Department of Neuroscience and Biomedical Engineering /
Department of Computer Science

# Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression

**Arno Solin**

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
**DOCTORAL DISSERTATIONS** 50/2016

# Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression

**Arno Solin**

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall F239a of the school on 8 April 2016 at 12.

**Aalto University**
**School of Science**
**Department of Neuroscience and Biomedical Engineering /**
**Department of Computer Science**
**Bayesian Methodology**

**Supervising professor**
Prof. Jouko Lampinen, Aalto University, Finland

**Thesis advisor**
Prof. Simo Särkkä, Aalto University, Finland

**Preliminary examiners**
Dr. Guido Sanguinetti, University of Edinburgh, United Kingdom
Dr. Finn Lindgren, University of Bath, United Kingdom

**Opponent**
Prof. Manfred Opper, Technical University of Berlin, Germany

Aalto University publication series
**DOCTORAL DISSERTATIONS** 50/2016



ISBN 978-952-60-6710-0 (printed)
ISBN 978-952-60-6711-7 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6711-7

Unigrafia Oy
Helsinki 2016

Finland

441 697
Printed matter

Aalto University

# Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

**Author**
Arno Solin

**Name of the doctoral dissertation**
Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression

**Publisher** School of Science

**Unit** Department of Neuroscience and Biomedical Engineering / Department of Computer Science

**Series** Aalto University publication series DOCTORAL DISSERTATIONS 50/2016

**Field of research** Computational Science

**Manuscript submitted** 15 December 2015 **Date of the defence** 8 April 2016

**Permission to publish granted (date)** 2 February 2016 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

**Abstract**

Gaussian processes (GPs) are widely used tools for non-parametric probabilistic modelling in machine learning, spatial statistics, and signal processing. Their strength lies in flexible model specification, where prior beliefs of the model functions are encoded by the GP model. This way they can also be interpreted as specifying a probability distribution over the space of functions. In signal processing GPs are typically represented as state-space models, whereas the kernel (covariance function) representation is favoured in machine learning. Under the kernel formalism, the naïve solution to a GP regression problem scales cubically in the number of data points, which makes the approach computationally infeasible for large data sets.

This work explores the link between the two representations, which enables the use of efficient sequential Kalman filtering based methods for solving the inference problem. These methods have linear time complexity with respect to the number of data points. The interest is in presenting an explicit connection between a large class of covariance functions and state-space models. This is done for one-dimensional (temporal) covariance functions and linear time-invariant stochastic differential equations. This class of models covers a wide range of both stationary and non-stationary GP models for encoding, for example, continuity, smoothness, or periodicity. The framework also extends to spatio-temporal models, where the GP is represented as an evolution type stochastic partial differential equation and inference conducted by infinite-dimensional Kalman filtering methods. Both separable and non-separable models are considered, and implementation techniques for numerical solutions are discussed.

The link between stochastic differential equations and standard covariance functions widens the applicability of Gaussian processes in combination with mechanistic physical differential equation models. Temporal and spatio-temporal Gaussian process models are useful in a multitude of data-intensive applications. Examples in this work include brain image analysis, weather modelling, financial forecasting, and tracking applications.

**Keywords** Stochastic differential equation, Gaussian process, state-space model, spatio-temporal data

**ISBN (printed)** 978-952-60-6710-0 **ISBN (pdf)** 978-952-60-6711-7

**ISSN-L** 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

**Location of publisher** Helsinki **Location of printing** Helsinki **Year** 2016

**Pages** 135 **urn** http://urn.fi/URN:ISBN:978-952-60-6711-7

A! Aalto-yliopisto

**Tiivistelmä**

**Aalto-yliopisto, PL 11000, 00076 Aalto** www.aalto.fi

**Tekijä**
Arno Solin

**Väitöskirjan nimi**
Stokastisia differentiaaliyhtälömenetelmiä spatiotemporaaliseen regressioon gaussisilla prosesseilla

**Julkaisija** Perustieteiden korkeakoulu

**Yksikkö** Neurotieteen ja lääketieteellisen tekniikan laitos / Tietotekniikan laitos

**Sarja** Aalto University publication series DOCTORAL DISSERTATIONS 50/2016

**Tutkimusala** Laskennallinen tiede

**Käsikirjoituksen pvm** 15.12.2015 **Väitöspäivä** 08.04.2016

**Julkaisuluvan myöntämispäivä** 02.02.2016 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

**Tiivistelmä**

Gaussiset prosessit (GP) ovat ei-parametrisia tilastollisia työkaluja, joita käytetään yleisesti koneoppimisessa, spatiaalisessa tilastotieteessä sekä signaalinkäsittelyssä. Niiden avulla mallifunktioita koskevat oletukset voidaan määritellä joustavasti. Gaussisen prosessin voi ajatella määrittelevän todennäköisyysjakauman funktioavaruuden yli. Signaalinkäsittelyssä gaussiset prosessit määritellään yleensä tila-avaruusmallin avulla, kun taas koneoppimisessa suositaan ytimeen (kovarianssifunktio) perustuvaa esitystapaa. Jälkimmäisessä esitystavassa GP-regressio-ongelman ratkaisun laskennallinen vaativuus skaalautuu kuutiollisesti aineiston koon suhteen, mikä tekee lähestymistavasta laskennallisesti raskaan suurilla tietoaineistoilla.

Tässä työssä tarkastellaan tila-avaruusmallien ja kovarianssifunktioiden yhteyttä, jonka avulla voidaan käyttää tilastollisessa päättelyssä tehokkaita vaiheittaisia Kalman-suotimeen perustuvia menetelmiä. Nämä menetelmät mahdollistavat lineaarisen laskennallisen skaalautuvuuden aineiston koon suhteen. Aluksi yhteys esitetään yksiulotteisille (temporaalisille) kovarianssifunktioille, jotka voidaan esittää lineaarisina aikainvariantteina stokastisina differentiaaliyhtälöinä. Käytetyt kovarianssifunktiot sisältävät sekä stationaarisia että ei-stationaarisia GP-malleja, joiden avulla mallioletuksiin voidaan sisällyttää esimerkiksi jatkuvuutta, sileyttä tai jaksollisuutta. Tämän jälkeen menetelmäkehystä laajennetaan spatiotemporaalisiin malleihin, joissa GP esitetään evoluutiotyyppisenä stokastisena osittaisdifferentiaaliyhtälönä ja päättely tehdään ääretönulotteista Kalman-suodinta käyttämällä. Työssä käsitellään separoituvia ja ei-separoituvia malleja sekä tarkastellaan laskennallisia toteutustapoja.

Stokastisten differentiaaliyhtälöiden ja yleisesti käytettyjen kovarianssifunktioiden yhteyttä voidaan käyttää myös GP-mallien yhdistämisessä mekanistisiin fysikaalisiin differentiaaliyhtälöihin. Temporaalisia ja spatiotemporaalisia gaussisia prosesseja voidaan käyttää monenlaisissa tietomääriltään suurissa sovelluksissa. Tässä työssä käytetään esimerkkeinä sovelluksia aivokuvannuksen, sään mallintamisen, markkinoiden ennustamisen sekä paikannuksen aloilta.

**Avainsanat** Stokastinen differentiaaliyhtälö, gaussinen prosessi, tila-avaruusmalli, spatiotemporaaliaineisto

**ISBN (painettu)** 978-952-60-6710-0 **ISBN (pdf)** 978-952-60-6711-7

**ISSN-L** 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

**Julkaisupaikka** Helsinki **Painopaikka** Helsinki **Vuosi** 2016

**Sivumäärä** 135 **urn** http://urn.fi/URN:ISBN:978-952-60-6711-7

# Preface

The research presented in this work has been carried out in the Bayesian Methodology group at Aalto University during years 2012–2016. The group was part of the Department of Biomedical Engineering and Computational Science (BECS), which was later on reorganized into the Department of Neuroscience and Biomedical Engineering and the Department of Computer Science. I gratefully acknowledge the financial support provided by the Academy of Finland and the Finnish foundation for Technology Promotion (TES).

I wish to express my gratitude to my supervisor Prof. Jouko Lampinen for providing support when needed, but leaving a lot of freedom to explore. I thank my thesis advisor Prof. Simo Särkkä for his constant enthusiasm and encouragement. He is one of the most efficient researchers I know, and we have been most productive together. I also thank Dr. Guido Sanguinetti and Dr. Finn Lindgren for pre-examining this thesis and providing insightful comments.

I am grateful for the help and expertise provided by my co-authors Dr. Jouni Hartikainen, Prof. Aki Vehtari, Prof. Fa-Hsuan Lin, Dr. Toni Auranen, Dr. Aapo Nummenmaa, and Dr. Simo Vanni.

Being a part of Prof. Vehtari's Bayes group at BECS and later on a member of the Probabilistic Machine Learning group at the Department of Computer Science has been integral for my everyday research. I wish to thank Prof. Vehtari and Prof. Samuel Kaski for providing an inspiring atmosphere for research.

During my years at Aalto University, I have had the privilege of working with many talented individuals. Various collaborations, sidetracks, and inspiring discussions have brought me together with Dr. Tomi Peltola, Juho Kokkala, Dr. Pasi Jylänki, Juho Piironen, Dr. Enrico Glerean, Dr. Jaakko Riihimäki, Dr. Mari Myllymäki, Ville Tolvanen, Janne Oja-

nen, Ville Väänänen, Olli-Pekka Koistinen, Dr. Tommi Mononen, Jukka Koskenranta, Christian Ritter, Manon Kok, Ernesto Ulloa, and Eric Malmi.

As a part of my doctoral studies, I spent a month as a visiting researcher in Prof. Neil Lawrence's group in Sheffield. I truly enjoyed my stay, and in addition to Prof. Lawrence I wish to thank Dr. James Hensman and Alan Saul for fruitful discussions and comments on my manuscripts.

I also thank my colleagues at IndoorAtlas Ltd., especially Dr. Esa Rahtu and Prof. Juho Kannala.

Family and friends have had an important role in supporting me throughout my studies. I thank my parents Merja and Olof for their encouragement and help. Finally and most importantly, I wish to thank my wife Kaisa for all her support during this project.

Helsinki, March 10, 2016,

Arno Solin

# Contents

# List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

**I** Simo Särkkä, Arno Solin, and Jouni Hartikainen. Spatiotemporal Learning via Infinite-Dimensional Bayesian Filtering and Smoothing. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

**II** Arno Solin and Simo Särkkä. Infinite-Dimensional Bayesian Filtering for Detection of Quasiperiodic Phenomena in Spatiotemporal Data. *Physical Review E*, 88(5):052909, 2013.

**III** Arno Solin and Simo Särkkä. Explicit Link Between Periodic Covariance Functions and State Space Models. In *Proceedings of the 17$^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR W&CP, volume 33, pages 904–912. Reykjavik, Iceland, 2014.

**IV** Arno Solin and Simo Särkkä. Gaussian Quadratures for State Space Approximation of Scale Mixtures of Squared Exponential Covariance Functions. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. Reims, France, 2014.

**V** Arno Solin and Simo Särkkä. State Space Methods for Efficient Inference in Student-$t$ Process Regression. In *Proceedings of the 18$^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS)*, JMLR W&CP, volume 38, pages 885–893. San Diego, CA, USA, 2015.

**VI** Simo Särkkä, Arno Solin, Aapo Nummenmaa, Aki Vehtari, Toni Auranen, Simo Vanni, and Fa-Hsuan Lin. Dynamic Retrospective Filtering of Physiological Noise in BOLD fMRI: DRIFTER. *NeuroImage*, 60(2):1517–1527, 2012.

# Author's Contribution

**Publication I: "Spatiotemporal Learning via Infinite-Dimensional Bayesian Filtering and Smoothing"**

Särkkä had the main responsibility in writing the article while Hartikainen and Solin designed the experiments and Solin derived the examples.

**Publication II: "Infinite-Dimensional Bayesian Filtering for Detection of Quasiperiodic Phenomena in Spatiotemporal Data"**

Solin had the main responsibility in writing the article. He also implemented all the models and methods used in the article as well as designed and ran the experiments.

**Publication III: "Explicit Link Between Periodic Covariance Functions and State Space Models"**

Solin had the main responsibility in writing the article. He also implemented all the models and methods used in the article as well as designed and ran the experiments.

**Publication IV: "Gaussian Quadratures for State Space Approximation of Scale Mixtures of Squared Exponential Covariance Functions"**

Solin had the main responsibility in writing the article. He also implemented all the models and methods used in the article as well as designed and ran the experiments.

### Publication V: "State Space Methods for Efficient Inference in Student-$t$ Process Regression"

Solin had the main responsibility in writing the article. He also implemented all the models and methods used in the article as well as designed and ran the experiments.

### Publication VI: "Dynamic Retrospective Filtering of Physiological Noise in BOLD fMRI: DRIFTER"

Särkkä had the main responsibility in writing the article while Nummenmaa and Vehtari provided technical expertise and Auranen, Vanni, and Lin neurological expertise. Solin implemented all the models and methods and ran the experiments.

# 1. Introduction

This thesis unites work from two distinctive fields in probabilistic modelling, namely time-series models from signal processing and Gaussian process based machine learning. The implications of this aim are twofold: to stimulate cross-talk between the fields by presenting a direct link between the notations, and to borrow efficient and established solution methods from one field to another.

The models and inference methods in this work are interpreted through a probabilistic viewpoint, which leans on the Bayesian methodology of modelling uncertainty as randomness and knowledge by assigning probability distributions to the modelled quantities (see, *e.g.*, Gelman *et al.*, 2013). Gaussian processes (O'Hagan, 1978) are prominent tools for probabilistic inference and learning. They generalize the concept of multivariate Gaussian distributions to infinite-dimensional spaces, where the quantities being modelled are functions and inference is conducted on their statistical properties.

This work focuses on Gaussian process models for phenomena which occur over space and time. Data of this kind is frequently encountered throughout fields of science, and a shared feature in this type of data is the tendency for the number of data points to be large in the temporal direction. In this type of modelling, the term 'space' should not be interpreted as restricted to geographical domains alone, but given a more general meaning in terms of arbitrary collections of observations at a specific snapshot of time.

This work aims at providing insights on the connection between spatio-temporal Gaussian process (regression) models and stochastic differential equations describing the evolution of the process over the temporal domain. For one-dimensional (temporal) processes this results in stochastic differential equation models driven by white noise. In the spatio-temporal

case they become evolution type stochastic partial differential equation models, which are driven by a spatial noise process. The motivation for forming this connection is that in the reformulated model the inference can be conducted sequentially with linear time complexity with respect to the number of distinct observations in the temporal domain. This is a remarkable boost compared to the cubic complexity associated with a batch solution of the same inference problem.

This thesis consists of Publications I–VI accompanied by this overview. The methodological contributions are covered in the original publications and the role of this overview is to provide a unifying view of their topics as a whole. Even though several types of applications are covered in the original publications, this overview concentrates on the methodological content and only refers to the publications for application examples.

The rest of this overview is structured as follows. Section 2 covers the necessary background related to Gaussian processes and batch learning. Section 3 explains the link between one-dimensional (temporal) Gaussian processes and state-space models given in terms of stochastic differential equations. The batch approach and sequential methods from the earlier sections are combined in Section 4 by writing the model as a stochastic partial differential equation. A summary of the publications is given as a part of the Discussion in Section 5, and finally concluding remarks are included in Section 6.

Source codes for the methods presented in this thesis are available online as part of several software packages: the GPstuff toolbox (Publications III–IV, `http://research.cs.aalto.fi/pml/software/gpstuff/`), the DRIFTER toolbox for modelling periodic noise in functional magnetic resonance imaging data (Publication VI, `http://becs.aalto.fi/en/research/bayes/drifter/`), and as supplementary material on the author's web page (`http://arno.solin.fi`).

# 2. Gaussian Processes in Machine Learning

In machine learning, Gaussian processes (GPs, O'Hagan, 1978; MacKay, 1998; Rasmussen and Williams, 2006) are a probabilistic paradigm for specifying probability distributions over functions. They are also a central part of several other disciplines, including signal processing, spatial statistics, and statistical inverse problems. The terminology varies between different fields of study; the 'process' viewpoint stems from signal processing, whereas the term Gaussian 'random field' is common in spatial statistics when the GP is two- or multi-dimensional.

A Gaussian process $f(\mathbf{x})$ is a random function in a $d$-dimensional input space. A way of defining a GP is through the following kernel (covariance function) formalism (Rasmussen and Williams, 2006). The process $f(\mathbf{x})$ is a Gaussian process on $\mathbb{R}^d$ with a mean function $\mu : \mathbb{R}^d \to \mathbb{R}$ and a covariance function (kernel) $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, which is denoted as

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')). \tag{2.1}$$

This notation implies that any finite collection of function values has a joint multivariate Gaussian distribution such that $(f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n)) \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{K})$, where $\mathbf{K}_{i,j} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ defines the covariance (Gram) matrix and $\boldsymbol{\mu}_i = \mu(\mathbf{x}_i)$ the mean vector, for $i, j = 1, 2, \ldots, n$.

The mean and covariance functions represent how the random functions behave on average and how the different points in the input space co-vary with respect to each other:

$$\begin{aligned} \mu(\mathbf{x}) &= \mathrm{E}[f(\mathbf{x})], \\ \kappa(\mathbf{x}, \mathbf{x}') &= \mathrm{E}[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))]. \end{aligned} \tag{2.2}$$

The covariance function thus encodes a correlation structure which introduces dependencies between function values at different inputs.

