
cdf	Cumulative distribution function
EEG	Electroencephalogram, electroencephalography
EP	Evoked potential
ERP	Event-related potential
EWA	Exponentially weighted average
GM	Gauss-Markov
IC	Independent component
ICA	Independent component analysis
KF	Kalman filter
KKT	Karush-Kuhn-Tucker
KS	Kalman smoother
LMS	Least mean square
LMS	Linear mean square
LS	Least squares
MA	Moving average
MAP	Maximum a posteriori
MEG	Magnetoencephalogram
ML	Maximum likelihood
MS	Mean square
MWA	Moving window average
NLMS	Normalized least mean square
PC	Principal component
PCA	Principal component analysis
pdf	Probability density function
RLS	Recursive least squares
RMSE	Root mean square error
SNR	Signal-to-noise ratio
STD	Standard deviation
SVD	Singular value decomposition
TR	Tikhonov regularization
UC	Uniform cost

Notations

$\mathbb{R}$	Real numbers
$\mathbb{C}$	Convex set
$\mathbb{R}^n$	n-dimensional space
$(\cdot)^T$	Transpose
$\lvert \cdot \rvert$	Absolute value
$\lVert \cdot \rVert$	Euclidean norm

$\|\cdot\|_{\cdot}$	Unspecified, general norm
$\|\cdot\|_P$	Quadratic norm, such as $\|x\|_P^2 = x^T P x$
I	Identity matrix
A^{-1}	Matrix inverse
det A	Determinant of matrix A
trace (A)	Trace of matrix A, sum of diagonal elements
diag $(a_1, a_2, \ldots, a_n)$	Diagonal matrix
$\inf\{\cdot\}$	Infimum, e.g. $\inf\{x \in \mathbb{R} : 0 < x < 1\} = 0$
$\succeq$	Matrix inequality, e.g. $A \succ 0$ denotes a symmetric positive definite matrix
$\nabla_x f$	Gradient vector
$\nabla_X f$	Gradient matrix
$\nabla_x^2 f$	Hessian matrix
d, D	Direction, for example for a vector iteration $x^{i+1} = x^i + a_i d^i$
$f^i(x; d)$	i-th directional derivative of f at x in the direction d
J_f	Jacobian of function f
J_i, J^i	Jacobian of a function at the i-th iteration
λ, ν	Lagrange multipliers
$L(x, \lambda, \nu)$	Lagrangian function
$P(\cdot)$	Probability of an event
$p(x)$	Probability density function of x
$p(x, y)$	Joint probability density of x and y
$p(x\|y)$	Conditional probability density of x given y
$p(x; y)$	Probability density of x that depends on parameters y
$E\{\cdot\}$	Mathematical expectation
$E_x\{\cdot\}$	Mathematical expectation over x
$\bar{E}\{\cdot\}$	Empirical expectation, sample mean
η_x	Expected value of x, i.e. $\eta_x = E\{x\}$
$\eta_{x\|y}$	Conditional mean of x given y
σ_x^2	Variance of x, or if x is vector then $C_x = \sigma_x^2 I$
$\gamma_x(t_1, t_2)$	Auto-covariance of the stochastic process x_t
C_x	Covariance matrix of x
C_{xy}	Cross-covariance matrix of x and y
$C_{x\|y}$	Conditional covariance of x and y
R_x	Correlation matrix of x
R_{xy}	Cross-correlation matrix of x and y
μ_k	k-th central moment
κ_k	k-th cumulant
$\Phi(\omega)$	Characteristic function, Fourier transform
$\Phi(z)$	Moment generating function
$H(x)$	Differential entropy
$H(x\|y)$	Conditional entropy
$I(x)$	Mutual information
$J(x)$	Negentropy
$\delta(p_x, p_y)$	Kullback Leibler divergence
z	Measurement vector

υ	Observation noise vector
θ, ϕ	Parameter vector
$\hat{\theta}$	Parameter estimate
$\tilde{\theta}$	Parameter estimation error
α	Regularization parameter
$B(\hat{\theta})$	Bayes cost or risk function
$C(\theta, \hat{\theta})$	Cost function
$R(\theta, \hat{\theta})$	Risk function
D_d	Difference matrix of order d
H	Observation matrix
W	Weighting matrix
L, R	Regularization matrix, $W = L^T L$
$\mathcal{E}$	Related to least squares criteria
l	Functional related to likelihoods or posterior densities
ω	State noise vector
F_t	State transition matrix
K_t	Kalman gain matrix
U	Matrix of eigenvectors or left singular vectors
Σ	Matrix of singular values

Chapter I

Introduction

A great challenge in biomedical engineering is to non-invasively gather information about the function of different parts of the human body. Various physiological systems which cannot be observed directly are subject to variation. However, a limited view of their behavior can be accessed by means of biosignals. The term biosignal can be used for any time-varying quantity that can be measured from the human body. A well known electrical biosignal is the electroencephalogram (EEG), which reflects activity in the central nervous system. EEG is broadly used for the study of different neurophysiological states and disorders. However, individual channel recordings represent superimposed signals generated by different neural or non-neural sources. Therefore, advanced signal processing methods are required in order to enhance brain related activity and to analyze complicated mental processes.

In the study of biosignals, three different approaches can be distinguished: analysis of transient events related to some physical stimulation, analysis of spontaneous effects describing the general activity and function of the physical mechanism, and correlation analysis of two or more biosignals of different nature and origin. The main benefit of an event-related analysis is that the system can be investigated under specific experimental conditions in an action-reaction scheme. The study of evoked potentials (EPs) is one such example. The aim is to analyze those parts of the EEG signal that are related to some stimulation of the central nervous system.

EPs are voltage changes of the brain's electrical activity due to stimulation. The measured signals are observed relative to an event, the timing of which can be reliably assessed. These are often considered as the combination of electric activity generated by different brain areas, which are active in association with the eliciting event. In addition, significant contributions exist from ongoing brain activity, and non-neural sources, such as eye blinks and other artifacts. In that sense, EP analysis focuses on estimating, enhancing, and categorizing stimulus evoked brain activity. In practice, EP measurements are usually performed with multiple electrodes. This means that spatial information is also included in the measurements, which can be used for the study of EPs.

Of special interest is the development of methods that enable access to in-

formation about single events, and thus, overcome traditional analysis involving simple averaging of stimulus-locked EEG data epochs. Traditional averaging aims primarily to detect a common pattern in the EP waveform which is hidden into random noise. In many situations, EPs have time-varying characteristics. Thus averaging implies a great loss of information about the hidden physiological mechanism. Therefore, EP research focuses on methods that can provide additional information about stimulus to stimulus characteristics. Trial-to-trial variability can, for example, be used to study the ability of the brain to sense, recognize, and store information. The applicability and performance of different estimation methods relate to successful mathematical modeling, identification of realistic assumptions, and more effective use of prior information.

The uncertainty of event-related phenomena allows probabilistic arguments for their description. In the simplest case, the EEG epochs can be considered as random vectors sampled from the same joint probability density function. Then, all the measurements can be used to access statistical information. The EPs can also be considered to have individual stimulus characteristics. Then information obtained from the ensemble can still be used for the enhancement of single-trials. Of special interest is the situation when some parameters of the potentials change dynamically form stimulus to stimulus. This kind of situation can be a trend-like change in amplitude or latency of some specific component of the EPs. Such a dynamic behavior can be modeled in a state-space mathematical formulation. Bayesian recursive mean square estimation methods, i.e. Kalman filtering and smoothing, can then be applied to investigate dynamic features.

EEG signals are incomplete observations of hidden mental processes. They can be used for the study of different brain dynamics. Processing this vast amount of data calls for efficient and reliable mathematical methods to extract features of interest, while suppressing different disturbances. The problem of separating and estimating source waveforms from sensor signals, without knowing the mixing system and the exact nature of the sources, can be addressed by several related methods. These include Independent Component Analysis (ICA) and Blind Source Separation (BSS), which over the last decade have been extensively used for EEG analysis [85, 39].

The aims and contents of the thesis

The aim of the thesis is to present novel methods for estimating dynamic features present in EP measurements by means of state-space modeling, signal subspace identification, and Bayesian methodology. Optimal estimates in the mean square sense are obtained with Kalman filter and smoother algorithms.

State-space modeling for EP estimation was originally proposed in [117, 116]. In these studies, a Kalman filter algorithm was used. The method was further developed in [64], and its applicability was systematically demonstrated and discussed. A Kalman smoother algorithm for trial-to-trial estimation of EPs was briefly introduced in [63].

In this thesis, the applicability of Kalman filtering and smoothing for dynamical estimation of EPs is described, and related assumptions and limitations are

further clarified. Emphasis is given to a signal subspace based method for dynamic EP estimation. The applicability of this method is demonstrated with simulated and real EP measurements under different noise conditions. The performance of Kalman filter and smoother algorithms for EP estimation is compared with different computer simulations. Different parametrizations of the problem are also considered and presented. A method for including prior information in the state-space model for dynamic estimation of EPs is also introduced. A state-space identification method for the improvement of the tracking capabilities is hereby developed and demonstrated. Since EP measurements are highly contaminated by different artifacts, BSS methods are also considered. ICA for BSS of EEG is used for artifact correction. Great effort is given to the presentation of the theoretical base of the related methods, by considering probabilistic and deterministic mathematical arguments for the problem.

The main issues covered in the thesis are summarized chapter by chapter as follows:

- Gradients and optimization methods (Chapter 2). Different estimation methods used in the thesis lead to the minimization of a specific multivariate objective function. Therefore, mathematical optimization based on vector and matrix gradients is considered. This include different iterative methods for nonlinear problems. Since prior information often constraints the solution of an estimation problem, constrained optimization is also considered.

- Probability theory (Chapter 3). ICA is related to the concept of independence. This is a theoretical concept understood by terms of probability theory. Different properties of independence are presented. Formally, statistical independence can be defined through the use of conditional probabilities. These are the foundation of Bayesian estimation. Different properties of Gaussian random vectors are also considered.

- Estimation theory (Chapter 4). The description of the theoretical properties of some estimation methods that are able to take into account in a meaningful and computationally feasible way prior knowledge about the nature of the parameters is considered. These include Bayesian and regularization methods in connection to least squares problems.

- Recursive estimation and Kalman filtering (Chapter 5). Kalman filter and smoother algorithms are presented from a Bayesian mean square estimation point of view. The smoothing problem is also treated as a constrained least squares problem, and the connection with smoothness priors regularization is emphasized. A method for including prior information in the state-space model is presented. A state-space identification method is also presented.

- Independent component analysis (Chapter 6). Different estimation principles exist for the problem of BSS, and in fact, some minimal prior knowledge is required or assumed for a specific method to be applied. Thus, the quality of source estimates and their statistical properties are naturally dependent

on the method and the assumptions used in the estimation procedure. ICA is presented from a Bayesian point of view.

- Estimation of EPs (Chapter 7). Different characteristics of EP signals are discussed, focusing on cognitive auditory potentials. The applicability of ICA for BSS of EEG is discussed and an artifact removal example is demonstrated. The artifact corrected EEG is also used in Chapter 8. Some methods for single-trial estimation of EPs are briefly discussed. The connection of Kalman filtering and smoothing with other dynamical estimation methods for EPs is discussed.

- Tracking dynamic changes (Chapter 8). The applicability of the proposed methods for tracking dynamic features in EPs is demonstrated and discussed. For demonstration an auditory experiment is considered (the auditory oddball paradigm). In that respect, different simulations are made to resemble the P300 peak, though the methods can also be applied to other EPs. The benefits of the smoothing algorithm versus the filtering algorithm are underlined. Different parametrizations of the problem are also considered, in relation to prior information about the smoothness of EPs, and other estimation needs. The developed state-space identification method for the improvement of the tracking performance is demonstrated.

- Chapter 9 contains an overall discussion and conclusions of the thesis.

Chapter II

Gradients and Optimization Methods

In this chapter, an overview of mathematical optimization, focusing on gradient based methods, is given. The concepts presented here are useful for obtaining linear and nonlinear optimization procedures for specific problems treated in later chapters of the thesis. Therefore, gradients of multivariate vector and matrix scalar valued functions and the related Taylor approximations are considered. Another useful concept discussed in this chapter is convexity that enables relatively easy solutions for different classes of optimization problems. Finally, focus is given on gradient based iterative procedures for nonlinear optimization like steepest descent and Newton's method. The chapter is primarily based on [23], though other classical references on optimization theory include [178, 51, 21, 18].

2.1 Basic concepts and definitions

A *mathematical optimization problem* has the form [23]

$$\begin{array}{ll} \text{minimize}_x & f_0(x) \\ \text{subject to} & f_i(x) \leq c_i, \quad i = 1, \ldots, k. \end{array} \tag{2.1}$$

The vector $x = (x_1, x_2, \ldots, x_n)^T$, where $(\cdot)^T$ denotes transpose, is the optimization variable of the problem. The function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the *objective function* or the *cost function*, the functions $f_i : \mathbb{R}^n \to \mathbb{R}$ define the (inequality) *constraints*, and the scalar constants c_i are the limits or bounds for the constraints. If exact equality holds for some constraint, then we have an equality constraint. If there are no constraints, the problem is called unconstrained. The set of points for which the objective function and all the constraints are defined is the *domain* $\mathbb{D} \subseteq \mathbb{R}^n$ of the optimization problem, i.e. $\mathbb{D} = \bigcap_{i=1}^{k} \{x \in \mathbb{R}^n : f_i(x) \leq c_i\} = \{x \in \mathbb{R}^n : f_i(x) \leq c_i, i = 1, \ldots, k\}$. A point is feasible if it satisfies the constraints.

The optimization problem (2.1) is an abstraction of the problem of making possible choice of a vector from a set of candidates. The constraints represent firm requirements, specifications, assumptions, or prior information for the nature of the optimal choice that limit the possible choices. They might even represent necessary compromises to be made so that the problem can have optimal or approximated optimal solution. The objective function can represent the cost of

choosing a particular solution. If the problem is a maximization problem, which can be considered as a minimization problem for the function $-f_0(x)$ subject to the constraints, then $f_0(x)$ can be considered to represent a gain or profit.

A vector x^{opt} is called optimal, or a solution for the problem (2.1), if it has the smallest objective value among all vectors that satisfy the constraints. Thus, for any x for which $f_i(x) \leq c_i, \quad i = 1, \ldots, k$, it holds $f_0(x) \geq f_0(x^{\text{opt}})$. The optimal value of the problem f^* is defined as

$$f^* = \inf\{f_0(x) : f_i(x) \leq c_i, i = 1, \ldots, k\}, \tag{2.2}$$

where f^* is allowed to take the extended values $\pm\infty$ [23]. If the problem is infeasible we have $f^* = \infty$, and if $f^* = -\infty$ the problem is unbounded below. A point x^{opt} is an optimal point if it is feasible and $f_0(x^{opt}) = f^*$. If there exist such an optimal point, then the problem is solvable. Note also, that there may exist more than one optimal points. A feasible point y is locally optimal if there is an $r > 0$ such that

$$f_0(y) = \inf\{f_0(x) : f_i(x) \leq c_i, i = 1, \ldots, k, \quad \|x - y\| \leq r\}, \tag{2.3}$$

meaning that y minimizes f_0 over nearby points in the feasible set. The i-th inequality constraint is said to be active at x, if x is feasible and $f_i(x) - c_i = 0$. If $f_i(x) - c_i < 0$ then the constraint is said to be inactive. Equality constraints are active at all feasible points. Finally, an inequality constraint of the form $f_i(x) \geq 0$ can be expressed in standard form as $-f_i(x) \leq 0$.

A *solution method* for a class of optimization problems is a procedure that computes an exact solution of the problem, i.e. an analytical formula, or more often an algorithm that gives an approximation up to some measured accuracy. The effectiveness of different procedures varies, and depends on the formulation and nature of the problem, as well as on compromises related to reduced computation time in the expense of the possibility of not reaching a satisfactory solution. In *local* optimization, the compromise is to give up seeking the optimal x, which minimizes the objective over all feasible points (*global* optimization). Instead, a locally optimal point is searched, which minimizes the objective function among all feasible points that are near to it, but it is not guaranteed to have a lower objective value over all the feasible points.

2.2 Optimality conditions for unconstrained optimization

Derivatives of the objective function and constraints are usually needed for the derivation of local optimality conditions for optimization problems, and for the derivation of gradient based solution methods.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable real scalar valued function of n variables. We have the notation $f(x) = f(x_1, x_2 \ldots, x_n) \in \mathbb{R}$, where $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$. The gradient of f with respect to its vector domain is the n-dimensional column vector of partial derivatives

$$\nabla_x f(x) = \nabla f(x) = \frac{\partial f(x)}{\partial x} = \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)^T. \tag{2.4}$$

While, the second-order gradient of a twice differentiable function with respect to its vector domain is an $n \times n$ symmetric matrix (the Hessian of f at x)

$$\nabla_x^2 f(x) = \nabla^2 f(x) = \frac{\partial^2 f(x)}{\partial x^2} = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}. \tag{2.5}$$

A differentiable function is said to be continuously differentiable at x, if there is a neighborhood of x such that the partial derivatives in (2.4) are continuous functions over the neighborhood. If the partial derivatives in (2.5) are also continuous, then f is twice continuously differentiable.

Let $f : \mathbb{X} \to \mathbb{R}$, where $\mathbb{X} \subset \mathbb{R}^n$ is open. Assuming f has continuous first-, second-, and third-order gradients over the open set $\mathbb{X}$, then for $x \in \mathbb{X}$ and any $d \in \mathbb{R}^n$ the function can be expanded on some open interval of $\tau \in \mathbb{R}$ as

$$f(x + \tau d) = f(x) + \tau f^1(x; d) + \frac{1}{2!}\tau^2 f^2(x; d) + o(\tau^3), \tag{2.6}$$

or for any $\bar{x} \in \mathbb{R}^n$ on some open interval of $\|\bar{x}\|$

$$f(\bar{x}) = f(x) + f^1(x; \bar{x} - x) + \frac{1}{2!} f^2(x; \bar{x} - x) + o(\|\bar{x}\|^3). \tag{2.7}$$

The reminder terms $o(\cdot)$ approach to zero when the approximation is close to x. The mean value theorem is what ensures the finite order of the expansion (see for example [180], Chap. 5). For some functions, one can show that the reminder $o(\|\bar{x}\|^i)$ approaches zero as i approaches infinity. Those functions can be expressed as Taylor series in a neighborhood of the point x, and are called analytic having derivatives of any order. The function $f^i(x; d) \in \mathbb{R}$ is the i-th *directional derivative* of f at x in the direction d defined for $i = 1, 2$ by

$$\begin{aligned} f^1(x; d) &= \lim_{\Delta\tau \to 0} \frac{f(x + \Delta\tau d) - f(x)}{\Delta\tau} && (2.8) \\ &= \frac{\partial f(x + \tau d)}{\partial \tau}\Big|_{\tau=0} && (2.9) \\ &= \nabla f(x)^T d, && (2.10) \end{aligned}$$

$$\begin{aligned} f^2(x; d) &= \frac{\partial^2 f(x + \tau d)}{\partial \tau^2}\Big|_{\tau=0} && (2.11) \\ &= \nabla_x f^1(x; d)^T d = \nabla_x (\nabla_x f(x)^T d)^T d = d^T \nabla^2 f(x) d, && (2.12) \end{aligned}$$

where $\tau \in \mathbb{R}$. The first directional derivative may be understood as the change in f at x when change in x is equal in magnitude and direction to d. The second directional derivative describes the local curvature of f.

A direction d defines a *descent* direction of f at x if

$$f(x + \tau d) < f(x) \quad \text{for all} \quad \tau > 0 \quad \text{and sufficiently small.} \tag{2.13}$$

Consider a first order approximation of a differentiable function f at x. It holds $f(x+\tau d) - f(x) = \tau \nabla f(x)^T d + o(\tau^2)$, with $o(\tau^2) \to 0$ as $\tau \to 0$. If

$$\nabla f(x)^T d < 0, \tag{2.14}$$

then for all $\tau > 0$ and sufficiently small, $f(x+\tau d) < f(x)$, and hence d is a descent direction of f at x. If x is a local minimum of f, then x must satisfy $\nabla f(x) = 0$. If it was $\nabla f(x) \neq 0$, then $d = -\nabla f(x)$ would be a descent direction, whereby x would not be a local minimum. Thus a necessary condition for local optimality is

$$\nabla f(x) = 0, \tag{2.15}$$

which defines the stationary points of the function f. Similarly, one can show that for a twice differentiable function at x a necessary condition for local optimality is $\nabla f(x) = 0$ and

$$d^T \nabla^2 f(x) d \geq 0, \tag{2.16}$$

for every direction d or, in other words, the Hessian at x to be positive semidefinite. Finally, it can be shown that a sufficient condition for local optimality states that if $\nabla f(x) = 0$ and the Hessian is positive definite, then x is a strict local minimum. Equivalently, if $\nabla f(x) = 0$ and the Hessian is negative definite, then x is a local maximum. When $\nabla f(x) = 0$ and the Hessian is positive semidefinite we cannot be sure if x is a local minimum. It could be a stationary point of inflection (saddle point), where the curvature of f changes sign (concavity). As an example consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^3$ at $x = 0$. For a thorough treatment of the optimality conditions see for example [19].

The concepts generalize to vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$, i.e. m-dimensional vectors of the form $f(x) = (f_1(x), \ldots, f_m(x))^T$, whose elements $f_i(x)$ are scalar functions of $x \in \mathbb{R}^n$. The Jacobian of f with respect to x is the matrix

$$J_f(x) = \frac{\partial f(x)}{\partial x} = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix}. \tag{2.17}$$

Thus is the $m \times n$ matrix with rows the gradient vectors of the scalar functions $f_i(x)$. The best linear approximation of f near the point x is given by

$$f(\bar{x}) \approx f(x) + J_f(x)(\bar{x} - x). \tag{2.18}$$

2.3 Convexity

Important concept in mathematical optimization is *convexity*, since it allows the derivation of efficient solution methods for global optimization. A *convex* optimization problem is one of the form (2.1), where all the functions f_0 and f_i are convex. *Nonlinear* optimization is the term used to describe an optimization problem when the objective and constraint functions are not linear, and in general not known to be convex. In general, there are no effective methods for solving every non linear problem.

2.3.1 Convex sets

A set $\mathbb{C} \subseteq \mathbb{R}^n$ is *affine* if the line through any two distinct points in $\mathbb{C}$ lies in $\mathbb{C}$, i.e., if for any $x_1, x_2 \in \mathbb{C}$ and $a \in \mathbb{R}$, it holds $ax_1 + (1-a)x_2 \in \mathbb{C}$. In other words, $\mathbb{C}$ contains the linear combinations of any two points in $\mathbb{C}$, provided the coefficients in the linear combination sum to one. A set $\mathbb{C}$ is *convex* if the line segment between any two points in $\mathbb{C}$ lies in $\mathbb{C}$, i.e., if for any $x_1, x_2 \in \mathbb{C}$ and any a with $0 \leq a \leq 1$, we have

$$ax_1 + (1-a)x_2 \in \mathbb{C}. \tag{2.19}$$

Clearly, every affine set is also convex. The empty set $\emptyset$, any single point x_0, and the whole space $\mathbb{R}^n$ are affine, hence convex sets of $\mathbb{R}^n$. Any line is convex and affine, but a line segment is only convex. Also convexity is preserved under intersection, i.e. if $\mathbb{C}_1, \mathbb{C}_2$ are convex then $\mathbb{C}_1 \cap \mathbb{C}_2$ is convex, and extends to the intersection of an infinite number of sets (see for example [23], section 2.3.1).

A vector of the form $a_1x_1 + \ldots + a_kx_k$, where $a_1 + \ldots + a_k = 1$, and $a_i \geq 0, \quad i = 1, \ldots k$, is a convex combination of the vectors $x_1, \ldots, x_k$. A set then is convex, if it contains every convex combination of its points. The concept of convex combination can be generalized to include infinite sums and integrals (e.g. [23], section 2.1.4). Let $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies $f(x) \geq 0$ for all $x \in \mathbb{C}$, where $\mathbb{C} \subseteq \mathbb{R}^n$ is convex, and $\int_{\mathbb{C}} f(x)dx = 1$, then, if the integral exists, it holds

$$\int_{\mathbb{C}} f(x)xdx \in \mathbb{C}. \tag{2.20}$$

An (Euclidean) ball in $\mathbb{R}^n$ has the form

$$\begin{aligned} S(x_c, r) &= \{x \in \mathbb{R}^n : \|x - x_c\| \leq r\} && (2.21)\\ &= \{x \in \mathbb{R}^n : (x - x_c)^T(x - x_c) \leq r^2\} && (2.22)\\ &= \{x_c + ru : \|u\| \leq 1\}, && (2.23) \end{aligned}$$

where, $r > 0$ and $\|\cdot\|$ denotes the Euclidean norm, i.e., $\|u\| = (u^Tu)^{1/2}$. The vector x_c is the *center* and the scalar r is its *radius*, and $S(x_c, r)$ consists off all the points within a distance r from the center x_c. $S(x_c, r)$ is convex, since if $\|x_1 - x_c\| \leq r$, $\|x_2 - x_c\| \leq r$, and $0 \leq a \leq 1$, then $\|ax_1 + (1-a)x_2 - x_c\| = \|a(x_1 - x_c) + (1-a)(x_2 - x_c)\| \leq a\|x_1 - x_c\| + (1-a)\|x_2 - x_c\| \leq r$. This is due the homogeneity property and triangle inequality for the Euclidean norm. Equivalently, the set $\{x \in \mathbb{R}^n : \|x - x_c\|. \leq r\}$, where $\|\cdot\|.$ is any norm on $\mathbb{R}^n$ is convex. Additionally, the set $\{(x, r) : \|x\|. \leq r\} \subseteq \mathbb{R}^{n+1}$ is convex set called a *norm cone*.

A related family of convex sets are the *ellipsoids*, which have the form

$$\{x \in \mathbb{R}^n : (x - x_c)^T P^{-1}(x - x_c) \leq 1\}, \tag{2.24}$$

where $P = P^T \succ 0$, i.e. the matrix P is symmetric and positive definite. The vector $x_c \in \mathbb{R}^n$ is the center of the ellipsoid and the matrix P determines how far

the ellipsoid extends in every direction from the center; the lengths of the semi-axes are given by $\sqrt{\lambda_i}$, where λ_i are the eigenvalues of P. A ball is an ellipsoid with $P = r^2 I$. Another representation of an ellipsoid is given by $\{x_c + Au : \|u\| \leq 1\}$, where A is square and non-singular. In this representation, we can assume without loss of generality that A is symmetric and positive definite. By taking $A = P^{1/2}$ this representation gives the ellipsoid (2.24). When the matrix A is symmetric positive semidefinite but singular, the set is called a degenerate ellipsoid and its affine dimension is equal to the rank of A. Degenerate ellipsoids are also convex ([23], section 2.2.2).

2.3.2 Convex functions and optimality conditions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if its domain $\mathbb{C}$ is a convex set, and if for all $x, y \in \mathbb{C}$ and for $0 \leq a \leq 1$ it holds (*Jensen's inequality*)

$$f(ax + (1-a)y) \leq af(x) + (1-a)f(y). \tag{2.25}$$

Geometrically, this inequality means that the line segment between $(x, f(x))$ and $(y, f(y))$, which is the chord from x to y, lies above the graph of f. A function is *strictly convex* if strict inequality holds whenever $x \neq y$ and $0 < a < 1$. A function f is called *concave* if $-f$ is convex. Inequality (2.25) is extended to convex combinations of more than two points. If f is convex, $x_1, \ldots, x_k \in \mathbb{C}$, and $a_1, \ldots, a_k \geq 0$ with $a_1 + \ldots + a_k = 1$, then

$$f(a_1 x_1 + \ldots + a_k x_k) \leq a_1 f(x_1) + \ldots + a_k f(x_k). \tag{2.26}$$

Additionally, the inequality extends to infinite sums and integrals (e.g. [23] section 3.1.8). For example, if $g(x) \geq 0$ and $\mathbb{S} \subseteq \mathbb{C}$, $\int_{\mathbb{S}} g(x)dx = 1$, then, if the integrals exist

$$f\left(\int_{\mathbb{S}} g(x)x dx\right) \leq \int_{\mathbb{S}} f(x)g(x)dx. \tag{2.27}$$

A function is convex if and only if it is convex when restricted to any line that intersects its domain, i.e. for all d the function $g(t) = f(x + td)$ is convex for every t for which $x + td \in \mathbb{C}$ ([23], section 3.1.1). This property is useful for investigating the convexity of some function. In fact, it can be used to derive the following important property for convex functions ([23], section 3.1.3) . If f is differentiable, then f is convex if and only if $\mathbb{C}$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{2.28}$$

holds for every $x, y \in \mathbb{C}$. The linear function of y given by $f(x) + \nabla f(x)^T (y - x)$ is the first order Taylor approximation of f near x. Thus, for a convex function the first order approximation is a global under-estimator of the function. As a consequence if $\nabla f(x) = 0$, then for all $y \in \mathbb{C}$ it holds $f(y) \geq f(x)$, i.e. x is a global minimizer of the function. With $x \neq y$ and strict inequality in (2.28) we have strict convexity. Equivalently, for a concave function it holds $f(y) \leq f(x) + \nabla f(x)^T (y - x)$.

If now we assume that f is twice differentiable, that is, the Hessian $\nabla^2 f$ exists at each point in its domain $\mathbb{C}$, which is open, then f is convex if and only if $\mathbb{C}$ is convex and its Hessian is positive semidefinite for all $x \in \mathbb{C}$ ([23], section 3.1.4)

$$\nabla^2 f(x) \succeq 0. \tag{2.29}$$

This condition can be interpreted geometrically as the requirement that the graph of the function has positive (upward) curvature at x. Similarly, f is concave if $\mathbb{C}$ is convex and $\nabla^2 f \preceq 0$ for all $x \in \mathbb{C}$. If $\nabla^2 f(x) \succ 0$ for all $x \in \mathbb{C}$ then f is strictly convex, but the converse is not true; for example the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^4$ is strictly convex, but has zero second derivative at $x = 0$.

2.4 Constrained optimization and optimality conditions

Consider the constrained, not necessary convex, optimization problem

$$\begin{array}{rll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0, & i = 1, \ldots, k \\ & g_i(x) = 0, & i = 1, \ldots, l. \end{array} \tag{2.30}$$

Because of the constraints, stationary points of $f_0(x)$ alone may not be solutions to the constrained problem, since they may not satisfy the constraints. An approach to study constrained problems relates to the theory of Lagrange multipliers.

The *Lagrangian* function $L : \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ associated with the problem is defined as ([23], p. 215)

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{i=1}^{l} \nu_i g_i(x) \tag{2.31}$$

where λ_i, ν_i are called *Lagrange multipliers* associated with the inequality and equality constraints respectively. The vectors λ, ν are also called the *dual variables* associated with the problem (2.30). The *Lagrange dual function* $g : \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ is defined as the minimum value of the Lagrangian over $x \in \mathbb{D}$, where $\mathbb{D}$ is the domain of the optimization problem, thus

$$g(\lambda, \nu) = \inf_{x \in \mathbb{D}} L(x, \lambda, \nu). \tag{2.32}$$

When the Lagrangian is unbounded below in x the dual function takes on the value $-\infty$. Let $\bar{x}$ be a (primal) feasible point, i.e. $f_i(\bar{x}) \leq 0$ and $g_i(\bar{x}) = 0$, and let $\lambda_i \geq 0$ (dual feasible). Then from (2.30, 2.31) we have $L(\bar{x}, \lambda, \nu) \leq f_0(\bar{x})$, and thus $g(\lambda, \nu) \leq L(\bar{x}, \lambda, \nu) \leq f_0(\bar{x})$. Since the inequality holds for every feasible point, then it holds also for the optimal value of the primal problem, i.e. $g(\lambda, \nu) \leq f^*$. The dual function gives a nontrivial lower bound on f^* only when $\lambda_i \geq 0$ and ν_i such as $g(\lambda, \nu) > -\infty$ ([23], section 5.1.3). Thus, we have a lower bound that depends on some parameters λ, ν.

Instead of deriving strict proofs related to the Lagrange multipliers and dual problem theory, we give an intuitively easier interpretation of the lower bound

property as a linear approximation of penalty functions approach. The original problem can be rewritten as an unconstrained problem ([23], p. 218)

$$\text{minimize} \quad f_0(x) + \sum_{i=1}^{k} I_-(f_i(x)) + \sum_{i=1}^{l} I_0(g_i(x)), \tag{2.33}$$

where I_- is an indicator function $I_-(u) = \begin{cases} 0 & u \leq 0 \\ \infty & u > 0 \end{cases}$, and similarly I_0 is the indicator function for 0. Thus they express our displeasure associated with a violation of a constraint. For example, for a inequality constraint our displeasure rises sharply from zero to infinity, as $f_i(x)$ moves from non positive to positive. Now, if in the formulation (2.33) we replace the sharp transitions with soft linear ones, i.e. $I_-(u)$ with $\lambda_i u$, where $\lambda_i \geq 0$, and $I_0(u)$ with $\nu_i u$, the objective becomes the Lagrangian . Thus in this formulation for the inequality constraints our displeasure is zero when $f_i(x) = 0$, and it is positive when $f_i(x) > 0$ (assuming $\lambda_i > 0$). So the displeasure grows as the constraint gets more violated. Unlike the original formulation, in which any non-positive value of $f_i(x)$ is acceptable, in the soft formulation we actually derive pleasure from constraints that have margin, i.e. $f_i(x) < 0$ ([23], p. 218). Clearly, the approximation of the indicator function $I_-(u)$ with a linear one $\lambda_i u$ is rather poor. But the linear function is at least an under-estimator of the indicator function, since $\lambda_i u \leq I_-(u)$ for $\lambda_i \geq 0$, and $\nu_i u \leq I_0(u)$ for all u. Thus, we see that the dual function yields a lower bound of the optimal value of the original problem.

The best lower bound that can be obtained from the Lagrange dual function is the solution of the optimization problem

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda_i \geq 0, \quad i = 1, \ldots, k. \end{array} \tag{2.34}$$

This problem is called the *Lagrange dual problem* associated with the problem (2.30) which is always convex since the objective is concave and the constraint is convex ([23] p. 223). If g^* is the optimal value of the dual problem then we have $g^* \leq f^*$ (weak duality). If the equality holds, then strong duality holds with zero duality gap, and the best bound that can be obtained from the Lagrange dual functions is tight. Strong duality does not in general hold. There are extra conditions (constraint qualifications) that establish strong duality. One such is *Slater's* condition ([23], p. 226) that states that for a convex problem, i.e. f_0, f_i are convex and g_i are linear ($Ax = b$), strong duality is guaranteed if it exists at least one strictly feasible point $\bar{x}$, i.e. it holds $f_i(\bar{x}) < 0, \quad i = 1, \ldots, k$, and $A\bar{x} = b$.

Suppose that strong duality holds for a general optimization problem. Let $\bar{x}$ be a primal optimal and $\bar{\lambda}, \bar{\nu}$, dual optimal points, then from the definition of strong duality and the definition of the dual function it holds ([23], p. 242)

$$f_0(\bar{x}) = g(\bar{\lambda}, \bar{\nu}) = \inf_x (f_0(x) + \sum_{i=1}^{k} \bar{\lambda}_i f_i(x) + \sum_{i=1}^{l} \bar{\nu}_i g_i(x)) \tag{2.35}$$

Now since the infimum of the Lagrangian over x is less than or equal its value at $x = \bar{x}$, we have

$$f_0(\bar{x}) \leq f_0(\bar{x}) + \sum_{i=1}^{k} \bar{\lambda}_i f_i(\bar{x}) + \sum_{i=1}^{l} \bar{\nu}_i g_i(\bar{x}) \leq f_0(\bar{x}). \tag{2.36}$$

The last inequality follows from $\bar{\lambda}_i \geq 0, f_i(\bar{x}) \leq 0$ and $g_i(\bar{x}) = 0$. Thus the inequalities in the chain hold with equality. So $\bar{x}$ is a minimizer (not necessary unique) of the Lagrangian $L(x, \bar{\lambda}, \bar{\nu})$ over x and it holds $\sum_{i=1}^{k} \bar{\lambda}_i f_i(\bar{x}) = 0$, which implies $\lambda_i f_i(\bar{x}) = 0$, since every element in the sum is non-positive.

Optimality conditions for any optimization problem with differentiable objective and constraint functions are given by the *Karush-Kuhn-Tucker* (KKT) conditions. They state that if strong duality can be obtained for any pair of primal $\bar{x}$ and dual $\bar{\lambda}, \bar{\nu}$ optimal points with zero duality gap, the following conditions must be satisfied ([23], p. 243)

$$\begin{aligned}
f_i(\bar{x}) &\leq 0, i = 1, \ldots, k && (2.37)\\
g_i(\bar{x}) &= 0, i = 1, \ldots, l && (2.38)\\
\bar{\lambda}_i &\geq 0, i = 1, \ldots, k && (2.39)\\
\bar{\lambda}_i f_i(\bar{x}) &= 0, i = 1, \ldots, k && (2.40)\\
\nabla f_0(\bar{x}) + \sum_{i=1}^{k} \bar{\lambda}_i \nabla f_i(\bar{x}) + \sum_{i=1}^{l} \bar{\nu}_i \nabla g_i(\bar{x}) &= 0. && (2.41)
\end{aligned}$$

The first two conditions state that $\bar{x}$ is primal feasible point and the third that is dual feasible point. The third condition is known as *complementary slackness*. This condition can be written as $\bar{\lambda}_i > 0 \Rightarrow f_i(\bar{x}) = 0$, and additionally $f_i(\bar{x}) < 0 \Rightarrow \lambda_i = 0$. Roughly speaking, this means that the optimal Lagrange multiplier associated with an inequality constraint is zero unless the constraint is active at the optimum. The last condition states that since $\bar{x}$ minimizes $L(x, \bar{\lambda}, \bar{\nu})$, it follows that its gradient must vanish at $\bar{x}$. Finally, if a convex optimization problem has a strictly feasible point, then the KKT conditions are also sufficient for the points to be primal and dual optimal with zero duality gap. In other words, x is the optimal solution of the convex problem if and only if there are λ, ν that all together satisfy the KKT conditions ([23], p. 244).

2.5 Descent methods for unconstrained optimization

Let f be a scalar valued multivariate differentiable function the minimum of which is searched. If f is convex, a necessary and sufficient condition for a point x^{opt} to be optimal is

$$\nabla f(x^{opt}) = 0. \tag{2.42}$$

Thus, solving the unconstrained problem is the same as finding a solution for (2.42). In a few special cases, an analytic solution can be found by solving the

optimality equation. Usually, the problem must be solved by an iterative algorithm. Thus, starting from a given vector $x^0 \in \mathbb{D}$, where $\mathbb{D}$ is the domain of f, an numerical method is searched that computes a sequence of points $x^i \in \mathbb{D}$ with $f(x^i) \to f(x^{opt})$ as $i \to \infty$. The algorithm should be terminated when $f(x^i) - f(x^{opt}) \leq \epsilon$, where $\epsilon > 0$ is some specified tolerance.

If f is strictly convex ($\nabla^2 f(x) \succ 0$) it can be shown that there are constants $m, M > 0$ such that for every $\{x : f(x) \leq f(x^0)\}$ ([23], p. 461)

$$mI \preceq \nabla^2 f(x) \preceq MI. \tag{2.43}$$

The ratio $k = M/m$ is an upper bound for the condition number of the Hessian matrix, i.e. the ratio of its largest eigenvalue to its smallest eigenvalue, which in general influences the speed of convergence of different algorithms (see for example [23], section 9.5.3). By considering second order Taylor approximation and convexity properties it can be shown ([23], section 9.1.2) that for the optimal value of the problem it holds

$$f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \leq f(x^{opt}) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|^2. \tag{2.44}$$

Thus, when the gradient is sufficiently small at a point then that point is nearly optimal, verifying intuition.

A general iterative procedure that updates the current point with a new one has the form

$$x^{i+1} = x^i + a_i d^i, \tag{2.45}$$

for $i = 0, 1, \ldots$, where the scalar parameter $a_i > 0$ gives the length of the step in the direction defined by the vector d^i. The step and the direction must guarantee, since we have a minimization problem, that (*descent methods*)

$$f(x^{i+1}) = f(x^i + a_i d^i) < f(x^i), \tag{2.46}$$

except when x^i is optimal (see also definition (2.13)). The procedure should converge to a vector that approximately minimizes f in a finite number of steps with good accuracy. Therefore, the success of a so-called line search optimization method depends on effective choices for the directions and steps.

A descent direction of f at x^i, for all $a_i > 0$ and sufficiently small, must satisfy (see eq. (2.14))

$$\nabla f(x^i)^T d^i < 0, \tag{2.47}$$

i.e., it must make an acute angle with the negative gradient. So the first part of an optimization algorithm involves the sequential selection of descent directions. The second part involves the selection of the step length a_i. It should formally be (*exact line search*)

$$a_i = \arg\min_{a>0} f(x^i + a d^i), \tag{2.48}$$

i.e. a_i is chosen to minimize f along the ray $\{x^i + a d^i : a > 0\}$. An exact line search is used when the computational cost of the minimization problem with

one variable is low compared to the cost of computing the search direction itself. Other line search methods are inexact. The step length is used to approximately minimize f along the ray, or to just reduce f "enough". Different inexact line search methods have been proposed as well as conditions, for example Armijo's condition, that effective methods should fulfill so that enough reduction to the value of f can be achieved and therefore to obtain convergence to the solution. One inexact line search method that is simple and effective (see for example [23], p. 464) is called *backtracking* line search, and it depends on two constants α, β with $0 < \alpha < 0.5$, and $0 < \beta < 1$. It starts with the selection of a descent direction d^i and $a_i^0 = 1$. Then the step size is reduced by the factor β, i.e. $a_i = a_i^k = \beta a_i^{k-1}$, until

$$f(x_i + a_i d^i) \leq f(x_i) + a_i \alpha \nabla f(x_i)^T d^i. \tag{2.49}$$

2.5.1 Gradient descent and steepest descent

Let x^i be a given point of the iteration and let the step size be constant. Without loss of generality let $a = 1$, and a direction d is searched such that $f(x^i+d) < f(x^i)$. We search for the direction that the rate of change of f is maximized. The rate of change of f at point x^i in the direction of the unit vector $\bar{d} = d/\|d\|$ is given by the directional derivative $\nabla f(x^i)^T \bar{d}$ at point x_i (see also eq. 2.9) for which it holds

$$|\nabla f(x^i)^T \bar{d}| = \|\nabla f(x^i)\| \|\bar{d}\| |\cos\varphi| \leq \|\nabla f(x^i)\|. \tag{2.50}$$

Thus, the maximum rate of change of the function f is given by $\|\nabla f\|$, and will occur in the direction ∇f for $\varphi = 0$, with maximum rate of increment $\|\nabla f(x)\|$, or $-\nabla f$ for $\varphi = \pi$, with maximum rate of decrement $-\|\nabla f(x)\|$. Since by keeping the step constant we search for an optimal direction for the minimization problem, the unnormalized direction

$$d_{gd}^i = -\nabla f(x^i) \tag{2.51}$$

points in a downhill direction where the function f gets decreased, and $\nabla f(x^i)^T d^i = -\nabla f(x^i)^T \nabla f(x^i) < 0$ as long as $\nabla f(x^i) \neq 0$. This is called the direction of the *steepest descent* (unnormalized, for euclidean norm) or *gradient descent* at the point x^i. The directions giving zero rate of change $\varphi = \pi/2$ are those orthogonal to $\nabla f(x^i)$. Thus, in a direction perpendicular to the gradient at point x^i, $f(x)$ remains unchanged. This is because the gradient vector is perpendicular to the level sets or equipotential surfaces of the function.

Finally, the gradient descent method can be formulated as

$$x^{i+1} = x^i - a_i \nabla f(x^i). \tag{2.52}$$

By equating the partial derivative of $f(x^{i+1})$ with respect to a to zero we have

$$\frac{\partial f(x^{i+1})}{\partial a} = \frac{\partial f(x^i - a\nabla f(x^i))}{\partial a} = \nabla f(x^{i+1})^T \frac{\partial (x^i - a\nabla f(x^i))}{\partial a} \tag{2.53}$$

$$= -\nabla f(x^{i+1})^T \nabla f(x^i) = 0. \tag{2.54}$$

So the gradient descent method with exact line search moves in directions orthogonal to previous steps. This provides in general a slow rate of convergence to the optimum (see for example [85], section 3.2.1).

The steepest descent direction can be defined in a more general way. The directional derivative $\nabla f(x)^T d$ gives the approximate change of f for a small d and it must be negative (2.14). Being linear in d it can be made as negative as possible, assuming d is descent, by taking d large (the magnitude measured by a suitable norm). The search for a steepest direction must be, therefore, reduced among descent directions (making acute angle with the negative gradient) that for example obey $\|d\|_{\cdot} \leq \epsilon$, where $\|\|_{\cdot}$ is a suitable norm. For appropriate $\epsilon > 0$, a normalized steepest descent direction (not necessarily unique) can be defined from the following optimization problem ([23], section 9.4)

$$\bar{d}_{sd} = \arg\min\{\nabla f(x)^T d : \|d\|_{\cdot} \leq \epsilon\}. \tag{2.55}$$

i.e., as the direction that extends furthest in the ball defined by $\|\|_{\cdot}$ and ϵ by making acute angle with the negative gradient. From the first-order approximation of f

$$f(x+d) \approx \bar{f}(x+d) = f(x) + \nabla f(x)^T d, \tag{2.56}$$

we can also see that $\bar{d}_{sd}$ is the step that gives the largest decrease in the linear approximation of f under the constraints.

If the norm is selected to be the Euclidean norm, then the steepest descent direction is simply the negative gradient, i.e. $d_{sd} = -\nabla f$. For example, lets select a quadratic norm $\|d\|_P = (d^T P d)^{1/2} = \|P^{1/2} d\|$, where P is symmetric and positive definite matrix. Thus from the KKT conditions we have that $\lambda > 0$ and

$$\nabla_d L(d, \lambda) = \nabla_d (\nabla f(x)^T d) + \lambda \nabla_d (d^T P d - \epsilon^2)) = 0 \tag{2.57}$$

yielding

$$\bar{d} = -\frac{1}{2\lambda} P^{-1} \nabla f(x), \tag{2.58}$$

where λ is defined by the constraint. Thus the unnormalized direction

$$d = -P^{-1} \nabla f(x) \tag{2.59}$$

points to the steepest direction (gives the largest degree of decrement in f) in the ellipsoid. Note, that for $P = I$ we have the Euclidean norm. This has an interesting geometrical interpretation. Define $x' = P^{1/2} x$, so that $\|x\|_P = \|x'\|$. Using this change of coordinates, the original problem of minimizing f can be solved by solving the equivalent problem of minimizing the function $g(x') = f(P^{-1/2} x') = f(x)$. By applying the gradient descent method to g, the search direction at point x', which corresponds to the point $x = P^{-1/2} x'$ for the original, is given by

$$d' = -\nabla_{x'} g(x') = -P^{-1/2} \nabla_x f(P^{-1/2} x') = -P^{-1/2} \nabla_x f(x). \tag{2.60}$$

This gradient search direction corresponds to the direction $d = P^{-1/2} d' = -P^{-1} \nabla f(x)$, for the original problem. In simpler words, the steepest descent

method in the quadratic norm defined by P can be thought as the gradient method applied to the problem after the change of coordinates ([23], p 477). More general, instead of defining a global change of coordinates for the problem, it can be defined a local change, meaning that the choice of norm can be different in every step. This generalizes (2.59) to

$$d = -P(x)^{-1}\nabla f(x). \tag{2.61}$$

The concept can be generalized even more, i.e. when the change of coordinates cannot exist, in terms of differential geometry [1]. Let $\mathbb{S} = \{x \in \mathbb{R}^n\}$ the parameter space where f is defined. When $\mathbb{S}$ is an Euclidean space with an orthonormal coordinate system for the vectors x, then the length of a change Δx can be measured by $\|\Delta x\|^2 = \sum_{i=1}^{n}(\Delta x_i)^2$ and for a non-orthonormal coordinate system by $\|\Delta x\|^2_{\cdot} = \sum_{i,j} g_{ij}\Delta x_i \Delta x_j$. But $\mathbb{S}$ can be a curved manifold with no orthonormal coordinate system. Such a space is a Riemannian space. Then, the matrix $G = \{g_{ij}\}$ is called the Riemannian metric tensor and in general depends on Δx. The steepest descent direction in that space is given by the *natural gradient* named by Amari [1], i.e.

$$d_{\text{nat}} = -G(x)^{-1}\nabla f(x), \tag{2.62}$$

The positive definitiveness of $G(x)^{-1}$ implies

$$-\nabla f(x)^T (G(x)^{-1}\nabla f(x)) < 0. \tag{2.63}$$

Thus d_{nat} is a descent direction, except of when x is optimal. The Riemannian structure of the parameter space is not trivially defined for every problem and a metric is chosen having some proper invariance [1].

2.5.2 Newton's method

Let now consider a second order approximation for the objective function f

$$f(x+d) \approx \bar{f}(x+d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d, \tag{2.64}$$

which is a convex quadratic function of d (assuming the Hessian is positive definite). The Newton direction can be defined from the following optimization problem ([23], section 9.5.1)

$$d_{Nt} = \arg\min\{\bar{f}(x+d)\}, \tag{2.65}$$

i.e., the Newton vector d_{Nt} is what should be added to the point x to minimize the second-order approximation of f at x. It can also be defined as the steepest descent direction in the ball defined by $P(x) = \nabla^2 f(x)$. By differentiating (2.64) with respect to d and setting the result equal to zero, we have the following expression for the Newton direction

$$d_{Nt} = -\nabla^2 f(x)^{-1}\nabla f(x), \tag{2.66}$$

where positive definitiveness of the Hessian implies that is a descent direction, unless x is optimal. If we linearize the optimality condition $\nabla f(x^{opt}) = 0$ near x we obtain

$$\nabla f(x+d) \approx \nabla f(x) + \nabla^2 f(x) d = 0, \tag{2.67}$$

which is a linear equation of d, with solution $d = d_{Nt}$. So the Newton vector is what must be added to x so that the linearized optimality condition holds. This suggests that when x is near x^{opt} the Newton update should be a very good approximation of x^{opt} ([23], p. 485).

2.5.3 Gauss-Newton method

Let us consider the case where the objective function has a specific form, i.e.

$$f(x) = \frac{1}{2}\sum_{i=1}^{m} h_i(x)^2, \tag{2.68}$$

where h_i are twice differentiable functions. The gradient and Hessian are given by

$$\nabla f(x) = \sum_{i=1}^{m} h_i(x)\nabla h_i(x) \tag{2.69}$$

$$\nabla^2 f(x) = \sum_{i=1}^{m} (\nabla h_i(x)\nabla h_i(x)^T + h_i(x)\nabla^2 h_i(x)). \tag{2.70}$$

The resulting minimization procedure based on the Newton's direction is called *Newton-Raphson* method. The *Gauss-Newton* method uses the direction

$$d_{GN} = -(\sum_{i=1}^{m} \nabla h_i(x)\nabla h_i(x)^T)^{-1}(\sum_{i=1}^{m} h_i(x)\nabla h_i(x)) \tag{2.71}$$

$$= -(J_h^T J_h)^{-1} J_h^T h(x), \tag{2.72}$$

where $h(x) = (h_1(x), h_2(x), \ldots, h_m(x))^T$. This direction can be considered as an approximate Newton direction obtained by dropping the second derivative terms from the Hessian of f. There is another interpretation, or derivation, for the Gauss-Newton direction. Using the first order approximation $h_i(x+d) \approx h_i(x) + \nabla h_i(x)^T d$ we obtain the approximation

$$f(x+d) \approx \frac{1}{2}\sum_{i=1}^{m} (h_i(x) + \nabla h_i(x)^T d)^2. \tag{2.73}$$

The Gauss-Newton direction is the minimizer of the approximation (see for example [23], p. 520). The algorithm has the benefit of not requiring the computation of second derivatives. Its performance might be different than that of Newton-Raphson method because it utilizes an approximation for the Hessian.

2.6 Gradient descent methods for functions with matrix argument

2.6.1 Matrix gradient and optimality conditions

Let $X = \{x_{ij}\}$ be a $m \times n$ matrix, and let f be a scalar function of X, i.e

$$f = f(X) = f(x_{11}, \ldots, x_{ij}, \ldots, x_{mn}). \tag{2.74}$$

A typical such a function is the determinant of X. As with the vector gradient, the matrix gradient is a matrix of the same size as X whose elements are the partial derivatives of f with respect to x_{ij}. Thus it is

$$\nabla_X f(X) = \frac{\partial f(X)}{\partial X} = \begin{pmatrix} \frac{\partial f(X)}{\partial x_{11}} & \cdots & \frac{\partial f(X)}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(X)}{\partial x_{m1}} & \cdots & \frac{\partial f(X)}{\partial x_{mn}} \end{pmatrix}. \tag{2.75}$$

The second-order gradient has representation in $R^{m\times n\times m\times n}$. Note, that any matrix can be represented as a vector by stacking together its elements. However, if the matrix X is known to have some specific structure (for example orthogonal) then this can be preserved in the gradient, see for example [20, 55, 33, 39].

In Chapter VI, the gradient of the determinant of an invertible $m \times m$ square matrix X is needed. A well known result for the inverse of a matrix states that

$$X^{-1} = \frac{1}{\det X}\mathrm{adj}(X) = \frac{1}{\det X}\begin{pmatrix} \chi_{11} & \cdots & \chi_{m1} \\ \vdots & \ddots & \vdots \\ \chi_{1m} & \cdots & \chi_{mm} \end{pmatrix}, \tag{2.76}$$

where $\mathrm{adj}(X)$ is the so-called adjoint matrix of X and the scalars χ_{ij} are called cofactors. There are obtained by first taking the $(m-1)\times(m-1)$ sub-matrices of X that remain when the ith row and jth column are excluded, then computing the determinant of the sub matrix, and finally multiplying by $(-1)^{i+j}$. The determinant of X can be expressed in terms of the cofactors as

$$\det X = \sum_{j=1}^{m} x_{ij}\chi_{ij}, \tag{2.77}$$

where the selected row i can be any row. Note, that in the determination of the cofactors χ_{ij} the elements of the ith row do not get involved. Thus, taking partial derivatives with respect to x_{ij} gives

$$\frac{\partial \det X}{\partial x_{ij}} = \chi_{ij}. \tag{2.78}$$

Thus from (2.76) and (2.78) it holds

$$\frac{\partial \det X}{\partial X} = \mathrm{adj}(X)^T = (X^T)^{-1}\det X. \tag{2.79}$$

Now, for the matrix gradient of function $f(X) = \log|\det X|$, where X is invertible, it is ([85], p. 61)

$$\frac{\partial \log|\det X|}{\partial X} = \frac{1}{|\det X|}\frac{\partial|\det X|}{\partial X} = (X^T)^{-1}. \tag{2.80}$$

Let $f : \mathbb{X} \to \mathbb{R}$, where $\mathbb{X} \subset \mathbb{R}^{m \times n}$ is open. Assuming f has continuous first-, second-, and third-order gradients over the open set $\mathbb{X}$, then for $X \in \mathbb{X}$ and any $D \in \mathbb{R}^{m \times n}$ the function can be expanded on some open interval of $\tau \in \mathbb{R}$ as

$$f(X + \tau D) = f(X) + \tau f^1(X; D) + \frac{1}{2!}\tau^2 f^2(X; D) + o(\tau^3), \tag{2.81}$$

where the directional derivatives are in $\mathbb{R}$ and

$$\begin{aligned} f^1(X; D) &= \frac{\partial f(X + \tau D)}{\partial \tau}\Big|_{\tau=0} && (2.82)\\ &= \text{trace}[\nabla f(X)^T D] && (2.83) \end{aligned}$$

$$\begin{aligned} f^2(X; D) &= \frac{\partial^2 f(X + \tau D)}{\partial \tau^2}\Big|_{\tau=0} && (2.84)\\ &= \text{trace}[\nabla_X f^1(X; D)^T D] && (2.85)\\ &= \text{trace}[\nabla_X (\text{trace}[\nabla_X f(X)^T D])^T D] && (2.86) \end{aligned}$$

where $\tau \in \mathbb{R}$. The optimality criteria are the same as in the vector case, for example for a stationary point we have $\nabla f(X) = 0$, but the second directional derivative does not have a general representation as in the vector case (2.12), so we can only write $f^2(X; D) \geq 0$.

2.6.2 Natural gradient

As for a function with vector argument the steepest descent in the Euclidean norm for the minimization procedure of the function $f(X)$ is given by the negative gradient ([41], p. 253)

$$D_{gd} = -\nabla_X f(X). \tag{2.87}$$

The Newton's method cannot be easily defined, since it depends on the second directional derivative. Let us then only consider the first order approximation

$$f(X + D) \approx f(X) + \text{trace}[\nabla f(X)^T D]. \tag{2.88}$$

Thus by keeping the length constant an optimal direction (descent) can be searched that is steepest for some appropriate norm.

Here we restrict the problem to the parameter space of non-singular square matrices, which is a Lie group, also called general linear group. In order to define the natural gradient (2.62) Amari introduced a Riemannien metric to the space of those matrices ([1], see also [39], p. 235, and [85], p. 67) in the following way.

