Linköping studies in science and technology. Dissertations.
No. 1710

Analytical Approximations for Bayesian Inference

Tohid Ardeshiri

LINKÖPING UNIVERSITY

Linköping studies in science and technology. Dissertations.
No. 1710

Analytical Approximations for Bayesian Inference

Tohid Ardeshiri

tohid@isy.liu.se
www.control.isy.liu.se
Division of Automatic Control
Department of Electrical Engineering
Linköping University
SE–581 83 Linköping
Sweden

ISBN : 978-91-7685-930-8 ISSN 0345-7524

Printed by LiU-Tryck, Linköping, Sweden 2015

To Adrian

Abstract

Bayesian inference is a statistical inference technique in which Bayes' theorem is used to update the probability distribution of a random variable using observations. Except for few simple cases, expression of such probability distributions using compact analytical expressions is infeasible. Approximation methods are required to express the *a priori* knowledge about a random variable in form of prior distributions. Further approximations are needed to compute posterior distributions of the random variables using the observations. When the computational complexity of representation of such posteriors increases over time as in mixture models, approximations are required to reduce the complexity of such representations.

This thesis further extends existing approximation methods for Bayesian inference, and generalizes the existing approximation methods in three aspects namely; prior selection, posterior evaluation given the observations and maintenance of computation complexity.

Particularly, the maximum entropy properties of the first-order stable spline kernel for identification of linear time-invariant stable and causal systems are shown. Analytical approximations are used to express the prior knowledge about the properties of the impulse response of a linear time-invariant stable and causal system.

Variational Bayes (VB) method is used to compute an approximate posterior in two inference problems. In the first problem, an approximate posterior for the state smoothing problem for linear state-space models with unknown and time-varying noise covariances is proposed. In the second problem, the VB method is used for approximate inference in state-space models with skewed measurement noise.

Moreover, a novel approximation method for Bayesian inference is proposed. The proposed Bayesian inference technique is based on Taylor series approximation of the logarithm of the likelihood function. The proposed approximation is devised for the case where the prior distribution belongs to the exponential family of distributions.

Finally, two contributions are dedicated to the mixture reduction (MR) problem. The first contribution, generalize the existing MR algorithms for Gaussian mixtures to the exponential family of distributions and compares them in an extended target tracking scenario. The second contribution, proposes a new Gaussian mixture reduction algorithm which minimizes the reverse Kullback-Leibler divergence and has specific peak preserving properties.

Populärvetenskaplig sammanfattning

Bayes sats är ett grundläggande verktyg inom statistik som kan användas för att förfina förkunskapen om en variabel med hjälp av observationer. Förkunskapen kallas prior och beskrivs matematiskt som en sannolikhetsfunktion för den okända variabeln, och observationen beskrivs av en s.k. likelihood-funktion. Bayes sats säger att den normaliserade produkten av dessa beskriver den så kallade posteriorn, dvs. fördelningen för variabeln som skattningen ska baseras på. Kärnproblemet i avhandlingen är att denna funktion i de flesta fall inte är analytisk, dvs kan skrivas som en matematiskt uttryck, och måste approximeras på ett eller annat sätt. Ett antal effektiva metoder presenteras i detta arbete.

Bilsäkerhet är en illustrativ tillämpning som studeras av författaren. Antag att mjukvaran i en kamera har upptäckt en cykel framför bilen den är monterad i, och att mjukvaran dessutom kan hitta de två hjulens positioner i bilden. Hjulen avbildas som ellipser i bilden, och man kan då med hjälp av ellipsernas form och Bayes sats förfina informationen om cykelns position och dessutom skatta hur cykeln kommer att ändra riktning, t.ex. om den håller på att svänga ut framför bilen. Bayes sats kan nämligen användas en gång till för att förutsäga var cykeln kommer befinna sig när nästa kamerabild tas. Dessa två steg för att förutsäga och skatta är grundkomponenter i olinjära filter, som är ett fokusområde i avhandlingen. Konceptet att modellera cykeln inte bara som en punkt, utan som en struktur med två hjul, går under benämningen utökat mål, och detta område har varit en motivation för många av resultaten i avhandlingen.

När man använder Bayes sats finns det ett antal kombinationer av prior och likelihood som faktiskt ger en analytisk posterior. En sådan prior som passar till en viss likelihood kallas konjugerad prior. En metod som föreslås är att approximera likelihooden så att den prior man har blir konjugerad. Med hjälp av detta trick blir posteriorn analytisk, och även på samma form som priorn. Det senare är viktigt när operationen ska upprepas många gånger, som t.ex. i ett olinjärt filter.

Ett annat exempel då en posterior har ett analytiskt uttryck är när både prior och likelihood är viktade summor av normalfördelningar. Sådana fördelningar är mycket flexibla och kan approximera vilken fördelning som helst godtyckligt väl. Posteriorn blir då också en viktad summa av normalfördelningar. Problemet är att den får allt fler komponenter varje gång Bayes sats används. Ett av avhandlingens bidrag tar fram konkreta algoritmer för att begränsa antalet komponenter genom smarta approximationer. Rent allmänt kan posteriorn alltid approximeras inom en given funktionsklass, och här studeras Kullback-Leibler som mått att optimera över. Detta används för att skatta impulssvar för dynamiska system. En metod som används i flera bidrag är Variational Bayes (VB). Här används VB till att hitta en produktform av posteriorn över två delmängder av variabler som ska skattas, vilket enligt VB kan göras stegvis med stora besparingar i beräkningskomplexitet.

Acknowledgments

On the 6th of November 1632, the Swedish king Gustav II Adolf, the founder of the Swedish Empire (1611-1721), was killed in the battle of Lützen in modern-day Germany. Gustav II Adolf was an extremely able commander (rather than an obedient soldier), was nearsighted and had a prominent nose. It is claimed by historians that in the thick mix of gun smoke and fog covering the battlefield, he was separated from his fellow riders and killed by several shots.

It is indeed just a coincidence that this thesis will be defended on the very same day 383 years later. Even the fact that I have well-known aspirations for becoming the king of Sweden does not worry me. Furthermore, I am not concerned about the fact that my opponent Dr Wolfgang Koch comes from a German defense institution which is situated only 500 km away from Lützen. *Let us stay objective when forming prior beliefs*. Furthermore, separation from my fellow riders is not expected to happen since I have written this thesis to enable me and my fellow riders to see through the fog of *noise* and smoke of *disturbances* using Bayes' rule.

I would never have been able to write such a dissertation without the support of my fellow riders. Here, I want to acknowledge their contributions to this thesis. 131 years after that fateful day reverend Thomas Bayes wrote the article "An Essay towards solving a Problem in the Doctrine of Chances " for which I am grateful. It took another 246 year till Lennart Ljung admitted me to the group and gave the opportunity to find my supervisor and my research subject. Thank you Lennart.

I want to thank my supervisor Fredrik Gustafsson for his engagement in the beginning and the end of my PhD studies and his patience along the way. I want to thank the head of division of Automatic Control Svante Gunnarsson for giving me the space to maneuver and making exceptions of traditions whenever I wished for them.

Umut Orguner possesses every quality one would desire in a supervisor. I offer my sincerest gratitude to Umut Orguner for sharing his vast knowledge, his patience and his generosity. Thank you Emre Özkan and Fredrik Gustafsson for proof reading this thesis.

During my graduate studies I had the opportunity to collaborate with many talented researchers. It was a pleasure to co-author papers with Jonas Sjöberg, Jonas Bärgman, Mathias Lidberg, Robert Thomson, Anders Hansson, Johan Löfberg, Mikael Norrlöf, Fredrik Larsson, Michael Felsberg, Fredrik Gustafsson, Thomas B. Schön, Christian Lundquist, Umut Orguner, Emre Özkan, Karl Granström, Tianshi Chen, Henri Nurminen, Robert Piché, Francesca P. Carli, Alessandro Chiuso, Lennart Ljung, and Gianluigi Pillonetto. Not all the research work resulted in publications; My conversations with Fredrik Lindsten, Saikat Saha, Michael Roth, Daniel Axehill, Martin Enqvist, Carsten Fritsche, Liam Ellis, and Claudio Altafini have been enlightening. I am very excited about my ongoing collaborations with Rafael Rui, Michael Roth, Henri Nurminen, and Mikhail Lifshits. Furthermore, I learned a lot from Anders Hansson and Torkel Glad through their graduate courses for which I am grateful.

I would like to thank division's secretaries Ulla Salaneck, Åsa Karmelind and Ninna Stensgård. Peter Rosander with whom I shared my office for three years is the kindest and most considerate Swedish man I have met so far. I want to thank Michael Roth and Kora Neupert for being such a wonderful friends. I want to thank Henrik Ohlsson for helping me get on my board so many times on those cold windy days.

I want to extend my gratitude to George Mathai, Niklas Wahlström, Alicia Pamela Tonnoli, Bram Dil, Ylva Jung, Fredrik Lindsten, Patrik Axelsson, Manon Kok, Martin Lindfors, Gustav Lindmark, Christian Andersson Naesseth , Hanna Nyqvist, Jonas Linder, Karl Granström, André Carvalho Bittencourt and Christian Lundquist for all the good time be it while playing board games, skiing in the Alps, and kite surfing at Swedish west coast.

I also want to thank Rikard Falkeborn, Daniel Petersson, Martin Skoglund and Christian Lyzell for patiently helping me with software related questions and especially Henrik Tidefelt for his contributions to the LaTeX template used for this thesis.

I offer my sincerest gratitude to the hardworking people of Sweden who gave me the possibility to study my PhD degree not worrying about funding and gave me the opportunity to have paid paternity leave. I wish to acknowledge the financial support from the frame project Extended Target Tracking and project Scalable Kalman Filters both funded by Swedish Research Council.

I want to thank my teachers at Sharif University of Technology for being exceptionally dedicated teachers, especially Mohammad Durali and Ali Meghdari. I want to thank Azadeh, Nazanin, Navid, Kavous, Hamed, Shahin, Shahla, Hossein and Hooshang for their support during my undergraduate studies.

Adrian! Your very existence, beautiful smiles, big hugs and kisses has been my shelter during these years. I love you more than "a thousand million and nine hundreds ninety one". Last but not the least, I want to thank my wonderful parents who taught me by example many things above all how to endure during difficult times.

Linköping, October 2015
Tohid Ardeshiri

Contents

II Publications

Notation

Abbreviations

Abbreviation	Meaning
ARMSE	Average Root Mean Square Error
ETT	Extended Target Tracking
EKF	Extended Kalman Filter
EM	Expectation Maximization
GLM	Generalized Linear Models
GM	Gaussian Mixture
GM-PHD	Gaussian Mixture Probability Hypothesis Density
GP	Gaussian Process
GPS	Global Positioning System
iid	independent and identically distributed
INLA	Integrated Nested Laplace Approximation
KF	Kalman Filter
KLD	Kullback-Leibler Divergence
LTI	Linear Time-Invariant
MAP	Maximum A Posteriori
MC	Monte-Carlo
MR	Mixture Reduction
MRA	Mixture Reduction Algorithm
MTT	Multiple Target Tracking
PDF	Probability Density Function
PHD	Probability Hypothesis Density
RKLD	Reverse Kullback-Leibler Divergence
RMSE	Root Mean Square Error
RTS	Rauch-Tung-Striebel
SS-1	First Order Stable Spline
STVBF	Skew-t Variational Bayes Filter
STVBS	Skew-t Variational Bayes Smoother
VB	Variational Bayes

Some sets

Notation	Meaning
$\mathbb{N}$	Set of natural numbers
$\mathbb{R}$	Set of real numbers
S_{++}^{d}	Set of $d \times d$ symmetric positive definite matrices

Probability

Notation	Meaning
$\mathbb{E}$	Expectation
$\mathbb{V}$	Variance
$D_{KL}(\cdot \Vert \cdot)$	Kullback-Leibler divergence
H	Differential entropy
$\overline{H}$	Differential entropy rate
$\sim$	Distributed according to or sampled from

Common distributions

Notation	Meaning
$\mathcal{N}(\mu, \Sigma)$	Multivariate Gaussian with mean μ and covariance Σ
$\mathcal{U}(a, b)$	Uniform over the interval $[a, b]$
$\text{Exp}(x; \lambda)$	Exponential with rate λ
$\text{Weibull}(\lambda, k)$	Weibull with scale λ and shape k
$\text{Laplace}(\mu, b)$	Laplace with location μ and scale b
$\text{Rayleigh}(\sigma)$	Rayleigh with scale σ
$\log-\mathcal{N}(\mu, \sigma)$	Log-normal with location μ and scale σ
$\text{Gamma}(\alpha, \beta)$	Gamma with shape α and rate β
$\text{IGamma}(\alpha, \beta)$	Inverse gamma with shape α and rate β
$\mathcal{W}_d(n, V)$	Wishart with degrees of freedom n and scale matrix $V \in S_{++}^{d}$
$\mathcal{IW}_d(\nu, \Psi)$	Inverse Wishart with degrees of freedom ν and scale matrix $\Psi \in S_{++}^{d}$
$t(\mu, \sigma^2, \nu)$	Student's t-distribution with location parameter μ, spread parameter σ, and degrees of freedom ν
$\text{T}(\cdot; 0, 1, \nu)$	Cumulative distribution function (CDF) of Student's t-distribution with degrees of freedom ν
$\text{ST}(z; \mu, \sigma^2, \delta, \nu)$	Skew t-distribution with location parameter μ, spread parameter σ, shape parameter δ and degrees of freedom ν
$\mathcal{N}_+(\mu, \Sigma)$	Truncated Multivariate Gaussian with closed positive orthant as support, location parameter μ and squared-scale matrix Σ

Operators and Symbols

Notation	Meaning
$\mathrm{tr}\,(A)$	Trace of matrix A
$\det(A)$, $\lvert A\rvert$	Determinant of matrix A
A^{T}	Transpose of matrix A
$\mathbf{vec}(A)$	Vectorized matrix A
$\mathrm{Diag}(\cdot)$	Diagonal matrix whose diagonal elements are the arguments of the operator
$x_{m:n}$	Sequence $(x_m, x_{m+1}, \cdots, x_n)$
I_d	d-dimensional identity matrix
$\mathcal{H}$	Hypothesis
argmin_λ	Minimizing argument with respect to λ
argmax_λ	Maximizing argument with respect to λ

Part I

Background

1

Introduction

This chapter introduces the research area that is considered in this thesis and summarizes the contributions that constitute this thesis. In Section 1.1, an introduction to the approximate Bayesian inference is given. In Section 1.2, the main contributions are summarized. In Section 1.3, the publications by the PhD candidate are listed. In Section 1.4 three applied research results produced by the PhD candidate are presented. In Section 1.5, the outline of the thesis is given.

1.1 Approximate Bayesian Inference

Bayesian inference is a statistical inference technique in which Bayes' theorem is used to update the probability distribution of a random latent variable using observations. This technique provides a mathematical tool for modeling systems where uncertainties of the model, as well as the system, are reflected by the probability distributions. The probabilistic models which are constructed by probability distributions that describe our knowledge about the system are determined using the rules of the probability calculus.

Probabilistic models describe the relation between the random latent variables, the deterministic parameters, and the measurements. Such relations are specified by prior distributions of the latent variables $p(\mathbf{x})$, and the likelihood function $p(\mathbf{y}|\mathbf{x})$ which gives a probabilistic description of the measurements given (some of) the latent variables. Using the probabilistic model and measurements the exact posterior can be expressed in a functional form using the Bayes' rule

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x})p(\mathbf{y}|\mathbf{x})}{\int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{x}}. \tag{1.1}$$

The prior knowledge about the latent variables and the parameters is expressed via prior distributions. Ideally, the prior distribution should express

this prior knowledge about the latent variables without any extra assumptions. The maximum entropy method (Jaynes, 1982) provides a tool to express the prior knowledge in form of prior distributions without further assumptions. Describing the prior knowledge about a random variable using compact analytical expressions is not always feasible. In such cases approximation methods are required. One of the contributions in this thesis concerns such approximations.

Determination of the posterior distribution of a latent variable $\mathbf{x}$ given the measurements (observed data) $\mathbf{y}$ is at the core of Bayesian inference using probabilistic models. The exact posterior distribution can be analytical. A subclass of cases where the posterior is analytical is when the posterior belongs to the same family of distributions as the prior distribution. In such cases the prior distribution is called a conjugate prior for the likelihood function. A well-known example where analytical posterior is obtained using conjugate priors is when the latent variable is *a priori* normal-distributed and the likelihood function given the latent variable as its mean is again normal.

Example 1.1

Let $\mathbf{x}$ have a normal prior distribution with mean μ and covariance Σ, i.e., $\mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mu, \Sigma)$. A measurement $\mathbf{y}$ with the likelihood function $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; H\mathbf{x}, R)$ is in hand where H is a matrix with proper dimensions and R is a covariance matrix. The posterior distribution of $\mathbf{x}$ can be obtained using the Bayes' rule given in (1.1);

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x})p(\mathbf{y}|\mathbf{x})}{\int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{x}} \tag{1.2}$$

$$= \frac{\mathcal{N}(\mathbf{x}; \mu, \Sigma)\mathcal{N}(\mathbf{y}; H\mathbf{x}, R)}{\int \mathcal{N}(\mathbf{x}; \mu, \Sigma)\mathcal{N}(\mathbf{y}; H\mathbf{x}, R)\, \mathrm{d}\mathbf{x}}. \tag{1.3}$$

The posterior distribution $p(\mathbf{x}|\mathbf{y})$ has an analytical solution and turns out to be the normal distribution $\mathcal{N}(\mathbf{x}; \mu', \Sigma')$ whose parameters can be computed via closed form expressions given here,

$$\mu' = \mu + K(\mathbf{y} - H\mu), \tag{1.4a}$$

$$\Sigma' = \Sigma - KH\Sigma, \tag{1.4b}$$

where

$$K = \Sigma H(H\Sigma H^{\mathrm{T}} + R)^{-1}. \tag{1.5}$$

The exact posterior distribution of a latent variable can not always be given a compact analytical expression. In the following, three examples of such cases will be given. In Example 1.2, a problem that is encountered in nonlinear filtering is presented.

Example 1.2

Let $\mathbf{x}$ have a normal prior distribution $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu, \Sigma)$. A measurement $\mathbf{y}$ with the likelihood function $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; h(\mathbf{x}), R)$ is in hand where $h(\cdot)$ is a vector valued nonlinear function and R is a covariance matrix. The posterior distribution of $\mathbf{x}$ can be expressed using the Bayes' rule given in (1.1). However, the posterior distribution does not necessarily have a compact analytical solution. A remedy can be obtained by approximation of the likelihood function via linearization of the function $h(\cdot)$ around the prior mean μ as in

$$h(\mathbf{x}) \approx h(\mu) + \widehat{H}(\mathbf{x} - \mu) \tag{1.6}$$

where

$$\widehat{H} \triangleq \nabla_{\mathbf{x}} h(\mathbf{x})|_{\mathbf{x}=\mu}. \tag{1.7}$$

Using the approximate likelihood $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \widehat{H}\mathbf{x}, R)$, the approximate posterior distribution can be computed using the analytical expressions given in Example 1.1.

One of the contributions in this thesis concerns the problem and the solution proposed in Example 1.2. In Example 1.3, a problem that is encountered in simulating a Markov chain with multi-modal transition density will be presented.

Example 1.3

Consider a Markov chain with transition density:

$$p(\mathbf{x}_{k+1}|\mathbf{x}_k) = w\mathcal{N}(\mathbf{x}_{k+1}; A\mathbf{x}_k, Q) + (1-w)\mathcal{N}(\mathbf{x}_{k+1}; \overline{A}\mathbf{x}_k, \overline{Q}), \tag{1.8}$$

where $0 < w < 1$, and factors A and $\overline{A}$ are two square matrices. We are interested in the marginal distribution of $\mathbf{x}_{1000}$, where $\mathbf{x}_1$ has the distribution $p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1; \mu_1, \Sigma_1)$. The marginal distribution of $\mathbf{x}_k$ can be obtained recursively by integration as in

$$p(\mathbf{x}_k) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}) \, \mathrm{d}\mathbf{x}_{k-1}. \tag{1.9}$$

The first step of the recursion is computed here.

$$p(\mathbf{x}_2) = \int \left(w\mathcal{N}(\mathbf{x}_2; A\mathbf{x}_1, Q) + (1-w)\mathcal{N}(\mathbf{x}_2; \overline{A}\mathbf{x}_1, \overline{Q}) \right) \mathcal{N}(\mathbf{x}_1; \mu_1, \Sigma_1) \, \mathrm{d}\mathbf{x}_1 \tag{1.10}$$

$$= w\mathcal{N}(\mathbf{x}_2; A\mu_1, Q + A\Sigma_1 A^{\mathrm{T}}) + (1-w)\mathcal{N}(\mathbf{x}_2; \overline{A}\mu_1, \overline{Q} + \overline{A}\Sigma_1 \overline{A}^{\mathrm{T}}) \tag{1.11}$$

Although the marginal density of $\mathbf{x}_2$ can be computed analytically, the complexity of $p(\mathbf{x}_2)$ has increased compared to $p(\mathbf{x}_1)$ due to increase in the number of Gaussian components needed to express $p(\mathbf{x}_2)$. The number of Gaussian components needed to express the marginal density of $\mathbf{x}_k$ grows exponentially with

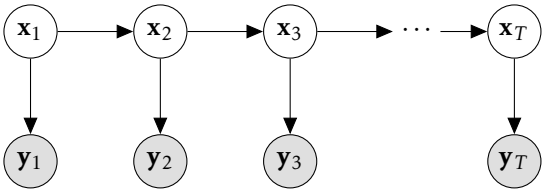


Figure 1.1: *A probabilistic graphical model for stochastic dynamical system with latent state $\mathbf{x}_k$ and measurements $\mathbf{y}_k$.*

respect to k and is 2^{k-1}. In order to maintain the computational complexity of the marginal distributions of $\mathbf{x}_k$ at a tractable level, the number of components needs to be reduced via approximation of the true marginal density of $\mathbf{x}_k$ with another distribution with less components. A candidate solution for the problem is minimizing a statistical distance between the true density of $\mathbf{x}_k$ and its approximation.