While a multivariate Gaussian distribution is concerned with a vector

(a collection of a finite number of elements), a Gaussian process is defined to deal with the continuum where the number of elements tends to infinity (the vector becoming an infinite-dimensional object, *i.e.*, a function). Thus the model functions $f(\mathbf{x}) \in \mathscr{H}(\mathbb{R}^d)$ are elements in an infinite-dimensional Hilbert space, and the finite indexing in the vector is replaced by the continuous-valued input variable $\mathbf{x}$. Even though the model is built upon continuous functions, data is finite and allows the computations to be performed on finite index sets. GP models are often referred to as 'non-parametric' because the number of parameters in the model is not fixed, but rather spanned by the number of data points.

## 2.1 Learning in Gaussian Processes

In Gaussian process based machine learning the interest is in the general framework for probabilistic inference in problems such as regression or classification. Given a set of data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ in form of input–output pairs, comprising the observations $y_i$ and the inputs (explanatory variable or covariate) $\mathbf{x}_i$, the task is to explain the hidden structure in the data or predict new observations. For notational convenience,s the set of inputs is denoted by $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ and the scalar outputs are combined into a vector $\mathbf{y} \in \mathbb{R}^n$.

GPs are used as non-parametric priors for the latent (unobserved) regression function, and the information given by the (often noise corrupted) data is propagated through a measurement model:

$$\begin{aligned} \text{GP prior:} \quad & f(\mathbf{x}) \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}')), \\ \text{Measurement model:} \quad & \mathbf{y} \mid \mathbf{f} \sim \prod_{i=1}^{n} p(y_i \mid f(\mathbf{x}_i)), \end{aligned} \tag{2.3}$$

where $\mathbf{f} \in \mathbb{R}^n$ denotes a vector of the latent values $f(\mathbf{x}_i)$, for $i = 1, 2, \ldots, n$.

The appeal of Gaussian process models comes from their flexibility and ease of encoding prior information into the model. The prior mean function is often, without loss of generality, fixed to zero; assumptions regarding the mean behaviour of the process can be encoded into the covariance function instead. The prior assumptions—such as continuity, smoothness, or periodicity—of the unknown functions are encoded into the covariance function, $\kappa_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}')$, and often parametrized (in terms of a characteristic length-scale, magnitude, period length, *etc.*) by a set of higher-level hyperparameters $\boldsymbol{\theta}$. In the spirit of hierarchical modelling, these parameters can then be assigned a hyperprior of their own, $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$. The same

applies to the parameters of the measurement model—such as the measurement noise variance.

The measurement model depends on the modelling task. The measurements are usually assumed conditionally independent given the latent function values $f(\mathbf{x})$ and thus the likelihood $p(\mathbf{y} \mid \mathbf{f})$ factorizes over the training data points.

Learning under the Gaussian process paradigm amounts to solving the posterior measure of $f(\mathbf{x})$ (distribution of $\mathbf{f}$) by combining the prior and likelihood by Bayes' rule:

$$p(\mathbf{f} \mid \mathcal{D}, \boldsymbol{\theta}) = \frac{p(\mathbf{f} \mid \mathcal{X}, \boldsymbol{\theta})\, p(\mathbf{y} \mid \mathbf{f})}{p(\mathbf{y} \mid \mathcal{X}, \boldsymbol{\theta})} = \frac{\mathrm{N}(\mathbf{f} \mid \mathbf{0}, \mathbf{K})}{p(\mathbf{y} \mid \mathcal{X}, \boldsymbol{\theta})} \prod_{i=1}^{n} p(y_i \mid f(\mathbf{x}_i)). \tag{2.4}$$

The normalization constant in the denominator is the marginal likelihood $p(\mathbf{y} \mid \mathcal{X}, \boldsymbol{\theta}) = \int p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f} \mid \mathcal{X}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{f}$ which provides a means of learning the hyperparameters $\boldsymbol{\theta}$ as will be explained in the later sections.

## 2.2 Computational Challenges and Special Structure

Their flexibility and the rather straight-forward probabilistic model building approach have made Gaussian processes popular for performing Bayesian inference on unknown latent functions. However, this generality and elegance come with some drawbacks. First, naïve implementations of the inference solution scale cubically in the number of training data points, $n$, which renders the modelling methodology challenging for large data sets. Second, the inference problem in Equation (2.4) has an analytically tractable form only for a Gaussian measurement model.

Both of these problems have been extensively studied in the Gaussian process community. Several approximation schemes tackling the intractable inference have been proposed. Typically they are either based on sampling methods such as Markov chain Monte Carlo (see, *e.g.*, Neal, 1998; Murray *et al.*, 2010) or analytical Gaussian approximations such as the Laplace approximation (see, *e.g.*, Williams and Barber, 1998; Rasmussen and Williams, 2006), expectation propagation (EP, see Minka, 2001), or variational methods (VB, see, *e.g.*, Gibbs and MacKay, 2000; Csató *et al.*, 2000). Even though Gaussian processes in machine learning are mostly concerned with inference and learning, this work concentrates on the representation problem for modelling the latent function. The focus will thus be on Gaussian measurement models (*i.e.* GP regression problems). The only exception is the mining disaster experiment in

Publication IV, where a Poisson likelihood is used in a log-Gaussian Cox process model and the approximate inference is conducted by the Laplace approximation.

Methods dealing with the computational complexity have gained a lot of attention. There exists a wide range of so-called sparse approximations which replace the full covariance matrix by a lower $m$-rank representation. This reduces the storage requirements from $\mathcal{O}(n^2)$ to $\mathcal{O}(nm)$ and the computational complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(nm^2)$ (Quiñonero-Candela and Rasmussen, 2005; Rasmussen and Williams, 2006; Smola and Bartlett, 2001; Csató and Opper, 2002; Seeger *et al.*, 2003; Titsias, 2009; Snelson and Ghahramani, 2006). These methods are typically based on the Nyström approximation (Baker, 1977; Williams and Seeger, 2001). Larger data sets can be covered by stochastic inference methods (Hensman *et al.*, 2013).

Special structure in data or the covariance function can be used for easing the computational burden. The following concepts are important in forthcoming sections, and each of them comes naturally to the 'process' point of view in Section 3.

**Stationarity** If the covariance function is *stationary* (homogeneous), $\kappa(\mathbf{x}, \mathbf{x}')$ only depends on the difference $\mathbf{x} - \mathbf{x}'$. This means that the covariance is translation invariant and thus the same regardless of the position in the input space (*cf.* Rasmussen and Williams, 2006). For stationary covariance functions the following one-input notation is used: $\kappa(\mathbf{x}, \mathbf{x}') \triangleq \kappa(\mathbf{r})$, where $\mathbf{r} = \mathbf{x} - \mathbf{x}'$.

Stationary models can be represented in terms of their spectral density. This arises from *Bochner's theorem* (see, *e.g.*, Akhiezer and Glazman, 1993; Da Prato and Zabczyk, 1992) which states that an arbitrary positive definite function $\kappa(\mathbf{r})$ can be represented as a Fourier transform of a positive measure. If the measure has a density, it is called the *spectral density* $S(\boldsymbol{\omega})$ corresponding to the covariance function $\kappa(\mathbf{r})$. This relation is the Fourier duality of covariance and spectral density, which is known as the *Wiener–Khintchin theorem* (see, *e.g.*, Rasmussen and Williams, 2006). It gives the identities

$$\kappa(\mathbf{r}) = \frac{1}{(2\pi)^d} \int S(\boldsymbol{\omega})\, e^{\mathrm{i}\, \boldsymbol{\omega}^\mathsf{T} \mathbf{r}}\, \mathrm{d}\boldsymbol{\omega} \quad \text{and} \quad S(\boldsymbol{\omega}) = \int \kappa(\mathbf{r})\, e^{-\mathrm{i}\, \boldsymbol{\omega}^\mathsf{T} \mathbf{r}}\, \mathrm{d}\mathbf{r}. \tag{2.5}$$

**Isotropy** If the value of the covariance function only depends on the distance (Euclidean norm) between the inputs, $\kappa(\mathbf{x}, \mathbf{x}') \triangleq \kappa(\|\mathbf{x} - \mathbf{x}'\|)$,

it is said to be *isotropic*. This means that the covariance is both translation and rotation invariant.

**Separability** The kernel is said to be *separable* if it can be rewritten as a product of two kernels with no shared inputs. For example, if $\mathbf{x} = (x_1, x_2)$, the kernel $\kappa(\mathbf{x}, \mathbf{x}') = \kappa_1(x_1, x_1')\,\kappa_2(x_2, x_2')$ is separable.

The structure might also stem from the input data. If the inputs are points on some regular grid with respect to one or several input dimensions, there exists a range of methods for combining the data with stationary or separable models. For stationary models, fast Fourier transform (FFT) methods (see, *e.g.*, Paciorek, 2007; Fritz *et al.*, 2009) can speed up the computations, or if the model is separable, Kronecker products can help reformulate the model (see, *e.g.*, Saatçi, 2012).

Low-dimensional inputs, where $d \leq 3$, are of interest in geostatistics (where GPs are known as kriging models, see, *e.g.*, Cressie, 1993) and physical inverse problems (Kaipio and Somersalo, 2005). Recent interest has been in utilizing efficient methods for Gauss–Markov random fields (Rue *et al.*, 2009; Lindgren *et al.*, 2011).

In signal processing, the interest is typically in models where the inputs are one-dimensional, $x \in \mathbb{R}$. If the data are points in time, this is regarded a time-series model (see, *e.g.*, Box *et al.*, 2008). The time-ordered structure in the data can be utilized in a sequential Gauss–Markov model also known as a *state-space* model. The analytical link between state-space models and Gaussian processes was already discussed by O'Hagan (1978), but has recently been vivified independently by both Hartikainen and Särkkä (2010) and Reece and Roberts (2010). The benefit is that fast sequential inference methods from signal processing can be used for solving the inference in linear time. This link will be the central theme of Section 3.

### 2.3 Batch Gaussian Process Regression

In GP regression (Rasmussen and Williams, 2006) the interest is in predicting an unknown scalar output $f(\mathbf{x}_\star)$ associated with a known input $\mathbf{x}_\star \in \mathbb{R}^d$, given a training data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. The model functions $f(\mathbf{x})$ are assumed to be realizations of a Gaussian random process prior and the observations corrupted by independent and identically dis-

tributed (i.i.d.) Gaussian noise

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}')), \\ y_i &\sim \mathrm{N}(y_i \mid f(\mathbf{x}_i), \sigma^2_{\text{noise}}). \end{aligned} \tag{2.6}$$

To ease the notation in the following sections, the equation below introduces a linear operator $\mathcal{H}$. It evaluates the latent function values $f(\mathbf{x})$ at the observation inputs such that

$$\mathbf{f} = \mathcal{H}\, f(x) = (f(\mathbf{x}_1) \quad f(\mathbf{x}_2) \quad \cdots \quad f(\mathbf{x}_n))^\mathsf{T}. \tag{2.7}$$

The operator is thus actually a functional from the function space to a finite vector space. $\mathcal{H}$ could be replaced with some more general linear operator (Särkkä, 2011), and the framework could used in solving arbitrary linear inverse problems (Kaipio and Somersalo, 2005). The GP regression problem can now be written as

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}')), \\ \mathbf{y} &= \mathcal{H}\, f(\mathbf{x}) + \boldsymbol{\varepsilon}, \end{aligned} \tag{2.8}$$

where $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2_{\text{noise}}\,\mathbf{I})$ is Gaussian measurement noise. Because both the prior (GP) and likelihood are Gaussian, the inference problem has an analytic solution. The joint distribution of $\mathbf{f}$ and the function value $f(\mathbf{x}_\star)$ at the test input, $p(\mathbf{f}, f(\mathbf{x}_\star) \mid \mathcal{X}, \mathbf{x}_\star, \boldsymbol{\theta})$, is also Gaussian. By using the conditioning properties of the multivariate Gaussian distribution, the conditional distribution for $f(\mathbf{x}_\star)$ given $\mathbf{f}$ can be written down analytically.

The conditional distribution $p(f(\mathbf{x}_\star) \mid \mathbf{f}, \mathcal{X}, \mathbf{x}_\star, \boldsymbol{\theta})$ still remains Gaussian. Integrating over the latent values $\mathbf{f}$ in the joint posterior distribution results in the posterior predictive distribution for $f(\mathbf{x}_\star)$:

$$p(f(\mathbf{x}_\star) \mid \mathcal{X}, \mathbf{y}, \mathbf{x}_\star, \boldsymbol{\theta}) = \int p(\mathbf{f}_\star \mid \mathbf{f}, \mathcal{X}, \mathbf{x}_\star, \boldsymbol{\theta})\, p(\mathbf{f} \mid \mathcal{X}, \mathbf{y}, \boldsymbol{\theta})\, \mathrm{d}\mathbf{f}. \tag{2.9}$$

By naïvely writing down the matrix equations, the solution to the GP regression problem $p(f(\mathbf{x}_\star) \mid \mathcal{D}, \mathbf{x}_\star, \boldsymbol{\theta}) = \mathrm{N}(\mathrm{E}[f(\mathbf{x}_\star)], \mathrm{V}[f(\mathbf{x}_\star)])$ can be computed in closed-form as (Rasmussen and Williams, 2006)

$$\begin{aligned} \mathrm{E}[f(\mathbf{x}_\star)] &= \mathbf{k}_\star^\mathsf{T} (\mathbf{K} + \sigma^2_{\text{noise}}\,\mathbf{I})^{-1} \mathbf{y}, \\ \mathrm{V}[f(\mathbf{x}_\star)] &= \kappa(\mathbf{x}_\star, \mathbf{x}_\star) - \mathbf{k}_\star^\mathsf{T} (\mathbf{K} + \sigma^2_{\text{noise}}\,\mathbf{I})^{-1} \mathbf{k}_\star, \end{aligned} \tag{2.10}$$

where $\mathbf{K}_{i,j} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}_\star$ is an $n$-dimensional vector with the $i$th entry being $\kappa(\mathbf{x}_\star, \mathbf{x}_i)$, and $\mathbf{y}$ is a vector of the $n$ observations. The cubic computational complexity comes from the $n \times n$ matrix inversion in (2.10).

The predictive distribution $p(y_\star \mid \mathcal{D}, \mathbf{x}_\star, \boldsymbol{\theta})$ of yet unseen new observations $y_\star$ can be obtained by integrating over the uncertainty related to

$f(\mathbf{x}_\star)$. Because the observation model is Gaussian, the expected value of the posterior predictive distribution remains the same as in (2.10) and the variance becomes $\mathrm{V}[f(\mathbf{x}_\star)] + \sigma^2_{\mathrm{noise}}$.

A probabilistic way of learning the hyperparameters $\boldsymbol{\theta}$ of the covariance function and the noise variance $\sigma^2_{\mathrm{noise}}$ is by maximizing the marginal likelihood function $p(\mathbf{y} \mid \mathcal{X}, \boldsymbol{\theta}, \sigma^2_{\mathrm{noise}})$ (see, *e.g.*, Rasmussen and Williams, 2006). In practice, the target function is usually the logarithm of the marginal likelihood function

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2_{\mathrm{noise}}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2_{\mathrm{noise}}\mathbf{I}| - \frac{1}{2}\mathbf{y}^{\mathsf{T}}(\mathbf{K}_{\boldsymbol{\theta}} + \sigma^2_{\mathrm{noise}}\mathbf{I})^{-1}\mathbf{y}. \quad (2.11)$$

A more principled Bayesian approach would require the prior distributions of the hyperparameters to be taken into account as well. A maximum *a posteriori* (MAP) estimation approach is to include them in the optimization target function.

Gaussian process regression can also be interpreted as a Bayesian viewpoint on regularization, where the inference corresponds to a minimization problem written in terms of the reproducing kernel Hilbert space (RKHS) norm associated with the covariance function (see, *e.g.*, Wahba, 1990; Berlinet and Thomas-Agnan, 2004).

# 3. Stochastic Differential Equation Representation of GPs

Despite its generality, computationally the batch GP methodology does not conveniently lend itself to long (or unbounded) time-series. This sort of data is typical in applications where the number of observations keeps growing as new data is acquired. Constraining the dimensionality of the model inputs $\mathrm{x}$ to one-dimensional, $\mathrm{x} \triangleq t \in \mathbb{R}$, makes the GP a stochastic process $f(t)$. For this class of Gaussian processes, instead of considering the kernel formalism, it is appealing to work with its mathematical dual, where the models are written out in terms of dynamical systems.