An inner product at X is defined by the squared norm of a deviation $D = \Delta X$

$$< D, D >_X = \|D\|_{\cdot}^2 = \sum_{i,j,k,l} g_{ijkl}(X) d_{ij} d_{kl}, \tag{2.89}$$

where d_{ij} are the elements of D. Additionally, a matrix X, by right multiplication with X^{-1}, is mapped to the unit matrix, i.e. $XX^{-1} = I$. With the same mapping

$X + D$ is mapped to $(X + D)X^{-1} = I + DX^{-1}$. This shows that a deviation D from X is equivalent to a deviation DX^{-1} from I by the correspondence given by multiplication with X^{-1}. According to [1] the Lie group invariance requires that the metric is kept invariant under this correspondence, that is the inner product of D at X is equal to the inner product of DY at XY for any Y. Therefore it must hold

$$< D, D >_X =< DY, DY >_{XY} . \tag{2.90}$$

This must also hold for $Y = X^{-1}$ $(XY = I)$, i.e with $A = \{a_{ij}\} = DX^{-1}$

$$< DX^{-1}, DX^{-1} >_I = \sum_{i,j} a_{ij}^2. \tag{2.91}$$

Thus Amari defined the Riemannian metric structure at point X as

$$< D, D >_X = \text{trace}[(DX^{-1})^T DX^{-1}] = \text{trace}[DX^{-1}(DX^{-1})^T]. \tag{2.92}$$

As in (2.55), we can now consider the following optimization problem for $\epsilon > 0$

$$\bar{D} = \arg\min\{\text{trace}[\nabla f(X)^T D] : \text{trace}[DX^{-1}(X^T)^{-1}D^T] \leq \epsilon^2\}. \tag{2.93}$$

The Lagrangian with $\lambda > 0$ is given by

$$L(D, \lambda) = \text{trace}[\nabla f(X)^T D] + \lambda(\text{trace}[D(X^T X)^{-1}D^T] - \epsilon^2), \tag{2.94}$$

By differentiating with respect to D (note that $\nabla_D \text{trace}(DBD^T) = D(B + B^T)$, for example see [39], p. 541) we have the condition

$$\nabla f(X) + 2\lambda D(X^T X)^{-1} = 0, \tag{2.95}$$

and the natural gradient descent direction (unnormalized) in the standard form is given by ([39], p. 235)

$$D_{nat} = -\nabla f(X) X^T X. \tag{2.96}$$

For extensions see ([39], section 6.2).

A related result was derived in [33] (see also [85], p. 68). From the first order approximation again a direction D of the form $D = AX$, i.e. is proportional to X, was considered. Then (2.88) can be written as

$$f(X + AX) \approx f(X) + \text{trace}[\nabla f(X)^T AX] = f(X) + \text{trace}[(\nabla f(X) X^T)^T A]. \tag{2.97}$$

The multiplier of A in the trace is called *relative* gradient by Cardoso. Therefore the largest decrement in the value of $f(X+D)-f(X)$ is happening at the direction $A = -\nabla f(X) X^T$ or at the direction $D = AX = -\nabla f(X) X^T X$, i.e. the natural gradient.

Chapter III

Probability Theory

In this chapter, different concepts of probability theory are discussed. Probabilistic models, making explicit reference to the nature and effects of chance phenomena, are the foundation upon which different statistical methods are based; for example for estimation and prediction. Of special interest are conditional densities and higher order statistical properties. Some issues of information theory are also treated. Special attention is given to the notion of independence. Significant are also different properties of Gaussian densities. Finally, some properties of stochastic processes are briefly discussed. Most of the concepts presented here can be found from standard probability theory books, for example [161]. For a thorough treatment see for example [59, 60, 11].

3.1 Basic concepts and definitions

An *experiment* can be intuitively described as a procedure that can be repeated under the same well defined conditions an unlimited number of times. After the completion of an experiment some outcome is observed. Experiments can be distinguished to *deterministic* and *random*. In deterministic experiments the selection of the conditions defines completely the outcome. On the other hand, for a random experiment the knowledge of the conditions just specify a set of possible outcomes. The set of all the possible outcomes of a random experiment is called *sample space*. Outcomes of a random experiment can be described in different ways, so that for a given experiment there can exist more than one sample spaces.

A random experiment is well defined if we know the set $\mathfrak{S}$ of all possible outcomes ς, and if we know for "enough" subsets S of $\mathfrak{S}$ the *probability* $P(S)$ that the outcome of the experiment belongs to S. It is not, for example, enough to know only the probabilities of the outcomes. The *sample space* $\mathfrak{S}$ provides a model of an ideal experiment in the sense that, by definition, every thinkable outcome of the experiment is completely described by one and only one sample point ([59], p. 14). Although not precise, this definition of experiment is considered broad enough to encompass the usual scientific experiments and other actions that are regarded simply as observations.

Mathematically, a random experiment is completely described by a sample

space $\mathfrak{S}$, a probability measure P, i.e. a rule for assigning probabilities, and a class $\mathbf{S}$ of admissible subsets of $\mathfrak{S}$ forming the domain set of the probability measure. The triad $(\mathfrak{S}, \mathbf{S}, P)$ is called a probabilistic model or probability space. The class $\mathbf{S}$ is a completely additive family of subsets of $\mathfrak{S}$, i.e. a σ-algebra of events, with the properties: i) $\mathfrak{S} \subset \mathbf{S}$, ii) if $S_k \subset \mathfrak{S}$ for $k = 1, 2, \ldots$, then $\cup_{k=1}^{n} S_k \subset \mathbf{S}$ for $n = 1, 2, \ldots$, and iii) if $S \subset \mathbf{S}$, then $\bar{S} \subset \mathbf{S}$, where $\bar{S}$ is the complement of S relative to $\mathfrak{S}$. An event S, i.e. a subset of the sample space belonging to $\mathbf{S}$, is said to have occurred if the experiment results in an outcome ς that is an element of S, i.e $\varsigma \in S \in \mathbf{S}$. Equivalently the union of events $\cup_{i \in I} S_i$ occurs if at least one of the events $S_i, i \in I$ occurs, and the event $\cap_{i \in I} S_i$ occurs if all events occur.

The probability of an event S, denoted $P(S)$ is a number assigned to this event. In order to have a mathematically well defined theory of probability, that does not depend on the rule used for assigning probabilities to events, Kolmogorov established a definition of probability measure based on *axioms* (e.g. [161], p. 5). According to this definition, a probability measure or simply probability is a set function $P : \mathbf{S} \to \mathbb{R}$ with the properties: i) $P(\mathfrak{S}) = 1$, i.e. the event $\mathfrak{S}$ is certain, ii) is no negative, i.e. $P(S) \geq 0$ for every $S \subset \mathbf{S}$, iii) is σ-additive, i.e. if $S_i \cap S_j = \emptyset$ for $i \neq j$, then $P(\cup_{k=1}^{n} S_k) = \sum_{k=1}^{n} P(S_k)$, with $n = 1, 2, \ldots$ and $\emptyset$ is the empty or null set with $P(\emptyset) = 0$.

In many cases, the probability of occurrence of an event S_i may depend on the occurrence of a related event S_j. The probability of the event S_i given that the S_j is known to have occurred is called *conditional probability* $P(S_i|S_j)$. The probability of an event $\cap_{i \in I} S_i$ is called joint probability of the events $S_i, i \in I$. Then in terms of joint probability, if $P(S_j) \neq 0$, the conditional probability of S_i given S_j is defined by

$$P(S_i|S_j) = \frac{P(S_i \cap S_j)}{P(S_j)}. \tag{3.1}$$

If we also require that $P(S_i) \neq 0$ then

$$P(S_j|S_i) = \frac{P(S_j \cap S_i)}{P(S_i)}, \tag{3.2}$$

and we have *(Bayes' rule)*

$$P(S_i|S_j)P(S_j) = P(S_j|S_i)P(S_i). \tag{3.3}$$

It can be shown that the set function $P(\cdot|S_j)$ forms a probability measure (e.g. [161], p. 28, and [59], chapter V). Let $(\mathfrak{S}, \mathbf{S}, P)$ be a probability space and S_j a simple member of $\mathbf{S}$. Then the information that an event occurred is equivalent to consider that the original probability space is replaced by $(S_j, \mathbf{S}_j, P(\cdot|S_j))$. Clearly, it holds $P(S_j|S_j) = 1$. In that sense, the concept of conditional probability allows the identification of the probability of events based on the occurrence of other evens. If we now consider that the events $S_i, i \in I \subseteq (1, 2, \ldots)$ form a partition of the sample space $\mathfrak{S}$, i.e. $S_i \cap S_j, i \neq j$ and $\cup_{i \in I} S_i = \mathfrak{S}$, with $P(S_i) \neq 0, i \in I$, then for every event S it holds (Theorem of total probability)

$$P(S) = P(S \cup \mathfrak{S}) = \sum_{i \in I} P(S \cup S_i) = \sum_{i \in I} P(S|S_i)P(S_i). \tag{3.4}$$

This form for the probability $P(S)$ is often called marginal probability. If we also require that $P(S) \neq 0$ then we have for every i (Bayes Theorem)

$$P(S_i|S) = \frac{P(S|S_i)P(S_i)}{\sum_{i \in I} P(S|S_i)P(S_i)}. \tag{3.5}$$

An important concept of probability theory is that of *independence*. The events $S_i, i = 1, \ldots, n$ are called independent if for every events $S_{i_1}, \ldots, S_{i_k}$ with $1 < k \leq n$ it holds

$$P(S_{i_1} \cap \ldots \cap S_{i_k}) = P(S_{i_1}) \cdots P(S_{i_k}). \tag{3.6}$$

Therefore, there are $2^n - n - 1$ conditions which must be satisfied ([59], p. 128). The events $S_i, i = 1, \ldots, n$ are called pairwise independent if $P(S_i \cap S_j) = P(S_i)P(S_j)$ for every $i \neq j$. Clearly, if S_i, S_j are independent, from (3.1) we have that

$$P(S_i|S_j) = P(S_i). \tag{3.7}$$

Thus, two events are independent if the occurrence of one does not permit any inference about the occurrence of the other. Note, that although presented oppositely, the basic definition of independence is equation (3.7) ([59], p.125), and (3.6) is a rather technical equivalent definition or direct consequence. Definition (3.6) is usually preferred since it allows $P(S_j) = 0$, in which case $P(S_i|S_j)$ is not defined. It must be noted that the concept of independence, and the related concept of conditional probability, are considered fundamental. In the sense that they justify the mathematical development of probability, not merely as a topic in measure theory, but as a separate discipline ([161], p. 35, and [59], Chapter V).

It is often useful to describe the outcome of a random experiment by a real number. A function defined on a sample space is called a *random variable* ([59], p. 212). Thus a random variable x is a function whose domain is the set of outcomes $\varsigma \in \mathfrak{S}$ and whose range is $\mathbb{R}$. This function must guarantee that for every $S \subset \mathbf{S}$ there is a corresponding set $B \subset \mathbb{R}$ called the image under x of S. Another requirement is that the inverse mapping exists and for every well defined set (Borel set) $B \subset \mathbb{R}$ there exists the inverse image $x^{-1}(B)$ that belongs to $\mathbf{S}$. Thus, by considering a probability space $(\mathfrak{S}, \mathbf{S}, P)$ and a random variable x, it can be defined for every Borel set $B \in \mathbf{B}$ the set function $P_x(B) = P(x^{-1}(B)) = P(\varsigma \in \mathfrak{S} : x(\varsigma) \in B)$, which is a valid probability measure. This probability measure is called probability distribution function or simply distribution of the random variable x. Formally, a random variable x is every real function $x(\varsigma)$ that for every $\bar{x} \in \mathbb{R}$ the set $\{\varsigma : x(\varsigma) \leq \bar{x}\}$ is an event belonging to $\mathbf{S}$. The concept of random variable is analogous extended to define n-dimensional random variables, or simpler random vectors, as vectors whose components are random variables.

3.2 Distribution and density of a random vector

Let $x = (x_1, x_2, \ldots, x_n)^T$ be a random vector with values in $\mathbb{R}^n$. The *cumulative distribution function* (cdf) of x is defined by $P_x(\bar{x}) = P(x \leq \bar{x})$, where $\bar{x}$ is

some constant value of the random vector x. The cdf is a non negative, non-decreasing (often monotonically increasing) continuous function whose values lie in the interval $0 \leq P_x(x) \leq 1$. The multivariate *probability density function* (pdf) $p_x(x) = p_{x_1,x_2,\ldots,x_n}(x_1, x_2, \ldots, x_n)$ is defined as the derivative of the distribution function with respect to all the components of the random vector

$$p_x(\bar{x}) = \frac{\partial}{x_1}\frac{\partial}{x_2}\ldots\frac{\partial}{x_n}P_x(x)|_{x=\bar{x}}, \tag{3.8}$$

hence $P_x(\bar{x}) = \int_{-\infty}^{\bar{x}} p_x(x)dx$. The probability density functions of each random variable $x_i, i = 1, \ldots, n$, named the *marginal* density functions of the random vector x, are obtained by integration

$$p_{x_i}(x_i) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} p_x(x_1, x_2, \ldots, x_n)dx_1\cdots dx_{i-1}dx_{i+1}\cdots dx_n. \tag{3.9}$$

In the same way the *joint* probability function $p_{x,y}(x, y)$ of the random vectors x and y is defined as the multivariate probability density function of the stacked vector $z = (x^T, y^T)^T$. The *conditional* density of x given y is defined as

$$p_{x|y}(x|y) = \frac{p_{x,y}(x, y)}{p_y(y)}, \tag{3.10}$$

whenever $p_y(y) > 0$ and 0 otherwise. Clearly, we can also write

$$p_{y|x}(y|x) = \frac{p_{x,y}(x, y)}{p_x(x)}, \tag{3.11}$$

and we have the *Bayes' rule* for probability densities

$$p_{x|y}(x|y)p_y(y) = p_{y|x}(y|x)p_x(x). \tag{3.12}$$

Useful is also the following expression for the random vectors x, y, z

$$p(x, y|z) = p(x|y, z)p(y|z). \tag{3.13}$$

Assuming that the two n-dimensional random vectors x and y are related by the vector mapping $y = g(x)$ for which the inverse mapping exists and is unique, it can be shown that the probability density function $p_y(y)$ of the transformation y is obtained from the density $p_x(x)$ of x as follows ([161], p. 244)

$$p_y(y) = \frac{1}{|\det J_g(g^{-1}(y))|}p_x(g^{-1}(y)), \tag{3.14}$$

where J_g is the Jacobian matrix (2.17) of $g(x) = (g_1(x), g_2(x), \ldots, g_n(x))^T$. In the special case where the transformation is linear and non-singular, so that $y = Ax$, it holds

$$p_y(y) = \frac{1}{|\det A|}p_x(A^{-1}y). \tag{3.15}$$

3.3 Expectation operator and covariance matrices

Let $g(x)$ denote any quantity (scalar, vector or matrix) derived from the random vector x, then the *expectation* of $g(x)$ is defined by ([85], p. 20)

$$E\{g(x)\} = \int_{-\infty}^{\infty} g(x)p_x(x)dx, \tag{3.16}$$

if the integral exists. For a n-dimensional random vector x its mean or expected value $\eta_x \in \mathbb{R}^n$ is by definition the integral

$$\eta_x = E\{x\} = \int_{-\infty}^{\infty} xp_x(x)dx, \tag{3.17}$$

where the integral is computed over all the components of x, and each component of the vector η_x is given by $E\{x_i\} = \int_{-\infty}^{\infty} x_i p_{x_i}(x_i)dx_i$. Let $y = g(x)$ be a vector-valued function of x, then the expected value of the random vector y is

$$E\{y\} = \int_{-\infty}^{\infty} yp_y(y)dy = \int_{-\infty}^{\infty} g(x)p_x(x)dx = E\{g(x)\}. \tag{3.18}$$

Thus the expectations are equal even though the integrations are carried out over different probability density functions. Based on this theorem it appears that for the determination of the mean of y it is not necessary to find its density p_y. For the random vectors $x_i, i = 1, \ldots, m$, for the nonrandom matrices $A_i, i = 1, \ldots, m$ and for the functions g_i we have for the expectation of the weighted sum

$$E\{\sum_{i=1}^{m} A_i g_i(x_i)\} = \sum_{i=1}^{m} A_i E\{g_i(x_i)\}. \tag{3.19}$$

The conditional mean of the random vector x given y is a function of y and is defined as

$$\eta_{x|y} = E\{x|y\} = \int_{-\infty}^{\infty} xp(x|y)dx. \tag{3.20}$$

The expectation operation can be extended for functions $g(x, y)$ of two or more different random vectors as follows

$$E\{g(x,y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)p_{x,y}(x,y)dxdy. \tag{3.21}$$

With the subscript notation

$$E_x\{g(x,y)\} = \int_{-\infty}^{\infty} g(x,y)p_x(x)dx \tag{3.22}$$

is defined the expectation that is taken only over the random vector x. We also have from (3.21) and (3.10) that

$$\begin{aligned} E\{g(x,y)\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)p_{x|y}(x|y)p(y)dxdy & (3.23)\\ &= E_y\{E\{g(x,y)|y\}\}. & (3.24) \end{aligned}$$

The non-normalized *correlation* matrix of a random vector x is defined as

$$R_x = E\{xx^T\} = \begin{pmatrix} E\{x_1x_1\} & E\{x_1x_2\} & \cdots & E\{x_1x_n\} \\ E\{x_2x_1\} & E\{x_2x_2\} & \cdots & E\{x_2x_n\} \\ \vdots & \vdots & \ddots & \vdots \\ E\{x_nx_1\} & E\{x_nx_2\} & \cdots & E\{x_nx_n\} \end{pmatrix}, \tag{3.25}$$

that is the component-wise expectation of the outer product of x with itself. The correlation matrix is symmetric ($R_x = R_x^T$), positive semidefinite, i.e. ($a^T R_x a \geq 0$ for all the n-dimensional vectors a), and all its eigenvalues are real and non-negative. *Covariance* is the correlation of the random vector $x - \eta_x$

$$C_x = E\{(x-\eta_x)(x-\eta_x)^T\} = E\{xx^T\} - \eta_x\eta_x^T, \tag{3.26}$$

and the conditional covariance of x given y is

$$C_{x|y} = E\{(x-\eta_{x|y})(x-\eta_{x|y})^T|y\} = E\{xx^T|y\} - \eta_{x|y}\eta_{x|y}^T. \tag{3.27}$$

The *cross-correlation* of random vectors x and y is defined as

$$R_{xy} = E\{xy^T\}, \tag{3.28}$$

and the *cross-covariance* of x and y is defined as

$$C_{xy} = E\{(x-\eta_x)(y-\eta_y)^T\} = E\{xy^T\} - \eta_x\eta_y^T. \tag{3.29}$$

By definition, cross-covariance and cross-correlation measure the relation between the random vectors x and y.

For the cross-covariance using equation (3.24) we also have

$$C_{xy} = E_y\{E\{xy^T|y\}\} - \eta_x\eta_y^T = E_y\{E\{x|y\}y^T\} - \eta_x\eta_y^T, \tag{3.30}$$

since when y is given it is not any more random. Since $\eta_x = E\{x\} = E_y\{\eta_{x|y}\}$ we also have the following expressions

$$C_{xy} = E_y\{\eta_{x|y}y^T\} - \eta_x\eta_y^T = E_y\{\eta_{x|y}y^T\} - E_y\{\eta_{x|y}\}\eta_y^T. \tag{3.31}$$

Clearly, it also holds $R_{xy} = E_y\{\eta_{x|y}y^T\}$. The conditional mean $\eta_{x|y}$ is in general a nonlinear function of y. If we assume that it is a linear function, i.e $E\{x|y\} = Ay + a$, where A, a are constant matrix and vector respectively, then we have

$$C_{xy} = E_y\{(Ay+a)y^T\} - E_y\{(Ay+a)\}\eta_y^T = AC_y. \tag{3.32}$$

It can be easily seen that then $\eta_x = A\eta_y + a$, $A = C_{xy}C_y^{-1}$, and

$$\eta_{x|y} = C_{xy}C_y^{-1}y + a = \eta_x + C_{xy}C_y^{-1}(y-\eta_y). \tag{3.33}$$

For the conditional covariance we first observe that

$$E\{(x-\eta_x)(x-\eta_x)^T|y\} = E\{xx^T|y\} - \eta_x\eta_{x|y}^T - \eta_{x|y}\eta_x^T + \eta_x\eta_x^T. \tag{3.34}$$

Then from (3.27) we have

$$C_{x|y} = E\{(x-\eta_x)(x-\eta_x)^T|y\} - (\eta_{x|y}-\eta_x)(\eta_{x|y}-\eta_x)^T, \tag{3.35}$$

and by inserting (3.33) we have after some simple calculations that the conditional covariance is

$$C_{x|y} = E\{(x-\eta_x)(x-\eta_x)^T|y\} - A(y-\eta_y)(y-\eta_y)^T A^T, \tag{3.36}$$

that is a nonlinear function of y. Finally, if we consider expectation over y, by using (3.24), $A = C_{xy}C_y^{-1}$, and that $C_{xy}^T = C_{yx}$ we have

$$E_y\{C_{x|y}\} = E\{(x-\eta_{x|y})(x-\eta_{x|y})^T\} = C_x - C_{xy}C_y^{-1}C_{yx}. \tag{3.37}$$

It must be noted that if the conditional mean is not linear then it cannot anymore be easily identified and it requires the computation of the posterior density $p_{x|y}$.

3.4 Characteristic functions and higher-order statistics

For a random variable x with probability density function $p_x(x)$ the *k-th moment* m_k, $k = 1, 2, \ldots$, is defined by the expectation ([161], p. 146)

$$m_k = E\{x^k\} = \int_{-\infty}^{\infty} x^k p_x(x)dx, \tag{3.38}$$

and the *k-th central moment* μ_k by the expectation

$$\mu_k = E\{(x-m_1)^k\} = \int_{-\infty}^{\infty} (x-\eta_x)^k p_x(x)dx. \tag{3.39}$$

The first moment m_1 is the mean η_x of x, and the second central moment μ_2 is the variance σ_x^2 of x. Mean and variance give only a limited characterization of p_x. Knowledge of other moments provides additional information, that can, for example, be used to distinguish between two densities with the same η and σ^2. The third central moment

$$\mu_3 = E\{(x-\eta_x)^3\} = E\{x^3\} - 3E\{x^2\}\eta_x + 2\eta_x^3 \tag{3.40}$$

is called *skewness* and it is a measure of the asymmetry of the density function, being zero for probability densities that are symmetric around their mean.

The moments can be determined through the *characteristic function* $\Phi(\omega)$, which is defined as the continuous Fourier transform of the density $p_x(x)$ of x

$$\Phi(\omega) = E\{e^{i\omega x}\} = \int_{-\infty}^{\infty} e^{i\omega x} p_x(x)dx, \tag{3.41}$$

where $i^2 = -1$ and $|\Phi(\omega)| \leq \Phi(0) = 1$. Every distribution function is uniquely determined by its characteristic function and the inversion of the Fourier transform

gives $p_x(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\Phi(\omega)e^{-i\omega x}d\omega$ ([161], p. 153). If $i\omega$ is changed with z, the function $\Phi(z)$ is called the *moment generating function* of x. Clearly, by differentiating k times we obtain

$$\frac{d^k\Phi(z)}{dz^k} = E\{x^k e^{zx}\}. \tag{3.42}$$

The derivatives at zero equal the moments of x ([161], p. 154)

$$E\{x^k\} = m_k = \frac{d^k\Phi(z)}{dz^k}|_{z=0}. \tag{3.43}$$

By expanding the moment generating function into Taylor series we have

$$\Phi(z) = \sum_{k=0}^{\infty} m_k \frac{z^k}{k!}. \tag{3.44}$$

This is valid only if all the moments are finite and the series converges absolutely near zero. So since the characteristic function defines p_x, the density of a random variable is uniquely determined if all its moments are known. The second characteristic function $\varphi(\omega)$ of x or *cumulant generating function* is given by the natural logarithm of the first characteristic function. Then for the cumulant generating function we have ([161], p. 154)

$$\varphi(z) = \ln(\Phi(z)) = \sum_{k=0}^{\infty} \kappa_k \frac{z^k}{k!}, \tag{3.45}$$

where κ_k is by definition the k-th cumulant of x and is obtained as the derivative

$$\kappa_k = \frac{d^k\varphi(z)}{dz^k}|_{z=0}. \tag{3.46}$$

The first four cumulants are the most commonly used, and it holds ([85], p. 41)

$$\kappa_1 = E\{x\} = m_1 = \eta_x, \tag{3.47}$$
$$\kappa_2 = E\{x^2\} - \eta_x^2 = \mu_2 = \sigma_x^2, \tag{3.48}$$
$$\kappa_3 = E\{x^3\} - 3E\{x^2\}\eta_x + 2\eta_x^3 = \mu_3, \tag{3.49}$$
$$\kappa_4 = E\{x^4\} - 3(E\{x^2\})^2 - 4E\{x^3\}\eta_x + 12E\{x^2\}\eta_x^2 - 6\eta_x^4. \tag{3.50}$$

The forth cumulant κ_4 is called *kurtosis*. A distribution having zero kurtosis is called *mesokurtic*. A typical mesokurtic distribution is the Gaussian distribution. Distributions having a negative kurtosis are said to be *platykurtic (or subgaussian)*, and if the kurtosis is positive they are said to be *leptokurtic (or supergaussian)*.

The definition of moments can be extended in the multivariate case to define joint moments for random vectors. The joint characteristic function of a random vector x is again the Fourier transform of the joint density ([85], p. 42)

$$\Phi(\omega) = E\{e^{i\omega^T x}\} = \int_{-\infty}^{\infty} e^{i\omega^T x} p_x(x)dx, \tag{3.51}$$

where $i^2 = -1$ and ω is now a column vector with the same dimension as x. Moments and cumulants (cross-cumulants) of x are the coefficients of the Taylor series expansion of the first and second joint characteristic function respectively, defined by the appropriate partial derivatives. It can be shown that for a zero mean random vector $x = (x_1, x_2, \ldots, x_n)^T$ the second, third and fourth order cumulants are for every $i, j, k, l = 1, \ldots, n$ ([85], p. 42)

$$\begin{aligned} cum(x_i, x_j) &= E\{x_i x_j\}, && (3.52)\\ cum(x_i, x_j, x_k) &= E\{x_i x_j x_k\}, && (3.53)\\ cum(x_i, x_j, x_k, x_l) &= E\{x_i x_j x_k x_l\} - E\{x_i x_j\}E\{x_k x_l\} \\ &\quad -E\{x_i x_k\}E\{x_j x_l\} - E\{x_i x_l\}E\{x_j x_k\}. && (3.54) \end{aligned}$$

Therefore, the covariance matrix of a random vector contains all the cumulants of order two, and thus all second order properties of the joint density.

3.5 Uncorrelatedness and independence

Mathematically, independence is defined in terms of probabilities (3.6) or conditional probabilities (3.7). Therefore, the random variables x, y are called independent if the events $\{x \leq \bar{x}\}$ and $\{y \leq \bar{y}\}$ are independent. Similar definitions can be applied for the joint distributions, densities and conditional densities. The random vector x has mutually independent components or the random variables x_i are statistically independent if and only if equivalently ([161], p. 244)

$$\begin{aligned} P_x(x) &= \prod_i P_{x_i}(x_i), && (3.55)\\ p_x(x) &= \prod_i p_{x_i}(x_i). && (3.56) \end{aligned}$$

For the expectation of the product of independent variables $x_1, x_2, \ldots, x_n$ it holds

$$\begin{aligned} E\{x_1 x_2 \cdots x_n\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_1 x_2 \cdots x_n p_x(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_1 x_2 \cdots x_n p_{x_1}(x_1) p_{x_2}(x_2) \cdots p_{x_n}(x_n) dx_1 dx_2 \ldots dx_n. \end{aligned}$$

Thus we have the following property for independent variables

$$E\{x_1 x_2 \ldots x_n\} = E\{x_1\}E\{x_2\}\ldots E\{x_n\}. \quad (3.57)$$

If the random variables $x_1, x_2, \ldots, x_n$ are independent, then it can be shown that the random variables $g_1(x_1), g_2(x_2), \cdots, g_n(x_n)$ are also independent ([161], p. 245) and the previous property of independence is generalized for any absolutely integrable functions $g_1, g_2, \ldots, g_n$ ([85], p. 27). Note, that the random variables $x_1, x_2, \ldots, x_n$ are independent if and only if

$$E\{g_1(x_1)g_2(x_2)\ldots g_n(x_n)\} = E\{g_1(x_1)\}E\{g_2(x_2)\}\cdots E\{g_n(x_n)\}, \quad (3.58)$$

for all functions $g_1, g_2, \ldots, g_n$ that the integrals exist. For specific functions equation (3.58) is just a consequence of independence. Based on this alternative definition of independence (3.58) we have that for the function $e^{i\omega_j x_j}$ it holds

$$E\{e^{i(\omega_1 x_1 + \ldots + \omega_n x_n)}\} = E\{e^{i\omega_1 x_1}\} \cdots E\{e^{i\omega_n x_n}\}. \tag{3.59}$$

Hence, for the characteristic functions of independent random variables it holds

$$\Phi_x(\omega) = \prod_i \Phi_{x_i}(\omega_i) \tag{3.60}$$

Conversely, from the inversion of the Fourier transform, if (3.60) holds then the independence definition based on densities is obtained, and thus equation (3.60) provides an alternative definition for independence through the characteristic functions. From the definition of independence it is clear that every subset of a set of independent variables is a set of independent variables as well. The opposite is not true. For example, if the random variables x_i, $i = 1, \ldots, n$ are pairwise independent, i.e. $p_{x_i x_j}(x_i, x_j) = p_{x_i}(x_i) p_{x_j}(x_j)$, for every $i \neq j$, then they are not necessarily mutually independent. The components of the random vector x are mutually uncorrelated if

$$C_x = E\{(x - \eta_x)(x - \eta_x)^T\} = D, \tag{3.61}$$

where D is a diagonal matrix whose non-zero entries are the variances of the components of x. Clearly, independent components of a random vector are also uncorrelated, the opposite is not true since independence is a much stronger condition. For a zero mean mutually uncorrelated random vector all the cumulants of order two (3.52) vanish for every $i \neq j$. For a zero mean mutually independent random vector all the cumulants (except the expectations of the powers of the same component) vanish.

The independence between random vectors is defined in the same way. For example, the random vectors x, y are independent if

$$p(x|y) = p_x(x). \tag{3.62}$$

Let $x, y, z, \ldots$ independent random vectors then it holds

$$p_{x,y,z,\ldots}(x, y, z, \ldots) = p_x(x) p_y(y) p_z(z) \cdots \tag{3.63}$$

$$E\{g_x(x) g_y(y) g_z(z) \cdots\} = E\{g_x(x)\} E\{g_y(y)\} E\{g_z(z)\} \cdots . \tag{3.64}$$

Two random vectors x and y are uncorrelated if their cross-covariance matrix C_{xy} is a zero matrix

$$C_{xy} = E\{(x - \eta_x)(y - \eta_y)^T\} = 0, \tag{3.65}$$

or equivalently if

$$R_{xy} = E\{xy^T\} = E\{x\} E\{y^T\} = \eta_x \eta_y^T. \tag{3.66}$$

3.6 Entropy and mutual information

An alternative approach for characterizing random variables is given by *information theory*. Starting point of information theory is the concept of *entropy*. Although, the functional of entropy was derived from a number of postulates based on heuristic understanding of uncertainty, its properties are developed axiomatically within the framework of probability theory [161, 133]. Intuitively, entropy of a random variable can be interpreted as the degree of information that the observation of the variable gives. Since, entropy is a measure of uncertainty about the random variable, the more random it is the larger its entropy. For discrete random variables the entropy is positive and it is exactly used as a measure of uncertainty about the random variable. However, this is not so for the continuous case. Then the entropy can take any value from $-\infty$ to ∞, and it is used to measure changes in uncertainty. For a continuous random vector x the entropy or *differential entropy* $H(x)$ is defined by the expectation ([86], p. 108)

$$H(x) = -\int p_x(x) \log p_x(x) dx = -E\{\log p_x(x)\}. \tag{3.67}$$

For the joint entropy of the random vectors x, y we have

$$H(x,y) = -E\{\log(p(x,y))\} = -E\{\log(p(x|y)p_y(y))\} = H(x|y) + H(y), \tag{3.68}$$

where $H(x|y)$ is the *conditional* entropy over all values of y ([161], p. 656), i.e.

$$H(x|y) = -E_y\{E\{\log p(x|y)|y\}\}. \tag{3.69}$$

Considering the invertible transformation $y = g(x)$, the entropy of y is (3.14)

$$H(y) = -E\{\log p_y(y)\} = -E\{\log \frac{1}{|\det J_g(g^{-1}(y))|} p_x(g^{-1}(y))\}. \tag{3.70}$$

Thus, we obtain the following relation between the entropies

$$H(y) = H(x) + E\{\log |\det J_g(g^{-1}(y))|\}. \tag{3.71}$$

So the transformation increases the entropy. If the transformation does not have unique inverse then $H(y) \leq H(x) + E\{\log |\det J_g(x)|\}$ ([161], p. 660). In the case that the transform is linear $y = Ax$ and invertible we have (3.15)

$$H(y) = H(x) + \log |\det A|. \tag{3.72}$$

If the n-dimensional random vector has independent components then from (3.67) and (3.56) it holds

$$H(x) = \sum_{i=1}^{n} H(x_i). \tag{3.73}$$

Based on entropy, *mutual information* of a random vector x is defined as follows

$$I(x) = I(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} H(x_i) - H(x) = E_x\{\log \frac{p_x(x)}{\prod_{i=1}^{n} p_{x_i}(x_i)}\}, \tag{3.74}$$

where p_x is the joint density and p_{x_i} the marginal densities ([86], p. 110). Consider two random vectors x and y with densities $p_x(x)$ and $p_y(y)$ respectively . The *Kullback-Leibler divergence* [128] between two probability densities is defined by

$$\delta(p_x, p_y) = \int p_x(x) \log \frac{p_x(x)}{p_y(x)} dx = E\{\log \frac{p_x(x)}{p_y(x)}\}. \tag{3.75}$$

Kullback-Leibler divergence (or *relative entropy*) is not a proper distance measure because it is not symmetric, but it satisfies

$$\delta(p_x, p_y) \geq 0, \tag{3.76}$$

with equality if and only if $p_x = p_y$ (see for example [85], p.110). This property is due to the convexity of the logarithm and Jensen's inequality. By the form of Kullback-Leibler divergence we can obtain the mutual information of a random vector x

$$I(x) = \delta(p_x, \prod_{i=1}^{n} p_{x_i}) = \int p_x(x) \log \frac{p_x(x)}{\prod_{i=1}^{n} p_{x_i}(x_i)} dx. \tag{3.77}$$

From the properties of Kullback-Leibler divergence we have that mutual information vanishes if and only if the variables x_i are independent and it is strictly positive otherwise. So mutual information is a measure of dependence between the random variables x_i ([85], p. 111, and [39], p. 234). Based on that we also have that in general (3.74, 3.76)

$$H(x) \leq \sum_{i=1}^{n} H(x_i), \tag{3.78}$$

and ([161], p. 658)

$$H(x|y) \leq H(x), \tag{3.79}$$

with equality for the second equation when the vectors x, y are independent.

3.7 Gaussian probability density functions

3.7.1 Normal random variables

The function defined by $p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is called the *(standard) normal or Gaussian* density function, and its integral $\int_{-\infty}^{\bar{x}} p(x)dx$ is the standard normal distribution function ([59], p. 174). Then x is said to have standard normal $N(0,1)$ distribution with mean $\eta = 0$ and variance $\sigma^2 = 1$. The $N(\eta, \sigma^2)$ distribution is by definition that of $\eta + \sigma y$ where $y \sim N(0,1)$, and it holds ([161], p. 162)

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\eta)^2/2\sigma^2}, \tag{3.80}$$

$$\Phi(\omega_0) = e^{i\omega_0\eta} e^{-\omega_0^2\sigma^2/2}. \tag{3.81}$$

If x is zero mean then ([161], p. 148)

$$E\{x^n\} = \begin{cases} 0, & n = 2k+1 \\ 1 \cdot 3 \cdots (n-1)\sigma^n, & n = 2k \end{cases} \tag{3.82}$$

The odd moments of x are zero because $p(-x) = p(x)$. For the kurtosis it holds

$$\kappa_4 = E\{x^4\} - 3(E\{x^2\})^2 = 0. \tag{3.83}$$

Gaussian random variables have some properties that are not shared by every other distribution. An important property is that: If x_1 and x_2 are independent random variables and if two linear combinations $a_1x_1 + a_2x_2$ and $b_1x_1 + b_2x_2$ are also independent, where a_1, a_2, b_1, and b_2 represent nonzero coefficients, then all random variables are normally distributed. Thus if two nontrivial linear combinations of two independent random variables are also independent, then all of them represent normal random variables. This theorem is due to *Darmois and Skitovitch* (see for example [161], p. 217, [45], [60]).

Another important property of the normal distribution is given by the *central limit theorem*. If $x_1, x_2, \ldots$ are independent identically distributed random variables with finite mean $E\{x_i\} = \eta$ and finite variance $E\{(x_i - \eta)^2\} = \sigma^2$ then the limiting distribution of

$$y = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \eta}{\sigma} \tag{3.84}$$

is that of $N(0, 1)$ (e.g. [59], p. 244). This type of convergence is called convergence in distribution implying that for large n the distribution of y is close to standard normal. Several different forms of the theorem exist, where requirements on independence and identical distributions have been weakened [59, 60, 161].

3.7.2 Normal random vectors

An n-dimensional random vector $x = (x_1, x_2, \ldots, x_n)^T$ is said to be Gaussian or normal if its probability density function has the form

$$p_x(x) = \frac{1}{(2\pi)^{n/2}(\det C_x)^{1/2}} \exp(-\frac{1}{2}(x - \eta_x)^T C_x^{-1}(x - \eta_x)). \tag{3.85}$$

So the joint density is an exponential whose exponent is a negative quadratic. Clearly, the mean η_x and the covariance matrix C_x (assumed positive definite) are sufficient for defining the multivariate Gaussian density completely and this is denoted by $x \sim N(\eta_x, C_x)$. Also, the contours of the multivariate Gaussian density, i.e. $p(x) = c$ where c is constant, are ellipses centered at η_x, i.e. $(x-\eta_x)^T C_x^{-1}(x - \eta_x) = c'$. The principal axes of the ellipse are parallel to the eigenvectors of C_x and the eigenvalues λ_i are the respective variances.

Let x, y be jointly Gaussian, then for the joint density, i.e the density of the vector $(x^T, y^T)^T$, if we omit the means it holds

$$p(x, y) = \frac{\exp\left\{-\frac{1}{2}(x^T, y^T) \begin{pmatrix} C_x & C_{xy} \\ C_{yx} & C_y \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix}\right\}}{(2\pi)^{(n+m)/2}(\det C)^{1/2}}, \tag{3.86}$$

where n, m the dimensions of x, y, and $\det C$ the determinant of joint covariance block matrix which is

$$\begin{aligned}\det\begin{pmatrix} C_x & C_{xy} \\ C_{yx} & C_y \end{pmatrix} &= \det\left(\begin{pmatrix} I & C_{xy} \\ 0 & C_y \end{pmatrix}\begin{pmatrix} C_x - C_{xy}C_y^{-1}C_{yx} & 0 \\ C_y^{-1}C_{yx} & I \end{pmatrix}\right) \\ &= \det C_y \det(C_x - C_{xy}C_y^{-1}C_{yx}). \end{aligned} \tag{3.87}$$

The matrix inversion lemma [69] gives the inverse of the joint covariance of x, y

$$C^{-1} = \begin{pmatrix} C_x & C_{xy} \\ C_{yx} & C_y \end{pmatrix}^{-1} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{3.88}$$

where

$$C_{11} = (C_x - C_{xy}C_y^{-1}C_{yx})^{-1} = C_x^{-1} + C_x^{-1}C_{xy}C_{22}C_{yx}C_x^{-1} \tag{3.89}$$

$$C_{22} = (C_y - C_{yx}C_x^{-1}C_{xy})^{-1} = C_y^{-1} + C_y^{-1}C_{yx}C_{11}C_{xy}C_y^{-1} \tag{3.90}$$

$$C_{12} = C_{21}^T = -C_{11}C_{xy}C_y^{-1} = -C_x^{-1}C_{xy}C_{22}. \tag{3.91}$$

Based on this, the joint density can be written in the form

$$p(x, y) = q(y)h(x, y), \tag{3.92}$$

where

$$q(y) = \frac{\exp(-\frac{1}{2}y^T C_y^{-1} y)}{(2\pi)^{m/2}(\det C_y)^{1/2}}, \tag{3.93}$$

$$h(x, y) = \frac{\exp(-\frac{1}{2}(x - C_{xy}C_y^{-1}y)^T C_{11}(x - C_{xy}C_y^{-1}y))}{(2\pi)^{n/2}(\det C_{11}^{-1})^{1/2}}. \tag{3.94}$$

Clearly, the function $h(x, y)$ is for fixed y a Gaussian density. Now, the marginal density of y is given by

$$p_y(y) = \int p(x, y)dx = q(y) \int h(x, y)dx = q(y). \tag{3.95}$$

Thus, marginal densities of jointly Gaussian distributed vectors are also Gaussian. The Gaussian conditional density of x given y is

$$p(x|y) = \frac{p(x, y)}{p(y)} = h(x, y), \tag{3.96}$$

and is of the form

$$p(x|y) \propto \exp(-\frac{1}{2}(x - \eta_{x|y})^T C_{x|y}^{-1}(x - \eta_{x|y})), \tag{3.97}$$

where

$$\eta_{x|y} = C_{xy}C_y^{-1}y \tag{3.98}$$

$$C_{x|y} = C_x - C_{xy}C_x^{-1}C_{yx}. \tag{3.99}$$

So the conditional mean is always linear. This is exactly, if we include the means, the linear conditional mean derived in (3.33). This is an important property related to mean square estimation methods. Consider that the random vector x is Gaussian having uncorrelated components, thus $C_x = D$, where D is diagonal matrix with the variances of the components $\sigma_{x_i}^2$ on the diagonal. The matrix C_x^{-1} is also diagonal with elements $1/\sigma_{x_i}^2$ and the density is

$$p_x(x) = \frac{1}{(2\pi)^{n/2}}(\prod_{i=1}^{n}\frac{1}{\sigma_{x_i}})\exp(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\eta_{x_i})^2}{\sigma_{x_i}^2}) = \prod_{i=1}^{n} p_{x_i}(x_i). \tag{3.100}$$

So uncorrelated jointly distributed Gaussian variables are also independent.

For the multivariate Gaussian density, the marginal densities are also Gaussian $x_i \sim N(\eta_{x_i}, \sigma_{x_i}^2)$. So if $x_1, x_2, \ldots, x_n$ are jointly normal distributed, then all x_i have normal distributions. There is an alternative definition that express the normality of n random variables in terms of the normality of a single random variable. This states that ([161], p. 257) the random variables x_i are jointly Gaussian if and only if the sum

$$\sum_{i=1}^{n}\omega_i x_i = \omega^T x = y, \tag{3.101}$$

is a Gaussian random variable for every constant column vector ω with the same dimension as x. Based on that definition if x is a Gaussian vector then the random variable y is Gaussian $N(\omega^T\eta_x, \omega^T C_x \omega)$. From (3.81) and without loss of generality by setting $\omega_0 = 1$ we have for the joint characteristic function of x

$$\Phi(\omega) = E\{e^{i\omega^T x}\} = e^{i\omega^T\eta_x}e^{-\omega^T C_x\omega/2}. \tag{3.102}$$

The inversion theorem gives (3.85). Based on the definition (3.102), the random vector y defined by the linear invertible transform $y = Ax$, is also Gaussian with mean $\eta_y = A\eta_x$ and covariance matrix $C_y = AC_xA^T$ since ([169], p. 91)

$$\Phi_y(\omega) = E\{e^{i\omega^T y}\} = E\{e^{i\omega^T A^{-1}x}\} = E\{e^{\bar{\omega}^T x}\} = \Phi_x(\bar{\omega}). \tag{3.103}$$

Note, that with the definition (3.102), the covariance matrix is not required to be positive definite ([108] p. 73). If A is an orthogonal matrix and $x = A^T y$ then y is Gaussian with probability density independent of A, since

$$\begin{aligned} p_y(y) &= \frac{1}{(2\pi)^{n/2}(\det C_y)^{1/2}}\exp(-\frac{1}{2}(y-\eta_y)^T C_y^{-1}(y-\eta_y)) \\ &= \frac{1}{(2\pi)^{n/2}(\det AC_xA^T)^{1/2}}\exp(-\frac{1}{2}(x-\eta_x)^T A^T(AC_xA^T)^{-1}A(x-\eta_x)) \\ &= \frac{1}{(2\pi)^{n/2}(\det C_x)^{1/2}}\exp(-\frac{1}{2}(x-\eta_x)^T C_x^{-1}(x-\eta_x)). \end{aligned} \tag{3.104}$$

Thus any orthogonal linear transform of independent or uncorrelated Gaussian variables cannot be estimated because it does not influence the joint pdf (it does not appear at the pdf).

The entropy of a n-dimensional random vector x having a Gaussian distribution can be evaluated as ([85], p. 113)

$$\begin{aligned} H(x) &= -E\{\log \frac{1}{(2\pi)^{n/2}|(\det C_x)^{1/2}|} \exp(-\frac{1}{2}(x-\eta_x)^T C_x^{-1}(x-\eta_x))\} \\ &= E\{\log((2\pi)^{n/2}|(\det C_x)^{1/2}|)\} + E\{\frac{1}{2}(x-\eta_x)^T C_x^{-1}(x-\eta_x)\} \\ &= \frac{n}{2}\log 2\pi + \frac{1}{2}\log|\det C_x| + E\{\frac{1}{2}(x-\eta_x)^T C_x^{-1}(x-\eta_x)\} \\ &= \frac{1}{2}\log|\det C_x| + \frac{n}{2}(1+\log 2\pi). \end{aligned} \tag{3.105}$$

It can be shown ([85], p. 112, [161], p. 669) that the Gaussian distribution has maximum entropy among all distributions with a given mean and covariance matrix. Negentropy $J(x)$ of a random vector x is defined as ([85], p. 112)

$$J(x) = H(x') - H(x), \tag{3.106}$$

where x' is a Gaussian vector with the same mean and covariance matrix as x. Negentropy is always non-negative and it is zero if and only if x has a Gaussian distribution [45, 85]. It is also invariant by any linear invertible change of coordinates ([85], p. 113). Using negentropy there is another expression for mutual information given by [45]

$$I(x) = \sum_{i=1}^{n}(H(x_i') - J(x_i)) - H(x') + J(x) \tag{3.107}$$

$$= J(x) - \sum_{i=1}^{n} J(x_i) + \frac{1}{2}\log\frac{\prod_{i=1}^{n}\sigma_{x_i}^2}{\det C_x}. \tag{3.108}$$

3.8 Stochastic processes

A random or stochastic process is a family $\{x_t, t \in \mathbb{T}\}$ of random variables, all defined on the same probability space $(\mathfrak{S}, \mathbf{S}, P)$. The parameter t is usually referred as time and the set $\mathbb{T}$ is the parameter set of the process, and can be continuous or discrete defining respectively continuous or discrete stochastic processes. Here, stochastic processes with discrete parameter set, where the random variables x_t take continuous values, are only considered. A random process is a function of both time and outcomes ς, i.e $x_t = x(t, \varsigma)$. Thus, for fixed time and outcome $x(t_o, \varsigma_o)$ is the state that the process is at that moment, for fixed time $x(t_o, \varsigma)$ is just a random variable and for fixed outcome $x_t = x(t, \varsigma_o)$ is a function of time that describes one evolution of the process that is called a realization of the process. A collection of many realizations of the process is called an ensemble ([169], p. 101).

A k-th order distribution of a random process is the joint distribution function $P_{t_1,t_2,\ldots,t_k}$ of the random variables $x_{t_1}, x_{t_2}, \ldots, x_{t_k}$. For the determination of all the statistical properties of a stochastic process, knowledge is required of $P_{t_1,t_2,\ldots,t_k}$

for every $k = 1, 2, \ldots$ and for every $t_1, t_2, \ldots, t_k$ ([169], p. 104) . If the random variables x_t are independent and identically distributed (i.i.d) then the process is just a sequence of i.i.d random variables. Note also that if the parameter set is finite then the path of the process $\{x_t, t = 1, 2, \ldots, T\}$ can be considered as a random vector $x = (x_1, x_2, \ldots, x_T)^T$. Additionally, we can define vector valued random processes, where for fixed t, x_t is a n-dimensional random vector

$$x_t = (x_{1,t}, x_{2,t}, \ldots, x_{n,t})^T. \tag{3.109}$$

In many applications, only the second order properties, i.e. second order temporal statistics, of a process are needed which in general are functions of time. The mean and variance of a process are

$$\eta_{x_t} = E\{x_t\}, \quad \sigma^2_{x_t} = E\{(x_t - \eta_{x_t})^2\} = \gamma_x(t), \tag{3.110}$$

and the autocovariance $\gamma_x(t_1, t_2)$, i.e the covariance of the random variables x_{t_1} and x_{t_2}, is a function of t_1, t_2 and

$$\gamma_x(t_1, t_2) = E\{(x_{t_1} - \eta_{x_{t_1}})(x_{t_2} - \eta_{x_{t_2}})\}. \tag{3.111}$$

If we use the time lag τ between t_1 and t_2 the autocovariance function becomes

$$\gamma_x(t, \tau) = E\{(x_t - \eta_{x_t})(x_{t+\tau} - \eta_{x_{t+\tau}})\}. \tag{3.112}$$

Analogously is defined the cross-covariance function $\gamma_{xy}(t, \tau)$ between two processes x_t and y_t

$$\gamma_{xy}(t, \tau) = E\{(x_t - \eta_{x_t})(y_{t+\tau} - \eta_{y_{t+\tau}})\}. \tag{3.113}$$

3.8.1 Stationarity

A property that can characterize stochastic processes is that of stationarity. The class of stationary processes contains processes that their statistical properties depend only on the time difference, and are invariant to a shift of the origin. Mathematically, a stochastic process is strict-sense stationary if for every k, τ and for every $t_1, t_2, \ldots, t_k$ it holds $P_{t_1,t_2,\ldots,t_k} = P_{t_1+\tau,t_2+\tau,\ldots,t_k+\tau}$. A more general class is that of wide sense (or weakly, or second order) stationary processes containing finite variance processes that ([169], p. 106)

$$\eta_x = E\{x_t\}, \quad \gamma_x(\tau) = E\{(x_t - \eta_x)(x_{t+\tau} - \eta_x)\} \quad \forall t, \tag{3.114}$$

where for the $\tau = 0, 1, \ldots$ time lag autocovariances it can be shown that it holds $\gamma(-\tau) = \gamma(\tau) \leq \gamma(0)$ ([169], p. 109). Analogously, the autocorrelation sequence $r_x(\tau) = E\{x_t x_{t+\tau}\}$ is defined which coincides with the autocovariance sequence for zero mean processes. Assuming that the autocovariances or the autocorrelations are known up to a lag τ then they can be summarized in a matrix form

$$\Gamma_x = \begin{pmatrix} \gamma_x(0) & \gamma_x(1) & \cdots & \gamma_x(\tau) \\ \gamma_x(1) & \gamma_x(0) & \cdots & \gamma_x(\tau - 1) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_x(\tau) & \gamma_x(\tau - 1) & \cdots & \gamma_x(0) \end{pmatrix}. \tag{3.115}$$

This matrix, namely autocovariance matrix, has all the properties of covariance matrices defined earlier, furthermore it is Toeplitz having the same elements in the main diagonal and each sub-diagonal. It can be also interpreted as the covariance matrix of the $(\tau+1)$-dimensional random vector $x = (x_t, x_{t+1}, \ldots, x_{t+\tau})^T$ with the additional properties (3.114). Higher order statistics of a stationary process can be defined in an analogous manner. Applications of stationary process in time series analysis or signal processing can be found, for example in [160, 169].

3.8.2 Markovian processes

An important class of stochastic processes is that of *Markovian or Markov* processes, which are the simplest time dependent stochastic processes [59]. A Markov process is a process $\{x_t, t \in \mathbb{T}\}$ that given the present value of the process the families of the random variables $\{x_p : p < t\}$ (past) and $\{x_f : f > t\}$ (future) are statistically independent. Thus, they are used to describe stochastic systems that their future development depends only on the present state, and not on the the manner in which the present state was reached. This property is called Markovian or *memory less property*. Mathematically, in terms of conditional probabilities, the Markovian property, for $t_{i-1} < t_i$, is expressed as ([161], p. 695)

$$P\{x_{t_i} \leq x_i | x_t, t \leq t_{i-1}\} = P\{x_{t_i} \leq x_i | x_{t_{i-1}}\}. \tag{3.116}$$

From this it follows that if $t_1 < t_2 < \cdots < t_n$ then

$$P\{x_{t_n} \leq x_n | x_{t_{n-1}}, x_{t_{n-2}}, \ldots, x_{t_1}\} = P\{x_{t_n} \leq x_n | x_{t_{n-1}}\}. \tag{3.117}$$

For discrete-time, continuous-state Markovian processes, in terms of conditional density functions, there are many important properties. For an evolution $x_1, x_2, \ldots, x_t$ from (3.117) it is

$$p(x_t | x_{t-1}, \ldots, x_1) = p(x_t | x_{t-1}). \tag{3.118}$$

For the joint density of $x_1, x_2, \ldots, x_t$ we have by conditioning (chain rule) and (3.118) that

$$p(x_t, \ldots, x_1) = p(x_t | x_{t-1}, \ldots, x_1) p(x_{t-1} | x_{t-2}, \ldots, x_1) \cdots p(x_2 | x_1) p(x_1) \tag{3.119}$$

$$= p(x_t | x_{t-1}) p(x_{t-1} | x_{t-2}) \cdots p(x_2 | x_1) p(x_1). \tag{3.120}$$

For the conditional density $p(x_t | x_{t+1}, \ldots, x_{t+k})$ we have by applying the definition of the conditional density, (3.120) and Bayes rule in line

$$\begin{aligned} p(x_t | x_{t+1}, \ldots, x_{t+k}) &= \frac{p(x_t, x_{t+1}, \ldots, x_{t+k})}{p(x_{t+1}, \ldots, x_{t+k})} = \frac{p(x_{t+1} | x_t) p(x_t)}{p(x_{t+1})} \\ &= p(x_t | x_{t+1}). \end{aligned} \tag{3.121}$$

Thus a Markov process is also Markov if time is reversed. The same properties given here hold also in the case of a Markov random vector process.

Chapter IV

Estimation Theory

The subject of statistical inference, traditionally divided between estimation theory and hypothesis testing, is related to the process of deriving valuable conclusions from observations. In estimation theory, key concept is to obtain plausible values for parameters, called point estimates, or plausible ranges of values, called interval estimates, that describe realities from the observations. The available finite set of measurements may contain errors or noise and so the concept is then related to denoising, filtering and smoothing. In hypothesis testing, starting point is an assumption that specifies values for parameters and then the determination of whether the data are consistent with the hypothetical values is under consideration. Both topics are highly related and it could be said that they belong to the subject of decision theory. However, this thesis is only concentrated on the point estimation side of inference, and especially on estimation methods related to parametric models and mean square error criteria. The main references for this chapter are [184, 145, 161, 169, 108, 23].

4.1 Basic concepts and definitions

If we denote the observations or measurements with z, then what is under concern is the identification of an *estimator* $\hat{\theta}$ for some unknown parameters of interest, i.e. for a parameter vector θ that takes values in some parametric space Θ. A natural requirement is that the estimator is a function of the measurements, i.e.

$$\hat{\theta} = \hat{\theta}(z). \tag{4.1}$$

Every function of the measurements that does not contain the unknown parameters is also called a statistical function. So under consideration is the identification of an estimator, i.e a statistical function, with the desired property that for the observed measurements its value is an estimate close enough to the unknown parameters. Estimation error is then the difference of the actual and estimated values of the parameters, i.e. $\tilde{\theta} = \theta - \hat{\theta}$.

Different estimation theory concepts arise with the characterization of the parameters θ as random or not. If θ is treated as non-random, but still variable, then the measurements, random variables or vectors, can be assumed to have a

joint density function $p_z(z;\theta)$, which depends on the unknown parameters. Since, the unknown variables are assumed non-random this can be also treated as the conditional density of the measurements given the parameters $p(z|\theta)$. The characterization of a parameter as random, i.e. the so called Bayesian assumption, leads to *Bayesian estimation* and inference. Then, the parameters are directly assumed to have a joint density function with the measurements $p_{z,\theta}(z,\theta)$. Whatever the characterization of the parameters is, the estimator being a statistical function of random measurements is always random. Although, other methods do not require the interpretation of neither the measurements nor the parameters as random vectors. Well known such methods are different regression and regularization methods. It must be noted that different approaches can lead to identical estimators, but with different properties and interpretation of the related assumptions.

In general, assessing the quality of an estimate is related to the *estimation error* ([184], p. 87)

$$\tilde{\theta} = \theta - \hat{\theta} = \theta - \hat{\theta}(z). \tag{4.2}$$

Although this is a natural way to measure the quality of an estimate, it cannot be used since $\hat{\theta}(z)$ is a random variable and does not have any specific value, i.e. it depends on the realization z. Also, it is preferable to have a scalar criterion for choosing optimal estimators. For this reason, other estimation criteria can be used. A widely used criterion is the *mean square error* MSE, i.e. the expected value over z of the squared error norm (e.g. [169], p. 300)

$$R_{MS}(\theta,\hat{\theta}) = E\{\|\theta - \hat{\theta}\|^2\} = E\{\tilde{\theta}^T\tilde{\theta}\} = E\{C_{MS}(\theta,\hat{\theta})\}. \tag{4.3}$$

More general, a *cost* (or loss) function

$$C(\theta,\hat{\theta}) = C(\tilde{\theta}) \tag{4.4}$$

can be defined, that describes the cost (or loss) if we estimate θ with the value $\hat{\theta}$. Typical properties required for the cost function are that it is symmetric, i.e. $C(\tilde{\theta}) = C(-\tilde{\theta})$, convex, and, for convenience, that the cost corresponding to zero error is zero ([184], p. 159). The convexity property, i.e. $C(\lambda\tilde{\theta}_1 + (1-\lambda)\tilde{\theta}_2) \leq \lambda C(\tilde{\theta}_1) + (1-\lambda)C(\tilde{\theta}_2), \quad 0 \leq \lambda \leq 1$, guarantees that the loss function increases as the estimation error increases. These properties cover a wide range of cost functions, for example, the quadratic error cost function $C_{MS} = \tilde{\theta}^T\tilde{\theta}$ and the absolute error cost function $C_{ABS} = \sum|\tilde{\theta}_i|$, see for example [184, 23, 145].