Two of the contributions in this thesis concern the problem described in Example 1.3. Approximate Bayesian inference is particularly important when the measurements appear sequentially in time as in the filtering task for a stochastic dynamical system, whose probabilistic graphical model is presented in Figure 1.1. In Example 1.4, the Bayesian filtering recursion is introduced and the need for approximations is highlighted. Three contributions in this thesis concern problems of similar nature to Example 1.4.

Example 1.4

Consider a stochastic dynamical system represented by the following recursion

$$\mathbf{x}_1 \sim p(\mathbf{x}_1), \tag{1.12a}$$
$$\mathbf{y}_k \sim p(\mathbf{y}_k|\mathbf{x}_k), \tag{1.12b}$$
$$\mathbf{x}_{k+1} \sim p(\mathbf{x}_{k+1}|\mathbf{x}_k). \tag{1.12c}$$

The Bayesian filtering recursion corresponds to computing the posterior distributions $p(\mathbf{x}_k|\mathbf{y}_{1:k})$;

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}) = \frac{p(\mathbf{x}_k|\mathbf{y}_{1:k-1})p(\mathbf{y}_k|\mathbf{x}_k)}{\int p(\mathbf{x}_k|\mathbf{y}_{1:k-1})p(\mathbf{y}_k|\mathbf{x}_k)\,\mathrm{d}\mathbf{x}_k}. \tag{1.13}$$

The density $p(\mathbf{x}_k|\mathbf{y}_{1:k-1})$ in the numerator of (1.13) which is called the predicted density of $\mathbf{x}_k$ and is obtained by integration as in

$$p(\mathbf{x}_k|\mathbf{y}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1})\,\mathrm{d}\mathbf{x}_{k-1}. \tag{1.14}$$

In such filtering problems, the posterior to the last processed measurement is the prior distribution in the next time step. To be able to use the same inference algorithm in a recursive manner, the posterior distribution at each time step should

obtain the same form as the prior. When such a condition does not exist, approximations can be used. A class of such approximations is called the variational approximations, where the posterior is assumed to have a specific functional form (the same as the prior). Subsequently a statistical distance between the assumed posterior and the true posterior is minimized to find the hyper-parameters of the assumed (approximate) posterior.

Several methods for approximate inference over probabilistic models are proposed in the literature such as variational Bayes (Jordan et al., 1999), expectation propagation (Minka, 2001), integrated nested Laplace approximation (INLA) (Rue et al., 2009), generalized linear models (GLMs) (Nelder and Wedderburn, 1972) and, Monte-Carlo (MC) sampling methods (Hastings, 1970; Geman and Geman, 1984).

Variational Bayes (VB) and expectation propagation (EP) are two optimization-based solutions to the approximate Bayesian inference (Wainwright and Jordan, 2008). In these two approaches Kullback-Leibler divergence (Cover and Thomas, 2006) between the true posterior distribution and an approximate posterior is minimized. INLA is a technique to perform approximate Bayesian inference in latent Gaussian models (Hennevogl et al., 2001) using the Laplace approximation. GLMs are an extension of ordinary linear regression when errors belong to the exponential family.

Sampling methods such as Markov Chain Monte Carlo (MCMC) methods provide a general class of solutions to the approximate Bayesian inference problem.

In this thesis, the focus is on fast analytical approximations which are applicable to large-scale inference problems. These approximations propose solutions to the Bayesian inference problems where the vanilla versions are described in Examples 1.2, 1.3 and 1.4. These analytical approximations either involve minimization of a statistical divergence between the true distribution and its approximation or are based on expansion of a function with respect to a basis function.

In Section 1.2, the main contributions in this thesis are summarized. In Section 1.5, the connection between the problems highlighted here in form of Examples 1.2, 1.3 and 1.4 and their corresponding contributions in this thesis will be drawn.

1.2 Contributions

The contributions of this thesis address various aspects of Bayesian inference. These contributions can be categorized in three groups:

1. **Prior selection:** The prior information about a stochastic process in a Gaussian process regression problem can be encoded in the covariance function. The maximum entropy properties of a covariance function for Gaussian process regression referred to as the discrete-time first-order stable spline kernel is proven.

2. **Determination of the posterior distribution for dynamical systems:** Approximate posterior of two Bayesian inference problems are derived using the variational Bayes technique. Furthermore, an approximation method for general Bayesian inference problems using linearization of log-likelihood function is proposed. These contributions in this category concern problems such as those highlighted in Examples 1.2 and 1.4.

3. **Maintenance of computational complexity:** The contributions in this category concern with maintenance of computational complexity in problems such as the one introduced in Example 1.3.

1.3 Publications

The following papers, listed in reverse chronological order, are published

T. Ardeshiri, U. Orguner, and F. Gustafsson. Bayesian inference via approximation of log-likelihood for priors in exponential family. *ArXiv e-prints*, October 2015b. Submitted to Signal Processing, IEEE Transactions on.

T. Ardeshiri, E. Özkan, U. Orguner, and F. Gustafsson. Approximate Bayesian smoothing with unknown process and measurement noise covariances. *To appear in Signal Processing Letters, IEEE*, 2015.

T. Chen, T. Ardeshiri, F. P. Carli, A. Chiuso, L. Ljung, and G. Pillonetto. Maximum entropy properties of discrete-time first-order stable spline kernel. *To appear in Automatica*, 2015.

T. Ardeshiri, U. Orguner, and E. Özkan. Gaussian Mixture Reduction Using Reverse Kullback-Leibler Divergence. *ArXiv e-prints*, August 2015. To be Submitted to Signal Processing, IEEE Transactions on.

H. Nurminen, T. Ardeshiri, R. Piché, and F. Gustafsson. A NLOS-robust TOA positioning filter based on a skew-t measurement noise model. In *2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Banff, Alberta, Canada*, October 2015b.

H. Nurminen, T. Ardeshiri, R. Piché, and F. Gustafsson. Robust inference for state-space models with skewed measurement noise. *Signal Processing Letters, IEEE*, 22(11):1898–1902, Nov 2015a. ISSN 1070-9908. doi: 10.1109/LSP.2015.2437456.

T. Ardeshiri, K. Granström, E. Özkan, and U. Orguner. Greedy reduction algorithms for mixtures of exponential family. *Signal Processing Letters, IEEE*, 22(6):676–680, June 2015a. ISSN 1070-9908. doi: 10.1109/LSP.2014.2367154.

T. Ardeshiri and T. Chen. Maximum entropy property of discrete-time stable spline kernel. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 3676–3680, April 2015. doi: 10.1109/ICASSP.2015.7178657.

T. Ardeshiri and E. Özkan. An adaptive PHD filter for tracking with unknown sensor characteristics. In *Information Fusion (FUSION), 2013 16th International Conference on*, pages 1736–1743, July 2013.

T. Ardeshiri, U. Orguner, C. Lundquist, and T. Schön. On mixture reduction for multiple target tracking. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 692–699, July 2012.

T. Ardeshiri, F. Larsson, F. Gustafsson, T. Schön, and M. Felsberg. Bicycle tracking using ellipse extraction. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8, July 2011a.

T. Ardeshiri, M. Norrlöf, J. Löfberg, and A. Hansson. Convex optimization approach for time-optimal path tracking of robots with speed dependent constraints. In *Proceedings of the 18th IFAC World Congress, Milan, Italy*, pages 14648–14653, August 2011b.

T. Ardeshiri, S. Kharrazi, R. Thomson, and J. Bärgman. Offset eliminative map matching algorithm for intersection active safety applications. In *Intelligent Vehicles Symposium, 2006 IEEE*, pages 82–88, 2006b. doi: 10.1109/IVS.2006.1689609.

T. Ardeshiri, S. Kharrazi, J. Sjöberg, J. Bärgman, and L. M. Sensor fusion for vehicle positioning in intersection active safety applications. In *International Symposium on Advanced Vehicle Control*, 2006a.

1.4 Applications

In this section, a summary of three applied research results produced by the PhD candidate is presented.

1.4.1 Bicycle tracking using ellipse extraction

A new approach to track bicycles from imagery sensor data is proposed in (Ardeshiri et al., 2011a). It is based on detecting ellipsoids in the image as in Figures 1.2 and 1.3. These ellipses are treated these pair-wise using a dynamic bicycle model illustrated in Figure 1.4. One important application area is in automotive collision avoidance systems, where no dedicated systems for bicyclists yet exist and where very few theoretical studies have been published. Possible conflicts can be predicted from the position and velocity state in the model, but also from the steering wheel articulation and roll angle that indicate yaw changes before the velocity vector changes. An algorithm is proposed in (Ardeshiri et al., 2011a) which consists of an ellipsoid detection and estimation algorithm and a particle filter. A simulation study of three critical single target scenarios is presented, and the algorithm is shown to produce excellent state estimates. An experiment using a stationary camera and the particle filter for state estimation is performed and has shown encouraging results.

Figure 1.2: *The green ellipses indicate measurements obtained from the two bike wheels. The ellipse parameters are later fed through a particle filter framework in order to estimate the bicycle state.*

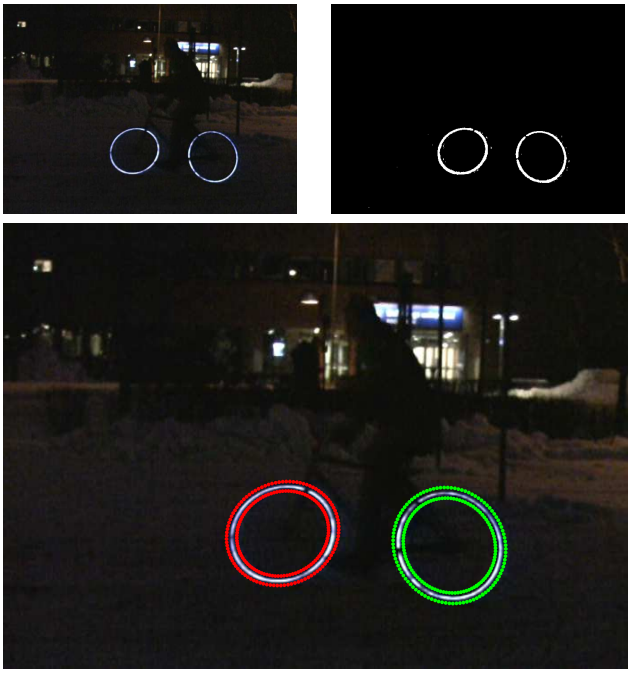
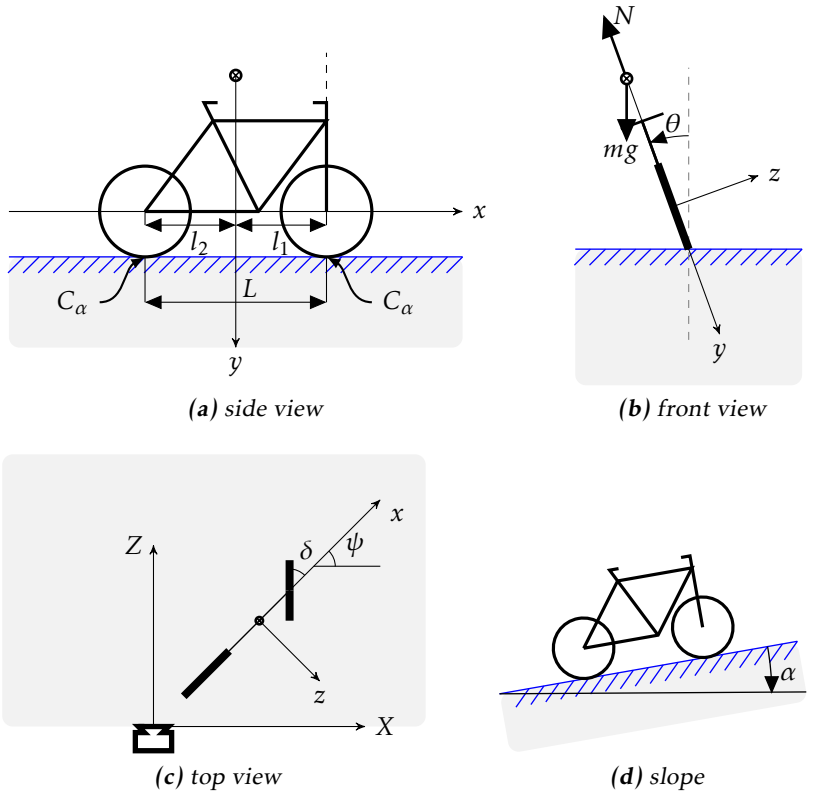


Figure 1.3: *Ellipse extraction. Top left: Query image, Top right: Query image after background subtraction. Bottom: Ellipses plotted with 0.9 and 1.1 times the estimated size, the actual estimated ellipses are halfway between the lines.*



(a) side view **(b)** front view

(c) top view **(d)** slope

Figure 1.4: *(a) Illustration of the coordinate system and the bicycle parameters. The wheelbase L and the distance of center of gravity to the wheel centers is denoted by l_1 and l_2. The y-axis goes through the center of gravity and the x-axis goes through the wheel centers. (b) Illustration of the inclination θ of the bicycle. The inclination angle can be calculated using Newton's second law of motion. The gravitational force is denoted by mg and the reaction force of the ground is denoted by N. (c) An extended bicycle model is used as motion model where ψ and δ are shown in this figure. The orientation of the camera at the origin of the global coordinate system is shown. (d) The slope of the bicycle's track is denoted by α.*

1.4.2 Positioning using ultra wide-band data

The skew-t variational Bayes filter (STVBF) (Nurminen et al., 2015a) is applied to indoor positioning with time-of-arrival (TOA) based distance measurements and pedestrian dead reckoning (PDR) in (Nurminen et al., 2015b). The proposed filter accommodates large positive outliers caused by occasional non-line-of-sight (NLOS) conditions by using a skew-t model of measurement errors. Real-data tests using the fusion of inertial sensors based PDR and ultra-wideband based TOA ranging show that the STVBF clearly outperforms the extended Kalman filter (EKF) in positioning accuracy with the computational complexity about three times that of the EKF. A tracking performance of one of the test tracks is illustrated in Figure 1.5.

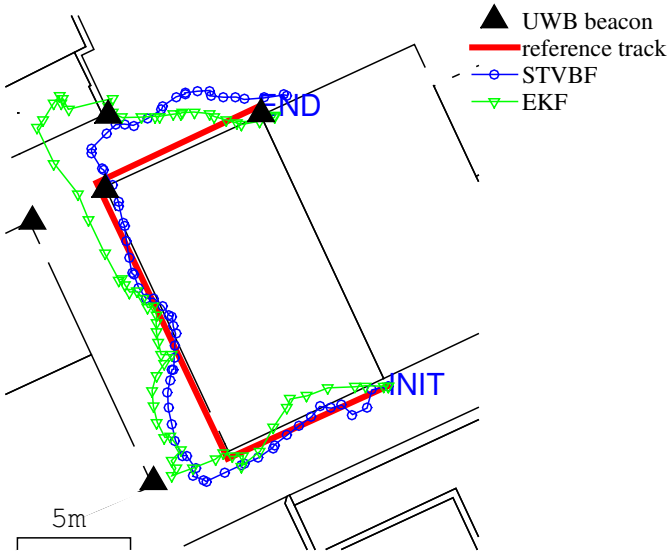


Figure 1.5: *Test track 1 consists of corridors and turns at corridor junctions.*

1.4.3 Path tracking for robots

The task of generating time optimal trajectories for a six degrees of freedom industrial robot is discussed in (Ardeshiri et al., 2011b) and an existing convex optimization formulation of the problem is extended to include new types of constraints. The new constraints are speed dependent and can be motivated from physical modeling of the motors and the drive system. It is shown how the speed dependent constraints should be added in order to keep the convexity of the overall problem. A method to, conservatively, approximate the linear speed dependent constraints by a convex constraint is also proposed (see Figure 1.6). A numerical example proves versatility of the extension proposed in (Ardeshiri et al., 2011b).

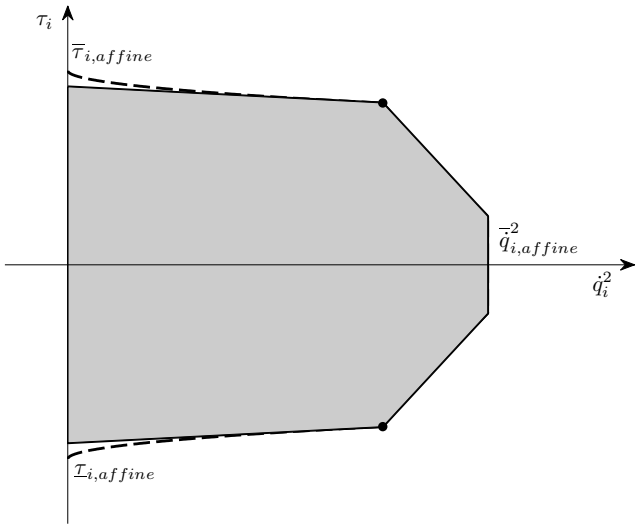


Figure 1.6: *The torque at a joint of a robotic arm is plotted versus the square of angular velocity of the same joint. The non-convex true feasible set is approximated by a set of affine constraints. The true actuator's constraints is represented by the dashed line. The approximation of the feasible set by a convex set is illustrated by the hatched area.*

1.5 Thesis outline

The thesis is divided into two parts. In the rest of the first part, background material for these contributions will be provided[1]. In the second part of the thesis, a compilation of six edited publications are presented.

1.5.1 Outline of Part I

In Chapter 2, the concepts of entropy, relative entropy, and maximum entropy priors, and their relation to the exponential family are introduced. Also a short introduction to variational Bayes method is given. The background material in Chapter 2 are intended to lay the theoretical foundation for the Papers A, B, C, D and E in the second part of this thesis.

In Chapter 3, a short introduction to the problem of identification of linear time-invariant, stable and causal systems using Gaussian process regression methods is given. This Chapter is intended to give an introduction to the problem addressed in Paper A which is about approximation of the prior knowledge for the purpose of devising a maximum entropy prior distribution.

[1] Parts of the material presented in the first part of the thesis is already published by the author in form of technical reports, conference papers and journal articles.

In Chapter 4, an introduction to the mixture reduction problem introduced in Example 1.3 is presented. The mixture reduction problem is addressed in the second part of this thesis by Papers E and F. Concluding remarks are given in Chapter 5.

1.5.2 Outline of Part II

Part II of the thesis is a compilation of six edited contributions which are summarized in the following.

Maximum entropy properties of discrete-time first-order stable spline kernel

Paper A

> T. Chen, T. Ardeshiri, F. P. Carli, A. Chiuso, L. Ljung, and G. Pillonetto. Maximum entropy properties of discrete-time first-order stable spline kernel. *To appear in Automatica*, 2015.

presents the maximum entropy properties of the discrete-time first-order stable spline kernel. The first order stable spline (SS-1) kernel (also known as the tuned-correlated kernel) is used extensively in regularized system identification, where the impulse response is modeled as a zero-mean Gaussian process whose covariance function is given by well designed and tuned kernels. In particular, the exact maximum entropy problem solved by the SS-1 kernel without Gaussian and uniform sampling assumptions is formulated. Under general sampling assumption, the special structure of the SS-1 kernel (e.g. its tridiagonal inverse and factorization have closed form expression) is derived. Also a maximum entropy covariance completion interpretation is given to it.