In signal processing and control engineering, a general framework for modelling dynamical processes is known as state-space methods (see, *e.g.*, Jazwinski, 1970). In a general sense, state-space models are concerned with modelling the states $\mathbf{f}(t_k) \triangleq \mathbf{f}_k, k = 1, 2, \ldots, n$, of a system at time instants $t_1, t_2, \ldots, t_n$ with

$$\begin{aligned} \mathbf{f}_k &\sim p(\mathbf{f}_k \mid \mathbf{f}_{k-1}), \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{f}_k), \end{aligned} \tag{3.1}$$

where $\mathbf{f}_k \in \mathbb{R}^m$ is the latent (unobserved) process at time instance $t_k$ and $\mathbf{y}_k$ is the measurement (observation) at the corresponding time step. The prior assumptions about the dynamical behaviour are encoded into the Markov model for the conditional state transition probability $p(\mathbf{f}_k \mid \mathbf{f}_{k-1})$. The measurement model $p(\mathbf{y}_k \mid \mathbf{f}_k)$ is defined in terms of the conditional relationship of the observations given the latent state of the system. The initial state is given by the prior $\mathbf{f}_0 \sim p(\mathbf{f}_0)$.

In the following sections a class of linear time-invariant (LTI) models for representing one-dimensional Gaussian processes is first defined. The substance is in giving an explicit model formulation for a wide range of standard GP models in terms of linear time-invariant continuous-time systems (relating to Publications I, III–V), and defining relations for combining models of this kind (Publication III). Methods for sequentially solv-

ing the GP regression problem are also presented. This modelling framework allows for a generalization of Gaussian processes to Student-$t$ processes (Publication V), which will be briefly covered at the end of this section.

### 3.1 Temporal GPs as Stochastic Differential Equations

For temporal Gaussian process regression problems, certain classes of covariance functions allow working with the following dual representation, where the Gaussian process is given in terms of a solution to an $m$th order linear stochastic differential equation (SDE, see, *e.g.*, Øksendal, 2003). For example, consider the linear differential equation

$$a_0\, f(t) + a_1 \frac{\mathrm{d}f(t)}{\mathrm{d}t} + \cdots + a_m \frac{\mathrm{d}^m f(t)}{\mathrm{d}t^m} = w(t) \tag{3.2}$$

driven by a zero-mean continuous-time (stochastic) white noise process $w(t)$. Every solution trajectory $f(t)$ of this stochastic differential equation is a random function—a stochastic process. The driving term $w(t)$ is the formal derivative of Brownian motion (also known as the Wiener process). Because $w(t)$ is a Gaussian process and Gaussianity is preserved under linear operations, the solution trajectory will also be a Gaussian process.

The higher-order linear SDE in Equation (3.2) can be rewritten as a first-order stochastic differential equation by defining a vector-valued function $\mathbf{f}(t) = (f(t),\, \mathrm{d}/\mathrm{d}t f(t), \ldots,\, \mathrm{d}^{m-1}/\mathrm{d}t^{m-1} f(t))$. Collecting the derivative terms gives a representation of the SDE known as the *companion form* (Grewal and Andrews, 2001)

$$\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & -a_1 & \ldots & -a_{m-1} \end{pmatrix} \mathbf{f}(t) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} w(t). \tag{3.3}$$

This model defines the coupled dynamics of $m$ Gaussian processes. However, the interest is in the scalar first component $f(t)$ of the state variable $\mathbf{f}(t) : \mathbb{R} \to \mathbb{R}^m$. Extracting this process from the state is a linear operation. The corresponding measurement model is

$$f(t) = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \mathbf{f}(t), \tag{3.4}$$

which will be denoted by $f(t) = \mathbf{H}\,\mathbf{f}(t)$.

The companion form is not the only way to represent the latent dynamics of $f(t)$. The state variables do not necessarily need to be the actual

time derivatives, but can be more general coupled processes. Furthermore, the class of linear time-invariant SDEs can be slightly generalized by allowing the driving white noise process to be multi-dimensional.

In the following sections the interest will be in classes of covariance functions which can be represented in terms of a dynamical model (the GP prior) and measurement model of the form:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} &= \mathbf{F}\,\mathbf{f}(t) + \mathbf{L}\,\mathbf{w}(t), \\ y_k &= \mathbf{H}\,\mathbf{f}(t_k) + \varepsilon_k, \end{aligned} \tag{3.5}$$

where $\mathbf{f}(t) = (f_1(t),\ f_2(t),\ \ldots,\ f_m(t))$ contains the $m$ stochastic processes, and $\mathbf{w}(t) \in \mathbb{R}^s$ is a multi-dimensional white noise process with spectral density matrix $\mathbf{Q}_\mathrm{c} \in \mathbb{R}^{s\times s}$. The model is defined by the feedback matrix $\mathbf{F} \in \mathbb{R}^{m\times m}$, the noise effect matrix $\mathbf{L} \in \mathbb{R}^{m\times s}$, and the initial state covariance $\mathbf{P}_0$. The measurements are assumed to be corrupted by i.i.d. Gaussian noise, $\varepsilon_k \sim \mathrm{N}(0, \sigma^2_\mathrm{noise})$.

The dynamical model in Equation (3.5) is an Itô stochastic differential equation written in terms of a driving white noise. In literature, the shorthand integral notation, $\mathrm{d}\mathbf{f}(t) = \mathbf{F}\,\mathbf{f}(t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}(t)$, is often used (see, *e.g.*, Karatzas and Shreve, 1991; Øksendal, 2003). The intuitive link between these two notations is that the white noise term $\mathbf{w}(t)$ can be seen as the formal derivative of the Brownian motion term $\mathrm{d}\boldsymbol{\beta}(t)$.

The form of continuous-time SDE models defined by (3.5) is well-suited for model specification, but for numerical implementations the model needs to deal with discrete-time inputs. In order to apply standard signal processing filtering and smoothing techniques (*cf.* Särkkä, 2013), the SDE model must be solved for the discrete time-instances corresponding to the input data and prediction points.

For discrete state values $\mathbf{f}_k = \mathbf{f}(t_k)$ the state-space model is

$$\begin{aligned} \mathbf{f}_k &\sim \mathrm{N}(\mathbf{f}_k \mid \mathbf{A}_{k-1}\,\mathbf{f}_{k-1}, \mathbf{Q}_{k-1}), \\ y_k &\sim \mathrm{N}(y_k \mid \mathbf{H}\,\mathbf{f}_k, \sigma^2_\mathrm{noise}), \end{aligned} \tag{3.6}$$

where $\mathbf{A}_k$ is the discrete-time state transition matrix between time-instance $t_k$ and $t_{k+1}$, and $\mathbf{Q}_k$ is the process noise covariance matrix. The model can be equivalently presented in the following form:

$$\begin{aligned} \mathbf{f}_k &= \mathbf{A}_{k-1}\,\mathbf{f}_{k-1} + \mathbf{q}_{k-1}, \quad \text{where} \quad \mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}_{k-1}), \\ y_k &= \mathbf{H}\,\mathbf{f}_k + \varepsilon_k, \qquad\qquad\ \ \text{where} \quad \varepsilon_k \sim \mathrm{N}(0, \sigma^2_\mathrm{noise}). \end{aligned} \tag{3.7}$$

The iteration is started from the initial state $\mathbf{f}_0 \sim \mathrm{N}(\mathbf{0}, \mathbf{P}_0)$. In general,

the discrete-time model matrices are given by

$$\mathbf{A}_k = \mathbf{\Phi}(\Delta t_k), \tag{3.8}$$

$$\mathbf{Q}_k = \int_0^{\Delta t_k} \mathbf{\Phi}(\Delta t_k - \tau)\, \mathbf{L}\, \mathbf{Q}_\mathrm{c}\, \mathbf{L}^\mathsf{T}\, \mathbf{\Phi}(\Delta t_k - \tau)^\mathsf{T}\, \mathrm{d}\tau, \tag{3.9}$$

where $\Delta t_k = t_{k+1} - t_k$ and $\mathbf{\Phi}(\tau) = \exp(\mathbf{F}\tau)$ is the matrix exponential of the feedback matrix. The integral (3.9) can be solved using the matrix fraction decomposition (see, *e.g.*, Särkkä, 2006, for implementation details).

If the model (and the corresponding covariance function) is stationary, the SDE has a stationary state $\mathbf{f}_\infty \sim \mathrm{N}(\mathbf{0}, \mathbf{P}_\infty)$. The stationary state corresponds to the state that the model stabilises to in infinity. It can be represented by the stationary covariance of $\mathbf{f}(t)$ that is the solution to

$$\frac{\mathrm{d}\mathbf{P}_\infty}{\mathrm{d}t} = \mathbf{F}\,\mathbf{P}_\infty + \mathbf{P}_\infty\,\mathbf{F}^\mathsf{T} + \mathbf{L}\,\mathbf{Q}_\mathrm{c}\,\mathbf{L}^\mathsf{T} = \mathbf{0}. \tag{3.10}$$

This matrix equation is known as the Lyapunov equation (a special case of the continuous-time Riccati equation).

The stationary state is invariant to the choice of input location and describes the state the process defaults to. Therefore, for stationary models the initial (prior) state is given by the stationary state covariance, $\mathbf{P}_0 := \mathbf{P}_\infty$. Provided that $\mathbf{P}_\infty$ exists and is known, the following relation is a computationally lightweight way of solving the integral (3.9) and obtaining the process noise covariance (an idea originally suggested by Davison and Man, 1968):

$$\mathbf{Q}_k = \mathbf{P}_\infty - \mathbf{A}_k\,\mathbf{P}_\infty\,\mathbf{A}_k^\mathsf{T}. \tag{3.11}$$

For these models (see Van Trees, 1971; Särkkä and Solin, 2014), the spectral density function $S(\omega)$ of the Gaussian process $f(t)$ can be written using the state-space representation as

$$S(\omega) = \mathbf{H}\,(\mathbf{F} - \mathrm{i}\,\omega\mathbf{I})^{-1}\,\mathbf{L}\,\mathbf{Q}_\mathrm{c}\,\mathbf{L}^\mathsf{T}\,\left[(\mathbf{F} + \mathrm{i}\,\omega\mathbf{I})^{-1}\right]^\mathsf{T}\,\mathbf{H}^\mathsf{T}. \tag{3.12}$$

The covariance function $\kappa(\tau)$ of $f(t)$ is the inverse Fourier transform of its spectral density, which in terms of the state-space matrices is

$$\kappa(\tau) = \begin{cases} \mathbf{H}\,\mathbf{P}_\infty\,\mathbf{\Phi}(\tau)^\mathsf{T}\,\mathbf{H}^\mathsf{T}, & \text{if } \tau \geq 0, \\ \mathbf{H}\,\mathbf{\Phi}(-\tau)\,\mathbf{P}_\infty\,\mathbf{H}^\mathsf{T}, & \text{if } \tau < 0, \end{cases} \tag{3.13}$$

where $\mathbf{\Phi}(\tau) = \exp(\mathbf{F}\tau)$ is the matrix exponential of the feedback matrix.

### 3.2 Sums and Products of Covariance Functions

The sum $\kappa(t,t') = \kappa_1(t,t') + \kappa_2(t,t')$ of two covariance functions is a valid covariance function (see, *e.g.*, Rasmussen and Williams, 2006). In terms of the state-space SDE model this corresponds to a stacked model $\mathbf{f}(t) = (\mathbf{f}_1(t), \mathbf{f}_2(t))$ of the two SDE models:

$$\begin{aligned} \mathbf{F} &= \mathrm{blkdiag}(\mathbf{F}_1, \mathbf{F}_2), & \mathbf{L} &= \mathrm{blkdiag}(\mathbf{L}_1, \mathbf{L}_2), \\ \mathbf{Q}_\mathrm{c} &= \mathrm{blkdiag}(\mathbf{Q}_{\mathrm{c},1}, \mathbf{Q}_{\mathrm{c},2}), & \mathbf{P}_0 &= \mathrm{blkdiag}(\mathbf{P}_{0,1}, \mathbf{P}_{0,2}), \end{aligned} \tag{3.14}$$

where $\mathrm{blkdiag}(\cdot,\cdot)$ defines a block-diagonal matrix, and the indices '1' and '2' refer to $\kappa_1(t,t')$ and $\kappa_2(t,t')$, respectively. The dynamics of the two state-space models are independent, but coupled by the measurement model $\mathbf{H} = (\mathbf{H}_1 \; \mathbf{H}_2)$. The state dimension becomes $m = m_1 + m_2$.

The product $\kappa(t,t') = \kappa_1(t,t')\,\kappa_2(t,t')$ of two covariance functions is also a valid covariance function (see, *e.g.*, Rasmussen and Williams, 2006). The corresponding product of two state-space SDEs must preserve the dynamics of both the models. In discrete-time, this would require the transition matrix to be $\mathbf{A}_k = \exp(\mathbf{F}_1 \Delta t_k) \exp(\mathbf{F}_2 \Delta t_k) = \exp(\mathbf{F} \Delta t_k)$. In order to factorize, the feedback matrices need to commute ($\mathbf{F}_1\,\mathbf{F}_2 = \mathbf{F}_2\,\mathbf{F}_1$) which ensures that the matrices preserve each others eigenspaces. The relation holds if the feedback is constructed as follows (Publication III):

$$\mathbf{F} = \mathbf{F}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{F}_2, \tag{3.15}$$

where '$\otimes$' denotes the Kronecker product of two matrices. This equation is known as the Kronecker sum of matrices $\mathbf{F}_1$ and $\mathbf{F}_2$, which makes the matrix exponential factor into the Kronecker product of the corresponding two matrix exponentials (see Higham, 2008, and Publication III). The noise effect, spectral density, initial state covariance, and measurement model matrices are then given as follows.

$$\begin{aligned} \mathbf{L} &= \mathbf{L}_1 \otimes \mathbf{L}_2, & \mathbf{Q}_\mathrm{c} &= \mathbf{Q}_{\mathrm{c},1} \otimes \mathbf{Q}_{\mathrm{c},2}, \\ \mathbf{P}_0 &= \mathbf{P}_{0,1} \otimes \mathbf{P}_{0,2}, & \mathbf{H} &= \mathbf{H}_1 \otimes \mathbf{H}_2. \end{aligned} \tag{3.16}$$

The state dimension becomes $m = m_1\,m_2$.

### 3.3 Covariance Functions as State-Space Models

Recall that the GP prior in Equation (2.8) in Section 2.3 was fully characterized by its covariance function. For one-dimensional problems, the

SDE representation of the Gaussian process in Equation (3.5) is defined by the model matrices $\mathbf{F}$, $\mathbf{L}$, $\mathbf{Q}_\mathrm{c}$, $\mathbf{P}_0$, and $\mathbf{H}$.

The connections between covariance functions and the corresponding SDE model matrices are summarized in the following subsections. They contain a non-exhaustive collection of SDE representations of common covariance functions. All these covariance functions either have an exact linear time-invariant state-space SDE representation or they can be approximated with such a model. Some of these models are degenerate (the corresponding covariance matrix has finite rank), some of the models are stationary and fall under the rational spectrum approach (Publication I), others obey some special structure (Publication III), or are mixture models (Publication IV). Figure 3.1 shows ten random realizations from twelve different GP priors, each of which has been simulated using the SDE representation. The parametrization of the covariance functions will follow the standard notation from Rasmussen and Williams (2006).

In the following, stationary covariance functions are given in terms of one input, such that $\kappa(t, t') = \kappa(\tau)$, where $\tau = |t - t'|$. The relation tends to hold for several input dimensions as well; multi-dimensional isotropic extensions of the covariance functions can be considered in the form $\kappa(\mathbf{x}, \mathbf{x}') = \kappa(\tau)$, where $\tau = \|\mathbf{x} - \mathbf{x}'\|$. These models are discussed further on in this work.

### Constant

Constant bias in the outputs can be modelled by including a constant kernel in the covariance structure of the model. The constant covariance function is given as

$$\kappa_{\mathrm{const.}}(t, t') = \sigma^2, \tag{3.17}$$

where $\sigma^2$ is a magnitude scale hyperparameter. Random draws from a GP with this covariance function correspond to constant functions. The covariance function is degenerate, meaning it has a finite rank one.

The corresponding state-space SDE can be given as follows: the feedback $F = 0$, noise effect $L = 1$, spectral density $Q_\mathrm{c} = 0$, observation model $H = 1$, and initial state variance $P_0 = \sigma^2$.

### Linear

A way of doing standard linear regression using GPs is to use a linear covariance function in the GP regression model. Both the constant and

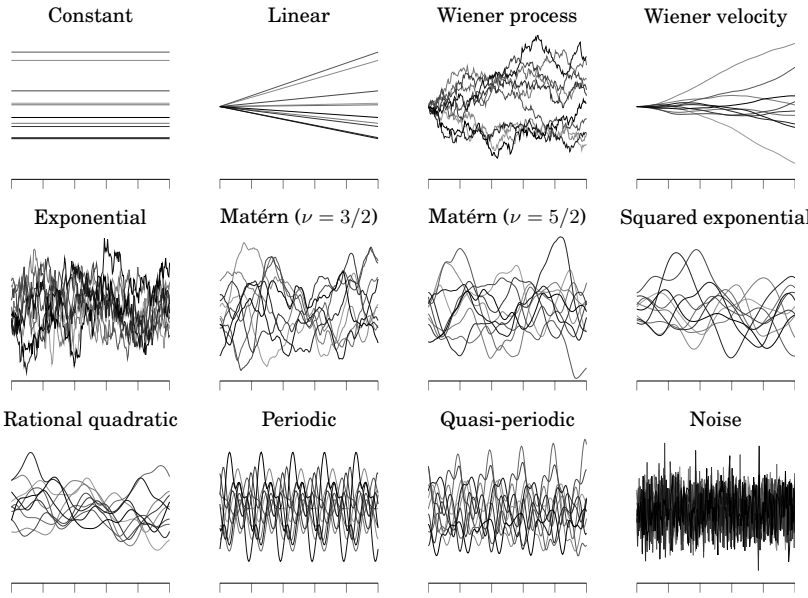


**Figure 3.1.** Random processes simulated from Gaussian process models using their state-space representations. The hyperparameters were fixed to unit values. The vertical and horizontal scales are arbitrary as they are subject to the choice of hyperparameter values.

linear covariance functions can be useful components in a more complicated GP model with sums and products of several covariance functions. The linear covariance function is non-stationary and defined by

$$\kappa_{\text{linear}}(t, t') = \sigma^2 \, t \, t', \tag{3.18}$$

where $\sigma^2$ is a magnitude scale hyperparameter. The corresponding state-space SDE can be given as follows. The feedback, noise effect, and initial state covariance matrices are

$$\mathbf{F} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{and} \quad \mathbf{P}_0 = \sigma^2 \begin{pmatrix} t_0^2 & t_0 \\ t_0 & 1 \end{pmatrix}. \tag{3.19}$$

The spectral density $Q_\text{c} = 0$ and observation model $\mathbf{H} = (1\,0)$. Similar constructions can be done for other (degenerate) models with finite rank by setting up a model for the evolution of the corresponding basis functions.