In order to obtain an error measure as a performance index or error criterion the expectation of the respective cost function is defined. This is also called the *risk function* of the estimator $\hat{\theta}$ ([184], p. 159)

$$R(\theta,\hat{\theta}) = E\{C(\theta,\hat{\theta})\}, \tag{4.5}$$

in relation to the respective cost function. Clearly, this is a function of the unknown variable θ and takes into account all the possible realizations of z, at least through

some density of the form $p(z;\theta)$. An estimator $\hat{\theta}_1$ is better than an estimator $\hat{\theta}_2$, respective to the risk function R, if

$$R(\theta, \hat{\theta}_1) \leq R(\theta, \hat{\theta}_2), \tag{4.6}$$

for every $\theta \in \Theta$, or $R(\theta_o, \hat{\theta}_1) < R(\theta_o, \hat{\theta}_2)$, for some $\theta_o \in \Theta$. Ideally, an optimal estimator should minimize a risk function for every value of θ. The identification of such an estimator does not always have solution, except for the trivial case that the parameter is constant. For example, consider the MS error criterion and let $\theta_1 \neq \theta_2$ both belonging to the parameter space $\Theta \subset \mathbb{R}$. Let us also assume that exists an estimator θ^* that is optimal for every θ. Then for every other estimator $\hat{\theta} = \hat{\theta}(z)$ it holds $E\{(\theta^* - \theta)^2\} \leq E\{(\hat{\theta} - \theta)^2\}$. So, for $\hat{\theta} = \theta_1$ it is $E\{(\theta^* - \theta_1)^2\} \leq 0$ and yields $\theta^* = \theta_1$. Equivalently, starting with $\hat{\theta} = \theta_2$, yields $\theta^* = \theta_2$, implying that $\theta_1 = \theta_2$, which is different from the assumption.

Different methods exist for solving the problem of non-existence of optimal estimator. One such a way is to restrict the search in a class of estimators that satisfy some intuitively logical condition. One such a condition is the property of unbiasedness. This implies that the expected value of the estimation error should be zero for every value of the parameter θ and measurements z. If the parameters are treated as non random then the condition leads to ([184], p. 88)

$$E\{\hat{\theta}\} = \int \hat{\theta}(z) p(z;\theta) dz = \theta, \tag{4.7}$$

for every $\theta \in \Theta$, and for random parameters to

$$E\{\hat{\theta}\} = \int\int \hat{\theta}(z) p(z,\theta) dz d\theta = \eta_\theta. \tag{4.8}$$

Note, that an estimator satisfying (4.7) is also called absolutely unbiased ([184], p. 88). If an estimator does not meet the unbiasedness conditions, it is said to be biased, and the mean value of the error $b = E\{\tilde{\theta}\}$ is defined as the bias of the estimator. If the bias approaches zero as the number of measurements grows infinitely large, the estimator is called asymptotically unbiased. Another reasonable requirement for a good estimator is that it should converge (in probability) to the true value of the parameter, when the number of measurements grows infinitely large. Estimators satisfying this property are called consistent ([184], p. 92, [169], p. 301).

For non-random parameters θ we have for the mean square error

$$\begin{aligned} E\{(\theta - \hat{\theta})^T (\theta - \hat{\theta})\} &= E\{\theta^T\theta - \theta^T\hat{\theta} - \hat{\theta}^T\theta + \hat{\theta}^T\hat{\theta}\} & (4.9)\\ &= \theta^T\theta - \theta^T E\{\hat{\theta}\} - E\{\hat{\theta}\}^T\theta + E\{\hat{\theta}^T\hat{\theta}\}. & (4.10) \end{aligned}$$

Similarly, it also holds

$$\begin{aligned} E\{\|\hat{\theta} - E\{\hat{\theta}\}\|^2\} &= E\{\hat{\theta}^T\hat{\theta} - \hat{\theta}^T E\{\hat{\theta}\} - E\{\hat{\theta}\}^T\hat{\theta} + E\{\hat{\theta}\}^T E\{\hat{\theta}\}\} & (4.11)\\ &= E\{\hat{\theta}^T\hat{\theta}\} - E\{\hat{\theta}\}^T E\{\hat{\theta}\}. & (4.12) \end{aligned}$$

Thus, from (4.10) and (4.12) we have another expression for (4.3)

$$\begin{aligned} R_{MS}(\tilde{\theta}) &= (\theta - E\{\hat{\theta}\})^T(\theta - E\{\hat{\theta}\}) + E\{(\hat{\theta} - E\{\hat{\theta}\})^T(\hat{\theta} - E\{\hat{\theta}\})\} \quad (4.13) \\ &= \|E\{\tilde{\theta}\}\|^2 + \text{trace}[E\{(\hat{\theta} - E\{\hat{\theta}\})(\hat{\theta} - E\{\hat{\theta}\})^T\}] \quad (4.14) \\ &= \|b\|^2 + \text{trace}[C_{\hat{\theta}}], \quad (4.15) \end{aligned}$$

where $C_{\hat{\theta}}$ is the covariance matrix of the estimator. Therefore, the mean square error criterion or risk function both measures the bias, i.e. the systematic error of the estimator, and a term related to the variance of the estimator, i.e. the accuracy of the estimator. So, with the unbiasedness constraint the MS error criterion searches the estimator with minimal sum of variance components.

Another closely related measure of the quality of an estimator is given by the covariance matrix of the estimation error

$$C_{\tilde{\theta}} = E\{(\theta - \hat{\theta})(\theta - \hat{\theta})^T\}, \quad (4.16)$$

which measures the errors of individual parameter estimates, while the MS error

$$R_{MS}(\tilde{\theta}) = E\{(\theta - \hat{\theta})^T(\theta - \hat{\theta})\} = \text{trace}[C_{\tilde{\theta}}] \quad (4.17)$$

is an overall scalar error measure for all the parameter estimates. An estimator that provides the smallest error covariance matrix among all unbiased estimators is the best one with respect to this quality criterion. Note, that a symmetric matrix A is said to be smaller than another symmetric matrix, if the matrix $B - A$ is positive definite. When the parameters are treated as non-random, such an estimator is called efficient, because it is considered to use optimally the information contained in the measurements. An important theoretical result states that there exists a lower bound for the error covariance of any unbiased estimator for deterministic or random parameters based on available measurements. This is provided by the Cramer-Rao lower bound [184]. A closely related concept for identifying MS optimal estimators is sufficiency. Informally speaking, a statistical function is sufficient if it retains all the information that the data contain. Precise definitions of the concept as well as conditions such as Neyman-Fisher factorization theorem can be found for example in [184]. Sufficiency can be used for the identification from the class of unbiased estimators one that minimizes the MS error for all θ (Rao-Blackwell and Lehmann-Scheffé theorems)[184].

Instead of searching for an estimator that minimizes a risk function $R(\theta, \hat{\theta})$ for every value $\theta \in \Theta$ in some class of estimators, it is convenient to require that $R(\theta, \hat{\theta})$ does not become large for many values (areas) of $\theta \in \Theta$. This can be stated as requiring minimization of the area (surface) $\int_\Theta R(\theta, \hat{\theta})d\theta$, or more general of

$$\int_\Theta R(\theta, \hat{\theta})w(\theta)d\theta, \quad (4.18)$$

where $w(\theta)$ is a weighting function that represents the meaning that is assigned to different values of θ. An estimator that minimizes (4.18) is called Bayesian

estimator. By treating the parameter θ as random and selecting $w(\theta) = p(\theta)$, i.e. a prior density for θ, and the quadratic cost function, we have the Bayesian mean square error criterion [184].

4.2 Estimation with observation model

In various situations, it might be available, or preselected, a model for the dependencies between observations and parameters. The most common observation model is the *additive noise model*

$$z = h(\theta) + \upsilon. \tag{4.19}$$

Where θ is an n-dimensional parameter vector, $z \in \mathbb{R}^M$ is the measurement vector, and $\upsilon \in \mathbb{R}^M$ is a vector whose components are unknown observation errors or measurement noise. The vector valued function $h(\cdot)$ is completely known, or at least its values for different values of θ are known. In this section, the parameters are treated as deterministic unknown quantities.

4.2.1 Ordinary and generalized linear least squares estimators

In the basic linear least squares method the data vector z is assumed to follow the linear additive noise model ([184], p. 33)

$$z = H\theta + \upsilon, \tag{4.20}$$

where H is a deterministic observation matrix which contains the basis vectors $\psi_1, \psi_2, \ldots, \psi_n$ of length M is its columns. The *least squares criterion* is defined as

$$\mathcal{E}_{LS} = \|\upsilon\|^2 = \|z - H\theta\|^2 = (z - H\theta)^T(z - H\theta). \tag{4.21}$$

Minimization of the criterion with respect to θ gives the ordinary least squares estimator. Indeed, the gradient of the quadratic function is

$$\frac{\partial \mathcal{E}_{LS}}{\partial \theta} = -2H^T(z - H\theta). \tag{4.22}$$

Then the least squares estimator satisfies

$$H^T(z - H\hat{\theta}_{LS}) = 0 \iff H^T H\hat{\theta}_{LS} = H^T z. \tag{4.23}$$

The equation in the right is called system of normal equations for determining the least squares estimate $\hat{\theta}_{LS}$ of θ. The matrix $H^T H$ corresponds to $\partial^2 \mathcal{E}_{LS}(\theta)/\partial\theta^2$, and if it is positive definite the quadratic function is strictly convex. Then the solution is guaranteed to be unique given by

$$\hat{\theta}_{LS} = (H^T H)^{-1} H^T z. \tag{4.24}$$

The matrix $H^+ = (H^T H)^{-1} H^T$ is called left pseudo-inverse of H (assuming H has maximal rank n and more rows than columns $M > n$). The estimated fit ($s = H\theta$) is

$$\hat{s}_{LS} = H\hat{\theta}_{LS} = H(H^T H)^{-1} H^T z. \tag{4.25}$$

The $M \times M$ matrix $P = H(H^T H)^{-1} H^T$ is a projection matrix projecting measurements onto the space spanned by the columns (the basis vectors) of the matrix H, and it holds $P^T = P, P^2 = P$, rank(P) = trace$[(H^T H(H^T H)^{-1}] = n$, and $PH = H$. Equation (4.23) implies that the residual $r = z - H\hat{\theta}_{LS}$, or fitting error, is orthogonal to all columns of the matrix H and

$$\begin{aligned} r = \hat{\upsilon} = z - \hat{s} &= (I - H(H^T H)^{-1} H^T) z && (4.26) \\ &= (I - H(H^T H)^{-1} H^T)(H\theta + \upsilon) && (4.27) \\ &= (I - H(H^T H)^{-1} H^T)\upsilon && (4.28) \\ &= H(\theta - \hat{\theta}) + \upsilon. && (4.29) \end{aligned}$$

The least squares estimation problem can be generalized by adding a symmetric and positive definite weighting matrix to the error criterion. The generalized least squares error criterion becomes ([184], p. 40)

$$\begin{aligned} \mathcal{E}_{GLS} &= (z - H\theta)^T W (z - H\theta) && (4.30) \\ &= ||Lz - LH\theta||^2, && (4.31) \end{aligned}$$

where $L^T L$=W. If W is diagonal the index is also called the weighted least squares index. For the minimization of the generalized least squares error criterion, by using equation (4.24) with $z' = Lz$ and $H' = LH$, it holds

$$\begin{aligned} \hat{\theta}_{GLS} &= (H'^T H')^{-1} H'^T z' = (H^T L^T LH)^{-1} H^T L^T Lz && (4.32) \\ &= (H^T W H)^{-1} H^T W z. && (4.33) \end{aligned}$$

The least squares method can be regarded as a deterministic approach for the estimation problem, where no assumptions on any probability distributions are necessary. However, statistical arguments can be used to justify the use of the method. For example, the observation error υ can be considered random ([184], p. 37). Thus the model (4.20) can define a deterministic vector $H\theta$ buried into random noise. If the noise has known mean η_υ and covariance matrix C_υ then the observations are necessarily random with mean η_z and covariance C_z given by

$$\begin{aligned} \eta_z &= H\theta + \eta_\upsilon, && (4.34) \\ C_z &= C_\upsilon. && (4.35) \end{aligned}$$

Then from (4.33, 4.34) the mean of $\hat{\theta}_{GLS}$ is

$$\begin{aligned} \eta_{\hat{\theta}_{GLS}} &= (H^T W H)^{-1} H^T W (H\theta + \eta_\upsilon) && (4.36) \\ &= \theta + (H^T W H)^{-1} H^T W \eta_\upsilon, && (4.37) \end{aligned}$$

thus, unless υ is zero mean, $\hat{\theta}_{GLS}$ is biased. For the unbiased estimator

$$\hat{\theta}_{U,GLS} = (H^T W H)^{-1} H^T W (z - \eta_\upsilon) \quad (4.38)$$

the estimation error covariance, or the covariance of the estimator would be

$$C_{\hat{\theta}_{U,GLS}} = C_{\tilde{\theta}_{U,GLS}} = (H^T W H)^{-1} H^T W C_\upsilon W H (H^T W H)^{-1}. \tag{4.39}$$

It turns out, see following section, that an optimal choice for the matrix W is the inverse of the covariance matrix of the measurements errors $W = C_\upsilon^{-1}$. Note also that $\hat{\theta}_{U,GLS}$ minimizes the least squares index

$$\mathcal{E}_{U,GLS} = ||L(z - z^* - H\theta)||^2, \tag{4.40}$$

where z^* is known and was selected to be η_υ.

4.2.2 Minimum variance linear unbiased estimator or Gauss-Markov estimator

Let us consider the linear observation model, that is

$$z = H\theta + \upsilon, \tag{4.41}$$

where υ is a random vector with known mean η_υ and covariance C_υ, assumed positive definite. The unknown parameters are considered non random. A linear unbiased estimator of θ is searched that minimizes the mean square error criterion (4.3). Because of the unbiasedness requirement, from (4.15) this is equivalent to the minimization of

$$R_{MS}(\tilde{\theta}) = \text{trace}[C_{\hat{\theta}}] \tag{4.42}$$

Since a linear estimator is searched, this must be of the form

$$\hat{\theta} = Kz + k. \tag{4.43}$$

Then it holds

$$E\{\hat{\theta}\} = E\{Kz + k\} = KE\{z\} + k = KE\{H\theta + \upsilon\} + k = KH\theta + K\eta_\upsilon + k. \tag{4.44}$$

For $\hat{\theta}$ to be unbiased, i.e. $E\{\hat{\theta}\} = \theta$ for every θ, K and k must satisfy

$$KH = I, \qquad k = -K\eta_\upsilon. \tag{4.45}$$

For the variance of the estimator it is

$$\begin{aligned}
C_{\hat{\theta}} &= E\{(\hat{\theta} - E\{\hat{\theta}\})(\hat{\theta} - E\{\hat{\theta}\})^T\} && (4.46)\\
&= E\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\} && (4.47)\\
&= E\{(Kz - K\eta_\upsilon - \theta)(Kz - K\eta_\upsilon - \theta)^T\} && (4.48)\\
&= E\{(KH\theta + K\upsilon - K\eta_\upsilon - \theta)(KH\theta + K\upsilon - K\eta_\upsilon - \theta)^T\} && (4.49)\\
&= KC_\upsilon K^T. && (4.50)
\end{aligned}$$

So with the linearity and unbiasedness constrains the covariance matrix of the estimator, or the covariance of the estimation error, is given by (4.50). Now, what

is under concideration is the identification of a matrix K so that the covariance is minimized for every $\theta \in \Theta$. For the matrix

$$K' = (H^T C_v^{-1} H)^{-1} H^T C_v^{-1} \tag{4.51}$$

it holds

$$K'H = I, \qquad \text{and} \qquad K' C_v K'^T = (H^T C_v^{-1} H)^{-1}. \tag{4.52}$$

So with this selection the estimator $\hat{\theta} = K'z$ is unbiased and the covariance does not depend on parameters θ. It must now be shown that this is the best form for K, that is based on the assumptions this is the minimum covariance matrix to be achieved, meaning that the matrix

$$KC_vK^T - K'C_vK'^T = KC_vK^T - (H^T C_v^{-1} H)^{-1} \tag{4.53}$$

is positive semidefinite for every K. It also holds $K'H = I$ and $KH = I$, thus

$$K'C_vK^T = (H^T C_v^{-1} H)^{-1}, \tag{4.54}$$

$$KC_vK'^T = (H^T C_v^{-1} H)^{-1}. \tag{4.55}$$

Thus it holds

$$\begin{aligned} KC_vK^T - K'C_vK'^T &= KC_vK^T - K'C_vK^T && (4.56)\\ &\quad -KC_vK'^T + K'C_vK'^T && (4.57)\\ &= (K-K')C_v(K-K')^T, && (4.58) \end{aligned}$$

where the matrix at the right of the last expression is positive semidefinite ([184], p. 157). Also, for the trace of the error covariance matrix it holds

$$\text{trace}[C_{\tilde{\theta}}] = \text{trace}[(K-K')C_v(K-K')^T] + \text{trace}[(H^T C_v^{-1} H)^{-1}], \tag{4.59}$$

and can be minimized for $K = K'$. The Gauss-Markov estimator can be written in the form

$$\hat{\theta}_{GM} = (H^T C_v^{-1} H)^{-1} H^T C_v^{-1} (z - \eta_v), \tag{4.60}$$

with covariance matrix or error covariance

$$C_{\hat{\theta}_{GM}} = C_{\tilde{\theta}_{GM}} = (H^T C_v^{-1} H)^{-1}, \tag{4.61}$$

and mean square error value

$$R_{MS}(\hat{\theta}_{GM}) = \text{trace}[(H^T C_v^{-1} H)^{-1}]. \tag{4.62}$$

The estimator is often referred as best linear unbiased estimator (BLUE) [184] for non random parameters. The vector v is usually assumed zero mean. In that case and if for the covariance of v it holds $C_v = \sigma^2 I$, i.e. $v(i)$ are zero mean uncorrelated with common variance σ^2, then clearly the estimator coincides with the standard least squares estimator. Furthermore, the GM estimator is exactly minimizing the generalized least squares index (4.30) with the selection $W = C_v^{-1}$.

For the estimator of $s = H\theta$ it holds

$$\hat{s}_{GM} = H\hat{\theta}_{GM}, \tag{4.63}$$

and

$$C_{\hat{s}} = H(H^T C_v^{-1} H)^{-1} H^T. \tag{4.64}$$

The estimator $\hat{s}_{GM} = H\hat{\theta}_{GM}$ is the Gauss Markov estimator of s and its covariance is minimal, among all the unbiased linear estimators of s of the form $H\theta$ when θ is restricted to be a linear function of the data.

4.2.3 Quadratic constraints and regularization

Least squares minimization with a quadratic inequality constraint is an estimation method that can be used in cases that the least squares problem needs to be regularized [23, 69]. Commonly such cases arise when attempting to fit a function to noisy data. The constrained least squares method considered here searches solution of the minimization problem

$$\min_{\theta} \|L_1(H\theta - z)\|^2 \quad \text{subject to} \quad \|L_2(\theta - \theta^*)\|^2 \leq c^2, \tag{4.65}$$

where $L_1^T L_1 = W_1$, $L_2^T L_2 = W_2$, and $c > 0$. The constraint defines an ellipsoid in $\mathbb{R}^n$ centered around an initial (prior) guess for the solution, and is usually chosen to damp out excessive oscillations of the fitting function. This can, for exampler, be done if L_2 is a discretized derivative operator. Another common selection is $L_2 = I$, that constrains the solution in a ball.

The Lagrangian of the problem is

$$L(\theta, \lambda) = (H\theta - z)^T W_1 (H\theta - z) + \lambda \left((\theta - \theta^*)^T W_2 (\theta - \theta^*) - c^2 \right), \tag{4.66}$$

where λ is the Lagrange multiplier associated with the constraint. By equating to zero the gradient of L with respect to θ we obtain a linear system of equations

$$(H^T W_1 H + \lambda W_2)\theta = H^T W_1 z + \lambda W_2 \theta^*, \tag{4.67}$$

with solution $\theta(\lambda)$ as a function of λ that must satisfy the constrain. From the rest of the KKT optimality conditions (see section 2.4) we have that if the constraint is satisfied at the optimal, i.e. $\|L_2(\theta(\lambda) - \theta^*)\|)^2 - c^2 < 0$, then it must be $\lambda = 0$ and the solution of (4.67) is then the generalized least squares estimator (4.33). When, $\lambda > 0$, then $\|L_2(\theta(\lambda) - \theta_o)\|)^2 - c^2 = 0$, i.e. the solution of the problem occurs on the boundary of the feasible set. A more thorough treatment of the problem can be found in [69].

Tikhonov regularization

Often there is not exact knowledge for the value of the side constraint c. Then the problem can be defined from ([23], p. 306)

$$\hat{\theta}_{TR}(\alpha) = \arg\min_{\theta} \{\|L_1(H\theta - z)\|^2 + \alpha^2 \|L_2(\theta - \theta^*\|^2\} = \arg\min_{\theta} \{\mathcal{E}_{TR}(\theta, \alpha)\} \tag{4.68}$$

This is recognized to be the generalized Tikhonov regularization solution [197]. By equating to zero the gradient of $\mathcal{E}_{TR}(\theta, \alpha)$ with respect to θ we obtain the linear system of equations (4.67), where λ is now replaced with α^2. The regularization parameter controls the weight given to the side constraint (i.e. keeping $\|L_2(\theta - \theta^*\|^2$ small) relative to the minimization of the weighted least squares index. Thus α controls the believe on the constraint and the estimator $\hat{\theta}_{TR}(\alpha)$ is a function of α.

The problem can be also written in the form

$$\hat{\theta}_{TR}(\alpha) = \arg\min_{\theta} \left\{ \left\| \begin{pmatrix} L_1 H \\ \alpha L_2 \end{pmatrix} \theta - \begin{pmatrix} L_1 z \\ \alpha L_2 \theta^* \end{pmatrix} \right\|^2 \right\} \tag{4.69}$$

$$= \arg\min_{\theta} \{ \|L'(H'\theta - z')\|^2 \} \tag{4.70}$$

where

$$H' = \begin{pmatrix} H \\ I \end{pmatrix}, \quad z' = \begin{pmatrix} z \\ \theta^* \end{pmatrix}, \quad L' = \begin{pmatrix} L_1 & 0 \\ 0 & aL_2 \end{pmatrix}. \tag{4.71}$$

Then the LS solution for the modified observation model is

$$\hat{\theta}_{TR}(\alpha) = (H'^T L'^T L' H')^{-1} H'^T L'^T L' z' \tag{4.72}$$

$$= (H^T W_1 H + \alpha^2 W_2)^{-1} (H^T W_1 z + \alpha^2 W_2 \theta^*). \tag{4.73}$$

Equivalently, a regularized solution for the least squares problem with k inequality constraints will be

$$\hat{\theta}_{TR}(\alpha_1, \alpha_2, \ldots, \alpha_k) = (H^T W H + \sum_{i=1}^{k} \alpha_i^2 W_i)^{-1} (H^T W z + \sum_{i=1}^{k} \alpha_i^2 W_i \theta_i^*) \tag{4.74}$$

The simplest form of regularization solution is given for $W_1 = I$, $W_2 = I$ and $\theta^* = 0$, then the estimator becomes

$$\hat{\theta} = (H^T H + \alpha^2 I)^{-1} H^T z. \tag{4.75}$$

The matrix $H^T H + \alpha^2 I$ is positive definite for every $\alpha^2 > 0$ ([23], p. 306). This can be seen from the singular value decomposition of H (e.g.[39], p. 58). Therefore, the regularized solution requires no rank assumptions on H.

Regularization is used in several contexts. It must be must noted that the quadratic norms used here is not the only case, see for example [23]. In an estimation setting, the extra term penalizing large $\|\theta\|$ can be interpreted as our prior knowledge requires $\|\theta\|$ not to be too large. In statistical literature, Tikhonov regularization is also called ridge regression [69, 142], though there is also the Bayesian interpretation [184, 108]. The constraint $\|\theta\|$ to be small can also reflect a modeling issue. It might, for example, be that $s = H\theta$ is only a good approximation of the true relationship $s = h(\theta)$ between s, θ. In order to have $h(\theta) \approx z$, it must $H\theta \approx z$, and also need θ small to ensure $h(\theta) \approx H\theta$ ([23], p. 306). Furthermore, regularization is also used in the case, for example, that the matrix H is square, and the goal is to solve the linear equations $H\theta = z$. In cases where H is poorly conditioned, or singular, regularization gives a compromise between solving the equations and keeping θ of reasonable size. Therefore, it have gained great popularity for the study of inverse and ill-conditioned problems [108, 74, 73, 149].

Robust least-squares

We consider an approximation problem with basic objective $\|z - H\theta\|^2$, but now for the matrix H there is some degree of uncertainty or variation. For more general settings see for example [23, 69, 65]. More specific we consider the statistical robust least square criterion ([23], p. 318)

$$E\{\|z - H'\theta\|^2\}, \tag{4.76}$$

where the parameters θ are still considered as non random. The only random term is the matrix H', which is assumed to be of the form $H' = H + U$, where H is a deterministic part and U is a random matrix with zero mean that describes the statistical variation. Then (4.76) becomes

$$E\{(z - H\theta - U\theta)^T(z - H\theta - U\theta)\} = (z - H\theta)^T(z - H\theta) + \theta^T E\{U^T U\}\theta. \tag{4.77}$$

Therefore the statistical robust approximation problem has the form of a regularized least squares problem with solution (by setting $P = E\{U^T U\}$)

$$\hat{\theta}_r = (H^T H + P)^{-1} H^T z. \tag{4.78}$$

If the observation matrix is subject to uncertainty, the vector $H'\theta$ will have more variation the larger θ is. This can be controlled with regularization ([23], p. 319).

Smoothing regularization

The idea of regularization can be extended in several directions. A useful extension is obtained when the regularization term is of the form $\|D_d\theta\|^2$, where D_d is a discrete approximation of the d-th order derivative. Methods using difference approximations can in general be called *smoothness priors* methods [122, 108]. In this case, L_2 is a banded matrix with full row rank. The most commonly used matrices are the 1st and 2nd difference matrices

$$D_1 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix} \in \mathbb{R}^{(n-1)\times n}, \tag{4.79}$$

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{(n-2)\times n}. \tag{4.80}$$

Then the problem becomes

$$\min_{\theta}(\|H\theta - z\|^2 + a^2\|D_d\theta\|^2), \tag{4.81}$$

which represents a trade-off between the measure of fit and the regularization term. For example, $\|D_2\theta\|^2$ is an approximation of the mean-square curvature of the parameter. Thus the parameter α^2 provides compromise between good fit and smoothness ([23], p. 307). There are situations that the parametrization of the problem does not give meaning to the smoothing problem (4.81). Instead, prior knowledge about the smoothness of $H\theta$ is to be enforced directly. Then the problem becomes

$$\min_{\theta}(\|H\theta - z\|^2 + \alpha^2\|D_d H\theta\|^2), \tag{4.82}$$

with the modified side constraint $\|D_2 H\theta\|^2$ and with solution

$$\hat{\theta}(\alpha) = (H^T H + \alpha^2 H^T D_d^T D_d H)^{-1} H^T z. \tag{4.83}$$

In the simplest case, when $H = I$, we have the model

$$z = s + v, \tag{4.84}$$

which represents a deterministic signal buried in noise. The noise is simply assumed to be unknown, small, and unlike the signal rapidly varying. The process of recovering s from z is called signal reconstruction or denoising. Often the operation of recovering s is smoothing. A simple reconstruction can be performed by using the quadratic smoothing penalty function ([23], p. 312)

$$\|D_1 s\|^2 = \sum_{i=2}^{M}(s_i - s_{i-1})^2. \tag{4.85}$$

Which implies that the signal s is assumed slowly varying. The estimator of s is given for different values of α^2 by

$$\hat{s} = (I + a^2 D_1^T D_1)^{-1} z = Gz. \tag{4.86}$$

The operator G is consider as a basic smoothing operator in this thesis.

4.2.4 Nonlinear least squares

Let us consider now the nonlinear additive noise observation model

$$z = h(\theta) + v, \tag{4.87}$$

where $h(\theta) = (h_1(\theta), h_2(\theta), \ldots, h_M(\theta))^T$ is a vector valued function of the parameters θ. The generalized non-linear least squares estimator with quadratic inequality constraint for the parameters is the solution of the minimization of the functional

$$\begin{aligned} \mathcal{E}(\theta) &= \|L_1(h(\theta) - z)\|^2 + \alpha^2\|L_2(\theta - \theta^*)\|^2 && (4.88)\\ &= (z - h(\theta))^T W_1 (z - h(\theta)) + \alpha^2(\theta - \theta^*)^T W_2(\theta - \theta^*). && (4.89) \end{aligned}$$

For a non-linear function $h(\theta)$ the minimization has to be done iteratively, for example with Gauss-Newton method. We consider here only the unconstrained problem. The Gauss-Newton method (2.72) then becomes ([184], p. 247)

$$\hat{\theta}^{i+1} = \hat{\theta}^i + a_i(J_i^T W J_i)^{-1}(J_i^T W(z - h(\hat{\theta}^i))), \tag{4.90}$$

where J_i is the Jacobian matrix of h evaluated at $\hat{\theta}^i$, and the step size $a_i > 0$ can, for example, be found with backtracking line search (2.49). In order to be a descent procedure $J^T W J$ must be positive definite. The Gauss-Newton method is approximating the Hessian by $J^T W J$, while the Levenberg-Marquardt enhancement [132, 141] adds to the approximation a positive definite matrix, in order to control ill conditioning. In the Levenberg-Marquardt algorithm, usually the line search strategy is replaced with a trust region strategy (e.g. [21]).

4.3 Maximum likelihood estimation

The method of maximum likelihood constitutes a general way for deriving estimators for an unknown parameter vector θ ([169], p. 305). The method is due to Fisher and considers the unknown θ as non-random. Is is based upon a principle stating that as soon as a realization of a random vector z is observed, then as an estimate of θ is chosen the value $\hat{\theta}_{ML}$ that maximizes for every $\theta \in \Theta$ the likelihood function (*maximum likelihood principle*). If z has a probability density function $p_z(z;\theta), \theta \in \Theta$, which is a parametric function of the parameters to be estimated, then the likelihood function is defined as

$$L(\theta) = p(z;\theta), \quad \theta \in \Theta, \tag{4.91}$$

that is the density evaluated at the observed value of z and considered as a function of θ. Thus, the maximum likelihood estimator is defined as

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} p(z;\theta). \tag{4.92}$$

Since many density functions contain an exponential function, it is often more convenient to deal with the log likelihood. The ML estimator is then found from the solution of the log likelihood equation

$$\frac{\partial}{\partial \theta} \ln p(z;\theta) = 0, \tag{4.93}$$

in the case that it is differentiable for $\theta \in \Theta$. The maximum might be obtained for an unique value $\hat{\theta}_{ML}$ or for many or it might not exist.

Let us now consider the observation model

$$z = h(\theta) + \upsilon, \tag{4.94}$$

where the random vector υ is considered zero mean with known density function p_υ. Thus from the model we have that the density of z is the same as that of υ

except of the mean, i.e $E\{z\} = h(\theta) + E\{\upsilon\} = h(\theta)$. For the likelihood of the observed z, given the model, we have

$$p(z;\theta) = p(h(\theta) + \upsilon;\theta) = p_\upsilon(z - h(\theta);\theta) \tag{4.95}$$

If we assume that the density of the random noise vector υ is Gaussian then

$$p(z;\theta) = \frac{1}{(2\pi)^{n/2}(\det C_\upsilon)^{1/2}} \exp(-\frac{1}{2}(z - h(\theta))^T C_\upsilon^{-1}(z - h(\theta))), \tag{4.96}$$

and by taking logarithms

$$\ln p(z;\theta) = \text{const} - \frac{1}{2}(z - h(\theta))^T C_\upsilon^{-1}(z - h(\theta)) \tag{4.97}$$

Thus, because of the quadratic form of the density, maximization of the likelihood is equivalent to the minimization of the functional

$$l_{ML} = (z - h(\theta))^T C_\upsilon^{-1}(z - h(\theta)). \tag{4.98}$$

This is identical to the generalized least squares index with the selection $W = C_\upsilon^{-1}$. In the linear case the estimator is

$$\hat{\theta}_{ML} = (H^T C_\upsilon^{-1} H)^{-1} H^T C_\upsilon^{-1} z, \tag{4.99}$$

which is exactly the Gauss Markov estimator (4.60), i.e. a minimum variance unbiased estimator for the non random parameters θ.

4.4 Bayesian estimation

The starting point in Bayesian estimation is the consideration of the unknown parameters θ as random ([108], p. 50). Then a joint density between parameters and measurements is assumed $p(z, \theta)$. Estimation is then based on the *posterior* density $p(\theta|z)$, which from the Bayes rule is given by

$$p_{\theta|z}(\theta|z) = \frac{p_{z|\theta}(z|\theta)p_\theta(\theta)}{p_z(z)}, \tag{4.100}$$

where the denominator is computed based on the law of total probability by integrating the numerator over all the possible values of θ ([108], p. 52)

$$p_z(z) = \int_{-\infty}^{\infty} p_{z|\theta}(z|\theta)p_\theta(\theta)d\theta. \tag{4.101}$$

The density $p(z|\theta)$ is recognized as the likelihood function defined earlier. The *prior* distribution $p(\theta)$ is assumed known. Although the interpretation of the parameters as random in many applications is rather technical, in Bayesian inference the probability functions for the parameters can be interpreted as degrees of belief

related to the values that the parameters can take, and they are the result of prior knowledge or consideration [22, 184, 108].

Note that in Bayesian methodology, if inference is required for the random parameters θ with prior density $p(\theta)$, it is considered that after observing z the knowledge that we have about θ has been naturally updated to the density $p(\theta|z)$. This is an old observation related to the meaning of inverse probability of Laplace and Bayes study on the doctrine of chances. Estimates for θ could then be searched from the posterior density, which is a density for θ, with reduced uncertainty because of the observations. In this thesis, we concentrate mainly to the Bayesian cost methodology and discuss some useful properties of optimality of the derived point estimators. However, a point estimator is not enough for a fully Bayesian inference since the inference should be based on the whole posterior density [108].

4.4.1 Bayes cost method

In Bayesian estimation, the selection of point estimates can be done by defining a cost function $C(\theta, \hat{\theta})$, which assigns an unique real valued cost to each combination of actual parameter value and estimate ([145], p. 182). The expected value of the cost is called Bayes cost or Bayes risk function and is given by ([145], p. 183)

$$\begin{aligned} B(\hat{\theta}) = E\{C(\theta, \hat{\theta}(z))\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(\theta, \hat{\theta}(z))p(\theta, z)dzd\theta && (4.102)\\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} C(\theta, \hat{\theta}(z))p(z|\theta)dz\right)p(\theta)d\theta && (4.103)\\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} C(\theta, \hat{\theta}(z))p(\theta|z)d\theta\right)p(z)dz. && (4.104) \end{aligned}$$

Since from the Bayesian assumption it holds for the joint density $p(\theta, z) = p(z|\theta)p(\theta)$ deriving (4.103), and $p(\theta, z) = p(\theta|z)p(z)$ giving (4.104). According to the Bayesian estimation criterion the optimal estimator $\hat{\theta}_B$ is the one that minimizes the Bayes cost for a given cost function, i.e.

$$B(\hat{\theta}_B) \leq B(\hat{\theta}) \tag{4.105}$$

for all $\hat{\theta}$. In general, the identification of the joint density is needed for deriving Bayesian estimators.

The inner integral in (4.103) is the conditional expectation of the cost given θ.

$$B(\hat{\theta}|\theta) = \int_{-\infty}^{\infty} C(\theta, \hat{\theta}(z))p(z|\theta)dz = E\{C(\theta, \hat{\theta})|\theta\}. \tag{4.106}$$

This is called the conditional Bayes cost ([145], p. 183), which for non random parameters is equal to the risk function $R(\theta, \hat{\theta})$ defined in (4.5). For the Bayes cost it holds

$$B(\hat{\theta}) = \int_{-\infty}^{\infty} B(\hat{\theta}|\theta)p(\theta)d\theta = E_\theta\{E\{C(\theta, \hat{\theta})|\theta\}\} = E_\theta\{B(\hat{\theta}|\theta)\}. \tag{4.107}$$

Similarly the conditional cost given z is

$$B(\hat{\theta}|z) = \int_{-\infty}^{\infty} C(\theta, \hat{\theta}(z)) p(\theta|z) d\theta = E\{C(\theta, \hat{\theta})|z\}, \tag{4.108}$$

and for the Bayes cost it holds

$$B(\hat{\theta}) = E_z\{E\{C(\theta, \hat{\theta})|z\}\} = E_z\{B(\hat{\theta}|z)\}. \tag{4.109}$$

When this form is applied, the minimization of the Bayes cost is equivalent to the minimization of the conditional cost given z since the outer integral does not depend on θ and $p(z)$ is a non-negative integrable function with $\int p(z)dz = 1$.

4.4.2 Bayesian mean square estimation

In mean square error based estimation, the cost function is the squared norm of the estimation error $C_{MS}(\theta, \hat{\theta}) = (\theta - \hat{\theta})^T(\theta - \hat{\theta})$. Then the Bayes mean square cost is defined as

$$\begin{aligned} B_{MS}(\hat{\theta}) &= E\{(\theta - \hat{\theta})^T(\theta - \hat{\theta})\} && (4.110)\\ &= \text{trace}[E\{(\theta - \hat{\theta})(\theta - \hat{\theta})^T\}] && (4.111)\\ &= \text{trace}[R_{\tilde{\theta}}], && (4.112) \end{aligned}$$

and for the error correlation matrix it holds

$$\begin{aligned} R_{\tilde{\theta}} &= E\{(\theta - \hat{\theta})(\theta - \hat{\theta})^T\} && (4.113)\\ &= E_z\{E\{(\theta - \hat{\theta})(\theta - \hat{\theta})^T|z\}\} && (4.114)\\ &= E_z\{E\{\theta\theta^T - \theta\hat{\theta}^T - \hat{\theta}\theta^T + \hat{\theta}\hat{\theta}^T|z\}\} && (4.115)\\ &= E_z\{E\{\theta\theta^T|z\} - E\{\theta|z\}\hat{\theta}^T - \hat{\theta}E\{\theta|z\}^T + \hat{\theta}\hat{\theta}^T\} && (4.116)\\ &= E_z\{C_{\theta|z} + \eta_{\theta|z}\eta_{\theta|z}^T - E\{\theta|z\}\hat{\theta}^T - \hat{\theta}E\{\theta|z\}^T + \hat{\theta}\hat{\theta}^T\} && (4.117)\\ &= E_z\{C_{\theta|z} + (\hat{\theta} - \eta_{\theta|z})(\hat{\theta} - \eta_{\theta|z})^T\} && (4.118)\\ &= E_z\{C_{\theta|z}\} + E_z\{(\hat{\theta} - \eta_{\theta|z})(\hat{\theta} - \eta_{\theta|z})^T\}. && (4.119) \end{aligned}$$

The Bayes cost can be written in the form

$$B_{MS}(\hat{\theta}) = E\{C_{MS}(\theta, \hat{\theta})\} = E_z\{B_{MS}(\hat{\theta}|z)\}, \tag{4.120}$$

where $B_{MS}(\hat{\theta}|z)$ is the conditional Bayes mean square cost ([184], p. 137)

$$\begin{aligned} B_{MS}(\hat{\theta}|z) &= E\{(\theta - \hat{\theta})^T(\theta - \hat{\theta})|z\} && (4.121)\\ &= \text{trace}[C_{\theta|z} + (\hat{\theta} - \eta_{\theta|z})(\hat{\theta} - \eta_{\theta|z})^T] && (4.122)\\ &= \text{trace}[C_{\theta|z}] + \|\hat{\theta} - \eta_{\theta|z}\|^2. && (4.123) \end{aligned}$$

The first term in (4.123) does not depend on $\hat{\theta}(z)$ and is positive. The second term, also positive, can be made zero by choosing $\hat{\theta} = \eta_{\theta|z}$. Therefore, the Bayesian

minimum mean square estimator, that minimizes the conditional Bayes cost and the Bayes cost is the function $\eta_{\theta|z}$, is the conditional mean ([184], p. 137)

$$\hat{\theta}_{MS} = \int_{-\infty}^{\infty} \theta p(\theta|z) d\theta = E\{\theta|z\} = \eta_{\theta|z}. \tag{4.124}$$

The result holds for all conditional densities $p(\theta|z)$ and $\hat{\theta}_{MS}$ is an uniquely defined optimal estimator, which is also called conditional mean estimator ([145], p. 184).

Let $\hat{\theta} = \hat{\theta}(z)$ be an arbitrary estimator of the random parameter vector θ, then for the expected value of the estimation error it holds

$$E\{(\theta - \hat{\theta})\} = E_z\{E\{(\theta - \hat{\theta})|z\}\} = E_z\{(E\{\theta|z\} - \hat{\theta})\} = E_z\{(\eta_{\theta|z} - \hat{\theta})\}. \tag{4.125}$$

Clearly, the Bayesian mean square estimator $\hat{\theta}_{MS} = \eta_{\theta|z}$ is unbiased and $E\{\hat{\theta}_{MS}\} = \eta_\theta$ or $E\{\tilde{\theta}_{MS}\} = 0$. For the estimation error covariance of any estimator $\hat{\theta}$ it holds from (4.119) and (4.125)

$$\begin{aligned} C_{\tilde{\theta}} &= E\{(\theta - \hat{\theta})(\theta - \hat{\theta})^T\} - E\{(\theta - \hat{\theta})\}E\{(\theta - \hat{\theta})\}^T && (4.126)\\ &= E_z\{C_{\theta|z}\} + E_z\{(\hat{\theta} - \eta_{\theta|z})(\hat{\theta} - \eta_{\theta|z})^T\} && (4.127)\\ &\quad -E_z\{(\eta_{\theta|z} - \hat{\theta})\}E_z\{(\eta_{\theta|z} - \hat{\theta})\}^T. && (4.128) \end{aligned}$$

The unbiased conditional mean estimator has estimation error covariance

$$C_{\tilde{\theta}_{MS}} = E\{(\theta - \eta_{\theta|z})(\theta - \eta_{\theta|z})^T\} = E_z\{C_{\theta|z}\}. \tag{4.129}$$

Additionally, it holds

$$C_{\tilde{\theta}} - C_{\tilde{\theta}_{MS}} = E_z\{[(\hat{\theta} - \eta_{\theta|z}) - E_z\{(\eta_{\theta|z} - \hat{\theta})\}][(\hat{\theta} - \eta_{\theta|z}) - E_z\{(\eta_{\theta|z} - \hat{\theta})\}]^T\}. \tag{4.130}$$

Thus, by considering the unknown parameters θ as random and by searching an estimator that has the smallest error covariance matrix, since the matrices in (4.130) are positive semidefinite, then the optimal estimator is identical to the Bayesian mean square estimator. So the Bayesian mean square estimator can be defined either as the estimator that minimizes the conditional Bayes cost given z, and therefore the Bayes cost, or as minimum error variance estimator for random parameters ([145], p. 185, [184], p. 140).

The previous results can be applied to a generalized mean square cost function

$$C_{GMS}(\theta, \hat{\theta}) = \tilde{\theta}^T W \tilde{\theta}, \tag{4.131}$$

where W is a symmetric positive semidefinite weighting matrix. The conditional Bayes cost for this cost function is then

$$\begin{aligned} B_{GMS}(\hat{\theta}|z) &= E\{(\theta - \hat{\theta})^T W (\theta - \hat{\theta})|z\} && (4.132)\\ &= E\{(\theta^T W \theta - \theta^T W \hat{\theta} - \hat{\theta}^T W \theta + \hat{\theta}^T W \hat{\theta})|z\} && (4.133)\\ &= (\hat{\theta} - \eta_{\theta|z})^T W (\hat{\theta} - \eta_{\theta|z}) + E\{\theta^T W \theta|z\} - \eta_{\theta|z}^T W \eta_{\theta|z}. && (4.134) \end{aligned}$$

This is minimized again by the conditional mean ([184], p. 137)

$$\hat{\theta}_{GMS} = \eta_{\theta|z}. \tag{4.135}$$

Note, that the mean square cost function can be written as

$$C_{MS}(\theta, \hat{\theta}) = \tilde{\theta}^T \tilde{\theta} = \sum_i \tilde{\theta}^T W_i \tilde{\theta}, \tag{4.136}$$

Where W_i is chosen such as only its (i,i)-th diagonal element is nonzero and equal to one. Thus the the conditional mean minimizes each squared error term $\tilde{\theta}_i^2$ individually ([184], p. 138).

Another important property is the orthogonality principle. Let $\xi = \xi(z)$ any function of the data z only. Then for the cross correlation of the estimation error and ξ we have

$$\begin{aligned} E_z\{(\theta - \hat{\theta}_{MS})\xi^T\} &= E_z\{E\{(\theta - \hat{\theta}_{MS})\xi^T|z\}\} && (4.137)\\ &= E_z\{E\{\theta - \hat{\theta}_{MS}|z\}\xi^T\} && (4.138)\\ &= E_z\{(E\{\theta|z\} - E\{\hat{\theta}_{MS}|z\})\xi^T\} && (4.139)\\ &= E_z\{(\hat{\theta}_{MS} - \hat{\theta}_{MS})\xi^T\} && (4.140)\\ &= 0. && (4.141) \end{aligned}$$

This indicates that the error in the minimum mean square estimator is orthogonal to any function of the data ([184], p. 139).

The Bayesian mean square estimator is theoretically very significant because of its conceptual simplicity and generality. However, actual computation of the Bayesian minimum mean square estimator is often very difficult. This is because in practice we only know or assume the prior density $p_\theta(\theta)$ and the conditional density of the observations $p_{z|\theta}(z|\theta)$ given the parameters (the likelihood). In order to obtain the conditional mean estimator the posterior density is required, which is obtained from Bayes rule. The computation of the conditional expectation then requires still another integration. These integrals are usually impossible to be evaluated analytically. Nevertheless, Monte-Carlo methods are often applied for the calculation of the mean square estimator [112, 108].

4.4.3 Linear Bayesian mean square estimators

The conditional mean $\eta_{\theta|z}$ is in general a nonlinear function of z and to be determined it requires the computation of the integral (4.124). However, there are special cases that it has a specific form. In the case that is linear, then from (3.33) the (linear) mean square Bayesian estimator is given by

$$\hat{\theta}_{LMS} = \eta_\theta + C_{\theta z} C_z^{-1}(z - \eta_z), \tag{4.142}$$

with error covariance

$$C_{\tilde{\theta}_{LMS}} = C_\theta - C_{\theta z} C_z^{-1} C_{z\theta}. \tag{4.143}$$

In a different way, we can restrict the search for optimal estimator for θ in the class of linear estimators of the form $\hat{\theta} = Kz$ by requiring that the optimal estimator should minimize the Bayes cost $B\{\hat{\theta}\} = \text{trace}[C_{\tilde{\theta}}]$. Then for the error covariance it holds (for simplicity z and θ are considered zero mean)

$$\begin{aligned} C_{\tilde{\theta}} &= E\{(\theta - \hat{\theta})(\theta - \hat{\theta})^T\} && (4.144)\\ &= E\{(\theta - Kz)(\theta - Kz)^T\} && (4.145)\\ &= C_\theta - KC_{z\theta} - C_{\theta z}K^T + KC_zK^T && (4.146)\\ &= C_\theta + (K - C_{\theta z}C_z^{-1})C_z(K - C_{\theta z}C_z^{-1})^T - C_{\theta z}C_z^{-1}C_{z\theta}. && (4.147) \end{aligned}$$

Only the second term on the right of the last equation depends on the matrix K, thus the trace and each term of the diagonal of the error covariance matrix can be minimized by choosing $K = C_{\theta z}C_z^{-1}$ ([184], p. 153).

Thus, we have two interpretations for the linear mean square estimator. It is a minimum error variance (covariance) estimator for random parameters optimal among other linear estimators and orthogonal to them. Additionally, it is overall optimal under the same criteria for the densities $p(\theta, z)$ that the conditional mean $\eta_{\theta|z}$ is a linear function of z. Such an example is the Gaussian conditional density. In fact, if z and θ are jointly Gaussian distributed, then the conditional density of θ given z is also Gaussian with the form (see section 3.7.2)

$$p_{\theta|z}(\theta) = \frac{1}{(2\pi)^{n/2}(\det C_{\theta|z})^{1/2}} \exp(-\frac{1}{2}(\theta - \eta_{\theta|z})^T C_{\theta|z}^{-1}(\theta - \eta_{\theta|z})), \qquad (4.148)$$

where

$$\begin{aligned} \eta_{\theta|z} &= \eta_\theta + C_{\theta z}C_z^{-1}(z - \eta_z) && (4.149)\\ C_{\theta|z} &= C_\theta - C_{\theta z}C_z^{-1}C_{z\theta}. && (4.150) \end{aligned}$$

This again underlines the fact that for the Gaussian distribution linear processing and knowledge of the first and second order statistical properties are usually sufficient to obtain optimal results (see for example [184], p. 149).

In many practical estimation problems, it is difficult to determine the cross covariance $C_{\theta z}$. Let us now constrain the observations to be a specific linear form of the parameters, i.e.

$$z = H\theta + \upsilon, \qquad (4.151)$$

where H is a deterministic observation matrix and θ, υ are random vectors with known means η_θ, η_υ and covariances C_θ, C_υ. Then, for the covariance matrix of the measurements we have

$$\begin{aligned} C_z &= E\{(z - \eta_z)(z - \eta_z)^T\} && (4.152)\\ &= E\{(H\theta + \upsilon)(H\theta + \upsilon)^T\} - (H\eta_\theta + \eta_\upsilon)(H\eta_\theta + \eta_\upsilon)^T && (4.153)\\ &= HC_\theta H^T + HC_{\theta\upsilon} + C_{\upsilon\theta}H^T + C_\upsilon. && (4.154) \end{aligned}$$

For the cross covariance $C_{\theta z}$ it holds

$$\begin{aligned} C_{\theta z} &= E\{\theta(H\theta + \upsilon)^T\} - \eta_\theta(H\eta_\theta + \eta_\upsilon)^T & (4.155)\\ &= C_\theta H^T + C_{\theta\upsilon}. & (4.156)\end{aligned}$$

The mean square estimator with linear observation model then becomes

$$\hat{\theta}_{LMS} = \eta_\theta + (C_\theta H^T + C_{\theta\upsilon})(HC_\theta H^T + HC_{\theta\upsilon} + C_{\upsilon\theta}H^T + C_\upsilon)^{-1}(z - H\eta_\theta - \eta_\upsilon), \quad (4.157)$$

with error covariance

$$C_{\tilde{\theta}_{LMS}} = C_\theta - (C_\theta H^T + C_{\theta\upsilon})(HC_\theta H^T + HC_{\theta\upsilon} + C_{\upsilon\theta}H^T + C_\upsilon)^{-1}(HC_\theta + C_{\upsilon\theta}). \quad (4.158)$$

A special case of interest is when θ and υ are uncorrelated, i.e. $C_{\theta\upsilon} = 0$. Then the previous equations become

$$C_z = HC_\theta H^T + C_\upsilon, \qquad C_{\theta z} = C_\theta H^T, \qquad C_{z\theta} = HC_\theta. \quad (4.159)$$

And the estimator is

$$\begin{aligned} \hat{\theta}_{LMS} &= \eta_\theta + C_\theta H^T(HC_\theta H^T + C_\upsilon)^{-1}(z - H\eta_\theta - \eta_\upsilon), & (4.160)\\ C_{\tilde{\theta}_{LMS}} &= C_\theta - C_\theta H^T(HC_\theta H^T + C_\upsilon)^{-1}HC_\theta. & (4.161)\end{aligned}$$

Yet another useful form is obtained by applying the matrix inversion lemma (3.89-3.91). First we have

$$\begin{aligned} C_{11} &= (C_\theta - C_\theta H^T(HC_\theta H^T + C_\upsilon)^{-1}HC_\theta)^{-1} = C_\theta^{-1} + H^T C_{22} H, & (4.162)\\ C_{22} &= (HC_\theta H^T + C_\upsilon - HC_\theta C_\theta^{-1} C_\theta H^T)^{-1} = C_\upsilon^{-1}. & (4.163)\end{aligned}$$

Thus it is

$$C_\theta - C_\theta H^T(HC_\theta H^T + C_\upsilon)^{-1}HC_\theta = (C_\theta^{-1} + H^T C_\upsilon^{-1} H)^{-1}. \quad (4.164)$$

Also we have

$$-C_{11}C_\theta H^T(HC_\theta H^T + C_\upsilon)^{-1} = -C_\theta^{-1}C_\theta H^T C_{22} = -H^T C_\upsilon^{-1}, \quad (4.165)$$

thus

$$C_\theta H^T(HC_\theta H^T + C_\upsilon)^{-1} = (C_\theta^{-1} + H^T C_\upsilon^{-1} H)^{-1} H^T C_\upsilon^{-1}. \quad (4.166)$$

For the estimator we have

$$\begin{aligned} \hat{\theta}_{LMS} &= C_\theta H^T(HC_\theta H^T + C_\upsilon)^{-1}(z - \eta_\upsilon)\\ &+ C_\theta C_\theta^{-1}\eta_\theta - C_\theta H^T(HC_\theta H^T + C_\upsilon)^{-1}HC_\theta C_\theta^{-1}\eta_\theta \end{aligned}$$

yielding finally (see for example [184], p. 155)

$$\begin{aligned} \hat{\theta}_{LMS} &= (C_\theta^{-1} + H^T C_\upsilon^{-1} H)^{-1}(H^T C_\upsilon^{-1}(z - \eta_\upsilon) + C_\theta^{-1}\eta_\theta) & (4.167)\\ C_{\tilde{\theta}_{LMS}} &= (C_\theta^{-1} + H^T C_\upsilon^{-1} H)^{-1}. & (4.168)\end{aligned}$$

This form can now be compared with the fully deterministic case of Tikhonov regularization (4.73) , and gives a Bayesian justification for the method. Furthermore, it shows how the constraints should be chosen (if we assume randomness), so that the variance of the estimation error is minimized. If we compare with the Gauss-Markov estimator (4.60) (optimal estimator for deterministic parameters), we can see that the prior density (first and second order statistics) serves as a base for optimal regularization. The absolutely unbiased optimal estimator (GM) is obtained by letting $C_\theta^{-1} = 0$, which is more convenient than assuming an infinite covariance matrix. This corresponds to the case that no prior information about θ is available ([184], p. 157).

Let us now consider the linear model

$$z = s + \upsilon. \tag{4.169}$$

Then, the Bayesian LMS estimator for s is (4.160)

$$\hat{s}_{LMS} = \eta_s + C_s(C_s + C_\upsilon)^{-1}(z - \eta_s - \eta_\upsilon). \tag{4.170}$$

If we assume that $s = H\theta$ then $\eta_s = H\eta_\theta$, $C_s = HC_\theta H^T$, and substitution gives

$$\hat{s}_{LMS} = H\hat{\theta}_{LMS}. \tag{4.171}$$

4.4.4 Maximum a posteriori estimation

The maximum a posteriori estimator is defined through another cost function, namely the uniform cost function ([145], p. 188)

$$C_{UC}(\theta, \hat{\theta}(z)) = \begin{cases} 0 & \text{if} \quad |\tilde{\theta}_k| < \epsilon, \quad \text{for every } k \\ 1 & \text{otherwise} \end{cases}, \tag{4.172}$$

where ϵ is a small constant. Thus, this cost function assigns zero cost if all the components of the n-dimensional estimation error vector $\tilde{\theta}$ are small, and unit penalty if any of the components is larger than ϵ; and this for every θ and measurement vector z. If we denote with $I = (-\epsilon, \epsilon) \times \ldots \times (-\epsilon, \epsilon) \subset \Theta$ the zero cost region, $\bar{I}$ the unit cost (penalty) region, i.e. the complement of I, then the Bayes conditional cost given z associated with the uniform cost function is

$$\begin{aligned} B_{UC}(\hat{\theta}|z) &= \int_\Theta C_{UC}(\theta, \hat{\theta}) p(\theta|z) d\theta && (4.173) \\ &= \int_{\tilde{\theta} \in \bar{I}} 1 \cdot p(\theta|z) d\theta && (4.174) \\ &= 1 - \int_{\tilde{\theta} \in I} p(\theta|z) d\theta && (4.175) \\ &= 1 - \int_{\hat{\theta}_1 - \epsilon}^{\hat{\theta}_1 + \epsilon} \ldots \int_{\hat{\theta}_n - \epsilon}^{\hat{\theta}_n + \epsilon} p(\theta_1, \ldots, \theta_n|z) d\theta_1 \ldots d\theta_n. && (4.176) \end{aligned}$$

Based on the mean value theorem for integrals, for the continuous function $p(\theta|z)$ there is a value $\dot{\theta}$, such as $\hat{\theta}_k - \epsilon \leq \dot{\theta}_k \leq \hat{\theta}_k + \epsilon$ for every k, for which it holds

$$\begin{aligned} B_{UC}(\hat{\theta}|z) &= 1 - p(\dot{\theta}|z) \int_{\hat{\theta}_1-\epsilon}^{\hat{\theta}_1+\epsilon} \dots \int_{\hat{\theta}_n-\epsilon}^{\hat{\theta}_n+\epsilon} 1 \quad d\theta_1 \dots d\theta_n && (4.177) \\ &= 1 - p(\dot{\theta}|z) \prod_{k=1}^{n} \int_{\hat{\theta}_k-\epsilon}^{\hat{\theta}_k+\epsilon} 1 \quad d\theta_k && (4.178) \\ &= 1 - p(\dot{\theta}|z)(2\epsilon)^n && (4.179) \\ &\approx 1 - p(\hat{\theta}|z)(2\epsilon)^n. && (4.180) \end{aligned}$$

So for small enough ϵ, it is $\dot{\theta} \approx \hat{\theta}$, and the above conditional Bayes cost can be minimized by maximizing the posterior density $p(\hat{\theta}|z)$. Thus, the optimal estimator $\hat{\theta}_{UC}$ related to the uniform cost function is defined as ([145], p. 189)

$$p(\hat{\theta}_{UC}|z) \geq p(\hat{\theta}|z), \quad \forall \hat{\theta}. \tag{4.181}$$

So since $\hat{\theta}_{UC}$ maximizes the density function $p(\theta|z)$, i.e. is the mode of the density, $\hat{\theta}_{UC}$ is also called the conditional mode estimator, or more commonly the maximum a posteriori (MAP) estimator $\hat{\theta}_{MAP}$ [145], and directly

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} p(\theta|z). \tag{4.182}$$

In order to obtain the MAP estimator the posterior density is required and is obtained from the Bayes' formula

$$p_{\theta|z}(\theta|z) = \frac{p_{z|\theta}(z|\theta)p_\theta(\theta)}{p_z(z)}, \tag{4.183}$$

where the denominator $p_z(z)$ does not depend on the parameter vector θ, and merely normalizes the posterior density $p_{\theta|z}(\theta|z)$ [108]. Hence, in order to obtain the MAP estimator of θ it suffices to find the value that maximizes the numerator, which is the joint density

$$p_{\theta,z}(\theta, z) = p_{z|\theta}(z|\theta)p_\theta(\theta). \tag{4.184}$$

Like in the maximum likelihood method, the MAP estimator can usually be found by considering the logarithm. This now has the form

$$\ln p(\theta, z) = \ln p(z|\theta) + \ln p(\theta). \tag{4.185}$$

By comparing with the respective log-likelihood equation we see that the MAP equation contains an additional term $\ln p(\theta)$ that takes into account prior information about the parameters θ. The solution is then regularized by that term. It can be shown, that if the prior density of θ is uniform for parameter values which include the ML estimator, then the MAP and ML estimators become the same. In other words, if we have no a priori knowledge concerning θ other than it is in a given region, then the ML and MAP estimates are equal ([145], p. 191). Clearly, $\hat{\theta}_{MAP} = \hat{\theta}_{MS}$ if the mode of the density $p(\theta|z)$ equals the mean $\eta_{\theta|z}$. This is the case that $p(\theta|z)$ is symmetric and unimodal, like in the Gaussian case.