Approximate Bayesian smoothing with unknown process and measurement noise covariances

Paper B

> T. Ardeshiri, E. Özkan, U. Orguner, and F. Gustafsson. Approximate Bayesian smoothing with unknown process and measurement noise covariances. *To appear in Signal Processing Letters, IEEE*, 2015.

presents an adaptive smoother for linear state-space models with unknown process and measurement noise covariances. The proposed method utilizes the variational Bayes technique to perform approximate inference. The resulting smoother is computationally efficient, easy to implement, and can be applied to high dimensional linear systems. The performance of the algorithm is illustrated on a target tracking example.

Robust inference for state-space models with skewed measurement noise

Paper C

> H. Nurminen, T. Ardeshiri, R. Piché, and F. Gustafsson. Robust inference for state-space models with skewed measurement noise. *Signal Processing Letters, IEEE*, 22(11):1898–1902, Nov 2015a. ISSN 1070-9908. doi: 10.1109/LSP.2015.2437456.

presents filtering and smoothing algorithms for linear discrete- time state-space models with skewed and heavy-tailed measurement noise. The algorithms use a variational Bayes approximation of the posterior distribution of models that have normal prior and skew-t-distributed measurement noise. The proposed filter and smoother are compared with conventional low-complexity alternatives in a simulated pseudorange positioning scenario. In the simulations the proposed methods achieve better accuracy than the alternative methods, the computational complexity of the filter being roughly 5 to 10 times that of the Kalman filter.

Bayesian inference via approximation of log-likelihood for priors in exponential family

Paper D

> T. Ardeshiri, U. Orguner, and F. Gustafsson. Bayesian inference via approximation of log-likelihood for priors in exponential family. *ArXiv e-prints*, October 2015b. Submitted to Signal Processing, IEEE Transactions on.

presents a Bayesian inference technique based on Taylor series approximation of the logarithm of the likelihood function. The proposed approximation is devised for the case where the prior distribution belongs to the exponential family of distribution and is continuous. The logarithm of the likelihood function is linearized with respect to the sufficient statistic of the prior distribution in exponential family such that the posterior obtains the same exponential family form as the prior. Similarities between the proposed method and the extended Kalman filter for nonlinear filtering are illustrated. Further, an extended target measurement update for target models where the target extent is represented by a random matrix having an inverse Wishart distribution is derived. The approximate update covers the important case where the spread of measurement is due to the target extent as well as the measurement noise in the sensor.

Greedy reduction algorithms for mixtures of exponential family

Paper E

> T. Ardeshiri, K. Granström, E. Özkan, and U. Orguner. Greedy reduction algorithms for mixtures of exponential family. *Signal Processing Letters, IEEE*, 22(6):676–680, June 2015a. ISSN 1070-9908. doi: 10.1109/LSP.2014.2367154.

presents a general framework for greedy reduction of mixture densities of exponential family. The performances of the generalized algorithms are illustrated both on an artificial example where randomly generated mixture densities are reduced and on a target tracking scenario where the reduction is carried out in the recursion of a Gaussian inverse Wishart probability hypothesis density (PHD) filter.

Gaussian mixture reduction using reverse Kullback-Leibler divergence

Paper F

> T. Ardeshiri, U. Orguner, and E. Özkan. Gaussian Mixture Reduction Using Reverse Kullback-Leibler Divergence. *ArXiv e-prints*, August 2015. To be Submitted to Signal Processing, IEEE Transactions on.

presents a greedy mixture reduction algorithm which is capable of pruning mixture components as well as merging them based on the Kullback-Leibler divergence (KLD). The algorithm is distinct from the well-known Runnalls' KLD based method since it is not restricted to merging operations. The capability of pruning (in addition to merging) gives the algorithm the ability of preserving the peaks of the original mixture during the reduction. Analytical approximations are derived to circumvent the computational intractability of the KLD which results in a computationally efficient method. The proposed algorithm is compared with Runnalls' and Williams' methods in two numerical examples, using both simulated and real world data. The results indicate that the performance and computational complexity of the proposed approach make it an efficient alternative to existing mixture reduction methods.

2

Entropy, Exponential Family, and Variational Bayes

Analytical approximations proposed in the second part of this thesis build upon the existing literature on maximum entropy priors, exponential family of distributions and variational Bayes. In this chapter some preliminary definitions and results relating to these contributions will be given. In Section 2.1, entropy and the relative entropy will be defined. Furthermore, maximum entropy distributions will be derived. The background material in section 2.1 will lay the foundations for Paper A in the second part of the thesis. Also the relationship between the maximum entropy priors and the exponential family will be explained. In Section 2.2, the exponential family of distributions and some of their properties will be given. These background material will be used in Papers E and D which are about approximate inference techniques relating to exponential family of distributions. In Section 2.3 the variational Bayes (VB) method is described. The VB method is used to derive approximate posteriors in Papers B and C.

2.1 Entropy

Entropy is a measure of the uncertainty of a random variable. In this thesis, only continuous random variables are considered. Consequently, only the aspects of the information theory which are related to continuous random variables will be covered. The definitions of the differential entropy and relative entropy will be given in the following.

Definition 2.1. For a distribution with its support on S with density $p(\cdot)$, the differential entropy is defined by (Cover and Thomas, 2012)

$$H(p) = -\int_S p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{2.1}$$

Example 2.2

For standard normal distribution in $\mathbb{R}^n$ where $p(\mathbf{x}) = (2\pi)^{-n/2} \exp\{-\sum_{j=1}^{n} \mathbf{x}_j^2/2\}$ and $\log p(\mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{n} \mathbf{x}_j^2$ the following holds.

$$\begin{aligned} H(p) &= -\mathbb{E}[\log p(\mathbf{x})] = n/2 \log(2\pi) + n/2 \\ &= n/2 \log(2\pi e), \end{aligned}$$

where e is the Euler number.

Definition 2.3. The relative entropy or the Kullback-Leibler divergence between two PDFs is defined by

$$D_{KL}(p\|q) = \mathop{\mathbb{E}}_{p(\mathbf{x})} \log \frac{p(\mathbf{x})}{q(\mathbf{x})}. \tag{2.2}$$

2.1.1 Maximum entropy prior distributions

By maximizing the differential entropy of a distribution subject to constraints imposed by prior knowledge, the probability distribution which encompasses the least assumptions about the data can be obtained. In the following, the maximum entropy distribution subject to some constraints expressed by equality constraints on expectation of some functions will be derived.

Example 2.4

Maximize the entropy $H(p)$ over all probability densities $p(\cdot)$ satisfying

1. $p(\mathbf{x}) \geq 0$, with equality outside the support set S,
2. $\int_S p(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1$,
3. $\int_S p(\mathbf{x}) T_i(\mathbf{x})\, \mathrm{d}\mathbf{x} = \alpha_i$, for $1 \leq i \leq m$.

The solution to the maximum entropy problem can be found using calculus (Cover and Thomas, 2012); The Lagrangian for the problem is given by

$$J(p) = -\int p \log p + \lambda_0 \int p + \sum_{i=1}^{m} \lambda_i \int T_i p. \tag{2.3}$$

Since the entropy is a concave function defined over a convex set we can compute the functional derivative and equate it to zero to obtain

$$\frac{\partial J}{\partial p(\mathbf{x})} = -\log p(\mathbf{x}) - 1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i T_i(\mathbf{x}) = 0. \tag{2.4}$$

Hence,

$$p(\mathbf{x}) = \exp\left(-1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i T_i(\mathbf{x})\right). \tag{2.5}$$

The result of the example above will be proven using the information inequality in the following theorem.

Theorem 2.5. *Let* $p^*(\mathbf{x}) = \exp\left(-1 + \lambda_0 + \sum_{i=1}^m \lambda_i T_i(\mathbf{x})\right)$, $\mathbf{x} \in S$, *where* $\lambda_0, \lambda_1, ..., \lambda_m$ *are chosen so that* p^* *satisfies (Cover and Thomas, 2012, Theorem 12.1.1)*

1. $p(\mathbf{x}) \geq 0$, *with equality outside the support set* S,
2. $\int_S p(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1$,
3. $\int_S p(\mathbf{x}) T_i(\mathbf{x})\, \mathrm{d}\mathbf{x} = \alpha_i$, *for* $1 \leq i \leq m$.

Then, p^* *uniquely maximizes* $H(p)$ *over all probability densities* p *satisfying the constraints.*

Proof: Proof is obtained using the information inequality. Let g satisfy the constraints. Then

$$\begin{aligned} H(g) &= -\int_S g \ln g = -\int_S g \ln \frac{g}{p^*} p^* = -D_{KL}(g\|p^*) - \int_S g \ln p^* \\ &\leq -\int_S g \ln p^* = -\int_S g\left(-1 + \lambda_0 + \sum_{i=1}^m \lambda_i T_i\right) \\ &= -\int_S p^*\left(-1 + \lambda_0 + \sum_{i=1}^m \lambda_i T_i\right) = -\int_S p^* \ln p^* = H(p^*) \end{aligned}$$

Note that the equality holds iff $D_{KL}(g\|p^*) = 0$ for all $\mathbf{x}$. Therefore, $g = p^*$ except for a set of measure 0. □

Example 2.6

The maximum entropy distribution on the support $S = (-\infty, \infty)$ satisfying the constraint $\mathbb{E}[\mathbf{x}] = \mu$, $\mathbb{E}[\mathbf{x}^2] = \sigma^2$ is $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu, \sigma^2)$.

Example 2.7

The maximum entropy distribution on the support $S = [0, +\infty)$ satisfying the constraint $\mathbb{E}[\mathbf{x}] = \lambda$, is $p(\mathbf{x}) = \mathrm{Exp}(\mathbf{x}; \lambda^{-1})$.

Example 2.8

The maximum entropy distribution on the support $S = [a, b]$ satisfying no other constraint than integrability, is $p(\mathbf{x}) = \mathcal{U}(\mathbf{x}; a, b)$.

The differential entropy for continuous random variables has some weaknesses compared to the discrete random variables which are listed here.

Remark 2.9. Differential entropy differs from the entropy of finely quantized version of the continuous random variable (the Shannon entropy) by the logarithm of the quantization resolution which is infinite in the limit. (Cover and Thomas, 2012, Theorem 8.3.1).

Remark 2.10. Differential entropy is not scale invariant on $\mathbb{R}^n$. That is, for a vector-valued random variable $X \in \mathbb{R}^n$ and a non-singular matrix $A \in \mathbb{R}^{n\times n}$ (Cover and Thomas, 2012, page 254)

$$H(AX) = H(X) + \log|\det(A)|. \tag{2.6}$$

Remark 2.11. Differential entropy can be negative. Hence, the well known relation between the information content of a distribution and the Shannon entropy does not hold for the differential entropy.

As we showed in theorem 2.5, the maximum entropy distribution subject to expectation constraints given in the theorem obtains the form $\exp\left(-1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i T_i(\mathbf{x})\right)$. In the following section the exponential family of distributions will be introduced. The members of this family arise naturally as the solution to the problem of finding maximum entropy distribution subject to the expectation constraint on their sufficient statistic $T(\mathbf{x})$.

2.2 Exponential Family

The exponential family of distributions (Wainwright and Jordan, 2008) includes many common distributions such as Gaussian, beta, Dirichlet, gamma and Wishart. The exponential family in its natural form can be represented by its natural parameters η, sufficient statistic $T(\mathbf{x})$, Log-partition function $A(\eta)$ and base measure $h(\mathbf{x})$ as in

$$q(\mathbf{x};\eta) = h(\mathbf{x})\exp(\eta \cdot T(\mathbf{x}) - A(\eta)), \tag{2.7}$$

where the natural parameter η belongs to the natural parameter space $\Omega = \{\eta \in \mathbb{R}^m | A(\eta) < +\infty\}$. Here $a \cdot b$ denotes the inner product of a and b. In Table 2.1 the sufficient statistic for some continuous members of the exponential family are given.

Definition 2.12. The set corresponding to all mean values for the sufficient statistics

$$\mathcal{M} = \{\mu \in \mathbb{R}^m | \exists p, \mathop{\mathbb{E}}_{p}[T(\mathbf{x})] = \mu\} \tag{2.8}$$

is called the mean parameter space (Wainwright and Jordan, 2008).

Definition 2.13. In a regular family of exponential family the domain Ω is an open set (Wainwright and Jordan, 2008).

Definition 2.14. In minimal representation of an exponential family a unique parameter vector is associated with each distribution (Wainwright and Jordan, 2008).

Table 2.1: *Some continuous exponential family distributions and their sufficient statistic are listed.*

Continuous Exp. Family Distribution	$T(\cdot)$
Exponential distribution	$\mathbf{x}$
Normal distribution with known variance σ^2	$\mathbf{x}/\sigma$
Normal distribution	$(\mathbf{x}, \mathbf{x}\mathbf{x}^{\mathrm{T}})$
Pareto distribution with known minimum x_m	$\log \mathbf{x}$
Weibull distribution with known shape k	$\mathbf{x}^k$
Chi-squared distribution	$\log \mathbf{x}$
Dirichlet distribution	$(\log \mathbf{x}_1, \cdots, \log \mathbf{x}_n)$
Laplace distribution with known mean μ	$\lvert\mathbf{x} - \mu\rvert$
Inverse Gaussian distribution	$(\mathbf{x}, 1/\mathbf{x})$
Scaled inverse Chi-squared distribution	$(\log \mathbf{x}, 1/\mathbf{x})$
Beta distribution	$(\log \mathbf{x}, \log(1-\mathbf{x}))$
Lognormal distribution	$(\log \mathbf{x}, (\log \mathbf{x})^2)$
Gamma distribution	$(\log \mathbf{x}, \mathbf{x})$
Inverse gamma distribution	$(\log \mathbf{x}, 1/\mathbf{x})$
Gaussian Gamma distribution	$(\log \tau, \tau, \tau\mathbf{x}, \tau\mathbf{x}^2)$
Wishart distribution	$(\log \lvert X\rvert, X)$
Inverse Wishart distribution	$(\log \lvert X\rvert, X^{-1})$

The formulas for representation of some probability distribution functions in the exponential family form are given in Appendix A.

2.3 Variational Bayes

Variational Bayes (VB) method is used to find an approximate solution to inference problems when an exact solution is not analytically tractable. Consider a Bayesian model in which prior distributions are assigned to all parameters and latent variables. We will denote all these parameters and latent variables by $\mathbf{x}$ where $\mathbf{x} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$. Now, consider the measurement vector $\mathbf{y}$ along with the joint posterior distribution $p(\mathbf{x}, \mathbf{y})$.

When there is no analytical solution for the posterior $p(\mathbf{x}|\mathbf{y})$ we can look for an approximate analytical solution using the following factorized variational approximation.

$$p(\mathbf{x}|\mathbf{y}) \approx q(\mathbf{x}) \tag{2.9}$$

$$\triangleq q_1(\mathbf{x}_1) q_2(\mathbf{x}_2) \cdots q_n(\mathbf{x}_n), \tag{2.10}$$

where the densities $q_1(\mathbf{x}_1)$, $q_2(\mathbf{x}_2)$, $\cdots$, $q_n(\mathbf{x}_n)$ are the approximate posterior densities for $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, respectively. Technique of VB (Bishop, 2006, Ch. 10),(Tzikas et al., 2008) chooses the estimates $\hat{q}_1(\mathbf{x}_1)$, $\hat{q}_2(\mathbf{x}_2)$, $\cdots$, $\hat{q}_n(\mathbf{x}_n)$ for the factors in (2.10)

using the following optimization problem

$$\hat{q}(\mathbf{x}) = \underset{q(\mathbf{x})}{\operatorname{argmin}} D_{KL}(q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{y})). \tag{2.11}$$

The optimal solution for the optimization problem satisfies the following set of equations.

$$\log \hat{q}_i(\mathbf{x}_i) = \underset{-i}{\mathbb{E}}[\log p(\mathbf{x}, \mathbf{y})] + \text{const.}, \ 1 \leq i \leq n \tag{2.12}$$

where the term const. is constant with respect to the variables $\mathbf{x}_i$ and the subscript $-i$ under the expectation operator means that the expectation is taken with respect to factors other than $q_i(\mathbf{x}_i)$.

The solution to (2.12) can be obtained via fixed-point iterations where only one factor in (2.10) is updated and all the other factors are fixed to their last estimated values (Bishop, 2006, Ch. 10). The iterations converge to a local optima of (2.11) (Bishop, 2006, Ch. 10), (Wainwright and Jordan, 2008, Ch. 3).

The posterior in (2.9) can be the smoothing distribution of the states and model parameters which my not be analytical. In Papers B and C, it is shown that the VB technique can be used to find an approximate posterior.

3

System Identification

This chapter concerns a maximum entropy prior for a specific approximate Bayesian inference problem. Particularly, the prior information about the impulse response of a linear time-invariant (LTI) stable and causal system will be described. The background material presented in this chapter lays the foundation for the contribution in Paper A in the second part of this thesis where the prior information is approximated to construct a maximum entropy kernel for Gaussian process regression. Parts of the back ground material is published in (Ardeshiri and Chen, 2015).

3.1 Impulse Response Identification

System identification is about how to construct mathematical models based on observed data, see e.g., (Ljung, 1999). For linear time-invariant (LTI) and causal systems, the identification problem can be stated as follows. Consider

$$y(t_i) = f * u(t_i) + v(t_i), \quad i = 0, 1, \cdots, N \tag{3.1}$$

where $t_i, i = 0, 1, \cdots, N$ are the time instants at which the measured input $u(t)$ and output $y(t)$ are collected, $v(t)$ is the disturbance, $f(t)$ is the impulse response with $t \in \mathbb{R}^+ \triangleq [0, \infty)$ for continuous-time systems and $t = t_i,\ i = 0, 1, \cdots$ for discrete-time systems, and $f * u(t_i)$ is the convolution of $f(\cdot)$ and $u(\cdot)$ evaluated at $t = t_i$. The goal is to estimate $f(t)$ as good as possible.

Recently, there have been increasing interests in system identification community to study system identification problems with machine learning methods, see e.g., (Ljung et al., 2011), (Pillonetto et al., 2014). An emerging trend among others is to apply Gaussian process regression methods for LTI stable and causal system identification problems, see (Pillonetto and Nicolao, 2010) and its follow up papers (Pillonetto et al., 2011), (Chen et al., 2012a), (Chen et al., 2014). Its

idea is to model the impulse response $f(t)$ with a suitably defined Gaussian process which is characterized by

$$f(t) \sim \mathrm{GP}(m(t), k(t,s)), \tag{3.2}$$

where $m(t)$ is the mean function and is often set to be zero, and $k(t,s)$ is the covariance function, also called the kernel function in machine learning and statistics, see e.g., (Rasmussen and Williams, 2006).

The kernel $k(t,s)$ is parametrized by a hyper-parameter β and further written as $k(t,s;\beta)$. The key issue is to design a suitable parametrization of $k(t,s;\beta)$, or in other words, the structure of $k(t,s;\beta)$, because it reflects our prior knowledge about the system to be identified. Several kernel structures have been proposed in the literature, e.g., the stable spline (SS) kernel in (Pillonetto and Nicolao, 2010) and the diagonal and correlated (DC) kernel in (Chen et al., 2012a).

Interestingly, (Pillonetto and Nicolao, 2011) shows based on a result in (Nicolao et al., 1998) that for continuous-time systems, the continuous-time first-order SS kernel (also derived by deterministic arguments in (Chen et al., 2012a) and called Tuned Correlated (TC) kernel):

$$k(t,s) = \min\{e^{-\beta t}, e^{-\beta s}\}, \quad t,s \in \mathbb{R}^+ \tag{3.3}$$

has a certain maximum entropy property. In Example 3.1 an the impulse response identification problem will be further illustrated.

Example 3.1

Consider the LTI system and the simulated input-output data presented in Figure 3.1.

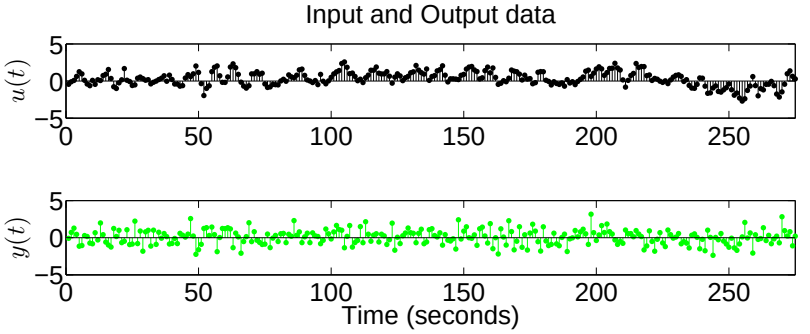


Figure 3.1: *Input (black) and output (green) data versus time.*

The impulse response f can be computed using the input-output data and, the prior knowledge about the form of the impulse response expressed by the kernel function,

$$k(t,s) = \min\{e^{-\beta t}, e^{-\beta s}\}, \quad t,s = t_i, i = 0,1,\cdots \tag{3.4}$$

where $f(t) \sim \mathrm{GP}(0, k(t,s))$. The estimated impulse response is presented in Figure 3.2.