### Wiener process and Wiener velocity models

The Wiener process (Brownian motion) is a fundamental non-stationary process which is widely used in stochastic modelling. It has the covariance

function (see, *e.g.*, Rasmussen and Williams, 2006)

$$\kappa_{\mathrm{WP}}(t, t') = \sigma^2 \min(t, t') \tag{3.20}$$

which is valid on the domain $t, t' \geq 0$. The corresponding state-space model is trivial: $F = 0$, $L = 1$, $Q_{\mathrm{c}} = \sigma^2$, $H = 1$, and $P_0 = 0$.

The Wiener velocity model is a commonly used model in tracking applications (see Bar-Shalom *et al.*, 2001, or Publication V for an example). It corresponds to a once integrated Wiener process, where the derivative (velocity) is modelled as a Wiener process. The covariance function can be given as

$$\kappa_{\mathrm{WV}}(t, t') = \sigma^2 \left( \frac{\min^3(t, t')}{3} + |t - t'| \frac{\min^2(t, t')}{2} \right) \tag{3.21}$$

and the corresponding state-space model feedback and noise effect as

$$\mathbf{F} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{L} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{3.22}$$

The process noise spectral density is controlled by $Q_{\mathrm{c}} = \sigma^2$ and $\mathbf{P}_0 = \mathbf{0}$. The measurement model equals $\mathbf{H} = (1 \ 0)$. In machine learning, higher-order integrated Wiener processes have recently been considered by Schober *et al.* (2014).

### Exponential

The exponential covariance function defines a stationary process, where the model functions are continuous but not differentiable. The covariance function is (Rasmussen and Williams, 2006)

$$\kappa_{\exp}(\tau) = \sigma^2 \exp\left( -\frac{\tau}{\ell} \right), \tag{3.23}$$

where $\sigma^2$ is a magnitude scale parameter and $\ell$ the characteristic length-scale parameter. The exponential covariance function is the covariance function of the so-called *Ornstein–Uhlenbeck process*.

The link to the Ornstein–Uhlenbeck process provides a direct representation in terms of an SDE model (see the discussion in Hartikainen and Särkkä, 2010, and Publication I). The SDE model is given as follows: the feedback $F = -1/\ell$, the noise effect $L = 1$, the spectral density of the white noise is $Q_{\mathrm{c}} = 2\sigma^2/\ell$, and the measurement model $H = 1$. The variance in the stationary state is $P_\infty = \sigma^2$.

### Matérn

A wide class of covariance functions is defined by the Matérn class (Matérn, 1960) which is given in its general form as

$$\kappa_{\text{Mat.}}(\tau) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu}\,\tau}{\ell} \right)^{\nu} \mathrm{K}_{\nu}\left( \frac{\sqrt{2\nu}\,\tau}{\ell} \right), \tag{3.24}$$

where $\sigma^2$ is a magnitude scale hyperparameter, $\ell$ the characteristic length-scale, and $\nu$ a smoothness parameter. $\mathrm{K}_{\nu}(\cdot)$ denotes the modified Bessel function of the second kind (Abramowitz and Stegun, 1970). For this class, the corresponding process is $k$-times differentiable if $\nu > k$ (Rasmussen and Williams, 2006)

The covariance function becomes simple for half-integer values of $\nu$. If $\nu = 1/2$, the exponential covariance function is recovered. For one-dimensional models the half-integer Matérn covariance functions have an exact SDE representation (see, *e.g.*, Hartikainen and Särkkä, 2010). For example, if $\nu = 3/2$, the processes are continuous and once differentiable. The covariance function becomes

$$\kappa_{\text{Mat.}}(\tau) = \sigma^2 \left( 1 + \frac{\sqrt{3}\,\tau}{\ell} \right) \exp\left( - \frac{\sqrt{3}\,\tau}{\ell} \right), \tag{3.25}$$

and has the SDE representation (Hartikainen and Särkkä, 2010)

$$\mathbf{F} = \begin{pmatrix} 0 & 1 \\ -\lambda^2 & -2\lambda \end{pmatrix}, \quad \mathbf{L} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{and} \quad \mathbf{P}_{\infty} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \lambda^2\sigma^2 \end{pmatrix}, \tag{3.26}$$

where $\lambda = \sqrt{3}/\ell$. The spectral density of the Gaussian white noise process $w(t)$ is $Q_{\text{c}} = 4\lambda^3\sigma^2$. The measurement model matrix is $\mathbf{H} = (1\ 0)$.

Higher-order half-integer SDE representations are given by Hartikainen and Särkkä (2010). For arbitrary values of the smoothness parameter $\nu$, the spectral density can be approximated by series expansions (see Roininen *et al.*, 2014, for analysis) giving a rational representation (see Publication I for methods).

### Squared exponential

The squared exponential (also known as the Gaussian, radial basis function, or exponentiated quadratic) covariance function is given as

$$\kappa_{\text{se}}(\tau) = \sigma^2 \exp\left( - \frac{\tau^2}{2\ell^2} \right), \tag{3.27}$$

where $\sigma^2$ is a magnitude scale hyperparameter and $\ell$ is the characteristic length-scale. It can be recovered in the limit of the Matérn smoothness

parameter $\nu$ going to infinity. The model functions are thus infinitely smooth (infnitely differentiable), which would require an infinite number of derivative trajectories stacked in the state of an exact SDE representation. However, the spectral density of the squared exponential can be efficiently approximated by a Taylor expansion (see Hartikainen and Särkkä, 2010, and Publication I) or Padé approximants (Särkkä and Piché, 2014). See Publication I for an example of how to construct an approximate state-space representation.

Figure 3.1 demonstrates several stationary covariance functions. The effect of smoothness (differentiability) is clearly visible on the middle row, where the realizations go from entirely non-differentiable to infinitely smooth.

**Rational quadratic**

The rational quadratic covariance function can be seen as a special case of a *scale mixture* kernel formulation, where a general class of kernels can be constructed as superpositions of squared exponential covariance functions with a distribution $p(\ell)$ over the length-scales (see, *e.g.*, Stein, 1999; Rasmussen and Williams, 2006):

$$\kappa_{\mathrm{sm}}(\tau) = \int_0^\infty \kappa_{\mathrm{se}}(\tau \mid \ell)\, p(\ell)\, \mathrm{d}\ell, \tag{3.28}$$

where $\kappa_{\mathrm{se}}(\tau \mid \ell)$ denotes the squared exponential kernel (3.27) with length-scale $\ell$. A general overview of how to compose the corresponding SDE models is presented in Publication IV.

The rational quadratic (RQ) covariance function (see, *e.g.*, Matérn, 1960; Rasmussen and Williams, 2006) corresponds to the scale-mixture model with a gamma prior distribution on the inverse squared length-scale. The analytic expression is

$$\kappa_{\mathrm{rq}}(\tau) = \sigma^2 \left(1 + \frac{\tau^2}{2\alpha\ell^2}\right)^{-\alpha}, \tag{3.29}$$

where $\alpha > 0$ is a shape parameter that defines the decay of the tail. The Cauchy covariance is a special case for $\alpha = 1$. As $\alpha \to \infty$, the covariance function converges to the squared exponential.

The rational quadratic covariance function can be approximated in terms of

$$\kappa_{\mathrm{rq}}(\tau) \approx \sum_{i=1}^{n} \kappa_{\mathrm{se}}(\tau \mid \sigma_i^2, \ell_i), \tag{3.30}$$

where the squared exponentials are evaluated with magnitudes $\sigma_i^2 = \sigma_{\mathrm{rq}}^2\, w_i / \Gamma(\alpha)$ and length-scales $\ell_i^2 = \ell_{\mathrm{rq}}^2 \alpha / x_i$ (Publication IV). The points

$x_i$ and weights $w_i$ are given such that $x_i, i = 1, 2, \ldots, n$ are the roots of the generalized Laguerre polynomial $\mathrm{L}_n^{\alpha-1}(x)$, and the weights $w_i$ are given as:

$$w_i = \frac{\Gamma(n+\alpha)\, x_i}{n!\,(n+1)^2\left[\mathrm{L}_{n+1}^{\alpha-1}(x_i)\right]^2}. \tag{3.31}$$

**Periodic**

MacKay (1998) discusses a way of constructing periodic covariance functions through warping. By mapping a one-dimensional input variable to two dimensions, $\mathbf{u}(t) = (\cos(\omega_0\, t), \sin(\omega_0\, t))$, and considering a squared exponential covariance function in $\mathbf{u}$-space, $\kappa(t, t') = \kappa_{\mathrm{se}}(\mathbf{u}(t), \mathbf{u}(t'))$, a periodic covariance function can be given as

$$\kappa_{\mathrm{periodic}}(\tau) = \sigma^2 \exp\left(-\frac{2\sin^2\left(\frac{\omega_0\,\tau}{2}\right)}{\ell^2}\right), \tag{3.32}$$

where $\sigma^2$ is the magnitude scale and $\ell$ the characteristic length-scale. The period length is determined by $\omega_0 = 2\pi/t_{\mathrm{period}}$. The periodic covariance function has a spectrum with a set of delta peaks appearing at the harmonic frequencies.

Publication III considers the link between SDE models and the above type of periodic covariance functions. The corresponding SDE is set up by a sum of $n$ two-dimensional SDE models of the following form. The feedback matrices are

$$\mathbf{F}_j = \begin{pmatrix} 0 & -\omega_0\, j \\ \omega_0\, j & 0 \end{pmatrix}, \tag{3.33}$$

noise effect matrices $\mathbf{L}_j = \mathbf{I}_2$ (a $2 \times 2$ identity matrix) and the stationary covariances $\mathbf{P}_{\infty,j} = q_j^2\, \mathbf{I}_2$. The process does not have a diffusion term, so $\mathbf{Q}_{\mathrm{c}} = \mathbf{0}$. The measurement models are $\mathbf{H}_j = (1\ 0)$. The coefficients (see Publication III) are given by $q_j^2 = 2\,\mathrm{I}_j(\ell^{-2})/\exp(\ell^{-2})$, for $j = 1, 2, \ldots, n$, and $q_0^2 = \mathrm{I}_0(\ell^{-2})/\exp(\ell^{-2})$, where $\mathrm{I}_\alpha(z)$ is the modified Bessel function of the first kind of order $\alpha$ (Abramowitz and Stegun, 1970).

If the periodic variation in the data is not exactly periodic, but quasi-periodic (almost periodic), the periodic covariance is too restrictive on its own. A common way to allow for seasonable variation is to consider the product of a long length-scale stationary covariance function and a periodic covariance function. This allows the model to decay away from exact periodicity (see, *e.g.*, Rasmussen and Williams, 2006). For the product formulation in Section 3.2 to hold, the process noise spectral density of the periodic SDE needs to be redefined as $\mathbf{Q}_{\mathrm{c},j} = q_j^2\, \mathbf{I}_2$ (see Publication III).

Figure 3.1 shows both periodic and quasi-periodic realizations (a product between a periodic and a Matérn covariance function). The periodic realizations are arbitrarily-shaped but exactly the same between periods. In the quasi-periodic realizations the shape slowly deforms between periods.

**White noise**

The white noise covariance function is (see Tarantola, 2005, for discussion)

$$\kappa_{\text{noise}}(t, t') = \sigma^2\, \delta(t - t'), \tag{3.34}$$

where $\sigma^2$ is a magnitude scale hyperparameter and $\delta(t - t')$ represents the Dirac delta function. This covariance function has a flat frequency spectrum. For an indexed set of inputs, the noise covariance is given by

$$\kappa_{\text{noise}}(t_i, t_j) = \sigma^2\, \delta_{i,j}, \tag{3.35}$$

where $\sigma^2$ is a magnitude scale hyperparameter and $\delta_{i,j}$ is the Kronecker delta function and each time point is independent from each other. It can be seen as the limit of other stationary kernels, when the characteristic length-scale tends to zero, $\ell \to 0$. Consider the Ornstein–Uhlenbeck process $\frac{\mathrm{d}f(t)}{\mathrm{d}t} = -\lambda f(t) + w(t)$, where $w(t)$ is a white noise process with spectral density $Q_{\text{c}} = 2\lambda\sigma^2$ and the stationary state covariance $P_\infty = \sigma^2$. Now if $\lambda \to \infty$ (the characteristics length-scale going to zero), this model converges pointwise to the independent discrete-time model defined by $A_k = 0$ and $Q_k = \sigma^2$. The initial state variance is $P_0 = \sigma^2$. However, in order for to process to converge to (3.34), the variance $\sigma^2$ would be required to go to infinity at the same time.

## 3.4 Sequential Inference

In state-space models, the Bayesian methodology of computing posterior distributions of the latent state based on a history of noisy measurements is known as *filtering* and *smoothing* (see, *e.g.*, Särkkä, 2013, for an overview). The data up to time instance $t_n$ is denoted by $\mathcal{D}_n = \{(t_i, y_i)\}_{i=1}^n$. In Bayesian filtering and smoothing for linear Gaussian systems the interest is in the following marginal distributions:

**Filtering distributions** are the result of the Kalman filter (Kalman, 1960). They are the marginal distributions of the state $\mathbf{f}_k$ given the

current and previous measurements up to time point $t_k$. The following shorthand notation is used:

$$\mathbf{f}_k \mid \mathcal{D}_k \sim \mathrm{N}(\mathbf{m}_{k|k}, \mathbf{P}_{k|k}).$$

**Prediction distributions,** computed from the prediction step of the Kalman filter, are the marginal distributions of the future state $\mathbf{f}_{k+j}$, for $j = 1, 2, \ldots$ steps following the previous observation:

$$\mathbf{f}_{k+j} \mid \mathcal{D}_k \sim \mathrm{N}(\mathbf{m}_{k+j|k}, \mathbf{P}_{k+j|k}).$$

**Smoothing distributions** are solved by the Rauch–Tung–Striebel smoother (Rauch *et al.*, 1965) and they are the marginal distributions of state $\mathbf{f}_k, k = 1, 2, \ldots, n$, given all the measurements in the entire interval:

$$\mathbf{f}_k \mid \mathcal{D}_n \sim \mathrm{N}(\mathbf{m}_{k|n}, \mathbf{P}_{k|n}).$$

The estimation is started from the Gaussian prior $p(\mathbf{f}_0) = \mathrm{N}(\mathbf{f}_0 \mid \mathbf{m}_{0|0}, \mathbf{P}_{0|0})$, where $\mathbf{m}_{0|0} = \mathbf{0}$ and $\mathbf{P}_{0|0} = \mathbf{P}_0$. For each input $k = 1, 2, \ldots, n$, the following *Kalman prediction* step gives the marginal distribution for step $k$ given the filtering outcome on the previous step:

$$\begin{aligned} \mathbf{m}_{k|k-1} &= \mathbf{A}_{k-1}\, \mathbf{m}_{k-1|k-1}, \\ \mathbf{P}_{k|k-1} &= \mathbf{A}_{k-1}\, \mathbf{P}_{k-1|k-1}\, \mathbf{A}_{k-1}^\mathsf{T} + \mathbf{Q}_{k-1}. \end{aligned} \tag{3.36}$$

The Kalman prediction step is evaluated for all inputs—including test inputs $t_\star$ at which there is no data. At points with data, the prediction is updated by the *Kalman update* step:

$$\begin{aligned} \mathbf{v}_k &= \mathbf{y}_k - \mathbf{H}_k\, \mathbf{m}_{k|k-1}, \\ \mathbf{S}_k &= \mathbf{H}_k\, \mathbf{P}_{k|k-1}\, \mathbf{H}_k^\mathsf{T} + \mathbf{R}_k, \\ \mathbf{K}_k &= \mathbf{P}_{k|k-1}\, \mathbf{H}_k^\mathsf{T}\, \mathbf{S}_k^{-1}, \\ \mathbf{m}_{k|k} &= \mathbf{m}_{k|k-1} + \mathbf{K}_k\, \mathbf{v}_k, \\ \mathbf{P}_{k|k} &= \mathbf{P}_{k|k-1} - \mathbf{K}_k\, \mathbf{S}_k\, \mathbf{K}_k^\mathsf{T}. \end{aligned} \tag{3.37}$$

For temporal models the measurement noise covariance $\mathbf{R}_k = \sigma^2_{\text{noise}}$ is one-dimensional, but the matrix notation is used for forward compatibility. The terms $\mathbf{v}_k$ and $\mathbf{S}_k$ are known as the innovation mean and covariance, and $\mathbf{K}_k$ is the Kalman gain.

