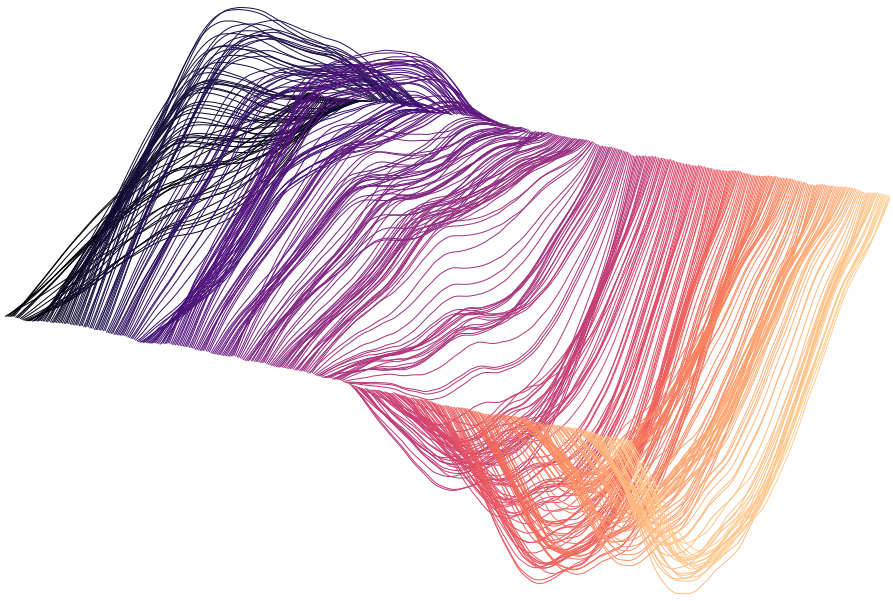
Department of Electrical Engineering and Automation

State-space deep Gaussian processes with applications

Zheng Zhao



Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 163/2021

State-space deep Gaussian processes with applications

Zheng Zhao

With the permission of the Aalto University School of Electrical Engineering, the defence of this doctoral thesis completed for the degree of Doctor of Science (Technology) will be held on 10 December 2021 at noon. The defence will take place simultaneously at a public examination in the lecture hall AS1 (Maarintie 8, Espoo) of the school and remotely via the link https://aalto.zoom.us/j/67529212279.

Aalto University
School of Electrical Engineering
Department of Electrical Engineering and Automation
Sensor Informatics and Medical Technology

Supervising professor
Prof. Simo Särkkä, Aalto University, Finland

Thesis advisor
Prof. Simo Särkkä, Aalto University, Finland

Preliminary examiners
Prof. Kody J. H. Law, The University of Manchester, United Kingdom
Prof. David Duvenaud, University of Toronto, Canada

Opponent
Prof. Manfred Opper, University of Birmingham, United Kingdom

Aalto University publication series
DOCTORAL DISSERTATIONS 163/2021

ISBN 978-952-64-0602-2 (printed)
ISBN 978-952-64-0603-9 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0603-9

Images: A realisation of a Wiener process taking value in a Sobolev space with Dirichlet boundary condition

Unigrafia Oy
Helsinki 2021

Finland

Printed matter
4041-0619

Aalto University

Abstract

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

Author
Zheng Zhao

Name of the doctoral dissertation
State-space deep Gaussian processes with applications

Publisher School of Electrical Engineering

Unit Department of Electrical Engineering and Automation

Series Aalto University publication series DOCTORAL DISSERTATIONS 163/2021

Field of research

Manuscript submitted 10 August 2021 **Date of the defence** 10 December 2021

Permission for public defence granted (date) 28 October 2021 **Language** English

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

Abstract

This thesis is mainly concerned with state-space approaches for solving deep (temporal) Gaussian process (DGP) regression problems. More specifically, we represent DGPs as hierarchically composed systems of stochastic differential equations (SDEs), and we consequently solve the DGP regression problem by using state-space filtering and smoothing methods. The resulting state-space DGP (SS-DGP) models generate a rich class of priors compatible with modelling a number of irregular signals/functions. Moreover, due to their Markovian structure, SS-DGPs regression problems can be solved efficiently by using Bayesian filtering and smoothing methods. The second contribution of this thesis is that we solve continuous-discrete Gaussian filtering and smoothing problems by using the Taylor moment expansion (TME) method. This induces a class of filters and smoothers that can be asymptotically exact in predicting the mean and covariance of stochastic differential equations (SDEs) solutions. Moreover, the TME method and TME filters and smoothers are compatible with simulating SS-DGPs and solving their regression problems. Lastly, this thesis features a number of applications of state-space (deep) GPs. These applications mainly include, (i) estimation of unknown drift functions of SDEs from partially observed trajectories and (ii) estimation of spectro-temporal features of signals.

Keywords Gaussian processes, stochastic differential equations, stochastic filtering and smoothing, state-space methods, signal processing, machine learning

ISBN (printed) 978-952-64-0602-2 **ISBN (pdf)** 978-952-64-0603-9

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2021

Pages 246 **urn** http://urn.fi/URN:ISBN:978-952-64-0603-9

Preface

The research work in this thesis has been carried out in the Department of Electrical Engineering and Automation, Aalto University, during the years 2018-2021. My doctoral studies officially started in April of 2018, while most of the pivotal work came in 2020-2021. During this time, my doctoral research was financially supported by Academy of Finland and Aalto ELEC Doctoral School. The Aalto Scientific Computing team and the Aalto Learning Center also provided useful computational and literature resources for my studies. I particularly enjoyed the Spring, Autumn, and Winter in Finland, which allowed me to find inner peace and focus on my research.

I would like to offer my greatest gratitude to Prof. Simo Särkkä who is my supervisor and mentor, and without whom this work would never have been possible. After finishing my master studies in Beijing University of Technology in 2017, I found myself lost in finding a "meaningful" way of life in the never-sleeping metropolis that is Beijing. This quest was fulfilled when Simo offered me the opportunity of pursuing a doctoral degree under his supervision. Disregarding my bewilderment on the research path in the beginning, Simo's patience and valuable guidance led me to a research area that I am fascinated in. Over the years, Simo's help, support, and friendship have helped me become a qualified and independent researcher. I think very highly of Simo's supervision, and I almost surely could not have found a better supervisor.

During my years in the campus, I owe a great thanks to Rui Gao (高 睿) who is a brilliant, learnt, and erudite researcher.

I would like to thank these few people that have accompanied me through joy and sorrow, I name: Adrien Corenflos and Christos Merkatas. I thank you for the friendship and relieving me from solitude[1].

During my years in Aalto university, I have shared my office with Marco Soldati, Juha Sarmavuori, Janne Myllärinen, Fei Wang (王 斐), Jiaqi Liu (劉 佳琦), Ajinkya Gorad, Masaya Murata (村田 眞哉), and Otto Kangasmaa.

[1] This was written under constraint.

I thank them all for filling the office with happiness and joy. I especially thank Marco Soldati who offered me honest friendship, lasagne, and taught me many useful Italian phrases. My thanks also go to Lauri Palva, Zenith Purisha, Joel Jaskari, Sakira Hassan, Fatemeh Yaghoobi, Abubakar Yamin, Zaeed Khan, Xiaofeng Ma (馬 曉峰), Prof. Ivan Vujaklija, Dennis Yeung, Wendy Lam, Prof. Ilkka Laakso, Marko Mikkonen, Noora Matilainen, Juhani Kataja, Linda Srbova, and Tuomas Turunen. All these amazing people made working at Aalto a real pleasure. I would also like to give my thanks to Laila Aikala who kindly offered me a peaceful place to stay in Espoo.

I warmly thank Prof. Leo Kärkkäinen for the collaboration on the AI in Health Technology course and our inspiring discussions on many Thursdays and Fridays. I particularly enjoyed the collaboration with Muhammad Fuady Emzir who offered me knowledge generously and with no reservations. Many thanks go to my coauthors Prof. Roland Hostettler, Prof. Ali Bahrami Rad, Filip Tronarp, and Toni Karvonen. I also appreciated the collaboration with Sarang Thombre and Toni Hammarberg from Finnish Geospatial Research Institute, Prof. Ville V. Lehtola from University of Twente, and Tuomas Lumikari from Helsinki University Hospital. I also thank Prof. Lassi Roininen and Prof. Arno Solin for their time and valuable advice.

Lastly, I would like to thank my parents and sister who support me persistently as always.

Helsinki, October 4, 2021,

Zheng Zhao
趙 正

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 Zheng Zhao, Toni Karvonen, Roland Hostettler, and Simo Särkkä. Taylor moment expansion for continuous-discrete Gaussian filtering. *IEEE Transactions on Automatic Control*, Volume 66, Issue 9, Pages 4460–4467, December 2020.

II Zheng Zhao, Muhammad Emzir, and Simo Särkkä. Deep state-space Gaussian processes. *Statistics and Computing*, Volume 31, Issue 6, Article number 75, Pages 1–26, September 2021.

III Zheng Zhao, Simo Särkkä, and Ali Bahrami Rad. Kalman-based spectro-temporal ECG analysis using deep convolutional networks for atrial fibrillation detection. *Journal of Signal Processing Systems*, Volume 92, Issue 7, Pages 621–636, April 2020.

IV Zheng Zhao, Filip Tronarp, Roland Hostettler, and Simo Särkkä. State-space Gaussian process for drift estimation in stochastic differential equations. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, Pages 5295–5299, May 2020.

V Zheng Zhao, Simo Särkkä, and Ali Bahrami Rad. Spectro-temporal ECG analysis for atrial fibrillation detection. In *Proceedings of the IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark, 6 pages, September 2018.

VI Sarang Thombre, Zheng Zhao, Henrik Ramm-Schmidt, José M. Vallet García, Tuomo Malkamäki, Sergey Nikolskiy, Toni Hammarberg, Hiski Nuortie, M. Zahidul H. Bhuiyan, Simo Särkkä, and Ville V. Lehtola. Sensors and AI techniques for situational awareness in

autonomous ships: a review. Accepted for publication in *IEEE Transactions on Intelligent Transportation Systems*, 20 pages, September 2020.

VII Zheng Zhao, Rui Gao, and Simo Särkkä. Hierarchical Non-stationary temporal Gaussian processes with L^1-regularization. Submitted to *Statistics and Computing*, May 2021.

Author's contribution

Publication I: "Taylor moment expansion for continuous-discrete Gaussian filtering"

Zheng Zhao wrote the article and produced the results. The stability analysis is mainly due to Toni Karvonen. Roland Hostettler gave useful comments. Simo Särkkä contributed the idea.

Publication II: "Deep state-space Gaussian processes"

Zheng Zhao wrote the article and produced the results. Muhammad Emzir and Simo Särkkä gave useful comments.

Publication III: "Kalman-based spectro-temporal ECG analysis using deep convolutional networks for atrial fibrillation detection"

Zheng Zhao wrote the article and produced the results. Ali Bahrami Rad helped with the experiments. Simo Särkkä came up with the spectro-temporal idea.

Publication IV: "State-space Gaussian process for drift estimation in stochastic differential equations"

Zheng Zhao wrote the article and produced the results. Filip Tronarp provided codes for the iterated posterior linearisation filter. Roland Hostettler gave useful comments. Idea was due to Simo Särkkä.

Publication V: "Spectro-temporal ECG analysis for atrial fibrillation detection"

Zheng Zhao wrote the article and produced the results. Ali Bahrami Rad helped with the experiments. Simo Särkkä came up with the spectro-temporal idea.

Publication VI: "Sensors and AI techniques for situational awareness in autonomous ships: a review"

Zheng Zhao wrote the reviews of AI techniques and produced corresponding results.

Publication VII: "Hierarchical Non-stationary temporal Gaussian processes with L^1-regularization"

Zheng Zhao wrote the article and produced the results. Rui Gao contributed the convergence analysis. Simo Särkkä gave useful comments.

Language check

The language of my dissertation has been checked by Adrien Corenflos, Christos Merkatas, and Dennis Yeung. I have personally examined and accepted/rejected the results of the language check one by one. This has not affected the scientific content of my dissertation.

Abbreviations

CD-FS	Continuous-discrete filtering and smoothing
DGP	Deep Gaussian process
GFS	Gaussian approximated density filter and smoother
GMRF	Gaussian Markov random field
GP	Gaussian process
Itô-1.5	Itô–Taylor strong order 1.5
LCD	Locally conditional discretisation
MAP	Maximum a posteriori
MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimation
NSGP	Non-stationary Gaussian process
ODE	Ordinary differential equation
PDE	Partial differential equation
RBF	Radial basis function
R-DGP	Regularised (batch) deep Gaussian process
R-SS-DGP	Regularised state-space deep Gaussian process
RTS	Rauch–Tung–Striebel
SDE	Stochastic differential equation
SS-DGP	State-space deep Gaussian process
SS-GP	State-space Gaussian process
TME	Taylor moment expansion

Symbols

a	Drift function of SDE
A	Drift matrix of linear SDE
$\mathscr{A}$	Infinitesimal generator
$\overline{\mathscr{A}}$	Multidimensional infinitesimal generator
b	Dispersion function of SDE
B	Dispersion matrix of linear SDE
c	Constant
$\mathscr{C}^k(\Omega;\Pi)$	Space of k times continuously differentiable functions on Ω mapping to Π
$C(t,t')$	Covariance function
$C_{\text{Mat.}}(t,t')$	Matérn covariance function
$C_{\text{NS}}(t,t')$	Non-stationary Matérn covariance function
$C_{1:T}$	Covariance/Gram matrix by evaluating the covariance function $C(t,t')$ on Cartesian grid $(t_1,\ldots,t_T)\times(t_1,\ldots,t_T)$
Cov	Covariance
$\text{Cov}[X \mid Y]$	Conditional covariance of random variable X given another random variable Y
$\text{Cov}[X \mid y]$	Conditional covariance of random variable X given the realisation y of random variable Y
d	Dimension of state variable
d_i	Dimension of the i-th GP element
d_y	Dimension of measurement variable

$\det$	Determinant
diag	Diagonal matrix
$\mathbb{E}$	Expectation
$\mathbb{E}[X \mid \mathscr{F}]$	Conditional expectation of X given sigma-algebra $\mathscr{F}$
$\mathbb{E}[X \mid Y]$	Conditional expectation of X given the sigma-algebra generated by random variable Y
$\mathbb{E}[X \mid y]$	Conditional expectation of X given the realisation y of random variable Y
f	Approximate transition function in discrete state-space model
f^M	M-order TME approximated transition function in discrete state-space model
$\check{f}$	Exact transition function in discrete state-space model
$\mathring{f}_j$	j-th frequency component
$\mathscr{F}$	Sigma-algebra
$\mathscr{F}_t$	Filtration
$\mathscr{F}_t^W$	Filtration generated by W and initial random variable
g	Transformation function
$\mathrm{GP}(0, C(t,t'))$	Zero-mean Gaussian process with covariance function $C(t,t')$.
h	Measurement function
H	Measurement matrix
$\mathrm{H}_x f$	Hessian matrix of f with respect to x
I	Identity matrix
J	Set of conditional dependencies of GP elements
$\mathrm{J}_x f$	Jacobian matrix of f with respect to x
K	Kalman gain
K_ν	Modified Bessel function of the second kind with parameter ν
ℓ	Length scale parameter

$\mathscr{L}^{\mathrm{A}}$	Augmented Lagrangian function
$\mathscr{L}^{\mathrm{B}}$	MAP objective function of batch DGP
$\mathscr{L}^{\mathrm{B-REG}}$	L^1-regularisation term for batch DGP
$\mathscr{L}^{\mathrm{S}}$	MAP objective function of state-space DGP
$\mathscr{L}^{\mathrm{S-REG}}$	L^1-regularisation term for state-space DGP
$m(t)$	Mean function
m_k^-	Predictive mean at time t_k
m_k^f	Filtering mean at time t_k
m_k^s	Smoothing mean at time t_k
M	Order of Taylor moment expansion
N	Order of Fourier expansion
$\mathrm{N}(x \mid m, P)$	Normal probability density function with mean m and covariance P
$\mathbb{N}$	Set of natural numbers
O	Big O notation
$p_X(x)$	Probability density function of random variable X
$p_{X \mid Y}(x \mid y)$	Conditional probability density function of X given Y taking value y
P_k^-	Predictive covariance at time t_k
P_k^f	Filtering covariance at time t_k
P_k^s	Smoothing covariance at time t_k
$P_k^{i,j}$	Filtering covariance of the i and j-th state elements at time t_k
$\mathbb{P}$	Probability measure
q_k	Approximate process noise in discretised state-space model at time t_k
$\check{q}_k$	Exact process noise in discretised state-space model at time t_k
Q_k	Covariance of process noise q_k

$R_{M,\phi}$	Remainder of M-order TME approximation for target function ϕ
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_{>0}$	Set of positive real numbers
$\mathbb{R}_{<0}$	Set of negative real numbers
sgn	Sign function
$\mathcal{S}_{m,P}$	Sigma-point approximation of Gaussian integral with mean m and covariance P
t	Temporal variable
tr	Trace
t_0	Initial time
T	Number of measurements
$\mathbb{T}$	Temporal domain $\mathbb{T} \coloneqq [t_0, \infty)$
U	(State-space) GP
$U^i_{j_i}$	(State-space) GP element in $\mathcal{V}$ indexed by i, and it is also a parent of the j_i-th GP element in $\mathcal{V}$
$U_{1:T}$	Collection of $U(t_1), U(t_2), \ldots, U(t_T)$
$\mathcal{U}^i$	Collection of parents of $U^i_{j_i}$
V	(State-space) deep GP
V_k	Shorthand of $V(t_k)$
$V_{1:T}$	Collection of $V(t_1), V(t_2), \ldots, V(t_T)$
$\mathcal{V}$	Collection of GP elements
Var	Variance
w	Dimension of Wiener process
W	Wiener process
X	Stochastic process
X_0	Initial random variable
X_k	Shorthand of $X(t_k)$
Y_k	Measurement random variable at time t_k

$Y_{1:T}$	Collection of $Y_1, Y_2, \ldots, Y_T$
γ	Dimension of the state variable of Matérn GP
Γ	Shorthand of $b(x)b(x)^\mathsf{T}$
$\mathit{\Gamma}$	Gamma function
Δt	Time interval $t - s$
Δt_k	Time interval $t_k - t_{k-1}$
η	Multiplier for augmented Lagrangian function
θ	Auxiliary variable used in augmented Lagrangian function
Θ_r	r-th polynomial coefficient in TME covariance approximation
$\lambda_{\min}$	Minimum eigenvalue
$\lambda_{\max}$	Maximum eigenvalue
$\Lambda(t)$	Solution of a matrix ordinary differential equation
$\mathbf{\Lambda}(t,s)$	Shorthand of $\Lambda(t)(\Lambda(s))^{-1}$
ξ_k	Measurement noise at time t_k
Ξ_k	Variance of measurement noise ξ_k
ρ	Penalty parameter in augmented Lagrangian function
σ	Magnitude (scale) parameter
Σ_M	M-order TME covariance approximant
ϕ	Target function
ϕ_{ij}	i,j-th element of ϕ
ϕ^{I}	$\phi^{\mathrm{I}}(x) \coloneqq x$
ϕ^{II}	$\phi^{\mathrm{II}}(x) \coloneqq x\,x^\mathsf{T}$
Φ	Sparsity inducing matrix
$\chi(\Delta t)$	Polynomial of Δt associated with TME covariance approximation
Ω	Sample space
$(\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P})$	Filtered probability space with sample space Ω, sigma-algebra $\mathscr{F}$, filtration $\mathscr{F}_t$, and probability measure $\mathbb{P}$

$\lvert\cdot\rvert$	Absolute value
$\lVert\cdot\rVert_p$	L^p norm or L^p-induced matrix norm
$\lVert\cdot\rVert_G$	Euclidean norm weighted by a non-singular matrix G
$\nabla_x f$	Gradient of f with respect to x
$\binom{\cdot}{\cdot}$	Binomial coefficient
$\langle\cdot,\cdot\rangle$	Inner product
$\circ$	Mapping composition
$:=$	By definition
$\times$	Cartesian product
$a \wedge b$	Minimum of a and b

1. Introduction

In signal processing, statistics, and machine learning, it is common to consider that noisy measurements/data are generated from a latent, unknown, function. In statistics, this is often regarded as a regression problem over the space of functions. Specifically, Bayesian statistics impose a prior belief over the latent function of interest in the form of a probability distribution. It is therefore of vital importance to choose the prior appropriately, since it will encode the characteristics of the underlying function. In recent decades, Gaussian processes[1] (GPs, Rasmussen and Williams, 2006) have become a popular family of prior distributions over functions, and they have been used successfully in numerous applications (Roberts et al., 2013; Hennig et al., 2015; Kocijan, 2016).

Formally, GPs are function-valued random variables that have Gaussian distributions fully determined by their mean and covariance functions. The choice of mean and covariance functions is in itself arbitrary, which allows for representing functions with various properties. As an example, Matérn covariance functions are used as priors to functions with different degrees of differentiability (Rasmussen and Williams, 2006). However, the use of GPs in practice usually involves two main challenges.

The first challenge lies in the expensive *computational cost* of training and parameter estimation. Due to the necessity of inverting covariance matrices during the learning phase, the computational complexity of standard GP regression and parameter estimation is cubic in the number of measurements. This makes GP computationally infeasible for large-scale datasets. Moreover, when the sampled data points are densely located, the covariance matrices that need inversion may happen to be numerically singular or close to singular, making the learning process unstable.

The second challenge is related to modelling of irregular functions, such as piecewise smooth functions, or functions that have time-varying features

[1] In the statistics and applied probability literature, Gaussian processes can also be found under the name of Gaussian fields, in particular when they are multidimensional in the input. Depending on the context, we may use one or the other terminology interchangeably.

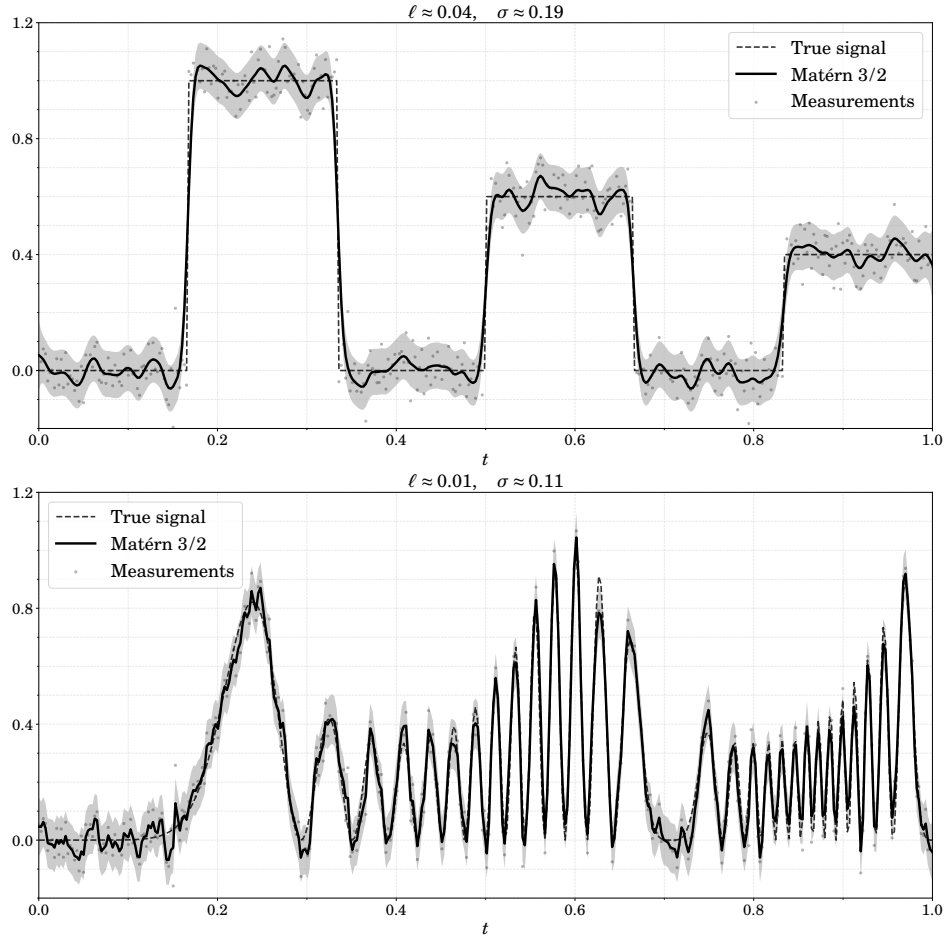


Figure 1.1. Matérn $\nu = 3/2$ GP regression on a magnitude-varying rectangular signal (top) and a composite sinusoidal signal (bottom). The parameters ℓ and σ are learnt by maximum likelihood estimation. The figures are taken from Zhao et al. (2021a).

(e.g., frequency or volatility). Many commonly-used GPs (e.g., with Matérn covariance functions) fail to cover these irregular functions mainly because their probability distributions are invariant under translation (i.e., they are said to be *stationary*). This behaviour is illustrated in Figure 1.1, where we show that a Matérn GP poorly fits two irregular functions (i.e., a rectangular signal and a composite sinusoidal signal), because the GP's parameters/features are assumed to be constant over time. Specifically, in the rectangular signal example, in order to model the discontinuities, the Matérn GP recovers a small global length scale ($\ell \approx 0.04$) which results in poor fitting in the continuous and flat parts. Similarly, in the composite sinusoidal signal example, the GP learns a small global length scale ($\ell \approx 0.01$) in order to model the high-frequency sections of the signal. This too results in poor fitting the low-frequency section of the signal.

The main aim of this thesis is thus to introduce a new class of non-stationary (Gaussian) Markov processes, that we name *state-space deep*

Gaussian processes (SS-DGPs)[2]. These are able to address the computational and non-stationarity challenges aforementioned, by hierarchically composing the state-space representations of GPs. Indeed, SS-DGPs are computationally efficient models due to their Markovian structure. More precisely, this means that the resulting regression problem can be solved in linear computational time (with respect to the number of measurements) by using Bayesian filtering and smoothing methods. Moreover, due to their hierarchical nature, SS-DGPs are capable of changing their features/characteristics (e.g., length scale) over time, thereby inducing a rich class of priors compatible with irregular functions. The thesis ends with a collection of applications of state-space (deep) GPs.