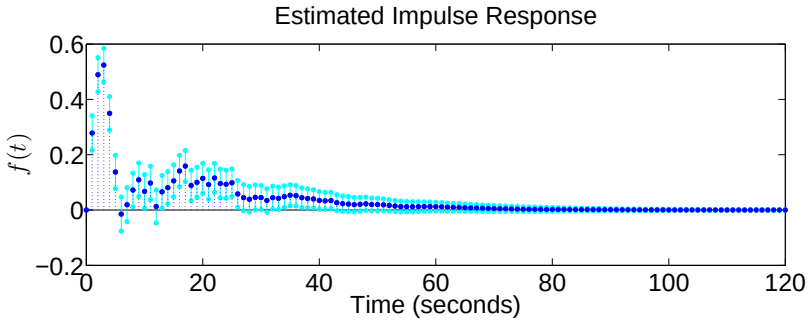


Figure 3.2: *The estimated impulse response (dark blue) along with one standard deviation band (cyan).*

In the following some characteristics of the impulse response of LTI stable and causal systems will be given in two separate sections one for the continuous-time case and another for the discrete-time case.

3.2 Continuous-time impulse response

The prior knowledge about the continuous-time impulse response of a stable and causal LTI system are

1. bounded input bounded output (BIBO) stability of the system and,
2. smoothness of the impulse response.

For continuous-time impulse response the BIBO stability is assured when the impulse response be absolutely integrable, i.e., its L^1-norm exists;

$$\int_{-\infty}^{\infty} |f(t)| \, \mathrm{d}t = \|f\|_1 < \infty. \tag{3.5}$$

The smoothness constraint on the *continuous-time* impulse responses can be addressed as in (Nicolao et al., 1998, Theorem 1) where the authors suggest that the smoothness of a signal can be imposed by assuming that the variances of its derivatives are finite.

$$\mathbb{V}\left[\frac{\mathrm{d}f}{\mathrm{d}t}\right] = \lambda, \quad \lambda < \infty \tag{3.6}$$

The impulse response of a continuous-time LTI system and its L^1-norm is illustrated in Figure 3.3.

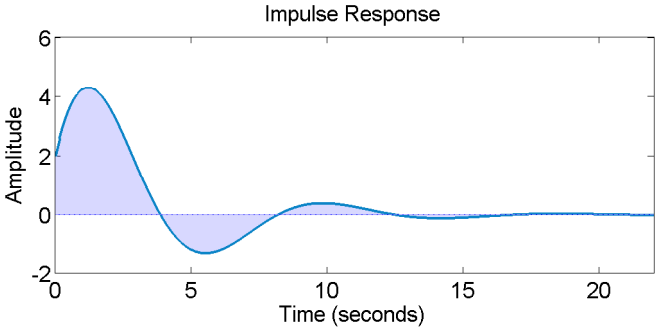


Figure 3.3: *The impulse response of a continuous-time LTI stable system. The shaded area under the impulse response should be finite.*

Some definitions which would be needed to solve the maximum entropy kernel estimation problem for continuous-time stochastic processes are given in the following.

Definition 3.2. ($\mathbb{L}_2$ differentiation) (Åström, 1970, page 37) A second order stochastic process f is said to be differentiable in the mean square at t if the limit

$$\lim_{s\to 0} \frac{f(t+s)-f(t)}{s} = f'(t) \tag{3.7}$$

exists in the sense of mean square convergence, that is, if

$$\lim_{s\to 0} \mathbb{E}\left[\frac{f(t+s)-f(t)}{s} - f'(t)\right]^2 = 0. \tag{3.8}$$

Recall that the derivative variances can be expressed via spectral measure by

$$\mathbb{E}\left[f^{(m)}(t)^2\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \omega^{2m} S(\omega)\,\mathrm{d}\omega \tag{3.9}$$

where, mth square mean derivative of f exists iff the integral in the right hand side of (3.9) is finite (Lifshits, 2014, page 107).

Definition 3.3. The differential entropy rate of a real-valued *continuous-time* stochastic process $f(\cdot)$ is defined in (Nicolao et al., 1998) as

$$\overline{H}(f) = \frac{1}{4\pi}\int_{-\infty}^{\infty} \log\Big(S(\omega)\Big)\,\mathrm{d}\omega. \tag{3.10}$$

3.3 Discrete-time impulse response

The prior knowledge about the discrete-time impulse response of a stable and causal LTI system are

1. bounded input bounded output (BIBO) stability of the system and,
2. smoothness of the impulse response.

BIBO stability is assured when the impulse response is absolutely summable, i.e., its ℓ^1 norm exists;

$$\sum_{n=-\infty}^{\infty} |f(n)| = \|f\|_1 < \infty. \tag{3.11}$$

The smoothness constraint on the discrete-time impulse responses can be imposed by assuming that the variances of its finite differences are proportional to the time increment over which the finite difference is computed;

$$\mathbb{V}[f(t_{i+1}) - f(t_i)] = \lambda(t_{i+1} - t_i), \quad \infty > \lambda > 0. \tag{3.12}$$

Some definitions which would be needed to solve the maximum entropy kernel estimation problem for discrete-time stochastic processes are given in the following.

Definition 3.4. (Differential entropy rate of a sequence) Let $\{X(n)\}$ be a sequence. Its differential entropy rate is defined as (Cover and Thomas, 2012)

$$\overline{H}(X) \triangleq \lim_{n\to\infty} \frac{1}{n} H(p(X(1), ..., X(n))), \tag{3.13}$$

when the limit exists.

Note that stationarity or even wide-sense stationarity are not required for definition 3.4 to hold.

Proposition 3.5. *Among all sequences with given covariance, Gaussian one has the maximal differential entropy rate (Cover and Thomas, 2012).*

Example 3.6

For independent and identically distributed (iid) standard Gaussian sequence we have

$$\overline{H}(X) = \log \sqrt{2\pi e}. \tag{3.14}$$

Let $X(\cdot)$ be a centered (zero-mean) discrete-time stationary sequence with autocorrelation $R(\cdot)$ where, $R(n) = \mathbb{E}[X(n)X(0)]$. Also, let $S(\cdot)$ denote the spectral density of $X(\cdot)$ on $[-\pi, \pi]$. The following hold for a spectral representation $S(\cdot)$ (Papoulis and Pillai, 2002, page 421),

$$R(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{in\omega} S(\omega)\, \mathrm{d}\omega. \tag{3.15}$$

Example 3.7

$X(\cdot)$ is the iid standard sequence with covariance function

$$R(n) = \begin{cases} 1 & \text{n=0} \\ 0 & \text{otherwise} \end{cases} \tag{3.16}$$

iff $S(\omega) = 1$ for $\omega \in [-\pi, \pi]$.

Theorem 3.8. *If a stationary sequence is Gaussian with spectral density $S(\omega)$, then (Papoulis and Pillai, 2002, page 663)*

$$\overline{H}(X) = \log \sqrt{2\pi e} + \frac{1}{4\pi} \int_{-\pi}^{\pi} \log S(\omega)\, \mathrm{d}\omega. \tag{3.17}$$

Example 3.9

Here, we will verify this theorem for iid Gaussian sequence with variance σ^2. From (3.14) we obtain

$$\overline{H}(X) = \log \sqrt{2\pi e} + \log \sigma. \tag{3.18}$$

From (3.17) we can obtain the same result as in

$$\begin{aligned} \overline{H}(X) &= \log \sqrt{2\pi e} + \frac{1}{4\pi} \int_{-\pi}^{\pi} \log \sigma^2 \, \mathrm{d}u \\ &= \log \sqrt{2\pi e} + \log \sigma. \end{aligned} \tag{3.19}$$

3.4 Maximum Entropy Kernel

In (Pillonetto and Nicolao, 2011), the maximum differential entropy rate *continuous-time* stochastic process subject to constraints on smoothness and bounded-input bounded-output (BIBO) stability is sought. In (Pillonetto and Nicolao, 2011) the Definition 3.3 for the differential entropy rate of a stationary *continuous-time Gaussian* process $g(t)$ with power spectrum $S(\omega)$ is adopted from (Nicolao et al., 1998). Furthermore, the following proposition is adopted from (Nicolao et al., 1998).

Proposition 3.10. *(Nicolao et al., 1998, Theroem 1) Let $g(t)$ be a zero-mean bandlimited stationary Gaussian process with power spectrum $S(\omega) = 0$ for $|\omega| > B$.*

Given finite λ_k^2, $k = 0, 1, \cdots, m$, *assume that there exist real numbers* α_j, $j = 0, 1, \cdots, m$ *such that*

$$\int_{-B}^{B} \frac{\omega^{2k}}{\sum_{j=0}^{m} \alpha_j \omega^{2j}} \, \mathrm{d}\omega = 2\pi\lambda_k^2, \qquad k = 0, 1, \cdots, m. \tag{3.20}$$

Under this assumption, if there exists $S(\omega)$ *that maximizes* $\overline{H}(g)$ *in*

$$\overline{H}(g) = \frac{1}{4\pi} \int_{-\infty}^{+\infty} \log\Big(S(\omega)\Big) \, \mathrm{d}\omega. \tag{3.21}$$

subject to constraints $\mathbb{V}[\frac{\mathrm{d}^k g(t)}{\mathrm{d}t^k}] = \lambda_k^2$, $k = 0, 1, \cdots, m$, *then the spectrum is given by* $S(\omega) = \frac{1}{\sum_{j=0}^{m} \alpha_j \omega^{2j}}$. *In particular, if there is no constraints on the first* $m - 1$ *order derivatives, then the spectrum becomes* $S(\omega) = \frac{1}{\alpha_m \omega^{2m}}$.

Deriving the maximum entropy process in continuous-time in (Nicolao et al., 1998) and (Pillonetto and Nicolao, 2011) is quite involved, due to the infinite-dimensional nature of of the problem and absence of a well-defined differential entropy rate for a generic continuous-time stochastic process.

In Paper A, we focus on discrete-time impulse responses (stochastic processes), and provide a simple and self-contained proof to show the maximum entropy properties of the discrete-time first-order SS kernel (3.4). The advantages of working in discrete-time domain include

1. The differential entropy rate is well-defined for discrete-time stochastic process.

2. Given a stochastic process, its finite difference process can be well-defined in discrete-time domain.

3. It is possible to show what maximum entropy property a zero-mean discrete-time Gaussian process with covariance function (3.4) has.

4

Mixture Reduction

The background material presented in this chapter introduces the mixture reduction problem and presents the background material which are related to papers E and F. Some of the material presented in this chapter are published by the PhD candidate in (Ardeshiri et al., 2012) and (Ardeshiri et al., 2014).

4.1 Mixture Reduction

A common problem encountered in Bayesian inference and particularly tracking is mixture reduction (MR). Examples of such circumstances are multi-hypotheses tracking (MHT)(Blackman and Popoli, 1999), Gaussian sum filter(Alspach and Sorenson, 1972), multiple model filtering (Blackman and Popoli, 1999), Gaussian mixture probability hypothesis density (GM-PHD) filter (Vo and Ma, 2006). In these algorithms the information about the state of a random variable is modeled as a mixture density.

A *mixture density* is a probability density which is a convex combination of (more basic) component probability densities, see e.g. (Bishop, 2006). A normalized mixture with N components is defined as

$$p(x) = \sum_{I=1}^{N} w^I q(x; \eta^I), \tag{4.1}$$

where the terms w^I are positive weights summing to unity, and η^I are the parameters of the component density $q(x; \eta^I)$. When the component density is a Gaussian density the mixture density is referred to as Gaussian mixture (GM). The *mixture reduction problem* (MRP) is to find an approximation of the original mixture density by a mixture density with fewer components. To be able to implement these algorithms for real time applications a mixture reduction

step is necessary. The aim of the reduction algorithm is to reduce the computational complexity into a predefined budget while keeping the inevitable error introduced by the approximation as small as possible.

4.2 Mixture Reduction for Target Tracking

This section concerns with the mixture reduction algorithms in multiple target tracking. The current mixture reduction convention in multiple target tracking (MTT) is to use exactly the same algorithm for reducing the computational load to a feasible level as for extracting the state estimates. In general, the mixture reduction for the state extraction should be much more aggressive than that for computational feasibility. For this reason, the number of components in the mixtures have to be reduced much more than what the computational resources actually allow for. This can result in coarser approximations than what is actually necessary. It is proposed in (Ardeshiri et al., 2012) to split the reduction step into two separate procedures according to:

- *Reduction in the loop* is a reduction step which must be performed at each point in time for computational feasibility of the overall target tracking framework. The objective for this algorithm is to reduce the number of components and to minimize the information loss.
- *Reduction for state extraction* aims at reducing the number of components so that the remaining components can be considered as state estimates in the target tracking framework.

This separation makes it possible to tailor these two algorithms to fulfill their individual objectives, which reduces the unnecessary approximations in the overall algorithm. A block diagram of the conventional mixture reduction method on a high level is shown in Figure 4.1.

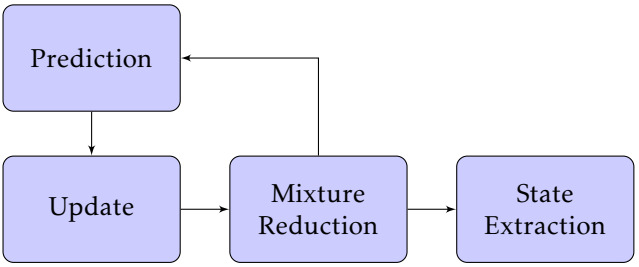


Figure 4.1: *The standard flowchart of the MTT algorithms has only one mixture reduction block.*

In the proposed implementation of MR for MTT in (Ardeshiri et al., 2012), the reduction algorithm is split into two subroutines each of which is tailored for its own purpose, see Figure 4.2. The first reduction algorithm, denoted *reduction*

in the loop, is designed to reduce the computational cost of the algorithm to the computational budget between the updates. In this reduction step, the number of components should be reduced to a number that is tractable by the available computational budget and minimal loss of information is in focus. The second reduction algorithm, denoted *reduction for extraction*, is designed to reduce the mixture to as many components as the number of targets. In this part of the algorithm, application dependent specifications and heuristics can enter into the picture. If the purpose of state extraction is only visualization, the second reduction does not have to be performed at the same frequency as the measurements are received and can be made less frequent. The advantages of the proposed algorithm are that the unnecessary loss of information in the reduction in the loop step will only be due to the finite computational budget rather than closeness of the components. Furthermore, some computational cost can be discounted if the state extraction does not have to be performed for every measurement update step.

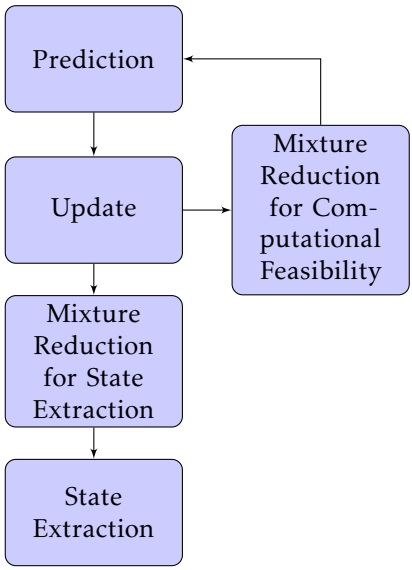


Figure 4.2: *The proposed block diagram of the MTT algorithm with two mixture reduction blocks; one tailored to keep the computational complexity within the computational budget and one tailored for state extraction.*

Another important advantage of the proposed algorithm in (Ardeshiri et al., 2012) is that the number of final components in both of the reduction algorithms is known since the computational budget is predefined in the reduction in the loop algorithm. Furthermore, the number of target states can be predetermined by summarizing the weights in e.g., a GM-PHD filter and utilized in the reduction for extraction algorithm. The clustering or optimization method selected for reduction can be executed more efficiently compared to a scenario where the

number of components is left to be decided by the algorithm itself.

4.3 Greedy mixture reduction

Ideally, the MRP is formulated as a nonlinear optimization problem where a divergence measure between a mixture and its approximation with a desired number of components is selected. The optimization problem is then solved by numerical solvers when the problem is not analytically tractable. The numerical optimization based approaches can be computationally quite expensive, especially for high dimensional data and they generally suffer from the problem of local optima. Hence, a common alternative solution to the MRP has been the greedy iterative approach. When the computational budget permits a numerical solution, the greedy approaches are used to initialize the global optimization approach (Williams and Maybeck, 2006).

In the *greedy* approach, the number of components in the mixture is reduced one at a time. By applying the same procedure over and over, a desired number of components can be reached. In order to reduce the number of components by one, two types of operations are considered, namely, pruning one component and merging of two components. These two operations will be given an official definition in the following.

Pruning which is the simplest operation for reducing the number of components in a mixture density is to remove one (or more) components of a mixture and rescaling the remaining components such that it integrates to unity. For example pruning component J from (4.1) results in the mixture density

$$p'(x) = (1 - w^J)^{-1} \sum_{I=1, I\neq J}^{N} w^I q(x; \eta^I). \tag{4.2}$$

The *merging* operation in a MRA approximates a subset of components in a mixture density with a single component of the same component density type. In general, an optimization problem minimizing the KLD between the normalized subset of the mixture and the single component is used for this purpose leading to a moment matching operation. More formally, approximation of a fraction of the mixture density (4.1) consisting of two components I and J; $w^I q(x; \eta^I) + w^J q(x; \eta^J)$ by a single weighted component $(w^I + w^J) q(x; \eta^{IJ})$ is referred to as merging components I and J, where

$$\eta^{IJ} = \arg\min D_{KL}\left(\frac{w^I q(x; \eta^I) + w^J q(x; \eta^J)}{w^I + w^J} \,\|\, q(x; \eta^{IJ}) \right).$$

When the component densities are Gaussian densities with mean μ and covari-

ance Σ the parameters of the approximate density are given by

$$\mu^{IJ} = \frac{1}{w^I + w^J}(w^I \mu^I + w^J \mu^J), \tag{4.3a}$$

$$\Sigma^{IJ} = \sum_{K \in \{I,J\}} \frac{w^K}{w^I + w^J} \left(\Sigma^K + (\mu^K - \mu^{IJ})(\mu^K - \mu^{IJ})^{\mathrm{T}}\right). \tag{4.3b}$$

There are two different types of greedy approaches in the literature, local and global approaches. The local approaches consider only the merging operation. The (two) components to be merged are selected among all possible pairs of components based on a divergence measure between the individual components and the divergence between the original mixture and its approximation is not (explicitly) taken into account. Well-known examples of local approaches are given in (Salmond, 1990; Granström and Orguner, 2012b).

In the global approach, each of the pruning or merging possibilities are considered to be a hypothesis. The decisions are then made by choosing the candidate hypothesis that minimizes a divergence measure involving the original mixture and the corresponding reduced mixtures (all of which has one less component).

In the global approach to mixture reduction, pruning or merging operations applicable on the original mixture $p(x)$ are considered to be hypotheses denoted by $\mathcal{H}$. The resulting mixtures that would be obtained if the I^{th} component is pruned, or if the I^{th} and J^{th} components are merged, are denoted by $p(x|\mathcal{H}_{0I})$ and $p(x|\mathcal{H}_{IJ})$, respectively. The single component obtained by merging the I^{th} and J^{th} components is denoted by $q(x, \eta^{IJ})$. If $p(x)$ has K components there are K pruning and $K(K-1)/2$ merging hypotheses. In order to decide on the candidate pruning and merging operations, all corresponding mixtures $p(x|\mathcal{H}_{0I})$, $p(x|\mathcal{H}_{IJ})$ and their associated divergence measure are calculated. The hypothesis which results in the smallest divergence measure, is most similar mixture to the original mixture, is selected.

More particularly, at the k^{th} stage of reducing mixture density of equation (4.1), $n_k = N - k + 1$ components are left and there are $\frac{1}{2} n_k \times (n_k - 1)$ possible merging decisions and n_k possible pruning decisions to choose from. Let the reduced density at the k^{th} stage be denoted by $p_k(x)$. We have a multiple hypotheses decision problem at hand where the hypotheses are formulated according to

$$\text{Pruning Hypotheses} \begin{cases} \mathcal{H}_{01} : x \sim p_k(x|\mathcal{H}_{01}), \\ \mathcal{H}_{02} : x \sim p_k(x|\mathcal{H}_{02}), \\ \qquad\qquad \vdots \\ \mathcal{H}_{0n_k} : x \sim p_k(x|\mathcal{H}_{0n_k}), \end{cases}$$

$$\text{Merging Hypotheses} \begin{cases} \mathcal{H}_{12} : x \sim p_k(x|\mathcal{H}_{12}), \\ \mathcal{H}_{13} : x \sim p_k(x|\mathcal{H}_{13}), \\ \qquad\qquad \vdots \\ \mathcal{H}_{(n_k-1)n_k} : x \sim p_k(x|\mathcal{H}_{(n_k-1)n_k}), \end{cases}$$

which is a decision problem with $n_k(n_k+1)/2$ hypotheses. The first n_k hypotheses account for pruning and the rest account for merging decisions. The subscript on hypotheses $\mathcal{H}_{\cdot}$ refers to the two components to be merged for merging hypotheses while in the case of pruning hypotheses the subscript refers to the label of the component to be pruned which is preceded by zero.