The computational complexity of the Kalman filtering scheme scales as $\mathcal{O}(nm^3)$, where $m$ is the dimensionality of the state. If the filter output

is stored, the memory requirements scale as $\mathcal{O}(nm^2)$. Possible numerical instability issues can be circumvented by using the so-called Joseph's formula for updating $\mathbf{P}_{k|k}$ or by resorting to square-root filtering (see Bierman, 1977).

The smoothing solution can be obtained by applying the following backward recursion on the filtering outcome starting from the filtering result on step $n$. This scheme is known as the Rauch–Tung–Striebel smoother (Rauch *et al.*, 1965):

$$\begin{aligned}
\mathbf{m}_{k+1|k} &= \mathbf{A}_k\, \mathbf{m}_{k|k}, \\
\mathbf{P}_{k+1|k} &= \mathbf{A}_k\, \mathbf{P}_{k|k}\, \mathbf{A}_k^\mathsf{T} + \mathbf{Q}_k, \\
\mathbf{G}_k &= \mathbf{P}_{k|k}\, \mathbf{A}_k^\mathsf{T}\, \mathbf{P}_{k+1|k}^{-1}, \\
\mathbf{m}_{k|n} &= \mathbf{m}_{k|k} + \mathbf{G}_k\, (\mathbf{m}_{k+1|n} - \mathbf{m}_{k+1|k}), \\
\mathbf{P}_{k|n} &= \mathbf{P}_{k|k} + \mathbf{G}_k\, (\mathbf{P}_{k+1|n} - \mathbf{P}_{k+1|k})\, \mathbf{G}_k^\mathsf{T},
\end{aligned} \tag{3.38}$$

where $\mathbf{G}_k$ denotes the smoother gain.

Predicting the latent value $f(t_\star)$ at an unseen test input $t_\star$ is done by including the test input in the prediction and smoothing steps. The prediction is given by $\mathrm{E}[f(t_\star)] = \mathbf{H}\, \mathbf{m}_{\star|n}$ and $\mathrm{V}[f(t_\star)] = \mathbf{H}\, \mathbf{P}_{\star|n}\, \mathbf{H}^\mathsf{T}$.

The filter and smoother work with the marginal distributions. However, it is possible to recover the full posterior covariance $\mathbf{\Sigma}$ from the smoother outcome. The diagonal elements (marginal variances) are given by $\mathbf{\Sigma}_{k,k} = \mathbf{H}\, \mathbf{P}_{k|n}\, \mathbf{H}^\mathsf{T}$, for $k = 1, 2, \ldots, n$, and the lower triangular of the symmetric covariance is given by

$$\mathbf{\Sigma}_{i,j} = \mathbf{H} \left( \prod_{k=j}^{i-1} \mathbf{G}_k \right) \mathbf{P}_{i|n}\, \mathbf{H}^\mathsf{T}, \quad \text{for} \quad i > j, \tag{3.39}$$

where $\mathbf{G}_k$ is the smoother gain in Equation (3.38). It is, however, not necessary to form the full posterior for drawing posterior samples from the latent process. This can be done directly by using the smoother solution as explained in Doucet (2010).

For training the hyperparameters, the logarithm of the marginal likelihood can be evaluated sequentially as a by-product of the filtering recursion (see, *e.g.*, Särkkä, 2013):

$$\mathcal{L}(\boldsymbol{\theta}, \sigma_{\text{noise}}^2) = \sum_{k=1}^{n} \left\{ -\frac{1}{2} \log |2\pi\, \mathbf{S}_k| - \frac{1}{2} \mathbf{v}_k^\mathsf{T}\, \mathbf{S}_k^{-1}\, \mathbf{v}_k \right\}, \tag{3.40}$$

where $\mathbf{v}_k$ and $\mathbf{S}_k$ are the innovation mean and covariance evaluated by the filter update step (3.37). The partial derivatives of the log marginal likelihood function (3.40) require derivatives of the entire filtering recursion to be calculated (see, *e.g.*, Mbalawata *et al.*, 2013).

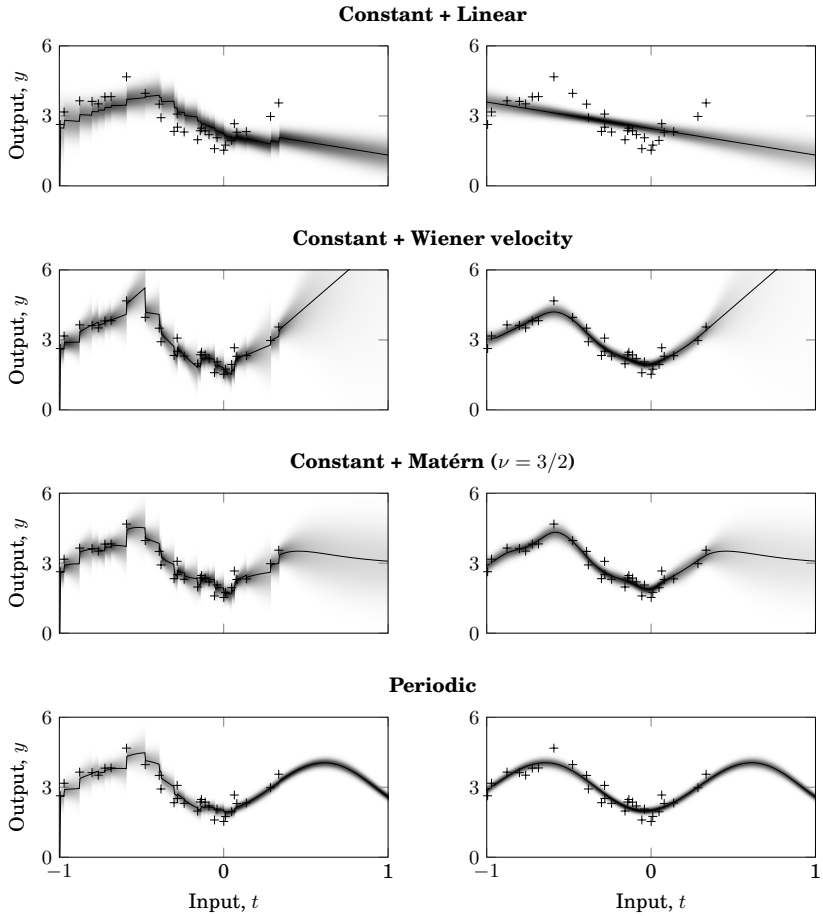


**Figure 3.2.** Filtering (left) and smoothing (right) solutions to Gaussian process regression problems with different prior models. The hyperparameters in each model were trained by maximizing the marginal likelihood. The training data is shown by the black markers, the black line shows the posterior mean, and the shading illustrates the marginal density.

Figure 3.2 demonstrates the filtering and smoothing outcome for a set of different GP priors. The data in each example is the same, but the prior assumptions differ. Because the data is clearly not zero-mean, a constant covariance function was added to three of the models. The hyperparameters in the models (including the measurement noise variance) were optimized with respect to the marginal likelihood. The filtering outcome in Figure 3.2 shows the sequential nature of the filter, where the marginal variance drops as new data points are encountered. The smoother outcomes on the right hand side correspond to the batch solutions given by Equation (2.10).

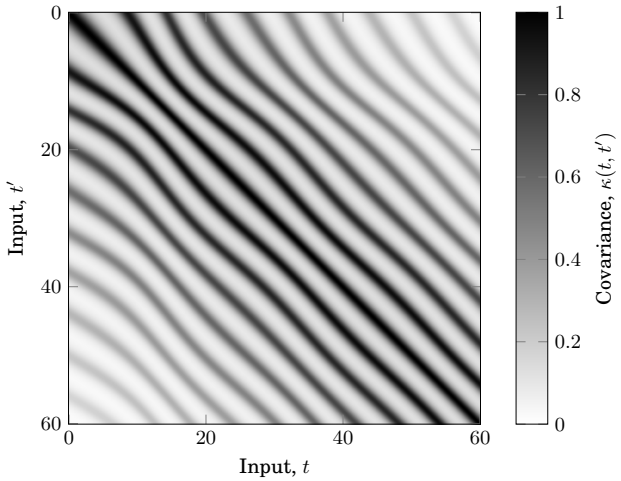


**Figure 3.3.** An illustrative example of the covariance associated with a quasi-periodic model with time-varying frequency similar to the model in Publication VI. The wiggling stripes encode the varying period length and the decay corresponds to quasi-periodicity.

In Publication VI the state-space model is not time-invariant, but the feedback model matrix $\mathbf{F}$ depends on time. This results in a highly non-stationary model for estimating quasi-periodic oscillations with time-varying frequencies. The time-varying periodicity makes the frequency trajectory a stochastic process as well. If the frequency is assumed constant between measurements and it only takes values in a discrete set, the time-dependent periodicity can be modelled as switchings between time-invariant periodic models. The interacting multiple models (IMM) filter (Bar-Shalom *et al.*, 2001) is a computationally efficient estimation algorithm for Markovian switching systems of this kind.

Figure 3.3 shows an example of what the corresponding covariance matrix looks like for a non-stationary quasi-periodic model similar to those in Publication VI. The model was constructed as a product of a long length-scale Matérn covariance function (seen as decay in the figure) and a periodic covariance function with time-varying frequency. The covariance matrix is evaluated from the relation (3.39) which can be used for evaluating the prior covariance if no Kalman updates are performed.

## 3.5 Student-$t$ Processes as an Extension of Gaussian Processes

The Gaussian process framework allows for a slight extension in terms of generalizing the methodology to more general processes with ellipti-

cal contours and heavy tails—most notably Student-$t$ processes (TPs, see Shah *et al.*, 2014). They borrow the framework for inference from Gaussian processes, but the multivariate Gaussian is replaced with a multivariate Student-$t$ distribution. They provide some additional flexibility and the heavy tails robustify inference in outlier prone data. In TPs the predictive covariance depends on the training observations, whereas in GPs it only depends on the training inputs. The following sections will shortly review the concept of one-dimensional TP regression and explain the link to state-space inference (Publication V).

Recently Shah *et al.* (2014) resurrected an interest in TP regression which is—alike GP regression—concerned with predicting an unknown scalar output $f(t_\star)$ associated with a known input $t_\star \in \mathbb{R}$, given a training data set $\mathcal{D} = \{(t_k, y_k)\}_{k=1}^n$. The model functions are assumed to be realizations of a zero-mean Student-$t$ random process prior with covariance function $\kappa_{\boldsymbol{\theta}}(t, t')$, degrees of freedom $\nu$, and the observations corrupted by an entangled Student-$t$ noise process (see Publication V for parametrizations):

$$\begin{aligned} f(t) &\sim \mathcal{TP}(0, \kappa(t, t'), \nu), \\ \mathbf{y} &= \mathcal{H}\, f(t), \end{aligned} \tag{3.41}$$

where the noise model is incorporated in the covariance function. The direct solution to the TP regression problem gives predictions for the latent function, $p(f(t_\star) \mid \mathcal{D}, t_\star) = \mathrm{MVT}(\mathrm{E}[f(t_\star)], \mathrm{V}[f(t_\star)], \nu + n)$. For the entangled noise model, there exists a closed-form solution (Shah *et al.*, 2014):

$$\begin{aligned} \mathrm{E}[f(t_\star)] &= \mathbf{k}_\star^\mathsf{T}\, \mathbf{K}^{-1}\, \mathbf{y}, \\ \mathrm{V}[f(t_\star)] &= \frac{\nu - 2 + \mathbf{y}^\mathsf{T}\, \mathbf{K}^{-1}\, \mathbf{y}}{\nu - 2 + n} \left( \kappa_{\boldsymbol{\theta}}(t_\star, t_\star) - \mathbf{k}_\star^\mathsf{T}\, \mathbf{K}^{-1}\, \mathbf{k}_\star \right), \end{aligned} \tag{3.42}$$

which in notation follows the GP solution. The noise model is included in the covariance function by adding a noise covariance function to the parametrized kernel: $\mathbf{K}_{i,j} = \kappa_{\boldsymbol{\theta}}(t_i, t_j) + \sigma_{\text{noise}}^2\, \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta. Consequently, the noise will be uncorrelated with the latent function, but not independent (Shah *et al.*, 2014). In the limit $\nu \to \infty$, this model tends to a GP regression model with independent Gaussian noise. The computational complexity of these equations remains cubic due to the matrix inversion.

Publication V introduced a link between one-dimensional TP regression, the corresponding state-space model formulation, and an associated sequential inference scheme. The formulation is based on representing the

entangled noise as an augmented noise process in the state. This was motivated by the connection of sums of covariance functions and stacked SDEs (*cf.* Sec. 3.2) and the formal representation of the noise covariance function (Eq. 3.35) as an SDE.

Approximative extensions of the Kalman filter to Student-$t$ distributions have been considered before (see Roth, 2013). However, the Student-$t$ filter and smoother solution corresponding to Equation (3.42) can be given in closed form, due to the entangled noise structure, which ignores the measurement noise ($\mathbf{R} = \mathbf{0}$) and augments the noise process instead (see Publication V for the algorithm).

# 4. Stochastic Partial Differential Equation Models

The previous two parts of this overview considered two different views on Gaussian process models: a general batch approach and a sequential model for one-dimensional inputs. This section combines these two into a spatio-temporal GP framework. The interest in this section is in the connection between spatio-temporal Gaussian processes and infinite-dimensional state-space models.

Modelling of spatio-temporal data is straight-forward under the GP regression paradigm. The space $\mathbf{r} \in \mathbb{R}^d$ and time $t \in \mathbb{R}$ variables jointly take the roles of input variables of the regression function: $\mathbf{x} = (\mathbf{r}, t)$. The batch GP regression formalism in Section 2.3 can be employed for solving models of this type. However, in applications of this kind, the cubic computational scaling easily becomes an issue. For example, in a data set with $n_\mathrm{r}$ observations made at $n_\mathrm{t}$ time instances, the computational complexity would scale as $\mathcal{O}((n_\mathrm{r} n_\mathrm{t})^3)$.

For reformulating the model, consider a set of spatio-temporal inputs $(\mathbf{r}_i, t_i)$ associated with outputs $y_i$ and regroup them such that for each $t_k$ there is a set of spatial inputs $\{\mathbf{r}_i\}_{i=1}^{n_k}$ and associated outputs $\mathbf{y}_k \in \mathbb{R}^{n_k}$. Under the measurement operator formulation in Section 2.3 (*cf.* Eq. 2.8) this would correspond to having a different measurement functional $\mathcal{H}_k$ for each $t_k, k = 1, 2, \ldots, n_\mathrm{t}$:

$$\begin{aligned} f(\mathbf{r}, t) &\sim \mathcal{GP}(0, \kappa(\mathbf{r}, t; \mathbf{r}', t')), \\ \mathbf{y}_k &= \mathcal{H}_k \, f(\mathbf{r}, t_k) + \boldsymbol{\varepsilon}_k, \end{aligned} \tag{4.1}$$

where $\boldsymbol{\varepsilon}_k \sim \mathrm{N}(\mathbf{0}, \sigma^2_\mathrm{noise} \, \mathbf{I})$ is the Gaussian measurement noise.

Recall that the regressors in the GP model are functions in an infinite-dimensional Hilbert space. In Publication I the rationale is that similarly as for the finite-dimensional state variables in Equation (3.5), the evolution of the infinite-dimensional quantities can be encoded into a state-space model. For spatio-temporal models, this is interpreted as a Gaus-

sian process (Gaussian random field) which evolves over time. The GP regression problem can be reformulated into a stochastic partial differential equation (SPDE, see Da Prato and Zabczyk, 1992) describing the temporal dynamics of the spatial process:

$$\begin{aligned} \frac{\partial \mathbf{f}(\mathbf{r}, t)}{\partial t} &= \boldsymbol{\mathcal{F}}\, \mathbf{f}(\mathbf{r}, t) + \mathbf{L}\, \mathbf{w}(\mathbf{r}, t), \\ \mathbf{y}_k &= \boldsymbol{\mathcal{H}}_k\, \mathbf{f}(\mathbf{r}, t_k) + \boldsymbol{\varepsilon}_k, \end{aligned} \tag{4.2}$$

where the process is driven by a spatio-temporal white noise process $\mathbf{w}(\mathbf{r}, t)$ with a spectral density function $\mathbf{Q}_\mathrm{c}(\mathbf{r}, \mathbf{r}')$.

The dynamical model (4.2) is an infinite-dimensional time-invariant linear stochastic partial differential equation. If the operator $\boldsymbol{\mathcal{F}}$ is a differential operator, the dynamic model is an evolution type SPDE (see, *e.g.*, Chow, 2007). The model formulation generalizes to a setup, where the operators are pseudo-differential operators (Shubin, 1987), such that the operator is constructed by manipulations of differential operators.