1.1 Bibliographical notes

In this section we provide a short and non-exhaustive review of related works in the GP literature. In particular we will focus on works that consider specifically reducing their computational complexity and allowing the non-stationarity in GPs.

Scalable Gaussian processes

We now give a list of GP methods and approximations that are commonly used to reduce the computational costs of GP regression and parameter learning.

Sparse approximations of Gaussian processes
Sparse GPs approximate full-rank GPs with sparse representations by using, for example, inducing points (Snelson and Ghahramani, 2006), subsets of data (Snelson and Ghahramani, 2007; Csató and Opper, 2002), or approximations of marginal likelihoods (Titsias, 2009), mostly relying on so-called pseudo-inputs. These approaches can reduce the computational complexity to quadratic in the number of pseudo-inputs and linear in the number of data points. In practice, the number and position of pseudo-inputs used in sparse representation must either be assigned by human experts or learnt from data (Hensman et al., 2013). For more comprehensive reviews of sparse GPs, see, for example, Quiñonero-Candela and Rasmussen (2005); Chalupka et al. (2013); Liu et al. (2020).

[2]Please note that although the name includes the term Gaussian, SS-DGPs are typically not Gaussian distributed, but instead hierarchically conditionally Gaussian, hence the name.

Gaussian Markov random fields
Gaussian Markov random fields (GMRFs, Rue and Held, 2005) are indexed collections of Gaussian random variables that have a Markov property (defined on graph). They are computationally efficient models because their precision matrices are sparse by construction. Methodologies for solving the regression and parameter learning problems on GMRFs can be found, for example, in Rue and Martino (2007); Rue et al. (2009). However, GMRFs are usually only approximations of Gaussian fields (see, e.g., Rue and Held, 2005, Chapter 5), although explicit representations exist for some specific Gaussian fields (Lindgren et al., 2011).

State-space representations of Gaussian processes
State-space Gaussian processes (SS-GPs) are (temporal) Markov GPs that are solutions of stochastic differential equations (SDEs, Särkkä et al., 2013; Särkkä and Solin, 2019). Due to their Markovian structure, probability distributions of SS-GPs factorise sequentially in the time dimension. The regression problem can therefore be solved efficiently in linear time with respect to the number of data points. Moreover, leveraging the sparse structure of the precision matrix (Grigorievskiy et al., 2017), or leveraging the associativity of the Kalman filtering and smoothing operations (Corenflos et al., 2021b) can lead to a sublinear computational complexity.

Other data-scalable Gaussian processes
Rasmussen and Ghahramani (2002); Meeds and Osindero (2006) form mixtures of GPs by splitting the dataset into batches resulting in a computational complexity that is cubic in the batch size. This methodology can further be made parallel (Zhang and Williamson, 2019). Lázaro-Gredilla et al. (2010) approximate stationary GPs with sparse spectral representations (i.e., trigonometric expansions). Gardner et al. (2018) and Wang et al. (2019) use conjugate gradients and stochastic trace estimation to efficiently compute the marginal log-likelihood of standard GPs, as well as their gradients with respect to parameters, resulting in a quadratic computational complexity in the number of data points.

Non-stationary Gaussian processes

In the below we give a list of methods that are introduced in order to induce non-stationarity in GPs.

Non-stationary covariance function-based Gaussian processes
Non-stationary covariance functions can be constructed by making their parameters (e.g., length scale or magnitude) depend on the data position. For instance, Gibbs (1997) and Higdon et al. (1999) present specific examples of covariance functions where the length scale parameter depends on the

spatial location. On the other hand, Paciorek and Schervish (2004, 2006) generalise these constructions to turn any stationary covariance function into a non-stationary one. There also exist some other non-stationary covariance functions, such as the polynomial or neural network covariance functions (Williams, 1998; Rasmussen and Williams, 2006) that can also give non-stationary GPs, but we do not review them here as they are not within the scope of this thesis.

Composition-based Gaussian processes
Sampson and Guttorp (1992); Schmidt and O'Hagan (2003); Rasmussen and Williams (2006) show that it is possible to construct a non-stationary GP as the pullback of an existing stationary GP by a non-linear transformation. Formally, given a stationary GP $U : E \to \mathbb{R}$, one can find a suitable transformation $\Upsilon : \mathbb{T} \to E$, such that the composition $U \circ \Upsilon : \mathbb{T} \to \mathbb{R}$ is a non-stationary GP on $\mathbb{T}$. For example, Calandra et al. (2016) and Wilson et al. (2016) choose Υ as neural networks.

Warping-based Gaussian processes
Conversely to the composition paradigm above, it is also possible to transform GPs the other way around, that is, to consider that GPs are the transformations of some non-Gaussian processes by non-linear functions (Snelson et al., 2004). Computing the marginal log-likelihood function of these warped GPs is then done by leveraging the change-of-variables formula for Lebesgue integrals (when it applies). However, the warping can be computationally demanding as the change-of-variables formula requires computing the inverse determinant of the transformation Jacobian. This issue can be mitigated, for example, by writing the the warping scheme with multiple layers of elementary functions which have explicit inverses (Rios and Tobar, 2019).

Deep Gaussian processes

The deterministic constructions for introducing non-stationarity GPs can be further extended in order to give a class of non-stationary non-Gaussian processes that can also represent irregular functions. While they are different in structure, the three subclasses of models presented below are usually all referred as deep Gaussian processes (DGPs) in literature.

Composition-based deep Gaussian processes
Lázaro-Gredilla (2012) extends the aforementioned pullback idea by taking $\Upsilon : \mathbb{T} \to E$ to be a GP instead of a deterministic mapping in order to overcome the overfitting problem. Resulting compositions of the form $U \circ \Upsilon : \mathbb{T} \to \mathbb{R}$ may not necessarily be GPs anymore but may provide a more flexible family of priors than that of deterministic compositions. This construction can be

done recursively leading to a subclass of DGPs (Damianou and Lawrence, 2013). However, the training of these DGPs is found to be challenging and requires approximate inference methods (Bui et al., 2016; Salimbeni and Deisenroth, 2017a). Moreover, Duvenaud (2014); Duvenaud et al. (2014) show that increasing the depth of DGPs can lead to a representation pathology, where samples of DGPs tend to be flat in high probability and exhibit sudden jumps. This problem can be mitigated by making their latent GP components explicitly depend on their original inputs (Duvenaud et al., 2014).

Hierarchical parametrisation-based deep Gaussian processes
A similar idea to compositional DGPs is to model the parameters of GPs as latent GPs. The posterior distribution of the joint model can then be computed by successive applications of Bayes' rule. As an example, Roininen et al. (2019) consider putting a GP prior on the length scale parameter of a Matérn GP and use Metropolis-within-Gibbs to sample from the posterior distribution. Similarly, Salimbeni and Deisenroth (2017b) model the length scale parameter of the non-stationary covariance function introduced by Paciorek and Schervish (2004) as a GP, but use a variational approximation to approximate its posterior distribution. Other sampling techniques to recover the posterior distribution of these models can be found, for example, in Heinonen et al. (2016); Monterrubio-Gómez et al. (2020).

Zhao et al. (2021a) and Emzir et al. (2020) show that this hierarchy in parametrisation can be done recursively, leading to another subclass of DGPs that can be represented by stochastic (partial) differential equations. The relationship between the composition-based and parametrisation-based DGPs is also briefly discussed in Dunlop et al. (2018).

1.2 Reproducibility

In order to allow for reproducibility of our work, we provide the following implementations.

- Taylor moment expansion (Chapter 3). Python and Matlab codes for it are available at https://github.com/zgbkdlm/tme.
- State-space deep Gaussian processes (Chapter 4). Python and Matlab codes for it are available at https://github.com/zgbkdlm/ssdgp.
- The Python codes for reproducing Figures 1.1, 4.3, 4.8, and 5.2, as well as the simulations in Examples 3.9, 3.10, 4.16, and 4.17 are available at https://github.com/zgbkdlm/dissertation.

1.3 Outline of the thesis

This thesis consists of seven publications and overviews of them, and the thesis is organised as follows.

In Chapter 2 we review stochastic differential equations (SDEs) and Bayesian continuous-discrete filtering and smoothing (CD-FS) problems. This chapter lays out the preliminary definitions and results that are needed in the rest of the thesis.

Chapter 3 (related to Publication I) shows how to solve Gaussian approximated CD-FS problems by using the Taylor moment expansion (TME) method. This chapter also features some numerical demonstrations and analyses the positive definiteness of TME covariance approximations as well as the stability of TME Gaussian filters.

Chapter 4 (related to Publications II and VII) introduces SS-DGPs. In particular, after defining DGPs formally, we introduce their state-space representations. Secondly, we present how to sample from SS-DGPs by combining discretisation and numerical integration. Thirdly, we illustrate the construction of SS-DGPs in the Matérn sense. Fourhtly, we represent SS-DGP regression problems as CD-FS problems that we can then solve using the methods introduced in Chapter 2. Finally, we explain how DGPs can be regularised in the L^1 sense, in particular to promote sparsity at any level of the DGP component hierarchy.

Chapter 5 (related to Publications IV, III, V, II, and VI) introduces various applications of state-space (deep) GPs. These include estimation of the drift functions in SDEs, probabilistic spectro-temporal signal analysis, as well as modelling real-world signals (from astrophysics, human motion, and maritime navigation) with SS-DGPs.

Finally, Chapter 6 offers a summary of the contributions of the seven publications presented in this thesis, and concludes with a discussion of unsolved problems and possible future extensions.

2. Preliminaries

The main scope of this thesis is to reduce deep Gaussian process (DGP) regression problems into continuous-discrete filtering and smoothing problems by representing DGPs as stochastic differential equations. In this chapter we focus on introducing the technical materials that will be necessary in constructing and solving such representations. Section 2.1 is concerned with introducing stochastic differential equations and their properties. Section 2.2 focuses on continuous-discrete filtering and smoothing problems as well as algorithms to solve them. Additionally, for the sake of completeness, we list several intermediate results that will be used in the course of this thesis in Section 2.3.

2.1 Stochastic differential equations (SDEs)

Solutions to stochastic differential equations (SDEs) are a large class of continuous-time Markov processes that are commonly used to model physical, biological, or financial dynamic systems (Kloeden and Platen, 1992; Braumann, 2019). In this section, we introduce SDEs via their stochastic integral equation interpretations, and we thereupon present a few important concepts and results, including, the notion of existence and uniqueness of their solutions, their Markovian nature, and Itô's formula. For more comprehensive reviews of SDEs, we refer the reader to, for example, Chung and Williams (1990); Karatzas and Shreve (1991); Ikeda and Watanabe (1992); Øksendal (2007).

2.1.1 Stochastic integral equations

One may think of SDEs as ordinary differential/integral equations with additional stochastic driving terms. Wiener processes, which are also known as Brownian motions, are the *de facto* choice for modelling these driving terms as they allow to represent a rich class of stochastic processes with varying characteristics.

Definition 2.1 (Wiener process)**.** *A stochastic process* $W\colon \mathbb{T}\times\Omega\to\mathbb{R}$ *on some probability space* $(\Omega,\mathscr{F},\mathbb{P})$ *is called an* $\mathbb{R}$*-valued Wiener process on* $\mathbb{T}\coloneqq[t_0,\infty)$*, if*

- $W(t_0)=0$ *almost surely,*
- $t\mapsto W(t)$ *is continuous almost surely,*
- *for every integer* $k\geq 1$ *and real numbers* $t_1\leq t_2\leq\cdots\leq t_k\in\mathbb{T}$*, the increments* $W(t_k)-W(t_{k-1}),\ldots,W(t_2)-W(t_1)$ *are mutually independent,*
- *and, for every* $t>s\in\mathbb{T}$*, the increment* $W(t)-W(s)\sim\mathrm{N}(0,t-s)$ *is Gaussian distributed of mean zero and covariance* $t-s$*,*

where $W(t)$ *is a shorthand for the random variable* $\omega\mapsto W(t,\omega)$*.*

There are several ways to construct Wiener processes. The first rigorous construction of Wiener processes is due to Nobert Wiener (Wiener, 1923) who construct the Wiener process by considering the space of real-valued continuous functions on an interval (i.e., $\mathscr{C}([0,1];\mathbb{R})$), and equipping it with a canonical measure (called Wiener measure) that corresponds to the law of the Wiener process (Schilling and Partzsch, 2012; Kuo, 1975, 2006). The space of continuous functions equipped with the Wiener measure is called the classical/canonical Wiener space.

Nobert Wiener and Raymond Paley also show that one can construct the Wiener process by representing it with a trigonometric orthonormal basis on $[0,1]$, and independent identically distributed Gaussian random variables (Paley and Wiener, 1934, Chapter IX). This approach was further generalised by Paul Lévy and Zbigniew Ciesielski for any orthonormal basis of the Hilbert space of square integrable functions $L^2([0,1])$. This is known as the Lévy–Ciesielski's construction (Karatzas and Shreve, 1991). For more comprehensive reviews on the existence/construction of Wiener processes, see, for example, Schilling and Partzsch (2012, Chapter 3) or Mörters and Peres (2010).

Definition 2.1 defines scalar-valued Wiener processes. In order to generalise Wiener processes to $\mathbb{R}^w$, it is common to think of $\mathbb{R}^w$-valued Wiener processes as vectors that are a collection of w mutually independent Wiener processes (Koralov and Sinai, 2007, Definition 18.5). As for function-valued Wiener processes, such as Q-Wiener processes[1], the generalisation often leverages infinite-dimensional Gaussian measures (Kuo, 1975; Bogachev, 1998; Prato and Zabczyk, 2014; Lord et al., 2014).

The key ingredient to defining solutions of SDEs are stochastic integrals of the form

$$\int_{t_0}^{t} b(s,\omega)\,\mathrm{d}W(s,\omega), \tag{2.1}$$

[1]The cover of the thesis illustrates a realisation of a Q-Wiener process taking value in a Sobolev space with homogenous Dirichlet boundary condition.

where b is any suitable adapted process in the sense that $\omega \mapsto b(t,\omega)$ is measurable with respect to a filtration to which the Wiener process is adapted (Kuo, 2006, Chapter 4). However, due to the fact that $t \mapsto W(t)$ has infinite first order variation almost surely (Øksendal, 2007, Chapter 3), one cannot define the integral above in the classical Stieltjes sense. There exist multiple interpretations of such stochastic integral, and the two most popular constructions are due to Itô (1944) and Stratonovich (1966). In Itô's construction, this leads to an integral being a (local) martingale with respect to the filtration that W is adapted to (Kuo, 2006). In particular, when the integrand b does not depend on ω (i.e., is non-random), the integral (2.1) reduces to a Gaussian process (Kuo, 2006).

Remark 2.2. *This thesis is only concerned with Itô's construction of stochastic integrals.*

The multidimensional extension of Itô integrals is defined as follows. Suppose that W is a w-dimensional Wiener process, and b is an $\mathbb{R}^{d\times w}$-valued process, then the i-th element of a d-dimensional Itô integral is defined as

$$\sum_{j=1}^{w} \int_{t_0}^{t} b_{ij}(s,\omega)\,\mathrm{d}W_j(s,\omega), \tag{2.2}$$

where ij and j above stand for the usual element selection notations (Karatzas and Shreve, 1991, Page 283).

With Itô integrals defined, we can then formally interpret SDEs. Consider a w-dimensional Wiener process W and a stochastic process $X\colon \mathbb{T} \to \mathbb{R}^d$ that satisfies the stochastic integral equation (SIE)

$$\begin{aligned} X(t) &= X(t_0) + \int_{t_0}^{t} a(X(s),s)\,\mathrm{d}s + \int_{t_0}^{t} b(X(s),s)\,\mathrm{d}W(s), \\ X(t_0) &= X_0, \end{aligned} \tag{2.3}$$

on some probability space. The differential shorthand

$$\begin{aligned} \mathrm{d}X(t) &= a(X(t),t)\,\mathrm{d}t + b(X(t),t)\,\mathrm{d}W(t), \\ X(t_0) &= X_0, \end{aligned} \tag{2.4}$$

of the SIE in Equation (2.3) is called a *stochastic differential equation*. The SDE coefficients $a\colon \mathbb{R}^d \times \mathbb{T} \to \mathbb{R}^d$ and $b\colon \mathbb{R}^d \times \mathbb{T} \to \mathbb{R}^{d\times w}$ are called the *drift* and *dispersion* functions, respectively.

2.1.2 Existence and uniqueness of SDEs solutions

One fundamental question is whether an SDE admits a solution and, if so, what the properties (e.g., uniqueness and continuity) of the solution(s) are. In literature, the solution analysis of SDEs is usually described in the sense of strong or weak solutions. In this thesis we are mostly concerned with strong solutions that we detail in the following definition.

Definition 2.3 (Strong solution)**.** *Let* $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P})$ *be a filtered probability space,* $W : \mathbb{T} \to \mathbb{R}^w$ *be a* w*-dimensional Wiener process defined on this space, and let* $X_0 \in \mathbb{R}^d$ *be a random variable independent of* W*. Also let* $\mathscr{F}_t^W$ *be the filtration generated by* $W(t)$ *and* X_0*. Then a continuous process* $X : \mathbb{T} \to \mathbb{R}^d$ *is said to be a strong solution of the SDE* (2.4) *if the following four conditions are satisfied.*

I. $X(t)$ *is adapted to* $\mathscr{F}_t^W$*.*

II. $\mathbb{P}$*-almost surely* $X(t)$ *solves Equation* (2.3) *for all* $t \in \mathbb{T}$*.*

III. $\mathbb{P}$*-almost surely* $\int_{t_0}^{t} |a_i(X(s), s)| + (b_{ij}(X(s), s))^2 \, \mathrm{d}s < \infty$ *holds for all* $i = 1, 2, \ldots, d$*,* $j = 1, 2, \ldots, w$*, and* $t \in \mathbb{T}$*.*

IV. $\mathbb{P}$*-almost surely* $X(t_0) = X_0$*.*

The above definition is found in Karatzas and Shreve (1991, Definition 2.1) or Chung and Williams (1990, Section 10.4), but for simplicity, here we omit to augment $\mathscr{F}_t^W$ with the null sets of Ω. This definition means that if we are given a probability space which carries W and X_0, the solution $X(t)$ must be adapted to the generated filtration $\mathscr{F}_t^W$. In other words, W and X_0 should completely characterise X, and one can write the strong solution as a function of W and X_0 only.

The third condition in Definition 2.3 is important to keep in mind as it makes the solutions continuous semimartingales (Chung and Williams, 1990; Rogers and Williams, 2000).

The notion of strong solution might not always be useful because the condition of being adapted to the generated filtration is sometimes too strict. For example, in Tanaka's equation (Øksendal, 2007, Example 5.3.2), one cannot find such an $\mathscr{F}_t^W$-adapted solution therefore, the equation does not admit a strong solution. To relax this restriction, we can allow flexibility on the Wiener process, and seek pairs (X, W) solutions of the SDE (2.4), instead of simply seeking X (Chung and Williams, 1990; Øksendal, 2007). Such pairs are called weak solutions and are closely related to the martingale problem (Stroock and Varadhan, 1969, 1979; Rogers and Williams, 2000). Moreover, strong solutions are weak solutions but the converse is not true. However, since this thesis is not concerned with weak solutions, we refer the reader to, for example, Chung and Williams (1990) or Karatzas and Shreve (1991, Definition 3.1) for technical expositions of these.

Remark 2.4. *In the remainder of this thesis, unless mentioned otherwise, we will be solely concerned with strong solutions of SDEs (although some results may hold in the weak sense too). Moreover, strong solutions of SDEs will be referred to as Itô processes.*

Pathwise and weak uniqueness of SDE solutions are defined as follows (see, Karatzas and Shreve, 1991, Chapter 5.3 or Chung and Williams, 1990, Page 247).

Definition 2.5 (Pathwise uniqueness)**.** *The pathwise uniqueness holds for the SDE in Equation* (2.4) *if for all solutions* $\overline{X}$ *and* $\widetilde{X}$ *that share the same probability space, Wiener process, and initial condition, we have*

$$\mathbb{P}\left(\left\{\omega\colon \left|\overline{X}(t)-\widetilde{X}(t)\right| = 0, \text{ for all } t\in\mathbb{T}\right\}\right) = 1. \tag{2.5}$$

Notice that the "for all $t \in \mathbb{T}$" condition in Equation (2.5) can be moved outside of the probability because $\left\{\omega\colon \left|\overline{X}(t)-\widetilde{X}(t)\right| = 0, \text{ for all } t\in\mathbb{T}\right\}$ includes $\left\{\omega\colon \left|\overline{X}(t)-\widetilde{X}(t)\right| = 0\right\}$ for all $t \in \mathbb{T}$, and the converse is true as well due to the continuity of the solutions.

Definition 2.6 (Weak uniqueness)**.** *The weak uniqueness holds for the SDE in Equation* (2.4)*, if all solutions are identical in law.*

Furthermore, a classical result by Yamada and Watanabe (1971) shows that the pathwise uniqueness implies the weak uniqueness.

2.1.3 Markov property of SDE solutions

One of the main purposes of using SDEs is to construct continuous-time Markov processes. Hence, it is necessary to examine if solutions of SDEs admit the Markov property defined as follows.

Definition 2.7 (Markov process)**.** *Let* $\mathscr{F}_t$ *be a given filtation on* Ω*, and* $X(t)$ *be an* $\mathscr{F}_t$*-adapted process. Then* X *is said to be a Markov process (with respect to* $\mathscr{F}_t$*) if*

$$\mathbb{E}[\varphi(X(t+s)) \mid \mathscr{F}_t] = \mathbb{E}[\varphi(X(t+s)) \mid X(t)] \tag{2.6}$$

for every $t \in \mathbb{T}, s \in \mathbb{R}_{\geq 0}$*, and bounded Borel measurable function* φ*.*

It can be shown that Itô processes are indeed Markov processes. Proofs can be found, for example, in Øksendal (2007, Theorem 7.1.2), Kuo (2006, Theorem 10.6.2), Schilling and Partzsch (2012, Theorem 18.13), Gall (2016, Theorem 8.6), and Chung and Williams (1990, Lemma 10.10).

Remark 2.8. *Thanks to the martingale-problem method (Stroock and Varadhan, 1969), the Markov property for SDEs can be proved in more general context than strong solutions of SDEs, if the associated martingale problem is well-posed. For details, see, for example, Rogers and Williams (2000); Ethier and Kurtz (1986, Theorem 21.1).*

The Markov property is useful in the sense that it allows for predicting the future given some past information (i.e., $\mathscr{F}_t$) by only using the present (i.e., $X(t)$). This feature makes many applications – such as Bayesian filtering and smoothing (Särkkä, 2013) – computationally efficient. To see this, let $p_{X(t_1),\ldots,X(t_k),X(t_{k+1})}(x_1,\ldots,x_k,x_{k+1})$ be the finite-dimensional probability density function of $\left\{X(t_1),\ldots,X(t_k),X(t_{k+1})\right\}$ for any integer $k \geq 1$ and $t_1 \leq \cdots \leq t_k \leq t_{k+1} \in \mathbb{T}$. The Markov property implies that

$$p_{X(t_{k+1})\mid X(t_k),\ldots,X(t_1)}(x_{k+1} \mid x_k,\ldots,x_1) = p_{X(t_{k+1})\mid X(t_k)}(x_{k+1} \mid x_k) \tag{2.7}$$

and

$$p_{X(t_k)\,|\,X(t_i)}(x_k \mid x_i) = \int p_{X(t_k)\,|\,X(t_j)}(x_k \mid x_j)\, p_{X(t_j)\,|\,X(t_i)}(x_j \mid x_i)\,\mathrm{d}x_j \tag{2.8}$$

hold for every $t_i \leq t_j \leq t_k \in \mathbb{T}$. The conditional density $p_{X(t_{k+1})\,|\,X(t_k)}(x_{k+1} \mid x_k)$ and Equation (2.8) are known as the transition probability density function and the Chapman–Kolmogorov equation, respectively. In particular, the Chapman–Kolmogorov equation means that the joint probability density function of a Markov process at times $t_1, \ldots, t_k$ factorises with respect to its transition densities. Therefore, one can compute Markov processes marginal distributions sequentially with linear complexity in time. This is particularly useful in the context of Bayesian filtering and smoothing which will be the subject of Section 2.2.

2.1.4 Itô's formula

Suppose that $X \colon \mathbb{T} \to \mathbb{R}$ is a deterministic smooth function, and that $\phi \in \mathscr{C}^1(\mathbb{R};\mathbb{R})$ is another smooth function. Then by Newton–Leibniz formula/chain rule, we have

$$\phi(X(t)) = \phi(X(t_0)) + \int_{t_0}^{t} \frac{\partial \phi}{\partial X} \frac{\partial X}{\partial t}(s)\,\mathrm{d}s.$$

Unfortunately the rule above does not generally hold when X is a stochastic process. As an example, if X is a Wiener process then the derivative $\partial X / \partial t$ does not exist in the usual limit definition (Schilling and Partzsch, 2012, Chapter 14).

The differentiation rule for continuous semimartingales is given by the so-called Itô's formula (see, e.g., Gall, 2016, Theorem 5.10). In the special case when X is an Itô process, Itô's formula takes the following form.

Theorem 2.9 (Itô's formula)**.** *Let $\phi \colon \mathbb{R}^d \times \mathbb{T} \to \mathbb{R}$ be a function that is twice-differentiable in the first argument and differentiable in the second argument. Suppose that $X \colon \mathbb{T} \to \mathbb{R}^d$ is an Itô process solving the SDE in Equation* (2.4)*, then*

$$\begin{aligned} \phi(X(t),t) = \phi(X(t_0),t_0) &+ \int_{t_0}^{t} \frac{\partial \phi}{\partial t}(X(s),s)\,\mathrm{d}s \\ &+ \int_{t_0}^{t} (\nabla_X \phi)^{\mathsf{T}}\, a(X(s),s) + \frac{1}{2}\operatorname{tr}\left(\Gamma(X(s),s)\,\mathrm{H}_X\,\phi\right)\mathrm{d}s \\ &+ \int_{t_0}^{t} (\nabla_X \phi)^{\mathsf{T}}\, b(X(s),s)\,\mathrm{d}W(s), \end{aligned} \tag{2.9}$$

where $\Gamma(X(s),s) := b(X(s),s)\,b(X(s),s)^{\mathsf{T}}$, and ∇ and H denote the gradient and Hessian operators, respectively. Moreover, $t \mapsto \phi(X(t),t)$ is also an Itô process.