CHAPTER V

Recursive Estimation and Kalman filtering

In this chapter, recursive mean square estimation methods are discussed. This involves state-space modeling and Kalman filters and smoothers for state estimation. The related concepts are approached through Bayesian estimation theory and much less through adaptive Wiener filtering theory. Main references include [145, 184, 185, 108] see also [7, 75, 71, 161, 160, 25].

5.1 Basic concepts and definitions

Consider that we observe T realizations of a random vector, for which we assume the additive noise models

$$z_t = H_t\theta_t + \upsilon_t, \quad t = 1, \ldots, T, \tag{5.1}$$

where H_t are known matrices. If we assume that the random parameters θ_t and the noise vectors υ_t are uncorrelated for every t, estimates can be obtained by the optimal linear mean square estimator as (4.167)

$$\hat{\theta}_t = (H_t^T C_{\upsilon_t}^{-1} H_t + C_{\theta_t}^{-1})^{-1}(H_t^T C_{\upsilon_t}^{-1} z_t + C_{\theta_t}^{-1}\eta_{\theta_t}), \quad t = 1, \ldots, T, \tag{5.2}$$

where we have assumed that υ_t are zero mean for every t. Thus, the identification of the first and second order properties of the prior densities $p(\theta_t)$ are needed. If all the involved densities are Gaussian, then the estimator is overall optimal, not only among linear estimators, and is also the MAP estimator. For the parameters θ_t we can also assume a common prior density $p(\theta_t) = p(\theta)$ for every t. Then estimates for the T realizations of the random vector θ can be obtained by

$$\hat{\theta}_t = (H_t^T C_{\upsilon_t}^{-1} H_t + C_{\theta}^{-1})^{-1}(H_t^T C_{\upsilon_t}^{-1} z_t + C_{\theta}^{-1}\eta_{\theta}), \quad t = 1, \ldots, T, \tag{5.3}$$

Estimates for $s_t = H_t\theta_t$ are obtained in both cases as

$$\hat{s}_t = H_t\hat{\theta}_t, \quad t = 1, \ldots, T. \tag{5.4}$$

We can consider also the model

$$z_t = H_t\theta + \upsilon_t, \quad t = 1, \ldots, T, \tag{5.5}$$

i.e. all the measurements are about the same parameter θ. The parameter can still be considered random. We can then form the model (or the posterior if we consider Gaussianity)

$$z = H\theta + \upsilon, \tag{5.6}$$

where $z = (z_1^T, \ldots, z_T^T)^T$ are the measurements, $\upsilon = (\upsilon_1^T, \ldots, \upsilon_T^T)^T$ is the noise and $H = (H_1^T, \ldots, H_T^T)^T$ is the observation model. If we also assume that the noise vectors υ_t are mutually independent, then the estimator becomes

$$\hat{\theta} = (H^T C_\upsilon^{-1} H + C_\theta^{-1})^{-1}(H^T C_\upsilon^{-1} z_t + C_\theta^{-1}\eta_\theta) \tag{5.7}$$

$$= (\sum_{t=1}^{T} H_t^T C_{\upsilon_t}^{-1} H_t + C_\theta^{-1})^{-1}(\sum_{t=1}^{T} H_t C_{\upsilon_t}^{-1} z_t + C_\theta^{-1}\eta_\theta). \tag{5.8}$$

Note, that the prior information becomes negligible when T grows, see also for example ([184], p. 203). If now we drop the prior the estimator reduces to the Gauss-Markov estimator (or maximum likelihood for Gaussian noise vectors). If we also consider $H_t = I$ and $C_{\upsilon_t} = C$ for every t, i.e. the noise vectors are independent identically distributed, then the estimator is the conventional average of the measurements $\hat{\theta} = \hat{s} = \frac{1}{T}\sum_{t=1}^{T} z_t$.

In the following we are interested in the first case (5.1), where the distribution of the parameters has different characteristics in every realization, and especially for the situation that the parameters θ_t have time correlations. Then the order of the measurements is important and $\{z_t\}, \{\theta_t\}$ can be considered as vector stochastic processes. Estimates for the parameters can be obtained recursively in terms of Kalman filtering [110, 111].

5.2 State-space modeling

5.2.1 State-space observation model

Consider that the time-varying parameters θ_t are indeed time instances of a vector valued stochastic process $\{\theta_t\}$. The simplest non-stationary process that can serve as a model is the first order Markov process. Consider a process that satisfies the recursion

$$\theta_t = f_t(\theta_{t-1}, \omega_t), \tag{5.9}$$

where f_t is a time-varying vector valued function of θ_{t-1} and ω_t. The noise process ω_t is a sequence of independent random vectors with time-varying distributions. The vector θ_t is determined in terms of θ_{t-1} and ω_t only, hence it is conditionally independent of θ_k for $k < t-1$, so the process is Markov having the memoryless property

$$p(\theta_t|\theta_{t-1}, \theta_{t-2}, \ldots, \theta_0) = p(\theta_t|\theta_{t-1}). \tag{5.10}$$

Where we considered a random starting point θ_0 independent of ω_t for every t. From the independence of $\{\omega_t\}$ and the model we have that θ_{t-1} depends on all

the past of the noise process, i.e. ω_k for $k = 1, \ldots, t-2, t-1$, and is independent of ω_k for $k = t$. For example, for the function $f_t(\theta_{t-1}, \omega_t) = \theta_{t-1} + \omega_t$ we have

$$\theta_t = \theta_{t-1} + \omega_t = \ldots = \theta_0 + \omega_1 + \omega_2 + \ldots + \omega_t. \tag{5.11}$$

Thus, the sum of independent random vectors or variables is a Markov sequence. This process is called random walk and is non-stationary. Also, if ω_t are zero mean and a sample path is given up to θ_{t-1}, then from equations (5.10) and (5.11) it holds

$$E\{\theta_t|\theta_{t-1}, \theta_{t-2}, \ldots, \theta_0\} = E\{\theta_{t-1} + \omega_t|\theta_{t-1}\} = \theta_{t-1}. \tag{5.12}$$

In general, a random sequence with the previous property is called martingale [59].

Consider now a random process $\{z_t\}$ that relates to the process (5.11) through the relation $z_t = \theta_t + \upsilon_t$, where υ_t is a sequence of independent random vectors independent of ω_t and not necessarily identically distributed. We observe that the process z_t is not Markov, since

$$z_t = \theta_t + \upsilon_t = \theta_{t-1} + \omega_t + \upsilon_t = z_{t-1} - \upsilon_{t-1} + \upsilon_t + \omega_t, \tag{5.13}$$

and the vector $\xi_t = -\upsilon_{t-1} + \upsilon_t + \omega_t$ depends on ξ_{t-1}, because they have the common term υ_{t-1}. After these observations we can now generalize.

A random process $\{z_t\}$ is observed and is assumed to relate to another process $\{\theta_t\}$ through the model

$$\theta_t = f_t(\theta_{t-1}, \omega_t) \tag{5.14}$$

$$z_t = h_t(\theta_t, \upsilon_t) \tag{5.15}$$

for every $t = 1, 2, \ldots$. Equation (5.14) defines the state or time evolution for a not directly observed Markov process $\{\theta_t\}$ and (5.15) the observation equation that relates the hidden process to the measurements. This an example of a *state-space observation model* ([108], p. 119). The assumptions of the model are summarized as follows

- the functions f_t, h_t are well defined known vector valued functions for all t,
- $\{\omega_t\}$ is a sequence of independent random vectors with different distributions and defines the *state noise process*,
- $\{\upsilon_t\}$ is also a white random vector process that represents *observation noise*,
- the random vectors ω_t and υ_t are mutually independent for every t,
- the distributions of ω_t, υ_t are known (or preselected),
- there is an initial vector θ_0 with known distribution independent of ω_t and υ_t for every t.

Obviously, the simplest state evolution is given by the random walk model. The observation model can be linear or non-linear depending on the application. Finally, the estimation problem related to the state-space model is to make inference about $\{\theta_t\}$ by observing $\{z_t\}$.

5.2.2 The evolution observation pair

The previous estimation problem can be described in a different way. Let $\{z_t\}$ be an observable sequence of data that contain information about an unobserved physical mechanism or system that has a stochastic behavior. For the description of the system a vector of parameters θ_t, i.e. the process of parameters of interest, is selected that naturally depends on the measurements. With natural dependency of parameters and measurements we only assume that the likelihood densities $p(z_t|\theta_t)$ have meaning. The processes $\{\theta_t\}$, $\{z_t\}$ form a (first order) *evolution observation pair* if the following properties hold ([108], p. 118)

- the hidden state evolution process $\{\theta_t\}$ has the memoryless property

$$p(\theta_t|\theta_{t-1}, \theta_{t-1}, \ldots, \theta_0) = p(\theta_t|\theta_{t-1}) \tag{5.16}$$

  for some random starting point θ_0 and some evolution up to t,

- the observed process $\{z_t\}$ has the memoryless property with respect to the history of $\{\theta_t\}$, that is

$$p(z_t|\theta_t, \theta_{t-1}, \ldots, \theta_0) = p(z_t|\theta_t), \tag{5.17}$$

- the parameter process $\{\theta_t\}$ depends on the past only through its own history, that is

$$p(\theta_t|\theta_{t-1}, z_{t-1}, z_{t-2} \ldots, z_1) = p(\theta_t|\theta_{t-1}). \tag{5.18}$$

Notice, that as soon as a state-space model is defined for an evolution observation pair, then the assumptions of the model come in parallel with the above definitions.

Let $\{\theta_k\}, \{z_k\}, k = 1, \ldots, t$ a path of the evolution observation pair. Then we can regroup the join density

$$\begin{aligned} p(z_t, \ldots, z_1, \theta_t, \ldots, \theta_1) &= p(z_t, \theta_t, \ldots, z_1, \theta_1) && (5.19)\\ &= p(z_1, \theta_1) \prod_{k=2}^{t} p(z_k, \theta_k|z_{k-1}, \theta_{k-1}, \ldots, z_1, \theta_1) && (5.20) \end{aligned}$$

Where we applied the chain rule for conditioning over the pairs $\{z_t, \theta_t\}$. For the conditional densities by applying (3.13) we have

$$\begin{aligned} p(z_k, \theta_k|z_{k-1}, \theta_{k-1}, \ldots, z_1, \theta_1) &= p(z_k|\theta_k, \theta_{k-1}, \ldots, \theta_1, z_{k-1}, \ldots, z_1) \\ &\quad \cdot p(\theta_k|\theta_{k-1}, \ldots, \theta_1, z_{k-1}, \ldots, z_1) && (5.21)\\ &= p(z_k|\theta_k)p(\theta_k|\theta_{k-1}) && (5.22)\\ &= p(z_k|\theta_k, \theta_{k-1})p(\theta_k|\theta_{k-1}, z_{k-1}) && (5.23)\\ &= p(z_k, \theta_k|z_{k-1}, \theta_{k-1}). && (5.24) \end{aligned}$$

Where we applied (3.13) again. Equation (5.22) holds because of (5.16-5.18). Also from (5.15) and from the mutual independence of υ_t, if θ_t is given then z_t is

conditionally independent of the previous observations. From the last equation we have that the stochastic process defined by the pairs $\{z_t, \theta_t\}$ have the memoryless property. Also by using equations (5.22), (5.20) and (5.17) it holds

$$p(z_t, \theta_t, \ldots, z_1, \theta_1, \theta_0) = p(z_t|\theta_t)p(\theta_t|\theta_{t-1}) \ldots p(z_1|\theta_1)p(\theta_1|\theta_0)p(\theta_0) \quad (5.25)$$

$$= p(\theta_0) \prod_{k=1}^{t} p(z_k|\theta_k)p(\theta_k|\theta_{k-1}). \quad (5.26)$$

Thus, for the evolution observation pair to be completely specified we need the probability density of the initial state $p(\theta_0)$, the transition conditional densities $p(\theta_t|\theta_{t-1})$ and the likelihood functions $p(z_t|\theta_t)$, for all $t = 1, 2, \ldots$ ([108], p. 119).

5.2.3 Bayesian formulation and related estimation problems

Let $\{\theta_t\}$, $\{z_t\}$ be an evolution observation pair, then the following problems are under concideration

- prediction, i.e. the determination of the predictive conditional densities $p(\theta_t|z_1, z_2, \ldots, z_{t-1})$ or more general $p(\theta_t|z_1, \ldots, z_{t-p})$ for $p \geq 1$ (p-step),
- filtering, i.e. the determination of $p(\theta_t|z_1, \ldots, z_t)$,
- fixed lag smoothing, i.e. the determination of $p(\theta_t|z_1, \ldots, z_{t+p})$ for $p \geq 1$ (p-lag),
- fixed interval smoothing, i.e when a measurement sequence is obtained for $t = 1, \ldots, T$ the determination of the conditional densities $p(\theta_t|z_1, \ldots, z_T)$.

Based on these conditional or posterior densities, estimators can be defined in a Bayesian framework. It can be noticed, that all the above mentioned problems are computationally related to the one step prediction problem as an intermediate step. In general, prediction problems are important for estimating future evolutions based on current information, thus forecasting. Filtering or (p-lag) smoothing can be used, for example, in situations that on-line processing is important as in optimal control problems. However, they can be used for batch processing as well, but the fixed interval smoothing usually yields smaller estimation errors.

Filtering and prediction distributions

Let us now consider the posterior density for θ_t given past and present measurements up to time instant t. Based on that conditional density, Bayesian MAP or conditional mean estimators can be defined. If we denote $Z_t = \{z_1, z_2, \ldots, z_t\}$ the compound measurements up to t, we have for the posterior in terms of the likelihood and the prior density of θ_t that

$$p(\theta_t|z_1, z_2, \ldots, z_t) = p(\theta_t|Z_t) = \frac{p(Z_t|\theta_t)p(\theta_t)}{p(Z_t)}. \quad (5.27)$$

For the likelihood by applying (3.13) we have

$$p(Z_t|\theta_t) = p(z_t, Z_{t-1}|\theta_t) = p(z_t|Z_{t-1}, \theta_t)p(Z_{t-1}|\theta_t) \quad (5.28)$$
$$= p(z_t|\theta_t)p(Z_{t-1}|\theta_t), \quad (5.29)$$

since if θ_t is given then z_t is conditionally independent to its past. Furthermore, (5.27) becomes [77]

$$p(\theta_t|Z_t) = \frac{p(z_t|\theta_t)p(Z_{t-1}|\theta_t)p(\theta_t)}{p(z_t, Z_{t-1})} \quad (5.30)$$
$$= \frac{p(z_t|\theta_t)p(\theta_t|Z_{t-1})p(Z_{t-1})}{p(z_t|Z_{t-1})p(Z_{t-1})} \quad (5.31)$$
$$= \frac{p(z_t|\theta_t)p(\theta_t|Z_{t-1})}{p(z_t|Z_{t-1})}, \quad (5.32)$$

where we applied again the Bayes rule. Thus, the posterior density is proportional to the product of the likelihood function $p(z_t|\theta_t)$ based on current information for the parameters and a compound prediction-prior density, which contains information from the state evolution process and past measurements since

$$p(\theta_t|Z_{t-1}) = \int p(\theta_t, \theta_{t-1}|Z_{t-1})d\theta_{t-1} = \int p(\theta_t|\theta_{t-1})p(\theta_{t-1}|Z_{t-1})d\theta_{t-1}. \quad (5.33)$$

The last equation provides the base for the different estimation problems. Equations (5.32), (5.33), see also (5.26), suggest a recursive estimation procedure for the determination of the density. For the denominator of (5.32) we can write

$$p(z_t|Z_{t-1}) = \int p(z_t, \theta_t|Z_{t-1})d\theta_t = \int p(z_t|\theta_t)p(\theta_t|Z_{t-1})d\theta_t, \quad (5.34)$$

which is the the integral of the nominator, i.e. a normalization constant, and it holds

$$p(\theta_t|Z_t) = \frac{\int p(z_t|\theta_t)p(\theta_t|\theta_{t-1})p(\theta_{t-1}|Z_{t-1})d\theta_{t-1}}{\int\int p(z_t|\theta_t)p(\theta_t|\theta_{t-1})p(\theta_{t-1}|Z_{t-1})d\theta_t d\theta_{t-1}} \quad (5.35)$$
$$\propto p(z_t|\theta_t)p(\theta_t|Z_{t-1}). \quad (5.36)$$

Smoothing distribution

Let us now consider the smoothing problem $p(\theta_t|Z_T)$. For this posterior we have

$$\begin{aligned} p(\theta_t|Z_T) &= \int p(\theta_t, \theta_{t+1}|Z_T)d\theta_{t+1} && (5.37)\\ &= \int p(\theta_t|\theta_{t+1}, Z_T)p(\theta_{t+1}|Z_T)d\theta_{t+1} && (5.38)\\ &= \int p(\theta_t|\theta_{t+1}, Z_t)p(\theta_{t+1}|Z_T)d\theta_{t+1} && (5.39)\\ &= \int \frac{p(\theta_{t+1}, Z_t|\theta_t)p(\theta_t)p(\theta_{t+1}|Z_T)}{p(\theta_{t+1}, Z_t)}d\theta_{t+1} && (5.40)\\ &= \int \frac{p(\theta_{t+1}|\theta_t, Z_t)p(Z_t|\theta_t)p(\theta_t)p(\theta_{t+1}|Z_T)}{p(\theta_{t+1}|Z_t)p(Z_t)}d\theta_{t+1} && (5.41)\\ &= p(\theta_t|Z_t)\int \frac{p(\theta_{t+1}|\theta_t)p(\theta_{t+1}|Z_T)}{p(\theta_{t+1}|Z_t)}d\theta_{t+1} && (5.42) \end{aligned}$$

Equation (5.39) is due to (3.121) and (5.18), and denotes a backward treatment for the process. The last form suggests again a recursive estimation procedure for the determination of the conditional density [10]. It is thus possible to compute the filtering and prediction distributions in a forward (filtering) recursion, i.e calculating $p(\theta_t|Z_t)$ and $p(\theta_{t+1}|Z_t)$ from (5.32), (5.33), and then execute a backward recursion with each smoothed distribution $p(\theta_t|Z_T)$ relying upon the quantities calculated in the forward run and the previous (in reverse time) smoothed distributions $p(\theta_{t+1}|Z_T)$. This property enables the formulation of the *forward-backward* method for the smoothing problem [174, 7], which gives the smoother estimates as corrections of the filter estimates.

Let us now write the joint density $p(\theta_t, \theta_{t+1}|Z_T)$ in a different form, i.e.

$$\begin{aligned} p(\theta_t, \theta_{t+1}|Z_T) &= \frac{p(Z_T|\theta_t, \theta_{t+1})p(\theta_t, \theta_{t+1})}{p(Z_T)} && (5.43)\\ &= \frac{p(Z_t, z_{t+1}, \ldots, z_T|\theta_t, \theta_{t+1})p(\theta_{t+1}, \theta_t)}{p(Z_T)} && (5.44)\\ &= \frac{p(Z_t|z_{t+1}, \ldots, z_T, \theta_t, \theta_{t+1})p(z_{t+1}, \ldots, z_T|\theta_t, \theta_{t+1})p(\theta_{t+1}, \theta_t)}{p(Z_T)} && (5.45)\\ &= \frac{p(Z_t|\theta_t)p(z_{t+1}, \ldots, z_T|\theta_{t+1})p(\theta_{t+1}|\theta_t)p(\theta_t)}{p(z_{t+1}, \ldots, z_T|Z_t)p(Z_t)} && (5.46)\\ &= \frac{p(\theta_t|Z_t)p(z_{t+1}, \ldots, z_T|\theta_{t+1})p(\theta_{t+1}|\theta_t)}{p(z_{t+1}, \ldots, z_T|Z_t)}. && (5.47) \end{aligned}$$

This form will be used for the derivation of the forward backward smoother as in [174]. Furthermore, by applying twice the Bayes rule we have from the last

expression

$$p(\theta_t|Z_T) = \int p(\theta_t, \theta_{t+1}|Z_T)d\theta_{t+1} \quad (5.48)$$

$$\propto p(\theta_t|Z_t)\int \frac{p(\theta_{t+1}|z_{t+1},\ldots,z_T)p(\theta_t|\theta_{t+1})}{p(\theta_t)}d\theta_{t+1}, \quad (5.49)$$

where $p(\theta_t) = \pi_t(\theta_t)$ is an artificial prior distribution for θ_t. This form suggests a *two-filter* smoothing procedure, such that the first filter recurs in the forward direction computing $p(\theta_t|Z_t)$ and the second filter recurs in the backward direction computing $\bar{p}(\theta_t|z_t,\ldots,z_T)$ (*backward information filter*). The smoothing distribution is then computed as an optimal combination of the two independent filters [61, 144, 122, 71].

Finally, we have from (5.26)

$$p(\theta_1,\ldots,\theta_T|z_1,\ldots,z_T) = \frac{p(\theta_1,z_1,\ldots,\theta_T,z_T)}{p(Z_T)} \quad (5.50)$$

$$\propto \prod_{t=1}^{T} p(z_t|\theta_t)\xi_1(\theta_1)\prod_{t=2}^{T-1} p(\theta_t|\theta_{t-1})\xi_T(\theta_T), \quad (5.51)$$

where $\xi_1(\theta_1) = p(\theta_1) = \bar{p}(\theta_1|\theta_0) = \bar{p}(\theta_0|\theta_1)\bar{p}(\theta_1)$, and $\xi_T(\theta_T) = p(\theta_T|\theta_{T-1}) = p(\theta_{T-1}|\theta_T)p(\theta_T)/p(\theta_{T-1})$. This reflects the fact that there is not enough prior information from the model about the distribution of θ_T [61], since no information obtained after the end of the measurements is used for the terminal state. This can also be understood by considering initialization for the backward recursions for both the previous discussed cases. However, it must be noted that this is of small practical interest when long measurements are available, since estimates for initialization can be obtained by backward and forward filtering.

5.3 Recursive mean square estimation

5.3.1 The linear Gaussian case

Let us consider the case of the linear state-space model, i.e.

$$\theta_t = F_t\theta_{t-1} + G_t\omega_t, \quad (5.52)$$

$$z_t = H_t\theta_t + \upsilon_t, \quad (5.53)$$

where H_t, F_t and G_t are assumed to be known matrices. Without loss of generality, we can assume that the noise processes are zero mean. For determining the posterior (5.32) we need an error distribution to define the likelihood $p(z_t|\theta_t)$ and a recursive procedure to define (5.33). Since the MAP estimator takes the same form with the linear conditional mean estimator for Gaussian densities we are going to derive it based on the Gaussianity assumption.

A Gaussian density for θ_0 and Gaussian densities for ω_t directly give from the state evolution that θ_t are Gaussian distributed for every t as linear combinations of Gaussian vectors. Thus it is

$$\theta_t \sim N(F_t\eta_{\theta_{t-1}}, F_tC_{\theta_{t-1}}F_t^T + G_tC_{\omega_t}G_t^T) \tag{5.54}$$

and $\theta_0 \sim N(\eta_{\theta_0}, C_{\theta_0})$. Additionally for the conditional density of θ_t given θ_{t-1} it holds

$$\theta_{t|t-1} \sim N(F_t\theta_{t-1}, G_tC_{\omega_t}G_t^T). \tag{5.55}$$

Also υ_t is assumed Gaussian, thus leading to a Gaussian likelihood model, and the covariances $C_{\omega_t}, C_{\upsilon_t}$ are assumed known.

5.3.2 Kalman filter

We search for the value that maximizes the density

$$p(\theta_t|Z_t) = \frac{p(z_t|\theta_t)p(\theta_t|Z_{t-1})}{p(z_t|Z_{t-1})}. \tag{5.56}$$

Thus, we search for the estimator

$$\hat{\theta} = E\{\theta_t|Z_t\}, \tag{5.57}$$

based on the linear model and the Gaussian assumptions. For Gaussian densities we only need to derive the means and covariances. For the mean and covariance of the density $p(z_t|\theta_t)$ we have

$$\eta_{z_t|\theta_t} = E\{z_t|\theta_t\} = E\{H_t\theta_t + \upsilon_t|\theta_t\} = H_t\theta_t \tag{5.58}$$

$$C_{z_t|\theta_t} = E\{(H_t\theta_t + \upsilon_t - \eta_{z_t|\theta_t})(H_t\theta_t + \upsilon_t - \eta_{z_t|\theta_t})^T|\theta_t\} = C_{\upsilon_t}. \tag{5.59}$$

Thus, the Gaussian density $p(z_t|\theta_t)$, or the likelihood function, is the density of υ_t and is of the form

$$p(z_t|\theta_t) \propto \exp\left(-\frac{1}{2}(z_t - H_t\theta_t)^TC_{\upsilon_t}^{-1}(z_t - H_t\theta_t)\right). \tag{5.60}$$

For the density $p(\theta_t|Z_{t-1})$ we have

$$\eta_{\theta_t|Z_{t-1}} = E\{F_t\theta_{t-1} + G_t\omega_t|Z_{t-1}\} \tag{5.61}$$

$$= F_t\hat{\theta}_{t-1} = \hat{\theta}_{t|t-1}, \tag{5.62}$$

$$C_{\theta_t|Z_{t-1}} = E\{(\theta_t - \eta_{\theta_t|Z_{t-1}})(\theta_t - \eta_{\theta_t|Z_{t-1}})^T|Z_{t-1}\} \tag{5.63}$$

$$= E\{(\theta_t - F_t\hat{\theta}_{t-1})(\theta_t - F_t\hat{\theta}_{t-1})^T|Z_{t-1}\} = C_{\tilde{\theta}_{t|t-1}}, \tag{5.64}$$

where $\hat{\theta}_{t|t-1}$ is the estimate (or prediction) for the state θ_t given the past observations Z_{t-1}, and $C_{\tilde{\theta}_{t|t-1}}$ is the error covariance of the prediction. The density $p(\theta_t|Z_{t-1})$ is then of the form

$$p(\theta_t|Z_{t-1}) \propto \exp\left(-\frac{1}{2}(\theta_t - \hat{\theta}_{t|t-1})^TC_{\tilde{\theta}_{t|t-1}}^{-1}(\theta_t - \hat{\theta}_{t|t-1})\right). \tag{5.65}$$

Thus, the posterior has the Gaussian form

$$p(\theta_t|Z_t) \propto \exp\left(-\frac{1}{2}\|z_t - H_t\theta_t\|^2_{C_{\upsilon_t}^{-1}} - \frac{1}{2}\|\theta_t - \hat{\theta}_{t|t-1}\|^2_{C^{-1}_{\tilde{\theta}_{t|t-1}}}\right). \tag{5.66}$$

In order to obtain the MAP estimator first the logarithm of the posterior is considered. By setting to zero the gradient of the log posterior with respect to θ_t we have

$$H_t^T C_{\upsilon_t}^{-1}(z_t - H_t\hat{\theta}_t) - C^{-1}_{\tilde{\theta}_{t|t-1}}(\hat{\theta}_t - \hat{\theta}_{t|t-1}) = 0, \tag{5.67}$$

which gives the form (e.g. [145], p. 247)

$$\hat{\theta}_t = (H_t^T C_{\upsilon_t}^{-1} H_t + C^{-1}_{\tilde{\theta}_{t|t-1}})^{-1}(H_t^T C_{\upsilon_t}^{-1} z_t + C^{-1}_{\tilde{\theta}_{t|t-1}}\hat{\theta}_{t|t-1}). \tag{5.68}$$

This has exactly the same form as the Bayesian mean square estimator (4.167), or the MAP estimator for Gaussian variables. Prior information is used in the estimator sequentially in terms of the prediction prior density (5.65). Another formulation of the estimator as in (4.160), by applying the matrix inversion lemma, is given by

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_{t|t-1} + K_t(z_t - H_t\hat{\theta}_{t|t-1}) && (5.69)\\ &= F_t\hat{\theta}_{t-1} + K_t(z_t - H_t F_t\hat{\theta}_{t-1}), && (5.70) \end{aligned}$$

where the matrix K_t is called *Kalman Gain* matrix given by (e.g. [145], p. 247)

$$K_t = C_{\tilde{\theta}_{t|t-1}} H_t^T (H_t C_{\tilde{\theta}_{t|t-1}} H_t^T + C_{\upsilon_t})^{-1}. \tag{5.71}$$

Thus, the estimate at time t is obtained by correcting the prediction $F_t\hat{\theta}_{t-1}$, based on the state model and past observations, with the one step prediction error $z_t - H_t F_t\hat{\theta}_{t-1}$, or innovation vector, that represents the new information in the observed data z_t (e.g. [75] chapter 7). The estimation error can be written as

$$\begin{aligned} \tilde{\theta}_t &= \theta_t - \hat{\theta}_t && (5.72)\\ &= \theta_t - \hat{\theta}_{t|t-1} - K_t(z_t - H_t\hat{\theta}_{t|t-1}) && (5.73)\\ &= \theta_t - \hat{\theta}_{t|t-1} - K_t(H_t\theta_t + \upsilon_t - H_t\hat{\theta}_{t|t-1}) && (5.74)\\ &= \tilde{\theta}_{t|t-1} - K_t(H_t\tilde{\theta}_{t|t-1} + \upsilon_t) && (5.75)\\ &= (I - K_t H_t)\tilde{\theta}_{t|t-1} - K_t\upsilon_t, && (5.76) \end{aligned}$$

and has covariance

$$\begin{aligned} C_{\tilde{\theta}_t} &= (I - K_t H_t) C_{\tilde{\theta}_{t|t-1}} (I - K_t H_t)^T + K_t C_{\upsilon_t} K_t^T && (5.77)\\ &= (I - K_t H_t) C_{\tilde{\theta}_{t|t-1}}, && (5.78) \end{aligned}$$

where we could have also used directly equations (4.161) or (4.168). Thus, the following expression is equivalent

$$C_{\tilde{\theta}_t} = (C^{-1}_{\tilde{\theta}_{t|t-1}} + H_t^T C_{\upsilon_t}^{-1} H_t)^{-1}. \tag{5.79}$$

In order to be complete, the Kalman filter algorithm requirs the recursive computation of the error covariance $C_{\tilde{\theta}_{t|t-1}}$. For the prediction error it holds

$$\tilde{\theta}_{t|t-1} = \theta_t - \hat{\theta}_{t|t-1} = F_t\theta_{t-1} + G_t\omega_t - F_t\hat{\theta}_{t-1} = F_t\tilde{\theta}_{t-1} + G_t\omega_t \tag{5.80}$$

Thus, $C_{\tilde{\theta}_{t|t-1}}$ can be written as

$$\begin{aligned} C_{\tilde{\theta}_{t|t-1}} &= E\{(F_t\tilde{\theta}_{t-1} + G_t\omega_t)(F_t\tilde{\theta}_{t-1} + G_t\omega_t)^T\} && (5.81)\\ &= F_tC_{\tilde{\theta}_{t-1}}F_t^T + G_tC_{\omega_t}G_t^T, && (5.82)\end{aligned}$$

since, $\hat{\theta}_{t-1}$ and ω_t are uncorrelated. Thus, we have a recursive formula to update the prediction error covariance and the derivation of Kalman filter is now complete. It can take at least two different forms depending of the formulation for the mean square estimator. We can summarize one of them as

$$\begin{aligned} C_{\tilde{\theta}_{t|t-1}} &= F_tC_{\tilde{\theta}_{t-1}}F_t^T + G_tC_{\omega_t}G_t^T && (5.83)\\ K_t &= C_{\tilde{\theta}_{t|t-1}}H_t^T(H_tC_{\tilde{\theta}_{t|t-1}}H_t^T + C_{\upsilon_t})^{-1} && (5.84)\\ \hat{\theta}_t &= F_t\hat{\theta}_{t-1} + K_t(z_t - H_tF_t\hat{\theta}_{t-1}) && (5.85)\\ C_{\tilde{\theta}_t} &= (I - K_tH_t)C_{\tilde{\theta}_{t|t-1}}. && (5.86)\end{aligned}$$

In order to initialize the algorithm, prior knowledge for $\hat{\theta}_0 = E\{\theta_0\}$ and $C_{\tilde{\theta}_0} = C_{\theta_0}$ is required.

If the state-space model contains known (input or control) deterministic vectors x_t, y_t, so that it becomes

$$\begin{aligned} \theta_t &= F_t\theta_{t-1} + G_t\omega_t + x_t && (5.87)\\ z_t &= H_t\theta_t + \upsilon_t + y_t, && (5.88)\end{aligned}$$

the recursive mean square estimator for the state becomes

$$\hat{\theta}_t = F_t\hat{\theta}_{t-1} + x_t + K_t(z_t - H_tF_t\hat{\theta}_{t-1} - H_tx_t - y_t). \tag{5.89}$$

The covariance and gain equations remain the same ([145], p. 249). The same form is obtained for the case that ω_t and υ_t have non-zero means. Their time varying means can, for example, be treated as known inputs. See also equation (4.160) for the general form of the mean square estimator.

Finally, consider a different state-space model, where state and observation noise are correlated. Thus there are extra known matrices for the cross correlations $C_{\omega_t,\upsilon_{t-1}} = C_t \neq 0$. The rest of the previous assumptions remain the same. We can transform this model to the case that Kalman filter can be used ([145], p. 250). We rewrite the state equation by adding a zero valued term, i.e.

$$\begin{aligned} \theta_t &= F_t\theta_{t-1} + G_t\omega_t + A_t(z_{t-1} - H_{t-1}\theta_{t-1} - \upsilon_{t-1}) && (5.90)\\ &= (F_t - A_tH_{t-1})\theta_{t-1} + (G_t\omega_t - A_t\upsilon_{t-1}) + A_tz_{t-1} && (5.91)\\ &= F_t'\theta_{t-1} + \omega_t' + x_t && (5.92)\end{aligned}$$

We also have that

$$C_{\omega'_t,\upsilon_{t-1}} = E\{(G_t\omega_t - A_t\upsilon_{t-1})\upsilon_{t-1}^T\} = F_tC_t - A_tC_{\upsilon_{t-1}}. \tag{5.93}$$

Thus, with the selection $A_t = F_tC_tC_{\upsilon_{t-1}}^{-1}$, the cross correlation in (5.93) becomes zero and the modified state space model has the necessary properties and Kalman filter can be applied. Different transformations of an original state-space model can be applied in order to bring the problem in the standard form for Kalman filtering, for example, for the fixed-lag smoothing problem as well as for higher order Markov models [108].

5.3.3 Fixed interval smoother

Similar to the filtering problem we search for the value that maximizes the density

$$p(\theta_t|Z_T) = p(\theta_t|z_1, z_2, \ldots, z_T), \tag{5.94}$$

where $Z_T = \{z_1, z_2, \ldots, z_T\}$ is a given measurement path. Thus, we search for the estimator (smoother)

$$\hat{\theta}_t^s = E\{\theta_t|Z_T\}, \tag{5.95}$$

based on the linear state-space model and the Gaussian assumptions. From (5.47) we have for the joint density $p(\theta_t, \theta_{t+1}|Z_T)$

$$p(\theta_t, \theta_{t+1}|Z_T) \propto p(\theta_t|Z_t)p(z_{t+1}, \ldots, z_T|\theta_{t+1})p(\theta_{t+1}|\theta_t). \tag{5.96}$$

Since first a forward filter recursion is considered, the densities $p(\theta_t|Z_t)$ are estimated to be Gaussian with means and variances

$$\eta_{\theta_t|Z_t} = \hat{\theta}_t, \tag{5.97}$$

$$C_{\theta_t|Z_t} = C_{\tilde{\theta}_t}. \tag{5.98}$$

For the Gaussian evolution density $p(\theta_{t+1}|\theta_t)$ we have from the model

$$\eta_{\theta_{t+1}|\theta_t} = F_{t+1}\theta_t, \tag{5.99}$$

$$C_{\theta_{t+1}|\theta_t} = G_{t+1}C_{\omega_{t+1}}G_{t+1}^T. \tag{5.100}$$

Lets us denote $\hat{\theta}_t^s$ the (smooth) estimator or the value of θ_t that maximizes the required density $p(\theta_t|Z_T)$. Then $\hat{\theta}_t^s$, $\hat{\theta}_{t+1}^s$ are the values that maximize the joint density $p(\theta_t, \theta_{t+1}|Z_T)$. We search for the maximum with respect to θ_t, θ_{t+1} of

$$\log p(\theta_t, \theta_{t+1}|Z_T) \propto -\|\theta_{t+1} - F_{t+1}\theta_t\|^2_{(G_{t+1}C_{\omega_{t+1}}G_{t+1}^T)^{-1}} - \|\theta_t - \hat{\theta}_t\|^2_{C_{\tilde{\theta}_t}^{-1}} + C, \tag{5.101}$$

where C represents terms arising from $p(z_{t+1}, \ldots, Z_T|\theta_{t+1})$ that do not depend on θ_t. Now, since we try to define the backward sequential procedure, the estimator $\hat{\theta}_{t+1}^s = \arg\max p(\theta_{t+1}|Z_T)$ is considered to be already obtained. It follows immediately that θ_t^s is the one that minimizes the expression

$$l = (\hat{\theta}_{t+1}^s - F_{t+1}\theta_t)^T(G_{t+1}C_{\omega_{t+1}}G_{t+1}^T)^{-1}(\hat{\theta}_{t+1}^s - F_{t+1}\theta_t) + (\theta_t - \hat{\theta}_t)^TC_{\tilde{\theta}_t}^{-1}(\theta_t - \hat{\theta}_t). \tag{5.102}$$

By considering the gradient with respect to θ_t and equating to zero and then applying the matrix inversion lemma the optimal estimator for the smoothing problem can be written in the original form in a backward recursion [174]

$$\begin{aligned} \hat{\theta}_t^s &= \hat{\theta}_t + A_t(\hat{\theta}_{t+1}^s - F_{t+1}\hat{\theta}_t) && (5.103)\\ &= \hat{\theta}_t + A_t(\hat{\theta}_{t+1}^s - \hat{\theta}_{t+1|t}), && (5.104) \end{aligned}$$

where

$$A_t = C_{\tilde{\theta}_t} F_{t+1}^T C_{\tilde{\theta}_{t+1|t}}^{-1} \tag{5.105}$$

for $t = T-1, T-2, \ldots, 1$, which relates the smoothing estimates with the stored filter and predictive distributions as it was indicated from (5.42). For the initialization of the backward steps the filter estimate $\hat{\theta}_T$ is used, i.e. $\hat{\theta}_T^s = \hat{\theta}_T$. For the error covariances, similar recursive procedure can be derived by examining the estimation error $\tilde{\theta}_t^s$ of the smoothed estimates. This results in the recursion [174]

$$C_{\tilde{\theta}_t^s} = C_{\tilde{\theta}_t} + A_t(C_{\tilde{\theta}_{t+1}^s} - C_{\tilde{\theta}_{t+1|t}})A_t^T, \tag{5.106}$$

which completes the solution of the fixed-interval smoothing problem. The computation is initiated by specifying $C_{\tilde{\theta}_T^s}$ for example based on $C_{\tilde{\theta}_T}$.

A Non-Recursive Formulation

In order to obtain a non-recursive formulation for the smoother, we consider the posterior $p(\theta_1, \ldots, \theta_T | Z_T)$ and equation (5.51). Based on the Gaussianity assumptions and the model the smoothing estimator can be obtained by minimizing the functional (after considering the logarithm)

$$l(\theta_1, \ldots, \theta_T) = \sum_{t=1}^{T} \|z_t - H_t\theta_t\|^2_{C_{v_t}^{-1}} + \sum_{t=2}^{T} \|\theta_t - F_t\theta_{t-1}\|^2_{C_{\omega_t}^{-1}} + \|\theta_1 - F_1\eta_{\theta_0}\|^2_{C_{\omega_1'}^{-1}}. \tag{5.107}$$

Where for simplified notation we considered $G_t = I$ or $\omega_t^{\text{new}} = G_t\omega_t$. Also from (5.54) we considered a gaussian density for θ_0 and $C_{\omega_1'} = F_1 C_{\theta_0} F_1^T + C_{\omega_1}$. The smoothing algorithm that we presented is a recursive estimation procedure for the above problem and the connections are also discussed in the original work [174]. Additionally, we can write the above problem in a more suggestive form as a constrained least square problem

$$\min_{\theta_1, \theta_2, \ldots, \theta_T} \sum_{t=1}^{T} \|z_t - H_t\theta_t\|^2_{C_{v_t}^{-1}} \tag{5.108}$$

subject to the constraints or prior information given by the state evolution model and the initial Gaussian density. In order to find an alternative, non-recursive solution, for the problem, we first write the augmented observation model for all the measurements, that is

$$\underbrace{\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_T \end{pmatrix}}_{z} = \underbrace{\begin{pmatrix} H_1 & & & \\ & H_2 & & \\ & & \ddots & \\ & & & H_T \end{pmatrix}}_{H} \underbrace{\begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_T \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \upsilon_1 \\ \upsilon_2 \\ \vdots \\ \upsilon_T \end{pmatrix}}_{\upsilon}. \tag{5.109}$$

The Gauss-Markov estimator or the ML estimator for the augmented model $z = H\theta + \upsilon$ minimizes (5.108) and is given by (4.99), where C_υ is a block diagonal matrix with elements C_{υ_t}. In the spirit of equations (4.69)-(4.73) we form a new model, that takes into account the constraints (see the state-space model), as

$$\begin{pmatrix} 0 \\ z_1 \\ 0 \\ z_2 \\ \vdots \\ z_{T-1} \\ 0 \\ z_T \end{pmatrix} = \begin{pmatrix} -I & & & & & & \\ H_1 & & & & & & \\ F_2 & -I & & & & & \\ & H_2 & & & & & \\ & & \ddots & \ddots & & & \\ & & & & H_{T-1} & & \\ & & & & F_T & -I & \\ & & & & & H_T & \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{T-1} \\ \theta_T \end{pmatrix} + \begin{pmatrix} \omega_1' \\ \upsilon_1 \\ \omega_2 \\ \upsilon_2 \\ \vdots \\ \upsilon_{T-1} \\ \omega_T \\ \upsilon_T \end{pmatrix}, \tag{5.110}$$

where for simplicity we considered $\eta_{\theta_0} = 0$. The smoothing estimates are obtained as ML estimates for the model H' and the extended measurement vector z' from

$$\hat{\theta}^s = (H'^T C_{\upsilon'}^{-1} H')^{-1} H'^T C_{\upsilon'}^{-1} z', \tag{5.111}$$

where

$$C_{\upsilon'} = \begin{pmatrix} F_1 C_{\theta_0} F_1^T + C_{\omega_1} & & & & & & \\ & C_{\upsilon_1} & & & & & \\ & & C_{\omega_2} & & & & \\ & & & \ddots & & & \\ & & & & C_{\omega_T} & & \\ & & & & & C_{\upsilon_T} & \end{pmatrix}. \tag{5.112}$$

By making the rearrangement

$$H' = \begin{pmatrix} H \\ L \end{pmatrix}, \quad C_{\upsilon'} = \begin{pmatrix} C_\upsilon & \\ & C_\omega \end{pmatrix}, \quad z' = \begin{pmatrix} z \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{5.113}$$

where

$$L = \begin{pmatrix} -I & & & & \\ F_2 & -I & & & \\ & F_3 & -I & & \\ & & \ddots & \ddots & \\ & & & F_T & -I \end{pmatrix}, \quad C_\upsilon = \begin{pmatrix} C_{\upsilon_1} & & \\ & \ddots & \\ & & C_{\upsilon_T} \end{pmatrix}, \tag{5.114}$$

and C_ω equivalently, we obtain a Bayesian MS estimator

$$\hat{\theta}^s = (H^T C_\upsilon^{-1} H + L^T C_\omega^{-1} L)^{-1} H^T C_\upsilon^{-1} z. \tag{5.115}$$

In this form, it can be perhaps easier seen that the prior information in the estimation procedure is related to a smoothing density $N(0, (L^T C_\omega^{-1} L)^{-1})$ that relates to the state evolution model. Note, that in the state space model L, C_ω, C_υ are assumed known and exactly of the form (5.114). However, this form for the smoothing estimator is not so useful for practical applications with long measurements. Actually, block Gaussian elimination (an operation that turns out to be equivalent to the Kalman filter forward algorithm), followed by block back substitution (equivalent to the backward smoothing operation) provides a computationally optimal solution for problem, i.e. the smoothing algorithm described earlier [216]. Another issue that can be observed from this formulation is that the initial condition constraint is not necessary for the problem to have solution. We can assume no prior information, and thus remove the first blocks from the matrices L and C_ω.

If we now assume that $C_{\upsilon_t} = \sigma_\upsilon^2 I$ for every t, i.e. time invariant and diagonal, then the estimator becomes

$$\hat{\theta}^s = (H^T H + \sigma_\upsilon^2 L^T C_\omega^{-1} L)^{-1} H^T z, \tag{5.116}$$

from which we can roughly see that the performance of the estimator is influenced only by the ratio $\sigma_\upsilon^2 C_\omega^{-1}$, and not by the exact values. Thus, by setting the variance term $\sigma_\upsilon^2 = 1$, the performance of the estimator can be controlled by, in some sense, the modified state covariances. Another observation for (5.115) is that, since the right term under the inverse comes clearly as a regularizer, the magnitude of the observation model H influence the performance of the regularization. For example, if we consider fixed the regulirizer and we change the observation model with another of the form $H' = cH$.

Let us now consider a useful example. Starting with a random walk model, i.e. $F_t = I$, $H_t = I$, and a scalar case, that is $\theta_t \in \mathbb{R}$, we have the state space model

$$\theta_t = \theta_{t-1} + \omega_t \tag{5.117}$$

$$z_t = \theta_t + \upsilon_t. \tag{5.118}$$

The smoothing estimator optimal in the mean square sense for the parameter vector is given by

$$\hat{\theta}^s = (I + \sigma_\upsilon^2 L^T C_\omega^{-1} L)^{-1} z, \tag{5.119}$$

where now

$$L = \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}, \tag{5.120}$$

and $C_\omega = \mathrm{diag}(\sigma^2_{\omega_1'}, \ldots, \sigma^2_{\omega_T})$, by the uncorrelatedness conditions of the model, and $\sigma^2_{\omega_1'} = \sigma^2_{\omega_1} + \sigma^2_{\theta_0}$. Thus we can write

$$\hat{\theta} = (I + L^T \Phi L)^{-1} z, \tag{5.121}$$

where

$$\Phi = \mathrm{diag}(\phi_1^2, \phi_2^2, \ldots, \phi_T^2), \tag{5.122}$$

and

$$\phi_t^2 = \frac{\sigma_v^2}{\sigma^2_{\omega_t}}. \tag{5.123}$$

This underlines that the smoothing density involves the determination of T parameters. Furthermore, if we set $\Phi = \alpha^2 I$ and if we assume no prior information for the initialization, then we have the smoothing priors regularization method (4.86). The first order difference operator comes in accordance to the first order Markov model (random walk) and the smoothing problem can also be treated deterministically.

5.4 Priors for the state evolution and a state-space identification scheme

We have seen that the state evolution model in Bayesian filtering forms the prior information for estimation in filtering and smoothing problems. The order of the Markov model characterize the predictive densities and the strength of the prior relates to the selection of the state noise and observation noise covariance matrices. Significant is also the role of the state evolution matrices F_t that relate directly to the parametrization of the problem and to other properties of the parameters θ_t. Especially, in estimation problems that there is not a clear and unique a priori parametrization , i.e. the observation model, the choices for H_t and F_t can come in parallel.

For example, there might be prior information outside the context of the time evolution. Indeed, if we were to estimate θ_t based purely on measurements z_t we might have (or want to enforce) prior knowledge in terms of a prior density. The parametrization can, for example, allow to use extra smoothing criteria. This prior information can be embedded in the state evolution model by adding spatial regularization in the observation model [13, 109, 127], that is

$$\begin{bmatrix} z_t \\ 0 \end{bmatrix} = \begin{bmatrix} H_t \\ R \end{bmatrix} \theta_t + \begin{bmatrix} v_t \\ 0 \end{bmatrix}. \tag{5.124}$$

Let us assume that we already have a state-space model for an estimation problem and we want to enforce some extra prior knowledge in terms of a prior density $p_{\rm pr}(\theta_t)$. In terms of the extra prior, for the transition density we have from Bayes' rule ([108], p. 134) that

$$p(\theta_t|\theta_{t-1}) = \frac{p(\theta_{t-1}|\theta_t)p_{\rm pr}(\theta_t)}{p(\theta_{t-1})}, \tag{5.125}$$

where

$$p(\theta_{t-1}) = \int p(\theta_{t-1}|\theta_t)p_{\rm pr}(\theta_t)d\theta_t. \tag{5.126}$$

The density $p(\theta_{t-1}|\theta_t)$ is a back transition density. Let

$$p_{\rm pr}(\theta_t) \propto \exp(-\frac{1}{2}\theta_t^T C_t^{-1}\theta_t) \tag{5.127}$$

be the prior density and let

$$\theta_t = \theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, C_{\omega_t}) \tag{5.128}$$

be the random walk model ($F_t = I$) that we want to enforce. For this model we have

$$p(\theta_{t-1}|\theta_t) \propto \exp(-\frac{1}{2}\|\theta_{t-1} - \theta_t\|^2_{C_{\omega_t}^{-1}}). \tag{5.129}$$

The forward transition density (5.125) becomes

$$p(\theta_t|\theta_{t-1}) \propto \frac{1}{p(\theta_{t-1})}\exp(-\frac{1}{2}(\theta_t - \theta_{t-1})^T C_{\omega_t}^{-1}(\theta_t - \theta_{t-1}) - \frac{1}{2}\theta_t^T C_t^{-1}\theta_t), \tag{5.130}$$

the denominator being the integral of the numerator. From the equation

$$\begin{aligned}(\theta_t - (C_t^{-1} + C_{\omega_t}^{-1})^{-1}C_{\omega_t}^{-1}\theta_{t-1})^T(C_t^{-1} + C_{\omega_t}^{-1})(\theta_t - (C_t^{-1} + C_{\omega_t}^{-1})^{-1}C_{\omega_t}^{-1}\theta_{t-1}) \quad &- \\ \theta_{t-1}^T C_{\omega_t}^{-1}(C_t^{-1} + C_{\omega_t}^{-1})^{-1}C_{\omega_t}^{-1}\theta_{t-1} + \theta_{t-1}^T C_{\omega_t}^{-1}\theta_{t-1} \quad &= \\ (\theta_t - \theta_{t-1})^T C_{\omega_t}^{-1}(\theta_t - \theta_{t-1}) + \theta_t^T C_t^{-1}\theta_t &\end{aligned}$$

we have that (5.130) by considering (5.127-5.129) can be written as

$$p(\theta_t|\theta_{t-1}) \propto \exp(-\frac{1}{2}(\theta_t - A_t C_{\omega_t}^{-1}\theta_{t-1})^T A_t^{-1}(\theta_t - A_t C_{\omega_t}^{-1}\theta_{t-1})), \tag{5.131}$$

where

$$A_t = (C_t^{-1} + C_{\omega_t}^{-1})^{-1}. \tag{5.132}$$

This transition density corresponds to the modified state evolution model, see (5.55), given by (see also [108] p. 135)

$$\theta_t = A_t C_{\omega_t}^{-1}\theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, A_t) \tag{5.133}$$

or

$$\theta_t = (I + C_{\omega_t} C_t^{-1})^{-1} \theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, (I + C_{\omega_t} C_t^{-1})^{-1} C_{\omega_t}). \tag{5.134}$$

Note, that the observation model is the same. For example, if we consider a random walk model $F_t = I$ with $C_{\omega_t} = \sigma_t^2 I$ and prior density with covariance $(D_d^T D_d)^{-1}$, then we obtain the model ([108] p. 135)

$$\theta_t = (I + \sigma_t^2 D_d^T D_d)^{-1} \theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, \sigma_t^2 (I + \sigma_t^2 D_d^T D_d)^{-1}). \tag{5.135}$$

The operator $G = (I + \sigma^2 D_d^T D_d)^{-1}$ is a smoothing operator, thus removing non smoothness of θ_t, while the noise ω_t is distributed according to a Gaussian smoothness density. Of interest is also the prior

$$\theta_t \sim N(0, (D_1^T \Phi_t D_1)^{-1}), \quad \text{where} \quad \Phi_t = \text{diag}(\phi_{1,t}^2, \phi_{2,t}^2, \ldots, \phi_{n,t}^2). \tag{5.136}$$

This has the extra interpretation of an embedded random walk model based smoother that connects individual parameters within a time instant t (section 5.3.3).

We can consider that the prior covariances C_t depend on some unknown parameter vectors ϕ_t. Then we can reformulate the smoothing problem (5.107, 5.110) as

$$\min_{\theta,\phi} \|z' - H'(\phi)\theta\|^2_{C_{\upsilon'}^{-1}(\phi)}, \tag{5.137}$$

where z', H', C'_υ are as in (5.110), and $\theta = (\theta_1^T, \ldots, \theta_T^T)^T$, $\phi = (\phi_1^T, \ldots, \phi_T^T)^T$. Or from (5.108, 5.109)

$$\min_{\theta,\phi} \|z - H\theta\|^2_{C_\upsilon^{-1}}, \tag{5.138}$$

subject to the constraints. For a fixed value of ϕ the value of θ that satisfies the constraints and minimizes (5.138) is given by (5.115)

$$\hat{\theta}^s(\phi) = (H^T C_\upsilon^{-1} H + L(\phi)^T C_\omega(\phi)^{-1} L(\phi))^{-1} H^T C_\upsilon^{-1} z \tag{5.139}$$

$$= A(\phi)^{-1} H^T C_\upsilon^{-1} z. \tag{5.140}$$

Then (5.138) becomes

$$\min_\phi \|(I - HA(\phi)^{-1} H^T C_\upsilon^{-1}) z\|^2_{C_\upsilon^{-1}} = \min_\phi \|z - \hat{s}(\phi)\|^2_{C_\upsilon^{-1}}, \tag{5.141}$$

which is a nonlinear minimization problem of ϕ only and could be solved for example with Gauss-Newton method, see also [68, 21]. The optimal value for ϕ can then be used to find θ.

Solution for the problem (5.138) subject to the constraints can be searched by using a minimization procedure by alternating between minimization of the two sets of variables. Kalman smoother estimates $\hat{\theta}^s = \hat{\theta}^s(\phi)$ can be obtained by the forward-backward smoothing algorithm and locally optimal values for ϕ from the numerical solution of the nonlinear problem in the right of (5.141). By determining an initial estimate for $\phi = \phi^0$, for example the one that gives a random walk model (no prior information), we have the algorithm

- $\phi = \phi^i$ (initialize for some $\phi = \phi^0$)
- compute $\hat{\theta}^s = \hat{\theta}^s(\phi^i)$, i.e. the smoother estimates with the forward-backward smoother algorithm and obtain $\hat{s} = H\hat{\theta}_s$
- estimate the Jacobians J_s^i (numerically) as follows:

  for every parameter ϕ_j ($j = 1, \ldots,$ number of parameters) introduce a small perturbation δ. Then for every j compute Kalman smoother estimates $\hat{\theta}^s(\phi^i + \delta e_j)$, where $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T$, i.e a vector whose j-th component is only non zero and is equal to 1. Approximate the j-th column of J_s^i with

$$J_s^i(j) = \delta^{-1} \left(\hat{s}(\phi^i + \delta e_j) - \hat{s}(\phi^i) \right) \tag{5.142}$$

- compute Gauss-Newton direction (or some other gradient based direction)

$$d^i = (J_s^{iT} W J_s^i)^{-1} J_s^{iT} W (z - \hat{s}(\phi^i)) \tag{5.143}$$

- compute optimal step size a_i, for example, with backtracking line search (2.49) by computing new Kalman smoother estimates until improvement with ($\phi = \phi^i + a d^i$)
- update

$$\phi^{i+1} = \phi^i + a_i d^i \tag{5.144}$$

- repeat until convergence to obtain $\hat{\phi}$ and for the estimated state-space model optimal mean square parameters $\hat{\theta}^s$.