The model (4.2) bears strong resemblance to the SDE models (3.5) discussed in the previous section, but now the feedback is defined by a matrix $\boldsymbol{\mathcal{F}}$ of (spatial) linear operators. The benefit of this is that the inference can now be done by infinite-dimensional Kalman filtering and smoothing (Curtain, 1975) which also goes under the name 'distributed parameter estimation' (Tzafestas, 1978). The computational scaling becomes $\mathcal{O}(n_\mathrm{t} n_\mathrm{r}^3)$, which means linear time complexity with respect to the number of temporal observations.

The following sections build upon the methods that have already been presented for finite-dimensional models. A formal generalization of the finite-dimensional sequential inference scheme to infinite-dimensions is briefly covered in the next section. More practical computational aspects are then discussed on a general level, and this section is concluded by presenting means of forming evolution SPDEs of common covariance function models.

### 4.1 Infinite-Dimensional Filtering and Smoothing

In order to do sequential inference in the family of evolution type SPDEs, the Markov structure in the model needs to be explicit. This can be accomplished by forming the corresponding discrete-time infinite-dimensional state-space model. The solution to the linear time-invariant SPDE can be

given as follows (see Särkkä and Hartikainen, 2012, for details):

$$\begin{aligned} \mathbf{f}(\mathbf{r}, t_k) &= \boldsymbol{\mathcal{A}}_{k-1}\, \mathbf{f}(\mathbf{r}, t_{k-1}) + \mathbf{q}_{k-1}(\mathbf{r}), \\ \mathbf{y}_k &= \boldsymbol{\mathcal{H}}_k\, \mathbf{f}(\mathbf{r}, t) + \varepsilon_k, \end{aligned} \tag{4.3}$$

where $\boldsymbol{\mathcal{A}}_k = \exp(\Delta t_k\, \boldsymbol{\mathcal{F}})$ is the evolution operator given in terms of the operator exponential function. The time step length is the difference between the previous and next step $\Delta t_k = t_{k+1} - t_k$. The process noise term $\mathbf{q}_{k-1}(\mathbf{r})$ is a zero-mean multivariate Gaussian process with covariance function $\mathbf{Q}_{k-1}(\mathbf{r}, \mathbf{r}')$.

The discrete-time model gives the formal solution to an evolution SPDE. This solution is sometimes referred to as the *mild solution*, which is a somewhat weaker solution concept than the *weak solution* (see Da Prato and Zabczyk, 1992). It satisfies the equation in distribution sense, which suffices for this probabilistic inference setup.

Similarly as the Kalman filter is the closed-form solution to the finite-dimensional filtering problem, the infinite-dimensional filtering problem has a solution in terms of an *infinite-dimensional Kalman filter* (Curtain, 1975; Tzafestas, 1978; Omatu and Seinfeld, 1989; Cressie and Wikle, 2002). The state estimates are given in terms of a multivariate Gaussian process with mean $\mathbf{m}(\mathbf{r})$ and covariance function $\mathbf{P}(\mathbf{r}, \mathbf{r}')$. The iteration is initialized by the prior $\mathbf{f}(\mathbf{r}, t_0) \sim \mathcal{GP}(\mathbf{m}_0(\mathbf{r}), \mathbf{P}_0(\mathbf{r}, \mathbf{r}'))$. The infinite-dimensional *prediction step* is:

$$\begin{aligned} \mathbf{m}_{k|k-1}(\mathbf{r}) &= \boldsymbol{\mathcal{A}}_{k-1}\, \mathbf{m}_{k-1|k-1}(\mathbf{r}), \\ \mathbf{P}_{k|k-1}(\mathbf{r}, \mathbf{r}') &= \boldsymbol{\mathcal{A}}_{k-1}\, \mathbf{P}_{k-1|k-1}(\mathbf{r}, \mathbf{r}')\, \boldsymbol{\mathcal{A}}_{k-1}^* + \mathbf{Q}_{k-1}(\mathbf{r}, \mathbf{r}'), \end{aligned} \tag{4.4}$$

where $(\cdot)^*$ denotes an adjoint. The operator adjoint can be seen as an operator version of a matrix transpose which effectively reverses the direction of the operation.

The infinite-dimensional Kalman *update step* becomes:

$$\begin{aligned} \mathbf{v}_k &= \mathbf{y}_k - \boldsymbol{\mathcal{H}}_k\, \mathbf{m}_{k|k-1}(\mathbf{r}), \\ \mathbf{S}_k &= \boldsymbol{\mathcal{H}}_k\, \mathbf{P}_{k|k-1}(\mathbf{r}, \mathbf{r}')\, \boldsymbol{\mathcal{H}}_k^* + \mathbf{R}_k, \\ \mathbf{K}_k(\mathbf{r}) &= \mathbf{C}_{k|k-1}(\mathbf{r}, \mathbf{r}')\, \boldsymbol{\mathcal{H}}_k^*\, \mathbf{S}_k^{-1}, \\ \mathbf{m}_{k|k}(\mathbf{r}) &= \mathbf{m}_{k|k-1}(\mathbf{r}) + \mathbf{K}_k(\mathbf{r})\, \mathbf{v}_k, \\ \mathbf{P}_{k|k}(\mathbf{r}, \mathbf{r}') &= \mathbf{P}_{k|k-1}(\mathbf{r}, \mathbf{r}') - \mathbf{K}_k(\mathbf{r})\, \mathbf{S}_k\, \mathbf{K}_k^*(\mathbf{r}). \end{aligned} \tag{4.5}$$

Note that the innovation mean $\mathbf{v}_k$ and covariance $\mathbf{S}_k$ are finite-dimensional, and $\mathbf{R}_k = \sigma^2_{\text{noise}}\, \mathbf{I}$. The filtering solution is given by $\mathbf{f}(\mathbf{r}, t_k) \mid \mathcal{D}_k \sim \mathcal{GP}(\mathbf{m}_{k|k}(\mathbf{r}), \mathbf{P}_{k|k}(\mathbf{r}, \mathbf{r}'))$. Figure 4.1 gives an illustrative example of

the marginal mean and variance (standard deviation) produced by the infinite-dimensional Kalman filter.

The infinite-dimensional Rauch–Tung–Striebel smoother equations are similar to the finite-dimensional smoothing equations. The backward recursion is started from the final filtering step and iterated for $k = n_\mathrm{t} - 1, n_\mathrm{t} - 2, \ldots, 1$ such that

$$\begin{aligned}
\mathbf{m}_{k+1|k}(\mathbf{r}) &= \boldsymbol{\mathcal{A}}_k \, \mathbf{m}_{k|k}(\mathbf{r}), \\
\mathbf{P}_{k+1|k}(\mathbf{r}, \mathbf{r}') &= \boldsymbol{\mathcal{A}}_k \, \mathbf{P}_{k|k}(\mathbf{r}, \mathbf{r}') \, \boldsymbol{\mathcal{A}}_k^* + \mathbf{Q}_k(\mathbf{r}, \mathbf{r}'), \\
\boldsymbol{\mathcal{G}}_k &= \mathbf{P}_{k|k}(\mathbf{r}, \mathbf{r}') \, \boldsymbol{\mathcal{A}}_k^* \, [\mathbf{P}_{k+1|k}(\mathbf{r}, \mathbf{r}')]^{-1}, \\
\mathbf{m}_{k|n}(\mathbf{r}) &= \mathbf{m}_{k|k}(\mathbf{r}) + \boldsymbol{\mathcal{G}}_k \, [\mathbf{m}_{k+1|n}(\mathbf{r}) - \mathbf{m}_{k+1|k}(\mathbf{r})], \\
\mathbf{P}_{k|n}(\mathbf{r}, \mathbf{r}') &= \mathbf{P}_{k|k}(\mathbf{r}, \mathbf{r}') + \boldsymbol{\mathcal{G}}_k \, [\mathbf{P}_{k+1|n}(\mathbf{r}, \mathbf{r}') - \mathbf{P}_{k+1|k}(\mathbf{r}, \mathbf{r}')] \, \boldsymbol{\mathcal{G}}_k^*,
\end{aligned} \tag{4.6}$$

where $(\cdot)^{-1}$ denotes an operator inverse. The smoother gain operator $\boldsymbol{\mathcal{G}}_k$ has an interpretation as a linear operator with a kernel defined in terms of the covariance functions of the filtering results. The marginal posterior Gaussian processes representing the smoothing solutions are given in terms of $\mathbf{f}(\mathbf{r}, t_k) \mid \mathcal{D}_n \sim \mathcal{GP}(\mathbf{m}_{k|n}(\mathbf{r}), \mathbf{P}_{k|n}(\mathbf{r}, \mathbf{r}'))$. Figure 4.1 also shows the smoother outcome, which coincides with the solution which would have been obtained by the naïve batch approach.

Predicting the latent values $\mathbf{f}(\mathbf{r}_\star, t_\star)$ in an unseen spatial test point $\mathbf{r}_\star$ at time $t_\star$, can be done similarly as in the one-dimensional case. The new time step follows the time-ordering such that $t_k < t_\star < t_{k+1}$ and it is included in the filter prediction and smoother iterations. The spatial input defines the point in which the measurement model evaluates the function. The prediction is given as

$$p(\mathbf{f}(\mathbf{r}_\star, t_\star) \mid \mathcal{D}_n) = \mathrm{N}(\mathbf{f}(\mathbf{r}_\star, t_\star) \mid \boldsymbol{\mathcal{H}}_\star \, \mathbf{m}_{\star|n}(\mathbf{r}), \boldsymbol{\mathcal{H}}_\star \, \mathbf{P}_{\star|n}(\mathbf{r}, \mathbf{r}') \, \boldsymbol{\mathcal{H}}_\star^*). \tag{4.7}$$

The infinite-dimensional filter and smoother are abstractions of their finite-dimensional counterparts. The extra abstraction layer limits their practical applicability, an issue which will be covered in the next section. However, the marginal likelihood $p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2_\mathrm{noise})$ can be directly evaluated using the innovation mean and covariance in Equation (4.5) and the likelihood function in Equation (3.40).

## 4.2 Computational Approaches

The infinite-dimensional filtering equations give the theoretical backdrop for implementation on computers with finite resources. For most cases

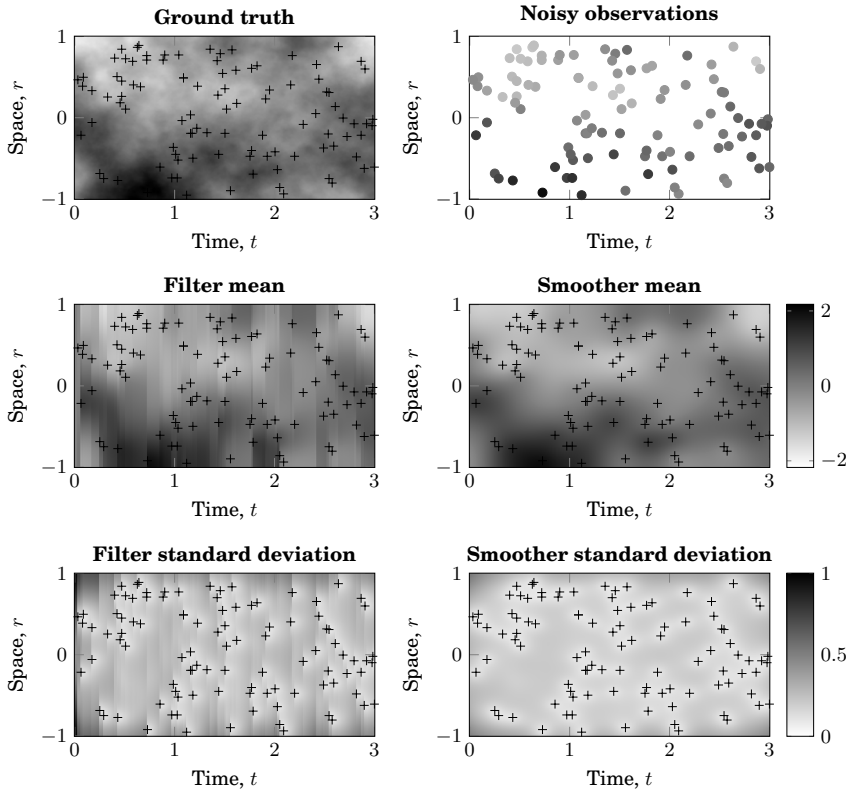


**Figure 4.1.** An illustrative example of spatio-temporal filtering and smoothing. The simulated ground truth and the noisy observations of it are visualized on the top row. The marginal filtering (left) and smoothing (right) mean and standard deviation fields are shown on the bottom rows.

the infinite-dimensional sequential filters and smoothers cannot be given exact implementations, but the equations need to be solved numerically. In practice, this means that some approximations need to be employed.

There exists a wide range of numerical methods arising from partial and pseudo-differential equation models. General methods are typically based on discretization of the domain. In finite-difference methods (FDM) the domain is discretized into a grid and the derivatives are approximated by the corresponding finite differences. The finite element method (FEM) uses more general subdivision techniques for representing the domain as a graph of simple elements and uses a variational approximation (such as the Galerkin method) for representing the problem in each element. This framework has been successfully used by Lindgren *et al.* (2011). They represent the Matérn covariance model as a spatial SPDE and establish a link to the corresponding Gauss–Markov random field model. In the resulting model the precision matrix is sparse and it allows for efficient

solutions by FEM solvers in arbitrarily shaped domains.

When the number of inputs is very large, it is common to employ some form of dimension reduction. In this type of models it is practical to use spectral representations (basis function approximations) in terms of some orthogonal functions (*e.g.* Fourier, orthogonal polynomials, or eigenvectors from a covariance function), wavelets, splines, bisquares, or discrete kernel convolutions (see Cressie and Wikle, 2011, for discussion). More recent methods have, for example, been considered by Dewar *et al.* (2009); Scerri *et al.* (2009), and Zammit-Mangion *et al.* (2012b).

For the type of problems considered in this work, efficient solutions can be formulated in terms of special basis function approximations, which allow to keep the finite-dimensional representation low-dimensional. These methods are also generally known as Hilbert space methods (Showalter, 2010), and they build upon expanding the infinite-dimensional operator equations to truncated series approximations. This is based on the Hilbert–Schmidt theorem which is also known as the eigenfunction expansion theorem.

The basis function approach for representing the spatio-temporal evolution equation is based on forming a representation of the process, such that the model corresponds to a finite truncation of the sum in

$$f(\mathbf{r}, t) = \sum_{j=1}^{\infty} f^{(j)}(t)\, \phi_j(\mathbf{r}), \tag{4.8}$$

where $\phi_j(\mathbf{r})$ is a suitable basis function and $f^{(j)}(t)$ is the associated weight. The weights depend on time and are Gaussian processes of their own.

If the covariance function is stationary, the corresponding evolution SPDE can be expressed in terms of the Laplace operator $\nabla^2 = \frac{\partial^2}{\partial r_1} + \frac{\partial^2}{\partial r_2} + \cdots + \frac{\partial^2}{\partial r_d}$ (see Publication I for discussion). The eigendecomposition of the negative Laplace operator is given by

$$-\nabla^2 \phi_j(\mathbf{r}) = \lambda_j \phi_j(\mathbf{r}), \tag{4.9}$$

where $\mathbf{r} \in \Omega$ is a spatial coordinate, $\phi_j(\mathbf{r})$ is an eigenfunction, and $\lambda_j$ is the corresponding eigenvalue for each $j = 1, 2, \ldots$. These eigendecompositions can be given as closed-form expressions on various simple domains, subject to known boundary conditions (see Solin, 2012, for a review and applications to infinite-dimensional filtering).

If the feedback operator $\boldsymbol{\mathcal{F}}$ is defined by an $m \times m$ matrix of linear operators such that each of the operators can be expressed with the help of the

Laplace operator, the eigenfunctions $\phi_j(\mathbf{x})$ of $-\nabla^2$ are also eigenfunctions of each operator in $\mathcal{F}$.

Särkkä and Hartikainen (2012) go through a procedure for representing the evolution equation as a finite SDE model. The change of basis yields that the evolution is not subject to an infinite-dimensional Hilbert space but to a (truncated) finite-dimensional space defined by the orthonormal basis $\{\phi_1(\mathbf{r}), \phi_2(\mathbf{r}), \ldots, \phi_m(\mathbf{r})\}$. This transforms the stochastic evolution equation to

$$\begin{aligned} \frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} &= \tilde{\mathbf{F}}\,\mathbf{f}(t) + \tilde{\mathbf{L}}\,\mathbf{w}(t), \\ \mathbf{y}_k &= \tilde{\mathbf{H}}_k\,\mathbf{f}(t) + \boldsymbol{\varepsilon}_k, \end{aligned} \tag{4.10}$$

where $\tilde{\mathbf{F}}$ is a block-diagonal matrix where each block represents the $j$th approximation of $\mathcal{F}$ subject to eigenvalue $\lambda_j$, the noise effect $\tilde{\mathbf{L}} = \mathbf{I} \otimes \mathbf{L}$, and $\mathbf{w}(t)$ is a multi-dimensional white noise process with a coupled spectral density matrix following a block structure defined by the projection

$$[\mathbf{Q}_\mathrm{c}]_{i,j} = \int_\Omega \int_\Omega \phi_i(\mathbf{r})\,\mathbf{L}\,\mathbf{Q}_\mathrm{c}(\mathbf{r}, \mathbf{r}')\,\mathbf{L}^\mathsf{T} \phi_j(\mathbf{r}')\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{r}'. \tag{4.11}$$

For projections of this kind, Solin and Särkkä (2014) present efficient methods for approximations to stationary covariance functions written in terms of the Laplace operator eigenbasis.