The divergence measures used for the aforementioned decision problem are presented in the following Section.

4.4 Divergence measures

A divergence measure is a function which establishes the distance of one probability distribution to the other on a statistical manifold (Minka, 2005). A divergence measure is a weaker form of a metric, in particular the divergence does not need to be symmetric and does not need to satisfy the triangle inequality.

4.4.1 Integral square error

ISE is a divergence measure between two densities which is defined as

$$\mathrm{ISE}(p\|q) = \int |p(x) - q(x)|^2 \, \mathrm{d}x \tag{4.4}$$

for two densities $p(x)$ and $q(x)$. ISE has all properties of a metric.

ISE is used by Williams and Maybeck in (Williams and Maybeck, 2006) as a divergence measure for mixture reduction. The cost of the hypothesis $\mathcal{H}_K$ obeys

$$\mathrm{ISE}(\mathcal{H}_K) = \int |p(x) - p_k(x|\mathcal{H}_K)|^2 \, \mathrm{d}x. \tag{4.5}$$

In this approach, the hypothesis which gives the smallest ISE will be chosen at each step of the reduction i.e., the decision rule based on ISE becomes "decide $\mathcal{H}_K$ if $\mathrm{ISE}(\mathcal{H}_K) < \mathrm{ISE}(\mathcal{H}_L)$ for all $L \neq K$", where K and L are permissible indices of the hypotheses.

An attractive property of the ISE as a divergence measure is that the ISE between two Gaussian mixtures has an analytical solution.

4.4.2 Kullback-Leibler Divergence

The global approach to mixture reduction problem can be posed as a multiple hypothesis testing problem [1]. Suppose that we have a mixture $p(x)$ with N components as in (4.1). Suppose we have a number of reduced mixtures $\{p(x|\mathcal{H}_j)\}_{j=1}^{K}$ and we would like to select one of them. Assuming that we have the data $\{x_i\}_{i=1}^{S}$ sampled from $p(\cdot)$, the selection of the best reduced mixture can be posed as

[1] For a short introduction to multiple hypothesis testing and maximum *a posteriori* decision rule see Appendix B.

a multiple hypothesis testing problem where the test statistics become the log-likelihood of the data given as

$$\log p(\{x_i\}_{i=1}^S|\mathcal{H}_j) = \sum_{i=1}^{S} \log p(x_i|\mathcal{H}_j) \tag{4.6}$$

and the decision is made to select $\mathcal{H}_{j^*}$ where

$$j^* \triangleq \arg\max_j \log p(\{x_i\}_{i=1}^S|\mathcal{H}_j). \tag{4.7}$$

When we let the number of the samples S go to ∞, we see that

$$\lim_{S\to\infty} \frac{1}{S} \log p(\{x_i\}_{i=1}^S|\mathcal{H}_j) = \mathop{\mathbb{E}}_{p(\cdot)}\Big[\log p(x|\mathcal{H}_j)\Big] \tag{4.8}$$

by the law of large numbers. Kullback-Leibler divergence $D_{KL}(p(\cdot)\|p(\cdot|\mathcal{H}_j))$ between $p(\cdot)$ and $p(\cdot|\mathcal{H}_j)$ is given as

$$D_{KL}(p(\cdot)\|p(\cdot|\mathcal{H}_j)) = -H(p(x)) - \mathop{\mathbb{E}}_{p(\cdot)}\Big[\log p(x|\mathcal{H}_j)\Big] \tag{4.9}$$

where $H(\cdot)$ is the entropy of its argument density. Therefore, the optimization (4.7) is equivalently given as

$$j^* \triangleq \arg\min_j D_{KL}(p(x)\|p(x|\mathcal{H}_j)). \tag{4.10}$$

The cost function in (4.10) can not be analytically evaluated when one of the arguments in the KLD is a Gaussian mixture. Runnalls in (Runnalls, 2007) used a nice analytical approximation of the KLD between two mixtures which can only be used for evaluating the merging hypotheses. The approximation is in fact an upper bound on $D_{KL}(p(x)\|p(x|\mathcal{H}_{IJ}))$, which is the cost of merging two components I and J, and is denoted by $\mathcal{B}(I,J)$ and defined by

$$\mathcal{B}(I,J) \triangleq w^I D_{KL}(q(x;\eta^I)\|q(x;\eta^{IJ})) + w^J D_{KL}(q(x;\eta^J)\|q(x;\eta^{IJ})). \tag{4.11}$$

Runnalls has shown that $\mathcal{B}(I,J) \geq D_{KL}(p(x)\|p(x|\mathcal{H}_{IJ}))$ in (Runnalls, 2007). The greedy MR algorithm suggested by (Runnalls, 2007) will be referred to as approximate Kullback-Leibler (AKL) algorithm in the rest of this thesis.

4.4.3 α-Divergences

A generalization of the KLD called the α-divergence is a family of divergences defined over a range of continuous hyper-parameter $\alpha \in (-\infty, \infty)$ by

$$D_\alpha(p\|q) \triangleq \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{(1+\alpha)/2} q(x)^{(1-\alpha)/2}\, \mathrm{d}x\right).$$

Some special cases of of the α-divergence are

$$\lim_{\alpha \to 1} D_\alpha(p\|q) = D_{KL}(p\|q) \tag{4.12a}$$

$$\lim_{\alpha \to -1} D_\alpha(p\|q) = D_{KL}(q\|p) \tag{4.12b}$$

$$D_0(p\|q) = D_H(p\|q) \tag{4.12c}$$

where, $D_H(p\|q)$ is the Hellinger distance (Bishop, 2006).

We will analyze the divergence measures given above in Example 4.1.

Example 4.1

Using an example given in (Minka, 2005) the effect of changing the hyper-parameter α in the divergence measure is illustrated and compared with the ISE distance. Consider the Gaussian mixture $p(x) = 0.6\ \mathcal{N}(x; -2, 1) + 0.4\ \mathcal{N}(x; 2, 0.16)$, and its approximation $q(x)$ which is a Gaussian distribution with unknown mean and standard deviation. In Figure 4.3 the minimizing argument of $D_\alpha(p\|q)$ over q is given for various values of α alongside the minimizing argument of ISE$(p\|r)$. The parameters of q (mean and standard deviation) are given in Figure 4.4. The parameters of q vary smoothly with α except when $-1 \leq \alpha \leq 1$. When $\alpha \ll -1$ the solution is mode seeking (the mode with the largest mass) and when $\alpha \gg 1$ the optimal solution distributes the probability mass over the support and where there is considerable probability mass in original distribution. The observations made in this simple example are general and can be conveniently explained by the definition of the α-divergence. The minimizing argument of ISE$(p\|r)$ in this example does not have the same general interpretation; in this example ISE is rather similar to D_α for $\alpha = 1$, if the two Gaussian densities where far away from each other, ISE becomes more similar to D_α for $\alpha = -1$. Another observation that is made is that the minimizing argument of $D_\alpha(p\|q)$ over q does not vary so much for values of α outside the interval of $[-1, 1]$. Therefore, we will only study the α-divergence in the limit as $\alpha \to 1$ where it corresponds to the KLD and as $\alpha \to -1$ where it corresponds to to the reversed Kullback-Leibler divergence (RKLD). In applications where the mode seeking property of the solution is desired RKLD is suitable. On the other hand, when the solution should preserve the statistical moments of a mixture density KLD is the most appropriate.

In Paper F a Gaussian mixture reduction algorithm using the RKLD is proposed which has the mode seeking property illustrated in Example 4.1.

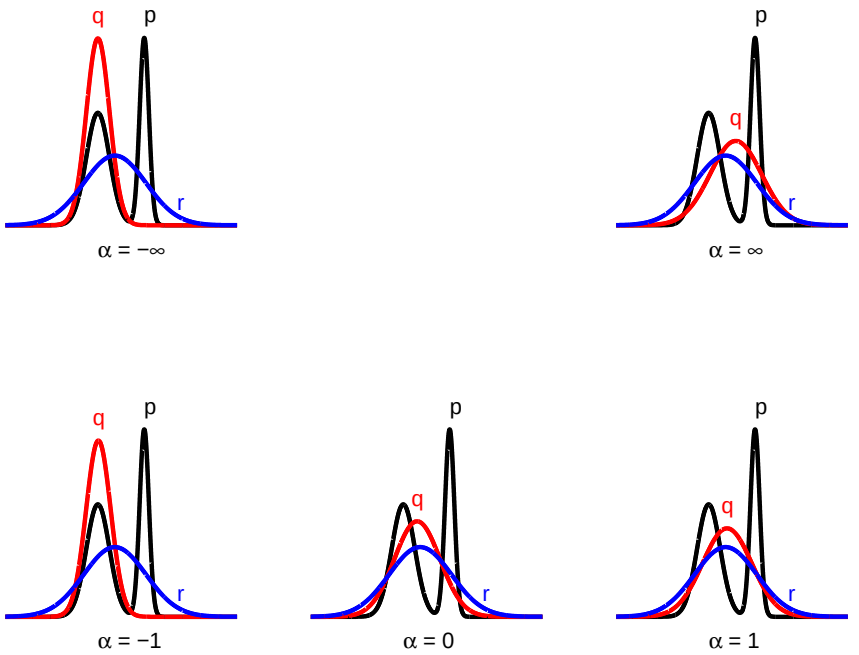


Figure 4.3: *The Gaussian mixture p (black) is approximated by two Gaussian densities q (red) and r (blue). q minimizes the α-divergence for different values of α and r minimizes the ISE.*

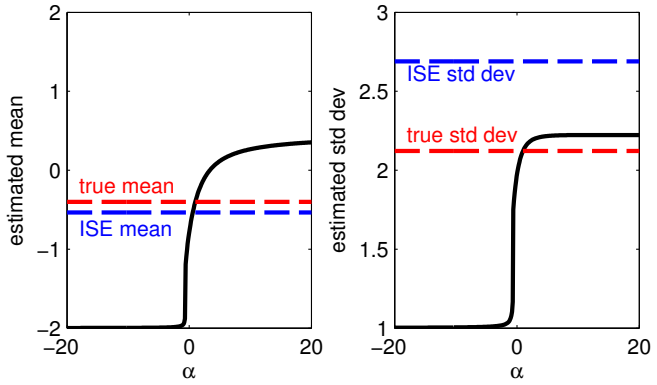


Figure 4.4: *The mean and standard deviation of the Gaussian density q which minimizes the α-divergence to p for different values of α (black). For $\alpha = 1$ the mean and standard deviation of q matches those of p (red). Mean and standard deviation of the Gaussian density r which minimizes the* ISE *to p is given for comparison (blue).*

4.5 Numerical comparison of mixture reduction algorithms

In Paper E three mixture reduction algorithms for mixtures of the exponential family are given and are evaluated in simulations. In these algorithms the ISE approach and AKL approach are compared with a local approach referred to as Symmetrized Kullback-Leibler Divergence. The Symmetrized Kullback-Leibler Divergence is used for the comparison of the merging hypotheses in local algorithms such as (Kitagawa, 1994), (Chen et al., 2012b), (Granström and Orguner, 2012b) and (Granström and Orguner, 2012a). The symmetrized KLD (SKL) for two component densities is defined as

$$D_{SKL}(I, J) = D_{KL}(q_{\eta^I} \| q_{\eta^J}) + D_{KL}(q_{\eta^J} \| q_{\eta^I}). \tag{4.13}$$

This approach is referred to as SKL and is used in the numerical simulation intended for comparison of different MR algorithms in the following.

In this section, eight mixture reduction examples are illustrated. In Figures 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11 and 4.12 mixture densities of exponential, Weibull, Rayleigh, Log-Normal, gamma, inverse gamma and Gaussian distribution are reduced, respectively. In each figure, a mixture density with 25 components along with its reduced approximations with 3 components using three reduction algorithms AKL, SKL and ISE are plotted. In these figures the original mixture density (black solid line) and its components (black dashed line) are given. In the subfigures AKL, SKL and ISE are used to approximate the original mixture which has 25 component densities with mixtures with 3 component densities. The approximate densities (thick dashed lines) and their components (thin dashed line) are drawn in different colors; red(AKL), green(SKL) and blue(ISE). AKL is used in the left sub-figure, SKL is used in the center sub-figure and ISE is used in the right sub-figure. The reduced mixture in the right sub-figure is not rescaled after possible pruning steps and is plotted as it is used in the ISE algorithm. For implementation aspects of the ISE approach see Appendix C.

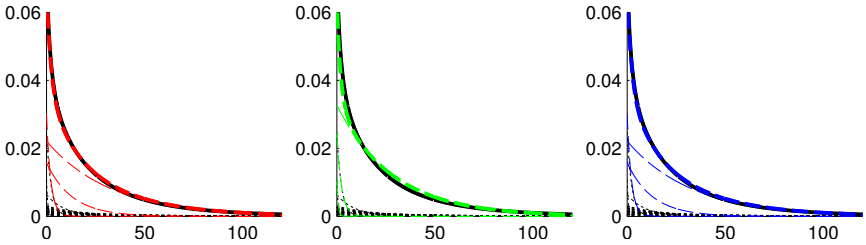


Figure 4.5: *Exponential Distribution*

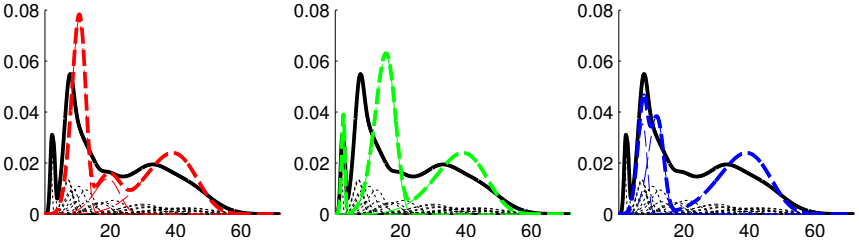


Figure 4.6: *Weibull Distribution with known shape k*

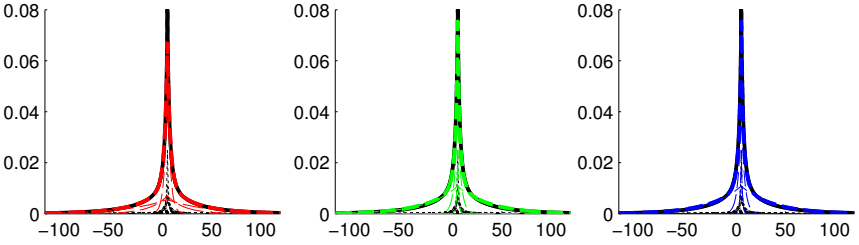


Figure 4.7: *Laplace Distribution with known mean μ*

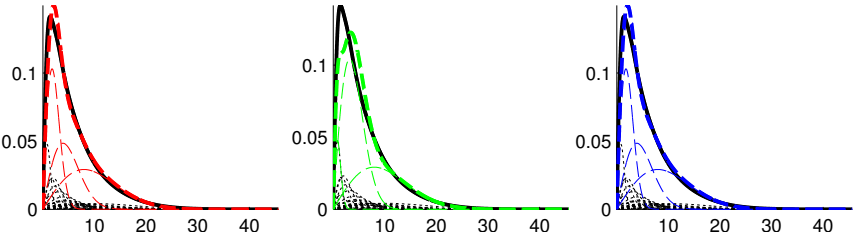


Figure 4.8: *Rayleigh Distribution*

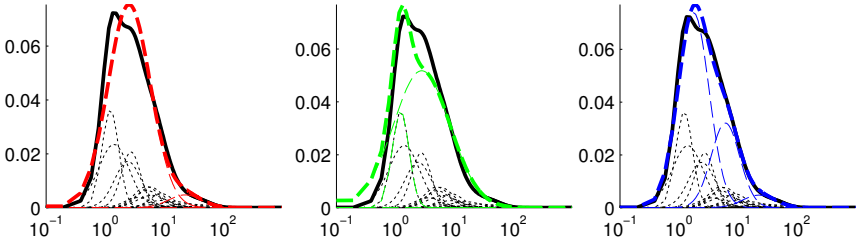


Figure 4.9: *Log-normal Distribution*

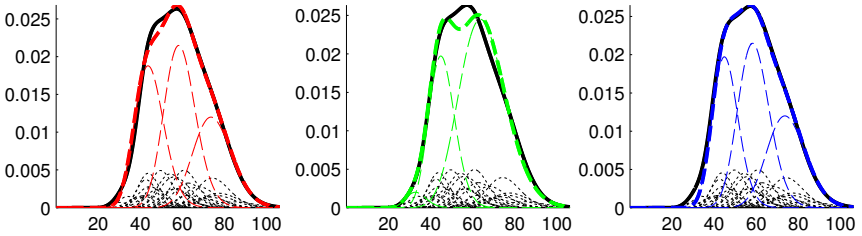


Figure 4.10: *Gamma Distribution*

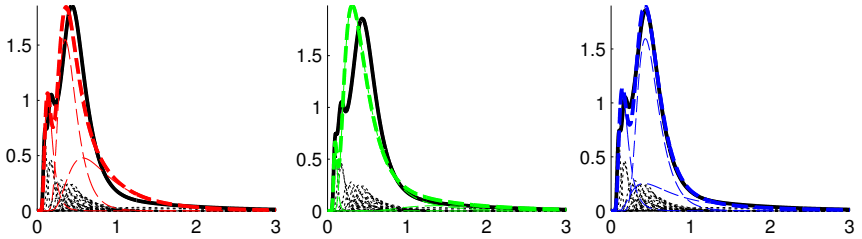


Figure 4.11: *Inverse Gamma Distribution*

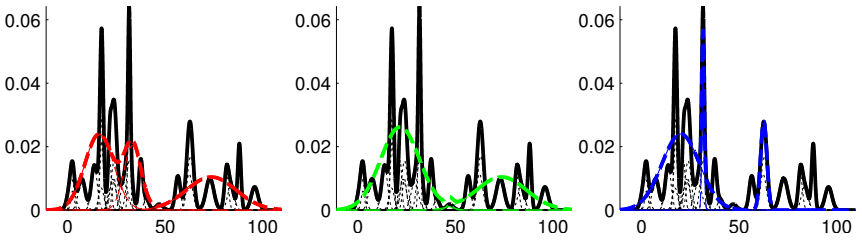


Figure 4.12: *Univariate Gaussian Distribution*

5

Concluding remarks

This chapter concludes the first part of this thesis. An overall summary of the contributions given in the second part of the thesis and some directions for further research will be given here. For more detailed discussion on the each contribution see the discussions and concluding remarks at the end of each contribution.

In paper A, the maximum entropy properties of the first-order stable spline kernel for identification of linear time-invariant stable and causal systems are shown. Analytical approximations are used to express the prior knowledge about the properties of the impulse response of a linear time-invariant stable and causal system. Future work on the subject includes studying maximum entropy interpretation of other kernels used for regression using Gaussian processes. Furthermore, the maximum entropy approach can be used to construct new kernels for system identification.

In papers B, variational Bayes (VB) method is used to compute an approximate posterior for the state smoothing problem for linear state-space models with unknown and time-varying noise covariances. The VB method gives an approximate posterior for the unknown noise covariances. Nevertheless, the Variational Bayes type algorithms approximate the posterior by minimizing Kullback-Leibler divergence in zero forcing mode, meaning that if there are multiple modes in the true posterior, the algorithm approximates only one of the modes. Hence the posterior covariance might underestimate the true covariance significantly in such cases. Computing a better estimate of estimation uncertainly for the noise covariances can be a future work. Theoretical comparison of the proposed VB method with expectation maximization and maximum likelihood estimate of the noise covariances is another possible future work.

In paper C, the VB method is used for approximate inference in state-space models with skewed measurement noise. A filter and a smoother that take into account the skewness and heavy-tailedness of the measurement noise are pro-

posed where skew-t distribution is used to model the distribution of measurement noise. Future research on the subject includes learning the skewness and spread parameters of the measurement noise from the data. Further research on the subject can include studying a class of hierarchical models for modeling the noise parameters and devising algorithms for learning the parameters of such a model from the data.

In paper D, a novel approximation method for Bayesian inference is proposed. The proposed Bayesian inference technique is based on Taylor series approximation of the logarithm of the likelihood function. The proposed approximation is devised for the case where the prior distribution belongs to the exponential family of distributions. The linearization of the log-likelihood is performed with respect to the sufficient statistic of the prior distribution. Extension of the proposed method for prior distributions outside the exponential family of distribution can be a future research direction. The comparison of possible choices for the linearization point and linearization methods with respect to the sufficient statistic are among the future research problems.