2.2 Continuous-discrete filtering and smoothing

In this section, we review Bayesian filtering and smoothing algorithms for continuous-discrete state-space models (Jazwinski, 1970; Särkkä, 2013; Särkkä and Solin, 2019).

2.2.1 Continuous-discrete state-space models

Consider a system

$$\begin{aligned} \mathrm{d}X(t) &= a(X(t),t)\,\mathrm{d}t + b(X(t),t)\,\mathrm{d}W(t), \quad X(t_0) = X_0, \\ Y_k &= h(X_k) + \xi_k, \quad \xi_k \sim \mathrm{N}(0, \Xi_k), \end{aligned} \tag{2.10}$$

where $X\colon \mathbb{T} \to \mathbb{R}^d$, $X_k := X(t_k)$, $Y_k \in \mathbb{R}^{d_y}$, $\Xi_k \in \mathbb{R}^{d_y \times d_y}$, and $h\colon \mathbb{R}^d \to \mathbb{R}^{d_y}$. Models represented by the combination of an SDE and a discrete-time measurement model as per Equation (2.10) are called *continuous-discrete state-space models*, or simply continuous-discrete models. These are ubiquitous in physics and engineering (see, e.g., Example 3.20 for manoeuvring target tracking). We call X_k and Y_k the *state* and *measurement*, respectively, of $X(t_k)$ at t_k.

Let $Y_{1:T} = \{Y_k : k = 1,2,\ldots,T\}$ be a collection of measurement variables and $y_{1:T} = \{y_k : k = 1,2,\ldots,T\}$ be the corresponding data at times $t_1 \le t_2 \le \cdots \le t_T \in \mathbb{T}$. The continuous-discrete filtering and smoothing problem for model (2.10) aims at solving the filtering posterior marginal densities

$$p_{X_k \mid Y_{1:k}}(x_k \mid y_{1:k}) \tag{2.11}$$

and the smoothing posterior marginal densities

$$p_{X_k \mid Y_{1:T}}(x_k \mid y_{1:T}), \tag{2.12}$$

for $k = 1,2,\ldots,T$ (Särkkä and Solin, 2019). Although in principle the filtering and smoothing problems aim at more general posterior densities (i.e., $p_{X(t) \mid Y_{1:T}}(x,t \mid y_{1:T})$ for all $t \in \mathbb{T}$), for the sake of simplicity of exposition, we restrict ourselves to estimating the marginal filtering and smoothing distribution at the data points $\{t_k : k = 1,2,\ldots,T\}$ only.

Since solutions of SDEs are Markov processes, we can use the Markov property (see, Section 2.1.3) to sequentially solve the filtering and smoothing posterior densities for $k = 1,2,\ldots,T$ (Särkkä, 2013). To see this, suppose that the filtering density $p_{X_{k-1} \mid Y_{1:k-1}}(x_{k-1} \mid y_{1:k-1})$ at t_{k-1} is known[2]. Then by leveraging Bayes' rule, the filtering density at t_k reads

$$p_{X_k \mid Y_{1:k}}(x_k \mid y_{1:k}) = \frac{p_{Y_k \mid X_k}(y_k \mid x_k)\, p_{X_k \mid Y_{1:k-1}}(x_k \mid y_{1:k-1})}{\int p_{Y_k \mid X_k}(y_k \mid x_k)\, p_{X_k \mid Y_{1:k-1}}(x_k \mid y_{1:k-1})\,\mathrm{d}x_k}, \tag{2.13}$$

[2]We define $p_{X_0 \mid Y_{1:0}}(x_0 \mid y_{1:0}) := p_{X_0}(x_0)$ at t_0.

where the predictive density

$$\begin{aligned} &p_{X_k \mid Y_{1:k-1}}(x_k \mid y_{1:k-1}) \\ &= \int p_{X_k \mid X_{k-1}}(x_k \mid x_{k-1}) p_{X_{k-1} \mid Y_{1:k-1}}(x_{k-1} \mid y_{1:k-1}) \mathrm{d}x_{k-1} \end{aligned} \tag{2.14}$$

needs to be computed by propagating $p_{X_{k-1} \mid Y_{1:k-1}}(x_{k-1} \mid y_{1:k-1})$ through the SDE. One can then obtain the filtering densities sequentially for $k = 1, 2, \ldots, T$ starting from a known/given initial condition.

The smoothing densities are solved backward for $k = T, \ldots, 1$ by using the filtering results. Suppose that the smoothing density $p_{X_{k+1} \mid Y_{1:T}}(x_{k+1} \mid y_{1:T})$ at t_{k+1} is known, then again by Bayes' rule (Kitagawa, 1987; Särkkä, 2013), the smoothing density at t_k is

$$\begin{aligned} &p_{X_k \mid Y_{1:T}}(x_k \mid y_{1:T}) \\ &= p_{X_k \mid Y_{1:k}}(x_k \mid y_{1:k}) \int \frac{p_{X_{k+1} \mid X_k}(x_{k+1} \mid x_k) p_{X_{k+1} \mid Y_{1:T}}(x_{k+1} \mid y_{1:T})}{p_{X_{k+1} \mid Y_{1:k}}(x_{k+1} \mid y_{1:k})} \mathrm{d}x_{k+1}. \end{aligned} \tag{2.15}$$

Unfortunately, for non-linear state-space models, Equations (2.13), (2.14), and (2.15) are rarely solvable in closed-form. In practice, one often needs to use approximation schemes, such as Taylor expansion, numerical integration, or particle approximations (Särkkä, 2013). However, if the SDE and measurement model happen to be linear (and also starting from a Gaussian initial condition), then the filtering and smoothing densities are exactly Gaussian and their means and covariances can be computed in closed-form sequentially. This is known as the (continuous-discrete) Kalman filtering and Rauch–Tung–Striebel smoothing (Särkkä and Solin, 2019), the details of which are given in the next section.

2.2.2 Rauch–Tung–Striebel smoothing

Consider a linear continuous-discrete model

$$\begin{aligned} \mathrm{d}X(t) &= A(t)X(t)\mathrm{d}t + B(t)\mathrm{d}W(t), \quad X(t_0) = X_0, \\ y_k &= H_k X_k + \xi_k, \quad \xi_k \sim \mathrm{N}(0, \Xi_k), \end{aligned} \tag{2.16}$$

where $X_0 \sim \mathrm{N}(m_0, P_0)$ is a Gaussian random variable of mean m_0 and covariance P_0. Here the coefficients $A \colon \mathbb{T} \to \mathbb{R}^{d \times d}$, $B \colon \mathbb{T} \to \mathbb{R}^{d \times w}$, and $H_k \in \mathbb{R}^{d_y \times d}$ are deterministic matrix-valued functions and a constant, respectively. In this case, the filtering and smoothing densities in Equations (2.13) and (2.15) can be solved exactly by using Kalman filters and Rauch–Tung–Striebel (RTS) smoothers as follows (cf. Särkkä and Solin, 2019).

Algorithm 2.10 (Continuous-discrete Kalman filter and RTS smoother)**.** *Let* $p_{X_k \mid Y_{1:k}}(x_k \mid y_{1:k}) = \mathrm{N}(x_k \mid m_k^f, P_k^f)$ *and* $p_{X_k \mid Y_{1:T}}(x_k \mid y_{1:T}) = \mathrm{N}(x_k \mid m_k^s, P_k^s)$ *be the Gaussian parametrisations of the filtering and smoothing posterior*

densities, respectively, at t_k. Also let $m_0^f := m_0$ and $P_0^f := P_0$ at t_0. At each step for $k = 1,2,\ldots,T$, the Kalman filter first obtains the predictive density $\mathrm{N}(x_k \mid m_k^-, P_k^-)$ by solving the system of ordinary differential equations (ODEs)

$$\begin{aligned} \frac{\mathrm{d}m(t)}{\mathrm{d}t} &= A(t)m(t), \\ \frac{\mathrm{d}P(t)}{\mathrm{d}t} &= A(t)P(t) + P(t)A(t)^\mathsf{T} + B(t)B(t)^\mathsf{T}, \end{aligned} \tag{2.17}$$

at t_k, starting from the initial values m_{k-1}^f and P_{k-1}^f at time t_{k-1}. Then, it updates the predictive density to get the filtering posterior mean m_k^f and covariance P_k^f at time t_k by computing

$$\begin{aligned} K_k &= P_k^- H_k^\mathsf{T} (H_k P_k^- H_k^\mathsf{T} + \Xi_k)^{-1}, \\ m_k^f &= m_k^- + K_k (y_k - H_k m_k^-), \\ P_k^f &= P_k^- - K_k (H_k P_k^- H_k^\mathsf{T} + \Xi_k) K_k^\mathsf{T}. \end{aligned} \tag{2.18}$$

Let $m_T^s := m_T^f$ and $P_T^s := P_T^f$. At each step for $k = T-1,\ldots,1$, the RTS smoother computes m_k^s and P_k^s at t_k by solving the system of ODEs

$$\begin{aligned} \frac{\mathrm{d}m(t)}{\mathrm{d}t} &= A(t)m(t) + B(t)B(t)^\mathsf{T} \left(P^f(t)\right)^{-1} \left(m(t) - m^f(t)\right), \\ \frac{\mathrm{d}P(t)}{\mathrm{d}t} &= \left[A(t) + B(t)B(t)^\mathsf{T} \left(P^f(t)\right)^{-1}\right] P(t) \\ &\quad + P(t) \left[A(t) + B(t)B(t)^\mathsf{T} \left(P^f(t)\right)^{-1}\right]^\mathsf{T} - B(t)B(t)^\mathsf{T}, \end{aligned} \tag{2.19}$$

starting from the initial values m_{k+1}^s and P_{k+1}^s at time t_{k+1}, where $m^f(t)$ and $P^f(t)$ stand for the filtering mean and covariance at time t, respectively.

Furthermore, if the SDE coefficients in Equation (2.16) do not depend on time (i.e., A and B are constant matrices), then the continuous-discrete filtering and smoothing problem can be reformulated in an equivalent discrete-discrete problem of the form

$$\begin{aligned} X_k &= F_{k-1} X_{k-1} + q_{k-1}, \\ Y_k &= H_k X_k + \xi_k, \end{aligned} \tag{2.20}$$

where $q_{k-1} \sim \mathrm{N}(0, Q_{k-1})$. The coefficients $F_{k-1} \in \mathbb{R}^{d \times d}$ and $Q_{k-1} \in \mathbb{R}^{d \times d}$ are in turn determined by

$$\begin{aligned} F_{k-1} &= e^{(t_k - t_{k-1})A}, \\ Q_{k-1} &= \int_{t_{k-1}}^{t_k} e^{(t_k - s)A} B B^\mathsf{T} \left(e^{(t_k - s)A}\right)^\mathsf{T} \mathrm{d}s. \end{aligned} \tag{2.21}$$

Provided one can numerically compute Equations (2.21) (see, e.g., Axelsson and Gustafsson, 2015; Särkkä and Solin, 2019, for how to do so in practice), one can then apply standard Kalman filters and RTS smoothers (Särkkä, 2013, Theorems 4.2 and 8.2) to the discretised state-space model.

2.2.3 Gaussian approximate smoothing

In this section, we review the Gaussian approximated density filtering and smoothing for non-linear continuous-discrete state-space models (Itô and Xiong, 2000; Särkkä and Sarmavuori, 2013). The idea of Gaussian filtering and smoothing is to approximate the filtering and smoothing densities by

$$\begin{aligned} p_{X_k \mid Y_{1:k}}(x_k \mid y_{1:k}) &\approx \mathrm{N}\big(x_k \mid m_k^f, P_k^f\big), \\ p_{X_k \mid Y_{1:T}}(x_k \mid y_{1:T}) &\approx \mathrm{N}\big(x_k \mid m_k^s, P_k^s\big). \end{aligned} \tag{2.22}$$

Then, by applying Gaussian identities, the general Bayesian filtering and smoothing formulations in Equations (2.13) and (2.15) admit closed-form approximations. We therefore have the following algorithm (cf. Särkkä and Solin, 2019, Chapter 10).

Algorithm 2.11 (Continuous-discrete Gaussian filter and smoother)**.** *Let $p_{X_k \mid Y_{1:k}}(x_k \mid y_{1:k}) \approx \mathrm{N}\big(x_k \mid m_k^f, P_k^f\big)$ and $p_{X_k \mid Y_{1:T}}(x_k \mid y_{1:T}) \approx \mathrm{N}\big(x_k \mid m_k^s, P_k^s\big)$ be approximate filtering and smoothing densities. Also consider a Gaussian approximation to the initial density $p_{X_0}(x_0) \approx \mathrm{N}(x_0 \mid m_0, P_0)$. The Gaussian filter obtains $\big\{m_k^f, P_k^f : k = 1, 2, \ldots, T\big\}$ by computing the following prediction and update steps sequentially for $k = 1, 2, \ldots, T$.*

I. Prediction:

$$\begin{aligned} m_k^- &= \int x_k \, p_{X_k \mid Y_{1:k-1}}(x_k \mid y_{1:k-1}) \, \mathrm{d}x_k, \\ P_k^- &= \int (x_k - m_k^-)(x_k - m_k^-)^\mathsf{T} \, p_{X_k \mid Y_{1:k-1}}(x_k \mid y_{1:k-1}) \, \mathrm{d}x_k. \end{aligned} \tag{2.23}$$

II. Update:

$$\begin{aligned} S_k &= \mathbb{E}\left[\big(h(X_k) - \mathbb{E}[h(X_k)]\big)\big(h(X_k) - \mathbb{E}[h(X_k)]\big)^\mathsf{T}\right] + \Xi_k, \\ K_k &= \mathbb{E}\left[\big(X_k - m_k^-\big)\big(h(X_k) - \mathbb{E}[h(X_k)]\big)^\mathsf{T}\right] S_k^{-1}, \\ m_k^f &= m_k^- + K_k(y_k - \mathbb{E}[h(X_k)]), \\ P_k^f &= P_k^- - K_k S_k K_k^\mathsf{T}. \end{aligned} \tag{2.24}$$

Note that the expectations above are taken with respect to the predictive density $p_{X_k \mid Y_{1:k-1}}(x_k \mid y_{1:k-1})$. In addition, if the measurement model is linear, then the update step above reduces to Equation (2.18)*.*

Let $m_T^s := m_T^f$ and $P_T^s := P_T^f$. The Gaussian smoother obtains $\big\{m_k^s, P_k^s : k =$

$1,2,\ldots,T-1\}$ *by sequentially computing*

$$\begin{aligned} D_{k+1} &= \operatorname{Cov}\left[X_k, X_{k+1}^\mathsf{T} \mid y_{1:k}\right], \\ G_k &= D_{k+1}\left(P_{k+1}^-\right)^{-1}, \\ m_k^s &= m_k^f + G_k(m_{k+1}^s - m_{k+1}^-), \\ P_k^s &= P_k^f + G_k(P_{k+1}^s - P_{k+1}^-)G_k^\mathsf{T}, \end{aligned} \tag{2.25}$$

for $k = T-1, T-2, \ldots, 1$.

In order to compute the integrals/expectations in Algorithm 2.11, it is often necessary to approximate the transition density by

$$p_{X_k \mid X_{k-1}}(x_k \mid x_{k-1}) \approx \mathrm{N}\big(x_k \mid \mathbb{E}[X_k \mid X_{k-1}], \operatorname{Cov}[X_k \mid X_{k-1}]\big). \tag{2.26}$$

There are various approaches to approximate the mean and covariance in the transition density above. One popular approach is linearising the SDE (or its discretisation) by using, for example, Taylor expansions. This leads to (continuous-discrete) extended Kalman filters and smoothers (Jazwinski, 1970). Another commonly used approach is to solve the ODEs (see, e.g., Equation (3.2)) that characterise the mean and covariance functions of the SDE (Sancho, 1970; Jazwinski, 1970; Maybeck, 1982; Särkkä and Sarmavuori, 2013). However this ODE approach requires to compute expectations with respect to the probability measure of SDEs, which in practice requires further approximation schemes (such as Monte Carlo).

We can also approximate the SDE by a Gaussian increment-based discretisation defined as

$$X_k \approx f_{k-1}(X_{k-1}) + q_{k-1}(X_{k-1}), \tag{2.27}$$

where $q_{k-1}(X_{k-1}) \sim \mathrm{N}(0, Q_{k-1}(X_{k-1}))$. In particular, $\mathbb{E}[X_k \mid X_{k-1}] \approx f_{k-1}(X_{k-1})$ and $\operatorname{Cov}[X_k \mid X_{k-1}] \approx Q_{k-1}(X_{k-1})$. The choice of the functions $f_{k-1}\colon \mathbb{R}^d \to \mathbb{R}^d$ and $Q_{k-1}\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$ depends on the discretisation method used for the approximation.

Example 2.12. *For instance, the Euler–Maruyama scheme gives*

$$\begin{aligned} f_{k-1}(X_{k-1}) &= X_{k-1} + a(X_{k-1}, t_{k-1})(t_k - t_{k-1}), \\ Q_{k-1}(X_{k-1}) &= b(X_{k-1}, t_{k-1})\, b(X_{k-1}, t_{k-1})^\mathsf{T}(t_k - t_{k-1}). \end{aligned} \tag{2.28}$$

Furthermore, in Chapter 3 we illustrate a Taylor moment expansion-based approach generalising the Euler–Maruyama scheme for approximating the transition coefficients in Equation (2.26)*.*

Recall that the expectations in Algorithm 2.11 are usually hard to compute exactly for non-linear models. However, we can use quadrature methods, for example, Gauss–Hermite quadrature (Davis and Rabinowitz, 1984;

Arasaratnam et al., 2007), unscented transform (Julier and Uhlmann, 2004), spherical cubature (Arasaratnam and Haykin, 2009; Särkkä and Solin, 2012), or sparse-grid quadratures (Jia et al., 2012; Radhakrishnan et al., 2016) to compute them numerically.

2.2.4 Non-Gaussian approximate smoothing

Despite the simplicity and efficiency of Gaussian approximated filtering and smoothing, these might lead to poor approximations for densities that are, for example, multi-modal or skewed (Särkkä, 2013). Moreover, Zhao et al. (2021a) show that, for many SS-DGPs constructions, the Kalman gain (i.e., K_k in Algorithm 2.11) of Gaussian approximated filters and smoothers converge to zero as $t \to \infty$. This can be problematic as a zero Kalman gain means that no further information from data is used for updating the posterior distributions. This issue is detailed in Section 4.8. Hence, the aim of this section is to briefly review some other non-linear filters and smoothers that could be useful for solving the continuous-discrete model in Equation (2.10) without relying on Gaussian approximations.

One way to compute the general filtering and smoothing densities is by using sequential Monte Carlo (SMC) methods (Chopin and Papaspiliopoulos, 2020). This class of methods considers Monte Carlo approximations of the integrals in Equations (2.13) and (2.15) instead of Gaussian quadrature ones. They sequentially propose new Monte Carlo samples that they then weight via a potential function, and use a resampling step in order to keep weight distribution non-degenerate (Doucet et al., 2000; Godsill et al., 2004; Andrieu et al., 2010). These result in two generic classes of methods called particle filters and particle smoothers retaining linear complexity at the cost of losing the closed-form interpretation. These methods can be customised to the problem at hand so as to provide better approximations of the distributions (Chopin and Papaspiliopoulos, 2020). In particular, in the context of SS-DGPs, Zhao et al. (2021a) show that they result in a better approximation of the posterior density for regression problems such as the rectangular signal in Figure 1.1. However, parameter learning in particle filters can be problematic, as the resampling procedure, in general, makes their loss functions non-differentiable. This can be addressed by using smooth resampling methods, such as the one in Corenflos et al. (2021a).

Another way to compute the filtering and smoothing densities is to think of them as solutions of ODEs/partial differential equations (PDEs). These connections are well-known for continuous-continuous state-space models (i.e., where instead of the discrete measurements in Equation (2.10) we have a continuous measurement modelled as an SDE depending on the state), such as the Kalman–Bucy filter (Kálmán and Bucy, 1961) for linear models. More generally, for non-linear continuous state-space mod-

els, the filtering density (Kushner, 1964; Zakai, 1969; Bain and Crisan, 2009; Särkkä, 2013) is governed by the Kushner–Stratonovich equation or Zakai's equation[3]. For the PDEs that characterise the continuous smoothing solutions, see, for example, Särkkä and Solin (2019, Algorithm 10.30) or Anderson (1972).

Analogously to the continuous filtering and smoothing, it is also possible to obtain continuous-discrete posterior densities by solving certain PDEs or ODEs. For example, Jazwinski (1970); Beard et al. (1999); Challa and Bar-Shalom (2000) show that one can combine the Fokker–Planck–Kolmogorov equation and Bayes' rule in order to compute the filtering solution. More specifically, Fokker–Planck–Kolmogorov equation is used to predict the state in Equation (2.14), while Bayes' rule is then used to update the predicted state into the filtered state as per the filtering formulation in Equation (2.13). In a different flavour, Brigo et al. (1998); Koyama (2018) consider the projection filter and smoother, which consist in projecting the filtering and smoothing solutions (of certain families of probability densities) on the space of their density parameters (e.g., the natural parameters of the exponential family). This transforms the problem in a system of ODEs in their density parameters that one then can solve instead of solving the original problem.

Archambeau et al. (2007, 2008); Li et al. (2020) show that one can also approximate the filtering/smoothing solution by another SDE. The idea is to use a parametrised SDE (e.g., a linear SDE is used in Archambeau et al., 2007, 2008) to approximate the filtering/smoothing solution and learn the SDE parameters by minimising the Kullback–Leibler (KL) divergence from the true filtering/smoothing distribution. Once this approximate SDE is learnt, the statistical properties (e.g., mean or covariance) of the filtering/smoothing solution can be computed in closed-form from the approximate linear SDE or by simulating trajectories from the approximate SDE (if the SDE is non-linear). Recall that solutions of SDEs are Markov processes. This SDE-based variational filtering/smoothing method is indeed reasonable in the sense that the optimal variational distribution (among a family of parametric variational distributions) for minimising the KL divergence admits the Markov property as shown in Courts et al. (2021, Lemma 1).

For more comprehensive reviews of non-linear filtering and smoothing methods, we refer the reader to, for example, Jazwinski (1970); Maybeck (1982); Särkkä (2013); Bain and Crisan (2009); Law et al. (2015); Evensen (2009); Doucet et al. (2001); Särkkä and Solin (2019).

[3]Note that Zakai's equation gives *unnormalised* filtering densities.

2.3 Some theorems

For the sake of self-containedness, in this section we list several intermediate results that will be used in the course of the thesis.

Theorem 2.13 (Cauchy product)**.** *Let $\sum_{i=0}^{\infty} \alpha_i x^i$ and $\sum_{i=0}^{\infty} \beta_i x^i$ be two power series of x with convergence radius $D_\alpha > 0$ and $D_\beta > 0$. Then their product is a power series*

$$\left(\sum_{i=0}^{\infty} \alpha_i x^i\right)\left(\sum_{i=0}^{\infty} \beta_i x^i\right) = \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} \alpha_j \beta_{k-j}\right) x^k \tag{2.29}$$

on an open disk of radius $D \geq \min(D_\alpha, D_\beta)$ (see, e.g., Canuto and Tabacco, 2014, Theorem 2.37).

We use the Cauchy product in Theorem 3.5 to truncate the product of two finite power series.

Theorem 2.14 (Weyl's inequality)**.** *Let A and B be Hermitian matrices of size $n \times n$. Also let $\lambda_1 \geq \cdots \geq \lambda_n$ denote the ordered eigenvalues of any $n \times n$ Hermitian matrix. Then*

$$\lambda_i(A) + \lambda_n(B) \leq \lambda_i(A+B) \leq \lambda_i(A) + \lambda_1(B), \tag{2.30}$$

for $i = 1, \ldots, n$.

Weyl's inequality was originally posed by Weyl (1912), and it can also be found, for example, in Bernstein (2009, Theorem 8.4.11), Horn and Johnson (1991), or Helmke and Rosenthal (1995, Section 5). Weyl's inequality is used in Theorem 3.5 to form a lower bound on the minimum eigenvalue of a covariance approximation.

Theorem 2.15 (Langenhop (1960))**.** *Let $u : \mathbb{T} \to \mathbb{R}_{\geq 0}$ and $f : \mathbb{T} \to \mathbb{R}_{\geq 0}$ be continuous functions, and let $v : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a continuous non-decreasing function with $v > 0$ on $\mathbb{R}_{>0}$. Now consider the invertible function*

$$G(r) = \int_{r_0}^{r} \frac{1}{v(\tau)}\,\mathrm{d}\tau, \quad r > 0, \quad r_0 > 0, \tag{2.31}$$

and its inverse function G^{-1} defined on domain E. Suppose that there is a $t_2 \in \mathbb{T}$ such that $G(u(s)) - \int_s^t f(\tau)\,\mathrm{d}\tau \in E$ for all $s, t \in \mathbb{T}$ and $s \leq t \leq t_2$. If the following inequality is verified,

$$u(t) \geq u(s) - \int_s^t f(\tau) v(u(\tau))\,\mathrm{d}\tau, \quad s, t \in \mathbb{T}, \quad s \leq t, \tag{2.32}$$

then

$$u(t) \geq G^{-1}\left(G(u(s)) - \int_s^t f(\tau)\,\mathrm{d}\tau\right), \quad s, t, t_2 \in \mathbb{T}, \quad s \leq t \leq t_2. \tag{2.33}$$

Remark 2.16. *Note that Theorem 2.15 is independent of the choice of $r_0 > 0$.*

Langenhop's inequality was originally derived in Langenhop (1960). A more modern presentation can be found, for example, in Pachpatte (1998, Theorem 2.3.2). This theorem is used in Remark 4.28 to obtain a positive lower bound on the variance of an SDE solution.