The measurement model both evaluates the process at the spatial locations and projects the weights onto the associated basis functions. Let $\boldsymbol{\Phi}$ be a matrix of the basis functions evaluated at the observation locations $\mathbf{r}_i$ observed at $t_k$ such that each row $[\boldsymbol{\Phi}_k]_i = (\phi_1(\mathbf{r}_i)\ \phi_2(\mathbf{r}_i)\ \cdots\ \phi_m(\mathbf{r}_i))$. The observation model matrix is thus given by $\tilde{\mathbf{H}}_k = \boldsymbol{\Phi}_k \otimes \mathbf{H}_k$, where $\mathbf{H}_k$ is a finite representation of the measurement model functional.

For example, the non-separable Matérn fields in Figure 4.1 and 4.2 were simulated by a basis function approach. The basis functions were obtained as the closed-form solution to the negative Laplace operator eigendecomposition in the spatial domain, $r \in [-1,\ 1]$, subject to Neumann boundary conditions (zero normal derivatives at the boundary). This approach also generalizes to inference on manifolds. An example is given in Publication II, where the domain is the surface of the Earth.

## 4.3 Spatio-Temporal GPs as SPDEs

Much like the SDE representation, the SPDE approach allows for rewriting a wide range of GP models in terms of an evolution equation. In the

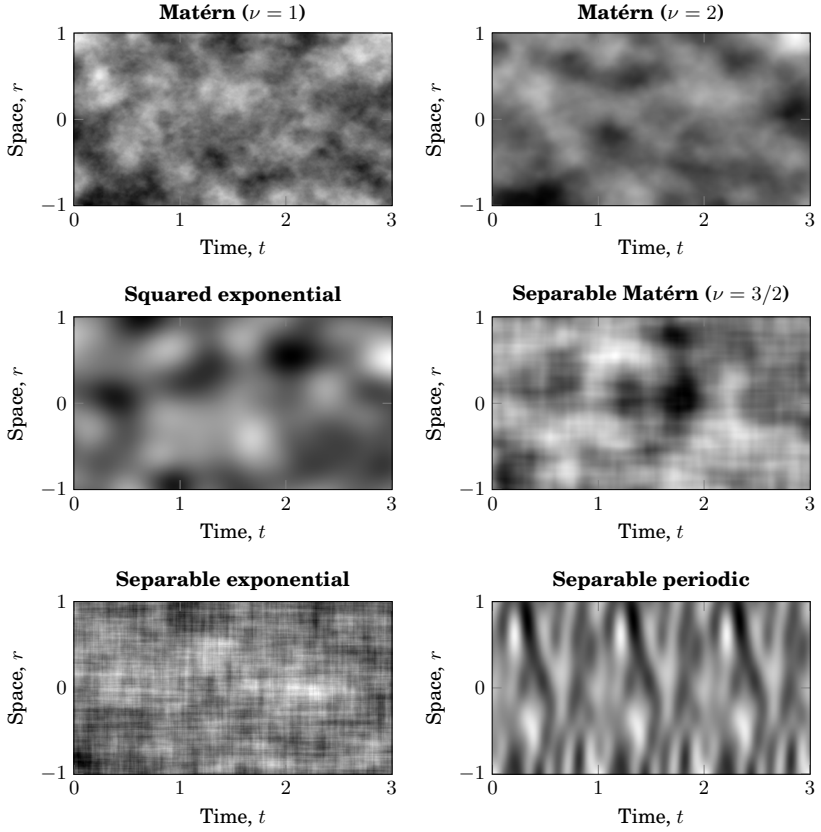


**Figure 4.2.** Examples of spatio-temporal Gaussian processes simulated by the infinite-dimensional state-space equations. The separable periodic model follows a squared exponential covariance functions in space and the periodic model ($t_{\text{period}} = 1$) in time. The spatial length-scales were fixed to $\ell = 0.25$ in all models, the magnitude is arbitrary and subject to the magnitude scale parameter.

following, a general overview of the methods will be presented both for non-separable and separable models. Figure 4.2 shows a collection of random draws from spatio-temporal models with different spatial and temporal structure. The vertical axis represents the spatial dimension, and time runs from left to right. In the example visualizations, the spatial domain is restricted to one-dimension.

### Non-separable stationary covariance functions

Transforming spatio-temporal GP models with stationary covariance functions into stochastic evolution type SPDEs was presented by Särkkä and Hartikainen (2012). The approach is further generalized in Publication I to cover the general case of approximating any stationary co-

variance function by a rational spectrum representation. This enables to solve a reformulated stationary GP regression problem with infinite-dimensional sequential methods.

As an example from Publication I, reconsider the Matérn covariance function that was presented under the SDE models in Section 3.3. The isotropic multi-dimensional ($\mathbf{x} \in \mathbb{R}^d$) Matérn covariance function has the form (3.24), such that $\kappa(\mathbf{x}, \mathbf{x}') = \kappa_{\text{Mat.}}(\|\mathbf{x} - \mathbf{x}'\|)$. The corresponding spectral density is

$$S(\boldsymbol{\omega}_{\text{x}}) = \frac{\gamma}{(\lambda^2 + \|\boldsymbol{\omega}_{\text{x}}\|^2)^{\nu+d/2}} = \frac{\gamma}{(\lambda^2 + \|\boldsymbol{\omega}_{\text{r}}\|^2 + \omega_{\text{t}}^2)^{\nu+d/2}}, \tag{4.12}$$

where $\lambda = \sqrt{2\nu}/\ell$. The scaling factor $\gamma = \sigma^2(4\pi)^{d/2}\Gamma(\nu + d/2)/\Gamma(\nu)\lambda^{2\nu}$ depends on the hyperparameters. The direct dependence on $d$ means that the same spectral density structure can give rise to differently behaving processes for different numbers of input dimensions. The roots of the expression in the denominator are given by $(\mathrm{i}\,\omega_{\text{t}}) = \pm\sqrt{\lambda^2 - \|\mathrm{i}\,\boldsymbol{\omega}_{\text{r}}\|^2}$, which means that the transfer function of the stable Markov process corresponds to (see Publication I for details)

$$G(\mathrm{i}\,\boldsymbol{\omega}_{\text{r}}, \mathrm{i}\,\omega_{\text{t}}) = (\mathrm{i}\,\omega_{\text{t}} + \sqrt{\lambda^2 - \|\mathrm{i}\,\boldsymbol{\omega}_{\text{r}}\|^2})^{-(\nu+d/2)}. \tag{4.13}$$

If $\nu + d/2$ is an integer, the term can easily be expanded and the coefficient matched to the corresponding feedback structure. For a two-dimensional ($d = 2$) visualization, let $\nu = 1$. The evolution type SPDE thus becomes

$$\frac{\partial \mathbf{f}(r,t)}{\partial t} = \begin{pmatrix} 0 & 1 \\ \nabla^2 - \lambda^2 & -2\sqrt{\lambda^2 - \nabla^2} \end{pmatrix} \mathbf{f}(r,t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w(r,t), \tag{4.14}$$

where $r \in \mathbb{R}$ is the spatial variable and and $\nabla^2 = \frac{\partial^2}{\partial r^2}$. The white noise process $w(r,t)$ has a spectral density with scale factor $\gamma$. The one-dimensional SDE in Equation (3.26) can be seen as a special case of this model. Similarly, if $\nu = 2$, the SPDE becomes

$$\frac{\partial \mathbf{f}(r,t)}{\partial t} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -(\lambda^2 - \nabla^2)^{3/2} & -3(\lambda^2 - \nabla^2) & -3\sqrt{\lambda^2 - \nabla^2} \end{pmatrix} \mathbf{f}(r,t) + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} w(r,t). \tag{4.15}$$

The top row in Figure 4.2 shows two random draws from spatio-temporal Gaussian process priors with non-separable Matérn covariance functions ($\nu = 1$ and $\nu = 2$).

### Separable models

In many applications it is desired to model spatial and temporal interactions through a separable model, where the correlation structure obeys different dependencies through space and time.

If the covariance function is separable over space and time, such that $\kappa(\mathbf{r}, t; \mathbf{r}', t') = \kappa_\mathrm{r}(\mathbf{r}, \mathbf{r}')\,\kappa_\mathrm{t}(t, t')$, the evolution SPDE simplifies. If the models are stationary, the spectral density factorizes as well, $S(\boldsymbol{\omega}_\mathrm{r}, \omega_\mathrm{t}) = S(\boldsymbol{\omega}_\mathrm{r})\,S(\omega_\mathrm{t})$. The resulting model becomes an SDE of the form

$$\frac{\partial \mathbf{f}(\mathbf{r}, t)}{\partial t} = \mathbf{F}\,\mathbf{f}(\mathbf{r}, t) + \mathbf{L}\,\mathbf{w}(\mathbf{r}, t), \tag{4.16}$$

where the feedback operator is a matrix and the only spatial coupling comes from the spectral density of $\mathbf{w}(\mathbf{r}, t)$. Most conveniently, for a finite collection of spatial points of interest $\{\mathbf{r}_i\}_{i=1}^{n_\mathrm{r}}$, the state becomes $\mathbf{f}(t) = (\mathbf{f}(\mathbf{r}_1, t), \mathbf{f}(\mathbf{r}_2, t), \ldots, \mathbf{f}(\mathbf{r}_{n_\mathrm{r}}, t))$. This means that the state-space model is finite-dimensional (a multi-output model) and the dynamics are uncoupled:

$$\mathbf{F} = \mathbf{I} \otimes \mathbf{F}_\mathrm{t}, \quad \mathbf{L} = \mathbf{I} \otimes \mathbf{L}_\mathrm{t}, \quad \mathbf{H} = \mathbf{I} \otimes \mathbf{H}_\mathrm{t}, \tag{4.17}$$

where the subscript denotes the temporal model matrices. The spectral density has a spatial structure $\mathbf{Q}_\mathrm{c} = \mathbf{K} \otimes \mathbf{Q}_\mathrm{c,t}$, as does the initial state covariance $\mathbf{P}_0 = \mathbf{K} \otimes \mathbf{P}_\mathrm{0,t}$. The elements in the covariance matrix are $\mathbf{K}_{i,j} = \kappa_\mathrm{r}(\mathbf{r}_i, \mathbf{r}_j)$. This model allows for sequential inference to be done directly by the standard Kalman filter.

Figure 4.2 shows random draws from four separable models. The spatio-temporal squared exponential covariance function (*cf.* Eq. 3.27) is separable per definition (see Publication I for the corresponding representation). As for other models, in this setting any temporal model presented in Section 3.3 can be combined with virtually any spatial covariance function $\kappa_\mathrm{r}(\mathbf{r}, \mathbf{r}')$. The three other models in Figure 4.2 are the product of a spatial and temporal Matérn ($\nu = 3/2$), the product of a spatial and temporal exponential, and the product of a spatial squared exponential and temporal periodic covariance function.

# 5. Discussion

In the following the main contributions in the six publications are briefly summarized and discussed.

## 5.1 Temporal Models (I, III–V)

This work builds upon exploiting the Markovian structure in certain type of Gaussian process models, where the covariance function is either directly rewritten into a Markovian representation or approximated by one. The origin of considering this (both temporal and spatio-temporal) connection dates back as far as the concept of GP regression itself. Already in the discussion part of O'Hagan (1978), the comments bring up both Kalman filtering and that *'there is indeed a direct link with so-called distributed parameter linear systems theory'*. Even so, until recently this link has been largely overlooked in machine learning or only exploited by a few. For one-dimensional temporal GP models, the interest was revived by Hartikainen and Särkkä (2010) and Reece and Roberts (2010).

A common motivator throughout the publications is the 'big $n$' problem related to the $\mathcal{O}(n^3)$ computational complexity in GP models. In the reformulated models, the computational benefits are remarkable. In a Kalman filter solution the computational complexity scales as $\mathcal{O}(nm^3)$, where $n$ is the number of data points and $m$ is the dimensionality of the state. If $m \ll n$, this formulation becomes very beneficial. An illustration of practical computational times is shown, for example, in Publication III (Fig. 5). However, the overhead related to recursive evaluation of small matrix operations and highly optimized matrix libraries for large matrices make the computational benefits visible only when $n$ becomes large.

Sparse and other rank-reduction approximation methods (see Sec. 2.2) typically have a computational scaling of $\mathcal{O}(nm^2)$. However, for these

methods the number of inducing inputs $m$ (matrix rank) needs to be large enough to cover the input space. Therefore $m$ tends to grow with $n$ in the temporal dimension. Under the state-space approach, the state dimensionality is truly independent of $n$. On the other hand, the state dimensionality grows with model complexity as each model component needs to be assigned a dynamical model of its own.

Publication I considers a general framework for the connection between GP regression and state-space SDE models. It presents the general workflow for converting any stationary covariance function that has a rational spectral density (or can be approximated by one) to a corresponding state-space model. This work builds upon spectral factorization methods which date back to Wiener (1950). Thus the approach is limited to covariance functions with a spectral density (*i.e.* stationary models). These models are of central interest in time-series studies in machine learning. Methods that use the spectral density representation of stationary covariance functions have proven to be an efficient way of inferring the internal structure of time-series data (see, *e.g.*, Wilson and Adams, 2013; Tobar *et al.*, 2015).

Not all covariance functions are stationary nor have a spectral density of convenient form. Publication III extends the SDE link beyond models with a rational spectral density. The spectral density of the canonical periodic covariance function comprises delta peaks at the base and harmonic frequencies. The corresponding model has deterministic dynamics, where the stochastic behaviour only enters the system through the initial state.

State-space inference for periodic GP models has also been proposed by Reece *et al.* (2014). They set up their model by a periodic basis function approximation. However, their approximation scheme does not generalize well to quasi-periodicity as they lose the linearity of the state-space model. This means that it can no longer be solved analytically. Another state-space formulation is given by Luttinen *et al.* (2014), who present a method and extension which in methodological sense relates to the approach in Publication VI, where the frequency becomes time-varying.

A strength of building GP models is the possibility of creating new models from sums and products of old ones. Publication III presents a general link for doing this in terms of the corresponding SDE models. As a real-life example of a modelling task with sums and products of covariance functions, Publication III considers an example data set from Gelman *et al.* (2013) (featured on the cover of the third edition) for modelling the short-

and long-term trends and periodic yearly and weekly effects in the number of child births in the United States. Another standard example is the Mauna Loa $CO_2$ concentration data (see, *e.g.*, Rasmussen and Williams, 2006), for which the GP model is also converted and solved in Publication III.

The intuition for combining covariance functions is furthered in Publication IV which is concerned with scale-mixtures of squared-exponential kernels—such as the rational quadratic (RQ) covariance function. They can be approximated by a weighted set of state-space models of squared exponentials. The weights and hyperparameters of the sub-models are given by a Gaussian quadrature scheme. This link is interesting primarily because of its generality, while a direct approach to SDE representations of the RQ covariance could be written in terms of the approach in Publication I.

Publication V presents an extension to temporal GP regression in terms of Student-$t$ processes (TPs) with an entangled noise process. This approach is based on rewriting the TP inference problem presented by Shah *et al.* (2014) as a state-space model and deriving a sequential inference scheme tailored for this model. The inference scheme is not an approximation, but the closed-form solution to the TP regression problem similarly as the Kalman filter equations are the solution to the one-dimensional GP regression problem. In signal processing, heavy-tailed models have recently been studied by Piché *et al.* (2012) and Roth (2013), who have derived several approximative inference schemes for them. The exact algorithm for entangled noise in Publication V can be seen as a special case of the approximative scheme in Roth (2013).

The main interest of the TP regression model is in the filtering solution, where the marginal variance now depends on the observed values. The examples in Shah *et al.* (2014) and Publication V suggest that the marginal likelihood is less prone to the effect of outliers during hyperparameter optimization. This is demonstrated in electricity consumption and stock price forecasting, and tracking applications.

In TP regression, as the number of data points grows, the degrees of freedom tend to infinity and the model quickly starts resembling a GP. For actual robust inference, the measurement model should be changed. Extending the state-space inference scheme to general likelihoods (non-Gaussian measurement models) enables tasks such as robust inference or classification to be performed. However, then the inference prob-

lem no longer has a closed-form solution. The inference must be conducted by some approximative method—such as the Laplace approximation, expectation propagation, or variational methods (as discussed briefly in Sec. 2.2). The coal mining accident example (Publication IV, Fig. 4) demonstrates the combination of the SDE approach in a log-Gaussian Cox process with a Poisson likelihood. Non-linear filtering methods such as the extended Kalman filter or sigma-point schemes (see, *e.g.*, Särkkä, 2013) can be combined in the SDE framework equally well.

### 5.2 Spatio-Temporal Models (I–II, VI)

Sequential methods have been integrated into the statistical tool set for spatio-temporal modelling (see, *e.g.*, Cressie and Wikle, 2011). For example, Zammit-Mangion *et al.* (2012a) model the logarithm of the event intensity in the Afghan conflict during 2004–2009 as a spatio-temporal Gaussian process. The model is carefully crafted for the data, which makes it fit for the purpose but requires expertise and understanding in setting up every part of the model.