CHAPTER VI

Independent Component Analysis

A significant problem in statistics and related areas is the identification of a suitable representation or transformation of a multivariate data set. The aim is that its essential structure is made more visible and more accessible for further investigation and analysis. *Independent component analysis (ICA)* is a class of decomposition methods in which the desired representation is the one that minimizes the statistical dependencies of higher order between the components of the representation. The concept of ICA may be seen as an extension of *principal component analysis (PCA)*, which impose independence up to the second order by using the information contained in the covariance matrix of the data. ICA is related with the problem of *blind signal separation (BSS)*. BSS consists in recovering unobserved signals or sources from several observed mixtures of them by using minimal prior information about the mixing system and by taking advantage of different statistical properties of the sources. Main references of the chapter are [85, 39, 103].

6.1 Basic concepts and definitions

Consider T realizations of the random vector $x = (x_1, x_2, \ldots, x_n)^T$. These can be summarized in a matrix notation as

$$X = \begin{pmatrix} x_1(1) & x_1(2) & \cdots & x_1(T) \\ x_2(1) & x_2(2) & \cdots & x_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ x_n(1) & x_n(2) & \cdots & x_n(T) \end{pmatrix}. \tag{6.1}$$

The matrix X can be considered to represent a cluster of T points in the vector space $\mathbb{R}^n$. If these points are represented by the use of the common orthonormal base $\mathbf{e} = \{e_1, e_2, \ldots, e_n\}$ of $\mathbb{R}^n$, then the values of the data are the coordinates of the points, i.e. $x(t) = e_1x_1(t)+e_2x_2(t)+\cdots+e_nx_n(t)$, where $e_i = (0, \ldots, 1, \ldots, 0)^T$. The data points could be represented through another base $\mathbf{u} = \{u_1, u_2, \ldots, u_n\}$ (coordinate system) not necessary orthogonal, as

$$x(t) = u_1y_1(t) + u_2y_2(t) + \cdots + u_ny_n(t), \tag{6.2}$$

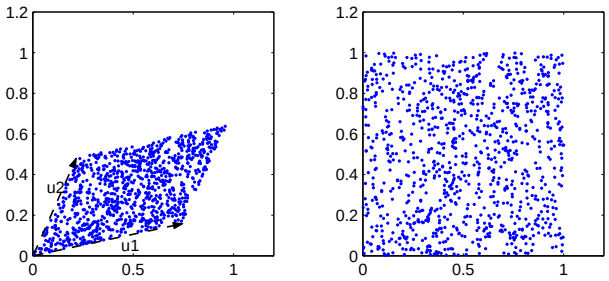

Figure 6.1: An example of a different representation of two dimensional data. Under the representation u_1, u_2 the new variables (right) are independent.

where $y_i(t)$ are the new coordinates. So, what is under consideration, is the identification of a better representation or transformation $y = Bx$ that reveals some information that is hidden into the data set. The representation could be based on the statistical properties of the resulted components, see Fig. 6.1.

A suitable representation of multivariate data can be searched through the concept of independence [85]. Thus, given a set of T realizations of a random vector x, ICA consists of estimating a new coordinate system (6.2) to decompose x, so that the resulting components y are as statistically independent as possible. If there is an accurate measure of independence, other considerations are not necessary for the identification of a new representation for every random vector. In that spirit, ICA can be understood as as extension of PCA, in which the independence condition is imposed up to uncorrelatedness.

The problem of finding a suitable representation of a set of observations could also be seen from a different point of view. Consider that some physical sources generate a set of continuous signals $s_j(t)$. These source signals can not be observed directly, but we observe an instantaneous mixture of them. Thus, at each time point t we observe several signals $x_i(t)$ such us

$$x_i(t) = \sum_j a_{ij} s_j(t). \tag{6.3}$$

In general, the problem of BSS consists of recovering the unobserved source signals $s_j(t)$ by observing only the signals $x_i(t)$, i.e. without information about the mixing system [39]. In other words, if the mixing matrix A is assumed square and non-singular, then we are looking for a matrix $B = A^{-1}$ to recover $s(t) = Bx(t)$. In practice, the mixed signals are observed at different time instances $t = 1, \ldots, T$. If we assume that the signals $x_i(t)$ are the realizations of a random vector x, then the problem of blind source separation is related to the problem of defining a suitable transformation to the observations. Formally, what is under consideration is the estimation of both the matrix B and the source signals $s_i(t)$. It turns out that such a problem is not in general uniquely defined and the problem can be reduced

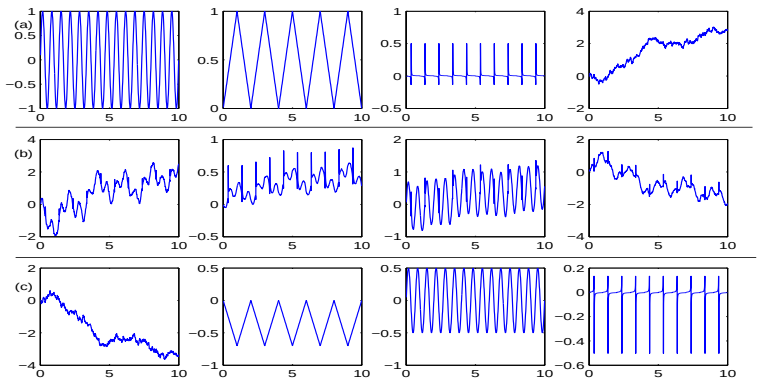

Figure 6.2: An illustration of source separation: (a) the original source signals, (b) the observed linear mixtures, (c)the separated signal, where the separation could be achieved up to permutation and scale change.

up to the estimation of a matrix $B' = SPB$, where P is a permutation matrix and S a scale matrix (non-singular diagonal matrix), see Fig. 6.2.

The lack of a priori knowledge for signal separation can be compensated by a statistically strong, but often plausible, assumption of independence between the source signals [29, 85, 39]. In this case, given a set of T realizations of a random vector $x = (x_1, x_2, \ldots, x_n)^T$, such as $x = As$ is a mixture of independent sources, where the mixing matrix A is assumed non-singular and the sources $s = (s_1, s_2, \ldots, s_n)^T$ are unknown, ICA-BSS consists of estimating a matrix B such as $y = Bx$, where $y_1, y_2, \ldots, y_n$ are as independent as possible and $y = SPs$. It turns out that the identification of a non-orthogonal transformation so that the transformed variables are as independent as possible can solve the problem of BSS, under some weak conditions and up to permutation and scale change, when the sources are independent and at most one source is Gaussian distributed [45]. For an intuitive illustration of the ICA-BSS problem see also Fig. 6.3. Some extensions related to identifiability and separability for ICA-BSS problems can be found in [27, 56, 195, 196].

6.2 Principal component analysis

PCA is a linear transformation that transforms the data to a new orthogonal coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest on the second coordinate and so on. PCA can be used for dimensionality reduction in a dataset by retaining those characteristics in the dataset that contribute most to its variance.

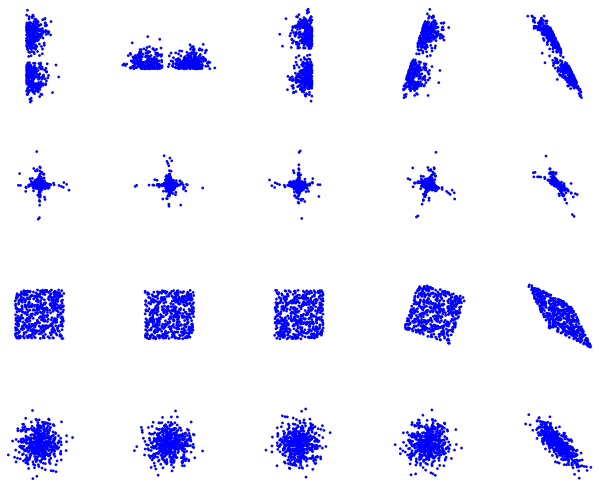

Figure 6.3: Examples of sample distributions of two random variables (four pairs) for five different transformation matrices. From left to right: the identity transform, a permutation, sign change, a rotation, and a generic linear transform.

6.2.1 Principal components

Consider T independent realizations of a random vector $x = (x_1, x_2, \ldots, x_n)^T$, summarized in a matrix X. Every column of X represents a point of $\mathbb{R}^n$. Without loss of generality we can assume that the vector is zero mean, because the vector $x' = x - \eta_x$ is a zero mean random vector and the mean can be estimated from the data. Let y_1 be a linear combination of the elements of x, i.e.

$$y_1 = u^T x, \tag{6.4}$$

for which we search a vector u, such that the variance the variance of y_1 is maximized. It is clear that the maximum will not be achieved for finite u unless a normalization constraint is imposed. A convenient constraint is $\|u\| = 1$ or $u^T u = 1$. Thus, we have the optimization problem

$$u_1 = \arg\max E\{(u^T x)^2\} = \arg\max(u^T E\{xx^T\}u), \tag{6.5}$$

subject to the constraint $u^T u - 1 = 0$. By considering the Lagrangian and taking the gradient with respect to u and setting to zero we have for the optimality condition

$$(C_x - \lambda I)u_1 = 0, \tag{6.6}$$

where λ is the Lagrange multiplier. Equation (6.6) suggests that u_1 is an eigenvector of the matrix C_x and that the optimal Lagrange multiplier is the corresponding eigenvalue. For that selection, we have that the variance becomes

$E\{y_1^2\} = u_1^T C_x u_1 = \lambda$, and since it should be as large as possible we have that $\lambda = \lambda_1$ is the largest eigenvalue of C_x and u_1 is the corresponding eigenvector. The second principal component is defined by the maximization $u^T C_x u$ subject to being uncorrelated to $u_1^T x$ and $u^T u = 1$. The solution is the eigenvector corresponding to the second largest eigenvalue and accordingly the m-th principal component is given by selecting u_m to be the m-th dominant eigenvector [103]. Note, that from the eigenvalue decomposition of $C_x = U_n D U_n^T$, where $U_n = (u_1, u_2, \ldots, u_n)$ we have that

$$y = U_n^T x \tag{6.7}$$

is a vector with uncorrelated components.

6.2.2 Mean square error compression

Principal component analysis can be seen in connection to the mean square error. We search for a set of m n-dimensional orthonormal basis vectors $u_1, u_2, \ldots, u_m$, spanning a m dimensional subspace, such that the mean square error between the random vector x and its projection on the subspace is minimal. Thus for the projection

$$\hat{x} = U_m (U_m^T U_m)^{-1} U_m^T x = U_m U_m^T x, \tag{6.8}$$

and for the mean square error

$$\begin{aligned} E\{\|x - \hat{x}\|^2\} &= E\{\|x\|^2\} - E\{\|\hat{x}\|^2\} & (6.9)\\ &= \text{trace}(C_x) - \sum_{i=1}^{m} u_i^T C_x u_i, & (6.10) \end{aligned}$$

it can be shown that the minimum of (6.9) is given by any orthonormal basis of the PCA subspace spanned by the first m eigenvectors [53]. However, the criterion does not specify the basis of this subspace at all. Any orthonormal basis of the subspace will give the same optimal compression. More general, consider in equation (6.8) the matrix $A_m^T = T U_m^T$, where T is not necessarily orthogonal but invertible.

In practice, principal component analysis can be done by using the eigenvalue decomposition of the data covariance matrix. The data correlation matrix can also be used to extract eigenvectors [153]. Then the eigenvectors will also reflect the mean. This approach is analogous to the one using covariance matrix with the exception that the quadratic $E\{y_k^2\}$ is maximized to define PCs instead of the variance.

6.2.3 Whitening

A random vector $z = (z_1, z_2, \ldots, z_n)^T$ is said to be white if its elements are uncorrelated and have unit variances ([39], p. 130). Since the eigenvalues of the matrix C_x contained in the diagonal matrix D correspond to the variances of the transformed variables $y = U^T x$, clearly the transformation

$$z = D^{-1/2} U^T x = V x, \tag{6.11}$$

where $D^{-1/2} = \text{diag}(\lambda_1^{-1/2}, \ldots, \lambda_n^{-1/2})$ transforms x to a white vector, i.e.

$$E\{zz^T\} = D^{-1/2}U^T E\{xx^T\}UD^{-1/2} = D^{-1/2}U^T UDU^T UD^{-1/2} = I. \quad (6.12)$$

If z is a white random vector, then the vector $z' = Az$, where A is an orthogonal matrix, is also white since

$$E\{z'z'^T\} = AE\{zz^T\}A^T = AIA^T = I. \quad (6.13)$$

So whiteness or uncorrelatedness cannot solve the BSS problem because we cannot say if the independent components are given from z or z'.

Whitening can transform the original mixing problem $x = Ay$ into a new one

$$z = Vx = VAy = \bar{A}y. \quad (6.14)$$

Since, as we have discussed, the exact variances of the sources cannot be identified without prior information for example about the mixing system, we can consider that y is white as well (having independent components with unit variances) and then we have

$$I = E\{zz^T\} = \bar{A}E\{yy^T\}\bar{A}^T = \bar{A}\bar{A}^T. \quad (6.15)$$

So the matrix $\bar{A}$ is an orthogonal matrix. This means that the search for the mixing matrix can be restricted to the space of orthogonal matrices and

$$y = \bar{A}^{-1}z = \bar{A}^{-1}Vx \quad (6.16)$$

or

$$A^{-1} = \bar{A}^{-1}V. \quad (6.17)$$

Note that the definitions of ICA given up to now imply no ordering of the independent components which is in contrast to PCA. However, it is possible to introduce some ordering between the independent components.

Blind source separation can be based on independence, but independence can not be reduced to the simple decorrelation condition. However, second order information (correlations between the variables and decorrelation) can be used to reduce the ICA problem to a simpler form [85]. In the first column of Fig. 6.4 three different sample distributions from two subgaussian, two supergaussian and two Gaussian independent random variables are plotted. The second column shows the sample distributions after a linear mixture of them. The third column shows the PCA representation of the mixed variables and the last a whitening transform. The arrows show the directions where the kurtosis of every component is maximized. Clearly, in the first two cases after the whitening transform the recovering of the original shape of the sample distribution is simplified to a simple rotation. Though, information about the nature of sources, i.e. sub-/super-Gaussian is required for separation. After whitening, for the Gaussian variables it is not possible any further rotation due two the total symmetry of the distribution. This is because uncorrelated Gaussian variables are also independent. Considering that any orthogonal linear transform of uncorrelated Gaussian variables does not influence

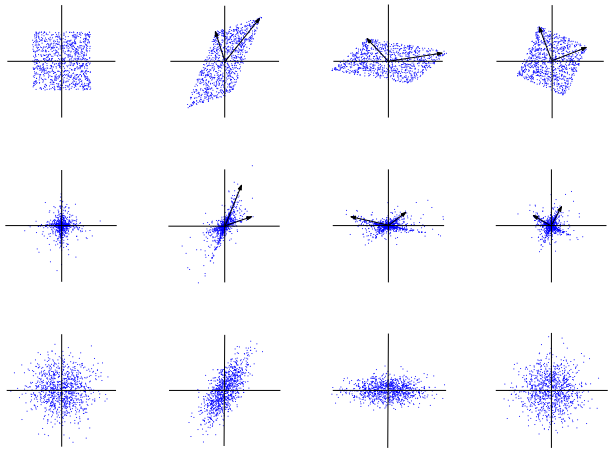

Figure 6.4: Different two dimensional sample distributions.

the joint probability density function it is clear that ICA for jointly distributed Gaussian variables reduces to a whitening procedure. Considering now the problem of blind source separation it is clear that ICA for BSS requires the assumption about non-Gaussian source variables (at most one is allowed [45]).

6.3 ICA by maximum likelihood identification

ICA relates to the identification of a matrix that creates independence between the transformed components. Independence is a theoretical concept, and thus the identification of a suitable measure of independence or an objective function that leads to independence up to some level is common to all different approaches for ICA [45, 29, 83, 88, 130, 37]. Optimization of such an objective is then required. So an ICA estimation method can be seen as a two stage procedure. First, the identification of the suitable objective and then the implementation of a suitable algorithm that optimizes it. Most of the ICA-BSS methodological directions can be found in [39, 85, 129, 67], but see also [89, 90, 91, 92, 93, 94] for different applications. Different algorithms for use can be found in [95, 96, 97, 98].

6.3.1 Bayesian formulation of the problem

In order to investigate the assumptions related to the basic ICA model, we shall start with a more general model and define the solution in a Bayesian framework. Therefore, let us consider the additive noise model for ICA

$$x = As + \upsilon, \tag{6.18}$$

where s is a vector with independent components. For the matrix A we do not have to assume anything yet, for example, about its rank. We search estimates

for both s and A. Then from Bayes rule the posterior density is

$$p(A, s|x) = \frac{p(x|A, s)p(A, s)}{p(x)}. \tag{6.19}$$

The source signals must be independent of the mixing system so we can write

$$p(A, s|x) \propto p(x|A, s)p(A)p(s). \tag{6.20}$$

From the model (6.18) we can also write for the likelihood of the observed x

$$p(x|A, s) = p_\upsilon(x - As|A, s) \tag{6.21}$$

for some noise distribution. The sources are assumed independent so we have the following general form for estimation

$$\begin{aligned} p(A, s|x) &\propto p_\upsilon(x - As|A, s)p(A)p(s) &(6.22)\\ &\propto p_\upsilon(x - As|A, s)p(A)p_1(s_1), \cdots, p_m(s_m), &(6.23)\end{aligned}$$

where $p_i(s_i)$ are prior densities for the random variables s_i. We can try to reduce the number of parameters for the estimation problem (6.22). In the sense that, if we knew some value for A, we could use it to estimate s. This can be done by marginalizing over all possible values of s

$$\begin{aligned} p(A|x) &= \int p(A, s|x)ds &(6.24)\\ &\propto \int p(x|A, s)p(A)p(s)ds &(6.25)\\ &\propto p(A)\int p(x|A, s)p(s)ds. &(6.26)\end{aligned}$$

Correspondingly for s it is

$$p(s|x) = \int p(A, s|x)dA = p(s)\int p(x|A, s)p(A)dA, \tag{6.27}$$

with the assumption of independence (6.20) and the model (6.21). Alternatively, if we had an $\hat{A}$ from the model (6.18) it is

$$p(s|x) \propto p_\upsilon(x - \hat{A}s|s)p(s). \tag{6.28}$$

Furthermore, we observe T realizations of a random vector $x = (x_1, x_2, \ldots, x_n)^T$, summarized in the matrix X. Therefore, we have the model

$$X = AS + \Upsilon. \tag{6.29}$$

If we assume $\upsilon(t)$ independent and identically distributed, the likelihood becomes

$$p(X|A, S) = \prod_{t=1}^{T} p_\upsilon(x(t) - As(t)|A, s(t)). \tag{6.30}$$

6.3.2 The likelihood of the ICA model

Until now, we have a general description of the estimation problem of ICA with minimum amount of assumptions. Now we reduce the problem to the basic ICA model

$$x = As, \tag{6.31}$$

where A is further considered to be an invertible square matrix. For no noise in the model, we can model the likelihood densities (6.21) with a δ function, i.e.

$$p_v(x - As|A, s) = \delta(x - As) \tag{6.32}$$

which is one when the model holds and zero otherwise. Then, (6.26) becomes

$$p(A|x) \propto p(A) \int \delta(x - As)p(s)ds. \tag{6.33}$$

By introducing a change of variables $\xi = x - As$ or by considering the model and (3.15) this becomes [125] (see also [85], p. 203)

$$p(A|x) \propto p(A)\frac{1}{|\det A|}p_s(A^{-1}x). \tag{6.34}$$

Clearly, if we have a value for $\hat{A}$, then we can obtain estimate for s as

$$y = \hat{A}^{-1}x = Bx. \tag{6.35}$$

Furthermore, we can consider the source distributions time invariant. So the vectors $s(t), t = 1, \ldots, T$ are assumed identically distributed and have independent components. Then the prior joint density of S becomes

$$p(S) = p_s(s(1)) \ldots p_s(s(T)) = \prod_{t=1}^{T}\prod_{i=1}^{n} p_i(s_i(t)). \tag{6.36}$$

Substituting the result to (6.34) and considering the logarithm we have

$$\log p(A|X) = C + \log p(A) - T \log |\det A| + \sum_{t=1}^{T}\sum_{i=1}^{n} \log p_i(b_i^T x(t)), \tag{6.37}$$

where $A^{-1} = B = (b_1, \ldots, b_n)^T$. By using that posterior, the optimal mixing matrix A can be searched by considering the gradient with respect to A. Note that the posterior $p(A^{-1}|x)$ is not the same with the posterior $p(A|x)$ [125]. We can assign a uniform prior distribution for $A = (a_{ij})$, for example $p(A) = c$, if $a_{min} \leq a_{ij} \leq a_{max}$ and zero elsewhere to denote lack of prior information.

The maximum likelihood method for ICA defines an estimator for the mixing matrix $B = A^{-1}$ from the following functional ([85], p. 204)

$$\log L(B) = \log p(X; B) = T \log |\det B| + \sum_{t=1}^{T}\sum_{i=1}^{n} \log p_i(b_i^T x(t)) \tag{6.38}$$

or from

$$l_T(B) = \frac{1}{T}\log L(B) = \log|\det B| + \bar{E}\{\sum_{i=1}^{n}\log p_i(b_i^T x(t))\}, \tag{6.39}$$

where $\bar{E}$ denotes average. We can also write that

$$l_T(B) \xrightarrow[T\to\infty]{} E\{\sum_{i=1}^{n}\log p_i(b_i^T x)\} + \log|\det B| \tag{6.40}$$

Considering the densities of the independent components as exactly known and B as nonrandom, maximization of (6.39) gives the maximum likelihood estimator for B. Though in many real data applications the densities p_i are unknown. It is then preferable to consider them as prior model densities as in the Bayesian formulation. Note, that with Gaussian prior densities and orthogonality constraints we obtain the PCA criterion.

6.3.3 Gradient optimization methods

The gradient of $l_T(B)$ with respect to B is given by ([85], p. 207)

$$\frac{\partial l_T(B)}{\partial B} = (B^T)^{-1} + \bar{E}\{g(Bx)x^T\} \tag{6.41}$$

and defines a gradient ascent direction for optimization. The function $g(y) = (g_1(y_1), g_2(y_2), \ldots, g_n(y_n))^T$ is given from

$$g_i(y_i) = \frac{\partial}{\partial y_i}\log p_i(y_i) = \frac{p_i'(y_i)}{p_i(y_i)}, \tag{6.42}$$

and it is based on the prior densities g_i. For the gradient of $\log|\det B|$ see section 2.6.1 and (2.80). From (2.82) we have

$$\begin{aligned}
\frac{\partial}{\partial\tau}\bar{E}\{\sum_{i=1}^{n}\log p_i(b_i^T x + \tau d_i^T x)\}|_{\tau=0} &= \bar{E}\{\sum_{i=1}^{n} g_i(b_i^T x)d_i^T x\} && (6.43)\\
&= \bar{E}\{\mathrm{trace}(x\sum_{i=1}^{n} g_i(b_i^T x)d_i^T)\} && (6.44)\\
&= \mathrm{trace}((\bar{E}\{g(Bx)x^T\})^T D) && (6.45)
\end{aligned}$$

From (2.83) we have the gradient. At every iteration estimates for the independent components are obtained from the current value of B as $y = Bx$. This algorithm is also referred as infomax derived through information theory criteria [15].

An algorithm that avoids matrix inversions and has better properties is obtained by the natural gradient, which becomes (2.95)

$$D_{nat} = ((B^T)^{-1} + \bar{E}\{g(Bx)x^T\})B^T B = (I + \bar{E}\{g(y)y^T\})B. \tag{6.46}$$

See section 2.6, and [6, 4, 5, 1]. The algorithm has been independently proposed also in [33], though a similar algorithm was also derived in [43]. In order to resolve the indeterminacy of scales, the basic natural gradient algorithms impose some constraints on the magnitudes of the recovered signals, for example $E\{g_i(y_i)y_i\} = 1$. A least in terms of better stability the following direction is proposed ([39], p. 238)

$$D = \left(\bar{E}\{g(y)y^T\} - \mathrm{diag}(\bar{E}\{g_i(y_i)y_i\})\right) B, \tag{6.47}$$

where $\mathrm{diag}(\bar{E}\{g_i(y_i)y_i\})$ is a diagonal matrix with elements $\bar{E}\{g_i(y_i)y_i\}$ for $i = 1, \ldots, n$ in the main diagonal, see also [3]. For prewhitened data, the gradient of the functional (6.39) is not any more given by (6.41) since the matrix B is restricted to be orthogonal. Then it can be shown that the natural gradient direction that takes approximately into account the orthogonality constraint (up to first order Taylor approximation) is given by ([1, 33], see also [39], p. 239)

$$D = \left(\bar{E}\{g(y)y^T\} - \bar{E}\{yg(y)^T\}\right) B. \tag{6.48}$$

Yet another algorithm, derived as a fixed point iteration and having characteristics of Newton's method uses the direction [82, 88]

$$D = \mathrm{diag}\left(\frac{1}{\bar{E}\{g_i(y_i)y_i\} - E\{g'(y_i)\}}\right)\left(\bar{E}\{g(y)y^T\} - \mathrm{diag}(\bar{E}\{g_i(y_i)y_i\})\right) B, \tag{6.49}$$

where at every step the matrix B is projected on the set of whitening matrices by

$$B' = (BC_xB^T)^{-\frac{1}{2}}B. \tag{6.50}$$

The algorithm is referred as FastICA for ML estimation ([85], p. 210).

The prior densities $p_i(s_i)$ are unknown for different real data applications. Therefore, they could be also parametrized, i.e. $p_i = p_i(s_i; \phi_i)$. We could have included extra parameters for the total posterior in the Bayesian methodology. Though, instead of optimizing over these parameters they can be replaced with some estimates $\hat{\phi}$. Many ICA methods use such estimates, usually, obtained by the method of moments. If the prior densities are assumed smooth, unimodal and symmetric, a characterization of the density is obtained from the kurtosis, to differentiate between super and sub-Gaussian densities [66, 40, 131]. More general, a parametric density family is considered and estimates for the parameters are obtained in every iteration [57, 36, 215, 119, 120, 121]. In the following it is sketched a general algorithm for the basic ICA model

- center the data to make its mean zero (it also possible to consider prewhitening possibly combined with PCA dimension reduction)
- set $B = B^j$ (for some $B = B^0$)
- compute new estimates $Y = B^jX$
- find new estimates for the parametrization of the prior densities $\hat{\phi}_i = \hat{\phi}_i(Y_i)$ and $g_i = g_i(y_i; \hat{\phi}_i)$ from (6.42) (note, that only the nonlinear functions (6.42) are needed and not the exact densities, and the variances are unknown)

- compute new direction $D^j = D^j(B^j, Y, \hat{\phi})$
- update
$$B^{j+1} = B^j + a_j D^j \tag{6.51}$$
- repeat until convergence to obtain B and $Y = BX$.

A simple parametric density model related to the hyperbolic-Cauchy distribution for distinguishing between supergaussian and subgaussian sources respectively is given by [85, 39]

$$g_i(y_i) = -2\tanh(y_i) \tag{6.52}$$
$$g_i(y_i) = \tanh(y_i) - y_i. \tag{6.53}$$

Another possibility is $-y_i^3$ for subgaussian components. The selection between the two densities can be made based on the sign of the kurtosis of the estimated transformed variables. The choice between (6.52) and (6.53) can be also made based on the sign of the non polynomial moment [85]

$$E\{-\tanh(y_i)y_i + (1 - \tanh(y_i)^2)\}. \tag{6.54}$$

When the algorithm (6.49), (6.50) is used, then the nonlinear function $g_i(y_i) = \tanh(y_i)$ can be selected for all the components. This is possible because in the algorithm there is embedded information about the nature (sub- or supergaussian) of the components ([85], p. 211).

6.4 Connection with other estimation principles

Source separation methods have emerged from the field of neural networks and they were related to the concept of nonlinear decorrelation, which leads to independence up to some level [76, 107, 46, 186, 42, 44, 43], see also equation (3.58). Note that any nonlinear function introduces, through its Taylor series expansion, higher order statistical properties of the underlined variables, see also equations (3.60) and (3.44). Other related methods and generalizations are given by the estimating functions approach [4, 39], see also [33, 155, 5, 2]. Some related concepts are also given by nonlinear PCA methods [113, 114, 154, 115, 155, 159]. The interpretation of nonlinear PCA criteria as maximum likelihood estimation has been presented in [115], see also [85].

Early approaches based on maximum likelihood include [62, 167, 164]. An estimation principle for ICA that is very closely related to maximum likelihood formalism is the infomax principle [15]. This approach is based on maximizing the output entropy, or information flow, of a neural network with nonlinear outputs. If the related nonlinear functions are chosen to be of the form (6.42), then the criterion becomes equivalent to maximum likelihood method [163, 28, 152]. Then Bayesian interpretations and generalizations followed [123, 176, 124, 146, 179], see also [12, 58, 35, 203, 79, 177]. For modeling prior information about the mixing matrix see [86, 125, 201]. A distinct advantage of the Bayesian formulation is that

it breaks the problem into three parts, i.e. the signal model, the cost function and the optimization algorithm.

Besides entropy maximization other information theoretic concepts have been applied for ICA-BSS. Mutual information is a measure of independence (see section 3.6). Therefore, minimization of mutual information gives independent components, see [45] where an approximation based on cumulants was considered together with the definition of ICA through an appropriate function that measures independence. Another measure of independence is given by negentropy (3.106) [45]. It turns out that independent components y_i are those that their marginal negentropy is maximized or those that are maximally far from being Gaussian [85]. Approximations of negentropy and other measures of nongaussianity (for example kurtosis) enable a deflationary approach, i.e. one-by-one estimation of the independent components, by searching maxima of nongaussianity of a single projection $b^T x$ [87, 80, 81, 82, 88], see also [39] for other related algorithms. The connections between all the information theory criteria for ICA-BSS as well as maximum likelihood identification are established through the definition of Kullback-Leibler divergence (3.75) [45, 28, 29, 152, 36, 37]. In fact the relevant assumptions are better understood and generalized as with the Bayesian formulation.

The source separation methods described until now exploit primarily spatial heterogeneity and homogeneity, i.e. investigating structure across the sensors and not across time. The consequence of ignoring any time structure is that information contained in the data is represented solely by the sample distributions of the observations. In many applications, this is enough to separate sources from mixtures even for non stationary source signals. However, signals with Gaussian or elliptical contoured joint sample distributions, for example Gaussian autoregressive stochastic processes, can not be separated by those methods. Also mixtures of sinusoidal signals, that have bimodal amplitude histograms, are better separated by other BSS methods that exploit time structure. These limitations hold also for some algebraic methods for ICA [32], which impose some level of independence by approximately canceling higher order correlations estimated by cross-cumulants (3.51-3.54), see also [30, 212, 39] for related methods. From the methods using time structure of the signals, some assume stationarity and use time delayed cross-correlation matrices for separation, see also equations (3.25), (3.113) and (3.114). By canceling time correlations between signals source separation can be achieved [198, 147, 17, 217], see also [39]. Note that those algorithms are not strictly speaking ICA methods since they exploit second order information for stochastic processes. However, they can separate, for example, time correlated Gaussian processes when they have distinct normalized autocorrelations (3.115) [166]. Finally, some approaches can take into account the non-stationarity of the signals for separation, for example [143, 16, 165], see also [67, 31, 39, 162, 79, 84].

Chapter VII

Estimation of EPs

In this chapter, different characteristics of electric brain signals are briefly described and difficulties arising from the complexity of EEG measurements and evoked potentials are discussed. The applicability of ICA-BSS methods for EEG separation and analysis is also considered. An application of ICA for BSS of EEG is presented, aiming to underline different characteristics of EEG signals. The artifact corrected EEG measurements are used in chapter VIII, aiming to show how the EP signal subspace is affected by strong disturbances. Some single-channel EP estimation methods are also discussed, in relation to chapters IV and V. Finally, dynamical estimation methods, and some noise considerations for EP modeling are presented. More about the origin of brain potentials, research methods, and applications can be found, for example, in [151, 168, 26, 8, 52, 209, 140].

7.1 Basic concepts and definitions

Many anatomical and functional brain imaging methods have been developed to study the brain noninvasively, and is now possible to collect vast amounts of data from the living human brain. Structural information can be reached by means of static pictures of living tissues through, for example, computer-assisted x-ray tomography (CT) and magnetic resonance imaging (MRI). Another class of methods provides information about the function of the brain with limited time resolution. This includes single-photon-emission computed tomography (SPECT), positron-emission tomography (PET), and functional MRI (fMRI). Noninvasive methods that provide information about the neural dynamics on a millisecond scale are EEG and magnetoencephalography (MEG). They measure effects of brain electrical activity. Disadvantage of these methods is that spatial resolution is usually poor, due to noise contamination, instrumentation, and modeling issues. EEG is measured using electrodes on the scalp. For MEG recordings superconducting quantum interference devices (SQUIDs) are used. The MEG measurements need to be carried out inside a magnetically shielded room. Therefore, instrumentation needed for MEG is more sophisticated [72].

The electric activity of the brain consists of ionic currents generated by biochemical reactions at the cellular level. The human brain consists of about 10^{10}

neurons in the outermost layer, i.e. the cerebral cortex. The neurons are the basic information processing units. EEG mainly represents the summation of synchronized activity of different groups of cortical neurons, though, sub-cortical areas, such as thalamus, may also contribute [8]. For the estimated levels of conductivity of brain and surrounding tissues the propagation velocity is very high, so that, practically, changes in activity may be detected simultaneously on the scalp. Additionally, this important statement leads to the conclusion that no charge is accumulated at any time in the brain and the potentials recorded at a given point on the scalp are the summation of potentials induced by several generators [151, 140].

Both in EEG and MEG, the measured signals are generated by changes of synchronized neural activity in the brain. The detection of temporal and spatial changes in active cortical areas gives the means to investigate higher mental functions. Of importance are also the geometry and conductivity of different parts of the head. It can be discussed that EEG is more sensitive to the morphology of the head. However, the use of realistic geometries based on MRI seems to be necessary for both methods for accurate modeling [156, 211].

A significant problem in EEG and MEG analysis is related to the fact that brain signals are seriously contaminated by noise. Strong electro-physiological signals are generated by the heart and by different skeletal muscles, and they may contaminate the measurements. Eye blinks and movements are also creating significant artifacts. Additionally, different brain areas are under continuous activation at any given time by handling many body or mental functions, and by processing information from the environment. In that sense, MEG and EEG analysis aims to investigate some meaningful signals buried into artifacts, ongoing brain activity, and other disturbances. Separation, estimation and denoising methods are necessary for the investigation of complex mental processes.

7.2 Electroencephalography

It is well know that the characteristics of EEG change in many situations, for example, with the level of vigilance (alertness, drowsiness, rest, sleep and dreaming), anxiety, and emotional tension. EEG is also used for the detection of abnormalities, such us epileptic spikes (spike wave complexes of 3Hz), different brain diseases, injuries and damages, and to monitor surgery recovery. Particular mental tasks also alter the pattern of the waves. Finally, EEG has been used to study psychiatric disorders (dementia, schizophrenia, mood and personality disorders). More about EEG research methods and applications can be found in [151].

7.2.1 EEG measurements

EEG recordings are usually taken by placing electrodes on the scalp, and the electrode skin interface is filled with conductive gel. Brain potentials can be also recorded through depth electrodes implanted in the brain. The electrode placement should conform the international 10-20 system [52] shown in Fig. 7.1. Usually, as a first preprocessing step, EEG signals are bandpass filtered in the interval

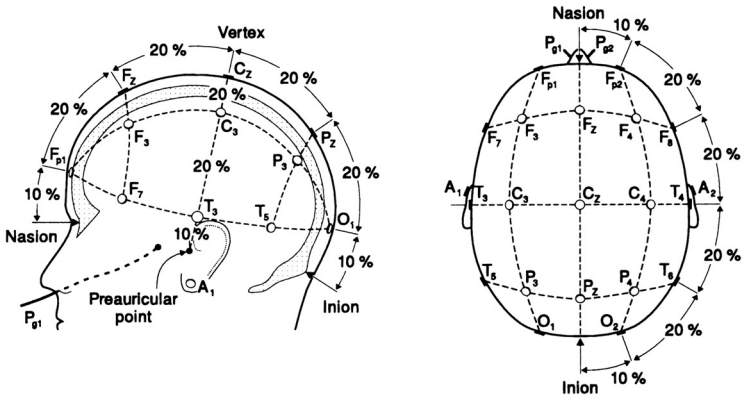

Figure 7.1: International 10-20 EEG electrode system (redrawn from [140]). A = ear lobe, C = central, Pg = nasopharyngal, P = parietal, F = frontal, Fp = frontal polar, and O = occipital.

around 0.05Hz to 70Hz, which is the band where significant brain activity could be observed. Though, narrower bands are often used for analysis.

The amplitude of EEG signals is of a few microvolts and changes stochastically. Different frequency bands of EEG have been empirically labeled with names such as alpha (8-13 Hz), beta (13-30 Hz), theta (4-8 Hz), and delta (0.5-4 Hz). Activity in a given frequency band may have time-varying characteristics (e.g. [190]), and different physiological origin depending on the particular situation (e.g. [150]).

7.2.2 Evoked potentials

Evoked potentials reflect changes of brain electric activity due to external (physical world), or internal (triggering mental process) stimulation of the central nervous system. EPs are observed as EEG epochs, time locked to a stimulus or event, the timing of which can be reliably assessed [52] (see Fig. 7.2). Significant advantage of EP research is the fact that cortical reactivity and function can be assessed with high temporal resolution. Therefore, they are used to study changes of brain function, for example levels of sedation [213], and to explain cognitive processes such as memory [151].

EPs can be divided into two categories: exogenous and endogenous. Exogenous potentials are determined by the type of stimulation (e.g. auditory, somatosensory, visual), the neurological path, and by physical characteristics of the stimulus (for example frequency, intensity, duration for auditory stimulation). In the acoustic modality exogenous components comprise the acoustic brain stem evoked potentials and the mid-latency components. So exogenous potentials depend on the

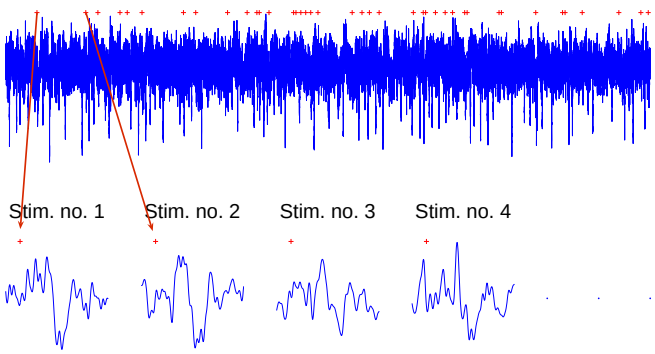

Figure 7.2: EPs during auditory stimulation. Upper part EEG measurements, lower part EEG epochs of 700ms length, crosses indicate stimulation time.

sensory system and general brain condition on receiving and decoding stimulus information. In contrast, endogenous EPs reflect higher mental procedures and are generated as a function of the mental processing allocated to the stimulus. Endogenous potentials are longer in latency (e.g. >100 ms after auditory stimulation). Late potentials can relate to recognition of novel stimulus, task requirements and difficulty, instructions, memory, or preparation for some upcoming motor action. Sometimes, it is difficult to separate purely sensory potentials from cognitive ones and they might overlap in time. The term event-related potentials (ERPs) is mainly used for evoked potentials that are elicited by cognitive activities [52].

The measured potentials are often considered as voltage changes resulted by multiple brain generators active in association with the eliciting event, and background noise, which is brain activity not related to the event. Additionally, there are contributions from non-neural sources, such as ocular artifacts (for an illustration see Fig. 7.3). In relation to the ongoing EEG, EPs exhibit very small amplitudes starting from few μV, and thus it is difficult to detect them straight from EEG. Therefore, traditional research and analysis of EPs requires an improvement of the signal-to-noise ratio by repeating stimulation, considering unchanged experimental conditions, and finally, averaging time locked EEG epochs. For example, in EEG source localization studies it is common to use averaged signals in order to identify cortex locations that are responsible to a specific mental task.

Evoked potentials are assumed to be generated either separately of ongoing brain activity, or through stimulus-induced reorganization of ongoing activity. For example, it might be possible that during the performance of an auditory oddball discrimination task, the brain activity is being restructured as attention is focused on the target stimulus [99]. Phase synchronization of ongoing brain activity is one

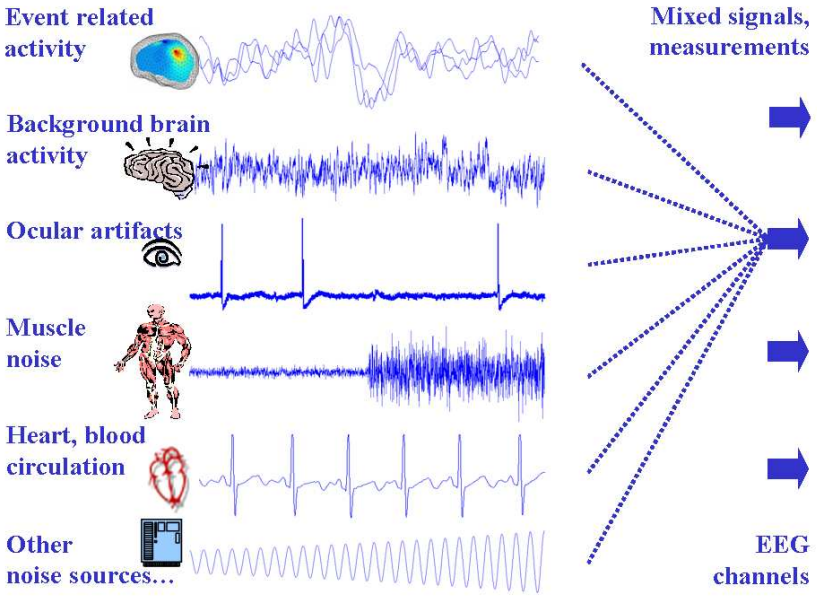

Figure 7.3: Noise sources in event related measurements.

possible mechanism for the generation of EPs. That is, following the onset of a sensory stimulus the phase distribution of ongoing activity changes from uniform to one which is centered around a specific phase [135]. Moreover, several studies have concluded that averaged EPs are not separate from ongoing cortical processes, but rather, are generated by phase synchronization and partial phase-resetting of ongoing activity [138, 102]. Though, phase coherence over trials observed with common signal decomposition methods (e.g. wavelets) can result both from a phase-coherent state of ongoing rhythms and from the presence of a phase-coherent EP which is additive to ongoing EEG [139, 135]. Furthermore, stochastic changes in amplitude and latency of different components of the EPs are able to explain the inter trial variability of the measurements [139, 200, 126]. Perhaps both type of variability may be present in EP signals.

A generally accepted EP terminology denotes the polarity of a detected component by the letters "N" for negative and "P" for positive, with a number indicating the typical latency. For example, for auditory evoked potentials (AEPs), N100 indicates a negative wave occurring with peak amplitude around 100ms after stimulation. It is usually followed by a positive deflection P200. However, the psychophysiological significance of the N100 depends on the context of its elicitation. When EPs are recorded during two different conditions, like in the oddball auditory paradigm where standard tones occurring in random times of about 1-2 Hz are interrupted by infrequent deviant tones, then other potentials arise (see

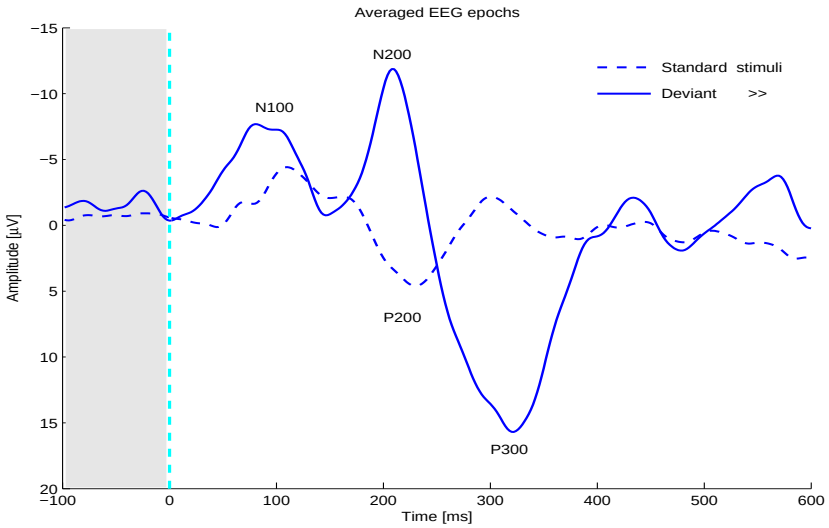

Figure 7.4: Average responses from an oddball paradigm.

Fig. 7.4).

Among all endogenous potentials, the P300 has attracted the widest interest in cognitive research. Even though there are still considerable controversies concerning its functional significance and generator structures, it is used in various areas of applied research and even in clinical examination. The difficulties in interpreting P300 "effects" in different paradigms have led to an enormous literature. Perhaps an explanation is that an unique P300 effect does not exist and with different experimental settings different subcomponents of it are involved. An important consideration of a P300 generation stems from the observation that intrusive or novel stimuli can produce an earlier positive peak P300a. This is to separate it from a later peak, the canonical P300 or P300b that is generated when the subject is required to respond to a designated stimuli. More information about the study of P300 potential can be found in [151].

Mental processes are sometimes considered as hierarchical steps and EPs are manifestations related to an elementary level in this hierarchy. For example, the auditory N100/P200 complex is believed to be associated with early discrimination of incoming stimulus and attention. N200 is believed to reflect manifestation of a memory retrieval system that stores physical characteristics of the regular stimuli. P300 amplitude could reflect attention to incoming stimulus information when memory representations are updated, and the latency is considered as a metric of stimulus classification speed and task difficulty.

EP research has to deal with several inherent difficulties. Firstly, the generation mechanisms of many components are not precisely known. Secondly, the separation in subcomponents is rather artificial, and a wide overlap of components

generated by different mechanisms might occur. Thirdly, traditional analysis is based on averaged data often by forming extra grand averages of different populations. Thus, trial-to-trial variability and individual subject characteristics are ignored. Therefore, the study of isolated components retrieved by averages might be misleading, or at least it is a simplification of the reality. For example, habituation may occur and the responses might be different from the beginning to the end of the recording session.

7.3 ICA for BSS of EEG

Individual EEG channels measure superimposed activity generated simultaneously from various brain sources. The behavior of the sources can be argued to be stochastic and generally non-stationary. In addition, artifact sources, such as eye blinks, can distort statistical properties of the signals. Therefore, the applicability of BSS and ICA methods relates to many EEG analysis fields.

The application of ICA to the analysis of EEG signals assumes that several conditions are verified, at least approximately. Although these restrictions of the basic ICA model could be argued to be more or less critical, the applicability of ICA and the gained separating results are considered useful for brain research. ICA for EEG analysis was first proposed in [134] and for MEG in [204]. Different applications of ICA and other BSS methods for EEG and different biosignals can be found in [89, 90, 91, 92, 93, 94], see also [85, 39]. In many studies on EEG and MEG signals it is shown that BSS methods can estimate noisy sources and potentially brain generated signals.

In this section, the applicability of ICA for BSS of EEG is discussed and an artifact removal example is demonstrated. The artifact corrected measurements are also used in chapter VIII.

7.3.1 Assumptions and applicability

For the problem of blind source separation of the multichannel EEG measurements, target is to recover the unobserved initial source signals of the brain by using only the available sensor data and some statistical properties assumed for the sources. Therefore, no physical model is usually used. For the same reason BSS and ICA methods do not aim and are not able to solve the EEG inverse problem and point out active locations within the brain. For example, the goal of ICA is to recover independent signals given only sensor observations (see chapter VI). Therefore, BSS and ICA methods for the analysis of EEG signals aim [101, 38]:

- to separate a set of measurements into their constituent components or source signals, and to provide information about the number of distinct sources producing the measurements,

- to provide the spatial distribution (on the sensors) of each estimated source along with the time varying characteristics of the source itself,

- to isolate components describing artifact signals and to identify signals reflecting distinct mental processes, and
- ultimately to track changes in the number, spatial distribution and morphology of sources over time.

The basic problem that BSS attempts to solve assumes a set of n measured data points at time instant t, i.e. $x(t) = (x_1(t), \ldots, x_n(t))^T$ to be a combination of m unknown sources $s(t) = (s_1(t), \ldots, s_m(t))^T$. For EEG n is the number of channels, and X is the matrix having the vectors x as its columns and different channel recordings as its rows (see (6.1)). A general formulation, without any assumptions about the nature of the data, will leave the problem of EEG separation intractable. Therefore, some basic assumptions are needed.

A first assumption for ICA and other BSS methods for EEG is the assumption of linear, time invariant, and without time delays mixing, i.e.

$$x(t) = As(t) + \upsilon(t), \tag{7.1}$$

where υ_t represents sensor noise. This is based on the plausible treatment of EEG recordings as linear sums of sources arising from spatially fixed, distinct, but overlapping brain or extra-brain areas, while the mixing mechanism is not changing. This may not be the case in all situations. For example, when electrocardiogram (ECG) is recorded on chest electrodes, the electrodes move over time due to breathing. An additional consideration stems from the fact that in biomedical recordings electrical impedance and therefore channel gain can change due to sweat production of the skin and pressure of electrode contact.

In order to use the basic ICA model for estimation, some other assumptions must be made, that is the noise free model

$$x(t) = As(t), \tag{7.2}$$

where the number of sensors is bigger or equal than the number of sources, i.e. $n \geq m$. This allows different algorithms to be applied, but the estimated components may remain contaminated by measurement noise. Reduction in noise for EEG signals can be done with high-pass, low-pass or band-pass linear filtering without altering the basic model ([85], p. 264). For example, if we assume the model (7.2) with $m = n$, low-pass filtering can be done by multiplying with a matrix M from right as

$$X' = XM = ASM + \Upsilon M = AS' + \Upsilon', \tag{7.3}$$

that can reduce the noise. Similarly high-pass filtering can remove slow trends that in general can distort statistical properties of the measurements. Thus, appropriate band-pass filtering aims to enhance the properties of the measurements (e.g. spatial dependencies) without destroying useful information for estimation.

If it is further assumed that $n > m$ and that the sensor random noise vectors are identically distributed having relatively small variances, then PCA dimension reduction can be used to reduce the noise and make the mixing system square.

Note that filtering and dimension reduction has as a consequence the loss of some measured source signals. This is the basic criticism for PCA dimension reduction, as well as that the number of retained dimensions is usually selected subjectively [101]. Though PCA will not eliminate relatively strong sources (at the sensors), especially for high-density measurement systems (see section 6.2). Additionally, when short data segments are analyzed with ICA, reducing dimensions prevents overlearning ([85], p. 268). This means that if the number of parameters in a statistical model is too large compared to the number of data points, then the estimated parameters depends less on the observed data and much more on the model assumptions. In addition, with band-pass filtered data a much stronger dimension reduction is necessary to prevent overlearning [85]. Though a very strong dimension reduction will not allow separation. Finally, if the sources have strong time-dependencies the sample size must be larger to avoid overlearning [85].

By far the most important assumption for ICA for BSS of EEG is that of independence of the sources. This assumption can be physiologically plausible in some situations, for example, between brain signals and ocular artifacts. But in general it could be argued that there is not enough evidence to support independence of brain generated signals in every situation. Furthermore, the ICA model considers the data as independent and identically distributed and requires non Gaussian sources. Thus, by ignoring time structure, the estimation is based solely on investigating structure across the sensors as estimated by the sample distribution of the measurements and the embedded density parametrization. Therefore, the model might not be able to separate every kind of source (e.g. stationary Gaussian random processes, bimodal sources etc.). However, in many situations predominant artifacts, for example blinks, show a highly kurtotic sample distribution that enables estimation. Therefore, ICA has found applicability in identifying and removing artifacts associated with blinks, eye-movements and muscle noise, see for example [207, 104, 105, 100]. For some other ocular correction methods see for example [47], and in comparison with ICA see also [208].

It must be noted that in contrast to other Factor Analysis methods ICA concentrates on finding the matrix B based on assumed properties of the components and not of the mixing system (e.g. sparseness). Though in many applications on EEG the mixing matrix A has proved to be sparse for distinct estimated source signals. In addition, some independent components tend to have clearer time/frequency characteristics than the original signals. Source localization methods applied to ICA components often suggest physiological plausible locations within the brain. These observations have encouraged the use of ICA for the analysis of brain generated signals. Therefore, ICA has been proposed for the study of brain signals in event related studies [182, 199, 137, 138, 136, 135, 106, 50, 49, 157, 48], where it is discussed the possible physiological origin of different estimated components. For related assumptions and considerations see also [14, 206, 205, 70]. For other BSS methods that take advantage of the time or frequency structure of EEG see for example [188, 189, 148, 9, 101, 38].

7.3.2 An artifact removal example

ICA methods carry ambiguities about the ordering, the overall amplitude, and sign of the estimated sources. The rows of the data matrix X are the EEG channel recordings. The estimated separating matrix B and the independent components matrix Y are given by $Y = BX$. Y describes time-varying characteristics of the sources. The matrix $A = B^{-1}$ mixes the estimated independent components, which have variance one, back to the sensors

$$X = AY. \tag{7.4}$$

The columns of A give the relative projection strengths of the respective components at each of the scalp sensors. An artifact source can be removed from the measurements by replacing the respective row of Y with zeros and using the previous equation to obtain denoised measurements at the sensors. Note that the dimension of the data is then reduced. The columns of A can be used to produce scalp maps that can help, together with the time, frequency, and time-frequency morphology, the characterization of components as artifacts. Eye movements and eye blinks project mainly to frontal sites (near electrodes FP1 and FP2), temporal muscle activity to the temporal sites (near T3 and T4), and occipital (rear head) movements to the back (near O1 and O2). The interpretation of scalp maps in any EEG analysis must be done with caution, since the location of maxima and minima on these maps does not locate the areas of maximal and minimal activity on the underlying cortex. If we denote $A = \{a_{ij}\}$, with $j = 1, \ldots, m$ being the column index, an ordering of the components can be based on the quantity [101]

$$p_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n} a_{ij}^2}, \tag{7.5}$$

which describes the power of the contribution of a unit variance component. Another possibility is to order the components according to their estimated kurtosis.

EEG measurements were obtained from a standard odd-ball paradigm with auditory stimulation (1 subject, 60 EEG channels, reference: ears). In the recording, 569 auditory stimuli were presented with an inter-stimulus interval of 1 second, 85% of the stimuli at 800Hz and randomly presented 15% deviant tones at 560 Hz. The subject was sitting in a chair and was asked to press a button every time he heard the deviant target tone. The sampling rate of the measurements was 500 Hz. Averaged EPs from the experiment (after artifact correction, channel CZ) can be seen in Fig. 7.4.

For the analysis, the data were digitally filtered in the range (1-40Hz). All the measurement length (about 10 minutes) was used for the estimation of the separating matrix. The dimension of the data was reduced with PCA to 31, $m = 31$, by keeping eigenvectors associated with eigenvalues larger than 1, $\lambda_i \geq 1$, resulting in more than 99.9% of explained variance, $\sum_i^m \lambda_i / \sum_i^n \lambda_i$. After whitening the data, the FastICA algorithm in parallel form (see 6.49, 6.50) was used for the estimation of independent components, which were shorted according to (7.5).

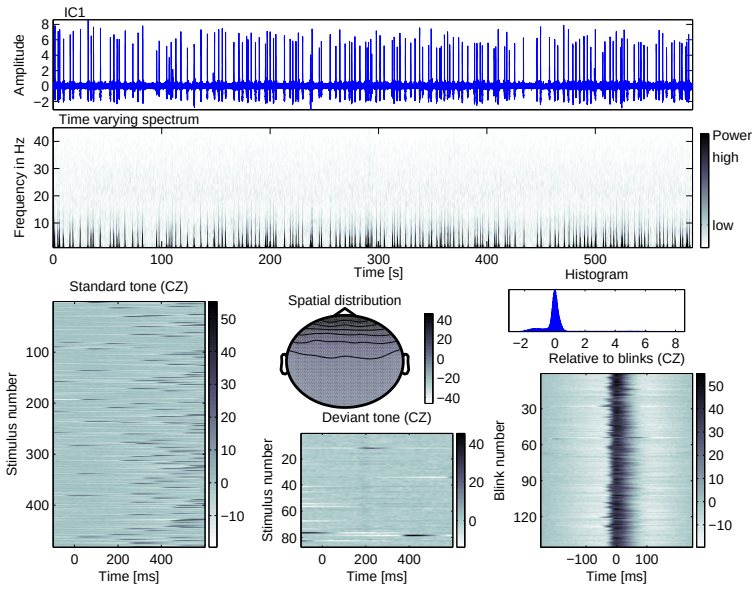

Figure 7.5: Strong blink related artifact.

For the evaluation and categorization of ICs time and time-frequency plots, histograms, and scalp maps were computed. For example, IC1 (Fig. 7.5) clearly represents a strong artifact associated with eye activity (blinks). From that component, estimates for the blink occurrence time were obtained. For every component epochs relative to blinks were sampled. Epochs were also sampled relative to the two stimulus types for every component. The epochs represent activity as it is observed from channel CZ (backprojection of the component). The epochs are also presented as image plots for the facilitation of ICs categorization.

IC1 is a clear blink artifact, easily detectable and highly kurtotic, giving confidence for a good separation result from brain related activity or other noise. IC19 (Fig. 7.6) is also an ocular related signal, but weaker and more noisy (having prominent activity between 20-30 Hz). Though it does not seem to be any stimuli related activity within and it correlates perfectly with blink occurrences. Both IC1 and IC19 represent most of the noise subspace related with blinks though it is difficult to characterize them as separate source signals since first order difference operation and rescaling of IC1 produce a signal similar (cleaner) to IC19. Those two signals were removed from further analysis.