Theorem 2.17 (Peano–Baker series)**.** *Consider linear ODE of the form*

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = A(t)x(t) + z(t), \quad x(t_0) = x_0 \in \mathbb{R}^d, \tag{2.34}$$

where the coefficients $A\colon \mathbb{T} \to \mathbb{R}^{d\times d}$ and $z\colon \mathbb{T} \to \mathbb{R}^d$ are locally bounded measurable functions. Then for every $t_0, t \in \mathbb{T}$, the ODE above has a unique solution of the form

$$x(t) = \mathbf{\Lambda}(t,t_0)\left(x_0 + \int_{t_0}^{t} \mathbf{\Lambda}(t,s)z(s)\,\mathrm{d}s\right). \tag{2.35}$$

If moreover A and z are continuous functions, then $\mathbf{\Lambda}$ can be represented by its Peano–Baker series

$$\mathbf{\Lambda}(t,t_0) = I + \int_{t_0}^{t} A(s)\,\mathrm{d}s + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)\,\mathrm{d}s_2\,\mathrm{d}s_1 + \cdots, \tag{2.36}$$

for all $t \in \mathbb{T}$.

While the continuity of A and z is not a necessary condition for the existence of $\mathbf{\Lambda}$ (cf. Theorem 2.18), the fact that its Peano–Baker series approximation is compactly convergent relies on the continuity of A and z (Baake and Schlägel, 2011; Brogan, 2011). Other constructions of $\mathbf{\Lambda}$ include, for example, Magnus expansion (Moan and Niesen, 2008) but they have somewhat stricter hypotheses.

Theorem 2.18 (Solution of linear SDEs)**.** *Let $A\colon \mathbb{T} \to \mathbb{R}^{d\times d}$ and $B\colon \mathbb{T} \to \mathbb{R}^{d\times w}$ be locally bounded measurable functions, and let $W\colon \mathbb{T} \to \mathbb{R}^w$ be a Wiener process. Then the solution of linear SDE of the form*

$$\mathrm{d}U(t) = A(t)U(t)\,\mathrm{d}t + B(t)\,\mathrm{d}W(t) \tag{2.37}$$

is given by

$$U(t) = \Lambda(t)U(t_0) + \Lambda(t)\int_{t_0}^{t} (\Lambda(s))^{-1}B(s)\,\mathrm{d}W(s), \tag{2.38}$$

where Λ is the unique solution of the matrix ODE

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = A(t)x(t), \quad x(t_0) = I_d \in \mathbb{R}^{d\times d}, \quad t \in \mathbb{T}. \tag{2.39}$$

Remark 2.19. *The Λ appearing in Theorem 2.18 is absolutely continuous on $\mathbb{T}$, and is such that $\Lambda(t)$ is a non-singular matrix for all $t \in \mathbb{T}$. Moreover, we have $\mathbf{\Lambda}(t,s) = \Lambda(t)(\Lambda(s))^{-1}$, for all $s,t \in \mathbb{T}$, where $\mathbf{\Lambda}$ is defined in Theorem 2.17.*

Theorem 2.18 can be found in Karatzas and Shreve (1991, Section 5.6). We used it in Theorem 4.11 in order to prove the strong existence of solutions to the SDE characterisation of SS-DGP as well as to give an explicit expression for the covariance functions of SS-DGP solutions. Noting that the conditions in Definitions 2.3 and 2.5 are verified, the process defined in Equation (2.38) is a strong solution, and the pathwise uniqueness holds for the SDE in Equation (2.37) (see, Karatzas and Shreve, 1991, Section 5.6).

3. Taylor moment expansion filtering and smoothing

This chapter is concerned with Publication I. More specifically, this chapter presents the Taylor moment expansion (TME) scheme for approximating the statistical properties of SDE solutions, such as their mean and covariance. Based on this, we thereupon present TME-based Gaussian filters and smoothers and analyse their stability.

The chapter starts with a general discussion on the motivation and background of the TME method. In Section 3.2, we briefly review diffusion processes and related infinitesimal generators which are the key ingredients of TME. Then, in Sections 3.3 we formally introduce TME, and in Section 3.4 we analyse the positive definiteness of their covariance approximants. Section 3.5 features several examples that illustrate how to use TME in practice. Finally, in Section 3.6, we present Gaussian approximated density filters and smoothers that leverage the TME method for approximating the predictive means and covariances of the system.

3.1 Motivation

Let $X(t)$ be an Itô process that satisfies the SDE given by Equation (2.4). In stochastic filtering and smoothing (Jazwinski, 1970; Bain and Crisan, 2009; Särkkä, 2013), it is often of interest to compute the conditional expectation of a given target function ϕ for any two time points $t \geq s \in \mathbb{T}$. For instance, as shown in Algorithm 2.11, Gaussian approximated density filters and smoothers require to be able to compute the predictive mean and covariance of SDE solutions. These conditional expectations take the form

$$\mathbb{E}[\phi(X(t)) \,|\, X(s)], \tag{3.1}$$

where different target functions result in different statistical quantities, such as mean, covariance or higher-order moments.

There exist several approaches to computing the expectation in Equation (3.1) numerically. One approach is based on forming an ODE that governs the conditional expectation in Equation (3.1) (see, e.g., Xiu, 2010;

Khasminskii, 2012; Särkkä and Solin, 2019). For example, let $\phi(x) = x$, then by Itô's formula we can obtain an ODE

$$\frac{\mathrm{d}\mathbb{E}[X(t) \mid X(s)]}{\mathrm{d}t} = \mathbb{E}[a(X(t), t) \mid X(s)] \tag{3.2}$$

starting from $s \in \mathbb{T}$. However, it is usually hard to solve the ODE in Equation (3.2) analytically. This is due to the fact that computing its driving term $\mathbb{E}[a(X(t), t) \mid X(s)]$ requires computing an expectation with respect to the SDE distribution, which is in general intractable analytically. One solution to this problem is to approximate the expectation using quadrature integration methods (Särkkä, 2007, 2010; Kulikov and Kulikova, 2014), but the approximation error can accumulate in time, resulting in unstable estimation. Another solution is to iteratively form ODEs that characterise their parent driving terms. Explicitly, one can choose $\phi = a$ in the above, so as to characterise $\mathbb{E}[a(X(t), t) \mid X(s)]$ by another ODE driven by some function a', then choose $\phi = a'$, and so on. However, this leads to a so-called closure problem as explained in Xiu (2010, Section 4.4.2).

It is also common to approximate Equation (3.1) using numerical discretisation methods, such as Euler–Maruyama, Milstein's method, or higher-order Itô–Taylor expansions (Kloeden and Platen, 1992). The upside of these methods is that if the function ϕ happens to be a polynomial function, then these methods can give analytical approximations of Equation (3.1). As an example, let $\phi(x) = (x - \mathbb{E}[X(t)])(x - \mathbb{E}[X(t)])^\mathsf{T}$. Then the Euler–Maruyama method gives the approximation

$$\mathbb{E}[\phi(X(t)) \mid X(s)] = \mathrm{Cov}[X(t) \mid X(s)] \approx (t - s)\, b(X(s), s)\, b(X(s), s)^\mathsf{T}.$$

However, for more general non-linear ϕ these approaches usually fail to give analytical approximations (see, e.g., Example 3.9), and one often needs to use Monte Carlo methods to approximate the expectation.

In the remainder of this section, we present the so-called Taylor moment expansion (Dacunha-Castelle and Florens-Zmirou, 1986; Florens-Zmirou, 1989; Kessler, 1997; Zhao et al., 2021b) approach for computing expectations of the form given in Equation (3.1). This method relies on approximating the expectation in Equation (3.1) in terms of a Taylor expansion up to a given order M that depends on the regularity of the coefficients of the SDE verified by X. The terms in this expansion are expressed as iterative applications of the infinitesimal generator of the SDE at hand on the target function ϕ (see, Section 3.2 for a formal definition). When the coefficients are infinitely smooth, this method offers asymptotically exact representations. We start by giving an overview of diffusion processes and the infinitesimal generator which is an essential part of the TME method.

3.2 Infinitesimal generator

Diffusion processes (Dynkin, 1965; Ikeda and Watanabe, 1992; Itô, 2004) are an important subclass of continuous-time Markov processes whose transition probability densities verify certain (infinitesimal) regularities (see, e.g., Kuo, 2006, Definition 10.8.3). These processes are entirely characterised by their infinitesimal generators which are defined as follows. Let $X : \mathbb{T} \to \mathbb{R}^d$ be a diffusion process starting from any $x \in \mathbb{R}^d$ at t_0 and let ϕ be a suitable function. The operator $\mathscr{A}$ defined by

$$\mathscr{A}\phi(x) = \lim_{t \downarrow t_0} \frac{\mathbb{E}[\phi(X(t)) \mid x] - \phi(x)}{t - t_0} \tag{3.3}$$

is called the *infinitesimal generator* of the diffusion X. Heuristically, the infinitesimal generator represents the expected rate of change of $\phi(X(t))$ around x.

There are many approaches to construct diffusion processes (with desired drift and diffusion coefficients), such as the semigroup approach, the PDE approach (i.e., Kolmogorov backward equation), and the (Itô's) SDE approach (Kuo, 2006; Schilling and Partzsch, 2012). In particular, if one considers diffusion processes that are solutions of SDE, then their infinitesimal generators can be expressed in terms of their SDE coefficients.

Theorem 3.1 (Infinitesimal generator in Itô's SDE representation)**.** *Let $X : \mathbb{T} \to \mathbb{R}^d$ be a diffusion process that is the solution of the following time-homogeneous SDE*

$$\mathrm{d}X(t) = a(X(t))\,\mathrm{d}t + b(X(t))\,\mathrm{d}W(t), \tag{3.4}$$

where $a : \mathbb{R}^d \to \mathbb{R}^d$, $b : \mathbb{R}^d \to \mathbb{R}^{d \times w}$, and $W : \mathbb{T} \to \mathbb{R}^w$ is a Wiener process. Also let us define $\Gamma(x) := b(x)\,b(x)^\mathsf{T}$. Then, the infinitesimal generator $\mathscr{A}$ defined in Equation (3.3) *is given by*

$$\begin{aligned} \mathscr{A}\phi(x) &= \sum_{i=1}^{d} a_i(x) \frac{\partial \phi}{\partial x_i}(x) + \frac{1}{2} \sum_{i,j=1}^{d} \Gamma_{ij}(x) \frac{\partial^2 \phi}{\partial x_i\, \partial x_j}(x) \\ &:= (\nabla_x \phi(x))^\mathsf{T}\, a(x) + \frac{1}{2} \operatorname{tr}\left(\Gamma(x)\, \mathrm{H}_x\, \phi(x)\right) \end{aligned} \tag{3.5}$$

for any suitable $\phi \in \mathscr{C}^2(\mathbb{R}^d; \mathbb{R})$. The drift and diffusion coefficients of the diffusion X are then given by a and Γ, respectively.

Proof. The proof can be found, for example, in Øksendal (2007, Theorem 7.3.3) or Kuo (2006, Theorem 10.9.11). □

Theorem 3.1 can be extended to the case of time-dependent $x, t \mapsto \phi(x, t)$ and SDE coefficients $x, t \mapsto a(x, t)$, $x, t \mapsto b(x, t)$. For details of this, see, for example, Särkkä and Solin (2019, Definition 5.3).

3.3 Taylor moment expansion (TME)

Recall that the aim of this section is to compute expectations of the form

$$\mathbb{E}[\phi(X(t)) \mid X(s)] \tag{3.6}$$

for $t \geq s \in \mathbb{T}$ and any given target function ϕ.

The idea of TME (Florens-Zmirou, 1989) is to approximate Equation (3.6) by means of a Taylor expansion

$$\mathbb{E}[\phi(X(t)) \mid X(s)] \approx \sum_{r=0}^{M} \frac{1}{r!} \frac{\mathrm{d}^r \mathbb{E}[\phi(X(t)) \mid X(s)]}{\mathrm{d}t^r}(s) \Delta t^r \tag{3.7}$$

centred at time s, where $\Delta t \coloneqq t - s$, and M is the expansion order. The right hand side of Equation (3.7) involves computing derivatives (when they exist) of the conditional expectation in Equation (3.6) when seen as a function of t. It turns out that these derivatives can be explicitly expressed as iterations of the infinitesimal generator in Equation (3.5). This is formally stated in the following theorem.

Theorem 3.2 (Taylor moment expansion)**.** *Let $M \geq 0$ be an integer and $X \colon \mathbb{T} \to \mathbb{R}^d$ be the solution of the SDE given in Equation* (3.4)*, where the SDE coefficients $a \colon \mathbb{R}^d \to \mathbb{R}^d$ and $b \colon \mathbb{R}^d \to \mathbb{R}^{d \times w}$ are M times differentiable. Suppose that the target function $\phi \in \mathscr{C}^{2(M+1)}(\mathbb{R}^d; \mathbb{R})$, then we have*

$$\mathbb{E}[\phi(X(t)) \mid X(s)] = \sum_{r=0}^{M} \frac{1}{r!} \mathscr{A}^r \phi(X(s)) \Delta t^r + R_{M,\phi}(X(s), \Delta t) \tag{3.8}$$

for every $t \geq s \in \mathbb{T}$, where $\Delta t \coloneqq t - s$, and

$$R_{M,\phi}(X(s), \Delta t) = \int_s^t \int_s^{\tau_1} \cdots \int_s^{\tau_M} \mathbb{E}\left[\mathscr{A}^{M+1} \phi(X(\tau)) \mid X(s)\right] \mathrm{d}\tau \tag{3.9}$$

is the remainder.

Proof. We prove that

$$\frac{\mathrm{d}^r \mathbb{E}[\phi(X(t)) \mid X(s)]}{\mathrm{d}t^r}(t) = \mathbb{E}[\mathscr{A}^r \phi(X(t)) \mid X(s)] \tag{3.10}$$

for every $r = 0, 1, \ldots, M$ by induction. When $r = 0$, the result trivially holds. When $r = 1$ by Itô's formula (see, Theorem 2.9) we obtain

$$\begin{aligned} \phi(X(t)) = \phi(X(s)) &+ \int_s^t (\nabla_X \phi)^\mathsf{T} a(X(\tau)) + \frac{1}{2} \operatorname{tr}\left(\Gamma(X(\tau)) \, \mathrm{H}_X \phi\right) \mathrm{d}\tau \\ &+ \int_s^t b(X(\tau)) \, \mathrm{d}W(\tau). \end{aligned} \tag{3.11}$$

Taking the expectation on both sides of the equation above yields

$$\mathbb{E}[\phi(X(t)) \mid X(s)] = \phi(X(s)) + \int_s^t \mathbb{E}[\mathscr{A} \phi(X(\tau)) \mid X(s)] \, \mathrm{d}\tau. \tag{3.12}$$

The fundamental theorem of calculus ensures that $\mathbb{E}[\phi(X(t)) \mid X(s)]$ is differentiable with respect to t because the integrand in the integral above is continuous. Therefore, we can interchangeably use its differential form

$$\frac{\mathrm{d}\mathbb{E}[\phi(X(t)) \mid X(s)]}{\mathrm{d}t}(t) = \mathbb{E}[\mathscr{A}\phi(X(t)) \mid X(s)] \tag{3.13}$$

when $\mathbb{E}[\phi(X(t)) \mid X(s)]$ is seen as function of t, so that the claim in Equation (3.10) holds for $r = 1$. Suppose now that Equation (3.10) holds for an $r > 1$, then by applying Itô's formula again on $\mathscr{A}^r\phi(X(t))$ we obtain

$$\mathscr{A}^r\phi(X(t)) = \mathscr{A}^r\phi(X(s)) + \int_s^t \mathscr{A}^{r+1}\phi(X(\tau))\,\mathrm{d}\tau. \tag{3.14}$$

Noting the fact that

$$\frac{\mathrm{d}^r\mathbb{E}[\phi(X(t)) \mid X(s)]}{\mathrm{d}t^r}(s) = \mathscr{A}^r\phi(X(s)),$$

we can take expectations on both sides of Equation (3.14), and substitute the resulting expression into Equation (3.10), we thus obtain

$$\begin{aligned}\frac{\mathrm{d}^r\mathbb{E}[\phi(X(t)) \mid X(s)]}{\mathrm{d}t^r}(t) &= \mathscr{A}^r\phi(X(s)) + \int_s^t \mathbb{E}\left[\mathscr{A}^{r+1}\phi(X(\tau)) \mid X(s)\right]\mathrm{d}\tau \\ &= \frac{\mathrm{d}^r\mathbb{E}[\phi(X(t)) \mid X(s)]}{\mathrm{d}t^r}(s) + \int_s^t \mathbb{E}\left[\mathscr{A}^{r+1}\phi(X(\tau)) \mid X(s)\right]\mathrm{d}\tau\end{aligned}$$

which is the integral form of the ordinary differential equation

$$\frac{\mathrm{d}^{r+1}\mathbb{E}[\phi(X(t)) \mid X(s)]}{\mathrm{d}t^{r+1}}(t) = \mathbb{E}\left[\mathscr{A}^{r+1}\phi(X(t)) \mid X(s)\right].$$

Hence, Equation (3.10) is proven.

Finally, by Taylor's theorem, we arrive at Equation (3.8). The remainder in Equation (3.9) is obtained by taking expectations on both sides of Equation (3.14) and substituting back into Equation (3.12) multiple times for $r = 1, \ldots, M$. The proof details can be found in Dacunha-Castelle and Florens-Zmirou (1986, Lemma 4) or Florens-Zmirou (1989, Lemma 1), for example. □

Note that even though the expansion is taken up an order $M \geq 0$, the TME method gives an exact representation of $\mathbb{E}[\phi(X(t)) \mid X(s)]$ for any suitable function ϕ. However, computing the remainder is infeasible in practice, and we usually approximate the representation by discarding the remainder[1]. This leads to a polynomial approximation with respect to Δt. However, please note that the order M cannot be chosen entirely arbitrarily because it depends on the smoothness of the SDE coefficients and function ϕ.

[1] If we discard the remainder, then the TME *approximation* only needs a and b to be $M-1$ times differentiable and ϕ to be $2M$ times differentiable.

In Gaussian filtering and smoothing we are particularly interested in estimating the conditional means and covariances of the process X. In order to do so, we introduce the following target functions

$$\begin{aligned} \phi^{\mathrm{I}}(x) &= x, \\ \phi^{\mathrm{II}}(x) &= x\,x^{\mathsf{T}}, \end{aligned} \tag{3.15}$$

corresponding to the first and second moments, respectively. Their TME representations are then given in Lemma 3.4.

Remark 3.3. *While generator $\mathscr{A}$ in Equation* (3.5) *is defined for scalar-valued target functions only, this definition can be extend to vector/matrix-valued target functions by introducing an elementwise operator $\overline{\mathscr{A}}$. Namely, let $\phi\colon \mathbb{R}^d \to \mathbb{R}^{m\times n}$, then $\overline{\mathscr{A}}$ is defined via*

$$\overline{\mathscr{A}}\phi(x) = \begin{bmatrix} \mathscr{A}\phi_{11} & \cdots & \mathscr{A}\phi_{1n} \\ \vdots & \ddots & \vdots \\ \mathscr{A}\phi_{m1} & \cdots & \mathscr{A}\phi_{mn} \end{bmatrix}(x), \tag{3.16}$$

where ϕ_{ij} stands for the i,j-th element of ϕ.

Lemma 3.4 (TME for first and second moments)**.** *The first and second conditional moments of X are given by*

$$\mathbb{E}[X(t) \mid X(s)] = \sum_{r=0}^{M} \frac{1}{r!}\overline{\mathscr{A}}\phi^{\mathrm{I}}(X(s))\Delta t^r + R_{M,\phi^{\mathrm{I}}}(X(s),\Delta t) \tag{3.17}$$

and

$$\mathbb{E}\left[X(t)X(t)^{\mathsf{T}} \mid X(s)\right] = \sum_{r=0}^{M} \frac{1}{r!}\overline{\mathscr{A}}\phi^{\mathrm{II}}(X(s))\Delta t^r + R_{M,\phi^{\mathrm{II}}}(X(s),\Delta t) \tag{3.18}$$

for all $s < t \in \mathbb{T}$, respectively.

Notice that if we choose $M = 1$ in the lemma above, then the resulting TME approximation $\sum_{r=0}^{1} \frac{1}{r!}\overline{\mathscr{A}}\phi^{\mathrm{I}}(X(s))\Delta t^r$ is exactly the same as the Euler–Maruyama approximation for the first moment. Moreover, the TME covariance approximation (formulated in the next section) will also coincide with the Euler–Maruyama approximation for the covariance when $M = 1$.

3.4 Covariance approximation by TME

This section shows how to use the TME method to approximate conditional covariances of the form in Equation (3.6). Based on the first and second

moment representations in Lemma 3.4, it seems that we can approximate the covariance $\mathrm{Cov}[X(t) \mid X(s)]$ by

$$\begin{aligned}\mathrm{Cov}[X(t) \mid X(s)] &= \mathbb{E}\left[X(t)X(t)^{\mathsf{T}} \mid X(s)\right] - \mathbb{E}[X(t) \mid X(s)]\mathbb{E}[X(t) \mid X(s)]^{\mathsf{T}} \\ &\approx \sum_{r=0}^{M} \frac{1}{r!} \overline{\mathscr{A}}^r \phi^{\mathrm{II}}(X(s))\Delta t^r \\ &\quad - \left(\sum_{r=0}^{M} \frac{1}{r!} \overline{\mathscr{A}}^r \phi^{\mathrm{I}}(X(s))\Delta t^r\right)\left(\sum_{r=0}^{M} \frac{1}{r!} \overline{\mathscr{A}}^r \phi^{\mathrm{I}}(X(s))\Delta t^r\right)^{\mathsf{T}}\end{aligned} \tag{3.19}$$

up to an order M. However, this approximation has two problems. First, the polynomial degree in this approximation is inconsistent with the approximations of the first and second moments. This is because the power of Δt in Equation (3.19) is now up to order $2M$ instead of M. Hence, we need to truncate the polynomial terms Δt^r for $r > M$ in Equation (3.19) for the sake of consistency.

The second problem is that the positive definiteness of the covariance approximation is not guaranteed as we discard the remainders (Iacus, 2008; Zhao et al., 2021b). To see this, let us consider a simple one-dimensional example as follows. Let $X : \mathbb{T} \to \mathbb{R}$ be an Itô process that solves the SDE (3.4). Suppose that its dispersion term is non-zero and let us also choose $M = 2$. Then the variance approximation in Equation (3.19), after truncating the polynomial terms Δt^r for $r > M$, becomes $\Gamma(X(s))\Delta t + \Gamma(X(s))\frac{\mathrm{d}a}{\mathrm{d}X}(X(s))\Delta t^2$. This approximation is not positive in general because it is positive if and only if $\frac{\mathrm{d}a}{\mathrm{d}X}(X(s)) > -1/\Delta t$. Moreover, if one requires the positivity hold uniformly for all $\Delta t \in \mathbb{R}_{>0}$ and all $X(s) \in \mathbb{R}$, then the function $\frac{\mathrm{d}a}{\mathrm{d}X}$ must be positive on its domain.

Therefore, in the following theorem we derive the TME approximation for the covariance $\mathrm{Cov}[X(t) \mid X(s)]$ by truncating the unnecessary polynomial terms of Δt in Equation (3.19), and we thereupon provide a sufficient criterion to ensure the positive definiteness of such approximation.

Theorem 3.5 (TME covariance approximation)**.** *Let $X : \mathbb{T} \to \mathbb{R}^d$ be the solution of the SDE that verifies Theorem 3.2. Let integer $M \geq 1$. The M-order TME approximation for* $\mathrm{Cov}[X(t) \mid X(s)]$ *is*

$$\Sigma_M(\Delta t) = \sum_{r=1}^{M} \frac{1}{r!} \Theta_r \Delta t^r, \tag{3.20}$$

where

$$\begin{aligned}\Theta_r &:= \Theta_{X(s),r} \\ &= \overline{\mathscr{A}}^r \phi^{\mathrm{II}}(X(s)) - \sum_{k=0}^{r} \binom{r}{k} \overline{\mathscr{A}}^k \phi^{\mathrm{I}}(X(s)) \left(\overline{\mathscr{A}}^{r-k} \phi^{\mathrm{I}}(X(s))\right)^{\mathsf{T}},\end{aligned} \tag{3.21}$$

and $\binom{r}{k}$ *denotes binomial coefficient. The approximation* $\Sigma_M(\Delta t)$ *is positive definite if the associated polynomial*

$$\chi(\Delta t) = \sum_{r=1}^{M} \frac{1}{r!} \lambda_{\min}(\Theta_r) \Delta t^r > 0. \tag{3.22}$$

Proof. Let us denote by ϕ^{II}_{ij} the i,j-th element of ϕ^{II}, and let us also denote by ϕ^{I}_i the i-th element of ϕ^{I}. Then the i,j-th element of the covariance approximation in Equation (3.19) is

$$\begin{aligned} &\sum_{r=1}^{M} \frac{1}{r!} \mathscr{A}^r \phi^{\mathrm{II}}_{ij}(X(s)) \Delta t^r \\ &\quad - \left(\sum_{r=1}^{M} \frac{1}{r!} \mathscr{A}^r \phi^{\mathrm{I}}_i(X(s)) \Delta t^r \right) \left(\sum_{r=1}^{M} \frac{1}{r!} \mathscr{A}^r \phi^{\mathrm{I}}_j(X(s)) \Delta t^r \right). \end{aligned} \tag{3.23}$$

Let $[\Sigma_M]_{ij}$ be the truncation of Equation (3.23) up to order M (i.e., eliminating terms with Δt^r for all $r > M$). Then, by Cauchy product (see, Theorem 2.13) we have

$$\begin{aligned} [\Sigma_M]_{ij} &= \sum_{r=1}^{M} \left[\frac{1}{r!} \mathscr{A}^r \phi^{\mathrm{II}}_{ij}(X(s)) - \left(\sum_{k=0}^{r} \frac{\mathscr{A}^k \phi^{\mathrm{I}}_i(X(s)) \mathscr{A}^{r-k} \phi^{\mathrm{I}}_j(X(s))}{k!(r-k)!} \right) \right] \Delta t^r \\ &= \sum_{r=1}^{M} \frac{1}{r!} \left[\mathscr{A}^r \phi^{\mathrm{II}}_{ij}(X(s)) - \sum_{k=0}^{r} \binom{r}{k} \mathscr{A}^k \phi^{\mathrm{I}}_i(X(s)) \mathscr{A}^{r-k} \phi^{\mathrm{I}}_j(X(s)) \right] \Delta t^r. \end{aligned}$$

Hence, by rearranging $[\Sigma_M]_{ij}$ into a matrix for $i,j = 1,\ldots,d$ we obtain Equation (3.20). Since $\Sigma_M(\Delta t)$ is symmetric by definition, its eigenvalues are real. Then by using Weyl's inequality (see, Theorem 2.14) we obtain

$$\lambda_{\min}(\Sigma_M(\Delta t)) \geq \sum_{r=1}^{M} \frac{1}{r!} \lambda_{\min}(\Theta_r) \Delta t^r. \tag{3.24}$$

Hence, $\Sigma_M(\Delta t)$ is positive definite if Equation (3.22) holds. Note that $\Theta_0 = 0$. □

Theorem 3.5 shows that $\Sigma_M(\Delta t)$ is a polynomial of Δt with coefficients determined by Hermitian matrices $\{\Theta_r : r = 1,\ldots,M\}$. These matrices depend on the starting condition $X(s)$. In order to guarantee the positive definiteness of $\Sigma_M(\Delta t)$, we use Weyl's inequality in order to find a lower bound on its minimum eigenvalue, resulting in another polynomial $\chi(\Delta t)$ of Δt. This reduces the problem of analysing the positive definiteness of $\Sigma_M(\Delta t)$ into the problem of analysing the positivity of polynomial $\chi(\Delta t)$.