The models in this work aim at being general and enable model specification in terms of standard covariance function families. Therefore the SDE approach is extended to models where the 'state' is interpreted as a spatial Gaussian process and the dynamical behaviour over the temporal domain is encoded in terms of a stochastic partial differential equation. The central message is that evolution type SPDEs allow for representing a wide range of models—most notably, separable combinations of any spatial covariance function and the temporal models from Section 3.3. Good examples of the potential of this approach are the separable Matérn fields in Figure 4.2, which are not approximations but exact realizations from the GP prior.

The non-separable Matérn fields in Figure 4.2 are exact in terms of the model, but approximate in terms of the numerical implementation. In spatio-temporal models, the sole problem not the number of data points, but the required spatial resolution can also become a challenge. The question of dealing with the infinite-dimensional quantities in an efficient manner is interesting. In this work the emphasis was on basis function approximations for models that were given in terms of the Laplace operator. However, it is rather model and problem specific to choose a method which provides a principled approach for approximating the oper-

ator equations. The related approaches in Rue *et al.* (2009), Lindgren *et al.* (2011), and Cseke *et al.* (2015) use the spatial Markov property which is well-suited for low-dimensional spatial domains. The basis functions can also be chosen deterministically; Zammit-Mangion *et al.* (2012a) consider a set of spatially distributed Gaussian radial basis functions with compact support. All these methods work well with low-dimensional inputs, where the curse of dimensionality is not a problem.

In machine learning, the spatial input domains tend to be high-dimensional. Hartikainen (2013) discusses how to combine standard inducing point methods with the spatio-temporal framework. The methods discussed in Section 2.2 are directly applicable to the spatio-temporal setup presented in this work. In practice, they take down the computational complexity from $\mathcal{O}(n_\mathrm{t}\, n_\mathrm{r}^3)$ to $\mathcal{O}(n_\mathrm{t}\, n_\mathrm{r}\, m_\mathrm{r}^2)$, where the $n_\mathrm{r}$ and $n_\mathrm{t}$ denote the number of distinct data points in the spatial and temporal dimensions, and $m_\mathrm{r}$ the number of inducing inputs. In this formulation the inducing inputs should not be seen as points, but inducing temporal trajectories (GPs).

Publication I generalizes the work of Särkkä and Hartikainen (2012) to general stationary covariance functions and defines the framework presented in this overview. Most importantly it also covers the transformation of non-separable covariance functions of the Matérn class into corresponding evolution type SPDE models. The efficiency of the approach is demonstrated by modelling the precipitation in Colorado (Publication I, Fig. 2) over a time-span of ten years.

Publication II presents an SPDE approach for modelling spatio-temporal oscillations. The model is built upon a harmonic oscillator model, where both the oscillations and the damping terms are affected by spatial coupling. It can be seen as a spatio-temporal variant of the methods discussed in Publications III and VI. This model is applied to explain spatio-temporal temperature variations on the surface of the Earth, and to modelling spatio-temporal oscillations in brain data (see Figs. 2 and 3 in Publication II).

The oscillator model is driven by a latent Gaussian process. Therefore Publication II also underlines an important aspect in the differential equation formalism of Gaussian processes: Having a representation of GP models in terms of differential equations makes it straight-forward to combine Gaussian processes with existing (often deterministic) physical models. This type of models are known as *latent force models* (Álvarez and

Lawrence, 2009; Hartikainen and Särkkä, 2011). They provide a principled and data-driven way of combining mechanistic ordinary differential equation models with Gaussian processes.

A more applied point of view is taken in Publication VI. It is concerned with tailoring the methodology to fit a particular purpose. The application is in identification and estimation of periodic physiological noise in functional magnetic resonance imaging (fMRI) brain data. These heartbeat and respiration induced confounders are a major source of signal variability. Even though the interest is primarily in providing a useful tool for removal of structured noise, two interesting methodological considerations are discussed in the paper.

First, the time-varying nature of the frequency of the periodic signals makes the periodic model highly non-stationary. As explained in Section 3.4, the time-varying frequency can be modelled as a switching system and the frequencies seen as piece-wise constant between each observation. This is similar to the further work by Luttinen *et al.* (2014).

Second, the spatio-temporal model is structured such that each time-series in the volumetric grid is modelled as an independent realization of the same GP prior. Because the predictive variance only depends on the training input times (shared between the spatial observations), the sequential inference can be conducted in an efficient parallel manner. In the filter and smoother, only the means need to be stored and updated, while the covariance matrix is shared. This is the key for the DRIFTER method to solve the GP inference problem for millions of data points within seconds.

# 6. Conclusion

This work has concentrated on temporal and spatio-temporal Gaussian process regression models. The main aim of this thesis has been to combine the intuitive model specification from machine learning with computationally efficient signal processing methods. Consequently, the inference can be done in linear time with respect to the number of temporal data points by Kalman filtering methods.

To accomplish this, a central part of this work has been to rewrite Gaussian process models in terms of state-space equations. These representations are not based on *ad hoc* approximations, but directly rewriting the covariance function into a corresponding SDE/SPDE model. Widely used stationary and non-stationary models have been covered, all of which have in common that they share a linear time-invariant representation. These models can then be combined and extended much like GP models under the kernel formalism making them widely applicable for practical modelling.

Furthermore, this overview has also gone through a wide range of sequential inference methods connecting standard batch techniques and sequential Kalman filtering. The sequential methods tend to be trickier to implement and optimize in high-level programming languages. This however pays off, because the gain in computation times can be remarkable.

Even though the relationship between state-space models and GP regression has been known for a long time, the methods for actually utilizing this link appear not to be widely used throughout the machine learning community. Still, emphasizing this link more would not only help speed up computations, but also make the models easily approachable for audiences in other disciplines such as physics or signal processing.

# Bibliography


Milton Abramowitz and Irene Stegun. *Handbook of Mathematical Functions*. Dover Publishing, New York, 1970.

Naum I. Akhiezer and Izrail' M. Glazman. *Theory of Linear Operators in Hilbert Space*. Dover, New York, 1993.

Mauricio Álvarez and Neil D. Lawrence. Latent force models. In *Proceedings of the 12$^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR W&CP*, pages 9–16, 2009.

Christopher T.H. Baker. *The Numerical Treatment of Integral Equations*. Clarendon Press, Oxford, 1977.

Yaakov Bar-Shalom, Xiao-Rong Li, and Thiagalingam Kirubarajan. *Estimation with Applications to Tracking and Navigation*. Wiley-Interscience, New York, 2001.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, MA, 2004.

Gerald J. Bierman. *Factorization Methods for Discrete Sequential Estimation*. Academic Press, New York, 1977.

George E.P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecast and Control*. Wiley series in probability and statistics. John Wiley & Sons, Hoboken, NJ, fourth edition, 2008.

Pao-Liu Chow. *Stochastic Partial Differential Equations*, volume 11 of *Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series*. Chapman & Hall/CRC Press, Boca Raton, FL / London, 2007.

Noel Cressie and Christopher K. Wikle. Space-time Kalman filter. In Abdel H. El-Shaarawi and Walter W. Piegorsch, editors, *Encyclopedia of Environmetrics*, volume 4, pages 2045–2049. John Wiley & Sons, Chichester, 2002.

Noel Cressie and Christopher K. Wikle. *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Hoboken, NJ, 2011.

Noel A.C. Cressie. *Statistics for Spatial Data*. Wiley series in probability and mathematical statistics. Wiley-Interscience, New York, corrected edition, 1993.

Lehel Csató and Manfred Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–668, 2002.

Lehel Csató, Ernest Fokoué, Manfred Opper, Bernhard Schottky, and Ole Winther. Efficient approaches to Gaussian process classification. In *Advances in Neural Information Processing Systems 13 (NIPS)*, pages 251–257. MIT Press, 2000.

Botond Cseke, Andrew Zammit-Mangion, Tom Heskes, and Guido Sanguinetti. Sparse approximate inference for spatio-temporal point process models. *Journal of the American Statistical Association*, pages 1–52, 2015. In press.

Ruth Curtain. A survey of infinite-dimensional filtering. *SIAM Review*, 17(3): 395–411, 1975.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 45 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992.

Edward J. Davison and Fu-Tin Man. The numerical solution of $A'Q + QA = -C$. *IEEE Transactions on Automatic Control*, 13(4):448–449, 1968.

Michael Dewar, Kenneth Scerri, and Visakan Kadirkamanathan. Data-driven spatio-temporal modeling using the integro-difference equation. *IEEE Transactions on Signal Processing*, 57(1):83–91, 2009.

Arnaud Doucet. A note on efficient conditional simulation of Gaussian distributions. Technical note. Departments of Computer Science and Statistics, University of British Columbia, Canada, 2010.

Jochen Fritz, Insa Neuweiler, and Wolfgang Nowak. Application of FFT-based algorithms for large-scale universal kriging problems. *Mathematical Geosciences*, 41(5):509–533, 2009.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, Boca Raton, FL, third edition, 2013.

Mark N. Gibbs and David J.C. MacKay. Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464, 2000.

Mohinder S. Grewal and Angus P. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB*. Wiley-Interscience, New York, second edition, 2001.

Jouni Hartikainen. *Sequential Inference for Latent Temporal Gaussian Process Models*. Doctoral dissertation, Aalto University, Finland, 2013.

Jouni Hartikainen and Simo Särkkä. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 379–384, 2010.

Jouni Hartikainen and Simo Särkkä. Sequential inference for latent force models. In *Proceedings of the 27$^{th}$ Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 311–318. AUAI Press, 2011.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the 29$^{th}$ Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 282–290. AUAI Press, 2013.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008.

Andrew H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, 1970.

Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. Number 160 in Applied Mathematical Sciences. Springer, New York, 2005.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82(1):35–45, 1960.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, New York, 1991.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Jaakko Luttinen, Tapani Raiko, and Alexander Ilin. Linear state-space model with time-varying dynamics. In *Machine Learning and Knowledge Discovery in Databases*, volume 8725 of *Lecture Notes in Computer Science*, pages 338–353. Springer, 2014.

David J.C. MacKay. Introduction to Gaussian processes. In Christopher M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *NATO ASI Series F Computer and Systems Sciences*, pages 133–166. Springer, Berlin, 1998.

Bertil Matérn. Spatial variation: Stochastic models and their applications to some problems in forest surveys and other sampling investigations. *Meddelanden från statens skogsforskningsinstitut*, 49:1–144, 1960.

Isambi S. Mbalawata, Simo Särkkä, and Heikki Haario. Parameter estimation in stochastic differential equations with Markov chain Monte Carlo and nonlinear Kalman filtering. *Computational Statistics*, 28(3):1195–1223, 2013.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17$^{th}$ Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 362–369. Morgan Kaufmann Publishers, 2001.

Iain Murray, Ryan P. Adams, and David J.C. MacKay. Elliptical slice sampling. In *Proceedings of the 13$^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *JMLR W&CP*, pages 541–548, 2010.

Radford M. Neal. Regression and classification using Gaussian process priors. In *Bayesian Statistics*, volume 6, pages 475–501. Oxford University Press, 1998.

Anthony O'Hagan. Curve fitting and optimal design for prediction (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1): 1–42, 1978.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin, sixth edition, 2003.

Sigeru Omatu and John H. Seinfeld. *Distributed Parameter Systems: Theory and Applications*. Clarendon Press / Ohmsha, New York, 1989.

Christopher J. Paciorek. Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package. *Journal of Statistical Software*, 19(2):1–38, 2007.

Robert Piché, Simo Särkkä, and Jouni Hartikainen. Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate Student-$t$ distribution. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2012.

Joaquin Quiñonero-Candela and Carl E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

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

Herbert E. Rauch, F. Tung, and Charlotte T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

Steven Reece and Stephen Roberts. An introduction to Gaussian processes for the Kalman filter expert. In *Proceedings of the 13$^{th}$ Conference on Information Fusion (FUSION)*. IEEE, 2010.

Steven Reece, Siddhartha Ghosh, Alex Rogers, Stephen Roberts, and Nicholas R. Jennings. Efficient state-space inference of periodic latent force models. *Journal of Machine Learning Research*, 15:2337–2397, 2014.

Lassi Roininen, Sari Lasanen, Mikko Orispää, and Simo Särkkä. Sparse approximations of fractional Matérn fields. *arXiv preprint arXiv:1410.2113*, 2014.

Michael Roth. *Kalman Filters for Nonlinear Systems and Heavy-Tailed Noise*. Licentiate thesis, Department of Electrical Engineering, Linköping University, Sweden, 2013.

Håvard Rue, Sara Martino, and Nicolas 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.

Yunus Saatçi. *Scalable Inference for Structured Gaussian Process Models*. PhD thesis, University of Cambridge, UK, 2012.

Simo Särkkä. *Recursive Bayesian Inference on Stochastic Differential Equations*. Doctoral dissertation, Helsinki University of Technology, Finland, 2006.

Simo Särkkä. Linear operators and stochastic partial differential equations in Gaussian process regression. In *Artificial Neural Networks and Machine Learning – ICANN 2011*, number 6792 in Lecture Notes in Computer Science, pages 151–158. Springer, 2011.

Simo Särkkä. *Bayesian Filtering and Smoothing*, volume 3 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, Cambridge, 2013.

Simo Särkkä and Jouni Hartikainen. Infinite-dimensional Kalman filtering approach to spatio-temporal Gaussian process regression. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22 of *JMLR W&CP*, pages 993–1001, 2012.

Simo Särkkä and Robert Piché. On convergence and accuracy of state-space approximations of squared exponential covariance functions. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2014.

Simo Särkkä and Arno Solin. Applied stochastic differential equations. Lecture notes available as a technical report (accessed: 2015-12-08). Aalto University, Finland, 2014. URL `http://becs.aalto.fi/~ssarkka/course_s2014/sde_course_booklet.pdf`.

Kenneth Scerri, Michael Dewar, and Visakan Kadirkamanathan. Estimation and model selection for an IDE-based spatio-temporal model. *IEEE Transactions on Signal Processing*, 57(2):482–492, 2009.

Michael Schober, David K. Duvenaud, and Philipp Hennig. Probabilistic ODE solvers with Runge–Kutta means. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 739–747. Curran Associates, 2014.

Matthias Seeger, Christopher K.I. Williams, and Neil D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2003.

Amar Shah, Andrew G. Wilson, and Zoubin Ghahramani. Student-$t$ processes as alternatives to Gaussian processes. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 33 of *JMLR W&CP*, pages 877–885, 2014.

Ralph E. Showalter. *Hilbert Space Methods in Partial Differential Equations*. Dover Publications, Mineola, NY, 2010.

Mikhail A. Shubin. *Pseudodifferential Operators and Spectral Theory*. Springer Series in Soviet Mathematics. Springer, Berlin, 1987.

Alex J. Smola and Peter Bartlett. Sparse greedy Gaussian process regression. In *Advances in Neural Information Processing Systems 13 (NIPS)*, pages 619–625. MIT Press, 2001.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18 (NIPS)*, pages 1259–1266. MIT Press, 2006.

Arno Solin. Hilbert space methods in infinite-dimensional Kalman filtering. Master's thesis, School of Science, Aalto University, Finland, 2012.

Arno Solin and Simo Särkkä. Hilbert space methods for reduced-rank Gaussian process regression. *arXiv preprint arXiv:1401.5508*, 2014.

Michael L. Stein. *Interpolation of Spatial Data*. Springer Series in Statistics. Springer, New York, 1999.

Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2005.

Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR W&CP*, pages 567–574, 2009.

Felipe Tobar, Thang D. Bui, and Richard E. Turner. Learning stationary time series using Gaussian processes with nonparametric kernels. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3483–3491. Curran Associates, 2015.

Spyros G. Tzafestas. Distributed parameter state estimation. In Willis H. Ray and Demetrios G. Lainiotis, editors, *Distributed Parameter Systems: Identification, Estimation and Control*. Marcel Dekker, New York, 1978.

Harry L. Van Trees. *Detection, Estimation, and Modulation Theory, Part II: Nonlinear Modulation Theory*. John Wiley & Sons, New York, 1971.

Grace Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

Norbert Wiener. *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. John Wiley & Sons, New York, 1950.

Christopher K.I. Williams and David Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

Christopher K.I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13 (NIPS)*, pages 682–688. MIT Press, 2001.

Andrew G. Wilson and Ryan P. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *JMLR W&CP*, pages 1067–1075, 2013.

Andrew Zammit-Mangion, Michael Dewar, Visakan Kadirkamanathan, and Guido Sanguinetti. Point process modelling of the Afghan war diary. *Proceedings of the National Academy of Sciences*, 109(31):12414–12419, 2012a.

Andrew Zammit-Mangion, Guido Sanguinetti, and Visakan Kadirkamanathan. Variational estimation in spatiotemporal systems from continuous and point-process observations. *IEEE Transactions on Signal Processing*, 60(7):3449–3459, 2012b.

Aalto-DD 50/2016

ISBN 978-952-60-6710-0 (printed)
ISBN 978-952-60-6711-7 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

**Aalto University**
**School of Science**
**Department of Neuroscience and Biomedical Engineering /**
**Department of Computer Science**
**www.aalto.fi**

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

**DOCTORAL**
**DISSERTATIONS**