Some other components representing artifacts are presented in the Appendix. Another distinct group of signals is given by IC7 (Fig. A.1) and IC26 (Fig. A.2) that maximally project on left and right temporal channels. They decrease in amplitude during the recording period and are active in the same frequency band. They probably represent artifacts (muscle or head phone related noise). IC25

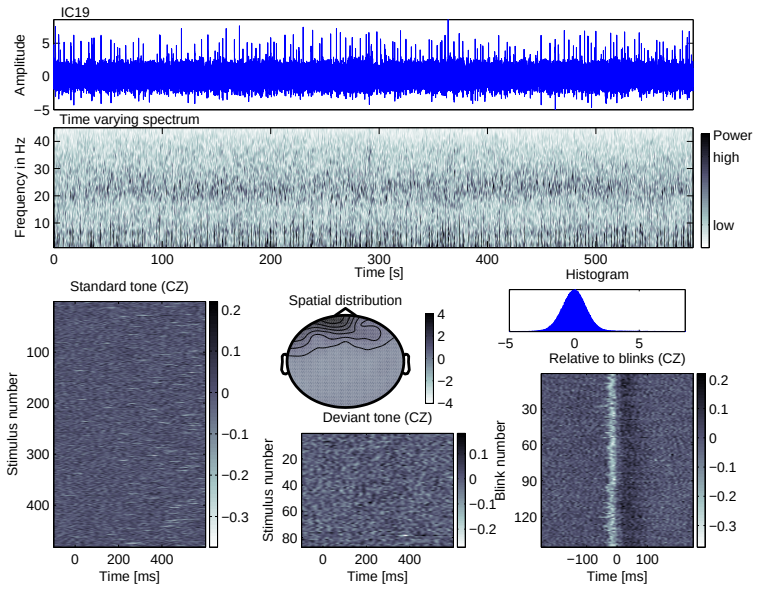

Figure 7.6: Second blink component.

(Fig. A.3) must be a rear head related artifact without any visible correlation to stimulation, or blinks, or the previous two characteristic signals. So these three signals were also removed from the measurements, as well as IC29 (Fig. A.4). Here, it must be noted that from the rejected signals primarily IC1 influences the quality of the estimates presented in chapter VIII.

In the Appendix, some interesting components originating from the brain are also presented. The characterization of brain signals is much more difficult. IC2 (Fig. A.5), IC4 (Fig. A.7), and IC5 (Fig. A.8) represent components of stimuli related activity, and seem to originate from different brain areas. IC3 (Fig. A.6) represents ongoing brain activity having large spectrum bandwidth. IC6 (Fig. A.9) has prominent alpha activity. It can be noticed that IC6 has some degree of correlation with stimulation and blinking.

When placed in physiological analysis the utility and assumptions of ICA and other BSS methods for the study of mental processes, individual application characteristics should be taken into account [158]. Then it is beneficial to use available physiological information for the selection of appropriate modeling method and for possibly embed prior information in the estimation. For a discussion see [101]. For this thesis, it is enough to conclude that EP signals can be the results of multiple overlapping event related activations and ongoing brain activity (alpha rhythms and others). Also it exists significant contamination from artifacts. Those observations can not be easily made straight from EEG, and especially from a single-channel.

7.4 Single-channel EP estimation

Several techniques have been proposed in order to denoise and estimate EPs. In general, the performance of the estimator or the filter is naturally dependent on the prior information imposed for the statistical properties of the EPs and the background noise. For an extensive discussion about different EP estimation methods see also [116]. Of special interest is the case when some parameter of the EP changes dynamically from stimulus to stimulus. This kind of situation can be, for example, a trend like change of amplitude or latency of some specific peak of the EPs. A method in which preceding or following measurements are used in the estimation procedure can give additional information about the EPs and improve single-trial estimation.

7.4.1 Single-trial estimation

The simplest model for EP estimation considers the EEG epochs to be the sum of an invariant signal (from trial-to-trial) and random noise. Then estimators of the form (5.8) can be used for selective or weighted averaging of EP epochs (see also [116]) . This model implies a loss of information about trial-to-trial variability, and time-varying features of event related phenomena [78, 213, 200]. In that sense, the investigation of the variability of EP parameters could reveal interesting hidden mental processes.

Digital filtering is sometimes used for single responses [181]. The main problem in linear time invariant filtering is that usually the spectra of EPs and background noise overlap heavily. Wiener filtering is also possible for single measurements, with some specific structure of the filter (e.g. [34]). Other approaches involve time-varying filtering of single-trials based on Wiener formalism [214]. A crucial problem in time-varying mean square error filtering is to obtain a good model for the cross covariances between the EPs and the measurements. This is a difficult estimation task, especially when there exists a significant level of correlation between the EPs and the noise. Later methods include, for example, time-frequency decomposition of the signal based on wavelet transform [171] and regularization based methods [118]. Some estimation methods also exist for the case that multiple channels are used in the analysis (e.g. [172]).

Different methods for single-trial EP analysis aim to decompose the measurements into relevant components or to explain the data through some parameters. The parametrization gives the means to investigate, for example, the changes that the stimuli caused to the ongoing EEG signal or that the repetition of the test caused to the responses. Most of the methods are based on some explicit model or on some assumptions for the EPs. Every decomposition then involves at least two main considerations. On the one hand, if the resulting estimates follow too closely the measurements, it is probable that some features are still going to be hidden by phenomena not related to the stimulation. On the other hand, if the estimates are not following closely the measurements, some features may have been lost or some extra features may have been created by the model itself. A balance between these considerations is necessary for the correct interpretation of a parametrization that

is able to reveal some specific features of the experiment.

7.4.2 Time-varying estimation with linear observation model

A sampled potential relative to the t-th stimulus from a single channel can be denoted with a column vector of length M, i.e. $z_t = (z_t(1), z_t(2), \ldots, z_t(M))^T$. Using these measurements, a vector of random parameters θ_t is to be estimated. The linear estimator that minimizes the mean square Bayes cost is (4.142)

$$\hat{\theta}_t = \eta_{\theta_t} + C_{\theta_t, z_t} C_{z_t}^{-1} (z_t - \eta_{z_t}), \tag{7.6}$$

where η_{z_t} is the mean and C_{z_t} the covariance of the measurement vector z_t, η_{θ_t} is the mean of the parameter vector, and C_{θ_t, z_t} is the cross-covariance of the measurements and the parameters to be estimated. The estimator is optimal among all possible estimators, not only linear, if θ_t and z_t are jointly Gaussian. The mean square optimality among linear estimators holds for every joint density of the form $p(z_t, \theta_t)$.

The selection or estimation of C_{θ_t, z_t} is in general difficult without some prior knowledge about the EPs. For the linear additive noise model we have

$$z_t = s_t + \upsilon_t. \tag{7.7}$$

The vector s_t corresponds to the part of the activity that is related to the stimulus and the rest of the activity υ_t can be considered to be independent of the stimulus and the EP. With this model, the EP s_t equals the parameter vector θ_t and equation (7.6) is the time-varying (within a trial) Wiener filter used, e.g., in [214]. The EPs can be further modeled as a linear combination of some pre-selected basis vectors. Then, the observation model takes the form

$$z_t = H_t \theta_t + \upsilon_t, \tag{7.8}$$

where H_t is a deterministic observation matrix, which contains the basis vectors $\psi_{t,1}, \ldots, \psi_{t,n}$ of length M in its columns, and θ_t is a parameter vector of length n. The estimated EP $\hat{s}_t$ can be obtained from the estimated parameters $\hat{\theta}_t$ as

$$\hat{s}_t = H_t \hat{\theta}_t. \tag{7.9}$$

The linear mean square estimator with observation model H_t, in the special case that θ_t and υ_t are uncorrelated, i.e. $C_{\theta_t, \upsilon_t} = 0$, is given by (4.167)

$$\hat{\theta}_t = (H_t^T C_{\upsilon_t}^{-1} H_t + C_{\theta_t}^{-1})^{-1} (H_t^T C_{\upsilon_t}^{-1} z_t + C_{\theta_t}^{-1} \eta_{\theta_t}), \tag{7.10}$$

where C_{θ_t} and η_{θ_t} are the covariance and the mean of θ_t (prior density), and C_{υ_t} is the covariance of the zero mean measurement noise. Again this estimator is optimal among all possible estimators, not only linear, if θ_t and z_t are jointly Gaussian, but with the requirement of uncorrelated parameters and noise.

Based on the Bayesian formalism, a simple estimation method for EP denoising that assumes only the smoothness of the EP signals is given by smoothness priors method (4.86)

$$\hat{s}_t = (I + \alpha_t^2 D_1^T D_1)^{-1} z_t \tag{7.11}$$

See also section 5.3.3. Of importance is the selection of the smoothing parameters $\alpha_t, t = 1, \ldots, T$.

7.4.3 Signal and noise subspaces

Several parametrizations have been used for single-trial estimation of EPs in different studies. Common ones are based on different types of Gaussian shaped components and exponentially damped sinusoidal functions [210]. Other choices are Fourier basis or several possibilities of wavelet basis. While these choices are not based on the measurements, it is possible to form decompositions based on the data. Singular value decomposition (SVD) has many theoretical and practical applications in signal processing and identification problems [69]. In relatively high signal-to-noise ratio conditions SVD of a data matrix can divide measurements into signal and noise subspaces. Alternatively it can also be understood in terms of principal component regression (PCR) as a combined method for signal enhancement and optimal model dimension reduction [103], see also section 6.2. The subspace method has been used to enhance stimulus phase-locked activity in different studies (e.g. [118, 41]).

The available data matrix $Z = [z_1, \ldots, z_T] \in \mathbb{R}^{M \times T}$ having as columns the EEG sampled epochs relative to the stimulation can be decomposed as

$$Z = U \Sigma V^T, \tag{7.12}$$

where $U \in \mathbb{R}^{M \times M}$ satisfies $U^T U = I$, $V \in \mathbb{R}^{T \times T}$ satisfies $V^T V = I$, and $\Sigma \in \mathbb{R}^{M \times T}$ is a pseudo-diagonal matrix with non-negative diagonal elements σ_i such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(M,T)} \geq 0$. If $M \leq T$ then Σ has the form $\Sigma = [\Sigma_1, 0]$, where $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_M)$ and 0 is a zero matrix. If $M > T$ then Σ has the form $\Sigma = \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix}$, where $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_T)$. Only r singular values are non zero, where $r = \mathrm{rank}(Z)$. For the additive noise model and relatively small noise the following decomposition can be considered

$$Z = [U_s, U_v] \begin{bmatrix} \Sigma_s & 0 \\ 0 & \Sigma_v \end{bmatrix} [V_s, V_v]^T. \tag{7.13}$$

The matrix Σ_s contains the k largest singular values and U_s the respective left singular vectors associated mainly with the signals s_t. Thus the matrices (U_s, Σ_s, V_s) represent a signal subspace and (U_v, Σ_v, V_v) represent primarily the noise subspace ([39], p. 118).

From the SVD of the matrix $Z = U \Sigma V^T$ we also have

$$Z Z^T = U \Sigma_1^2 U^T. \tag{7.14}$$

This means that the left singular vectors of Z are the eigenvectors of ZZ^T or of the data correlation matrix

$$\hat{R} = \frac{1}{M} ZZ^T. \tag{7.15}$$

If we denote with H_s the matrix with columns the k dominant eigenvectors, then the ordinary least squares estimator for the parameters θ_t becomes

$$\hat{\theta}_t = (H_s^T H_s)^{-1} H_s^T z_t = H_s^T z_t. \tag{7.16}$$

Estimates for the EPs can be obtained from (7.9). Quantitatively, the first basis vector is the best mean-square fit of a single waveform to the entire set of epochs. Thus, the first eigenvector is similar to the mean of the epochs and the corresponding parameters or principal component $\hat{\theta}_t(1)$ $(t = 1, 2, \ldots, T)$ reveal the contribution of the eigenvector to each epoch. The rest of the dominant eigenvectors model primarily amplitude differences between individual EP peak components and latency variations from trial-to-trial.

The estimator can be combined (preprocessing) with smoothness prior method. Estimates can then be obtained from

$$\hat{s}'_t = H_s H_s^T z_t, \tag{7.17}$$

where now the signal subspace (dominant eigenvectors) is estimated from the smoothed vectors (7.11) S'=$[\hat{s}_1, \ldots, \hat{s}_T]$, rather than the raw measurements. A related approach was proposed in [41], see also [187]. The effect of strong artifacts on the EP signal subspace is discussed in Chapter VIII.

7.4.4 Dynamical estimation

In this thesis, focus in given on the case that dynamic changes exist from stimulus to stimulus. Although some of the methods that are briefly mentioned here and in chapter IV, could be used to estimate dynamically changing features, the possibility that previous trials and estimates may contain relevant information to next trials and estimates, is not taken explicitly into account in the estimation procedure. State-space modeling (chapter V) can give a good theoretical base and estimation performance for tracking time-varying features of the EPs.

The most obvious way to handle time variations between single-trial measurements is sub-averaging of the measurements in groups. Sub-averaging is used, e.g., in [24] to demonstrate the decrease of amplitude in visual EPs. Sub-averaging could give optimal estimators if the EPs are assumed to be invariant within the sub-averaged groups. Another method for the dynamical estimation of EPs is the windowed averaging of the measurement vectors. This can also be called sliding window averaging. The estimator then takes the form of a moving average filter. In vector form the moving window average filter for dynamical estimation is

$$\hat{s}_t^{MWA} = \sum_{i=0}^{k-1} w_i z_{t-i}, \tag{7.18}$$

where t denotes the t-th stimulus. In [194], this filter average was used with equal weights $w_i = 1/k$. Another method which was used in [194, 54] is exponentially weighted average (EWA) in which the weights are of the form $w_i = \gamma^i / \sum_{j=0}^{k-1} \gamma^j$, for some $0 < \gamma < 1$. It can be shown that an equivalent form is given by

$$\hat{s}_t^{EWA} = \gamma z_t + (1-\gamma)\hat{s}_{t-1}^{EWA}, \tag{7.19}$$

The common disadvantage of these moving averages is that they cannot be adaptively tuned based on the data. Another disadvantage is that neither their statistical properties nor the assumptions imposed to the EPs and background EEG can be directly defined. However, their statistical properties can be investigated through the Kalman filter equations.

A more natural way to handle time variations is given by state-space modeling and recursive mean square estimation. In this formulation, the measured EP epochs are directly assumed to be vector valued stochastic processes. With the selections $F_t = I$, $G_t = I$ for every t the state-space equations are of the form (random walk)

$$\theta_{t+1} = \theta_t + \omega_t \tag{7.20}$$
$$z_t = H_t\theta_t + \upsilon_t. \tag{7.21}$$

If we denote the conditional covariance matrix of the parameter estimation error as $P_t = C_{\tilde{\theta}_t} + C_{\omega_t}$, the Kalman filter equations (5.83-5.86) can be written as

$$K_t = P_{t-1}H_t^T(H_tP_{t-1}H_t^T + C_{\upsilon_t})^{-1} \tag{7.22}$$
$$P_t = (I - K_tH_t)P_{t-1} + C_{\omega_t} \tag{7.23}$$
$$\hat{\theta}_t = \hat{\theta}_{t-1} + K_t(z_t - H_t\hat{\theta}_{t-1}), \tag{7.24}$$

where K_t is the Kalman-gain matrix. Estimators for the EPs can be directly obtained from $\hat{s}_t = H_t\hat{\theta}_t$. The last equation can be written in the following form

$$\hat{\theta}_t = \hat{\theta}_{t-1} + K_t\epsilon_t, \tag{7.25}$$

where the residual or prediction error $\epsilon_t = z_t - H_t\hat{\theta}_{t-1}$ is the estimator of the unknown noise vector υ_t. With different choices or assumptions for C_{ω_t}, C_{υ_t}, and P_0 several adaptive algorithms can be written in the form (7.22-7.24). Kalman gain matrices K_t and recursive covariance estimates P_t for different recursive algorithms, namely recursive least squares (RLS), least mean square (LMS), and normalized least mean square (NLMS), are presented in Table 7.1. In that sense, these adaptive algorithms can be optimal in the mean square sense if the specific choices or assumptions about the parameters are valid. The connections of RLS, LMS, and NLMS algorithms to Kalman filtering are discussed e.g in [183, 75]. Such methods have been proposed for EP estimation (especially brain stem potential tracking), see for example [170] and the references therein. For a comparison of different adaptive algorithms in EEG spectrum estimation see [193, 190].

Table 7.1: Kalman gain matrices K_t and conditional covariance matrices of parameter estimation error P_t for different recursive algorithms.

Kalman filter	
	$K_t = P_{t-1}H_t^T(H_tP_{t-1}H_t^T + C_{\upsilon_t})^{-1}$
	$P_t = (I - K_tH_t)P_{t-1} + C_{\omega_t}$
RLS	
	$K_t = P_{t-1}H_t^T(H_tP_{t-1}H_t^T + \lambda_t)^{-1}$
	$P_t = \lambda_t^{-1}(I - K_tH_t)P_{t-1}$
LMS	
	$K_t = \mu H_t^T$
	$P_t = \mu(I - \mu H_{t+1}^T H_{t+1})^{-1}$
NLMS	
	$K_t = \mu H_t^T(\mu H_t H_t^T + 1)^{-1}$
	$P_t = \mu I$

If we further choose $H_t = I$, from (7.24) it holds

$$\hat{s}_t = \hat{\theta}_t = \hat{\theta}_{t-1} + K_t(z_t - \hat{\theta}_{t-1}) = K_t z_t + (1 - K_t)\hat{s}_{t-1}. \tag{7.26}$$

If we compare now with equation (7.19), we can see that exponentially weighted average can be obtained by choosing some fixed value for the Kalman gain such as $K_t = \gamma I$. Actually, it is enough to choose $H_t = I$, $C_{\upsilon_t} = C_\upsilon$, $C_{\omega_t} = \gamma^2/(1 - \gamma)C_\upsilon$ for every t and $P_0 = \gamma/(1 - \gamma)C_\upsilon$ in the Kalman filter equations to obtain exponential weighted average. Through the connection of all the above mentioned methods with Kalman filtering theory we can conclude that state-space modeling and recursive Bayesian mean square estimation gives a good theoretical base for the investigation of realistic models for dynamical estimation of EPs.

Some smoothing methods have also been proposed for modeling trial-to-trial variability in EPs (e.g. [175, 202]). A Bayesian smoothing method that can be applied for smoothing EPs across trials is given by smoothness priors method. If $G = (I + \alpha^2 D_1^T D_1)^{-1}$, then estimates can be obtained in matrix notation as

$$\hat{S} = ZG, \tag{7.27}$$

where Z is the matrix having as columns the single trials. A pseudo two dimensional smoothing procedure is given by (see eq. (7.11))

$$\hat{S}' = G'ZG, \tag{7.28}$$

for appropriate selected dimensions. The connection of the above methods with Kalman smoother has been presented in chapter V.

Chapter VIII

Tracking dynamic changes

In this chapter, novel methods for estimating dynamic features present in evoked potential measurements are presented, and their applicability is discussed. The methods are based on state-space modeling and recursive Bayesian mean square estimation. These topics were treated in chapter V. Here, it is demonstrated the capability of the proposed methods to track time-varying changes due to repeated presentation of stimuli. The evaluation is based on simulated and real EPs. The same measurement set, obtained form an auditory oddball type of experiment, is used throughout the chapter.

8.1 Basic concepts and definitions

The sampled potential (from a single channel) relative to the t-th stimulus is denoted with a column vector of length M

$$z_t = \begin{pmatrix} z_t(1) \\ z_t(2) \\ \vdots \\ z_t(M) \end{pmatrix}, \quad t = 1, \ldots, T. \tag{8.1}$$

The observation at the t-th stimulus z_t depends on some unobserved parameters θ_t (state vector) through the model

$$z_t = H_t \theta_t + \upsilon_t, \tag{8.2}$$

where H_t is an observation matrix, which contains the basis vectors $\psi_{t,1}, \ldots, \psi_{t,n}$ of length M in its columns. The unknown quantity θ_t depends on the parametrization of the estimation problem. For the time evolution of the hidden process θ_t a linear first order Markov model is used, i.e.

$$\theta_t = F_t \theta_{t-1} + \omega_t. \tag{8.3}$$

The assumptions of the state-space model are given in section 5.2. The estimated EPs $\hat{s}_t$ can be obtained from the estimated parameters $\hat{\theta}_t$ as

$$\hat{s}_t = H_t \hat{\theta}_t, \tag{8.4}$$

and it represents stimulus triggered EEG activity. Optimal estimates $\hat{\theta}_t$ based on past measurements are obtained recursively by Kalman filter (KF) algorithm. If all the measurement vectors are available, i.e. $z_t, t = 1, \ldots, T$, then the fixed interval smoother (KS) can be used. The algorithms are summarized as follows:

- Initialization

$$C_{\tilde{\theta}_0} = C_{\theta_0}, \tag{8.5}$$
$$\hat{\theta}_0 = E\{\theta_0\}. \tag{8.6}$$

- Prediction step

$$\hat{\theta}_{t|t-1} = F_t \hat{\theta}_{t-1}, \tag{8.7}$$
$$C_{\tilde{\theta}_{t|t-1}} = F_t C_{\tilde{\theta}_{t-1}} F_t^T + C_{\omega_t}. \tag{8.8}$$

- Filtering step

$$K_t = C_{\tilde{\theta}_{t|t-1}} H_t^T (H_t C_{\tilde{\theta}_{t|t-1}} H_t^T + C_{\upsilon_t})^{-1}, \tag{8.9}$$
$$\hat{\theta}_t = \hat{\theta}_{t|t-1} + K_t(z_t - H_t \hat{\theta}_{t|t-1}), \tag{8.10}$$
$$C_{\tilde{\theta}_t} = (I - K_t H_t) C_{\tilde{\theta}_{t|t-1}}, \tag{8.11}$$

for $t = 1, \ldots, T$.

- Smoothing (backward recursion)

$$A_t = C_{\tilde{\theta}_t} F_{t+1}^T C_{\tilde{\theta}_{t+1|t}}, \tag{8.12}$$
$$\hat{\theta}_t^s = \hat{\theta}_t + A_t(\hat{\theta}_{t+1}^s - \hat{\theta}_{t+1|t}), \tag{8.13}$$
$$C_{\tilde{\theta}_t^s} = C_{\tilde{\theta}_t} + A_t(C_{\tilde{\theta}_{t+1}^s} - C_{\tilde{\theta}_{t+1|t}}) A_t^T, \tag{8.14}$$

for $t = T-1, T-2, \ldots, 1$. For the initialization of the backward recursion the filter estimates can be used, i.e. $\hat{\theta}_T^s = \hat{\theta}_T$.

A random walk model $F_t = I$, for every t, for the state evolution can be modified to (5.135)

$$\theta_t = (I + C_{\omega_t} C_t^{-1})^{-1} \theta_{t-1} + \omega_t, \quad \omega_t \sim N(0, (I + C_{\omega_t} C_t^{-1})^{-1} C_{\omega_t}), \tag{8.15}$$

in order to include extra prior information in the estimation procedure (e.g. extra smoothness). Though, the prior selection of C_t depends on the parametrization. Finally, the state-space identification algorithm presented in section 5.4 can also be used. This can improve the tracking capabilities of the methods by introducing time variation to the basic model.

8.2 Applicability

For the methods based on Kalman filtering, a common issue is the investigation of optimal choices for the covariance matrices C_{ω_t}, C_{v_t} in an unknown environment. The adaptation, or the speed-of-change of the parameters can be controlled by the selection of different state noise C_{ω_t}, and observation noise C_{v_t} covariance matrices. The matrices C_{v_t}, although assumed to be known until now, could be estimated based on the data. This is a difficult estimation task, since background EEG, even not related to the stimulation, is a non-stationary process. Therefore, its properties cannot be estimated accurately from a pre-stimulus sample or from ensemble data. Estimation based on the prediction error is one possibility, but it is related to the selection of the state noise covariance matrices. When recursive algorithms are applied for adaptive modeling of time series, some strategies exist for time varying selection. For some approach see [191, 192]. This thesis is primarily focused on the selection of C_{ω_t}, and of an appropriate state-space model for estimation. Since different statistical properties of the residuals depend at least on the selection of observation model, the selection $C_{v_t} = \sigma_v^2 I$, for all t, can be made. In that case the selection $\sigma_v^2 = 1$ follows, see equation (5.116). Then care can be given to the selection of C_{ω_t} in relation to the selected state-space model.

A common strategy in state-space modeling is the choice of diagonal matrices for C_{ω_t}. This implies that the vectors ω_t have uncorrelated components. For dynamical estimation of EPs, it can be further assumed that the interesting phenomena should be slowly varying from stimulus to stimulus. Then the selection $C_{\omega_t} = C_\omega$ for every t can be made. This assumes that every sudden variation is due to the background noise, and it should be filtered out. The simpler choice is then $C_{\omega_t} = \sigma_\omega^2 I$, i.e. all the parameters are allowed to change similarly from trial-to-trial. Naturally, the speed-of-change of the estimated parameters is tightly related to the selection of σ_ω^2 (in relation always to the observation model). When it is selected to be too large, then the estimates tend to be close to the least squares solution (e.g. see (5.116)). A very small selection forces the estimated parameters to be almost identical, and identical to the initialization (e.g. see 5.107).

The assumptions imposed to the background noise are important for dynamical estimation of EPs. The Gaussian assumption can be easily relaxed, if we only require the optimality among all the linear estimators. Furthermore, it is understood that EEG measures activity generated from different brain locations as well as activity arising from non-neural sources. Therefore, it can be considered that a significant part of the variability in EP data is created by superposition of a large number of random processes that are not relevant to the stimulation. Additionally, and much more important for the applicability of the proposed methods, some part of the activity must be uncorrelated from stimulus to stimulus. Then it is possible to eliminate all nearly Gaussian variability, and every other uncorrelated from trial-to-trial contributions.

In practice, it is difficult to decide which part of the activity is related to the eliciting event, and which parts represent just randomly occurring phenomena. From this arises the following intuitional approach for dynamic estimation of EPs

when slow changes of interest exist from stimulus to stimulus. With appropriate tuning of the relevant parameters state-space representation and recursive mean square estimation model the most dominant phenomena correlated to the stimulation. Any other nearly randomly occurring phenomena will be filtered out. Any significant violation of the assumptions may hide interesting results of the experiment.

State-space modeling for EP estimation was originally proposed in [117, 116], where the Kalman filter algorithm was considered. In those studies models of the form $H_t = H$, $F_t = I$, $C_{\upsilon_t} = I$, and $C_{\omega_t} = \sigma_\omega I$, for every $t = 1, \ldots, T$, where only considered. In [64] the method was further developed, and the applicability of Kalman filter was demonstrated based on the use of time shifted Gaussian shaped functions (generic observation model). Kalman smoother algorithm for EP estimation was briefly introduced in [63]. In this thesis, in the spirit of [64], Kalman filter and smoother algorithms are compared in order to demonstrate the better performance of the smoother, and to discuss the applicability of the method. Additionally, models of the form (8.15) are introduced for the enhancement of the random walk model for EP estimation. Finally, a state-space identification method is presented for the improvement of the tracking capabilities of the methods.

In section 7.4.3, it was concluded that the dominant eigenvectors of the data correlation matrix can form an observation model for EP estimation. Since this basis contains prior information about phase-locked characteristics of the EP signals the following state-space model for dynamical estimation can be considered

$$\theta_t = F_t\theta_{t-1} + \omega_t \tag{8.16}$$
$$z_t = H_s\theta_t + \upsilon_t, \tag{8.17}$$

with $H_t = H_s$ for all t. This observation model was also considered in [116]. The performance of the method relates on the quality of the signal subspace in low signal-to-noise ratio conditions, as well as on the assumption of hidden dynamical behavior. In this thesis, this model is further exploited and its usability discussed together with other generic parametrizations.

8.3 Simulations

Although the proposed methods are capable to estimate different kinds of EPs, data resembling the P300 peak were simulated. The P300 peak is one of the most extensively studied cognitive potential and there exist many studies where the trial-to-trial variability of the component is discussed.

Real EEG measurements were used as background EEG activity (noise) in the simulations. From the recordings, only the channel CZ was used, after preprocessing and artifact removal by ICA (see section 7.3.2). As background activity for the simulations, prestimulus EEG epochs from -700 ms to 0 ms relative to the deviant stimulus onset were sampled. Simulated EP measurements were then constructed according to the additive noise model by superimposing upon the selected real EEG epochs linear combinations of 3 Gaussian shaped functions of the form

$$s_t = \sum_{i=1}^{3} a_t(i) e^{-(\tau - b_t(i))^2 / c_t^2(i)}, \tag{8.18}$$

where $a_t(i)$ is the amplitude , $b_t(i)$ is the latency, and $c_t(i)$ is the width of the i-th Gaussian component at the t-th stimulus. The Gaussian components were computed over the points $\tau = \{-49, \ldots, 0, \ldots, 300\}$, such that the theoretical stimulus occurs at point zero. In order to be consistent to the real measurements sampling rate (500Hz), each pseudo-real EP vector has three Gaussian peaks, a negative around 100 ms after the stimulus, a negative around 200 ms, and a positive around 300 ms.

In order to evaluate the methods for different dynamic variations, four sets of simulations (87 EPs per set) for the noiseless third peak were created. In the first case (Case 1), trial-to-trial dynamic variations of the amplitude and latency of the third peak were set to be linear functions of time. In Case 2, dynamic variations were created by two different sinusoidal functions, one for the amplitudes and one for the latencies. Random variations uniformly distributed were further added to the amplitudes and latencies for both cases. In Case 3, for the amplitudes and latencies of the third peak uniform variations in a larger range were allowed. Case 4 simulates a sudden change in amplitude and latency after trial 45. The range of the random variability was set double compared to Case 1. The widths of the third peak in all the cases were also uniformly distributed but in a smaller interval. The amplitudes and latencies of the simulations are presented in Fig. 8.1 as functions of the stimulus number. The other two peaks were selected to be the same for the three cases having random variations for the amplitudes, latencies and widths, and linear trends in amplitudes. The background noise was the same for all the cases for better comparisons.

8.4 Kalman filter vs. smoother

For estimation the model (8.16), (8.17) was considered. For the state evolution the random walk model , i.e. $F_t = I$, was selected. In section 8.5.1, the problem of selecting the appropriate number of eigenvectors is treated. Here it is shown that a few eigenvectors are enough to model dynamic variability. Focus is given on the comparison of Kalman filter and smoother, as well as on the applicability of the methods for tracking different trends. In this section, the matrix H_s can be treated as preselected or fixed.

8.4.1 The model and practical considerations

From the eigenvalue decomposition of the matrix $\hat{R} = ZZ^T/T$, where Z is the matrix with columns the measurement vectors, a few dominant eigenvectors can be selected to form the columns of the time invariant observation matrix H_s. The first 5 dominant eigenvectors were here selected for every simulated set individually. For better estimation performance, before obtaining the eigenvectors, the noisy simulations have been smoothed with the smoothing operator in (4.86), i.e. $z'_t =$

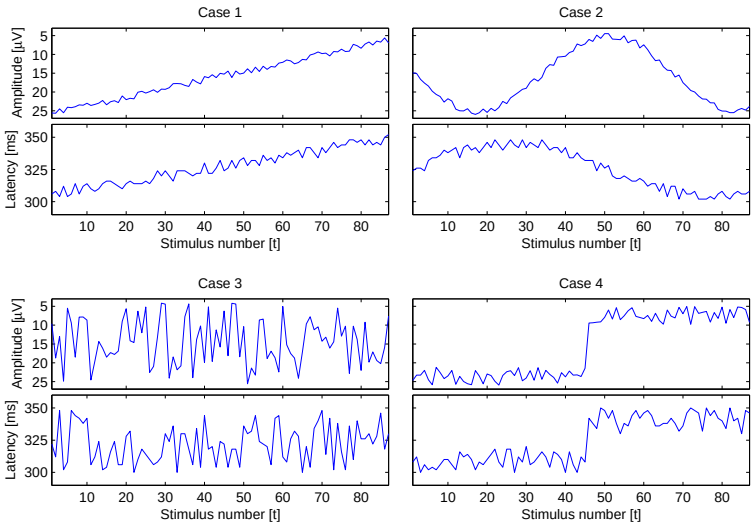

Figure 8.1: Amplitudes and latencies of the third peak for the (noiseless) simulations. Case 1: slow linear trends, Case 2: faster sinusoidal, Case 3: only random variations uniformly distributed, and Case4: a jump in amplitude and latency.

Gz_t, where $G = (I + \alpha^2 D_1^T D_1)^{-1}$, and D_1 is the first order difference matrix. The regularization parameter was selected as $\alpha^2 = 25$ by visual inspection of the eigenvectors. Therefore, the matrix H_s was formed based on z_t', but for estimation the original simulations were used. Estimates for the parameters θ_t were computed with Kalman filter and smoother algorithms, and for s_t as $\hat{s}_t = H_s \hat{\theta}_t$.

For the covariance matrices the selection $C_{\omega_t} = \sigma_\omega^2 I$ and $C_{\upsilon_t} = \sigma_\upsilon^2 I$ was made, for every $t = 1, \ldots, 87$. Therefore, the selection of σ_ω^2 is not essential since only the ratio $\sigma_\upsilon^2/\sigma_\omega^2$ has effect on the estimates. Then the choice $C_{\upsilon_t} = I$ can be made and care is given to the selection of σ_ω^2. If it is tuned too small, then the estimates have bias toward the previous estimates. If it is selected too big, then they have too much variance tending to be similar to the ordinary LS estimates. The same parameter controls the adaptivity or the over-smoothing of Kalman smoother. The selection can be based on experience, expected variability, and visual inspection of the estimates. If the variance of the background noise change randomly from trial-to-trial, a time-varying selection is not expected to improve the overall properties of the estimates. It may improve some individual single-trial estimates mainly related to much lower or much higher signal-to-noise ratio conditions. Intuitively, very noisy single-trials are better estimated based on the past, or past and future measurements, while less noisy based on the present. If the background noise is changing with a clear pattern, for example, if its variance grows from trial-to-trial, the model variances should be selected accordingly.

The influence of the initial values θ_0, C_{θ_0} can be reduced by using the algorithm first backwards in time. The last estimates of the backward run can then be used to initialize the forward run. Usually, around 10 backward steps are enough for a good starting point. Since we have a time invariant signal model, Kalman filter matrix equations can be iterated until an adequate convergence is observed. The speed of convergence relates to the initial conditions, properties of state-space model, and state and observation noise covariance matrices. More about the convergence of Kalman filters can be found for example in ([71], p. 286, [7]). The converged values can be used to initialize the backward run. The average vector of the ordinary LS estimates was used for initialization of the backward procedure (40 backward steps). With this initialization, in the limited case of a very small state noise variance parameter, $\hat{s}_t$ will vary slowly around the mean of the measurements. For bigger values the initialization is less important, and it is visible for a few first estimates. Though, the smoother is less sensitive to initializations. The same procedure was used for every initialization in the thesis.

8.4.2 Error comparison and state-noise variance parameter selection

In order to identify optimal values for the variance term $\sigma_\omega^2 = \sigma^2$, root mean square errors (RMSEs) were calculated, between the estimates based on the noisy data and the noiseless simulated EPs (for all the different cases, and for both Kalman filter and smoother). This also describes the performance of Kalman filter and smoother in relation to the parameter selection. In parallel, in order to investigate the performance of the methods when the noise is not present, new estimates were computed based on the noiseless data, and new RMSEs between these estimates and the noiseless simulations. The same observation model, based on the smoothed noisy data, was used for these estimates. This aims to describe the damage on the hidden variability due to over filtering or over smoothing. A balance between these two considerations is the guide for optimal selection. Since focus is given on the third peak, the RMSEs were computed over the time interval (250-400ms). The means of the RMSEs for different values of the variance term are presented in Fig. 8.2 for all cases.

In all situations, the smoothing method results in significantly smaller errors when the noisy measurements are used. For Cases 1, 2 and 4 the error measure takes clear minimum for both filter and smoother. Case 1, which simulates the slowest transition, is better estimated for values around 0.01. Case 2 and 4 require bigger values around 0.05. The damage is naturally bigger for the sudden transition (Case 4). These values give a compromise between noise reduction and tracking performance. Additionally, it can be observed (see also Section 8.4.4) that the difference in both estimation methods is due to two factors. Smoother reduces greater the noise having in the same time better tracking performance. For Case 3, although noise reduction is achieved, clear minimum cannot be observed. A value bigger than 1 is reasonable with respect to the damage of the random variability of the peak. Past and future measurements contain information about the shape and sign of the EPs that is used to reduce the noise. Though, in that case the damage is bigger for the smoother.

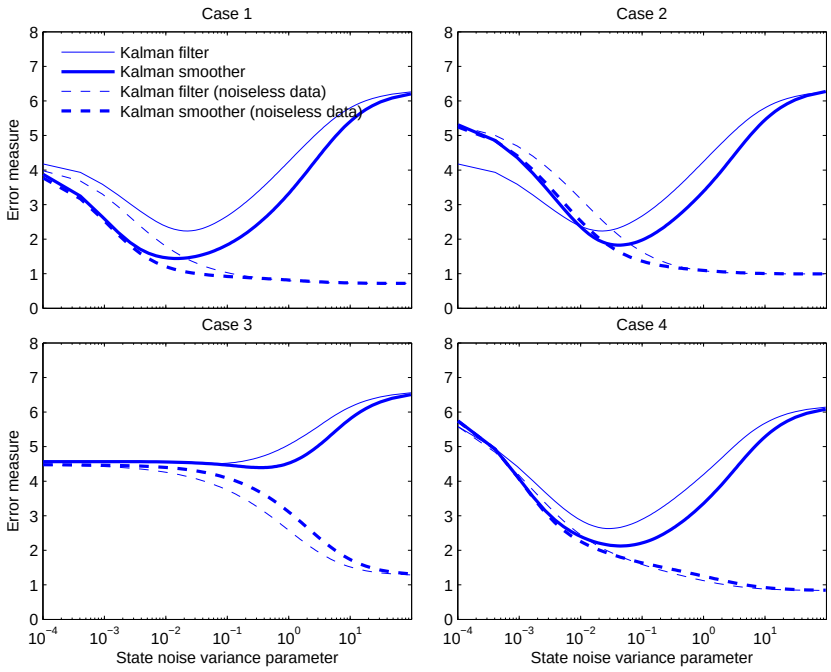

Figure 8.2: Mean of the root mean square errors for different values of the state noise variance parameter σ^2, where $C_{\omega_t} = \sigma^2 I$, and $C_{\upsilon_t} = I$ for every t. RMSEs were computed between the estimates and the noiseless simulated EPs in the interval 250-400 ms (third peak). The x-axis is in logarithmic scale. Observation model: 5 eigenvectors of the data correlation matrix.

8.4.3 Single-trial estimates and applicability revised

By construction, the synthetic datasets have similar means, but different higher order statistics and time correlations. Different plots describing the simulated EPs are presented for Case 1, Case 2, Case 3, and Case 4 in Figs. 8.3-8.6 respectively. The general structure and dynamic variability of the EPs is estimated by both Kalman filter and smoother. The extra noise reduction by the smoother is visible in all the image plots. Excellent estimates are obtained under poor signal to noise ration conditions (Case 1 after stimulus 70, Case 2 stimuli 40-70, second half of Case 4). With bigger values for the variance parameter noise is still highly present (Case 3). In that case also over filtering and especially over smoothing is observed. But even in that case the estimates preserve the general structure of the peak. The effect of the filter and smoother for different values of σ^2 can be observed by comparing the estimates for the first two peaks. The first two peaks are, by construction, better estimated in Case 1.

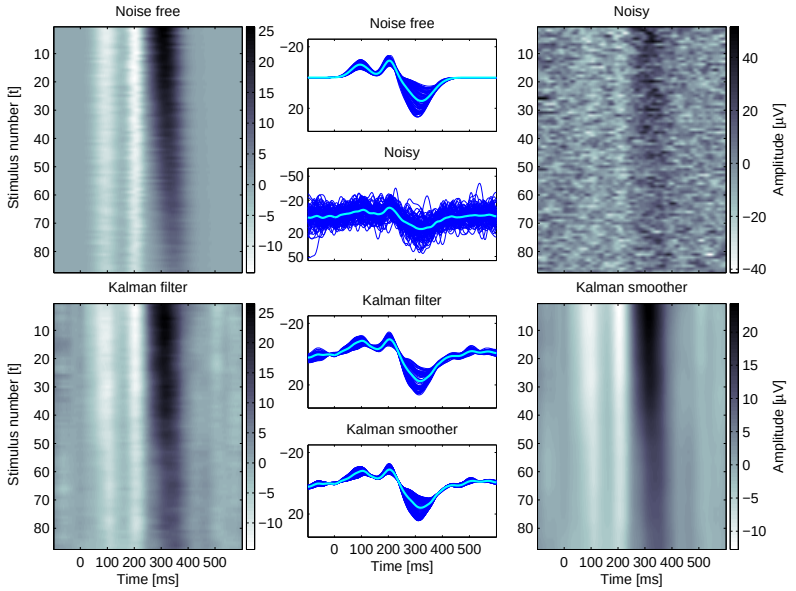

Figure 8.3: Case 1: Slow trends, $\sigma^2 = 0.01$.

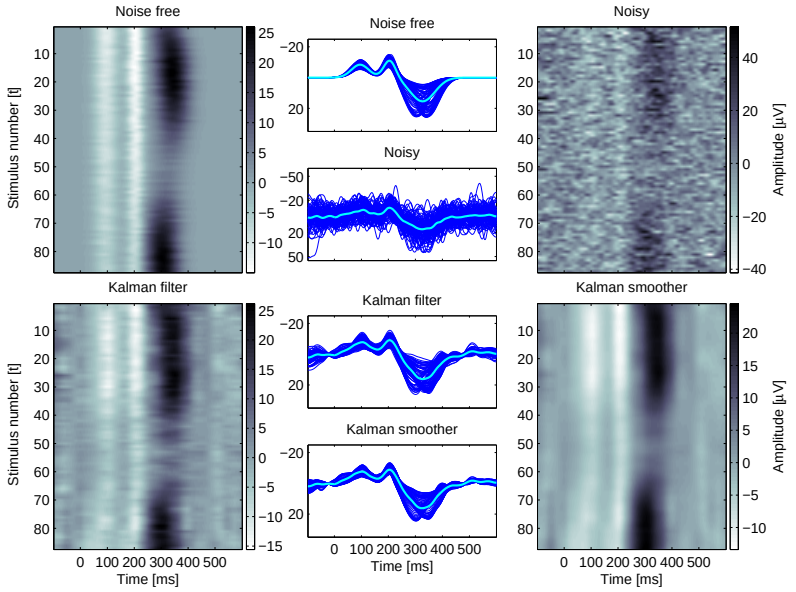

Figure 8.4: Case 2: Sinusoidal trends, $\sigma^2 = 0.05$.

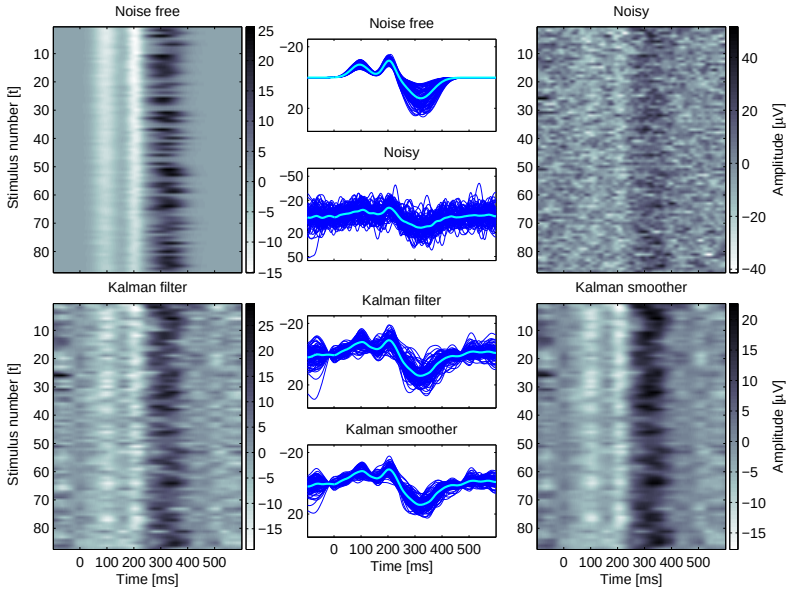

Figure 8.5: Case 3: Random variability, $\sigma^2 = 1$.

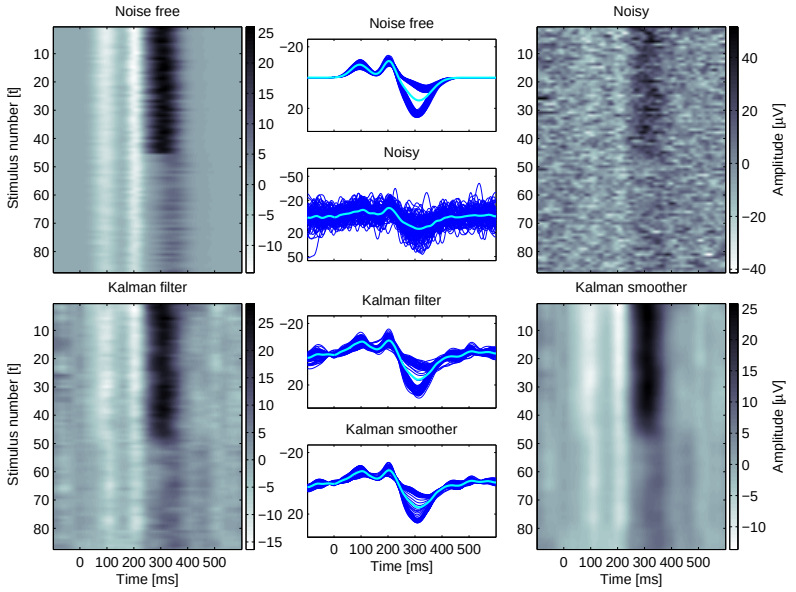

Figure 8.6: Case 4: Sudden jump, $\sigma^2 = 0.05$.

Estimates for the latency and amplitude of the third peak were also computed, by simply taking the maximum value in the interval 250-400ms. The estimates for the different cases are presented in Figs. 8.7-8.10. The tracking capability of the methods can be compared from these plots. The amplitude and latency estimates are presented on top of each figure, and estimates based on noiseless data on bottom. For the first two cases the reduction in noise is greater for Kalman smoother, which provides better estimates for the latencies when the signal-to-noise ratio is poor. Other significant observation is that it cancels the time delays that the filter is causing (see also e.g. [193, 190]). With that selection of parameter, for Case 2 the amplitudes are over smoothed towards the mean. For Case 4, Kalman filter tracks well until the jump and creates a time delay after that point for both amplitude and latency. The smoother tracks better the sudden latency change, but oversmooths the sudden amplitude transition.

Case 3, with only random variability, represents the worst performance for the methods. Although, there can be observed reasonable estimates for the latencies. This is partly due to the selected observation model, which naturally contains prior information about the range of the peak, and up to some level about the shape of the peak. This depends also on the number of the eigenvectors included in the model. Less eigenvectors preserve better the shape, but they might not be able to model late potentials, and long latency trends. More eigenvectors introduce noise. For small values of σ^2 the selection is less important since prior information of the observation model becomes inactive, as prior information from the state evolution becomes dominant (see also section 8.5.1). The amplitude range is underestimated and a bigger value for σ^2 is perhaps more appropriate. This can improve individual estimates for strong EPs, but it will keep hidden into the noise the smaller ones.

From the analysis of the simulations some general observations can be derived. Recursive estimation by Kalman filter can model dynamic variability and creates accurate estimates for the simulated EPs. Though, for small values of the state variance parameter creates time delays in the estimates. Kalman smoother cancels these delays, and in the same time removes greater portions of noise. Both have as an effect reduction in the mean square error. Though, the biggest part of the noise reduction is achieved during the filtering procedure. For slow variations, Kalman filter and smoother give very good estimates. Relatively slow variability permits the choice of small values for σ^2, which leads to highly improved signal-to-noise ratio. Bigger values for the state noise variance can still give good estimates, but more noisy. The clear superiority of Kalman smoother makes necessary the use of filter only in situations that the estimation is needed on-line (for example to monitor depth of sedation in clinical applications). Then improvement in the tracking capability is related to the fixed-lag smoother. When only sudden and large random variations exist, bigger values for the state noise covariance are appropriate in order to keep the random structure of the peak. Kalman filter seems to track better some sudden jumps in the expense of more noise in the estimates. In those situations extra prior information is needed for better estimation. In addition, time-varying models should improve the tracking performance.

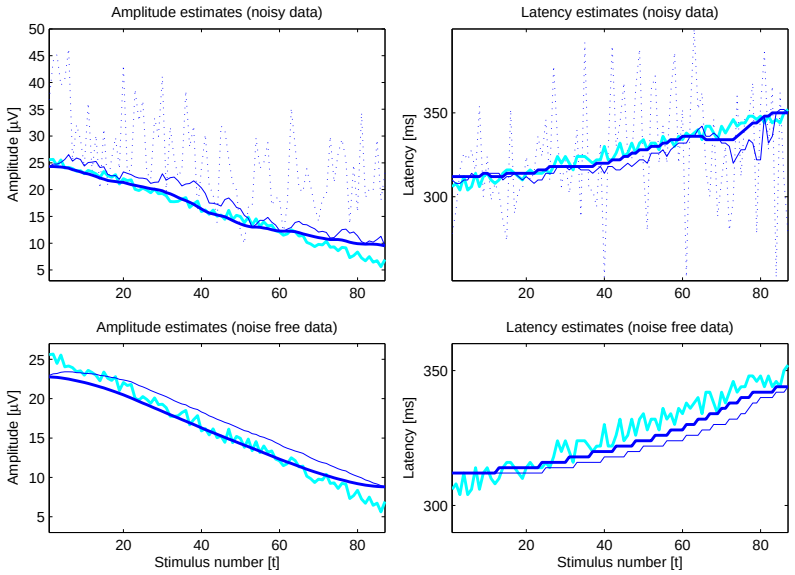

Figure 8.7: Case 1: amplitude and latency values, true values (gray), raw data values (dotted), Kalman filter estimates (thin), and smoother (thick).

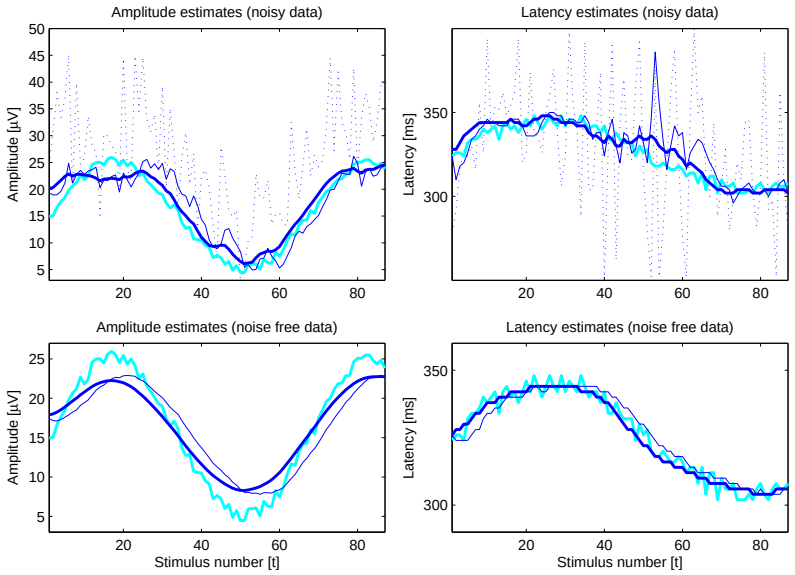

Figure 8.8: Case 2: line description as in Fig. 8.7.

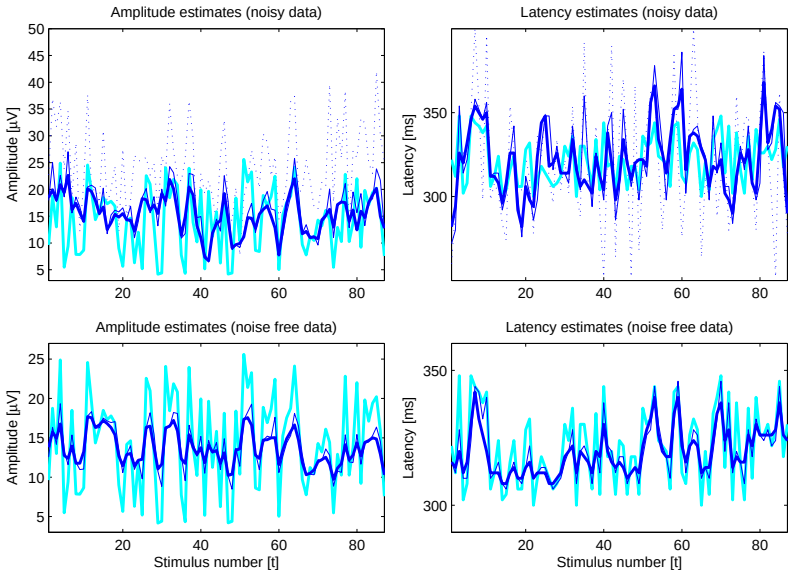

Figure 8.9: Case 3: line description as in Fig. 8.7.

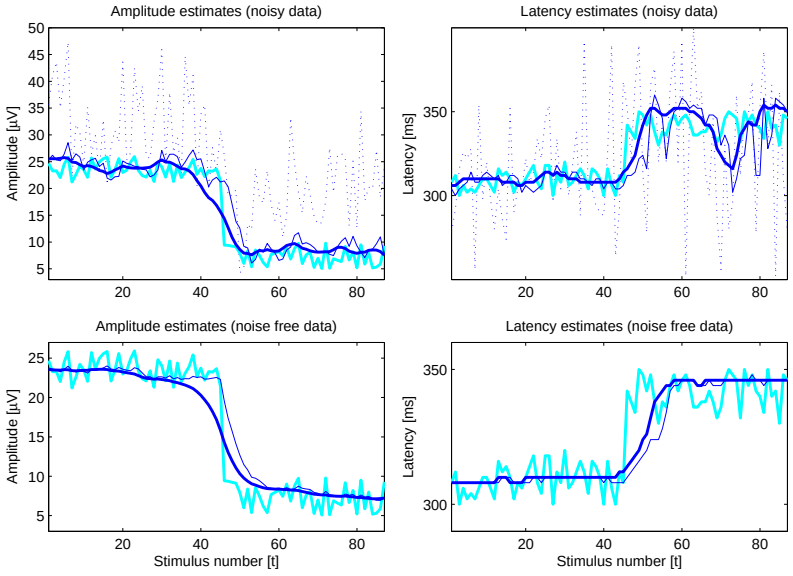

Figure 8.10: Case 4: line description as in Fig. 8.7.

8.5 On the selection of observation model

In the previous section, the applicability of recursive mean square methods in estimating dynamic variability present in the measurements was demonstrated. Since Kalman smoother is essentially based on the filtering procedure, its better performance do not seem to depend on the selection of observation model, neither to the state evolution model.

8.5.1 Number of eigenvectors

The number of eigenvectors included in the previous analysis contains prior information about the overall mean and latency range of the peaks. Note that by construction the simulations are covering the same range. In Figs. 8.11-8.14, there are presented the dominant eigenvectors (up to 9) for the four simulated cases. In the first column of each figure, there are the eigenvectors computed from the raw simulated measurements (top row), after conservative smoothing (middle row), and from the noise free simulations (bottom). On the other two columns there are error measures based on different number of eigenvectors (average RMSEs between estimates and noiseless simulations for the third peak). For small state variance parameter selection, prior information embedded in the observation model concerning individual peaks becomes inactive. Though, at least 2-3 eigenvectors are needed in order to estimate latency variation for the 3 peak.

If the assumption is that the EPs should be smooth transient waveforms, then too many eigenvectors do not contribute to the accuracy of the estimates. When big state noise variance parameter is selected, better estimates are obtained with less eigenvectors (preserving shape information). Since even perfect (noise free) eigenvectors can largely model the noise (in the least-squares sense). Less dominant eigenvectors model mainly random variability of the peaks, and details related to the width and exact shape. Therefore, the associate parameters have mainly uncorrelated from trial-to-trial behavior, and as a consequence they become easier eliminated in the filtering-smoothing procedure. Therefore, for small σ^2 the introduction of extra eigenvectors do not change the quality of the estimates. Additionally, less dominant eigenvectors, when are based on raw measurements, they largely correspond to the noise subspace. A combination of these considerations gives the number of eigenvectors for estimating a single peak.

In relation to the assumptions for the speed of change of a particular peak, someone can control the quality of estimates by introducing a diagonal matrix for the state noise covariance, with different values associated to different eigenvectors. For example, in Case 1, first and second noiseless eigenvector model the main amplitude decrease and latency shift. Therefore, they could be tuned with different values. Though, with noisy eigenvectors this is rather difficult (especially in real unknown measurements) since their order and shape do not always match to intuition. Even dominant eigenvectors can correspond largely to the noise subspace, while containing information about the EPs. Notice that in Case 1 and 2 the order of the second eigenvector has changed to third when noise is added.