To ensure the positivity of polynomial $\chi(\Delta t)$, one can trivially restrict all the coefficients $\{\lambda_{\min}(\Theta_r) : r = 1,\ldots,M\}$ to be positive, but this in turn significantly limit the SDE models that the TME approximation applies. Another solution is to let $\chi(\Delta t)$ have no real roots on $\mathbb{R}_{>0}$ and $\chi(\Delta t) > 0$

for some $\Delta t \in \mathbb{R}_{>0}$. For instance, one can bound/count the number of real roots of polynomial on any intervals by using Budan's theorem or Sturm's theorem (Basu et al., 2006).

The positive definiteness of $\Sigma_M(\Delta t)$ is entirely determined by the order M, the time interval Δt, the starting point $X(s)$, and the SDE coefficients. If Δt is somehow tunable, one can then let Δt be small enough to guarantee the positive definiteness numerically. This is true because the term $\Theta_1 = \Gamma(X(s))$ which is positive semi-definite by definition, dominates $\Sigma_M(\Delta t)$ in the limit $\Delta t \to 0$. This numerical approach is especially useful in Gaussian filtering and smoothing, as it is common to perform multiple integration steps with small Δt in the prediction steps (see, Algorithm 2.11).

However, it might not always be possible to tune Δt. For example, if we have limited computational resources, using multiple integration steps with smaller Δt in Gaussian filtering and smoothing may not be realistic. Hence, it is also important to show the positive definiteness conditions of $\Sigma_M(\Delta t)$ that are independent of the choice of Δt. A few results on these conditions are given in the following corollary.

Corollary 3.6. *The following results hold for all* $\Delta t \in \mathbb{R}_{>0}$.

I. $\Sigma_1(\Delta t)$ *is positive definite, if* $\Gamma(X(s))$ *is positive definite. Notice that* $\Gamma(X(s))$ *is always positive semi-definite by definition.*

II. $\Sigma_2(\Delta t)$ *is positive definite, if* Θ_2 *and* $\Gamma(X(s))$ *are positive semi-definite, and one of the two is positive definite.*

III. $\Sigma_3(\Delta t)$ *is positive definite, if* Θ_3 *is positive semi-definite and* $\lambda_{\min}(\Theta_2) > \frac{-2\sqrt{6}}{3}\sqrt{\lambda_{\min}(\Theta_1)\,\lambda_{\min}(\Theta_3)}$.

Proof. This corollary follows from Theorem 3.5 and the root conditions of quadratic and cubic polynomials (i.e., by letting $\chi(\Delta t)$ have no real roots on $\mathbb{R}_{>0}$). See, Zhao et al. (2021b, Proposition 5) for details. □

Remark 3.7. *For* $r = 0, 1, 2$ *we can immediately derive* $\Theta_0 = 0$, $\Theta_1 = \Gamma(X(s))$, *and* $\Theta_2 = \Gamma(X(s))\,\mathrm{J}_X a(X(s)) + (\Gamma(X(s))\,\mathrm{J}_X a(X(s)))^\mathsf{T}$. *For results in higher orders (and in one state dimension), see, Zhao et al. (2021b, Example 9).*

The approximation Σ_M has an important property that it does not *explicitly* depend on $X(s)$. More precisely, the expression of Σ_M only have $X(s)$ appearing inside the SDE coefficients and their derivatives. With a slight abuse of terminology, we say that Σ_M is $X(s)$-*homogeneous*. This property is meaningful in the sense that it is possible to ensure the positive definiteness of Σ_M independent of $X(s)$ by manipulating the SDE coefficients.

Lemma 3.8 ($X(s)$-homogeneity)**.** *Let* $b(X(t)) = b \in \mathbb{R}^{d\times w}$ *be a constant, hence* $\Gamma = b\,b^\mathsf{T}$. *Denote by* Θ_r^{uv} *the* u,v*-th element of* Θ_r *and* Γ_{ij} *the* i,j*-th element*

of Γ. *Also denote by* $\alpha_r^u := \mathscr{A}^r \phi_u^{\mathrm{I}}(X(s))$. *Then*

$$\begin{aligned}\Theta_r^{uv} &= \sum_{i,j=1}^{d} \sum_{k=0}^{r-1} \binom{r-1}{k} \frac{\partial \alpha_k^u}{\partial X_i(s)} \frac{\partial \alpha_{r-k-1}^v}{\partial X_j(s)} \Gamma_{ij} + \mathscr{A}\Theta_{r-1}^{uv} \\ &= \sum_{k=0}^{r-1} \overline{\mathscr{A}}^k \sum_{l=0}^{r-k-1} \binom{r-k-1}{l} \operatorname{tr}\left(\nabla_X \alpha_{r-k-l-1}^v \left(\nabla_X \alpha_k^u\right)^\top \Gamma\right)\end{aligned} \tag{3.25}$$

and $\Theta_0^{uv} = 0$, *for all* $r \geq 1$ *and* $u, v \leq d$. *Notice that* $\overline{\mathscr{A}}^r \phi^{\mathrm{I}}(X(s))$ *is* $X(s)$*-homogeneous for* $r \geq 0$.

Proof. Define $\beta_r^{uv} := \mathscr{A}^r \phi_{uv}^{\mathrm{II}}(X(s))$. Since $\Theta_r^{uv} = \beta_r^{uv} - \sum_{k=0}^{r} \binom{r}{k} \alpha_k^u \alpha_{r-k}^v$ by Equation (3.21), the task is to find an $X(s)$-homogeneous expression for Θ_r^{uv}. If we do a few initial trials for $r = 0, 1, \ldots$, we will find a pattern

$$\begin{aligned}\beta_0^{uv} &= \alpha_0^u \alpha_0^v, \\ \beta_1^{uv} &= \alpha_0^u \alpha_1^v + \alpha_0^v \alpha_1^u + \Gamma_{uv}, \\ &\vdots\end{aligned} \tag{3.26}$$

Hence, we want to prove that

$$\beta_r^{uv} = \sum_{k=0}^{r} \binom{r}{k} \alpha_k^u \alpha_{r-k}^v + \Theta_r^{uv}, \tag{3.27}$$

where

$$\Theta_r^{uv} = \sum_{i,j=1}^{d} \sum_{k=0}^{r-1} \binom{r-1}{k} \frac{\partial \alpha_k^u}{\partial X_i(s)} \frac{\partial \alpha_{r-k-1}^v}{\partial X_j(s)} \Gamma_{ij} + \mathscr{A}\Theta_{r-1}^{uv}. \tag{3.28}$$

Equation (3.27) holds for $r = 0$ and 1. Now let us suppose that they hold for an $r > 1$. Then, by the definition of β_r^{uv} we have

$$\begin{aligned}\beta_{r+1}^{uv} = \mathscr{A}^{r+1} \phi_{uv}^{\mathrm{II}}(X(s)) &= \mathscr{A}\beta_r^{uv} \\ &= \sum_{i=1}^{d} \frac{\partial \beta_r^{uv}}{\partial X_i(s)} a_i(X(s)) + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 \beta_r^{uv}}{\partial X_i(s) \partial X_j(s)} \Gamma_{uv}.\end{aligned} \tag{3.29}$$

Now, by substituting Equation (3.27) in Equation (3.29), Equation (3.29) becomes

$$\begin{aligned}&\beta_{r+1}^{uv} = \\ &\sum_{i=1}^{d} \sum_{k=0}^{r} \binom{r}{k} \left(\frac{\partial \alpha_k^u}{\partial X_i(s)} \alpha_{r-k}^v + \alpha_k^u \frac{\partial \alpha_{r-k}^v}{\partial X_i(s)} \right) a_i(X(s)) \\ &\quad + \frac{1}{2} \sum_{i,j=1}^{d} \left(\sum_{k=0}^{r} \binom{r}{k} \Big(\frac{\partial^2 \alpha_k^u}{\partial X_i(s) \partial X_j(s)} \alpha_{r-k}^v + \frac{\partial \alpha_k^u}{\partial X_i(s)} \frac{\partial \alpha_{r-k}^v}{\partial X_j(s)} + \frac{\partial \alpha_k^u}{\partial X_j(s)} \frac{\partial \alpha_{r-k}^v}{\partial X_i(s)} \right. \right. \\ &\qquad \left. \left. + \alpha_k^u \frac{\partial^2 \alpha_{r-k}^v}{\partial X_i(s) \partial X_j(s)} \Big) \right) \Gamma_{ij} + \mathscr{A}\Theta_r^{uv}.\end{aligned}$$

By the definition of generator $\mathscr{A}$ we have that $\sum_{i=1}^{d} \frac{\partial \alpha_k^u}{\partial X_i(s)} \alpha_{r-k}^v a_i(X(s)) + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 \alpha_k^u}{\partial X_i(s) \partial X_j(s)} \alpha_{r-k}^v = \mathscr{A} \alpha_k^u \alpha_{r-k}^v = \alpha_{k+1}^u \alpha_{r-k}^v$. Defining $\binom{r}{-1} = 0$, we arrive at

$$\begin{aligned}
\beta_{r+1}^{uv} &= \sum_{k=0}^{r} \binom{r}{k} \left(\alpha_{k+1}^u \alpha_{r-k}^v + \alpha_k^u \alpha_{r-k+1}^v \right) + \sum_{i,j=1}^{d} \sum_{k=0}^{r} \binom{r}{k} \frac{\partial \alpha_k^u}{\partial X_i(s)} \frac{\partial \alpha_{r-k}^v}{\partial X_j(s)} \Gamma_{ij} + \mathscr{A} \Theta_r^{uv} \\
&= \sum_{k=0}^{r+1} \binom{r}{k-1} \alpha_k^u \alpha_{r-k+1}^v + \sum_{k=0}^{r+1} \binom{r}{k} \alpha_k^u \alpha_{r-k+1}^v + \Theta_{r+1}^{uv} \\
&= \sum_{k=0}^{r+1} \binom{r+1}{k} \alpha_k^u \alpha_{r-k+1}^v + \Theta_{r+1}^{uv}
\end{aligned}$$

which is exactly Equation (3.27) at $r+1$. Thus, Equation (3.27) is proven by mathematical induction. Finally,

$$\begin{aligned}
\Theta_r^{uv} &= \beta_r^{uv} - \sum_{k=0}^{r} \binom{r}{k} \alpha_k^u \alpha_{r-k}^v \\
&= \sum_{i,j=1}^{d} \sum_{k=0}^{r-1} \binom{r-1}{k} \frac{\partial \alpha_k^u}{\partial X_i(s)} \frac{\partial \alpha_{r-k-1}^v}{\partial X_j(s)} \Gamma_{ij} + \mathscr{A} \Theta_{r-1}^{uv} \\
&= \sum_{k=0}^{r-1} \binom{r-1}{k} \operatorname{tr} \left(\nabla_X \alpha_{r-k-1}^v \left(\nabla_X \alpha_k^u \right)^\top \Gamma \right) + \mathscr{A} \Theta_{r-1}^{uv}.
\end{aligned}$$

Starting from $\Theta_0 = 0$, one can arrive at the last line in Equation (3.25) by iterating Θ_r^{uv} for $r \geq 1$. □

The homogeneity property does not hold for the first and second moment approximations in Lemma 3.4. For instance, the TME mean approximation reads $\mathbb{E}[X(t) \mid X(s)] \approx X(s) + a(X(s))\Delta t + \cdots$ which explicitly depends on $X(s)$.

3.5 Numerical examples of TME

In this section we present a few examples that apply the TME method for approximating expectations of the form in Equation (3.6). In addition, we compare the results of TME against some commonly-used methods, such as the Euler–Maruyama scheme and the Itô–Taylor strong order 1.5 (Itô-1.5) method (Kloeden and Platen, 1992). In Example 3.10, we present a concrete example showing how to use Theorem 3.5 to analyse the positive definiteness of a TME covariance approximation.

For simplicity we call TME-M the M-order TME approximation.

Example 3.9. *Consider an Itô process $X : \mathbb{T} \to \mathbb{R}$ which solves the Beneš model*

$$\mathrm{d}X(t) = \tanh(X(t))\,\mathrm{d}t + \mathrm{d}W(t), \tag{3.30}$$

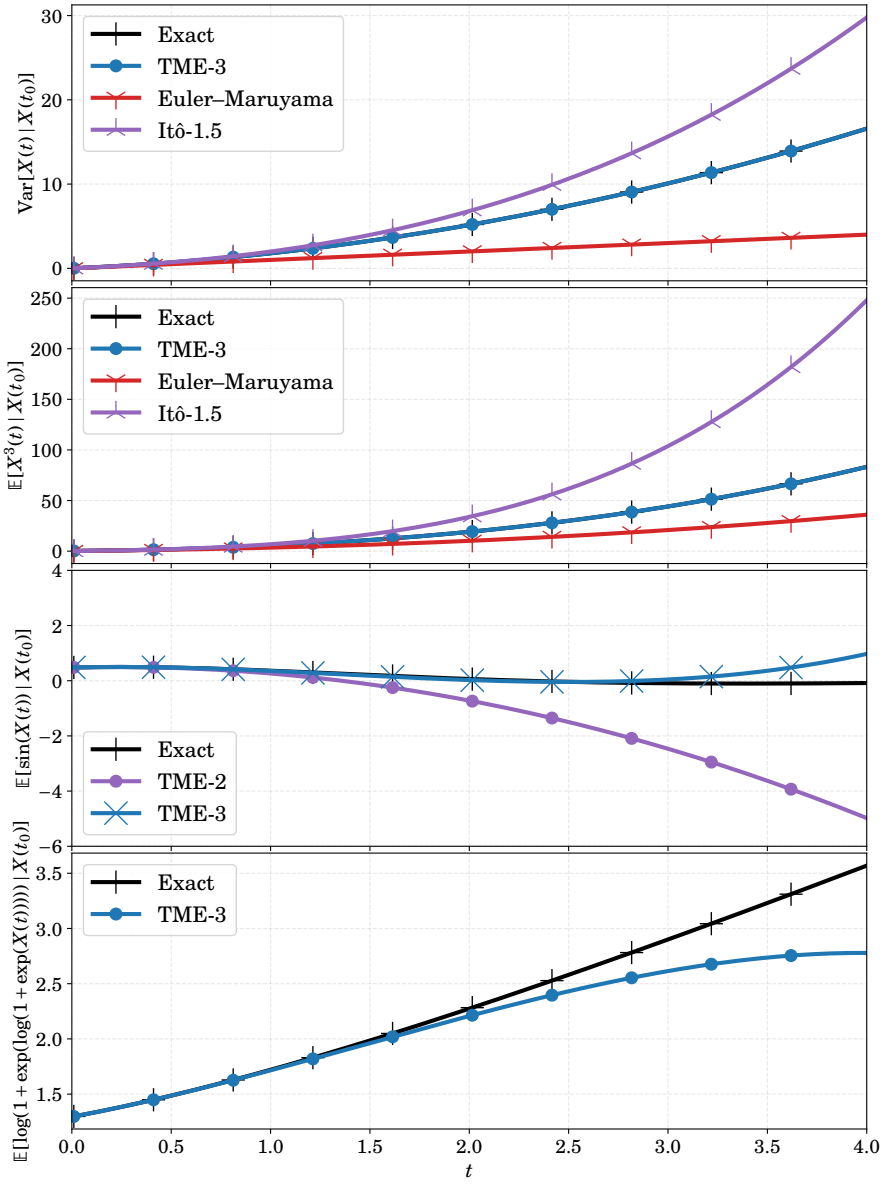


Figure 3.1. Expectation approximations in Example 3.9. The exact solutions are computed by numerical integration, as the transition density of the Beneš model is explicitly known (Särkkä and Solin, 2019, Equation 10.58).

starting from $X(t_0) = 0.5$. *We are interested in computing its variance* $\mathrm{Var}[X(t) \mid X(t_0)]$, *third moment* $\mathbb{E}[X(t)^3 \mid X(t_0)]$, *and two expectations*

$$\begin{aligned} &\mathbb{E}[\sin(X(t)) \mid X(t_0)], \\ &\mathbb{E}[\log(1+\exp(\log(1+\exp(X(t))))) \mid X(t_0)]. \end{aligned} \tag{3.31}$$

Notice that one can understand the last expectation above as a way to describe the propogation of X *through a neural network consiting of two single-neuron layers with Softplus activation functions.*

The TME-2 approximation for the variance $\mathrm{Var}[X(t) \mid X(t_0)]$ *is exact. Specifically,* $\Sigma_2(\Delta t) = \Delta t + (1 - \tanh(X(t_0))^2)\Delta t^2$ *is equal to* $\mathrm{Var}[X(t) \mid X(t_0)]$.

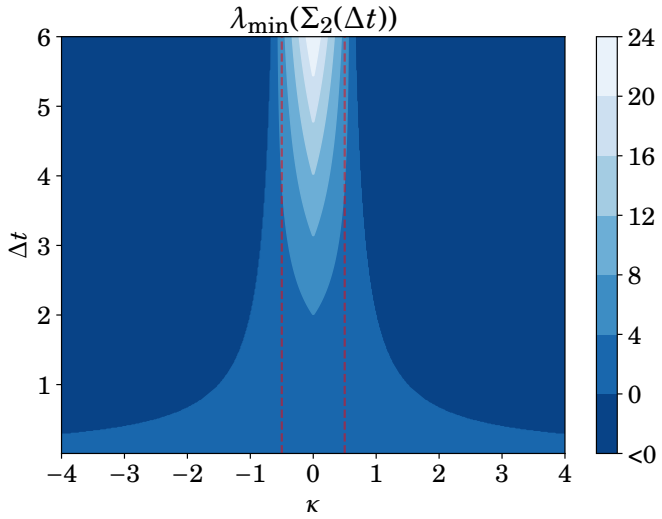


Figure 3.2. Contour plot of the minimum eigenvalue of Σ_2 with respect to Δt and κ in Example 3.10. Red dashed lines stand for $\kappa = -0.5$ and 0.5.

In Figure 3.1, we plot the results for the expectations in Example 3.9. In addition, we compare the TME method against the Euler–Maruyama and Itô-1.5 methods. From the figure, we see that the TME approach outperforms the Euler–Maruyama and Itô-1.5 methods significantly. Also, the TME approach can approximate the expectations in Equation (3.31) to a good extent within a small time span. Note that the Euler–Maruyama and Itô-1.5 schemes do not give closed-form approximations for the expectations in Equation (3.31), we thus omit the two methods for these expectations.

In the next example, we show how to use Theorem 3.5 and Corollary 3.6 in practice to analyse the positive definiteness of the TME covariance approximation of a non-linear multidimensional SDE.

Example 3.10. *Consider a two-dimensional SDE*

$$\begin{aligned} \mathrm{d}X^1(t) &= (\log(1+\exp(X^1(t))) + \kappa X^2(t))\,\mathrm{d}t + \mathrm{d}W_1(t), \\ \mathrm{d}X^2(t) &= (\log(1+\exp(X^2(t))) + \kappa X^1(t))\,\mathrm{d}t + \mathrm{d}W_2(t), \end{aligned} \tag{3.32}$$

where $\kappa \in \mathbb{R}$ is a tunable parameter. We want to ensure the positive definiteness of $\Sigma_2(\Delta t)$ for all $\Delta t \in \mathbb{R}_{>0}$ by tuning κ. In order to do so, we can first explicitly derive Θ_1 and Θ_2. It turns out that Θ_1 is an identity matrix and

$$\Theta_2 = 2 \begin{bmatrix} \frac{e^{X^1(t_0)}}{e^{X^1(t_0)}+1} & \kappa \\ \kappa & \frac{e^{X^2(t_0)}}{e^{X^2(t_0)}+1} \end{bmatrix}. \tag{3.33}$$

Then, by Corollary 3.6 it is sufficient to guarantee the positive definiteness of $\Sigma_2(\Delta t)$ for all $\Delta t \in \mathbb{R}_{>0}$ by ensuring that Θ_2 is positive semi-definite. Thus, one should let

$$|\kappa| \leq \sqrt{\frac{e^{X^1(t_0)+X^2(t_0)}}{(e^{X^1(t_0)}+1)(e^{X^2(t_0)}+1)}}.$$

Figure 3.2 plots the minimum eigenvalues of $\Sigma_2(\Delta t)$ *with respect to* Δt *and* κ *when* $X^1(t_0) = X^2(t_0) = 0$*. In this case,* $|\kappa|$ *should be less than* 0.5 *(red dashed lines in the figure) in order to guarantee the positive definiteness of* $\Sigma_2(\Delta t)$*. The figure shows that* $\Sigma_2(\Delta t)$ *is indeed positive definite for all* $\Delta t \in \mathbb{R}_{>0}$ *within the region* $|\kappa| \leq 0.5$*, and that this sufficient region is very close to the true region (i.e., the region of* κ *that* $\lambda_{\min}(\Sigma_2(\Delta t)) > 0$ *for all* Δt*).*

The positive definiteness analysis via Theorem 3.5 might be limited to simple SDEs and low orders of expansion. For $M \leq 3$, some explicit results can be stated as shown in Corollary 3.6, but higher-order expansions can result in complicated polynomials to be analysed.

3.6 TME Gaussian filter and smoother

In the pioneering works by Dacunha-Castelle and Florens-Zmirou (1986); Kessler (1997); Aït-Sahalia (2003), the TME method was originally introduced for estimating unknown parameters of SDEs. More specifically, they use TME to discretise SDEs in order to perform maximum likelihood estimations. In this section, we show that the TME method could also be applied for solving Gaussian filtering and smoothing problems (Zhao et al., 2021b; Zhao and Särkkä, 2021).

Consider a (time-homogeneous) continuous-discrete state-space model

$$\begin{aligned} \mathrm{d}X(t) &= a(X(t))\,\mathrm{d}t + b(X(t))\,\mathrm{d}W(t), \quad X(t_0) = X_0, \\ Y_k &= h(X_k) + \xi_k, \quad \xi_k \sim \mathrm{N}(0, \Xi_k), \end{aligned} \tag{3.34}$$

where the solution $X \colon \mathbb{T} \to \mathbb{R}^d$ is observed through a non-linear function $h \colon \mathbb{R}^d \to \mathbb{R}^{d_y}$ and additive Gaussian noises $\{\xi_k : k = 1, 2, \ldots\}$. Furthermore, we assume that the SDE coefficients satisfy the conditions in Theorem 3.2, so that we can apply the TME method.

As shown in Algorithm 2.11, a key procedure of Gaussian filtering is to propagate the previous filtering result through the SDE and compute the predictive mean m_k^- and covariance P_k^-. As for the Gaussian smoothing steps, one needs to compute the cross-covariance D_{k+1}^- in Algorithm 2.11. These quantities can be approximated by using the TME method as follows.

Let us denote by f^M and Q^M the M-order TME approximations to the conditional mean and covariance (see, Lemma 3.4 and Theorem 3.5), that are,

$$\begin{aligned} \mathbb{E}[X_k \mid X_{k-1}] &\approx f^M(X_{k-1}), \\ \mathrm{Cov}[X_k \mid X_{k-1}] &\approx Q^M(X_{k-1}). \end{aligned} \tag{3.35}$$

Then by substituting f^M and Q^M into the prediction step in Algorithm 2.11

we obtain the TME-approximated predictive mean and covariance

$$\begin{aligned}
&\int x_k \, p_{X_k \mid Y_{1:k-1}}(x_k \mid y_{1:k-1}) \mathrm{d}x_k \\
&= \iint x_k \, p_{X_k \mid X_{k-1}}(x_k \mid x_{k-1}) \, p_{X_{k-1} \mid Y_{k-1}}(x_{k-1} \mid y_{k-1}) \mathrm{d}x_{k-1} \mathrm{d}x_k \\
&\approx \int f^M(x_{k-1}) \mathrm{N}\left(x_{k-1} \mid m_{k-1}^f, P_{k-1}^f\right) \mathrm{d}x_{k-1} \\
&= m_k^-, \\
&\int (x_k - m_k^-)(x_k - m_k^-)^\mathsf{T} \, p_{X_k \mid Y_{1:k-1}}(x_k \mid y_{1:k-1}) \mathrm{d}x_k \\
&\approx \int \left(Q^M(x_{k-1}) + f^M(x_{k-1}) \left(f^M(x_{k-1})\right)^\mathsf{T}\right) \mathrm{N}\left(x_{k-1} \mid m_{k-1}^f, P_{k-1}^f\right) \mathrm{d}x_{k-1} \\
&\quad - m_k^- (m_k^-)^\mathsf{T} \\
&= P_k^-.
\end{aligned} \tag{3.36}$$

Similarly, for the cross-covariance D_{k+1} in the smoothing pass we have

$$\begin{aligned}
&\iint x_k \, x_{k+1}^\mathsf{T} \, p_{X_{k+1} \mid X_k}(x_{k+1} \mid x_k) \, p_{X_k \mid Y_{1:k}}(x_k \mid y_{1:k}) \mathrm{d}x_k \, \mathrm{d}x_{k+1} - m_k^f (m_{k+1}^-)^\mathsf{T} \\
&\approx \int x_k \left(f^M(x_k)\right)^\mathsf{T} \mathrm{N}\left(x_k \mid m_k^f, P_k^f\right) \mathrm{d}x_k - m_k^f (m_{k+1}^-)^\mathsf{T} \\
&= D_{k+1}.
\end{aligned} \tag{3.37}$$

We formally define the TME Gaussian filter and smoother in the following algorithm.