In papers E and F, two contributions are dedicated to the mixture reduction (MR) problem. The first contribution, generalizes the existing MR algorithms for Gaussian mixtures to the exponential family of distributions and compares them in an extended target tracking scenario. The second contribution, proposes a new Gaussian mixture reduction algorithm using the reversed Kullback-Leibler divergence which has specific peak preserving properties. Future research on these topics includes evaluation of these methods in real life scenarios with real measurements.

There is a general class of solutions to the Bayesian inference problem referred to by sampling methods which can obtain much better performance compared to proposed approximation methods with respect to accuracy when the computation time is not critical. Sampling methods are not covered in this thesis. That is why the approximations used in this thesis are specified as analytical approximations. The proposed analytical approximations however, can be used for initialization of the sampling based methods as well as selecting proposals in Monte-Carlo (MC) methods. Speeding up these MC methods using the proposed approximation methods is a general direction for future research.

Appendix

A

Expressions for some members of exponential family

Essential expressions and formula for reduction of mixture densities of common exponential family distributions are given in this section. These expressions can be found in (Ardeshiri et al., 2014) as well. Some functions which are used in the expressions such as the gamma function are defined here for completeness. The gamma function is defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} \exp(-x)\, \mathrm{d}x. \tag{A.1}$$

The multivariate gamma function which is a generalization of the gamma function is

$$\begin{aligned} \Gamma_d(t) &= \int_{S \succ 0} \exp(-\operatorname{Tr}(S)) |S|^{t-\frac{d+1}{2}}\, \mathrm{d}S \\ &= \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma\left(t + \frac{1-j}{2}\right). \end{aligned} \tag{A.2}$$

The digamma function is given as

$$\psi(t) = \frac{d}{dt} \log \Gamma(t) = \frac{\Gamma'(t)}{\Gamma(t)}. \tag{A.3}$$

The multivariate polygamma function of order n is defined as

$$\psi_d^{(n)}(t) = \frac{\mathrm{d}^{n+1}}{\mathrm{d}t^{n+1}} \log \Gamma_d(t) \tag{A.4}$$

$$= \sum_{j=1}^{d} \frac{\mathrm{d}^{n+1}}{\mathrm{d}t^{n+1}} \log \Gamma\left(t + \frac{1-j}{2}\right) \tag{A.5}$$

$$= \sum_{j=1}^{d} \psi^{(n)}\left(t + \frac{1-j}{2}\right). \tag{A.6}$$

The multinomial beta function in terms of the gamma function is given as

$$\mathrm{B}_K(\alpha) = \frac{\prod_{j=1}^{K} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{K} \alpha_j\right)}. \tag{A.7}$$

Exponential Distribution

$$\mathrm{Exp}(x;\lambda) = \lambda \exp(-\lambda x) \tag{A.8a}$$

$$\text{Support: } x \in [\,0, +\infty) \tag{A.8b}$$

$$\text{Parameter space: } \lambda \in (0, +\infty) \tag{A.8c}$$

$$\eta = -\lambda \tag{A.8d}$$

$$A(\eta) = -\log(-\eta) \tag{A.8e}$$

$$\nabla_\eta A = \frac{\partial A}{\partial \eta} = -\frac{1}{\eta} \tag{A.8f}$$

$$h(x) = 1 \tag{A.8g}$$

$$\mathbb{E}[h(x)] = 1 \tag{A.8h}$$

$$T(x) = x \tag{A.8i}$$

Solution to $\nabla_{\eta^L} A = Y$ is given by

$$\eta^L = -\frac{1}{Y}. \tag{A.9}$$

Weibull Distribution with known shape k

$$\text{Weibull}(x;\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left(-\frac{x^k}{\lambda^k}\right) \tag{A.10a}$$

$$\text{Support: } x \in [\,0,+\infty) \tag{A.10b}$$

$$\text{Parameter space: } \lambda \in (0,+\infty),\ k \in (0,+\infty) \tag{A.10c}$$

$$\eta = -\frac{1}{\lambda^k} \tag{A.10d}$$

$$A(\eta) = -\log(-\eta) - \log(k) \tag{A.10e}$$

$$\nabla_\eta A = \frac{\partial A}{\partial \eta} = -\frac{1}{\eta} \tag{A.10f}$$

$$h(x) = x^{k-1} \tag{A.10g}$$

$$\mathbb{E}[h(x)] = \Gamma\left(\frac{2k-1}{k}\right)(-\eta)^{\frac{1-k}{k}} \tag{A.10h}$$

$$T(x) = x^k \tag{A.10i}$$

Solution to $\nabla_{\eta^L} A = Y$ is given by

$$\eta^L = -\frac{1}{Y}. \tag{A.11}$$

The expression for $\mathbb{E}_{q(x;\eta)}[h(x)]$ is derived here

$$\begin{aligned}
\mathbb{E}[h(x)] &= \int_0^\infty x^{k-1}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left(-\frac{x^k}{\lambda^k}\right)\,\mathrm{d}x = \left\{z \triangleq \frac{x^k}{\lambda^k},\ \mathrm{d}z = k\frac{x^{k-1}}{\lambda^k}\,\mathrm{d}x\right\} \\
&= \int_0^\infty \lambda^{k-1} z^{\frac{k-1}{k}}\exp(-z)\,\mathrm{d}z \\
&= \lambda^{k-1}\Gamma\left(\frac{k-1}{k}+1\right) = \lambda^{k-1}\Gamma\left(\frac{2k-1}{k}\right) = \Gamma\left(\frac{2k-1}{k}\right)(-\eta)^{\frac{1-k}{k}}.
\end{aligned} \tag{A.12}$$

Laplace Distribution with known mean μ

$$\text{Laplace}(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \tag{A.13a}$$

$$\text{Support: } x \in (-\infty, \infty) \tag{A.13b}$$

$$\text{Parameter space: } b \in (0, +\infty),\ \mu \in \mathbb{R} \tag{A.13c}$$

$$\eta = -\frac{1}{b} \tag{A.13d}$$

$$A(\eta) = \log\left(-\frac{2}{\eta}\right) \tag{A.13e}$$

$$\nabla_\eta A = \frac{\partial A}{\partial \eta} = -\frac{1}{\eta} \tag{A.13f}$$

$$h(x) = 1 \tag{A.13g}$$

$$\mathbb{E}[h(x)] = 1 \tag{A.13h}$$

$$T(x) = |x-\mu| \tag{A.13i}$$

Solution to $\nabla_{\eta^L} A = Y$ is given by

$$\eta^L = -\frac{1}{Y}. \tag{A.14}$$

Rayleigh Distribution

$$\text{Rayleigh}(x;\sigma) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{A.15a}$$

$$\text{Support: } x \in [\,0, +\infty) \tag{A.15b}$$

$$\text{Parameter space: } \sigma \in (0, +\infty) \tag{A.15c}$$

$$\eta = -\frac{1}{2\sigma^2} \tag{A.15d}$$

$$A(\eta) = -\log(-2\eta) \tag{A.15e}$$

$$\nabla_\eta A = \frac{\partial A}{\partial \eta} = -\frac{1}{\eta} \tag{A.15f}$$

$$h(x) = x \tag{A.15g}$$

$$\mathbb{E}[h(x)] = \sqrt{\frac{\pi}{-2\eta}} \tag{A.15h}$$

$$T(x) = x^2 \tag{A.15i}$$

Solution to $\nabla_{\eta^L} A = Y$ is given by

$$\eta^L = -\frac{1}{Y}. \tag{A.16}$$

The expression for $\mathbb{E}_{q(x;\eta)}[h(x)]$ is derived here

$$\mathbb{E}[h(x)] = \int_0^\infty x\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)\,\mathrm{d}x = \sigma\sqrt{\frac{\pi}{2}} = \sqrt{\frac{\pi}{-2\eta}}. \tag{A.17}$$

Log-normal Distribution

$$\log -\mathcal{N}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\log x - \mu)^2\right) \tag{A.18a}$$

$$\text{Support: } x \in (0,+\infty) \tag{A.18b}$$

$$\text{Parameter space: } \sigma \in (0,+\infty),\ \mu \in \mathbb{R} \tag{A.18c}$$

$$\eta = (\eta_1, \eta_2) \tag{A.18d}$$

$$\eta_1 = \frac{\mu}{\sigma^2} \tag{A.18e}$$

$$\eta_2 = -\frac{1}{2\sigma^2} \tag{A.18f}$$

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) \tag{A.18g}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}\right) \tag{A.18h}$$

$$\frac{\partial A}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} \tag{A.18i}$$

$$\frac{\partial A}{\partial \eta_2} = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} \tag{A.18j}$$

$$h(x) = \frac{1}{x\sqrt{2\pi}} \tag{A.18k}$$

$$\mathbb{E}[h(x)] = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\eta_1}{2\eta_2} - \frac{1}{4\eta_2}\right) \tag{A.18l}$$

$$T(x) = \left(\log(x), (\log(x))^2\right) \tag{A.18m}$$

Solution to the system of equations $\nabla_{\eta^L} A = Y$ is given by

$$\eta_2^L = \frac{-2}{Y_2 - Y_1^2}, \tag{A.19a}$$

$$\eta_1^L = -2Y_1\eta_2^L. \tag{A.19b}$$

The expression for $\mathbb{E}_{q(x;\eta)}[h(x)]$ is derived here

$$\mathbb{E}[h(x)] = \frac{1}{\sqrt{2\pi}}\mathbb{E}\left[\frac{1}{x}\right] = \frac{1}{\sqrt{2\pi}}\exp\left(-\mu + \frac{\sigma^2}{2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{\eta_1}{2\eta_2} - \frac{1}{4\eta_2}\right). \tag{A.20}$$

Gamma Distribution

$$\text{Gamma}(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x) \tag{A.21a}$$

$$\text{Support: } x \in (0,+\infty) \tag{A.21b}$$

$$\text{Parameter space: } \alpha \in (0,+\infty),\ \beta \in (0,+\infty) \tag{A.21c}$$

$$\eta = (\eta_1,\eta_2) \tag{A.21d}$$

$$\eta_1 = \alpha - 1 \tag{A.21e}$$

$$\eta_2 = -\beta \tag{A.21f}$$

$$A(\eta) = \log\Gamma(\eta_1+1) - (\eta_1+1)\log(-\eta_2) \tag{A.21g}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}\right) \tag{A.21h}$$

$$\frac{\partial A}{\partial \eta_1} = \psi(\eta_1+1) - \log(-\eta_2) \tag{A.21i}$$

$$\frac{\partial A}{\partial \eta_2} = -\frac{\eta_1+1}{\eta_2} \tag{A.21j}$$

$$h(x) = 1 \tag{A.21k}$$

$$\mathbb{E}[h(x)] = 1 \tag{A.21l}$$

$$T(x) = (\log(x), x) \tag{A.21m}$$

To solve the system of equations $\nabla_{\eta^L} A = Y$, first let $Z = \log(Y_2) - Y_1$ and $u = \eta_1 + 1$. Then solve $\psi(u) - \log(u) + Z = 0$ numerically and obtain

$$\eta_1^L = u - 1, \tag{A.22a}$$

$$\eta_2^L = -\frac{u}{Y_2}. \tag{A.22b}$$

Inverse Gamma Distribution

$$\text{IGamma}(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right) \tag{A.23a}$$

$$\text{Support: } x \in (0,+\infty) \tag{A.23b}$$

$$\text{Parameter space: } \alpha \in (0,+\infty),\ \beta \in (0,+\infty) \tag{A.23c}$$

$$\eta = (\eta_1, \eta_2) \tag{A.23d}$$

$$\eta_1 = -\alpha - 1 \tag{A.23e}$$

$$\eta_2 = -\beta \tag{A.23f}$$

$$A(\eta) = \log\Gamma(-\eta_1 - 1) - (-\eta_1 - 1)\log(-\eta_2) \tag{A.23g}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}\right) \tag{A.23h}$$

$$\frac{\partial A}{\partial \eta_1} = -\psi(-\eta_1 - 1) + \log(-\eta_2) \tag{A.23i}$$

$$\frac{\partial A}{\partial \eta_2} = \frac{\eta_1 + 1}{\eta_2} \tag{A.23j}$$

$$h(x) = 1 \tag{A.23k}$$

$$\mathbb{E}[h(x)] = 1 \tag{A.23l}$$

$$T(x) = \left(\log(x), \frac{1}{x}\right) \tag{A.23m}$$

To solve the system of equations $\nabla_{\eta^L} A = Y$ first let $Z = \log(Y_2) + Y_1$ and $u = -\eta_1 - 1$. Then solve $\psi(u) + \log(u) - Z = 0$ numerically and obtain

$$\eta_1^L = -u - 1, \tag{A.24a}$$

$$\eta_2^L = -\frac{u}{Y_2}. \tag{A.24b}$$

Univariate Gaussian Distribution

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{A.25a}$$

$$\text{Support: } x \in \mathbb{R} \tag{A.25b}$$

$$\text{Parameter space: } \sigma \in (0,+\infty),\ \mu \in \mathbb{R} \tag{A.25c}$$

$$\eta = (\eta_1, \eta_2) \tag{A.25d}$$

$$\eta_1 = \frac{\mu}{\sigma^2} \tag{A.25e}$$

$$\eta_2 = -\frac{1}{2\sigma^2} \tag{A.25f}$$

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) \tag{A.25g}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}\right) \tag{A.25h}$$

$$\frac{\partial A}{\partial \eta_1} = -\frac{\eta_1}{2\eta_2} \tag{A.25i}$$

$$\frac{\partial A}{\partial \eta_2} = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} \tag{A.25j}$$

$$h(x) = \frac{1}{\sqrt{2\pi}} \tag{A.25k}$$

$$\mathbb{E}[h(x)] = \frac{1}{\sqrt{2\pi}} \tag{A.25l}$$

$$T(x) = \left(x, x^2\right) \tag{A.25m}$$

Solution to the system of equations $\nabla_{\eta^L} A = Y$ is given by

$$\eta_2^L = \frac{-2}{Y_2 - Y_1^2}, \tag{A.26a}$$

$$\eta_1^L = -2Y_1\eta_2^L. \tag{A.26b}$$

Multivariate Gaussian Distribution

$$\mathcal{N}(x; m, P) = (2\pi)^{-\frac{k}{2}} |P|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-m)^{\mathrm{T}} P^{-1}(x-m)\right) \tag{A.27a}$$

$$\text{Support: } x \in \mathbb{R}^k \tag{A.27b}$$

$$\text{Parameter space: } P \in \mathbb{R}^{k\times k} \text{ and } P = P^{\mathrm{T}} \succ 0,\ \mu \in \mathbb{R}^k \tag{A.27c}$$

$$\eta = (\eta_1, \eta_2) \tag{A.27d}$$

$$\eta_1 = P^{-1} m \tag{A.27e}$$

$$\eta_2 = -\frac{1}{2} P^{-1} \tag{A.27f}$$

$$A(\eta) = -\frac{1}{4}\eta_1^{\mathrm{T}} \eta_2^{-1} \eta_1 - \frac{1}{2}\log|-2\eta_2| \tag{A.27g}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}\right) \tag{A.27h}$$

$$\frac{\partial A}{\partial \eta_1} = -\frac{1}{2}\eta_1^{\mathrm{T}} \eta_2^{-1} \tag{A.27i}$$

$$\frac{\partial A}{\partial \eta_2} = \frac{1}{4}\eta_2^{-\mathrm{T}} \eta_1 \eta_1^{\mathrm{T}} \eta_2^{-\mathrm{T}} - \frac{1}{2}\eta_2^{-1} \tag{A.27j}$$

$$h(x) = (2\pi)^{-k/2} \tag{A.27k}$$

$$\mathbb{E}[h(x)] = (2\pi)^{-k/2} \tag{A.27l}$$

$$T(x) = \left(x, xx^{\mathrm{T}}\right) \tag{A.27m}$$

Solution to system of equations $\nabla_{\eta^L} A = Y$ is given by

$$\eta_2^L = -\frac{1}{2}(Y_2 - Y_1^{\mathrm{T}} Y_1)^{-1}, \tag{A.28}$$

$$\eta_1^L = \left(-2Y_1 \eta_2^L\right)^{\mathrm{T}}. \tag{A.29}$$

Gaussian Gamma Distribution

$$\text{GaussianGamma}(x, \tau; \mu, \lambda, \alpha, \beta) = \mathcal{N}\left(x; \mu, \frac{1}{\lambda\tau}\right)\text{Gamma}(\tau; \alpha, \beta) \tag{A.30a}$$

$$\text{Support: } x \in \mathbb{R},\ \tau \in (0, +\infty) \tag{A.30b}$$

$$\text{Parameter space: } \alpha \in (0, +\infty),\ \beta \in (0, +\infty),\ \lambda \in (0, +\infty),\ \mu \in \mathbb{R} \tag{A.30c}$$

$$\eta = (\eta_1, \eta_2, \eta_3, \eta_4) \tag{A.30d}$$

$$\eta_1 = \alpha - \frac{1}{2} \tag{A.30e}$$

$$\eta_2 = -\beta - \frac{\lambda\mu^2}{2} \tag{A.30f}$$

$$\eta_3 = \lambda\mu \tag{A.30g}$$

$$\eta_4 = -\frac{\lambda}{2} \tag{A.30h}$$

$$A(\eta) = \log\Gamma\left(\eta_1 + \frac{1}{2}\right) - \frac{1}{2}\log(-2\eta_4) - \left(\eta_1 + \frac{1}{2}\right)\log\left(-\eta_2 + \frac{\eta_3^2}{4\eta_4}\right) \tag{A.30i}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}, \frac{\partial A}{\partial \eta_3}, \frac{\partial A}{\partial \eta_4}\right) \tag{A.30j}$$

$$\frac{\partial A}{\partial \eta_1} = \psi\left(\eta_1 + \frac{1}{2}\right) - \log\left(-\eta_2 + \frac{\eta_3^2}{4\eta_4}\right) \tag{A.30k}$$

$$\frac{\partial A}{\partial \eta_2} = -\frac{\eta_1 + \frac{1}{2}}{-\eta_2 + \frac{\eta_3^2}{4\eta_4}} \tag{A.30l}$$

$$\frac{\partial A}{\partial \eta_3} = -\frac{2\eta_3 + \left(\eta_1 + \frac{1}{2}\right)}{4\eta_4\left(-\eta_2 + \frac{\eta_3^2}{4\eta_4}\right)} \tag{A.30m}$$

$$\frac{\partial A}{\partial \eta_4} = \frac{\eta_3^2\left(\eta_1 + \frac{1}{2}\right)}{4\eta_4^2\left(-\eta_2 + \frac{\eta_3^2}{4\eta_4}\right)} - \frac{1}{2\eta_4} \tag{A.30n}$$

$$h(x) = \frac{1}{\sqrt{2\pi}} \tag{A.30o}$$

$$\mathbb{E}[h(x)] = \frac{1}{\sqrt{2\pi}} \tag{A.30p}$$

$$T(x) = \left(\log(\tau), \tau, \tau x, \tau x^2\right) \tag{A.30q}$$

To solve the system of equations $\nabla_{\eta^L} A = Y$ first let $Z = \log(-Y_2) - Y_1$ and $u = \eta_1 + \frac{1}{2}$. Then solve $\psi(u) - \log(u) + Z = 0$ numerically and obtain

$$\eta_1^L = u - \frac{1}{2}, \tag{A.31a}$$

$$\eta_4^L = -\frac{1}{2}\left(\frac{Y_3^2}{Y_2} + Y_4\right)^{-1}, \tag{A.31b}$$

$$\eta_3^L = \frac{2\eta_4 Y_3}{Y_2}, \tag{A.31c}$$

$$\eta_2^L = \frac{\eta_3^2}{4\eta_4} + \frac{\eta_1 + \frac{1}{2}}{Y_2}. \tag{A.31d}$$

Dirichlet distribution

$$\mathrm{Dir}_K(x;\alpha) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \tag{A.32a}$$

$$\text{Support: } x_i \in [0,1] \text{ for } i = 1 \cdots K \text{ and } \sum_{i=1}^{K} x_i = 1 \tag{A.32b}$$

$$\text{Parameter space: } \alpha_i > 0 \text{ and } K \geq 2 \tag{A.32c}$$

$$\eta = (\eta_1, \cdots, \eta_K) \tag{A.32d}$$

$$\eta_i = \alpha_i - 1 \tag{A.32e}$$

$$A(\eta) = \sum_{i=1}^{K} \log \Gamma(\eta_i + 1) - \log \Gamma\left(\sum_{i=1}^{K} (\eta_i + 1) \right) \tag{A.32f}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}, \cdots, \frac{\partial A}{\partial \eta_K} \right) \tag{A.32g}$$

$$\frac{\partial A}{\partial \eta_i} = \psi(\eta_i + 1) - \psi\left(\sum_{i=1}^{K} (\eta_i + 1) \right) \tag{A.32h}$$

$$h(x) = 1 \tag{A.32i}$$

$$\mathbb{E}[h(x)] = 1 \tag{A.32j}$$

$$T(x) = (\log(x_1), \cdots, \log(x_K)) \tag{A.32k}$$

The system of equations $\nabla_{\eta^L} A = Y$ can be solved using a numerical method such as newton method where the Hessian is given by,

$$\frac{\partial^2 A}{\partial {\eta_i}^2} = \psi^{(1)}(\eta_i + 1) - \psi^{(1)}\left(\sum_{k=1}^{K} (\eta_k + 1) \right), \tag{A.33a}$$

$$\frac{\partial^2 A}{\partial \eta_{ij}} = -\psi^{(1)}\left(\sum_{k=1}^{K} (\eta_k + 1) \right). \tag{A.33b}$$