Another issue worth mentioning relates to the correlation between individual

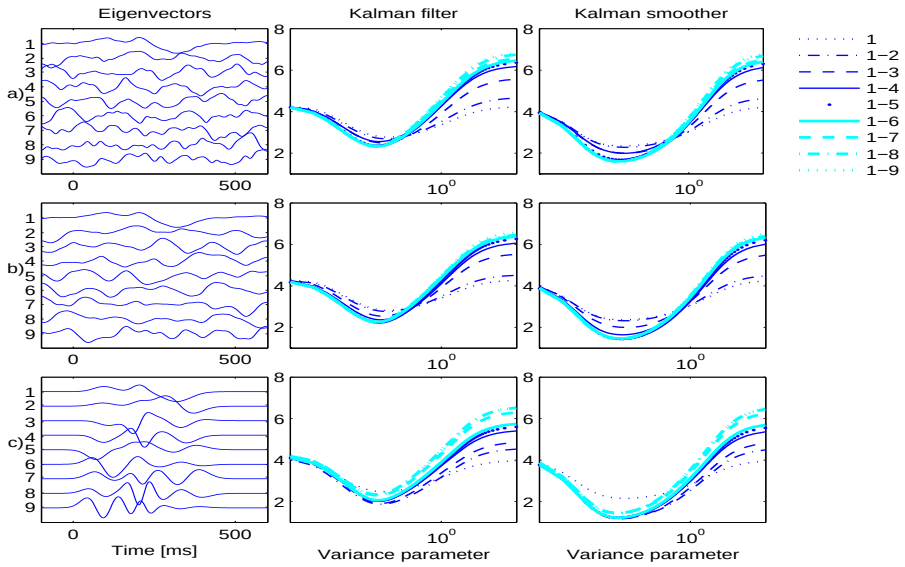

Figure 8.11: Case 1: Eigenvectors and means of the RMSEs for different dimensions of the observation model when noisy (top), smoothed (middle), and noise-free data (bottom) are used for computing the eigenvectors.

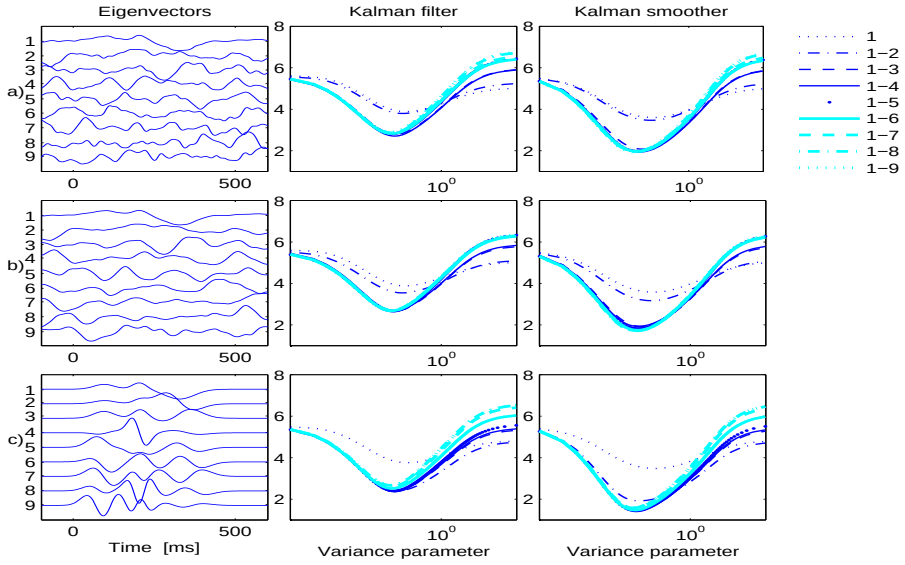

Figure 8.12: Case 2: figure description as in Fig. 8.11.

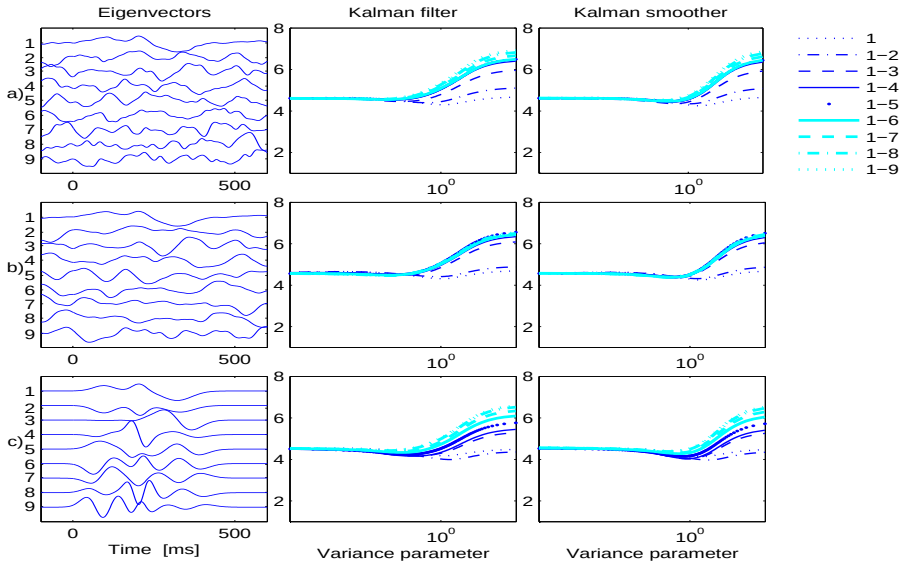

Figure 8.13: Case 3: figure description as in Fig. 8.11.

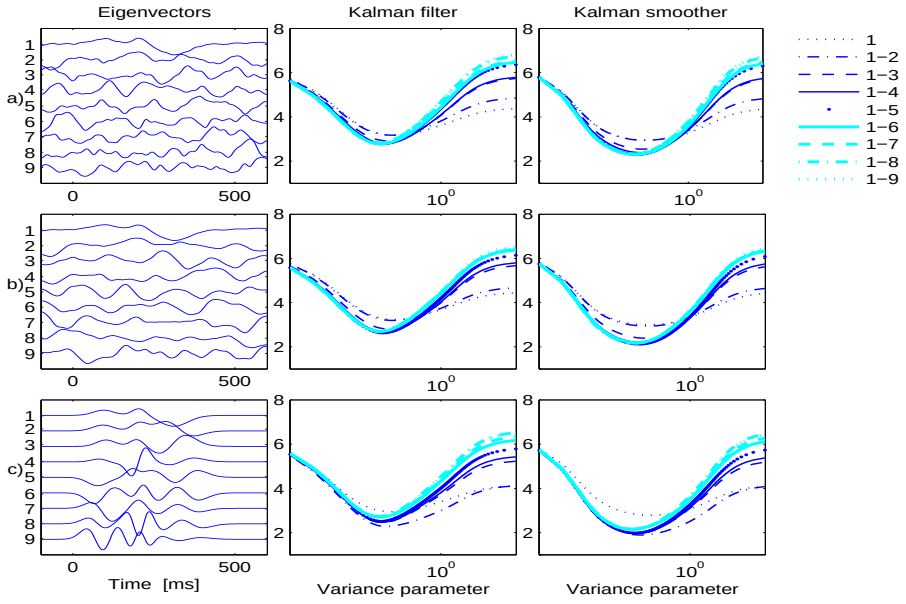

Figure 8.14: Case 4: figure description as in Fig. 8.11.

peaks. Naturally the eigenvectors contain prior information about the correlation structure between the wave-shapes of the hidden source signals. If the peaks are largely correlated in amplitude, and especially, in latency changes then perhaps eigenvectors are an optimal choice (in comparison to other generic observation models) for reducing the dimension of the problem and for dynamic estimation.

8.5.2 Generic vectors

Other generic parametrizations can also be used for estimating features of EPs. Here some examples are presented by considering only simulated Case 1. As a starting point for a demonstration, an observation model formed by time shifted and scaled (sampled) Gaussian functions was selected. The selection was not intended to be optimal in any sense, but reasonable by visual inspection of the obtained estimates. Thus the observation matrix becomes $H = [\psi_1, \ldots, \psi_n]$, where its columns are formed by

$$\psi_i = e^{-(\tau - a_i))^2/2b^2}, \tag{8.19}$$

The width parameter was chosen $b = 20$ and $n = 20$. The basis vectors are plotted in the upper left part of Fig. 8.15.

It is known in regularization theory that the effect of the regularization parameter on the estimates relates to the singular values of the observation matrix (e.g. [39], p. 59). For example, if the singular values of H are much larger than the regularization parameter then the regularization parameter has little effect to the solution. So at least the largest singular value must be considered, for example, in order to compare the performance of two different observation models. However, this is true when the model $F_t = I, C_{v_t} = I, C_{\omega_t} = \sigma^2 I$, is considered. Even the introduction of diagonal matrices can change the situation, making the comparison more difficult.

Furthermore, the singular value decomposition of the matrix H was considered, i.e. $H = U\Sigma V^T$, where $U = (u_1, u_2, \ldots, u_M) \in \mathbb{R}^{M\times M}$ and $V = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^{n\times n}$ are orthogonal matrices and Σ is a pseudo-diagonal M by n matrix whose top n rows contain the singular values $\mathrm{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ (ordered $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$) and whose bottom $(M - n)$ rows are all zero. From the SVD of H a second observation model was formed $H_u = [u_1, u_2, \ldots, u_n]$, with columns the left singular vectors. This matrix is shown on the top of Fig. 8.15 (b). Finally, a third observation model was formed as $H' = H_u V^T$, and it shown on the top of Fig. 8.15 (c). Note, that the matrices H, H_u, H' provide the same LS fitting error. In order to have a reasonable, but not exact, comparison with the original matrix, each column of H was normalized to unit norm.

The columns (length $M = 350$) of the observation models are shown in Fig. 8.15 (top row). In the second row there are plotted the LS parameter vector estimates $\hat{\theta}_{t,LS}$, for every $t = 1, \ldots, 87$, $k = 1, \ldots, 20$, obtained from the noiseless simulated Case 1. The parameters $\hat{\theta}_{t,LS}$ are given in absolute values for better visualization of the parameter space. Parameter estimates (based on the noisy simulated Case 1) obtained with Kalman smoother are given in the third row. The state noise variance was selected to be 0.01 for every observation model. Kalman

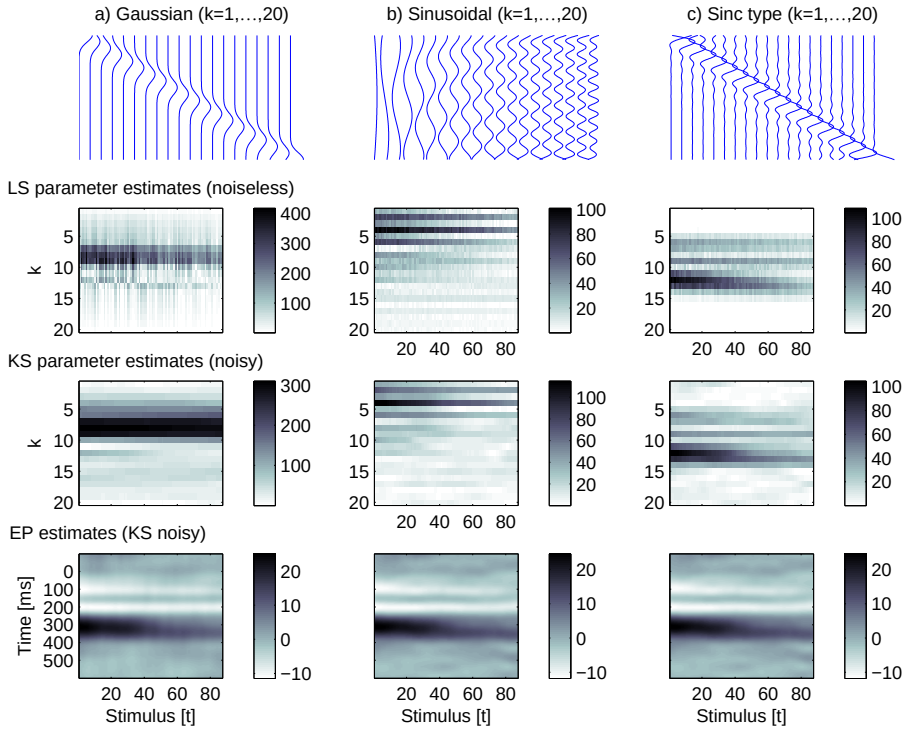

Figure 8.15: Case 1: Different parametrizations of the EP estimation problem. First row: basis vectors, Second row: LS parameter estimates (absolute values) based on noiseless simulations, Third row: KS parameter estimates (absolute values) based on noisy simulations, Last row: EP estimates (KS, noisy).

smoother estimates for the simulated EPs are give at the last row. For observation model b) and c) the EP estimates are identical. Note that the vectors in H_u, H' are orthonormal. The Gaussian humps a) have the ability to provide smoother results (bottom). However, the resulting estimates $\hat{s}_t$ for the 3 observation models tend to be the same for very big or very small values for the state noise variance parameter.

This example aimed to emphasize that although the LS fit can be the same for some observation models, when the state space model is considered the results can be different. Additionally, even though the base and parameter space look different (model b) and c)) the estimates for EPs can be exactly the same. These observations are important for imposing extra prior information for estimation or for state-space identification. The Gaussian functions can form a useful all purpose model, for example for online estimation. The base in column b), is of the

form of discrete sinusoidal type functions of different frequencies. Therefore, the parameter estimates describe in that example decreasing signals from trial-to-trial of different frequencies. The base in column (c), is of the form of time-shifted discrete sinc type functions, uncorrelated, with band pass type discrete Fourier transform. Roughly speaking, the parameters represent the time-varying behavior of the power in a frequency band within a trial and from to trial-to-trial. Therefore, different observation models (e.g. band pass filters) can be designed focusing on a particular application.

8.5.3 A smoothness priors evolution model

Different observation models can force smoothness for the EPs. Though they require the non trivial selection of wave shape and number of vectors in the observation model. Extra smoothness for the EPs can be enforced directly via the smoothness priors method. The simplest observation model for estimation is given by $H = I$. Thus, it is $\theta_t = s_t$. Estimates for Case 1 with the same variance parameter as before ($\sigma^2 = 0.01$) are shown on the left of Fig. 8.16. Prior information about the smoothness of the EPs can be introduced through the state evolution model (5.135)

$$F' = (I + \sigma^2 \alpha^2 D_1^T D_1)^{-1}, \quad C'_\omega = \sigma^2 F', \tag{8.20}$$

where D_1 is the first order difference operator. This model can be understood as a two dimensional smoothing method for EP estimation. For the selection $\sigma^2 = 0.01$ and $\alpha^2 = 1000$ the improved estimates are presented in the middle of Fig. 8.16. As a comparison, estimates obtained with observation model 5 dominant eigenvectors of the noiseless data correlation matrix are presented on the right of Fig. 8.16.

8.6 State-space identification

In the previous sections different parametrizations for state-space modeling and estimation of dynamic features in EP measurements were considered. For every parametrization improvement of the tracking capabilities is related to the introduction of time variation in the model. In this section, the performance of the state-space identification method presented in section 5.4 is demonstrated. Its form is general and can be modified according to the parametrization.

In order to continue the comparison based on the four simulated cases, as before the 5 dominant eigenvectors are used for estimation. The simplest time-varying prior for the enhancement of the random walk model ($F_t = I, C_\upsilon = I, C_\omega = \sigma^2 I$) is then (see section 5.4) $C_t = c_t^2 I$ and the model becomes

$$F'_t = \frac{c_t^2}{c_t^2 + \sigma^2} I = \frac{1}{1 + \sigma^2 \phi_t^2} I = f_t I, \tag{8.21}$$

where it was set $\phi_t^2 = \frac{1}{c_t^2}$ and

$$C'_{\omega_t} = \sigma^2 f_t I = \sigma^2 \frac{1}{1 + \sigma^2 \phi_t^2} I \tag{8.22}$$

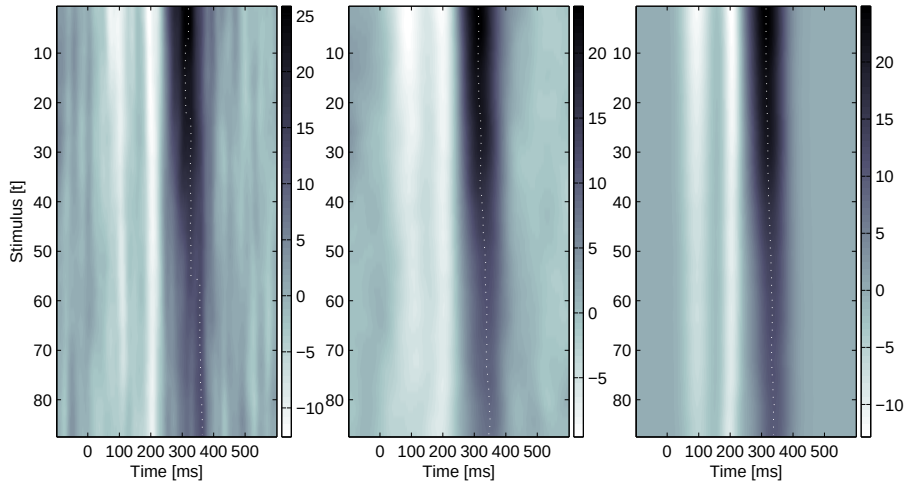

Figure 8.16: Case 1: Prior information only about the dynamic variability (left), extra smoothness (middle), 5 eigenvectors from noise free simulations (right).

Estimates for the parameters ϕ_t^2 for $t = 1, \ldots, T$ can be obtained with the algorithm presented in section 5.4. It must be noted that this parametrization is self constrained not to over-fit to the measurements, neither to over-smooth the data. Since for big values of ϕ_t^2, state noise variance becomes small (thus over smoothing) and the estimation criterion gets bigger, and small values of ϕ_t^2 lead to the random walk model. Initialization was based on the random walk model. Thus, a local minimum was searched in order to provide tracking improvements over the basic random walk model. Though, the prior selection of σ^2 is still necessary.

In Fig. 8.17, it is presented as before the error measure as a function of the state variance parameter σ^2 for the 4 simulated cases. The optimization was performed in the interval 0-500 ms. The error is smaller than the filter. Near the optimal choice for the parameter the presented state-space identification method and the introduction of time variation at the model seems to reduce the error (smoother, random walk model). The selection of the variance parameter seems to be less important near the optimal since the adaptivity is improved. For big value of σ^2 naturally bigger jumps are possible, which has as a consequence that the estimates may follow also the noise. Inevitably the performance of the method relates to the properties of the noise. Though the damage in the trends is always smaller.

For Case 1 it allows the use of even smaller values for the variance parameter which leads to greater noise reduction. For the same value as before ($\sigma^2 = 0.01$) the error is improved by reducing the noise, tracking better the linear trends and introducing some random variability in the estimates (Fig. 8.18). Though, for the latencies there are very small differences. For Case 2 (Fig. 8.19), similar

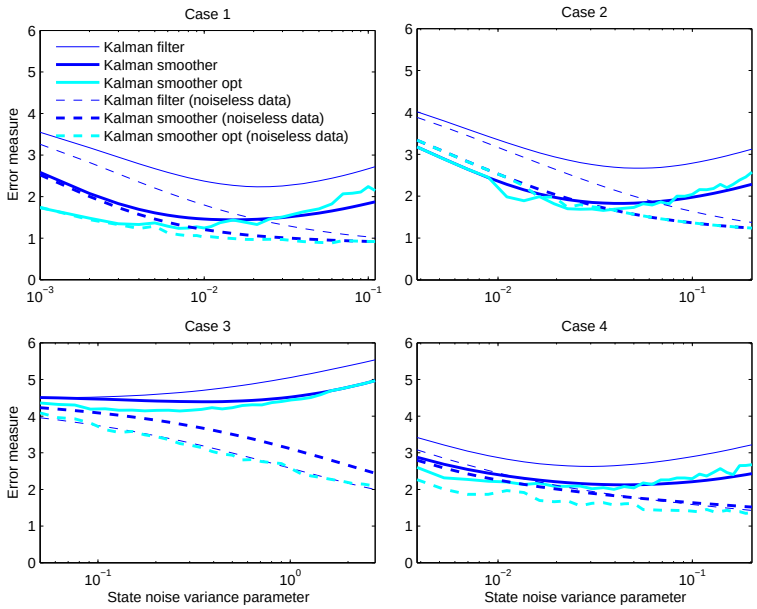

Figure 8.17: Error measures as a function of the variance parameter for the four simulated cases.

observations can be made ($\sigma^2 = 0.05$). For Case 3 (Fig. 8.20), the damage is comparable to the filter ($\sigma^2 = 1$). However, the identification method models faster changes with greater noise reduction than the filter and plain smoother. For Case 4, the estimates are overall improved (Fig. 8.21). For the selected parameter ($\sigma^2 = 0.05$), and the noise levels, the method could not estimate so sharp jump but rather broke the amplitude jump into two smaller ones. Though now in the epoch plots and in the image the sudden transition is more clear. The estimated models f_t are shown on the top left corner of Figs. 8.22-8.25. These figures also present three individual single-trial estimates. Visual observation of the raw data for the decision of amplitude and latency of peaks can be misleading (Stimulus 25) even in simple pseudo real simulated examples. A compromise between following too closely the measurements and respecting prior information gives better estimates.

Next, the performance of the methods is tested under different noises. In Fig. 8.26 mean RMSEs after repeated simulations for Case 1 (left) are presented. Different noise levels were considered (Gaussian noise with different variances) and the error measure was computed for least squares, Kalman filter, smoother and smoother optimized (the data were always smoothed before computing the eigenvectors). In this plot, the error measure is presented as a function of the standard deviation of the added observation noise. The state noise variance parameter for

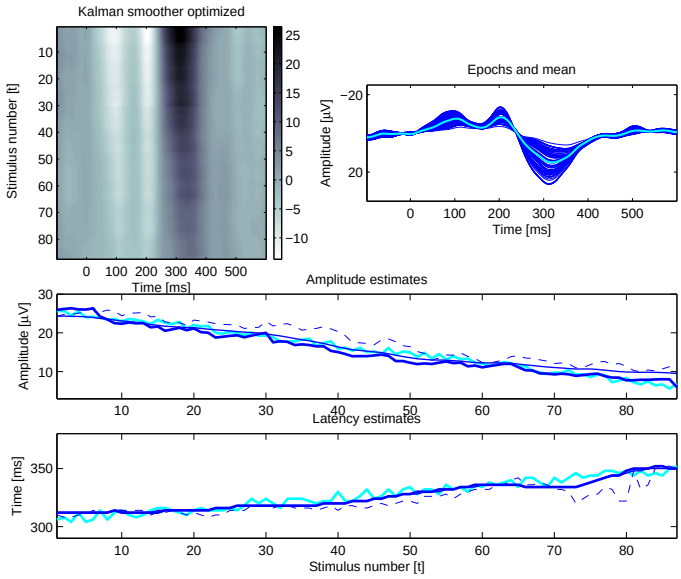

Figure 8.18: Case 1: estimates, Kalman filter (dashed lines), smoother (thin lines), smoother optimized (thick lines), and the simulated variability (gray).

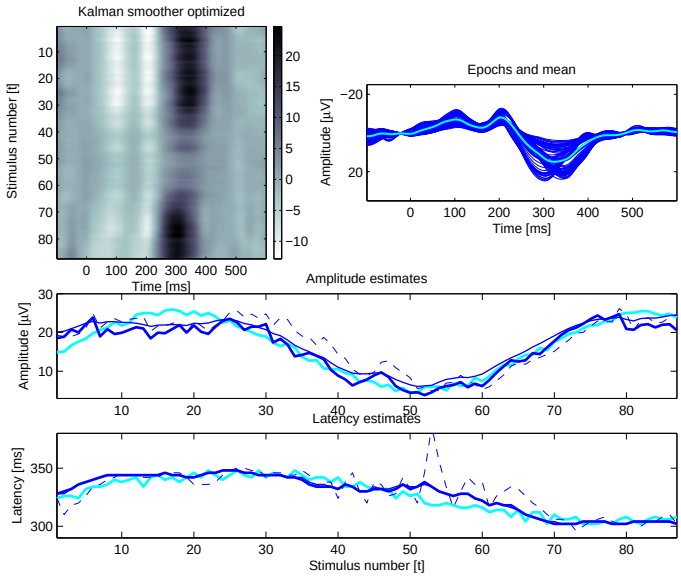

Figure 8.19: Case 2: line description as in Fig. 8.18.

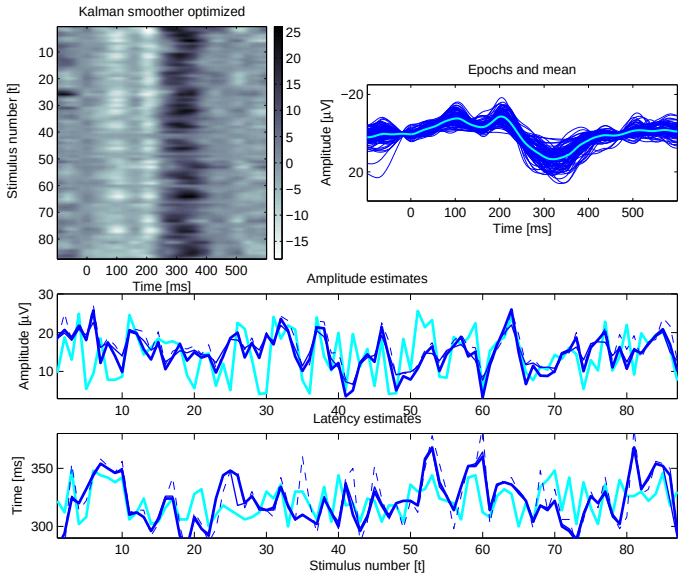

Figure 8.20: Case 3: line description as in Fig. 8.18.

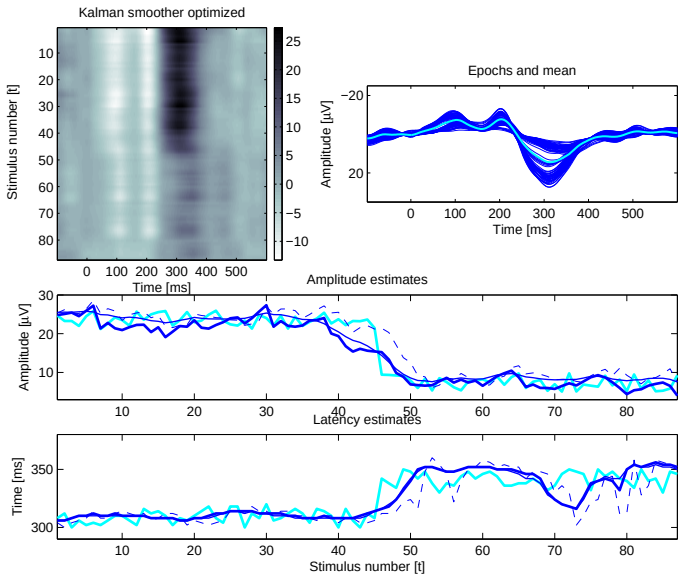

Figure 8.21: Case 4: line description as in Fig. 8.18.

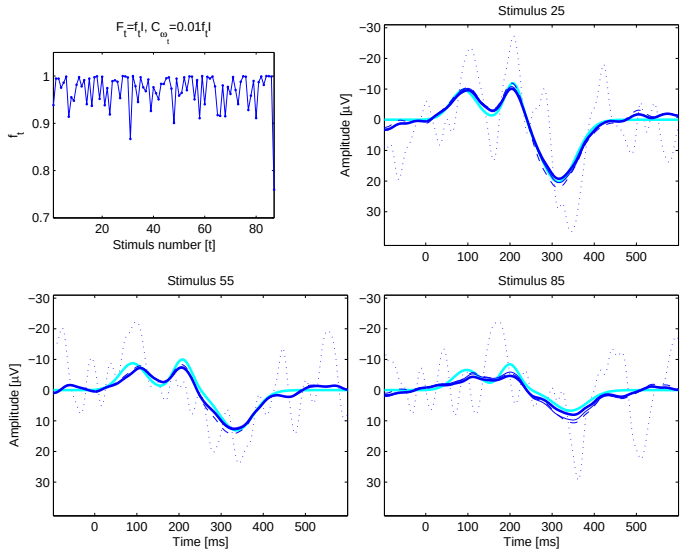

Figure 8.22: Case 1: estimated model F_t, C_{ω_t} and three single-trial estimates, noisy simulations (dotted), Kalman filter (dashed), smoother (thin), smoother optimized (thick), noise free simulations (gray).

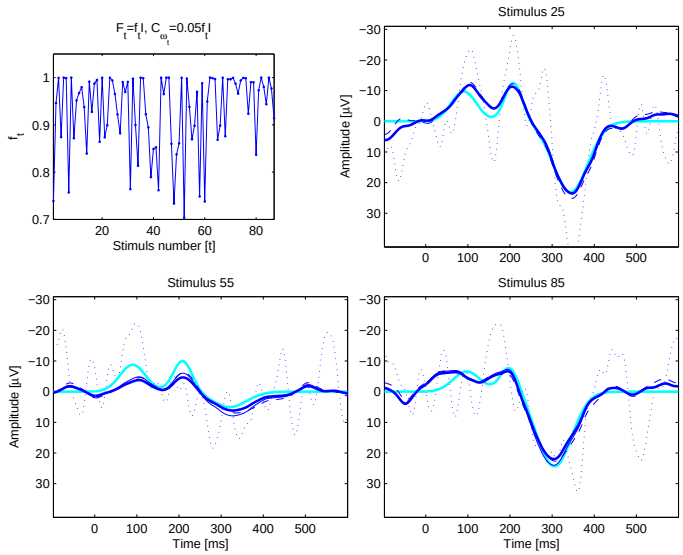

Figure 8.23: Case 2: line description as in Fig. 8.22.

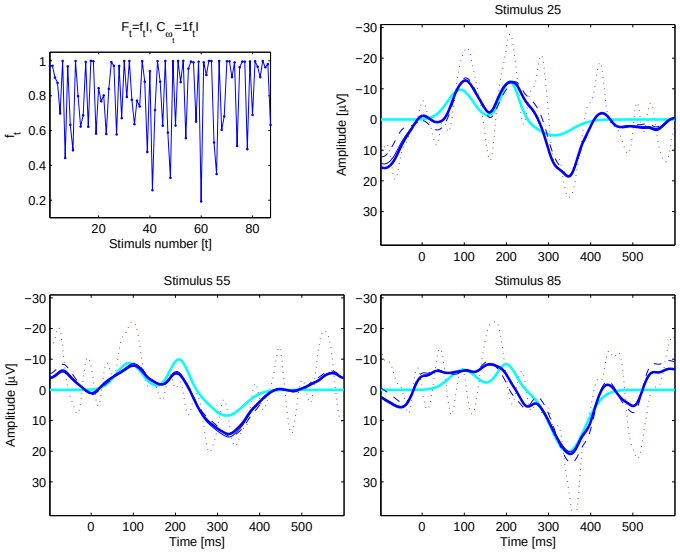

Figure 8.24: Case 3: line description as in Fig. 8.22.

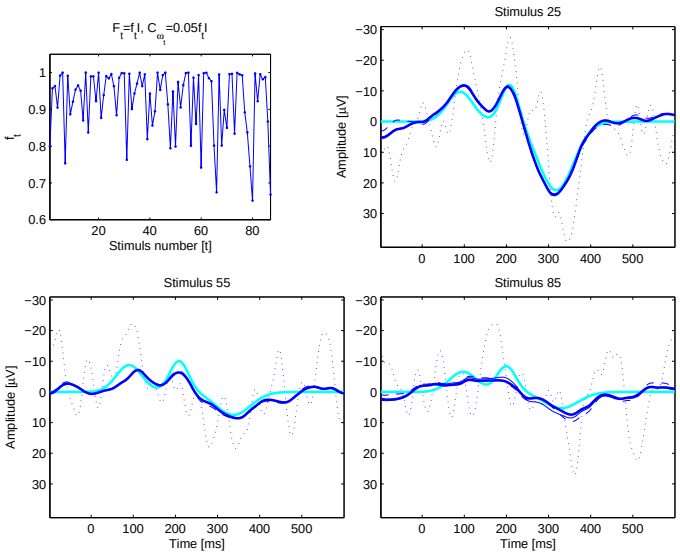

Figure 8.25: Case 4: line description as in Fig. 8.22.

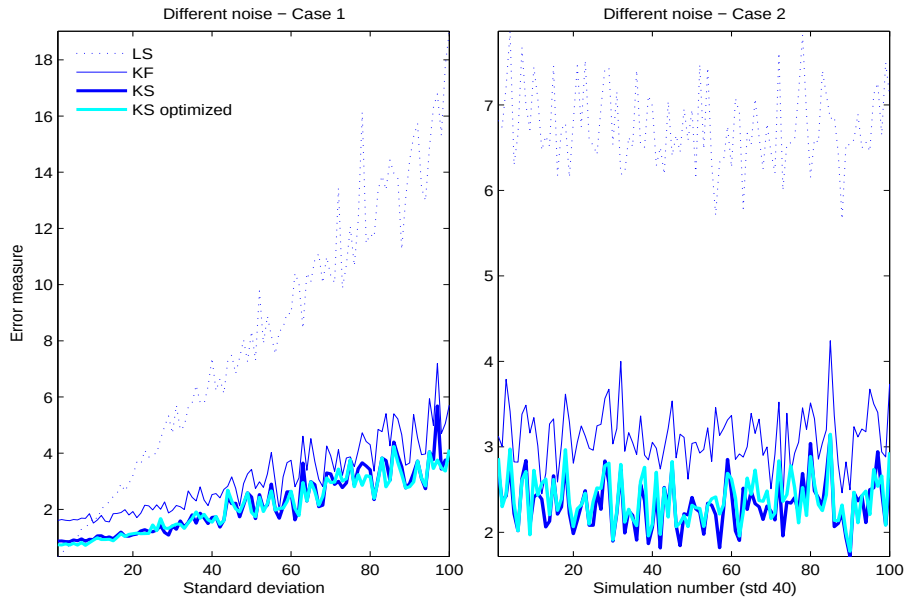

Figure 8.26: Error measure for different realizations of background noise.

estimation in all the simulations was fixed ($\sigma^2 = 0.01$). In the same Fig. 8.26 (right) estimates with ($\sigma^2 = 0.05$) were computed for Case 2 for different Gaussian noises (100 set of realizations) with fixed standard deviation (STD= 40). In Figure 8.27 it is presented the ensemble mean of all the estimates of amplitude and latency (thin lines) for Case 2 only. Dashed lines denote intervals of ± 2 times the (ensemble) standard deviation and thick lines the simulated trends. The time delay of the filter is visible, the state space method seems to over all correct the smoother. The variance of the estimates is kept small even in poor signal-to-noise ratio conditions. The state identification method seems also robust, and given that it fixes individual EPs, corrects the trends, and controls over smoothing it seems overall preferable. For higher signal-to-noise ratio conditions and for bigger values of the state variance parameter the method can model accurately sudden jumps of the state evolution (see Fig. 8.28).

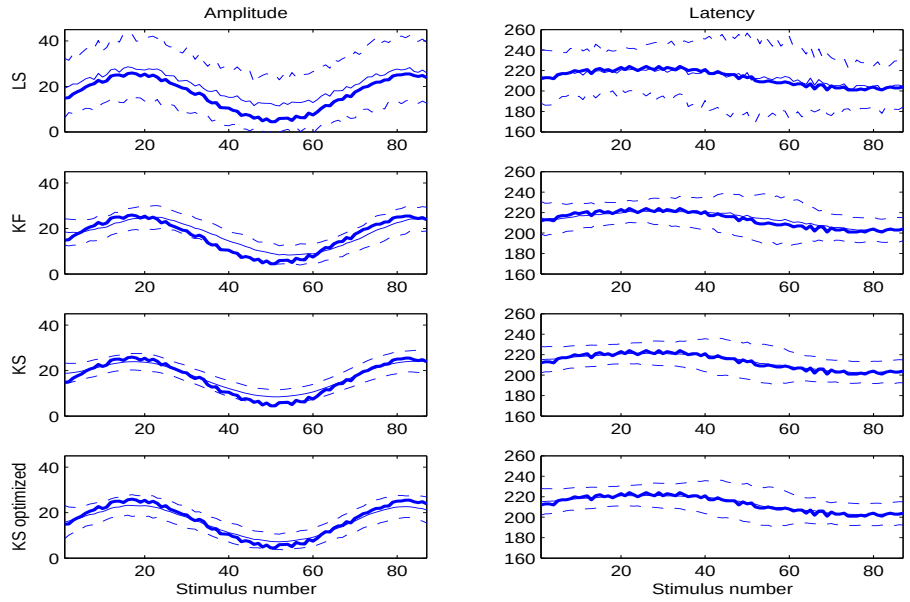

Figure 8.27: Case 2: amplitude and latency ensemble means $\pm 2STD$ from estimates based on simulations with different noises, see also Fig. 8.26 (right).

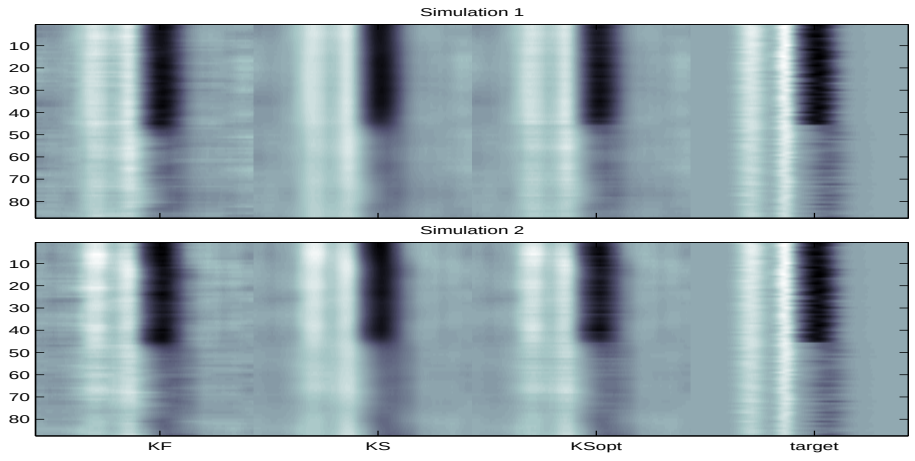

Figure 8.28: Case 4: two different Gaussian background noises with standard deviation 20 and $\sigma^2 = 0.1$. The third column shows the improvement due to the presented state-space identification method.

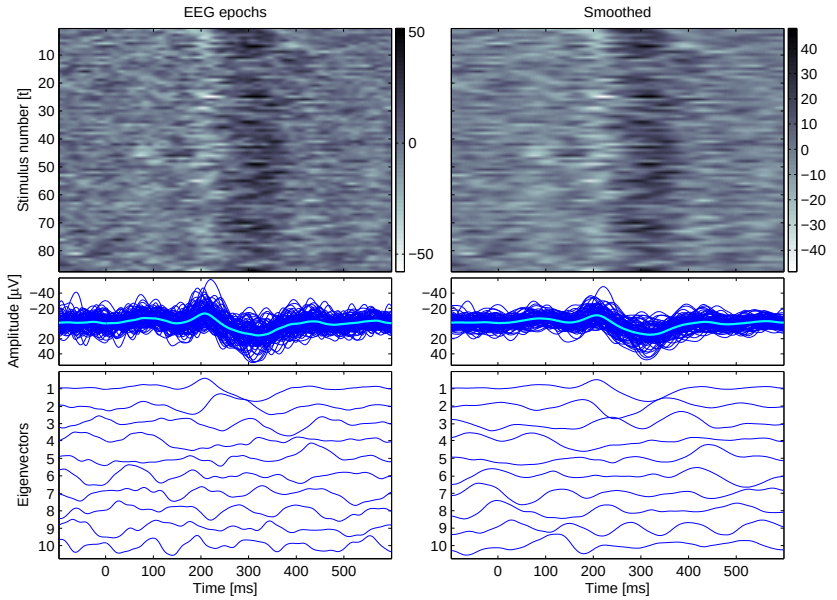

Figure 8.29: Deviant tone measurements and obtained eigenvectors.

8.7 Application to real EP data

In this section, the proposed single-trial estimation methods are applied to the real EP data used in the ICA example (section 7.3.2).

8.7.1 Deviant stimuli and P300 component

As in the simulated examples, 5 eigenvectors of the data correlation matrix of the smoothed measurements (after artifact removal by ICA) are used. Dominant eigenvectors and measurements are plotted in Fig. 8.29. Three sets of estimates were computed for different values of state noise variance parameter (σ^2 =0.05, 0.1, and 0.5). Estimates obtained with Kalman smoother and smoother optimized (in the interval 0-500ms) are presented in Fig. 8.30-8.32. By visual inspection of the estimates the value 0.05 over-smooths the data (Fig. 8.30). Though by adding time variability to the model better estimates are obtained. In Fig. 8.32 the estimates start to follow the noise. By considering the good quality of the eigenvectors (at least the first 3) and by observing that for small values of σ^2 the data still show trend like variations both in latency and amplitude, a value around 0.1 seems optimal for σ^2. For that value the optimized method seems also to give better results. For that value single-trial estimates are presented in Fig. 8.33.

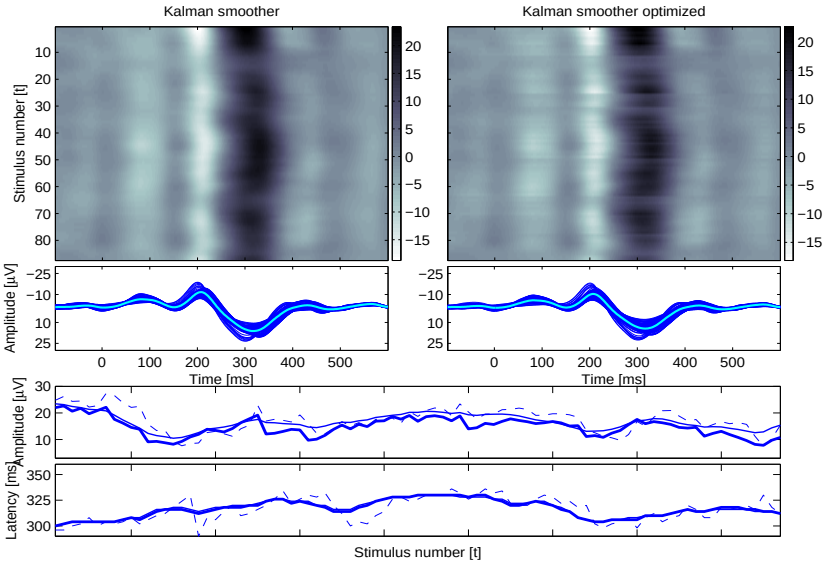

Figure 8.30: Estimates $\sigma^2 = 0.05$, KF (dashed), KS (thin), KS opt. (thick).

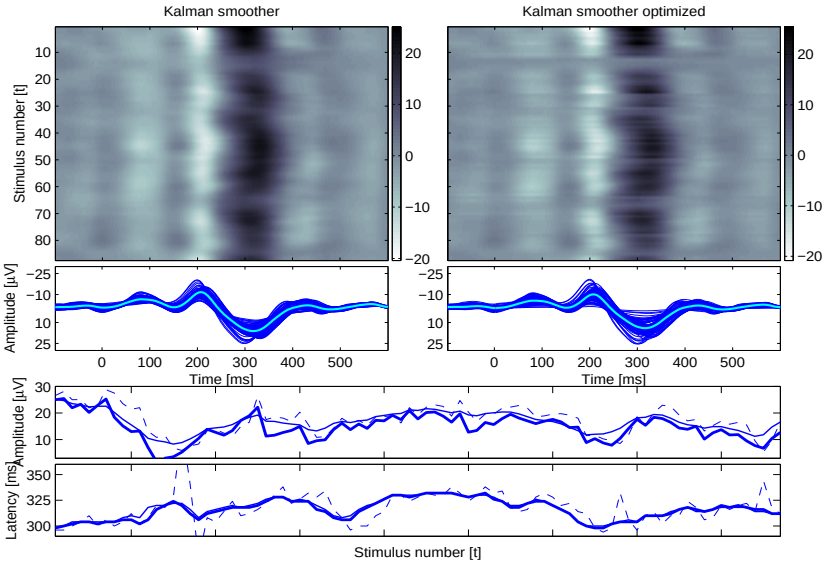

Figure 8.31: Estimates $\sigma^2 = 0.1$, KF (dashed), KS (thin), KS opt. (thick).

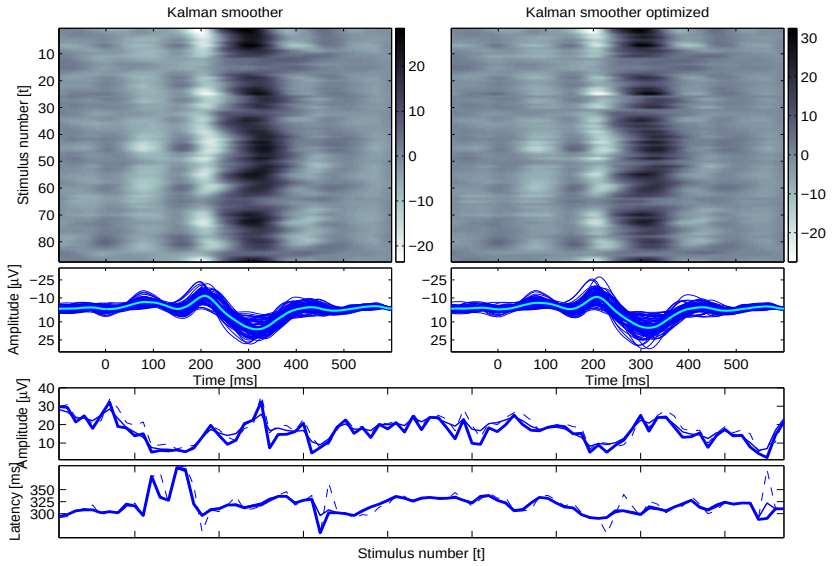

Figure 8.32: Estimates $\sigma^2 = 0.5$, KF (dashed), KS (thin), KS opt. (thick).

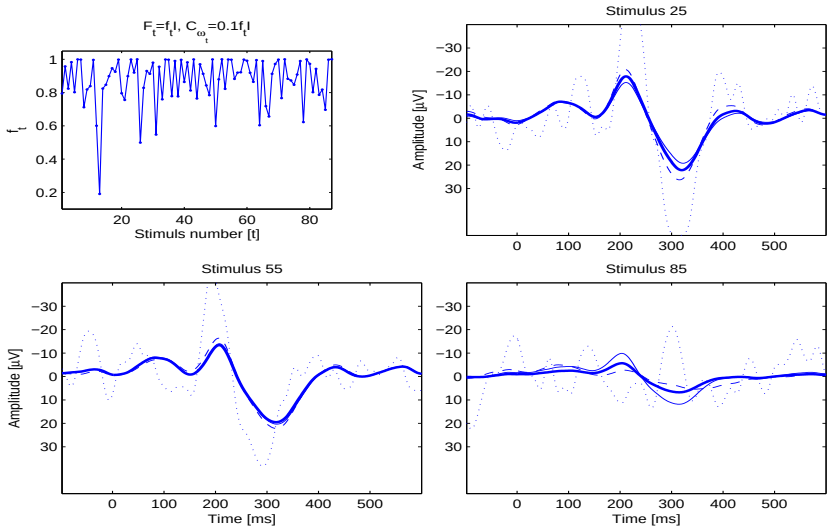

Figure 8.33: Deviant tone: estimated model F_t, C_{ω_t} and three single-trial estimates, noisy EPs (dotted), KF (dashed), KS (thin), KS opt. (thick).

8.7.2 Standard stimuli and N100/P200 complex

Estimates for the N100/P200 complex obtained from the standard tone were also computed. The EP measurements and eigenvectors are shown in Fig. 8.34. In the same way, 5 dominant eigenvectors of the data correlation matrix were used for estimation. By visual inspection the value $\sigma^2 = 0.01$ was selected. Estimates for the EPs are presented in Fig. 8.35, image plots (up), and estimates for the amplitude and latency of the P200 (bottom). Optimization was now considered in the interval (0-400ms).

In order to investigate the effect of artifacts in the accuracy of the estimates, dynamical estimation was applied to the measurements without artifact removal by ICA. EP measurements with artifacts are presented in Fig. 8.36. The blink artifacts are visible, especially at the end of the data epochs. The eigenvectors and the signal subspace is distorted and the most dominant eigenvectors model largely the blinks. Furthermore, estimates are presented in Fig. 8.37. Naturally, the estimates are of less quality. Since the blinks were more concentrated at the end of the epochs, and since there are randomly occurring, Kalman smoother has largely removed them providing, for example, comparable estimates for the amplitude of P200 (see Fig. 8.35 and Fig. 8.37). Though at the end of the measurements the latencies are distorted. In case of noisy eigenvectors, it is better to use other generic observation models, or to estimate the signal subspace by first rejecting some corrupted epochs. Single-trial estimates (without artifacts) are presented in Fig. 8.38, together with the estimated state-space model.

Furthermore, the blink artifacts are also visible in the averaged epochs. In the Appendix (Fig. A.10) there are presented the averaged epochs obtained for each channel separately before and after ICA. Clearly the blink contribution is largely visible and ICA and BSS methods, or other artifact removal methods, are important even for simple traditional analysis. In Fig. A.11 it is presented the revealed variability of the N100/P200 complex in terms of amplitudes and latencies of the two peaks. The estimates are based on the optimized smoother (channel CZ). Kalman smoother estimates were also computed for each channel separately (artifact corrected measurements). The observation model was 5 eigenvectors of the correlation matrix of each channel individually. Multichannel estimates, in form of scalp maps time locked at the latencies estimated from channel CZ for the P200 are presented in Fig. A.12. The scalp maps (row by row) correspond to steps of 10 stimuli. An overall decrease in amplitude is visible.

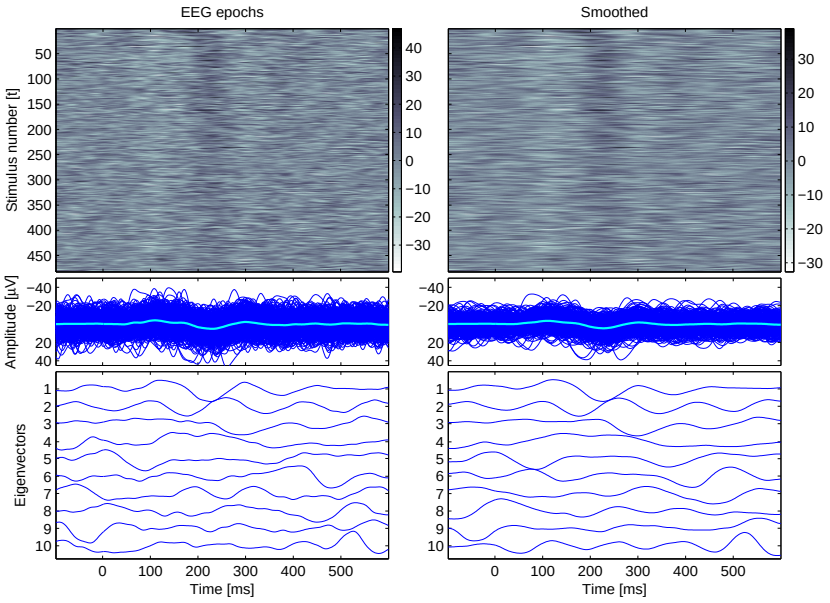

Figure 8.34: Standard tone measurements and obtained eigenvectors.

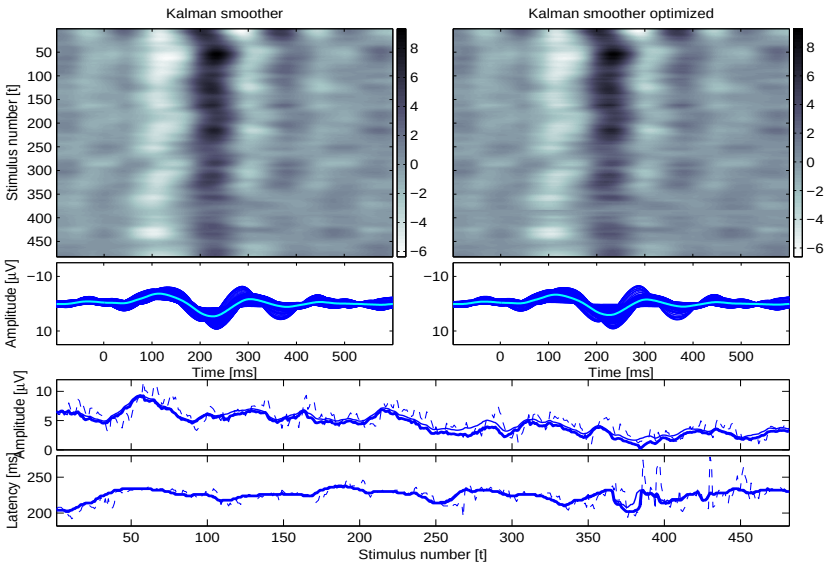

Figure 8.35: Estimates $\sigma^2 = 0.01$, KF (dashed), KS (thin), KS opt. (thick).

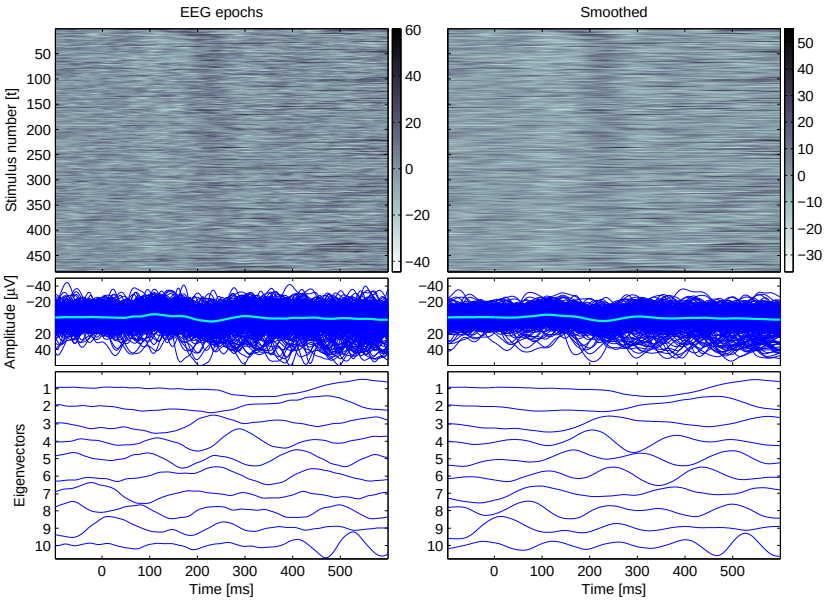

Figure 8.36: Blink corrupted measurements and eigenvectors.

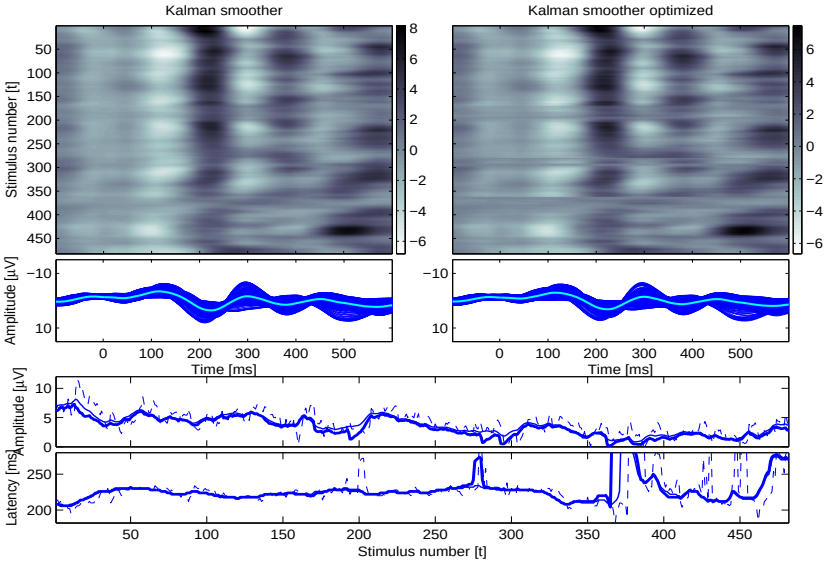

Figure 8.37: Estimates $\sigma^2 = 0.01$, KF (dashed), KS (thin), KS opt. (thick).

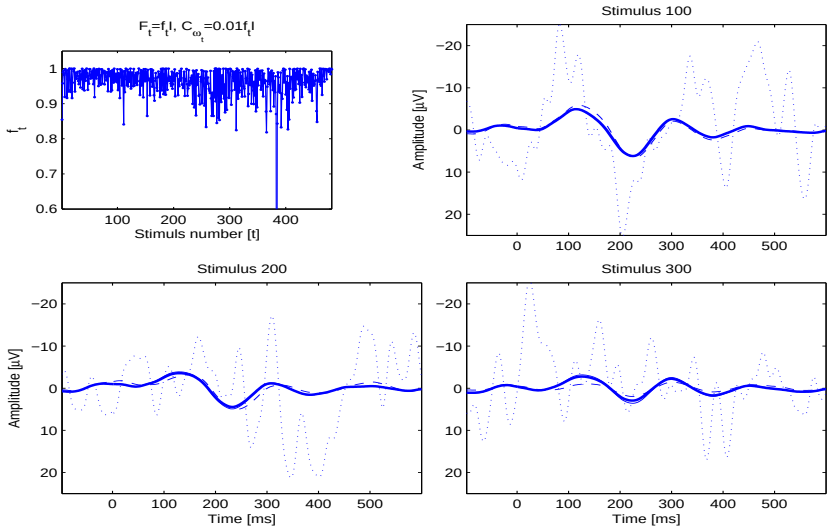

Figure 8.38: Standard tone: estimated model F_t, C_{ω_t} and three single-trial estimates, noisy EPs (dotted), KF (dashed), KS (thin), KS opt. (thick).

Chapter IX

Discussion and Conclusions

In this thesis, novel methods for EP denoising and enhancement were presented. The developed methods involve state-space modeling and identification techniques for dynamical estimation of EPs. Estimates for the EPs are obtained with Kalman filter and smoother algorithms. The performance of the methods relates to the quality of the EP signal subspace in low signal-to-noise ratio conditions, and especially to the assumption of hidden dynamic behavior from trial-to-trial. The estimates could, for example, be used to study different habituation effects, to detect changes in cognitive state, or to study cortical activity during anesthesia.