Algorithm 3.11 (TME Gaussian filter and smoother)**.** *The algorithm is the same as Algorithm 2.11, except that the computations for the prediction (i.e., m_k^- and P_k^-) and cross-covariance (i.e., D_{k+1}) are replaced by Equations* (3.36) *and* (3.37)*, respectively.*

The expectations in Equations (3.36) and (3.37) are usually computed by quadrature integration methods (e.g., sigma-point methods), since the approximations f^M and Q^M are usually non-linear functions.

3.6.1 Filter stability

The filter stability in this context refers to the error bound of the filtering estimates in the mean-square sense. For Kalman filters, some classical stability results are already shown, for example, by Jazwinski (1970) and Anderson and Moore (1981). As for non-linear filters, their stability analyse has also been studied extensively in recent decades. For example, Reif et al. (1999) analyse the stability of extended Kalman filters, while

the stability of more general Gaussian filters are found in Itô and Xiong (2000); Xiong et al. (2006). There are also stability analysis that are model-specific. For instance, Blömker et al. (2013) and Law et al. (2014) analyse the stability of a class of Gaussian filters on the Navier–Stokes equation and a Lorenz model, respectively. In the remainder of this section, we rely on the stability results in Karvonen et al. (2020) which apply for a wide class of non-linear filters and non-linear state-space models including ours.

In this section, we analyse the stability of the TME Gaussian filters (see, Algorithm 3.11) that use sigma-point integration methods for computing the expectations in Equation (3.36). This analysis is necessary, as it is important to know if the TME Gaussian filtering error – which accumulates in time – is in some sense bounded. The sources of the error include, for example, TME approximations, Gaussian approximations to the filtering posterior distributions, and numerical integration.

To proceed, let

$$X_k = \check{f}(X_{k-1}) + \check{q}(X_{k-1}) \tag{3.38}$$

stand for the *exact* discretisation of the SDE in Equation (3.34) for $k = 1,2,\ldots$, where $\check{f}(X_{k-1}) \coloneqq \mathbb{E}[X_k \mid X_{k-1}]$, and $\check{q}(X_{k-1})$ is a zero-mean random variable whose conditional covariance is $\check{Q}(X_{k-1}) \coloneqq \mathrm{Cov}[\check{q}(X_k) \mid X_{k-1}]$. The principle of TME Gaussian filters is such that the TME method approximates X_k via the discretisation

$$X_k \approx f^M(X_{k-1}) + q^M(X_{k-1}), \tag{3.39}$$

where $q^M(X_{k-1}) \sim \mathrm{N}(0, Q^M(X_{k-1}))$. By Theorem 3.2 or Lemma 3.4 we have

$$\check{f}(X_{k-1}) = f^M(X_{k-1}) + R_M(X_{k-1}), \tag{3.40}$$

where we abbreviate the remainder by $R_M(X_{k-1}) \coloneqq R_{M,\phi^{\mathrm{I}}}(X_{k-1}, \Delta t_k)$.

Now suppose that we perform TME Gaussian filtering on a linearly-observed state-space model

$$\begin{aligned} X_k &= \check{f}(X_{k-1}) + \check{q}(X_{k-1}), \\ Y_k &= H X_k + \xi_k, \quad \xi_k \sim \mathrm{N}(0, R), \end{aligned} \tag{3.41}$$

defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where $H \in \mathbb{R}^{d_y \times d}$ and $R \in \mathbb{R}^{d_y \times d_y}$ are constant matrices. Here, we limited ourselves to linear measurement models in order to use the preceding results by Karvonen et al. (2020).

In the following, sigma-point approximations of Gaussian integrals of the form $\int z(x)\mathrm{N}(x \mid m, P)\,\mathrm{d}x$ are denoted by $\mathscr{S}_{m,P}(z)$. The sigma-point TME Gaussian filter is such that the predictive mean in Algorithm 3.11 becomes $m_k^- = \mathscr{S}_{m_{k-1}^f, P_{k-1}^f}(f^M)$.

Remark 3.12. *Sigma-point approximations of the form $\mathscr{S}_{m,P}(z)$ are weighted summations of z evaluated at integration nodes that are determined by m, P, and their quadrature rules. For details of these, see, for example, Särkkä (2013).*

We show the stability of sigma-point TME Gaussian filters in the sense that

$$\sup_{k\geq 1} \mathbb{E}\left[\left\|X_k - m_k^f\right\|_2^2\right] < \infty. \tag{3.42}$$

Remark 3.13. *Note that if $m_k^f = \mathbb{E}[X_k \mid Y_{1:k}]$ is obtained exactly, then the mean-square in the equation above is minimised, since $\mathbb{E}[X_k \mid Y_{1:k}]$ is an orthogonal projection of X_k. But in practice, one can only hope for approximating $\mathbb{E}[X_k \mid Y_{1:k}]$ by using, for example, TME Gaussian filters. The stability analysis here is devoted to show that the TME Gaussian filtering error has a finite (contractive) bound that depends on step k.*

We use the following assumptions.

Assumption 3.14. *There exist constants $c_M \geq 0$, $c_{\check{q}} \geq 0$, and $c_P \geq 0$ such that $\sup_{k\geq 1}\|R_M(X_{k-1})\|_2 \leq c_M$ $\mathbb{P}$-almost surely, $\sup_{k\geq 1}\mathbb{E}\left[\operatorname{tr}\left(\check{Q}(X_{k-1})\right)\right] \leq c_{\check{q}}$, and $\sup_{k\geq 1}\mathbb{E}\left[\operatorname{tr}(P_k^f)\right] \leq c_P$.*

Assumption 3.15. *There exists $c_{\mathscr{S}} \geq 0$ such that*

$$\left\|f^M(x) - \mathscr{S}_{m,p}(f^M)\right\|_2^2 \leq \left\|\mathrm{J}_x f^M(x)\right\|_2^2 \|x-m\|_2^2 + c_{\mathscr{S}} \operatorname{tr}(P), \tag{3.43}$$

for all $x \in \mathbb{R}^d$, $m \in \mathbb{R}^d$, and positive semi-definite matrix $P \in \mathbb{R}^{d\times d}$.

Assumption 3.16. *There exists $c_K \geq 0$ such that $\sup_{k\geq 1}\|I - K_k H\|_2 \leq c_K$ $\mathbb{P}$-almost surely, and*

$$c_f^2 \coloneqq c_K^2 \sup_{x\in\mathbb{R}^d}\left\|\mathrm{J}_x f^M(x)\right\|_2^2 < \frac{1}{4}. \tag{3.44}$$

Indeed, the assumptions above are in some sense restrictive. In particular, the TME remainder and the covariance of the transition density are required to be bounded by c_M and $c_{\check{q}}$, respectively. In order to satisfy these assumptions, it is sufficient to require that the SDE coefficients are smooth enough and all their derivatives up to a certain order are uniformly bounded (e.g., the Beneš model in Example 3.9). For more detailed explanations of these assumptions can be found in Zhao et al. (2021b) and Karvonen et al. (2020).

The main result is shown in the following theorem.

Theorem 3.17 (TME Gaussian filter stability)**.** *Suppose that Assumptions 3.14 to 3.16 hold. Then the sigma-point TME Gaussian filter for system* (3.41) *is such that*

$$\begin{aligned}\mathbb{E}\left[\left\|X_k - m_k^f\right\|_2^2\right] &\leq \left(4c_f^2\right)^k \operatorname{tr}(P_0) \\ &\quad + \frac{4\left(c_K^2\left(c_{\mathscr{S}} c_P + c_M^2 + c_{\check{q}}\right) + \operatorname{tr}(R) c_P^2 \|H\|_2^2 \|R^{-1}\|_2^2\right)}{1 - 4c_f^2}.\end{aligned} \tag{3.45}$$

Proof. Define $Z_k \coloneqq I - K_k H$. By substituting the sigma-point TME Gaussian filtering steps and the model (3.41) in $X_k - m_k^f$, we get

$$\begin{aligned} X_k - m_k^f &= \check{f}(X_{k-1}) + \check{q}(X_{k-1}) - m_k^- - K_k (Y_k - H m_k^-), \\ &= Z_k \left(\check{f}(X_{k-1}) - \mathscr{S}_{m_{k-1}^f, P_{k-1}^f}(f^M) + \check{q}(X_{k-1}) \right) - K_k \xi_k \\ &= Z_k \left(f^M(X_{k-1}) - \mathscr{S}_{m_{k-1}^f, P_{k-1}^f}(f^M) \right) \\ &\quad + Z_k R_M(X_{k-1}) + Z_k \check{q}(X_{k-1}) - K_k \xi_k. \end{aligned} \tag{3.46}$$

Then

$$\begin{aligned} \mathbb{E}\left[\left\| X_k - m_k^f \right\|_2^2 \right] \leq 4\, \mathbb{E}\left[\left\| Z_k \left(f^M(X_{k-1}) - \mathscr{S}_{m_{k-1}^f, P_{k-1}^f}(f^M) \right) \right\|_2^2 \right] \\ + 4 \left(\mathbb{E}\left[\| Z_k R_M(X_{k-1}) \|_2^2 \right] + \mathbb{E}\left[\| Z_k \check{q}(X_{k-1}) \|_2^2 \right] + \mathbb{E}\left[\| K_k \xi_k \|_2^2 \right] \right). \end{aligned} \tag{3.47}$$

Now, by substituting the bounds

$$\begin{aligned} \mathbb{E}\left[\left\| Z_k \left(f^M(X_{k-1}) - \mathscr{S}_{m_{k-1}^f, P_{k-1}^f}(f^M) \right) \right\|_2^2 \right] &\leq c_f^2\, \mathbb{E}\left[\left\| X_{k-1} - m_{k-1}^f \right\|_2^2 \right] \\ &\quad + c_K^2\, c_{\mathscr{S}}\, c_P, \\ \mathbb{E}\left[\| Z_k R_M(X_{k-1}) \|_2^2 \right] &\leq c_K^2\, c_M^2, \\ \mathbb{E}\left[\| Z_k \check{q}(X_{k-1}) \|_2^2 \right] &\leq c_K^2\, c_{\check{q}}, \\ \mathbb{E}\left[\| K_k \xi_k \|_2^2 \right] &\leq \operatorname{tr}(R)\, c_P^2\, \|H\|_2^2\, \|R^{-1}\|_2^2, \end{aligned} \tag{3.48}$$

following Assumptions 3.14, 3.15, and 3.16 into Equation (3.47), we obtain the recursive inequality

$$\begin{aligned} \mathbb{E}\left[\left\| X_k - m_k^f \right\|_2^2 \right] \leq 4\, c_f^2\, \mathbb{E}\left[\left\| X_{k-1} - m_{k-1}^f \right\|_2^2 \right] \\ + 4 \left(c_K^2 \left(c_{\mathscr{S}}\, c_P + c_M^2 + c_{\check{q}} \right) + \operatorname{tr}(R)\, c_P^2\, \|H\|_2^2\, \|R^{-1}\|_2^2 \right). \end{aligned} \tag{3.49}$$

The assumption $4 c_f^2 < 1$ concludes the bound in Equation (3.45). □

Stability analysis of Gaussian smoothers that use the TME method can be found in Zhao and Särkkä (2021).

3.6.2 Signal estimation and target tracking examples

This section presents a few applications of TME Gaussian filters and smoothers on signal estimation and target tracking problems. In the examples below, we uniformly use the expansion order $M = 3$, and we use the Gauss–Hermite quadrature method (of order 3) to approximate the Gaussian expectations in Equations (3.36) and (3.37).

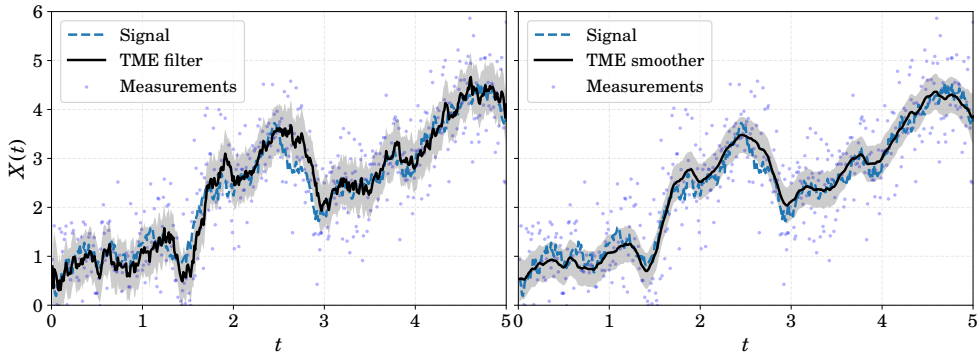


Figure 3.3. TME Gaussian filtering and smoothing for the Beneš model in Example 3.18. Shaded area stands for 0.95 confidence interval.

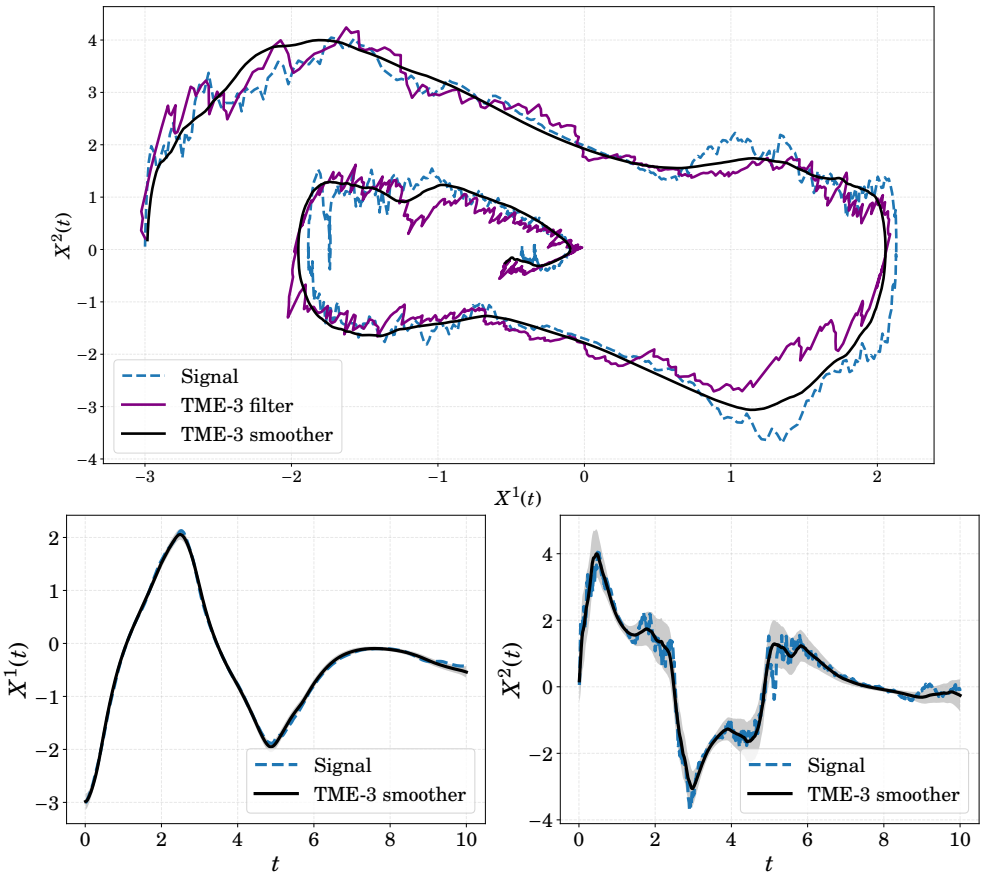


Figure 3.4. TME Gaussian filtering and smoothing for the Duffing–van der Pol model in Example 3.19.

Example 3.18 (Beneš)**.** *Consider the Beneš model in Example 3.9, and also consider a linear measurement model* $Y_k = X(t_k) + \xi_k$*, where* $\xi_k \sim \mathrm{N}(0, 0.5)$*. We simulate a pair of a signal and its measurements at times* $\{t_k = 0.01k : k = 0, 1, \ldots, 500\}$*, then we apply the TME-3 Gaussian filter and smoother to estimate the signal from the measurements. The results are plotted in Figure 3.3.*

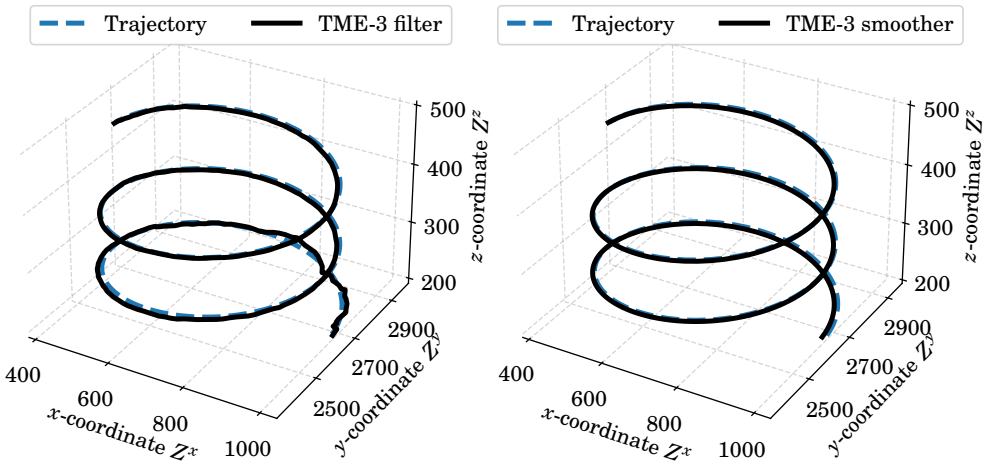


Figure 3.5. TME Gaussian filtering and smoothing for tracking a target moving as per the 3D coordinated turn model in Example 3.20.

Example 3.19 (Duffing–van der Pol)**.** *Consider a continuous-discrete state-space model*

$$\begin{aligned} \mathrm{d}X^1(t) &= X^2(t)\,\mathrm{d}t, \\ \mathrm{d}X^2(t) &= \left(X^1(t)\left(\kappa - \left(X^1(t)\right)^2\right) - X^2(t)\right)\mathrm{d}t + X^1(t)\,\mathrm{d}W(t), \\ Y_k &= X^1(t_k) + 0.1\,X^2(t_k) + \xi_k, \end{aligned} \tag{3.50}$$

starting from the initial values $X^1(t_0) = -3$ *and* $X^2(t_0) = 0$*, where* $\kappa = 2$ *and* $\xi_k \sim \mathrm{N}(0, 0.1)$*. The non-linear multiplicative SDE above is called a modified stochastic Duffing–van der Pol oscillator equation (Lord et al., 2014; Särkkä and Solin, 2019). We simulate a pair of a signal and its measurements at times* $\{t_k = 0.01k : k = 0, 1, \ldots, 1000\}$*. The results of the TME-3 Gaussian filtering and smoothing for this model is illustrated in Figure 3.4.*

It is worth mentioning that the Euler–Maruyama-based Gaussian smoothing methods on this model may encounter numerical problems because the Euler–Maruyama scheme gives singular covariance approximation.

Example 3.20 (3D coordinated turn tracking)**.** *Consider a continuous-discrete model*

$$\begin{aligned} \mathrm{d}Z(t) &= a^{\mathrm{CT}}(Z(t))\,\mathrm{d}t + b^{\mathrm{CT}}\,\mathrm{d}W(t), \\ Y_k &= h^{\mathrm{CT}}(Z(t_k)) + \xi_k, \end{aligned} \tag{3.51}$$

where the state $Z : \mathbb{T} \to \mathbb{R}^7 := \begin{bmatrix} Z^x(t) & \dot{Z}^x(t) & Z^y(t) & \dot{Z}^y(t) & Z^z(t) & \dot{Z}^z(t) & \vartheta(t) \end{bmatrix}^{\mathsf{T}}$ *stands for the 3D Cartesian coordinate and the turn rate of a target. The*

SDE coefficients and the measurement function are defined by

$$
\begin{aligned}
a^{\mathrm{CT}}(Z(t)) &= \begin{bmatrix} \dot{Z}^x(t) & -\vartheta(t)\dot{Z}^y(t) & \dot{Z}^y(t) & \vartheta(t)\dot{Z}^x(t) & \dot{Z}^x(t) & 0 & 0 \end{bmatrix}^{\mathsf{T}}, \\
h^{\mathrm{CT}}(Z(t_k)) &= \begin{bmatrix} \sqrt{(Z^x(t_k))^2 + (Z^y(t_k))^2 + (Z^z(t_k))^2} \\ \arctan(Z^y(t_k)/Z^x(t_k)) \\ \arctan\left(Z^z(t_k)/\sqrt{(Z^x(t_k))^2 + (Z^y(t_k))^2}\right) \end{bmatrix}.
\end{aligned}
\tag{3.52}
$$

For details of this model, we refer the reader to Zhao et al. (2021b). This model is widely used for manoeuvring target tracking and is very challenging for filtering and smoothing algorithms due to its high dimensionality and non-linearity (Arasaratnam et al., 2010; Bar-Shalom et al., 2002). A tracking example by using the TME-3 Gaussian filter and smoother is shown in Figure 3.5.

4. State-space deep Gaussian processes

In this chapter we introduce state-space deep Gaussian processes (SS-DGPs). The chapter starts with a brief review on Gaussian processes (GPs) and their state-space representations in Sections 4.1 and 4.2, respectively. Subsequently, in Section 4.3 deep Gaussian processes and their state-space representations (i.e., SS-DGPs) are defined. In Section 4.6, we introduce deep Matérn processes which are a subclass of SS-DGPs where each GP element in the SS-DGP hierarchy is conditionally a Matérn GP. Section 4.7 represents the SS-DGP regression problems as continuous-discrete filtering and smoothing problems. Finally, Section 4.9 illustrates how to solve L^1-regularised SS-DGP regression problems.

The content of this chapter is based on Publications II and VII.

4.1 Gaussian processes

Gaussian processes (GPs) are a class of stochastic processes with finite-dimensional Gaussian distributions. More precisely, an $\mathbb{R}^d$-valued stochastic process $U\colon \mathbb{T} \to \mathbb{R}^d$ is said to be a GP if the following definition is satisfied.

Definition 4.1 (Gaussian process)**.** *A stochastic process $U\colon \mathbb{T} \to \mathbb{R}^d$ on some probability space is called a Gaussian process on $\mathbb{T}$ if for every integer $k \geq 0$ and real numbers $t_1 < t_2 < \cdots < t_k \in \mathbb{T}$, the random variables $U(t_1), U(t_2), \ldots, U(t_k)$ are jointly Gaussian (see, e.g., Karatzas and Shreve, 1991, Section 2.9).*

Remark 4.2. *In the spirit of this thesis, we restrict Definition 4.1 to temporal GPs only, however, it is possible to define GPs on more general domains (Rasmussen and Williams, 2006).*

Since multivariate normal distributions are entirely determined by their means and covariances, Definition 4.1 is usually interpreted by the shorthand notation

$$U(t) \sim \mathrm{GP}\big(m(t), C(t,t')\big), \tag{4.1}$$

where $m\colon \mathbb{T} \to \mathbb{R}^d$ and $C\colon \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{d\times d}$ stand for the mean and covariance functions of the process, respectively. Under this notation, the finite-dimensional probability density function of U at time instances $t_1, t_2, \ldots, t_k \in \mathbb{T}$ is given by

$$
\begin{aligned}
&p_{U(t_1),U(t_2),\ldots,U(t_k)}(u_1, u_2, \ldots, u_k) \\
&= \mathrm{N}\left(\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix} \,\middle|\, \begin{bmatrix} m(t_1) \\ m(t_2) \\ \vdots \\ m(t_k) \end{bmatrix}, \begin{bmatrix} C(t_1,t_1) & C(t_1,t_2) & \cdots & C(t_1,t_k) \\ C(t_2,t_1) & C(t_2,t_2) & \cdots & C(t_2,t_k) \\ \vdots & \vdots & \ddots & \vdots \\ C(t_k,t_1) & C(t_k,t_2) & \cdots & C(t_k,t_k) \end{bmatrix}\right).
\end{aligned} \tag{4.2}
$$

There are numerous possible choices for the covariance function C, and researchers and practitioners can choose one or the other depending on their applications. One of the most popular family of covariance functions to model continuous functions with varying degrees of regularity is given by the Whittle–Matérn covariance function (Matérn, 1960)

$$
C_{\text{Mat.}}(t,t') = \frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,|t-t'|}{\ell}\right)^{\nu} \mathrm{K}_{\nu}\left(\frac{\sqrt{2\nu}\,|t-t'|}{\ell}\right), \tag{4.3}
$$

where ℓ and σ are scale parameters, Γ is the Gamma function, K_{ν} is the modified Bessel function of the second kind, and $\nu \in \left\{\frac{1}{2}, \frac{3}{2}, \ldots\right\}$. The smoothness of U is controlled by the value of ν. For example, if $\nu = \frac{3}{2}$, then $t \mapsto U(t)$ will be differentiable almost surely.

Without loss of generality, we assume from now on that $m(t) = 0$ for all $t \in \mathbb{T}$, that is

$$
U(t) \sim \mathrm{GP}\big(0, C(t,t')\big). \tag{4.4}
$$

The covariance function C thus entirely determines the properties of U, such as its continuity and stationarity.

Remark 4.3. *A stochastic process U is said to be stationary if its finite-dimensional distribution is invariant under translation. That is,*

$$
p_{U(t_1+\tau),\ldots,U(t_k+\tau)}(u_1, \ldots, u_k) = p_{U(t_1),\ldots,U(t_k)}(u_1, \ldots, u_k),
$$

for all $k \geq 1$, $t_1 < \cdots < t_k \in \mathbb{T}$, and $t_1 + \tau < \cdots < t_k + \tau \in \mathbb{T}$ (Karatzas and Shreve, 1991). Since GPs are characterised by their mean and covariance functions, we say that a zero-mean GP is stationary if $C(t+\tau, t'+\tau)$ does not depend on τ, or equivalently, $C(t,t')$ is only a function of the time difference $t - t'$.