Wishart Distribution

$$\mathcal{W}_d(X; n, V) = \frac{|X|^{\frac{1}{2}(n-d-1)} \exp \operatorname{Tr}\left(-\frac{1}{2} V^{-1} X\right)}{2^{\frac{1}{2}nd} \Gamma_d\left(\frac{1}{2}n\right) |V|^{\frac{1}{2}n}} \tag{A.34a}$$

$$\text{Support: } X \in \mathbb{R}^{d\times d} \text{ and } X = X^{\mathrm{T}} \succ 0 \tag{A.34b}$$

$$\text{Parameter space: } V \in \mathbb{R}^{d\times d} \text{ and } V = V^{\mathrm{T}} \succ 0,\ n \geq d \tag{A.34c}$$

$$\eta = (\eta_1, \eta_2) \tag{A.34d}$$

$$\eta_1 = \frac{1}{2}(n - d - 1) \tag{A.34e}$$

$$\eta_2 = -\frac{1}{2} V^{-1} \tag{A.34f}$$

$$A(\eta) = -\left(\eta_1 + \frac{d+1}{2}\right) \log|-\eta_2| + \log \Gamma_d\left(\eta_1 + \frac{d+1}{2}\right) \tag{A.34g}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}\right) \tag{A.34h}$$

$$\frac{\partial A}{\partial \eta_1} = -\log|-\eta_2| + \psi_d\left(\eta_1 + \frac{d+1}{2}\right) \tag{A.34i}$$

$$\frac{\partial A}{\partial \eta_2} = -\left(\eta_1 + \frac{d+1}{2}\right) \eta_2^{-1} \tag{A.34j}$$

$$h(X) = 1 \tag{A.34k}$$

$$\mathbb{E}[h(X)] = 1 \tag{A.34l}$$

$$T(X) = (\log|X|, X) \tag{A.34m}$$

To solve the system of equations $\nabla_{\eta^L} A = Y$ first let $Z = \log|Y_2| - Y_1$ and $u = \eta_1 + \frac{d+1}{2}$. Then solve $\psi_d(u) - d\log(u) + Z = 0$ numerically and obtain

$$\eta_1^L = u - \frac{d+1}{2}, \tag{A.35a}$$

$$\eta_2^L = -u\, Y_2^{-1}. \tag{A.35b}$$

Inverse Wishart Distribution

$$\mathcal{IW}_d(X;\nu,\Psi) = \frac{|\Psi|^{\frac{1}{2}(\nu-d-1)}\exp\operatorname{Tr}\left(-\frac{1}{2}\Psi X^{-1}\right)}{2^{\frac{1}{2}(\nu-d-1)d}\Gamma_d\left(\frac{1}{2}(\nu-d-1)\right)|X|^{\frac{1}{2}\nu}} \tag{A.36a}$$

$$\text{Support: } X \in \mathbb{R}^{d\times d} \text{ and } X = X^{\mathrm{T}} \succ 0 \tag{A.36b}$$

$$\text{Parameter space: } \nu > 2d\ \Psi \in \mathbb{R}^{d\times d},\ \Psi = \Psi^{\mathrm{T}} \succ 0 \tag{A.36c}$$

$$\eta = (\eta_1, \eta_2) \tag{A.36d}$$

$$\eta_1 = -\frac{1}{2}\nu \tag{A.36e}$$

$$\eta_2 = -\frac{1}{2}\Psi \tag{A.36f}$$

$$A(\eta) = \left(\eta_1 + \frac{d+1}{2}\right)\log|-\eta_2| + \log\Gamma_d\left(-\eta_1 - \frac{d+1}{2}\right) \tag{A.36g}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}\right) \tag{A.36h}$$

$$\frac{\partial A}{\partial \eta_1} = \log|-\eta_2| - \psi_d\left(-\eta_1 - \frac{d+1}{2}\right) \tag{A.36i}$$

$$\frac{\partial A}{\partial \eta_2} = \left(\eta_1 + \frac{d+1}{2}\right)\eta_2^{-1} \tag{A.36j}$$

$$h(X) = 1 \tag{A.36k}$$

$$\mathbb{E}[h(X)] = 1 \tag{A.36l}$$

$$T(X) = \left(\log|X|, X^{-1}\right) \tag{A.36m}$$

To solve the system of equations $\nabla_{\eta^L} A = Y$ first let $Z = -\log(Y_2) - Y_1$ and $u = -\eta_1 - \frac{d+1}{2}$. Then solve $-\psi_d(u) + d\log(u) + Z = 0$ numerically and obtain

$$\eta_1^L = -u - \frac{d+1}{2}, \tag{A.37a}$$

$$\eta_2^L = -u\,Y_2^{-1}. \tag{A.37b}$$

Gaussian Inverse Wishart Distribution

$$\mathrm{GIW}(x, X; m, P, \nu, \Psi) = \mathcal{N}(x; m, P)\,\mathcal{IW}_d(X; \nu, \Psi) \tag{A.38a}$$

$$\text{Support: } x \in \mathbb{R}^k,\ X \in \mathbb{R}^{d\times d} \text{ and } X = X^{\mathrm{T}} \succ 0 \tag{A.38b}$$

$$\text{Parameter space: } \nu > 2d\ \Psi \in \mathbb{R}^{d\times d},\ \Psi = \Psi^{\mathrm{T}} \succ 0,$$
$$P \in \mathbb{R}^{k\times k} \text{ and } P = P^{\mathrm{T}} \succ 0,\ \mu \in \mathbb{R}^k \tag{A.38c}$$

$$\eta = (\eta_1, \eta_2, \eta_3, \eta_4) \tag{A.38d}$$

$$\eta_1 = -\frac{1}{2}\nu \tag{A.38e}$$

$$\eta_2 = -\frac{1}{2}\Psi \tag{A.38f}$$

$$\eta_3 = P^{-1}m \tag{A.38g}$$

$$\eta_4 = -\frac{1}{2}P^{-1} \tag{A.38h}$$

$$A(\eta) = \left(\eta_1 + \frac{d+1}{2}\right)\log|-\eta_2| + \log\Gamma_d\left(-\eta_1 - \frac{d+1}{2}\right) - \frac{1}{4}\eta_3^{\mathrm{T}}\eta_4^{-1}\eta_3 - \frac{1}{2}\log|-2\eta_4| \tag{A.38i}$$

$$\nabla_\eta A = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2}, \frac{\partial A}{\partial \eta_3}, \frac{\partial A}{\partial \eta_4}\right) \tag{A.38j}$$

$$\frac{\partial A}{\partial \eta_1} = \log|-\eta_2| - \psi_d\left(-\eta_1 - \frac{d+1}{2}\right) \tag{A.38k}$$

$$\frac{\partial A}{\partial \eta_2} = \left(\eta_1 + \frac{d+1}{2}\right)\eta_2^{-1} \tag{A.38l}$$

$$\frac{\partial A}{\partial \eta_3} = -\frac{1}{2}\eta_3^{\mathrm{T}}\eta_4^{-1} \tag{A.38m}$$

$$\frac{\partial A}{\partial \eta_4} = \frac{1}{4}\eta_4^{-\mathrm{T}}\eta_1\eta_3^{\mathrm{T}}\eta_4^{-\mathrm{T}} - \frac{1}{2}\eta_4^{-1} \tag{A.38n}$$

$$h(x, X) = (2\pi)^{-k/2} \tag{A.38o}$$

$$\mathbb{E}[h(x, X)] = (2\pi)^{-k/2} \tag{A.38p}$$

$$T(x, X) = \left(\log|X|, X^{-1}, x, xx^{\mathrm{T}}\right) \tag{A.38q}$$

To solve the system of equations $\nabla_{\eta^L} A = Y$ first let $Z = -\log(Y_2) - Y_1$ and $u = -\eta_1 - \frac{d+1}{2}$. Then solve $-\psi_d(u) + d\log(u) + Z = 0$ numerically and obtain

$$\eta_1^L = -u - \frac{d+1}{2}, \tag{A.39a}$$

$$\eta_2^L = -u\, Y_2^{-1}, \tag{A.39b}$$

$$\eta_3^L = -\frac{1}{2}(Y_4 - Y_3^{\mathrm{T}} Y_3)^{-1}, \tag{A.39c}$$

$$\eta_4^L = \left(-2 Y_3 \eta_4^L\right)^{\mathrm{T}}. \tag{A.39d}$$

B

Multiple hypothesis testing

Here, the multiple hypothesis testing problem and the maximum *a posteriori* decision rule is given for the sake of completeness (Ardeshiri et al., 2014). For more complete treatment see (Kay, 1998).

Consider that we want to decide among M hypotheses $\{\mathcal{H}_1, \mathcal{H}_2, ..., \mathcal{H}_M\}$. Let the cost assigned to the decision to choose $\mathcal{H}_i$ when $\mathcal{H}_j$ is true is denoted by $\mathcal{C}_{ij}$ where

$$\mathcal{C}_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} . \tag{B.1}$$

The expected Bayes risk (Kay, 1998) becomes

$$\mathcal{R} = \sum_{i=1}^{M} \sum_{j=1}^{M} \mathcal{C}_{ij} P(\mathcal{H}_i|\mathcal{H}_j) P(\mathcal{H}_j). \tag{B.2}$$

We are looking for a decision rule that minimizes $\mathcal{R}$. Let us partition the space to regions R_i for $i = 1 : M$ so that

$$\begin{aligned} \mathcal{R} &= \sum_{i=1}^{M} \sum_{j=1}^{M} \mathcal{C}_{ij} \int_{R_i} p(x|\mathcal{H}_j) P(\mathcal{H}_j) \, \mathrm{d}x \\ &= \sum_{i=1}^{M} \int_{R_i} \sum_{j=1}^{M} \mathcal{C}_{ij} P(\mathcal{H}_j|x) p(x) \, \mathrm{d}x \\ &= \sum_{i=1}^{M} \int_{R_i} \mathcal{C}_i p(x) \, \mathrm{d}x \end{aligned} \tag{B.3}$$

where $\mathcal{C}_i(x) = \sum_{j=1}^{M} \mathcal{C}_{ij} P(\mathcal{H}_j|x)$. Since each data x should trigger only one decision, i.e. assigned to only one of the R_i partitions we should decide $\mathcal{H}_k$ for which $\mathcal{C}_i$ is minimum.

Since $\mathcal{C}_i(x) = \sum_{j=1}^{M} P(\mathcal{H}_j|x) - P(\mathcal{H}_i|x)$, $\mathcal{C}_i(x)$ is minimized if $P(\mathcal{H}_i|x)$ is maximized. Thus the decision rule is decide $\mathcal{H}_k$ if $P(\mathcal{H}_k|x) > P(\mathcal{H}_i|x)$ for $i \neq k$. For equal prior probabilities $P(\mathcal{H}_k) = P(\mathcal{H}_i)$ the decision rule will be to decide $\mathcal{H}_k$ if $p(x|\mathcal{H}_k) > p(x|\mathcal{H}_i)$ for $i \neq k$. This decision rule is also referred to as maximum a posteriori decision rule.

If the prior probabilities are not equal due to e.g., heuristics $P(\mathcal{H}_k) \neq P(\mathcal{H}_i)$, Bayes rule $P(\mathcal{H}_i|x) \propto p(x|\mathcal{H}_i)P(\mathcal{H}_i)$ can be used. This possibility is not exploited in this thesis.

C

Implementation aspects of the ISE approach

An advantage of the ISE metric is that, it can be computed analytically for many distributions (Ardeshiri et al., 2015a). In the ISE approach two parameters can be varied to create slightly different reduction algorithms as detailed below (Ardeshiri et al., 2014):

1. In the first variation, the ISE is calculated for each hypothesis according to $\mathrm{ISE}(\mathcal{H}_K) = \int |p(x) - p_k(x|\mathcal{H}_K)|^2 \, \mathrm{d}x$ and the density after pruning is re-normalized. This variation is consistent with the presentation of the ISE algorithm so far in this technical report.

2. In the second variation, as it is pointed out in (Williams and Maybeck, 2006), when the ISE is being calculated for a pruning hypothesis the rescaling can be skipped since re-normalizing the weights will increase the error value in parts of the support that are not affected by the pruning hypothesis. This choice also brings substantial computational savings.

3. In the third variation, instead of comparing $p(x|\mathcal{H}_K)$ with the original mixture $p(x)$, it is compared with the resulting mixture of the previous reduction step $p_k(x)$, as given here

$$\mathrm{ISE}(\mathcal{H}_K) = \int |p_k(x) - p_k(x|\mathcal{H}_K)|^2 \, \mathrm{d}x.$$

   In this way, the ISE metric for merging decision can be simplified to

$$\begin{aligned}\mathrm{ISE}(\mathcal{H}_{IJ}) = {} & (w^I)^2 Q(I,I) + (w^J)^2 Q(J,J) \\ & + (w^{IJ})^2 Q(IJ,IJ) + 2w^I w^J Q(I,J) \\ & - 2w^I w^{IJ} Q(I,IJ) - 2w^J w^{IJ} Q(J,IJ).\end{aligned}$$

where,

$$Q(I,J) = \int q(x;\eta^I)q(x;\eta^J)\,\mathrm{d}x. \tag{C.1}$$

$Q(I,J)$ can be calculated analytically for many basic densities of interest belonging to the exponential family such as Gaussian, gamma and Wishart distributions. For explicit expressions for the exponential family of distribution see (Ardeshiri et al., 2015a) and (Ardeshiri et al., 2014).

Similarly the ISE metric for pruning decision can be simplified as in

$$\mathrm{ISE}(\mathcal{H}_{0I}) = \left(\frac{w^I}{1-w^I}\right)^2 \left[Q(I,I) - 2\sum_{i=1}^{N} w^i Q(I,i) + \sum_{i=1}^{N}\sum_{j=1}^{N} w^i w^j Q(i,j)\right].$$

4. The fourth variant is similar to the third variant in terms of the choice of the reference density, but the mixture is not renormalized after each pruning which results in the expression

$$\mathrm{ISE}(\mathcal{H}_{0I}) = (w^I)^2 Q(I,I)$$

for pruning hypotheses.

Calculation of the ISE for each hypothesis at every step of the reduction is costly. A scheme is suggested here to cache the calculated quantities to reduce the computational cost of the reduction. The cost reduction scheme is given for the second type of implementation of the ISE approach, where the mixture density after pruning hypothesis is not re-normalized.

In the first step of the reduction of the mixture density (4.1) merging of all possible pairs of components results in $\frac{1}{2}N(N-1)$ hypotheses. For the evaluation of these hypotheses the resultant component of each merging should be calculated. To calculate the ISE of each hypothesis $Q(\cdot,\cdot)$ should be calculated for all pairs of components in the mixture as well as the pair of components where one component is among the merged components and the other one is among the existing components. All these quantities should be stored and can be reused in the future reduction steps.

At the k^{th} step of the reduction of the mixture density given in (4.1), the reduced density is denoted by $p_k(x)$. In order to keep the notation less cluttered, let the term q^J denote $w^J q(x;\eta^J)$; p denote $p(x)$ and p_k denote $p_k(x)$. Let us assume that the cost of the reduction hypotheses at the k^{th} stage denoted by $\mathrm{ISE}_k(\mathcal{H}_R)$ are stored in a vector Y_k and let $M = \operatorname{argmin} \mathrm{ISE}_k(\mathcal{H}_R)$ for all permissible values of R.

When M corresponds to a pruning hypothesis, for example $M = 0J$, the vector Y_{k+1} can be updated with less computations for next pruning hypotheses using

$$\begin{aligned}
\mathrm{ISE}_{k+1}(\mathcal{H}_{0S}|M = 0J) &= \int (p - p_k + q^J + q^S)^2 \, \mathrm{d}x \\
&= \int (p - p_k + q^S)^2 \, \mathrm{d}x + \int (q^J)^2 \, \mathrm{d}x + 2 \int q^J (p - p_k + q^S) \, \mathrm{d}x \\
&= \int (p - p_k + q^S)^2 \, \mathrm{d}x + \int (q^J)^2 \, \mathrm{d}x + 2 \int q^J (p - p_k) \, \mathrm{d}x + 2 \int q^J q^S \, \mathrm{d}x \\
&= \mathrm{ISE}_k(\mathcal{H}_{0S}) + \underbrace{\int (q^J)^2 \, \mathrm{d}x + 2 \int q^J (p - p_k) \, \mathrm{d}x}_{A(J)} + 2 \int q^J q^S \, \mathrm{d}x,
\end{aligned} \tag{C.2}$$

where, the quantity $\mathrm{ISE}_k(\mathcal{H}_{0S})$ is already known from the previous step and $A(J)$ is a part of the ISE added to elements of Y_k due to the pruning of the J^{th} component.

Similarly, when M corresponds to a pruning hypothesis, for example $M = 0J$, the vector Y_{k+1} can be updated with less computations for the next merging hypotheses using

$$\begin{aligned}
&\mathrm{ISE}_{k+1}(\mathcal{H}_{ST}|M = 0J) = \int (p - p_k + q^J + q^S + q^T - q^{ST})^2 \, \mathrm{d}x \\
&= \int (p - p_k + q^S + q^T - q^{ST})^2 \, \mathrm{d}x + \int (q^J)^2 \, \mathrm{d}x + 2 \int q^J (p - p_k + q^S + q^T - q^{ST}) \, \mathrm{d}x \\
&= \int (p - p_k + q^S + q^T - q^{ST})^2 \, \mathrm{d}x + \int (q^J)^2 \, \mathrm{d}x + 2 \int q^J (p - p_k) \, \mathrm{d}x \\
&+ 2 \int q^J (q^S + q^T - q^{ST}) \, \mathrm{d}x \\
&= \mathrm{ISE}_k(\mathcal{H}_{ST}) + A(J) + 2 \int q^J (q^S + q^T - q^{ST}) \, \mathrm{d}x.
\end{aligned} \tag{C.3}$$

After each pruning step all elements of vector Y_{k+1} corresponding to the pruned component will be eliminated from Y_{k+1}.

Using a similar approach, when M corresponds to a merging hypothesis, say $M = IJ$, the vector Y_{k+1} can be updated with less computations for the next prun-

ing hypotheses using

$$
\begin{aligned}
&\mathrm{ISE}_{k+1}(\mathcal{H}_{0S}|M=IJ) = \int (p - p_k + q^J + q^I - q^{IJ} + q^S)^2 \,\mathrm{d}x \\
&= \int (p - p_k + q^S)^2 \,\mathrm{d}x + \int (q^J + q^I - q^{IJ})^2 \,\mathrm{d}x \\
&+ 2\int (q^J + q^I - q^{IJ})(p - p_k + q^S) \,\mathrm{d}x \\
&= \int (p - p_k + q^S)^2 \,\mathrm{d}x + \int (q^J + q^I - q^{IJ})^2 \,\mathrm{d}x \\
&+ 2\int (q^J + q^I - q^{IJ})(p - p_k) \,\mathrm{d}x + 2\int (q^J + q^I - q^{IJ})q^S \,\mathrm{d}x \\
&= \mathrm{ISE}_k(\mathcal{H}_{0S}) + \underbrace{\int (q^J + q^I - q^{IJ})^2 \,\mathrm{d}x + 2\int (q^J + q^I - q^{IJ})(p - p_k) \,\mathrm{d}x}_{C(I,J)} \\
&+ 2\int (q^J + q^I - q^{IJ})q^S \,\mathrm{d}x,
\end{aligned}
\tag{C.4}
$$

and for the next merging hypotheses using

$$
\begin{aligned}
&\mathrm{ISE}_{k+1}(\mathcal{H}_{ST}|M=IJ) = \int (p - p_k + q^J + q^I - q^{IJ} + q^S + q^T - q^{ST})^2 \,\mathrm{d}x \\
&= \int (p - p_k + q^S + q^T - q^{ST})^2 \,\mathrm{d}x + \int (q^J + q^I - q^{IJ})^2 \,\mathrm{d}x \\
&+ 2\int (q^J + q^I - q^{IJ})(p - p_k + q^S + q^T - q^{ST}) \,\mathrm{d}x \\
&= \int (p - p_k + q^S + q^T - q^{ST})^2 \,\mathrm{d}x + \int (q^J + q^I - q^{IJ})^2 \,\mathrm{d}x \\
&+ 2\int (q^J + q^I - q^{IJ})(p - p_k) \,\mathrm{d}x \\
&+ 2\int (q^J + q^I - q^{IJ})(q^S + q^T - q^{ST}) \,\mathrm{d}x \\
&= \mathrm{ISE}_k(\mathcal{H}_{ST}) + C(I,J) + 2\int (q^J + q^I - q^{IJ})(q^S + q^T - q^{ST}) \,\mathrm{d}x.
\end{aligned}
\tag{C.5}
$$

When two components I and J are merged, the merged component labeled IJ will obtain the label of component I in the computation environment and all elements of Y_{k+1} corresponding to element J will be eliminated. The vector Y_{k+1}

should be updated for the new component as in

$$\begin{aligned}
&\mathrm{ISE}_{k+1}(\mathcal{H}_{(IJ)S}|M = IJ) = \int (p - p_k + q^J + q^I - q^{IJ} + q^S + q^{IJ} - q^{(IJ)S})^2 \,\mathrm{d}x \\
&= \int (p - p_k)^2 \,\mathrm{d}x + \int (q^J + q^I + q^S - q^{(IJ)S})^2 \,\mathrm{d}x \\
&+ 2 \int (p - p_k)(q^J + q^I + q^S - q^{(IJ)S}) \,\mathrm{d}x,
\end{aligned} \tag{C.6}$$

where, the first term is known from the last reduction step.