For the analysis of EPs, the clarification of different assumptions entering in the estimation procedure is critical. Different assumptions can be investigated or imposed within the framework of probability theory, based on the Bayesian formalism, or rather deterministically, based on regularization theory. This has been extensively discussed in this thesis. A mathematically elegant way to investigate dynamic variability in EPs is given by state-space modeling. The measured potentials are treated as vector valued stochastic processes, and the target is to recover slowly varying features buried into ongoing EEG and other interferences. Bayesian recursive mean square estimation methods can then be applied. In this thesis, the applicability, assumptions, and limitations of the approach were discussed and demonstrated. Furthermore, it was shown that the proposed methods are able to track dynamic variability from trial-to-trial with simulated and real EP data.

Kalman filter and smoother algorithms were systematically compared with computer simulations. The simulation studies provide also empirical ways for tuning relevant variance parameters. In this thesis, it was shown that the smoothing estimation procedure should be used when all the measurements are available and the filtering procedure is necessary only in cases when the estimation is required to be on-line. This is because the tracking capability is improved and the noise reduction is greater. This conclusion is neither restricted to the particular application (e.g. [193]), nor to the selected observation model [63]. Compromises between tracking capability and on-line estimation are given by fixed-lag smoothing algorithms (e.g. [108]).

The identification of an appropriate observation model for estimation was also

considered. Data-based observation models based on eigenvectors provide a good choice since they contain information about prominent characteristics of the signals; although even dominant eigenvectors can reflect artifacts and ongoing activity. The spatially dependent nature of multichannel EEG measurements can be exploited for identification and correction of different artifacts. One way to approach the problem is within the framework of blind source separation (e.g. [85, 39]). ICA for BSS of EEG was considered in order to demonstrate that the signal subspace, and therefore, the quality of the estimates can be affected by strong artifacts. However, the proposed methods are able to estimate EPs under very low signal-to-noise ratio conditions, as soon as slow changes from stimulus to stimulus exist. This was demonstrated with computer simulations and real EP measurements.

Some cases of generic observation models were also considered and demonstrated. However, the development of other models which can be tailored according to the application and estimation needs is still necessary. In parallel with the selection of the observation model, a method for enhancement of the state-evolution model for dynamical estimation of EPs was presented. The method imposes extra prior information for the parameters and can be tailored depending on the parametrization, for example, to provide smoothness. One such example was demonstrated. Based on the same formulation, a method for state-space identification was proposed that seems to perform well in difficult estimation problems. The method introducestime variability in the state-space model by estimating extra parameters from the data. Furthermore, it can better track sudden changes and prevents over-smoothing. Though, the prior selection of one parameter is still necessary for estimation.

In this thesis, the introduction of estimates for the observation noise variances was not considered. Estimation of extra non-stationary properties is a difficult task, but could improve tracking capabilities and estimation accuracy. More important, the fully Bayesian nature of the method is then going to be revealed by the provision of meaningful estimates for the parameter error covariances. Modeling of extra prior information obtained from the multichannel measurements or other relevant biosignals could also be investigated. For some approaches see [172, 173]. Finally, the introduction of extra control inputs in the estimation procedure may lead to further improvements.

Appendix I

Additional Figures

For the evaluation and categorization of ICs time and time-frequency plots, histograms, and activation maps were computed. From the strong blink artifact component estimates for the blink occurrence time were obtained, and for every IC epochs relative to blinks were sampled. Epochs were also sampled relative to the two stimulus types for every component, see also section 7.3.2. The Figs. A.10-A.12 are discussed in section 8.7.2.

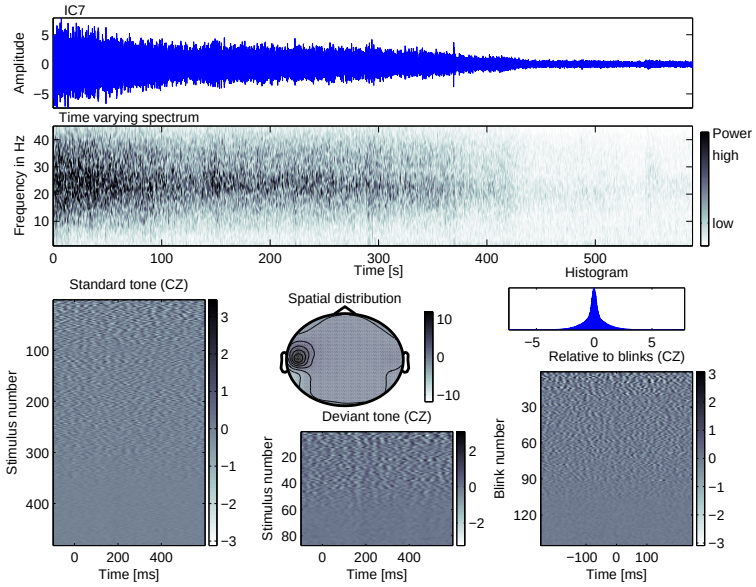

Figure A.1: Left temporal artifact.

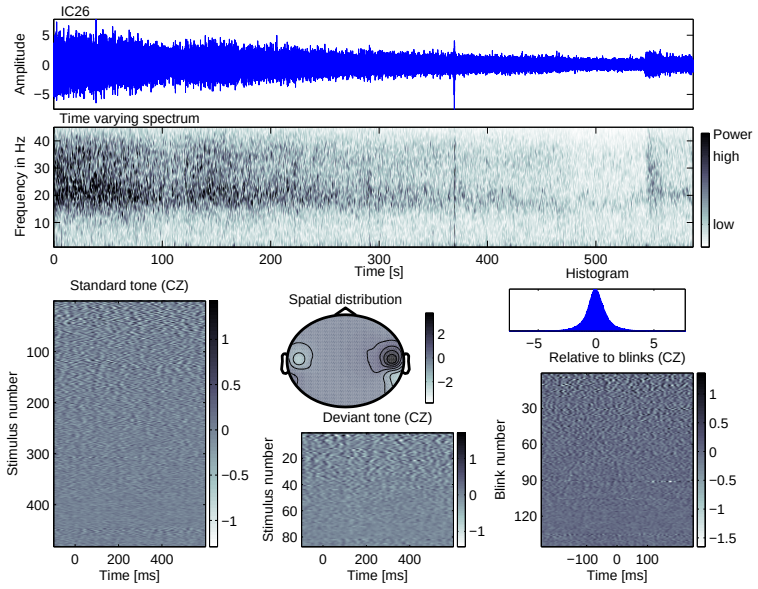

Figure A.2: Right temporal artifact.

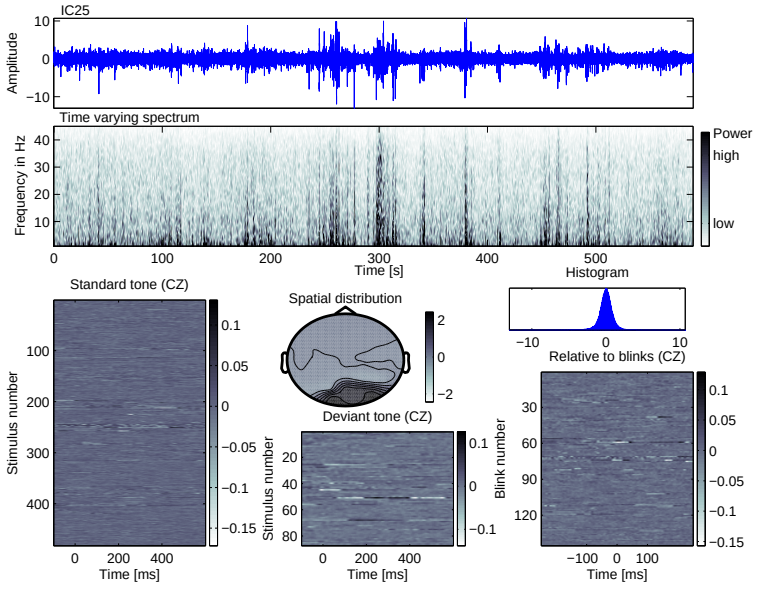

Figure A.3: Rear head muscle artifact.

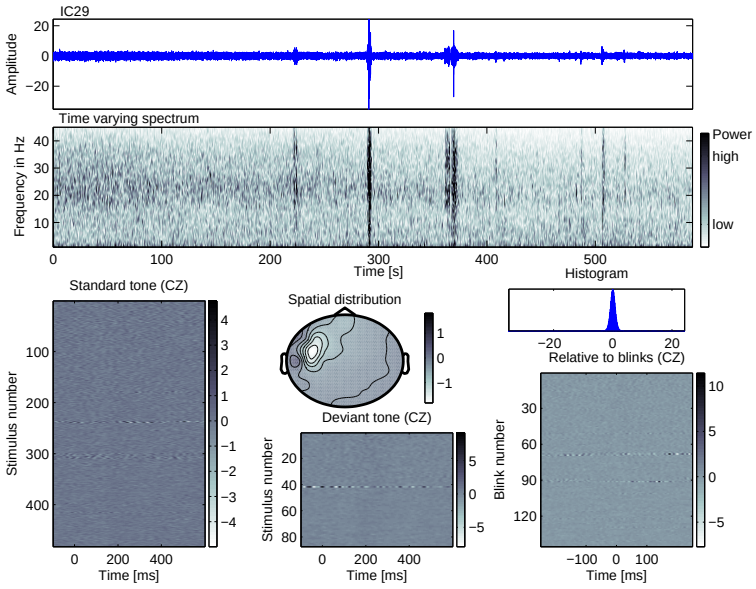

Figure A.4: Artifact.

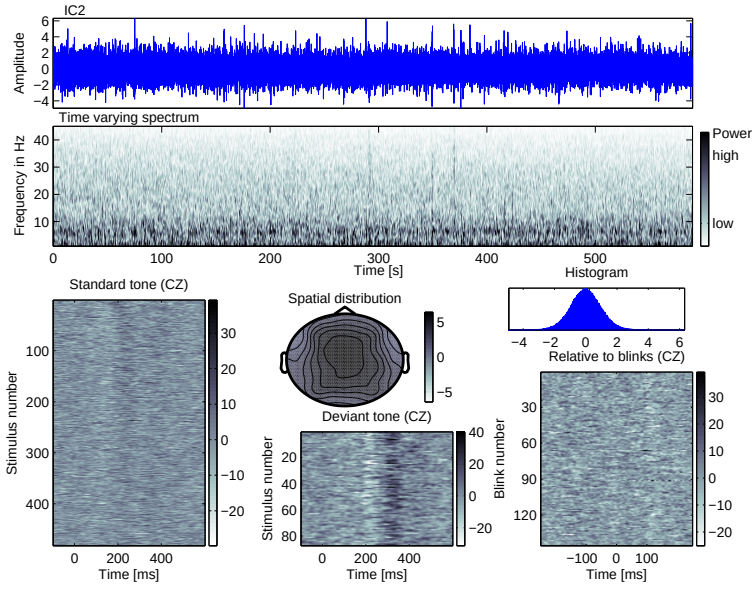

Figure A.5: Event related activity.

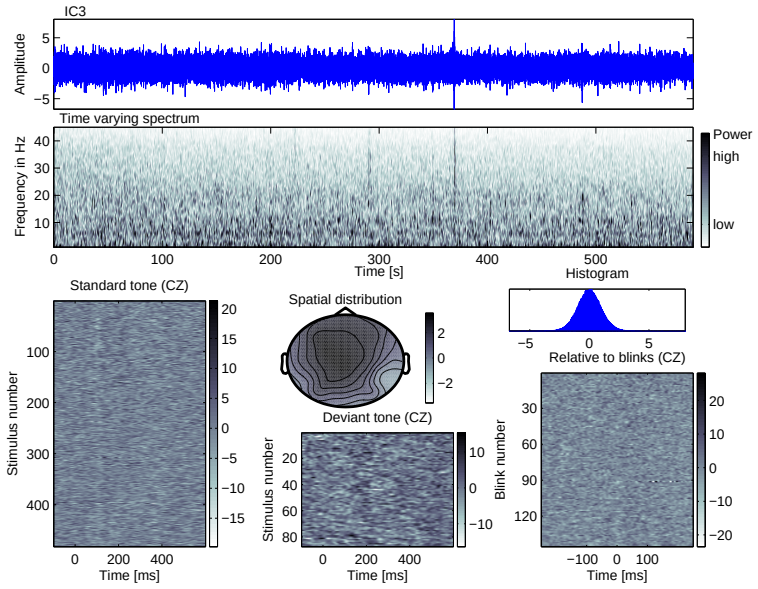

Figure A.6: Brain activity.

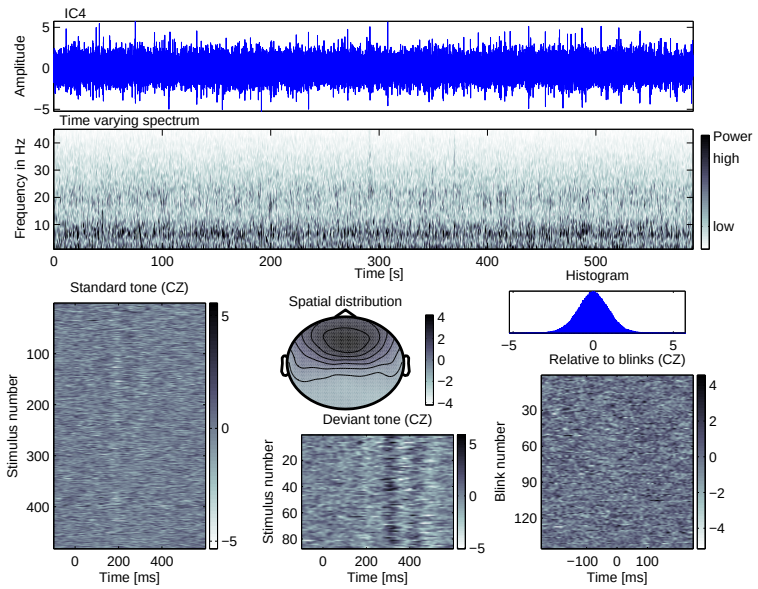

Figure A.7: Event related activity.

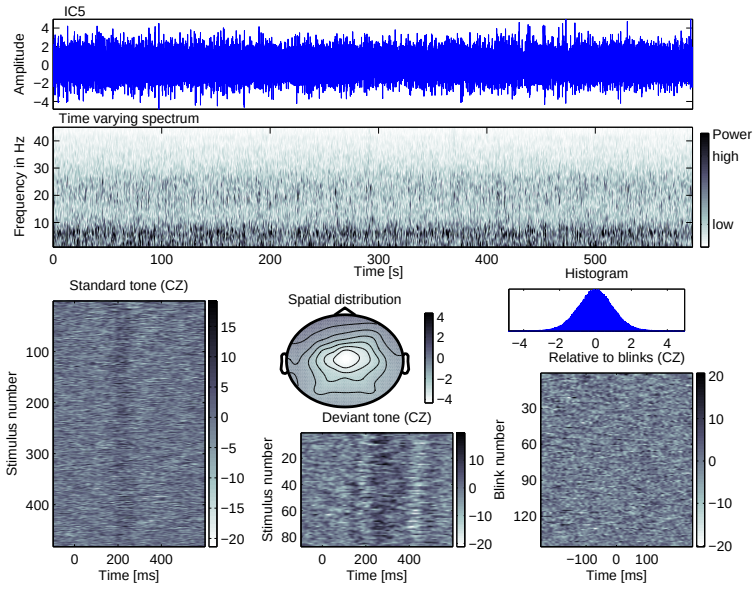

Figure A.8: Event related activity.

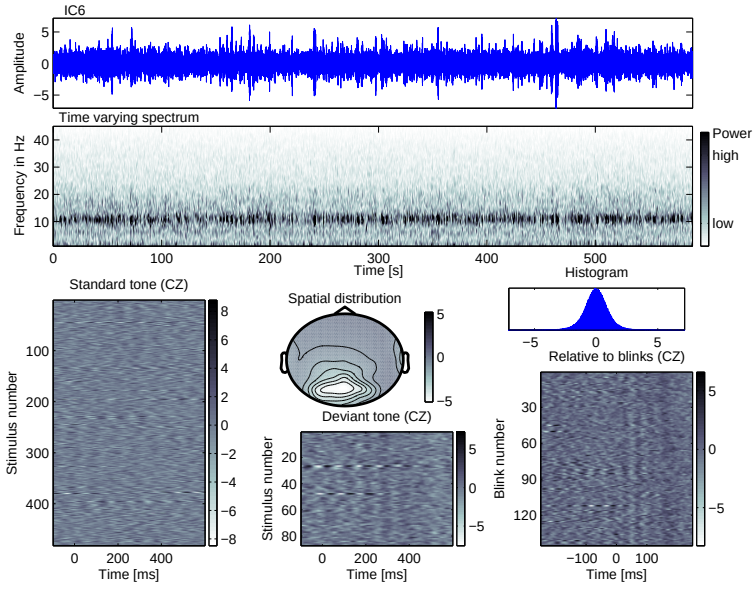

Figure A.9: Alpha activity component.

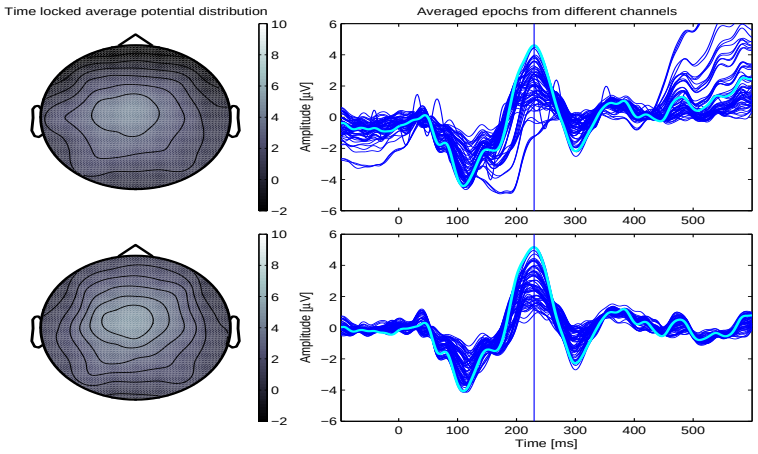

Figure A.10: Averaged epochs before and after ICA, P200 Peak, channel CZ (gray line).

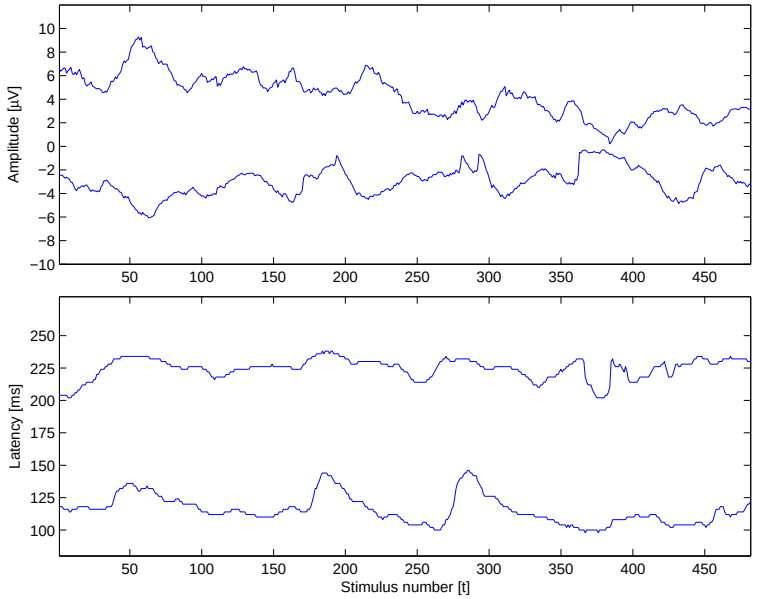

Figure A.11: Estimated stochastic variability of the N100 and P200 peaks in terms of time-varying amplitudes and latencies. The estimates are based on the state-space identification method for Kalman smoother.

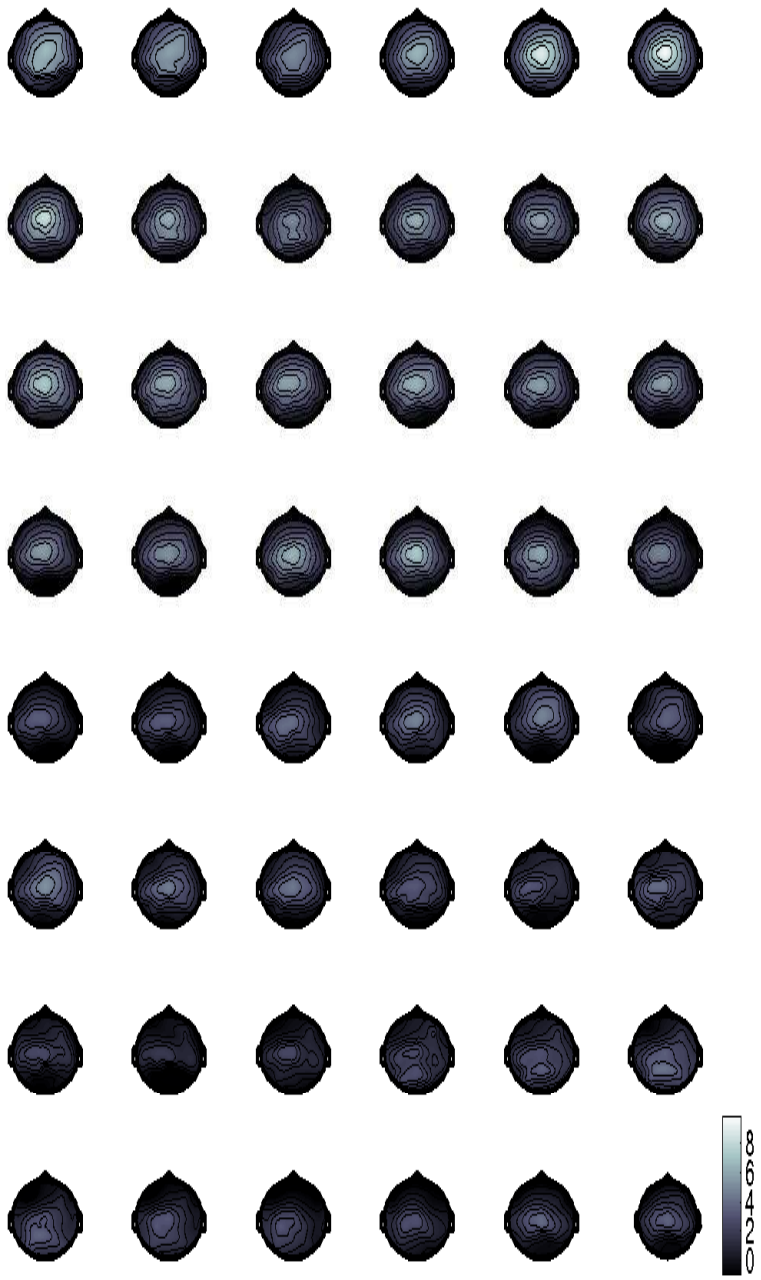

Figure A.12: Dynamic behavior of EPs (P200 Peak). Time-locked scalp maps (based on Kalman smoother estimates from each channel individually) at the latencies of the P200 peak (estimated from channel CZ). From left to right steps of 10 stimuli, continuing to next rows.

REFERENCES

[1] S. Amari. Natural gradiant works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[2] S. Amari. Superefficiency in blind source separation. *IEEE Transactions on Signal Processing*, 47(4):936–944, 1999.

[3] S. Amari. Nonholonomic orthogonal learning algorithms for blind source separation. *Neural Computation*, 12:1463–1484, 2000.

[4] S. Amari and J.-F. Cardoso. Blind source separation-semiparametric statistical approach. *IEEE Transactions on Signal Processing*, 45(11):2692–2700, 1997.

[5] S. Amari, T.-P. Chen, and A. Cichocki. Stability analysis of learning algorithms for blind source separation. *Neural Networks*, 10(8):1345–1351, 1997.

[6] S. Amari, A. Cichocki, and H.. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems (NIPS'95)*, volume 8, pages 757–763. MIT Press, 1996.

[7] B.D.O. Anderson and J.B. Moore. *Optimal Filtering*. Prentice Hall, 1979.

[8] J.L. Andreassi. *Psychophysiology Human Behavior and Physiological Response*. Lawrence Erlbaum Associates, Publishers, 4th edition, 2000.

[9] J. Anemueller, T.J. Sejnowski, and S. Makeig. Complex independent component analysis of frequency-domain electroencephalographic data. *Neural Networks*, 16(9):1311–1323, 2003.

[10] M. Askar ang H. Derin. A recursive algorithm for the Bayes solution of the smoothing problem. *IEEE Transactions on Automatic Control*, 26(2):558–561, 1981.

[11] R.B. Ash and C.A. Doleans-Dade. *Probability and Measure Theory*. Academic Press, 2nd edition, 2000.

[12] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[13] D. Baroudi, J.P. Kaipio, and E. Somersalo. Dynamical electric wire tomography: Time series approach. *Inverse Problems*, 14:799–813, 1998.

[14] A.K. Barros, R. Vigário, V. Jousmäki, and V. Ohnishi. Extraction of event-related signals from multichannel bioelectrical measurements. *IEEE Transactions on Biomedical Engineering*, 47(5):583–588, 2000.

[15] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[16] A. Belouchrani and M. Amin. Blind source separation based on time-frequency signal representation. *IEEE Transactions on Signal Processing*, 46(11):2888–2897, 1998.

[17] A. Belouchrani, J.-F. Cardoso K. Abed Meraim, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.
[18] D.P. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific MIT, 2003.
[19] D.P. Bertsekas, A. Nedic, and A.E. Ozdaglar. *Nonlinear Programming*. Athena Scientific MIT, 2nd edition, 1999.
[20] R. Bhatia. *Matrix Analysis, Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.
[21] Å. Björk. *Numerical Methods for Least Squares Problems*. SIAM Press, 1996.
[22] G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley, 1973.
[23] S.P. Boyd and L.I Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
[24] M.A.B. Brazier. Evoked responses recorded from the depths of the human brain. *Annals of the New York Academy of Sciences*, 112:33–59, 1964.
[25] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer, 1991.
[26] J.T. Cacioppo, L.G. Tassinary, and G.G. Berntson, editors. *Handbook of Psychophysiology*. Cambridge Univ. Press, 2nd edition, 2000.
[27] X.-R. Cao and R.W. Liu. General approach to blind source separation. *IEEE Transactions on Signal Processing*, 44(3):562–571, 1996.
[28] J.-F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114, 1997.
[29] J.-F. Cardoso. Blind signal separation: Statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.
[30] J.-F. Cardoso. Higher-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.
[31] J.-F. Cardoso. The three easy routes to independent component analysis; contrasts and geometry. In *Proc. ICA 2001, San Diego*, 2001.
[32] J.-F. Cardoso and B.H. Laheld. Blind beamforming for non gaussian signals. *IEE Proceedings-F, Radar and Signal Processing*, 140(6):362–370, 1993.
[33] J.-F. Cardoso and B.H. Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, 1996.
[34] S. Cerutti, V. Bersani, A. Carrara, and D. Liberati. Analysis of visual evoked potentials through Wiener filtering applied to a small number of sweeps. *Journal of Biomedical Engineering*, 9(1):3–12, 1987.
[35] K.-L. Chan, T.-W. Lee, and T.Sejnowski. Variational learning of clusters on undercomplete nonsymmetric independent components. *Journal of Machine Learning Research*, 3:99–114, 2003.
[36] S. Choi, A. Cichocki, and S. Amari. Flexible independent component analysis. *Journal of VLSI Signal Processing*, 26(1-2):25–38, 2000.
[37] S. Choi, A. Cichocki, H.M. Park, and S.-Y. Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing - Letters and Reviews*, 6(1):1–57, 2005.
[38] A. Cichocki. Blind signal processing methods for analyzing multichannel brain signals. *International Journal of Bioelectromagnetism*, 6(1), 2004.
[39] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing, Learning Algorithms and Applications*. Wiley, 2002.
[40] A. Cichocki, S.C. Douglas, and S. Amari. Robust techniques for independent component analysis with noisy data. *Neurocomputing*, 22:113–129, 1998.
[41] A. Cichocki, R.R. Gharieb, and T. Hoya. Efficient extraction of evoked potentials

by combination of Wiener filtering and subspace methods. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-2001, May 2001, Utah, USA*, volume 5, pages 3117–3120, 2001.

[42] A. Cichocki and L. Moszczynski. New learning algorithm for blind separation of sources. *Electronics Letters*, 28(21):1986–1987, 1992.

[43] A. Cichocki and R. Unbehaunen. Robust neural networks with on-line learning for blind identification and blind separation of signals. *IEEE Transactions on Circuits and Systems*, 43(11):894–906, 1996.

[44] A. Cichocki, R. Unbehaunen, and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30(17):1386–1387, 1994.

[45] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[46] P. Comon, C. Jutten, and J. Hérault. Blind separation of sources, part II: problems statment. *Signal Processing*, 24(1):11–20, 1991.

[47] R.J. Croft and R.J. Berry. Removal of ocular artifact form the EEG: a review. *Clinical Neurophysiology*, 30(1):5–19, 2000.

[48] S. Debener, S. Makeig, A. Delorme, and A.K. Engel. What is novel in the novelty oddball paradigm? functional significance of the novelty P3 event-related potential as revealed by independent component analysis. *Brain Research. Cognitive Brain Research*, 22(3):309–321, 2005.

[49] A. Delorme and S. Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

[50] A. Delorme, S. Makeig, M. Fabre-Thorpe M, and T.J. Sejnowski. From single-trial EEG to brain area dynamics. *Neurocomputing*, 44-46:1057–1064, 2002.

[51] J.E. Dennis and R.B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

[52] G. Deuschl and A.. Eisen, editors. *Recommendations for the Practice of Clinical Neurophysiology: Guidlines of the International Federation of Clinical Neurophysiology*. Elsevier, 1999.

[53] K.I. Diamantaras and S.Y. Kung. *Principal Component Neural Networks: Theory and Applications*. Wiley, 1996.

[54] C. Doncarli, L. Goering, and Guiheneuc. Adaptive smoothing of evoked potentials. *Signal Processing*, 28(1):63–76, 1992.

[55] A. Edelman, T. Arias, and S. Smith. The geomery of algorithms with orthogonality contraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1999.

[56] J. Ericsson and V. Koivunen. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

[57] R.M. Everson and S.J. Roberts. Independent component analsyis: a flexible nonlinearity and decorrelating manifold approach. *Neural Computation*, 11(8):1957–1983, 1999.

[58] R.M. Everson and S.J. Roberts. Blind source separation for non-stationary mixing. *Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, 26(8):15–24, 2000.

[59] W. Feller. *An Itroduction to Probability Theory and Its Applications*, volume I. Wiley, 3rd edition, 1968.

[60] W. Feller. *An Itroduction to Probability Theory and Its Applications*, volume II. Wiley, 2nd edition, 1971.

[61] D.C. Fraser and J.E. Potter. The optimum linear smoother as a combination of two optimum linear filters. *IEEE Transactions on Automatic Control*, 14(4):387–390, 1969.

[62] M. Gaeta and J.-L. Lacoume. Source separation without prior knowledge: the maximum likelihood solution. In *Proc. EUSIPCO'90*, pages 621–624, 1990.

[63] S.D. Georgiadis, P.O. Ranta-aho, M.P. Tarvainen, and P.A. Karjalainen. Recursive mean square estimators for single-trial event related potentials. In *Proc. Finnish Signal Processing Symposium - FINSIG'05*, Kuopio, Finland, 2005.

[64] S.D. Georgiadis, P.O. Ranta-aho, M.P. Tarvainen, and P.A. Karjalainen. Single-trial dynamical estimation of event related potentials: a Kalman filter based approach. *IEEE Transactions on Biomedical Engineering*, 52(8):1397–1406, 2005.

[65] L. El Ghaoui and H. Lebret. Robust solutions to least squares problems with uncertain data. In *Recent advances in total least squares techniques and errors-in-variables modeling (Leuven, 1996)*, pages 161–170. SIAM, Philadelphia, PA, 1997.

[66] M. Girolami. An alternative prespective on adaptive independent component analysis algorithms. *Neural Computation*, 10(8):2103–2114, 1998.

[67] M. Girolami, editor. *Advances in Independent Component Analysis*. Springer, 2000.

[68] G. Golub and V. Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems*, 19:R1–R26, 2003.

[69] G.H. Golub and C.F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, 1989.

[70] P.G. Gregoriev, A. Cichocki, and H. Bakardjian. Optimization techniques for independent component analysis with applications to EEG data analysis. In Pardalos et al., editor, *Quantitative Neuroscience Models, Algorithms, Diagnostics, and Therapeutic Applications*, pages 53–68. Kluwer Academic Publishers, 2004.

[71] F. Gustafsson. *Adaptive Filtering and Change Detection*. Wiley, 2000.

[72] M. Hämäläinen, R. Hari, R.J. Ilmoniemi, J. Knuutila, and O.V. Lounasmaa. Magnetoencephalography-theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews in Modern Physics*, 65(2):413–497, 1993.

[73] P.C. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, 1992.

[74] P.C. Hansen. *Regularization Tools: A Matlab Package for Analysis and Solution of Discrete Ill-Posed Problems*. UNI-C, Technical University of Denmark, 1992.

[75] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 2nd edition, 1991.

[76] J. Hérault and B. Ans. Circuits neuronaux a synapses modifiables: décodage de messages composites par apprentissage non supervisé. *C.-R de l' Acaémie des Sciences*, 299(III-13):525–528, 1984.

[77] Y.C. Ho and R.C.K. Lee. A bayesian approach to problems in stochastic estimation and control. *IEEE Transactions on Automatic Control*, 9:333–339, 1964.

[78] A. Holm, P.O. Ranta-aho, M. Sallinen, P.A. Karjalainen, and K. Müller. Relationship of P300 single trial responses with reaction time and preceding stimulus sequence. *International Journal of Psychophysiology*, 61(2):244–252, 2006.

[79] S. Hosseini, C. Jutten, and D.T. Pham. Markovian source separation. *IEEE Transactions on Signal Processing*, 51(12):3009–3019, 2003.

[80] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, volume 10. MIT Pess, 1998.

[81] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.
[82] A. Hyvärinen. The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters*, 10(1):1–5, 1999.
[83] A. Hyvärinen. Survey on independnet component analysis. *Neural Computing Surveys*, 2:94–128, 1999.
[84] A. Hyvärinen. A unifying model for blind separation of independent sources. *Signal Processing*, 85(7):1419–1427, 2005.
[85] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.
[86] A. Hyvärinen and R. Karthikesh. Imposing sparsity on the mixing matrix in independent component analysis. *Neurocomputing*, 49(1):151–162, 2002.
[87] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.
[88] A. Hyvärinen and Erkki Oja. Independnet component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.
[89] *ICA 1999 Proc. 1st. Int. Conf. on Independent Component Analysis and Blind Signal Separation (January 11-15, Aussois, France)*, 1999.
[90] *ICA 2000 Proc. 2nd. Int. Conf. on Independent Component Analysis and Blind Signal Separation (June 19-22, Helsinki, Finland)*, 2000.
[91] *ICA 2001 Proc. 3rd. Int. Conf. on Independent Component Analysis and Blind Signal Separation (December 9-12, San Diego, California, USA)*, 2001.
[92] *ICA 2003 Proc. 4th. Int. Conf. on Independent Component Analysis and Blind Signal Separation (April 1-4, Nara, Japan)*, 2003.
[93] *ICA 2004 Proc. 5th. Int. Conf. on Independent Component Analysis and Blind Signal Separation (September 22-24, Granada, Spain)*, 2004.
[94] *ICA 2006 Proc. 6th. Int. Conf. on Independent Component Analysis and Blind Signal Separation (March 5-8, Charleston, SC, USA)*, 2006.
[95] ICA algorithms in MATLAB online http://www.tsi.enst.fr/icacentral/.
[96] EEGLAB MATLAB package online http://www.sccn.ucsd.edu/eeglab/.
[97] FastICA MATLAB package online http://www.cis.hut.fi/projects/ica/fastica/.
[98] ICALAB MATLAB package online http://www.bsp.brain.riken.go.jp/icalab/.
[99] J. Intriligator and J. Polich. On the relationship between background EEG and the P300 event-related potential. *Biological Psychology*, 37(3):207–218, 1994.
[100] J. Iriarte, E. Urrestarazu, M. Valencia, M. Alegre, A. Malanda, C. Viteri, and J. Artieda. Independent component analysis as a tool to eliminate artifacts in EEG: A quantitative study. *Journal of Clinical Neurophysiology*, 20(4):249–257, 2003.
[101] C.J. James and C.W. Hesse. Independent component analysis for biomedical signals. *Physiological Measurements*, 26(1):15–39, 2005.
[102] B.H. Jansen, G. Agarwal, A. Hegde, and N.N. Boutros. Phase synchronization of the ongoing EEG and auditory EP generation. *Clinical Neurophysiology*, 114(1):79–85, 2003.
[103] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.
[104] T-P. Jung, S. Makeig, C. Humphries, T-W. Lee, M.J. McKeown, V. Iragui, and T.J. Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, 2000.
[105] T-P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T.J. Sejnowski. Removal of eye activity artifacts from visual event-related potentials in

normal and clinical subjects. *Clinical Neurophysiology*, 111(10):1745–1758, 2000.

[106] T-P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T.J. Sejnowski. Analysis and visualization of single-trial event-related potentials. *Human Brain Mapping*, 14(3):166–185, 2001.

[107] C. Jutten and J. Hérault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[108] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Applied Mathematical Sciences. Springer, 2005.

[109] J.P. Kaipio and E. Somersalo. Nonstationary inverese problems and state estimation. *Journal of Inverse and Ill-Posed Problems*, 7:273–282, 1999.

[110] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82:35–45, 1960.

[111] R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Transactions of the ASME, Journal of Basic Engineering*, 83:95–108, 1961.

[112] Malvin H. Kalos and Paula A. Whitlock. *Monte Carlo Methods*. John Wiley & Sons, Inc., 1986.

[113] J. Karhunen and J. Joutsensalo. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks*, 7(1):113–127, 1994.

[114] J. Karhunen and J. Joutsensalo. Generalizations of principal component analysis, optimization problems, and neural networks. *Neural Networks*, 8(4):549–562, 1995.

[115] J. Karhunen, P. Pajunen, and E. Oja. The nonlinear PCA criterion in blind source separation: Relations with other methods. *Neurocomputing*, 22(1-3):5–20, 1998.

[116] P.A. Karjalainen. *Regularization and Bayesian methods for evoked potential esimation*. PhD thesis, University of Kuopio, Department of Applied Physics, 1997.

[117] P.A. Karjalainen, J.P. Kaipio, A.S. Koistinen, and T. Kärki. Recursive Bayesian estimation of single trial evoked potentials. In *Proc. 18th Annual Conf. IEEE-EMBS*, Amsterdam, 1996.

[118] P.A. Karjalainen, J.P. Kaipio, A.S. Koistinen, and M. Vauhkonen. Subspace regularization method for the single trial estimation of evoked potentials. *IEEE Transactions on Biomedical Engineering*, 46(7):849–860, 1999.

[119] J. Karvanen, J. Eriksson, and V. Koivunen. Adaptive score functions for maximum likelihood ICA. *The Journal of VLSI Signal Processing Systems*, 32(1-2):83–92, 2002.

[120] J. Karvanen and V. Koivunen. Blind separation methods based on Pearson system and its extensions. *Signal Processing*, 82(4):663–673, 2002.

[121] J. Karvanen and V. Koivunen. Independent component analysis via optimum combining of kurtosis and skewness-based criteria. *Journal of the Franklin Institute*, 431(5):401–418, 2004.

[122] G. Kitagawa and W. Gersch. *Smoothness Priors Analysis of Time Series*. Springer, 1996.

[123] K.H. Knuth. Difficulties applying recent blind source separation techniques to EEG and MEG. In *Proc. Maximum Entropy and Bayesian Methods 1997, Boise, Idaho, USA*, pages 209–222. Dordrecht, Kluwer, 1998.

[124] K.H. Knuth. A Bayesian approach to source separation. In *Proc. First Int. Workshop on ICA and Signal Separation, ICA'99, Aussois,France*, pages 283–288, 1999.

[125] K.H. Knuth. Informed source separation: A Bayesian tutorial. In *Proc. of the 13th European Signal Processing Conference (EUSIPCO 2005), Antalya, Turkey*, 2005.

[126] K.H. Knuth, A. Shah, W. Truccolo, M. Ding, S.L. Bressler, and C.E. Schroeder. Differentially variable component analysis (dVCA): Identifying multiple evoked

components using trial-to-trial variability. *Journal of Neurophysiology*, 95(5):3257–3276, 2006.

[127] V. Kolehmainen, S. Prince, S.R. Arridge, and J.P. Kaipio. State-estimation approach to the nonstationary optical tomography problem. *Journal of the Optical Society of America A*, 20(5):876–889, 2003.

[128] S. Kullback and R.A Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[129] T.-W. Lee. *Independent Component Analysis: Theory and applications*. Kluwer, 1998.

[130] T.-W. Lee, M. Girolami, A.J. Bell, and T. Seijnowski. A unifying information-theoretic framework for independent component analysis. *Computers and Mathematics with applications*, 39(11):1–21, 2000.

[131] T.-W. Lee, M. Girolami, and T.J. Sejnowski. Independent component analysis using extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):417–441, 1999.

[132] K. Levenberg. A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168, 1944.

[133] D.J.C MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[134] S. Makeig, A.J. Bell, T-P. Jung, and T.J. Sejnowski. Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 145–151. MIT Press, 1996.

[135] S. Makeig, S. Debener, and A. Delorme. Mining event-related brain dynamics. *Trends in Cognitive Science*, 8(5):204–210, 2004.

[136] S. Makeig, A. Delorme, M. Westerfield, T-P. Jung T-P, J. Townsend, E. Courchesne, and T.J. Sejnowski. Electroencephalographic brain dynamics following manually responded visual targets. *PLOS Biology*, 2(6):742–762, 2004.

[137] S. Makeig, M. Westerfield, T-P. Jung, J. Covington, J. Townsend, T. J. Sejnowski, and E. Courchesne. Functionally independent components of the late positive event-related potential during visual spatial attention. *Journal of Neuroscience*, 19(7):2665–2680, 1999.

[138] S. Makeig, M. Westerfield, T-P. Jung T-P, S. Enghoff, J. Townsend, E. Courchesne, and T.J. Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295:690–694, 2002.

[139] V. Mäkinen, H. Tiitinen, and P. May. Auditory even-related responses are generated independently of ongoing brain activity. *NeuroImage*, 24(4):961–968, 2005.

[140] J. Malmivuo and R. Plonsey. *Bioelectromagnetism*. Oxford university press, New York, 1995.

[141] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM*, 11:431–441, 1963.

[142] D. W. Marquardt. A critique of some ridge regression methods: comment. *Journal of the American statistical association*, 75:87–91, 1980.

[143] K. Matsuoka and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419, 1995.

[144] D.Q. Mayne. A solution to the smoothing problem for linear dynamical systems. *Automatica*, 4:73–92, 1966.

[145] J.L. Melsa and D.L. Cohn. *Decision and Estimation Theory*. McGraw-Hill, 1978.

[146] A. Mohammad-Djafari. A bayesian approach to source separation. In *Proc. Max-*

imum Entropy and Bayesian Methods, Boise, USA, pages 221–244, 1999.

[147] L. Molgedey and G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23):3634–3637, 1994.

[148] R. Müller, R Vigário, F Meinecke, and A Ziehe. Blind source separation techniques for decomposing evoked brain signals. *International Journal of Bifurcation and Chaos*, 14(2):773–791, 2004.

[149] A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40(3):636–666, 1998.

[150] E. Niedermeyer. Alpha rhythms as physiological and abnormal phenomena. *International Journal of Psychophysiology*, 26(1-3):31–49, 1997.

[151] E. Niedermeyer and F. Lopes da Silva, editors. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Williams and Wilkins, 4th edition, 1999.

[152] D. Obradovic and G. Deco. Information maximization and independent component analysis: Is there a difference? *Neural Computation*, 10(8):2085–2101, 1998.

[153] E. Oja. *Subspace Methods in Pattern Recognition*. Wiley, 1983.

[154] E. Oja. The nonlinear PCA learning rule in independent component analysis. *Neurocomputing*, 17(1):25–46, 1997.

[155] E. Oja. From neural learning to independent components. *Neurocomputing*, 22(1):187–199, 1998.

[156] J. Ollikainen, M. Vauhkonen, P. Karjalainen, and J. Kaipio. Effects of local skull inhomogeneities on eeg source estimation. *Medical Engineering and Physics*, 21(3):143–154, 1999.

[157] J. Onton, A. Delorme, and S. Makeig. Frontal midline EEG dynamics during working memory. *Neuroimage*, 27(2):341–356, 2005.

[158] J. Onton and S. Makeig. Information-based modeling of event-related brain dynamics. *Progress in Brain Research*, 159:99–120, 2006.

[159] P. Pajunen and J. Karhunen. Least-squares methods for blind source separation methods based on nonlinear PCA. *International Journal of Neural Systems*, 8(5-6):601–612, 1997.

[160] A. Papoulis. *Signal Analysis*. McGraw-Hill, 1984.

[161] A. Papoulis and S.U. Pillai. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 4th edition, 2002.

[162] L. Parra and P. Sajda. Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Reasearch*, 4:1261–1269, 2003.

[163] B.A. Pearlmutter and L.C. Para. Maximum likelihood blind source separation: A context-sesitive generalization of ICA. In *Advances in Neural Information Processing Systems*, volume 9, pages 613–619. MIT Press, 1997.

[164] D.-T. Pham. Blind separation of instantaneus mixture of sources via an independent component analysis. *IEEE Transactions on Signal Processing*, 44(11):2768–2779, 1996.

[165] D.-T. Pham and J.-R. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Transactions on Signal Processing*, 49(9):1837–1848, 2001.

[166] D.-T. Pham and P. Garrat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Transactions on Signal Processing*, 45(7):1712–1725, 1997.

[167] D.-T. Pham, P. Garrat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO'90*, pages

771–774, 1992.

[168] T.W. Picton, editor. *Human Event Related Potential, EEG Handbook, revised series*, volume 3. Elsevier, 1988.

[169] M.B. Priestley. *Spectral Analysis and Time Series*. Academic Press, 1981.

[170] W. Qiu, C. Chang, W. Lie, P.W.F. Poon, F.K. Lam, R.P. Hamernik, G. Wei, and F.H.Y. Chan. Real-time data-reusing adaptive learning of a radial basis function network for tracking evoked potentials. *IEEE Transanctions on Biomedical Engineering*, 53(2):226–237, 2006.

[171] R. Quian Quiroga and H. Garcia. Single-trial evoked potentials with wavelet denoising. *Clinical Neurophysiology*, 114:376–390, 2003.

[172] P.O. Ranta-aho, A.S. Koistinen, J.O. Ollikainen, J.P. Kaipio, J. Partanen, and P.A. Karjalainen. Single-trial estimation of multichannel evoked-potential measurements. *IEEE Transactions on Biomedical Engineering*, 50(2):189–196, February 2003.

[173] P.O. Ranta-aho, M.P. Tarvainen MP, M. Valkonen-Korhonen, S.D. Georgiadis, J. Lehtonen, J-P. Niskanen, and P.A. Karjalainen PA. On correlation between single-trial ERP and GSR responses: a principal component regression approach. In *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, August 2006.

[174] H.E. Rauch, F. Tung, and C.T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3:1445–1450, 1965.

[175] J. Raz, B. Turetsky, and G. Fein. Selecting the smoothing parameter for estimation of slowly changing evoked potential signals. *Biometrics*, 45(3):745–762, 1989.

[176] S. J. Roberts. Independent component analysis: Source assessment and separation, a bayesian approach. *IEE Proceedings Vison, Image and Signal Processing*, 145(3):149–154, 1998.

[177] S. J. Roberts, E Roussos, and R. Choudrey. Hierarchy, priors and wavelets: Structure and signal modelling using ICA. *Signal Processing*, 84(2):283–297, 2004.

[178] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, reprint edition, 1996.

[179] D. B. Rowe. A Bayesian approach to blind source separation. *Journal of Interdisciplinary Mathematics*, 5(1):49–76, 2002.

[180] W. Rubin. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976.

[181] D.S. Ruchkin and E.M. Glaser. Simple digital filters for examining CNV and P300 on a single trial basis. In D.A. Otto, editor, *Multidisciplinary perpectives on event-related brain potential research*, pages 579–581. U.S. Governement Printing Office, Washington DC, 1978.

[182] W. Sato, T. Kochiyama, S. Yoshikawa, and M. Matsumura. Emotional expression boosts early visual processing of the face: ERP recording and its decomposition by independent component analysis. *Neuroreport*, 12(4):709–714, 2001.

[183] A. Sayed and T-Kailath. State-space approach to adaptive RLS filtering. *IEEE Signal Processing Magagine*, 11(3):18–60, 1994.

[184] H.W. Sorenson. *Parameter Estimation, Principles and Problems*, volume 9 of *Control and Systems Theory*. Marcel Dekker Inc., New York, 1980.

[185] H.W. Sorenson, editor. *Kalman Filtering: Theory and Applications*. IEEE Press, 1985.

[186] E. Sorouchyari. Blind separation of sources, part III: Stability analysis. *Signal Processing*, 24:21–29, 1991.

[187] G. Sparacino, S. Milani, E. Arslan, and C. Cobelli. A Bayesian approach to

estimate evoked potentials. *Computer Methods and Programs in Biomedicine*, 68(3):233–248, June 2002.

[188] A.C. Tang, B.A. Pearlmutter, N.A. Malaszenko, D.B. Phung, and B.C. Reeb. Independent components of magnetoencephalography: localization. *Neural Computation*, 14(8):1827–1858, 2002.

[189] A.C. Tang, M.T. Sutherland, and C.J. McKinney. Validation of SOBI components from high-density EEG. *Neuroimage*, 25(2):539–553, 2005.

[190] M.P. Tarvainen. *Estimation Methods for Nonstationary Biosignals*. PhD thesis, University of Kuopio, Department of Applied Physics, 2004. Available at http://it.uku.fi/biosignal.

[191] M.P. Tarvainen, S.D. Georgiadis, and P.A. Karjalainen. Time-varying analysis of heart rate variability with Kalman smoother algorithm. In *Proceedings of 27th Annual Int Conference of IEEE EMBS*, Shanghai, September 1-4 2005.

[192] M.P. Tarvainen, S.D. Georgiadis, P.O. Ranta-aho, and P.A. Karjalainen. Time-varying analysis of heart rate variability signals with Kalman smoother algorithm. *Physiological Measurements*, 27(3):225–239, 2006.

[193] M.P. Tarvainen, J.K. Hiltunen, P.O. Ranta-aho, and P.A. Karjalainen. Estimation of nonstationary EEG with Kalman smoother approach: an application to event-related synchronization (ERS). *IEEE Transaction on Biomedical Engineering*, 51(3):516–524, March 2004.

[194] N.V. Thakor, C.A. Vaz, R.W. McPherson, and D. F. Hanley. Adaptive Fourier series modeling of time-varying evoked potentials: Study of human somatosensory evoked response to etomidate anesthetic. *Electroencephalography and Clinical Neurophysiology*, 80(2):108–118, 1991.

[195] F.J. Theis. A new concept for separability problems in blind source separation. *Neural Computation*, 16:1827–1850, 2004.

[196] F.J. Theis. Uniqueness of complex and multidimensional independent component analysis. *Signal Processing*, 84(5):951–956, 2004.

[197] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. Winston and Sons, 1977.

[198] L. Tong, R.-W. Liu, V.C. Soon, and Y.-F. Huang. Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38(5):499–509, 1991.

[199] J. Townsend, M. Westerfield, E. Leaver, S. Makeig, T-P. Jung, K. Pierce, and E. Courchesne. Event-related brain response abnormalities in autism: evidence for impaired cerebello-frontal spatial attention networks. *Beain Research. Cognitive Brain Research*, 11(1):127–145, 2001.

[200] W.A. Truccolo, D. Mingzhou, K.H. Knuth, R. Nakamura, and S.L. Bressler. Trial-to-trial variability of cortical evoked responses: implications for the analysis of functional connectivity. *Clinical Neurophysiology*, 113(2):206–226, 2002.

[201] A.C. Tsai, M. Liou, T.-P. Jung, J. Onton, P.E. Cheng, C.-C. Huang, J.R. Duann, and S. Makeig. Mapping single-trial EEG records on the cortical surface through a spatiotemporal modality. *NeuroImage*, 32(1):195–207, 2006.

[202] B.I. Turetsky, J. Raz, and G. Fein. Estimation of trial-to-trial variation in evoked potential signals by smoothing across trials. *Psychophysiology*, 26(6):700–712, 1989.

[203] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

[204] R. Vigário, R.V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent com-

ponent analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing Systems*, volume 10, pages 229–235. MIT Press, 1997.

[205] R. Vigário and E. Oja. Independence: a new criterion for the analysis of the electromagnetic fields in the global brain? *Neural Networks*, 13(8-9):891–907, 2000.

[206] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, 47(5):589–593, 2000.

[207] R. N. Vigário. Extraction of ocular artifacts from EEG using independent component analysis. *Electroencephalography and Clinical Neurophysiology*, 103(3):395–404, 1997.

[208] G.L. Wallstrom, R.E. Kass, A. Miller, J.F. Cohn, and N.A. Fox. Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component based methods. *International Journal of Psychophysiology*, 53(2):105–119, 2004.

[209] J.C. Webster, editor. *Medical Instrumentation, application and design*. John Wiley and Sons, second edition, 1995.

[210] J.J. Westerkamp and J.I. Aunon. Optimum multielectrode a posteriori estimates of single-response evoked potentials. *IEEE Transactions on Biomedical Engineering*, 34:13–22, 1987.

[211] C.H. Wolters, A. Anwander, X. Tricoche, D. Weinstein, M.A. Koch, and R.S. MacLeod. Influence of tissue conductivity anisotropy on EEG/MEG field and return current computation in a realistic head model: A simulation and visualization study using high-resolution finite element modeling. *NeuroImage*, 30(3):813–826, 2006.

[212] A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Transactions on Signal Processing*, 50(7):1545–1553, 2002.

[213] H. Yppärilä, I. Korhonen, M. Tarvainen, T. Musialowicz, S. Jacob, and J. Partanen. Discrimination of sedation levels based on event-related potentials and electroencephalogram. *Journal of Clinical Monitoring and Computing*, 18(3):163–170, 2004.

[214] K. Yu and C.D. McGillem. Optimum filters for estimating evoked potential waveforms. *IEEE Transactions on Biomedical Engineering*, 30(11):730–737, 1983.

[215] L. Zhang, A. Cichocki, and S. Amari. Self-adaptive source separation based on activation functions adaptation. *IEEE Transactions on Neural Networks*, 15(2):233–244, 2004.

[216] Y. Zhang, A. Ghodrati, and D.H. Brooks. An analytical comparison of three spatio-temporal regularization methods for dynamical linear inverse problems in a common statistical framework. *Inverse Problems*, 21:357–382, 2005.

[217] A. Ziehe and K.R. Müller. TDSEP – an efficient algorithm for blind separation using time structure. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98), Skövde, Sweden*, pages 675–680, 1998.

Kuopio University Publications C. Natural and Environmental Sciences

C 196. Heijari, Juha. Seed origin, forest fertilization and chemical elicitor influencing wood characteristics and biotic resistance of Scots pine.
2006. 39 p. Acad. Diss.

C 197. Hakulinen, Mikko. Prediction of density, structure and mechanical properties of trabecular bone using ultrasound and X-ray techniques.
2006. 84 p. Acad. Diss.

C 198. Al Natsheh, Anas. Quantum Mechanics Study of Molecular Clusters Composed of Atmospheric Nucleation Precursors.
2006. 55 p. Acad. Diss.

C 199. Tarvainen, Tanja. Computational Methods for Light Transport in Optical Tomography.
2006. 123 p. Acad. Diss.

C 200. Heikkinen, Päivi. Studies on Cancer-related Effects of Radiofrequency Electromagnetic Fields. 2006. 165 p. Acad. Diss.

C 201. Laatikainen, Tarja. Pesticide induced responses in ectomycorrhizal fungi and symbiont Scots pine seedlings.
2006. 180 p. Acad. Diss.

C 202. Tiitta, Markku. Non-destructive methods for characterisation of wood material.
2006. 70 p. Acad. Diss.

C 203. Lehesranta, Satu. Proteomics in the Detection of Unintended Effects in Genetically Modified Crop Plants.
2006. 71 p. Acad. Diss.

C 204. Boman, Eeva. Radiotherapy forward and inverse problem applying Boltzmann transport equation.
2007. 138 p. Acad. Diss.

C 205. Saarakkala, Simo. Pre-Clinical Ultrasound Diagnostics of Articular Cartilage and Subchondral Bone.
2007. 96 p. Acad. Diss.

C 206. Korhonen, Samuli-Petrus. FLUFF-BALL, a Fuzzy Superposition and QSAR Technique - Towards an Automated Computational Detection of Biologically Active Compounds Using Multivariate Methods.
2007. 154 p. Acad. Diss.

C 207. Matilainen, Merja. Identification and characterization of target genes of the nuclear reseptors VDR and PPARs: implementing in silico methods into the analysis of nuclear receptor regulomes.
2007. 112 p. Acad. Diss.

C 208. Anttonen, Mikko J. Evaluation of Means to Increase the Content of Bioactive Phenolic Compounds in Soft Fruits.
2007. 93 p. Acad. Diss.

C 209. Pirkanniemi, Kari. Complexing agents: a study of short term toxicity, catalytic oxidative degradation and concentrations in industrial waste waters.
2007. 83 p. Acad. Diss.

C 210. Leppänen, Teemu. Effect of fiber orientation on cockling of paper.
2007. 96 p. Acad. Diss.