Stationarity is an important concept to keep in mind as many widely used covariance functions, such as the Matérn family and the radial basis function (RBF) lead to stationary GPs. However, as mentioned in Introduction, these stationary GPs might not be suitable priors in a number of applications.

Batch GP regression

Consider a GP regression model

$$\begin{aligned} U(t) &\sim \mathrm{GP}\big(0, C(t,t')\big), \\ Y_k &= U(t_k) + \xi_k, \quad \xi_k \sim \mathrm{N}(0, \Xi_k), \end{aligned} \tag{4.5}$$

where we have a set of measurement data $y_{1:T} = \{y_k : k = 1,2,\ldots,T\}$ at times $t_1, t_2, \ldots, t_T \in \mathbb{T}$. Let us denote by $C_{1:T}$ the (Gram) matrix obtained by evaluating the covariance function C on the Cartesian grid $(t_1, t_2, \ldots, t_T) \times (t_1, t_2, \ldots, t_T)$. Let us also define $\Xi_{1:T} := \mathrm{diag}(\Xi_1, \Xi_2, \ldots, \Xi_T)$ and $U_{1:T} := \big\{U(t_1), U(t_2), \ldots, U(t_T)\big\}$.

Using Bayes' rule, and Gaussian identities, one can prove that the joint batch posterior probability density $p_{U_{1:T} \mid Y_{1:T}}(u_{1:T} \mid y_{1:T})$ is Gaussian. More specifically, the mean and covariance of the batch posterior density are given by

$$\mathbb{E}[U_{1:T} \mid y_{1:T}] = C_{1:T}(C_{1:T} + \Xi_{1:T})^{-1} y_{1:T} \tag{4.6}$$

and

$$\mathrm{Cov}[U_{1:T} \mid y_{1:T}] = C_{1:T} - C_{1:T}(C_{1:T} + \Xi_{1:T})^{-1} C_{1:T}, \tag{4.7}$$

respectively. With a slight modification of the two equations above, the mean and covariance of the posterior density at test points (i.e., interpolation/extrapolation) can also be obtained in closed-form (see, e.g., Rasmussen and Williams, 2006, Section 2.2).

Remark 4.4. *The batch term in the name comes from the fact that the posterior density is solved jointly at $t_1, t_2, \ldots, t_T$ by using the full covariance matrix $C_{1:T}$.*

In Figure 1.1, we illustrate two examples of this batch GP regression using a Matérn $\nu = 3/2$ covariance function of the form in Equation (4.3).

It is worth pointing out two numerical problems of batch GP regressions. First, the computational complexity for computing the posterior mean and covariance is $O(T^3)$. This is due to the necessity of solving a system of equations of size T. This makes standard GP regression computationally expensive for large-scale datasets. This prompted researchers to introduce a number of alternatives (e.g., sparse GPs) that alleviate this prohibitive complexity. We refer the reader to Section 1.1 for a short review on this topic.

Another problem is that if the data times $t_1, t_2, \ldots, t_T$ are densely located (i.e., $t_k - t_{k-1}$ is numerically small for $k = 1,2,\ldots,T$), or when some of them are identical, then the covariance matrix $C_{1:T}$ might be numerically close to singular (see, e.g., Ababou et al., 1994; Ranjan et al., 2011). This numerical problem does not in general affect the numerical computation of Equations (4.6) and (4.7), as the minimum eigenvalue of $C_{1:T} + \Xi_{1:T}$ is greater than the minimum eigenvalue of $\Xi_{1:T}$. However, it affects any

procedure that needs to compute the matrix inverse of $C_{1:T}$ (e.g., maximum a posterior estimate of GP regression), or that the GP is observed without measurement noises (Ranjan et al., 2011). It may also affect making samples from GP by means of Cholesky decomposition of $C_{1:T}$.

State-space representations of GPs, as formulated in the following section, can be used to avoid the two problems above.

4.2 State-space Gaussian processes

In this section, we introduce state-space representations of GPs. Namely, we represent GPs as solutions of linear SDEs. In order to do this, let $U\colon \mathbb{T} \to \mathbb{R}^d$ be the solution of a linear SDE

$$\begin{aligned} \mathrm{d}U(t) &= A(t)U(t)\,\mathrm{d}t + B(t)\,\mathrm{d}W(t), \\ U(t_0) &= U_0, \end{aligned} \tag{4.8}$$

where coefficients $A\colon \mathbb{T} \to \mathbb{R}^{d\times d}$ and $B\colon \mathbb{T} \to \mathbb{R}^{d\times w}$ are deterministic time-dependent functions, $W\colon \mathbb{T} \to \mathbb{R}^w$ is a Wiener process, and $U_0 \sim \mathrm{N}(m_0, P_0)$. For the sake of simplicity, let us from now on assume that these coefficients are regular enough so that the SDE above is well-defined (see, e.g., Theorem 2.18 for sufficient conditions).

It turns out that the solution U of the SDE in Equation (4.8) verifies the axioms of Gaussian processes (given in Definition 4.1) on $\mathbb{T}$ (see, Karatzas and Shreve, 1991, Section 5.6). Moreover, its mean $t \mapsto \mathbb{E}[U(t)]$ and covariance $t \mapsto \mathrm{Cov}[U(t)]$ functions are solutions of the following linear ODEs

$$\begin{aligned} \frac{\mathrm{d}m(t)}{\mathrm{d}t} &= A(t)\,m(t), \\ \frac{\mathrm{d}P(t)}{\mathrm{d}t} &= A(t)P(t) + P(t)A(t)^\mathsf{T} + B(t)B(t)^\mathsf{T}, \end{aligned} \tag{4.9}$$

for every $t \in \mathbb{T}$ starting from the initial values $m(t_0) = m_0$ and $P(t_0) = P_0$. Note that if the initial mean $m_0 = 0$ then $m(t) = 0$ for all $t \in \mathbb{T}$, so that U will be a zero-mean GP.

Compared to the batch GP representation in Equation (4.1), state-space representations do not need to explicitly specify their mean and covariance functions. These functions are instead implicitly defined by the SDE coefficients. Finding the state-space representation of a GP with desired covariance function is possible as well (see, e.g., Hartikainen and Särkkä, 2010; Särkkä et al., 2013; Solin, 2016).

Suppose that the coefficients $A(t) = A$ and $B(t) = B$ are constants, and all the real parts of the eigenvalues of A are negative. Let $m_0 = 0$, and let P_0 solve the Lyapunov equation

$$AP + PA^\mathsf{T} + BB^\mathsf{T} = 0, \tag{4.10}$$

then

$$U(t) \sim \mathrm{GP}\big(0, \mathrm{Cov}[U(t), U(t')]\big)$$

is a zero-mean stationary GP, and its covariance function is given by

$$\mathrm{Cov}[U(t), U(t')] = \begin{cases} P_0\, e^{|t-t'|A^\mathsf{T}}, & t < t' \in \mathbb{T}, \\ e^{|t-t'|A}\, P_0, & t' \leq t \in \mathbb{T}. \end{cases}$$

See, for example, Karatzas and Shreve (1991, Theorem 6.7), Pavliotis (2014, Section 3.7), or Särkkä and Solin (2019, Section 6.5) for details.

State-space GP regression

Due to the fact that state-space GPs (SS-GPs) are solutions of SDEs, they verify the Markov property. This is key in allowing to perform GP regression sequentially for $k = 1, 2, \ldots, T$ without computing the full covariance matrix $C_{1:T}$. To see this, let us consider a GP regression problem in the state-space form

$$\begin{aligned} \mathrm{d}U(t) &= A(t)U(t)\,\mathrm{d}t + B(t)\,\mathrm{d}W(t), \quad U(t_0) = U_0, \\ Y_k &= H_k\, U(t_k) + \xi_k, \quad \xi_k \sim \mathrm{N}(0, \Xi_k). \end{aligned} \tag{4.11}$$

We aim to compute the posterior density $p_{U(t_k)\,|\,Y_{1:T}}(u_k \mid y_{1:T})$ for $k = 1, 2, \ldots, T$ instead of the joint posterior density $p_{U_{1:T}\,|\,Y_{1:T}}(u_{1:T} \mid y_{1:T})$. This state-space GP regression problem is equivalent to the continuous-discrete smoothing problem in Section 2.2.2 (Särkkä and Solin, 2019). Therefore, one can apply Kalman filters and RTS smoothers (see, Algorithm 2.10) to carry out the state-space GP regression at hand exactly. Figure 4.1 illustrates an example showing the equivalence between batch and state-space GP regression on a toy model.

The computational complexity of state-space GP regression is $O(T)$, whereas the batch GP regression is $O(T^3)$. As an example, the batch and state-space GP regression shown in Figure 4.1 take around 37 s and 0.1 s, respectively, on a computer with $T = 10{,}000$ measurements. Furthermore, by using prefix-sum algorithms, state-space GP regression can be solved in logarithmic $O(\log(T))$ time (Corenflos et al., 2021b; Särkkä and García-Fernández, 2021).

It is worth mentioning that not all GPs are Markov processes, hence, not all GPs have analytical state-space representations. As an example, Rozanov (1977, 1982) show that certain stationary Gaussian processes/fields are Markovian if and only if the reciprocal of their spectral densities are polynomials. For instance, GPs using the RBF covariance function are not Markovian, but it is possible to approximate them up to an arbitrary order by using their approximate state-space representations (Särkkä et al., 2013).

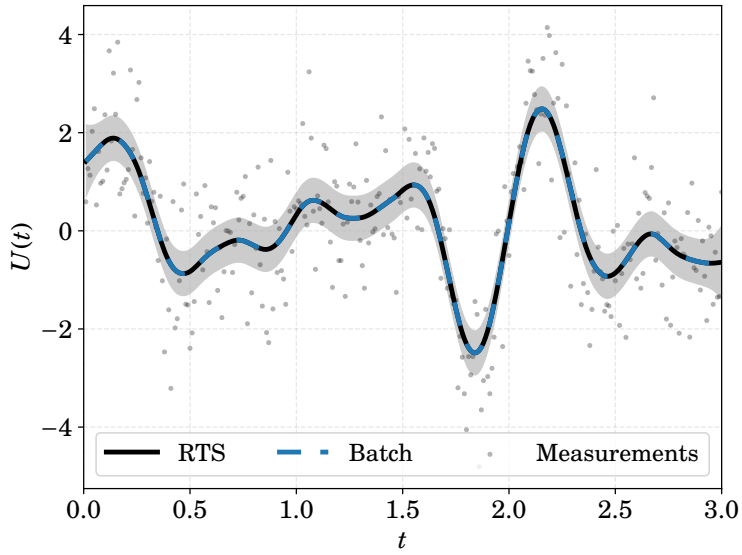


Figure 4.1. Batch and state-space GP regression on a toymodel with a Matérn $\nu = 3/2$ covariance function and zero mean function. These two regression methods recover the same posterior densities (the lines and shaded area stand for the posterior mean and 0.95 confidence interval, respectively).

4.3 State-space deep Gaussian processes (SS-DGPs)

State-space deep Gaussian processes (SS-DGPs) are stochastic processes that parametrise multiple conditional GPs hierarchically. This hierarchical construction makes SS-DGPs suitable priors for modelling irregular function in many applications. To see this, let us first consider a GP

$$U(t) \sim \mathrm{GP}\big(0, C(t,t';\ell(t))\big),$$

where the covariance function $C(t,t';\ell(t))$ has an unknown (random) parameter $\ell(t) \in \mathbb{R}_{>0}$ (i.e., a time-varying length scale). When the parameter $\ell(t)$ does not depend on t, it can be assigned by human experts or automatically learnt from data by, for example, maximum likelihood estimation (MLE), maximum a posteriori (MAP), variational inference, or Markov chain Monte Carlo (MCMC) (Rasmussen and Williams, 2006). However, the assumption that ℓ being independent of t might not be reasonable for a number of applications that exhibit time-varying features. A way to mitigate this issue is, for example, to consider putting another GP prior on the length scale parameter, that is

$$\ell(t) \sim \mathrm{GP}\big(0, C(t,t';\ell_2)\big),$$

where ℓ_2 is another length scale parameter. This hierarchical feature is meaningful in the sense that it allows the characteristics of U to change over time, since its length scale $t \mapsto \ell(t)$ now is a stochastic process of t. It is then of interest to ask if this hierarchical recursion can be continued up

to a given depth L:

$$\begin{aligned}
\ell_2(t) &\sim \mathrm{GP}\big(0, C(t,t';\ell_3(t))\big),\\
\ell_3(t) &\sim \mathrm{GP}\big(0, C(t,t';\ell_4(t))\big),\\
&\vdots\\
\ell_L(t) &\sim \mathrm{GP}\big(0, C(t,t';\ell_{L+1})\big),
\end{aligned} \tag{4.12}$$

where the final leaf ℓ_{L+1} is a constant. This construction leads to a class of deep Gaussian processes (DGPs, see, Section 1.1 for background).

In the rest of this chapter, we formulate the hierarchy in Equation (4.12) in more abstract form in order to define DGPs. Thereupon we leverage this definition to represent DGPs as solutions of SDEs in order to arrive at SS-DGPs.

Deep Gaussian processes

Equation (4.12) exemplifies a DGP where the length scale parameters *only* are considered as GPs. In graph theory, this type of DGP hierarchy corresponds to a path graph (Gross et al., 2018) where the length scale parameters are vertices that ordered in a line/path. This type of DGP construction is the most studied case in the parametrisation-based DGP community (Roininen et al., 2019; Salimbeni and Deisenroth, 2017b; Emzir et al., 2020).

However, in principle, a GP can take any number of parameters. Thus, in order to abstract DGPs, we need to think of a DGP as a joint process defined over a set of conditional GPs. These conditional GPs are not necessarily limited to representing length scale parameters only. In order to do so, we need to introduce an indexing system and a few notations. Let $U_j^i \colon \mathbb{T} \to \mathbb{R}^{d_i}$ denote a GP indexed by an integer $i \in \mathbb{N}$. This superscript i means that U_j^i is the i-th GP element in a (yet to be defined) collection of GPs. The subscript j in U_j^i means that the GP U_j^i is a parent of the j-th GP element (i.e., the j-th GP element is parametrised by the i-th element). The terminology "parent" follows from probabilistic graph model conventions (Koller and Friedman, 2009). The fact that GP element does not have any child means that it does not parametrise any other GP therefore, we define its subscript j to be $j = 0$. This is always true for the first element U_0^1 as we shall see later in the definition of the collection of these GP elements.

Additionally, in order to give a well-defined graph, we restrict $j < i$ so that a GP element can only parametrise *one* of its *preceding* elements. This implies that a GP element can have multiple parents but no more than one child. Without this restriction, one might have two elements, for instance, U_1^2 and U_2^1 depending on each other, that is not within the scope of this thesis.

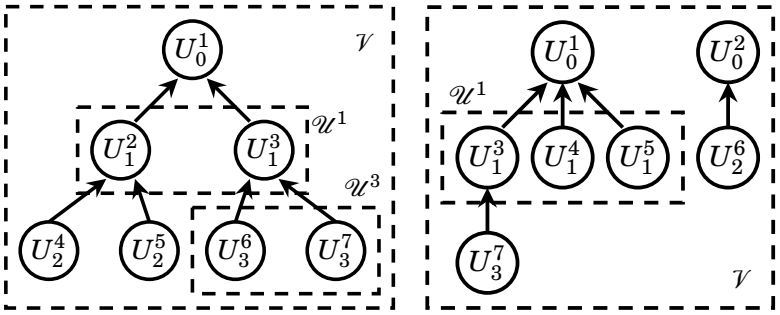


Figure 4.2. Two DGP ($L = 7$) examples in graph illustration.

Remark 4.5. *The set of dependencies between the conditional GPs can be thought of as a collection of directed trees where the head of each tree has j subscript $j = 0$, and the notation U_j^i implies that there is an edge pointing from U_j^i to U_k^j for some k. See, Figure 4.2 for an illustration.*

Suppose that we have L GPs $U_0^1, U_{j_2}^2, \ldots, U_{j_L}^L$ and a set $J = \{j_i \in \mathbb{N}: i = 1,2\ldots,L,\ 0 \le j_i < i\}$ that describes the conditional dependencies of these GPs. We define the collection of all these GPs as

$$\mathcal{V} := \mathcal{V}_J^L = \left\{U_{j_i}^i : i = 1,2,\ldots,L,\ j_i \in J\right\}, \tag{4.13}$$

and we will call these conditional GPs the GP elements of $\mathcal{V}$. Based on this collection, we define a DGP as a vector-valued process composed of all the GP elements in $\mathcal{V}$.

Definition 4.6 (Deep Gaussian process)**.** *Let $\mathcal{V}$ be a collection of L $\mathbb{R}^{d_i}$-valued conditional GPs defined by Equation* (4.13)*. An $\mathbb{R}^{\sum_{i=1}^L d_i}$-valued stochastic process $V : \mathbb{T} \to \mathbb{R}^{\sum_{i=1}^L d_i}$ is said to be a deep Gaussian process on $\mathbb{T}$ with respect to $\mathcal{V}$ if V is a permutation of all the elements of $\mathcal{V}$.*

Remark 4.7. *Note that in the special case $L = 1$, a DGP reduces to a standard GP.*

It is also natural to define another set

$$\mathcal{U}^i = \left\{U_i^k \in \mathcal{V} : k = 1,2,\ldots,L\right\} \tag{4.14}$$

that collects all the parent GPs of the i-th GP element in $\mathcal{V}$. It follows from Lemma 4.8 that all the collections of parent GPs form a partition of the set of all GP elements.

Lemma 4.8 (Partition)**.** *Let $\mathcal{U}^0, \mathcal{U}^1, \ldots, \mathcal{U}^L$ be collections of parent GPs as defined by Equation* (4.14)*. These collections satisfy the axiom of a partition.*

I. (Pairwise disjointness) For every $m, n \in \{0,1,\ldots,L-1\}$ and $m \ne n$,

$$\mathcal{U}^m \cap \mathcal{U}^n = \emptyset. \tag{4.15}$$

II. (Exhaustiveness)

$$\bigcup_{i=0}^{L-1} \mathscr{U}^i = \mathscr{V}. \tag{4.16}$$

Remark 4.9. *Note that $\mathscr{U}^L = \emptyset$ by construction.*

Proof. In order to prove the first property, suppose that there exists a pair $m, n \in \{0, 1, \ldots, L-1\}$ and $m \neq n$ such that $\mathscr{U}^m \cap \mathscr{U}^n$ is non-empty. This implies that there is a GP element pointing simultaneously to $U_{j_m}^m$ and to $U_{j_n}^n$, which violates the definition of a GP element.

Following Equations (4.13) and (4.14), we have $\bigcup_{i=0}^{L-1} \mathscr{U}^i \subseteq \mathscr{V}$. Suppose that there exists a GP element that is in $\mathscr{V}$ but not in $\bigcup_{i=0}^{L-1} \mathscr{U}^i$, then this GP element is not a parent of any GP elements (i.e., it must be in $\mathscr{U}^0$) which violates the hypothesis. □

We mention that the indexing system for DGPs here is simplified compared to Publication II which additionally used an unnecessary index denoting the depth of the GP element in the hierarchy. Figure 4.2 illustrates two graphical examples of DGPs to clarify the indexing and notations used here.

Batch representations of DGPs

Following Definition 4.6, we can use the following shorthand batch notation to represent a DGP $V : \mathbb{T} \to \mathbb{R}^{\sum_{i=1}^{L} d_i}$ with L conditional GPs:

$$\begin{aligned}
U_0^1(t) \,\big|\, \mathscr{U}^1 &\sim \mathrm{GP}\big(0, C^1(t, t'; \mathscr{U}^1)\big), \\
U_{j_2}^2(t) \,\big|\, \mathscr{U}^2 &\sim \mathrm{GP}\big(0, C^2(t, t'; \mathscr{U}^2)\big), \\
&\vdots \\
U_{j_i}^i(t) \,\big|\, \mathscr{U}^i &\sim \mathrm{GP}\big(0, C^i(t, t'; \mathscr{U}^i)\big), \\
&\vdots \\
U_{j_L}^L(t) &\sim \mathrm{GP}\big(0, C^L(t, t')\big),
\end{aligned} \tag{4.17}$$

where $C^i : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{d_i \times d_i}$ is the covariance function of $U_{j_i}^i$ parametrised by the GPs in $\mathscr{U}^i$, and

$$V(t) := \begin{bmatrix} U_0^1(t) & U_{j_2}^2(t) & \cdots & U_{j_L}^L(t) \end{bmatrix}^\mathsf{T}.$$

Thanks to the conditional hierarchy structure of the model, the probability density function

$$\begin{aligned} p_{V(t)}(v,t) &\coloneqq p_{U_0^1(t),\ldots,U_{j_L}^L(t)}\left(u_0^1,\ldots,u_{j_L}^L,t\right) \\ &= \prod_{i=1}^{L} p_{U_{j_i}^i(t)\,|\,\mathscr{U}^i}\left(u_{j_i}^i,t\,|\,\mathscr{U}^i\right) \end{aligned} \tag{4.18}$$

of V can factorise over the probability densities of the GP elements $p_{U_{j_i}^i(t)\,|\,\mathscr{U}^i}\left(u_{j_i}^i,t\,|\,\mathscr{U}^i\right)$ for $i=1,\ldots,L$. Notice that for the sake of readability, we slightly abused the notation in Equation (4.18), in the sense that $\mathscr{U}^i$, appearing in the argument of $p_{U_{j_i}^i(t)\,|\,\mathscr{U}^i}\left(u_{j_i}^i,t\,|\,\mathscr{U}^i\right)$, actually stands for the realisation of all the GPs contained in $\mathscr{U}^i$.

In order for the DGP V represented by Equation (4.17) to be well-defined, its covariance functions $C^1,\ldots,C^L$ must be chosen suitably. Many conventional covariance functions – such as the Matérn $C_{\text{Mat.}}$ in Equation (4.3) – mostly fail to be positive definite if one replaces their parameters with time dependent functions. To allow for time-varying parameters, a typical choice is to use

$$\begin{aligned} &C_{\text{NS}}(t,t';\ell,\sigma) \\ &= \frac{\sigma(t)\sigma(t')\left(\ell(t)\,\ell(t')\right)^{\frac{1}{4}}\sqrt{2}}{\Gamma(\nu)2^{\nu-1}\sqrt{\ell(t)+\ell(t')}}\left(\sqrt{\frac{8\nu(t-t')^2}{\ell(t)+\ell(t')}}\right)^{\nu}\mathrm{K}_{\nu}\left(\sqrt{\frac{8\nu(t-t')^2}{\ell(t)+\ell(t')}}\right), \end{aligned} \tag{4.19}$$

which is a non-stationary generalisation of the Matérn family by Paciorek and Schervish (2004, 2006). Gibbs (1997) introduces a similar formulation for constructing a non-stationary RBF covariance function. More non-stationary covariance function examples using time-varying parameters can also be found in, for example, Higdon et al. (1999); Snoek et al. (2014); Remes et al. (2017).

State-space representations of DGPs

Another way to represent a DGP as defined in Definition 4.6 is through the use of SDEs. The idea consists in forming a (non-linear) system of SDE representations of all the GP elements appearing in the hierarchy. More precisely, let $U_0^1, U_{j_2}^2, \ldots, U_{j_i}^i, \ldots, U_{j_L}^L$ be $\mathbb{R}^{d_i}$-valued GPs that satisfy the

following SDEs

$$\begin{aligned}
\mathrm{d}U_0^1(t) &= A^1\left(t;\mathscr{U}^1\right)U_0^1\,\mathrm{d}t + B^1\left(t;\mathscr{U}^1\right)\mathrm{d}W^1(t),\\
\mathrm{d}U_{j_2}^2(t) &= A^2\left(t;\mathscr{U}^2\right)U_{j_2}^2\,\mathrm{d}t + B^2\left(t;\mathscr{U}^2\right)\mathrm{d}W^2(t),\\
&\vdots\\
\mathrm{d}U_{j_i}^i(t) &= A^i\left(t;\mathscr{U}^i\right)U_{j_i}^i\,\mathrm{d}t + B^i\left(t;\mathscr{U}^i\right)\mathrm{d}W^i(t),\\
&\vdots\\
\mathrm{d}U_{j_L}^L(t) &= A^L(t)U_{j_L}^L\,\mathrm{d}t + B^L(t)\mathrm{d}W^L(t),
\end{aligned} \tag{4.20}$$

respectively. In Equation (4.20), $W^i\colon \mathbb{T}\to\mathbb{R}^{w_i}$ for $i=1,2,\ldots,L$ are w_i-dimensional Wiener processes, and $A^i\colon\mathbb{T}\to\mathbb{R}^{d_i\times d_i}$ and $B^i\colon\mathbb{T}\to\mathbb{R}^{d_i\times w_i}$ for $i=1,2,\ldots,L-1$ are stochastic processes that are parametrised by the GPs in $\mathscr{U}^i$. The L-th coefficients $A^L\colon\mathbb{T}\to\mathbb{R}^{d_L}$ and $B^L\colon\mathbb{T}\to\mathbb{R}^{d_L\times w_L}$, on the other hand, are deterministic, since $\mathscr{U}^L=\emptyset$ by definition. For the sake of simplicity, we collapse Equation (4.20) into a matricial form

$$\begin{aligned}
\mathrm{d}V(t) &= a(V(t))\,\mathrm{d}t + b(V(t))\,\mathrm{d}W(t),\\
V(t_0) &= V_0,
\end{aligned} \tag{4.21}$$

where $V(t)\coloneqq\begin{bmatrix}U_0^1(t) & U_{j_2}^2(t) & \cdots & U_{j_L}^L(t)\end{bmatrix}^\intercal\in\mathbb{R}^{\sum_{i=1}^L d_i}$, and the SDE coefficients are defined by

$$a(V(t))\coloneqq\begin{bmatrix}A^1\left(t;\mathscr{U}^1\right) & & & \\ & A^2\left(t;\mathscr{U}^2\right) & & \\ & & \ddots & \\ & & & A^L\left(t;\mathscr{U}^L\right)\end{bmatrix}V(t) \tag{4.22}$$

and

$$b(V(t))\coloneqq\begin{bmatrix}B^1\left(t;\mathscr{U}^1\right) & & & \\ & B^2\left(t;\mathscr{U}^2\right) & & \\ & & \ddots & \\ & & & B^L\left(t;\mathscr{U}^L\right)\end{bmatrix}. \tag{4.23}$$

The vector-valued Wiener process appearing in Equation (4.21) is similarly defined by $W(t)\coloneqq\begin{bmatrix}W^1(t) & W^2(t) & \cdots & W^L(t)\end{bmatrix}^\intercal\in\mathbb{R}^{\sum_{i=1}^L w_i}$.