Bibliography

D. Alspach and H. Sorenson. Nonlinear Bayesian estimation using Gaussian sum approximations. *Automatic Control, IEEE Transactions on*, 17(4):439–448, 1972. ISSN 0018-9286. doi: 10.1109/TAC.1972.1100034.

T. Ardeshiri and T. Chen. Maximum entropy property of discrete-time stable spline kernel. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 3676–3680, April 2015. doi: 10.1109/ICASSP.2015.7178657.

T. Ardeshiri and E. Özkan. An adaptive PHD filter for tracking with unknown sensor characteristics. In *Information Fusion (FUSION), 2013 16th International Conference on*, pages 1736–1743, July 2013.

T. Ardeshiri, S. Kharrazi, J. Sjöberg, J. Bärgman, and L. M. Sensor fusion for vehicle positioning in intersection active safety applications. In *International Symposium on Advanced Vehicle Control*, 2006a.

T. Ardeshiri, S. Kharrazi, R. Thomson, and J. Bärgman. Offset eliminative map matching algorithm for intersection active safety applications. In *Intelligent Vehicles Symposium, 2006 IEEE*, pages 82–88, 2006b. doi: 10.1109/IVS.2006.1689609.

T. Ardeshiri, F. Larsson, F. Gustafsson, T. Schön, and M. Felsberg. Bicycle tracking using ellipse extraction. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8, July 2011a.

T. Ardeshiri, M. Norrlöf, J. Löfberg, and A. Hansson. Convex optimization approach for time-optimal path tracking of robots with speed dependent constraints. In *Proceedings of the 18th IFAC World Congress, Milan, Italy*, pages 14648–14653, August 2011b.

T. Ardeshiri, U. Orguner, C. Lundquist, and T. Schön. On mixture reduction for multiple target tracking. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 692–699, July 2012.

T. Ardeshiri, K. Granström, E. Özkan, and U. Orguner. Greedy reduction algorithms for mixtures of exponential family. *Signal Processing Letters, IEEE*, 22 (6):676–680, June 2015a. ISSN 1070-9908. doi: 10.1109/LSP.2014.2367154.

T. Ardeshiri, U. Orguner, and F. Gustafsson. Bayesian inference via approximation of log-likelihood for priors in exponential family. *ArXiv e-prints*, October 2015b. Submitted to Signal Processing, IEEE Transactions on.

T. Ardeshiri, U. Orguner, and E. Özkan. Gaussian Mixture Reduction Using Reverse Kullback-Leibler Divergence. *ArXiv e-prints*, August 2015. To be Submitted to Signal Processing, IEEE Transactions on.

T. Ardeshiri, E. Özkan, U. Orguner, and F. Gustafsson. Approximate Bayesian smoothing with unknown process and measurement noise covariances. *To appear in Signal Processing Letters, IEEE*, 2015.

T. Ardeshiri, E. Özkan, and U. Orguner. On reduction of mixtures of the exponential family distributions. Technical Report LiTH-ISY-R-3076, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden, August 2014. URL `http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-100234`.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

S. Blackman and R. Popoli. *Design and analysis of modern tracking systems*. Artech House radar library. Artech House, 1999. ISBN 9781580530064.

T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and Gaussian processes - Revisited. *Automatica*, 48:1525–1535, 2012a.

T. Chen, T. Ardeshiri, F. P. Carli, A. Chiuso, L. Ljung, and G. Pillonetto. Maximum entropy properties of discrete-time first-order stable spline kernel. *To appear in Automatica*, 2015.

T. Chen, M. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto. System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *Automatic Control, IEEE Transactions on*, 59(11): 2933–2945, Nov 2014. ISSN 0018-9286. doi: 10.1109/TAC.2014.2351851.

X. Chen, R. Tharmarasa, M. Pelletier, and T. Kirubarajan. Integrated clutter estimation and target tracking using Poisson point processes. *Aerospace and Electronic Systems, IEEE Transactions on*, 48(2):1210–1235, April 2012b. ISSN 0018-9251. doi: 10.1109/TAES.2012.6178058.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

T. M. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, 2006.

M. Feldmann, D. Fränken, and W. Koch. Tracking of extended objects and group targets using random matrices. *Signal Processing, IEEE Transactions on*, 59(4): 1409–1420, April 2011.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984. ISSN 0162-8828.

K. Granström and U. Orguner. Estimation and maintenance of measurement rates for multiple extended target tracking. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 2170 –2176, july 2012a.

K. Granström and U. Orguner. On the reduction of Gaussian inverse Wishart mixtures. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 2162 –2169, july 2012b.

W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970. doi: 10.1093/biomet/57.1.97.

W. Hennevogl, L. Fahrmeir, and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. Springer New York, 2001. ISBN 9780387951874.

E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999. ISSN 0885-6125. doi: 10.1023/A:1007665907178.

S. Kay. *Fundamentals of Statistical Signal Processing: Detection theory*. Prentice Hall signal processing series. Prentice-Hall PTR, 1998. ISBN 9780135041352.

G. Kitagawa. The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics*, 46 (4):605–623, 1994.

M. Lifshits. *Random Processes by Example*. World Scientific Publishing Co. Pte. Ltd, 2014. ISBN 978-981-4522-28-1.

L. Ljung. *System Identification - Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., 2nd edition, 1999.

L. Ljung, H. Hjalmarsson, and H. Ohlsson. Four encounters with system identification. *European Journal of Control*, 17:449–471, 2011.

T. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 362–369, San Francisco, CA, 2001. Morgan Kaufmann.

T. Minka. Divergence measures and message passing. Technical report, Microsoft Research Ltd., Cambridge, UK, 2005.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384, 1972. ISSN 00359238.

G. D. Nicolao, G. Ferrari-Trecate, and A. Lecchini. MAXENT priors for stochastic filtering problems. In *Mathematical Theory of Networks and Systems*, Padova, Italy, July 1998.

H. Nurminen, T. Ardeshiri, R. Piché, and F. Gustafsson. Robust inference for state-space models with skewed measurement noise. *Signal Processing Letters, IEEE*, 22(11):1898–1902, Nov 2015a. ISSN 1070-9908. doi: 10.1109/LSP.2015.2437456.

H. Nurminen, T. Ardeshiri, R. Piché, and F. Gustafsson. A NLOS-robust TOA positioning filter based on a skew-t measurement noise model. In *2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Banff, Alberta, Canada*, October 2015b.

A. Papoulis and S. Pillai. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill series in electrical engineering: Communications and signal processing. Tata McGraw-Hill, 2002. ISBN 9780070486584.

G. Pillonetto and G. D. Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

G. Pillonetto and G. D. Nicolao. Kernel selection in linear system identification. Part I: A Gaussian process perspective. In *Proc. 50th IEEE Conference on Decision and Control*, pages 4318–4325, Orlando, Florida, 2011.

G. Pillonetto, A. Chiuso, and G. D. Nicolao. Prediction error identification of linear systems: a nonparametric Gaussian regression approach. *Automatica*, 47(2):291–305, 2011.

G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682, 2014.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

K. J. Åström. *Introduction to stochastic control theory*, volume 70 of *Mathematics in science and engineering*. Academic press, New York, London, 1970. ISBN 0-12-065650-7.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. ISSN 1467-9868.

A. Runnalls. Kullback-Leibler approach to Gaussian mixture reduction. *Aerospace and Electronic Systems, IEEE Transactions on*, 43(3):989–999, July 2007. ISSN 0018-9251. doi: 10.1109/TAES.2007.4383588.

D. J. Salmond. Mixture reduction algorithms for target tracking in clutter. In *Proceeding of SPIE, Signal and Data Processing of Small Targets*, volume 1305, pages 434–445, 1990.

D. G. Tzikas, A. C. Likas, and N. P. Galatsanos. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, November 2008.

B.-N. Vo and W.-K. Ma. The Gaussian mixture probability hypothesis density filter. *Signal Processing, IEEE Transactions on*, 54(11):4091–4104, nov. 2006. ISSN 1053-587X. doi: 10.1109/TSP.2006.881190.

M. Wainwright and M. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and trends in machine learning. Now Publishers, 2008. ISBN 9781601981844.

J. L. Williams and P. S. Maybeck. Cost-function-based hypothesis control techniques for multiple hypothesis tracking. *Mathematical and Computer Modelling*, 43(9-10):976–989, May 2006. ISSN 08957177. doi: 10.1016/j.mcm.2005.05.022.

Part II

Publications

Papers

PhD Dissertations
Division of Automatic Control
Linköping University

M. Millnert: Identification and control of systems subject to abrupt changes. Thesis No. 82, 1982. ISBN 91-7372-542-0.

A. J. M. van Overbeek: On-line structure selection for the identification of multivariable systems. Thesis No. 86, 1982. ISBN 91-7372-586-2.

B. Bengtsson: On some control problems for queues. Thesis No. 87, 1982. ISBN 91-7372-593-5.

S. Ljung: Fast algorithms for integral equations and least squares identification problems. Thesis No. 93, 1983. ISBN 91-7372-641-9.

H. Jonson: A Newton method for solving non-linear optimal control problems with general constraints. Thesis No. 104, 1983. ISBN 91-7372-718-0.

E. Trulsson: Adaptive control based on explicit criterion minimization. Thesis No. 106, 1983. ISBN 91-7372-728-8.

K. Nordström: Uncertainty, robustness and sensitivity reduction in the design of single input control systems. Thesis No. 162, 1987. ISBN 91-7870-170-8.

B. Wahlberg: On the identification and approximation of linear systems. Thesis No. 163, 1987. ISBN 91-7870-175-9.

S. Gunnarsson: Frequency domain aspects of modeling and control in adaptive systems. Thesis No. 194, 1988. ISBN 91-7870-380-8.

A. Isaksson: On system identification in one and two dimensions with signal processing applications. Thesis No. 196, 1988. ISBN 91-7870-383-2.

M. Viberg: Subspace fitting concepts in sensor array processing. Thesis No. 217, 1989. ISBN 91-7870-529-0.

K. Forsman: Constructive commutative algebra in nonlinear control theory. Thesis No. 261, 1991. ISBN 91-7870-827-3.

F. Gustafsson: Estimation of discrete parameters in linear systems. Thesis No. 271, 1992. ISBN 91-7870-876-1.

P. Nagy: Tools for knowledge-based signal processing with applications to system identification. Thesis No. 280, 1992. ISBN 91-7870-962-8.

T. Svensson: Mathematical tools and software for analysis and design of nonlinear control systems. Thesis No. 285, 1992. ISBN 91-7870-989-X.

S. Andersson: On dimension reduction in sensor array signal processing. Thesis No. 290, 1992. ISBN 91-7871-015-4.

H. Hjalmarsson: Aspects on incomplete modeling in system identification. Thesis No. 298, 1993. ISBN 91-7871-070-7.

I. Klein: Automatic synthesis of sequential control schemes. Thesis No. 305, 1993. ISBN 91-7871-090-1.

J.-E. Strömberg: A mode switching modelling philosophy. Thesis No. 353, 1994. ISBN 91-7871-430-3.

K. Wang Chen: Transformation and symbolic calculations in filtering and control. Thesis No. 361, 1994. ISBN 91-7871-467-2.

T. McKelvey: Identification of state-space models from time and frequency data. Thesis No. 380, 1995. ISBN 91-7871-531-8.

J. Sjöberg: Non-linear system identification with neural networks. Thesis No. 381, 1995. ISBN 91-7871-534-2.

R. Germundsson: Symbolic systems – theory, computation and applications. Thesis No. 389, 1995. ISBN 91-7871-578-4.

P. Pucar: Modeling and segmentation using multiple models. Thesis No. 405, 1995. ISBN 91-7871-627-6.
H. Fortell: Algebraic approaches to normal forms and zero dynamics. Thesis No. 407, 1995. ISBN 91-7871-629-2.
A. Helmersson: Methods for robust gain scheduling. Thesis No. 406, 1995. ISBN 91-7871-628-4.
P. Lindskog: Methods, algorithms and tools for system identification based on prior knowledge. Thesis No. 436, 1996. ISBN 91-7871-424-8.
J. Gunnarsson: Symbolic methods and tools for discrete event dynamic systems. Thesis No. 477, 1997. ISBN 91-7871-917-8.
M. Jirstrand: Constructive methods for inequality constraints in control. Thesis No. 527, 1998. ISBN 91-7219-187-2.
U. Forssell: Closed-loop identification: Methods, theory, and applications. Thesis No. 566, 1999. ISBN 91-7219-432-4.
A. Stenman: Model on demand: Algorithms, analysis and applications. Thesis No. 571, 1999. ISBN 91-7219-450-2.
N. Bergman: Recursive Bayesian estimation: Navigation and tracking applications. Thesis No. 579, 1999. ISBN 91-7219-473-1.
K. Edström: Switched bond graphs: Simulation and analysis. Thesis No. 586, 1999. ISBN 91-7219-493-6.
M. Larsson: Behavioral and structural model based approaches to discrete diagnosis. Thesis No. 608, 1999. ISBN 91-7219-615-5.
F. Gunnarsson: Power control in cellular radio systems: Analysis, design and estimation. Thesis No. 623, 2000. ISBN 91-7219-689-0.
V. Einarsson: Model checking methods for mode switching systems. Thesis No. 652, 2000. ISBN 91-7219-836-2.
M. Norrlöf: Iterative learning control: Analysis, design, and experiments. Thesis No. 653, 2000. ISBN 91-7219-837-0.
F. Tjärnström: Variance expressions and model reduction in system identification. Thesis No. 730, 2002. ISBN 91-7373-253-2.
J. Löfberg: Minimax approaches to robust model predictive control. Thesis No. 812, 2003. ISBN 91-7373-622-8.
J. Roll: Local and piecewise affine approaches to system identification. Thesis No. 802, 2003. ISBN 91-7373-608-2.
J. Elbornsson: Analysis, estimation and compensation of mismatch effects in A/D converters. Thesis No. 811, 2003. ISBN 91-7373-621-X.
O. Härkegård: Backstepping and control allocation with applications to flight control. Thesis No. 820, 2003. ISBN 91-7373-647-3.
R. Wallin: Optimization algorithms for system analysis and identification. Thesis No. 919, 2004. ISBN 91-85297-19-4.
D. Lindgren: Projection methods for classification and identification. Thesis No. 915, 2005. ISBN 91-85297-06-2.
R. Karlsson: Particle Filtering for Positioning and Tracking Applications. Thesis No. 924, 2005. ISBN 91-85297-34-8.
J. Jansson: Collision Avoidance Theory with Applications to Automotive Collision Mitigation. Thesis No. 950, 2005. ISBN 91-85299-45-6.
E. Geijer Lundin: Uplink Load in CDMA Cellular Radio Systems. Thesis No. 977, 2005. ISBN 91-85457-49-3.
M. Enqvist: Linear Models of Nonlinear Systems. Thesis No. 985, 2005. ISBN 91-85457-64-7.
T. B. Schön: Estimation of Nonlinear Dynamic Systems — Theory and Applications. Thesis No. 998, 2006. ISBN 91-85497-03-7.

I. Lind: Regressor and Structure Selection — Uses of ANOVA in System Identification. Thesis No. 1012, 2006. ISBN 91-85523-98-4.

J. Gillberg: Frequency Domain Identification of Continuous-Time Systems Reconstruction and Robustness. Thesis No. 1031, 2006. ISBN 91-85523-34-8.

M. Gerdin: Identification and Estimation for Models Described by Differential-Algebraic Equations. Thesis No. 1046, 2006. ISBN 91-85643-87-4.

C. Grönwall: Ground Object Recognition using Laser Radar Data – Geometric Fitting, Performance Analysis, and Applications. Thesis No. 1055, 2006. ISBN 91-85643-53-X.

A. Eidehall: Tracking and threat assessment for automotive collision avoidance. Thesis No. 1066, 2007. ISBN 91-85643-10-6.

F. Eng: Non-Uniform Sampling in Statistical Signal Processing. Thesis No. 1082, 2007. ISBN 978-91-85715-49-7.

E. Wernholt: Multivariable Frequency-Domain Identification of Industrial Robots. Thesis No. 1138, 2007. ISBN 978-91-85895-72-4.

D. Axehill: Integer Quadratic Programming for Control and Communication. Thesis No. 1158, 2008. ISBN 978-91-85523-03-0.

G. Hendeby: Performance and Implementation Aspects of Nonlinear Filtering. Thesis No. 1161, 2008. ISBN 978-91-7393-979-9.

J. Sjöberg: Optimal Control and Model Reduction of Nonlinear DAE Models. Thesis No. 1166, 2008. ISBN 978-91-7393-964-5.

D. Törnqvist: Estimation and Detection with Applications to Navigation. Thesis No. 1216, 2008. ISBN 978-91-7393-785-6.

P-J. Nordlund: Efficient Estimation and Detection Methods for Airborne Applications. Thesis No. 1231, 2008. ISBN 978-91-7393-720-7.

H. Tidefelt: Differential-algebraic equations and matrix-valued singular perturbation. Thesis No. 1292, 2009. ISBN 978-91-7393-479-4.

H. Ohlsson: Regularization for Sparseness and Smoothness — Applications in System Identification and Signal Processing. Thesis No. 1351, 2010. ISBN 978-91-7393-287-5.

S. Moberg: Modeling and Control of Flexible Manipulators. Thesis No. 1349, 2010. ISBN 978-91-7393-289-9.

J. Wallén: Estimation-based iterative learning control. Thesis No. 1358, 2011. ISBN 978-91-7393-255-4.

J. Hol: Sensor Fusion and Calibration of Inertial Sensors, Vision, Ultra-Wideband and GPS. Thesis No. 1368, 2011. ISBN 978-91-7393-197-7.

D. Ankelhed: On the Design of Low Order H-infinity Controllers. Thesis No. 1371, 2011. ISBN 978-91-7393-157-1.

C. Lundquist: Sensor Fusion for Automotive Applications. Thesis No. 1409, 2011. ISBN 978-91-7393-023-9.

P. Skoglar: Tracking and Planning for Surveillance Applications. Thesis No. 1432, 2012. ISBN 978-91-7519-941-2.

K. Granström: Extended target tracking using PHD filters. Thesis No. 1476, 2012. ISBN 978-91-7519-796-8.

C. Lyzell: Structural Reformulations in System Identification. Thesis No. 1475, 2012. ISBN 978-91-7519-800-2.

J. Callmer: Autonomous Localization in Unknown Environments. Thesis No. 1520, 2013. ISBN 978-91-7519-620-6.

D. Petersson: A Nonlinear Optimization Approach to H2-Optimal Modeling and Control. Thesis No. 1528, 2013. ISBN 978-91-7519-567-4.

Z. Sjanic: Navigation and Mapping for Aerial Vehicles Based on Inertial and Imaging Sensors. Thesis No. 1533, 2013. ISBN 978-91-7519-553-7.

F. Lindsten: Particle Filters and Markov Chains for Learning of Dynamical Systems. Thesis No. 1530, 2013. ISBN 978-91-7519-559-9.
P. Axelsson: Sensor Fusion and Control Applied to Industrial Manipulators. Thesis No. 1585, 2014. ISBN 978-91-7519-368-7.
A. Carvalho Bittencourt: Modeling and Diagnosis of Friction and Wear in Industrial Robots. Thesis No. 1617, 2014. ISBN 978-91-7519-251-2.
M. Skoglund: Inertial Navigation and Mapping for Autonomous Vehicles. Thesis No. 1623, 2014. ISBN 978-91-7519-233-8.
S. Khoshfetrat Pakazad: Divide and Conquer: Distributed Optimization and Robustness Analysis. Thesis No. 1676, 2015. ISBN 978-91-7519-050-1.