A DGP $V\colon\mathbb{T}\to\mathbb{R}^{\sum_{i=1}^L d_i}$ that is characterised as per SDE (4.21) is called a state-space deep Gaussian process (SS-DGP). Compared to batch representations of DGPs, one specifies the SDE coefficients A^i and B^i for $i=1,2,\ldots,L$ and the initial condition V_0 instead of explicitly specifying the covariance functions of DGPs. In Section 4.6 we present some concrete examples of how to select these SDE coefficients so that each GP element of the SS-DGPs is conditionally a Matérn GP.

4.4 Existence and uniqueness of SS-DGPs

In the previous sections, we have defined SS-DGPs as SDE represented DGPs. However, the solution existence and uniqueness of the SDE in Equation (4.20) has still not been proven. In this section, we provide sufficient conditions on the SDE coefficients in SDE (4.20) so that the strong existence and pathwise uniqueness hold for the SDE.

In particular, one must understand that the hierarchical nature of SS-DGPs makes a direct application of Theorem 2.18 slightly unsound. Indeed, the system of SDEs (4.20) is not a linear system when seen as a multidimensional SDE. However, the individual GP elements SDEs are (conditionally on their parents in the DGP hierarchy) linear.

Theorem 4.10. *Let $W^i : \mathbb{T} \to \mathbb{R}^{w_i}$ and $U^i(t_0)$ for $i = 1,2,\dots,L$ be Wiener processes and initial random variables defined on filtered probability spaces $\left(\Omega^i, \mathscr{F}^i, \mathscr{F}_t^i, \mathbb{P}^i\right)$ for $i = 1,2,\dots,L$, where their filtrations $\{\mathscr{F}_t^i : 1,2,\dots,L\}$ are generated by their Wiener processes and initial variables. Suppose that functions A^i and B^i for $i = 1,2,\dots,L$ in Equation* (4.20) *are locally bounded measurable, then the multidimensional SDE* (4.20)*, or equivalently,* (4.21) *has a strong solution and the pathwise uniqueness holds.*

Proof. By Theorem 2.18 and the conditions of this theorem, the SDEs in Equation (4.20) are exactly the same with the integral equations

$$
\begin{aligned}
U_0^1(t) &= \Lambda^1\left(t;\mathscr{U}^1\right) U_0^1(t_0) + \Lambda^1\left(t;\mathscr{U}^1\right) \int_{t_0}^t \left(\Lambda^1\left(s;\mathscr{U}^1\right)\right)^{-1} B^1\left(s;\mathscr{U}^1\right) \mathrm{d}W^1(s), \\
U_{j_2}^2(t) &= \Lambda^2\left(t;\mathscr{U}^2\right) U_{j_2}^2(t_0) + \Lambda^2\left(t;\mathscr{U}^2\right) \int_{t_0}^t \left(\Lambda^2\left(s;\mathscr{U}^2\right)\right)^{-1} B^2\left(s;\mathscr{U}^2\right) \mathrm{d}W^2(s), \\
&\vdots \\
U_{j_i}^i(t) &= \Lambda^i\left(t;\mathscr{U}^i\right) U_{j_i}^i(t_0) + \Lambda^i\left(t;\mathscr{U}^i\right) \int_{t_0}^t \left(\Lambda^i\left(s;\mathscr{U}^i\right)\right)^{-1} B^i\left(s;\mathscr{U}^i\right) \mathrm{d}W^i(s), \\
&\vdots \\
U_{j_L}^L(t) &= \Lambda^L(t) U_{j_L}^L(t_0) + \Lambda^L(t) \int_{t_0}^t \left(\Lambda^L(s)\right)^{-1} B^L(s) \mathrm{d}W^L(s),
\end{aligned}
\tag{4.24}
$$

where Λ^i for $i = 1,2,\dots,L$ are defined as per Theorem 2.18. Hence, the joint process $V(t) \coloneqq \begin{bmatrix} U_0^1(t) & U_{j_2}^2(t) & \cdots & U_{j_L}^L(t) \end{bmatrix}^\mathsf{T}$ is an $\mathscr{F}_t$-adapted process defined on the product space $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P})$, where $\Omega = \Omega^1 \times \cdots \times \Omega^L$, $\mathscr{F}$ and $\mathscr{F}_t$ are the product sigma-algebras and filtrations (Schilling, 2017), and $\mathbb{P}(E^1 \times \cdots \times E^L) = \mathbb{P}^1(E^1) \cdots \mathbb{P}^L(E^L)$ for every $E^1 \in \Omega^1, \dots, E^L \in \Omega^L$. Noting that the other properties in Definition 2.3 are also verified, Equation (4.24) is a strong solution of the multidimensional SDE (4.20). The pathwise uniqueness of

SDE (4.20) follows from the fact that the pathwise uniqueness holds for the linear SDEs of all the GP elements (see, Zhao et al., 2021c, Lemma 7). □

The theorem above shows that in order to give a well-defined SS-DGP we only needs to ensure the SDE coefficients be locally bounded measurable functions. This condition is substantially weaker compared to the classical ones, such as the global Lipschitz and linear growth conditions (Karatzas and Shreve, 1991; Friedman, 1975; Mao, 2008; Shen et al., 2006), because we have leveraged the hierarchical nature of SS-DGP. From now on, unless otherwise specified, we will assume that this condition holds whenever we construct an SS-DGP.

Thanks to the Markov property, probability densities of SS-DGPs can factorise in the time dimension. Suppose that we have temporal instances $t_1 \leq t_2 \leq \cdots \leq t_T \in \mathbb{T}$, then the probability density function of V on these time instances reads

$$p_{V_{1:T}}(v_{1:T}) = p_{V_1}(v_1) \prod_{k=1}^{T} p_{V_{k+1} \mid V_k}(v_{k+1} \mid v_k),$$

where we denote $V_{1:T} \coloneqq \{V(t_1), V(t_2), \ldots, V(t_T)\}$. We can also factorise the probability density above in the GP element variable like in Equation (4.18) as well.

Covariance functions of SS-DGPs

The equivalence between batch and state-space DGP representations can be stated in terms of equivalence of covariance functions. In particular, conditionally on its parents in the DGP hierarchy, we can express the covariance function of a GP element as a function of its SDE coefficients.

Theorem 4.11. *Let $V(t)$ be an SS-DGP governed by the SDE in Equation* (4.21) *on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Also let $\mathscr{F}^i \subset \mathscr{F}$ be the sub-sigma-algebra generated by the GPs in $\mathscr{U}^i(t)$ for all $t \in \mathbb{T}$. Then the covariance function of the i-th GP element is*

$$\begin{aligned} C^i_{\mathrm{SS}}(t, t'; \mathscr{U}^i) &\coloneqq \operatorname{Cov}\left[U^i_{j_i}(t), U^i_{j_i}(t') \mid \mathscr{F}^i\right] \\ &= \boldsymbol{\Lambda}^i(t, t_0) \operatorname{Cov}\left[U^i_{j_i}(t_0) \mid \mathscr{U}^i(t_0)\right] \left(\boldsymbol{\Lambda}^i(t, t_0)\right)^{\mathsf{T}} \\ &\quad + \int_{t_0}^{t \wedge t'} \boldsymbol{\Lambda}^i(t, s) B^i\left(s; \mathscr{U}^i(s)\right) B^i\left(s; \mathscr{U}^i(s)\right)^{\mathsf{T}} \left(\boldsymbol{\Lambda}^i(t, s)\right)^{\mathsf{T}} \mathrm{d}s, \end{aligned} \tag{4.25}$$

where $\boldsymbol{\Lambda}^i(t, s) = \Lambda^i(t) \left(\Lambda^i(s)\right)^{-1}$ for $t, s \in \mathbb{T}$, and $\Lambda^i(t)$ is generated by A^i as per Theorem 2.18.

Proof. By Itô's formula and Theorem 2.18, we have that

$$\begin{aligned} U_{j_i}^i(t) &= \Lambda^i(t) U_{j_i}^i(t_0) + \Lambda^i(t) \int_{t_0}^t \left(\Lambda^i(s)\right)^{-1} B^i\left(s; \mathscr{U}^i(s)\right) \mathrm{d}W^i(s) \\ &:= \mathbf{\Lambda}^i(t, t_0) U_{j_i}^i(t_0) + \int_{t_0}^t \mathbf{\Lambda}^i(t,s) B^i\left(s; \mathscr{U}^i(s)\right) \mathrm{d}W^i(s) \end{aligned} \tag{4.26}$$

with respect to $\mathscr{F}^i$. Note that $\Lambda(t_0) = I$ as per Equation (2.39). Hence, by Itô isometry and by substituting $U_{j_i}^i(t)$ into

$$\begin{aligned} &\mathrm{Cov}\left[U_{j_i}^i(t), U_{j_i}^i(t') \mid \mathscr{F}^i\right] \\ &= \mathbb{E}\left[U_{j_i}^i(t)\left(U_{j_i}^i(t')\right)^\mathsf{T} \mid \mathscr{F}^i\right] - \mathbb{E}\left[U_{j_i}^i(t) \mid \mathscr{F}^i\right]\left(\mathbb{E}\left[U_{j_i}^i(t') \mid \mathscr{F}^i\right]\right)^\mathsf{T}, \end{aligned} \tag{4.27}$$

we arrive at Equation (4.25). For details, see, Zhao et al. (2021c). □

Remark 4.12. *The matrix $\mathbf{\Lambda}^i(t,s)$ above is often referred to as the transition matrix in control theory (Brogan, 2011). Although in general $\mathbf{\Lambda}^i(t,s)$ does not have a closed-form representation, Peano–Baker series in Theorem 2.17 can be used to approximate it successively (Baake and Schlägel, 2011; DaCunha, 2005). One can also use Magnus expansions, if an exponential representation of transition matrix (i.e., $\mathbf{\Lambda}^i(t,s) = \exp(\cdot)$) is required, but the convergence usually requires strict conditions on A^i (Moan and Niesen, 2008).*

However, if A^i is self-commuting for all $t \in \mathbb{T}$, then the transition matrix simplifies to $\mathbf{\Lambda}^i(t,s) = \exp\left(\int_{t_0}^t A^i\left(s; \mathscr{U}^i(s)\right) \mathrm{d}s\right)$.

The converse of Theorem 4.11 is also available to some extent in the sense that the covariance functions in batch DGPs can be translated into the SDE coefficients of state-space DGPs. For how to proceed on this, we refer the reader to Hartikainen and Särkkä (2010); Särkkä et al. (2013).

4.5 Numerical simulation of SS-DGPs

In this section we discuss the numerical simulation of the SDEs describing SS-DGPs. In particular we present approximate discretisation methods that leverage the hierarchical nature of SS-DGPs, then we discuss alternatives that would result in exact simulations.

Discretisation of SDEs

In order to simulate SS-DGPs, it is very common to consider discretisations of their SDEs. In particular, we focus on the Gaussian increment-based explicit discretisations of the form

$$V_k \approx f_{k-1}(V_{k-1}) + q_{k-1}(V_{k-1}), \tag{4.28}$$

where $V_k \coloneqq V(t_k)$ and $q_{k-1} \sim \mathrm{N}(0, Q_{k-1}(V_{k-1}))$, and the functions f_{k-1} and Q_{k-1} depend on the discretisation scheme used.

Unfortunately, many commonly used discretisation methods fail to provide valid numerical schemes for SS-DGPs. For instance, the Euler–Maruyama method yields singular covariance Q_{k-1} for smooth Matérn SS-DGPs (see, e.g., Example 4.17). While higher-order Itô–Taylor expansions, such as Milstein's method, exist, they are only numerically efficient for constant, diagonal, or more generally, commutative dispersion function b (see, the definition of commutative noise in Kloeden and Platen, 1992, Chapter 10). However, dispersion functions of SS-DGPs may not always verify these conditions (e.g., Example 4.16).

The Taylor moment expansion (TME) method presented in Section 3.3 does not suffer from the problems of high-order Itô–Taylor expansions, but on the other hand it requires sufficient smoothness on the SDE coefficients. Moreover, the resulting covariance estimate Q_{k-1} used in the TME-based discretisation in Equation (4.28) may be singular. While the smoothness of the coefficients is a necessary price to pay, the possible singularity of the estimated covariance Q_{k-1} can be addressed. We refer the reader back to Section 3.4 for methods to do so.

In this thesis, we additionally present an ad-hoc discretisation approach by leveraging the hierarchical structure of SS-DGPs and explicit solutions of linear SDEs (e.g., Equation (4.24)). The idea relies on approximating the SDE of each GP element between two time steps t_{k-1} and t_k by a time-invariant SDE, the coefficients of which depend on the values of its parent GPs at t_{k-1}. This idea roots in the so-called local linearisation methods as in Ozaki (1993); Särkkä and Solin (2019). By using this approach, the transition matrix $\mathbf{\Lambda}$, as defined in Theorem 4.11, reduces to a matrix exponential. The following algorithm shows how this hierarchical discretisation can be used in practice.

Algorithm 4.13 (Locally conditional discretisation)**.** *Starting from any $t_{k-1} \in \mathbb{T}$, locally conditional discretisation (LCD) approximates the solution of SDEs* (4.21) *at time $t_k \in \mathbb{T}$ by*

$$U^i_{j_i}(t_k) \approx \widetilde{\mathbf{\Lambda}}^i(t_k, t_{k-1}) U^i_{j_i}(t_{k-1}) + \int_{t_{k-1}}^{t_k} \widetilde{\mathbf{\Lambda}}^i(t_k, s) B^i\big(t_{k-1}; \mathscr{U}^i(t_{k-1})\big)\, \mathrm{d}W^i(s), \quad (4.29)$$

for $i = 1, 2, \ldots, L$, where $\widetilde{\mathbf{\Lambda}}^i(t_k, s) \coloneqq \exp\big((t_k - s) A^i(t_{k-1}; \mathscr{U}^i(t_{k-1}))\big)$, and A^i and B^i are defined in Equation (4.20)*.*

Remark 4.14. *Except in special cases (such as Matérn SS-DGPs presented later), Equation* (4.29) *needs to be solved numerically. This can be done, for example, by using the methods highlighted around Equations* (2.20) *and* (2.21)*.*

It is worth mentioning that the non-stationary Gaussian state-space model introduced by Li (2020) coincides with the LCD approximation to the Matérn class of SS-DGPs (see, Section 4.6).

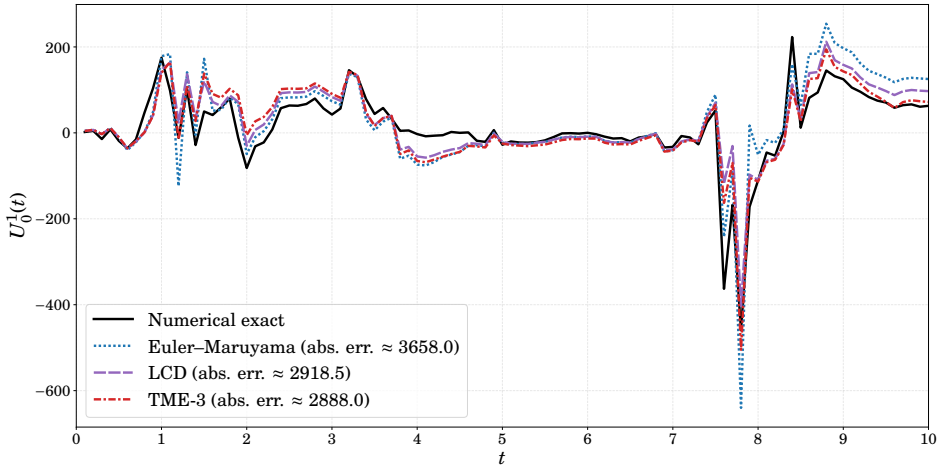


Figure 4.3. Comparison of different discretisation schemes on the Matérn $\nu = 1/2$ SS-DGP defined in Example 4.16, where we let parameters $\ell_2 = \ell_3 = 1$ and $\sigma_2 = \sigma_3 = 0.1$. The numbers in the legend are the cumulative absolute errors with respect to the (numerically) exact discretisation.

Figure 4.3 illustrates a comparison amongst the Euler–Maruyama, TME, and LCD methods on a Matérn SS-DGP. On this example, both LCD and TME methods outperform Euler–Maruyama substantially, especially in the "high-frequency" portions of this SDE trajectory.

Exact simulation methods

Apart from discretisation-based simulations, there also exist exact simulation methods (Beskos and Roberts, 2005; Kessler et al., 2012; Blanchet and Zhang, 2020). Although these methods can avoid the discretisation errors, they are usually limited to specific types of SDEs, which may not apply to all SS-DGPs. As an example, the method introduced by Beskos and Roberts (2005) requires that the dispersion coefficient be constant, which is usually not the case in SS-DGPs.

Finally, it is worth noting that each sub-SDE in Equation (4.21) is a linear SDE conditionally on its parent GPs. Hence, we could borrow the idea of Gibbs sampling (Robert and Casella, 2004) in order to sample from $U^i_{j_i}$ for $i = L, L-1, \ldots, 1$. While this method was not implemented in the context of this thesis, it is likely to improve on the LCD method and will therefore be a subject of future work.

4.6 Deep Matérn processes

In this section, we present SS-DGPs that are constructed in the Matérn sense. Specifically, we choose the SDE coefficients in Equation (4.20) in such a way that each GP element is a Matérn GP when conditioned on its parent GPs.

Let us start by considering linear SDEs of the form

$$\mathrm{d}U(t) = A\,U(t)\,\mathrm{d}t + B\,\mathrm{d}W(t), \tag{4.30}$$

where the initial condition $U(t_0) \sim \mathrm{N}(0, P_0)$ is a Gaussian random variable, and the Wiener process $W\colon \mathbb{T} \to \mathbb{R}$ takes value in $\mathbb{R}$. Let $\nu \in \left\{\frac{1}{2}, \frac{3}{2}, \ldots\right\}$ and $\gamma = \nu + \frac{1}{2}$. Suppose that the state $U\colon \mathbb{T} \to \mathbb{R}^\gamma$ verifies

$$U(t) = \begin{bmatrix} \overline{U}(t) & \frac{\mathrm{d}\overline{U}}{\mathrm{d}t}(t) & \cdots & \frac{\mathrm{d}^{\gamma-1}\overline{U}}{\mathrm{d}t^{\gamma-1}}(t) \end{bmatrix}^\mathsf{T}, \tag{4.31}$$

and that the coefficients in Equation (4.30) are given by

$$A = \begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ \vdots & & \ddots & \\ -\binom{\gamma}{0}\kappa^\gamma & -\binom{\gamma}{1}\kappa^{\gamma-1} & \cdots & -\binom{\gamma}{\gamma-1}\kappa \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \frac{\sigma\Gamma(\gamma)(2\kappa)^{\gamma-\frac{1}{2}}}{\sqrt{\Gamma(2\gamma-1)}} \end{bmatrix}, \tag{4.32}$$

where $\kappa = \sqrt{2\nu}/\ell$. Furthermore, suppose that the initial covariance P_0 solves the corresponding Lyapunov equation (see, Equation (4.10)) of the SDE. Then the process $\overline{U}\colon \mathbb{T} \to \mathbb{R}$ in Equation (4.31) is a zero-mean Matérn GP with the covariance function $C_{\text{Mat.}}$ defined in Equation (4.3) (Särkkä et al., 2013; Solin, 2016).

Remark 4.15. *The matrix A in Equation* (4.32) *is Hurwitz (Khalil, 2002) as all its eigenvalues have strictly negative real part. However, A is prone to be ill-conditioned if ν is large, resulting in numerically unstable SDEs. This can be addressed by using balancing algorithms (Osborne, 1960; Parlett and Reinsch, 1971).*

Based on the aforementioned Matérn SDE representation, we can now construct Matérn SS-DGPs by choosing their coefficients A^i and B^i for $i = 1, \ldots, L$, as per Equation (4.32). As an example, suppose that the i-th GP element $U^i_{j_i} \in \mathbb{R}^\gamma$ in Equation (4.20) has two parents $U^m_i\colon \mathbb{T} \to \mathbb{R}^{d_m}$ and $U^n_i\colon \mathbb{T} \to \mathbb{R}^{d_n}$, that encode the length scale and the magnitude parameters, respectively. Then, we can select two suitable transformation functions $g_m\colon \mathbb{R}^{d_m} \to \mathbb{R}_{>0}$ and $g_n\colon \mathbb{R}^{d_n} \to \mathbb{R}_{>0}$, and let

$$\ell_i(t) = g_m\left(U^m_i(t)\right) \tag{4.33}$$

and

$$\sigma_i(t) = g_n\left(U^n_i(t)\right). \tag{4.34}$$

Under these notations, the coefficient A^i of $U^i_{j_i}$ reads

$$A^i(t; \mathscr{U}^i) = A^i\left(U^m_i(t)\right) = \begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ \vdots & & \ddots & \\ -\binom{\gamma}{0}\kappa_i^\gamma(t) & -\binom{\gamma}{1}\kappa_i^{\gamma-1}(t) & \cdots & -\binom{\gamma}{\gamma-1}\kappa_i(t) \end{bmatrix}, \tag{4.35}$$

where $\kappa_i(t) = \sqrt{2\nu}/\ell_i(t)$. Likewise, one can derive the coefficient $B^i(t;\mathscr{U}^i) = B^i\left(U_i^m(t), U_i^n(t)\right) = \begin{bmatrix} 0 & 0 & \cdots & \sigma_i(t)\Gamma(\gamma)(2\kappa_i(t))^{\gamma-\frac{1}{2}}(\Gamma(2\gamma-1))^{\frac{1}{2}} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^\gamma$.

Transformation functions should also be chosen regular enough so that the solution of the related SDE is well-defined (see, Section 4.4). For example, in Zhao et al. (2021a,c), we use $g(u) = \exp(u)$, $g(u) = \arctan(u) + \pi/2$, or $g(u) = \log(1 + \exp(u))$.

SDEs of Matérn SS-DGPs are time-homogeneous by construction (i.e., the coefficients A^i and B^i for $i = 1,2,\ldots,L$ do not explicitly depend on time). Provided that the transformation functions are chosen suitably as per Øksendal (2007, Definition 7.1.1), the Matérn SS-DGPs are then Itô diffusions. This can bring many useful features, such as the strong Markov property (Ikeda and Watanabe, 1992).

In the following, we give some concrete examples of Matérn SS-DGPs and plot a few of their simulations.

Example 4.16 (Matérn $\nu = 1/2$ SS-DGP with three GP elements)**.** *Let $\nu = 1/2$. Consider the following SDEs*

$$\begin{aligned} \mathrm{d}U_0^1(t) &= -\frac{1}{\ell_1(t)} U_0^1(t)\,\mathrm{d}t + \frac{\sqrt{2}\sigma_1(t)}{\sqrt{\ell_1(t)}}\,\mathrm{d}W^1(t), \\ \mathrm{d}U_1^2(t) &= -\frac{1}{\ell_2} U_1^2(t)\,\mathrm{d}t + \frac{\sqrt{2}\sigma_2}{\sqrt{\ell_2}}\,\mathrm{d}W^2(t), \\ \mathrm{d}U_1^3(t) &= -\frac{1}{\ell_3} U_1^3(t)\,\mathrm{d}t + \frac{\sqrt{2}\sigma_3}{\sqrt{\ell_3}}\,\mathrm{d}W^3(t), \end{aligned} \tag{4.36}$$

where $\ell_1(t) = g\left(U_1^2(t)\right)$ and $\sigma_1 = g\left(U_1^3(t)\right)$ are the length scale and magnitude of $U_0^1(t)$, respectively. The solution $V(t) = \begin{bmatrix} U_0^1(t) & U_1^2(t) & U_1^3(t) \end{bmatrix}^\mathsf{T}$ is said to be a Matérn $\nu = 1/2$ SS-DGP.

Example 4.17 (Matérn $\nu = 3/2$ SS-DGP with three GP elements)**.** *Let $\nu = 3/2$. Consider the following SDEs*

$$\begin{aligned} \mathrm{d}U_0^1(t) &= \begin{bmatrix} 0 & 1 \\ -\left(\frac{\sqrt{3}}{\ell_1(t)}\right)^2 & \frac{-2\sqrt{3}}{\ell_1(t)} \end{bmatrix} U_0^1(t)\,\mathrm{d}t + \begin{bmatrix} 0 \\ 2\sigma_1(t)\left(\frac{\sqrt{3}}{\ell_1(t)}\right)^{\frac{3}{2}} \end{bmatrix} \mathrm{d}W^1(t), \\ \mathrm{d}U_1^2(t) &= \begin{bmatrix} 0 & 1 \\ -\left(\frac{\sqrt{3}}{\ell_2}\right)^2 & \frac{-2\sqrt{3}}{\ell_2} \end{bmatrix} U_1^2(t)\,\mathrm{d}t + \begin{bmatrix} 0 \\ 2\sigma_2\left(\frac{\sqrt{3}}{\ell_2}\right)^{\frac{3}{2}} \end{bmatrix} \mathrm{d}W^2(t), \\ \mathrm{d}U_1^3(t) &= \begin{bmatrix} 0 & 1 \\ -\left(\frac{\sqrt{3}}{\ell_3}\right)^2 & \frac{-2\sqrt{3}}{\ell_3} \end{bmatrix} U_1^3(t)\,\mathrm{d}t + \begin{bmatrix} 0 \\ 2\sigma_3\left(\frac{\sqrt{3}}{\ell_3}\right)^{\frac{3}{2}} \end{bmatrix} \mathrm{d}W^3(t), \end{aligned} \tag{4.37}$$

where $U_0^1(t) = \begin{bmatrix} \overline{U}_0^1(t) & \frac{\mathrm{d}\overline{U}_0^1}{\mathrm{d}t}(t) \end{bmatrix}^\mathsf{T}$, and similarly for $U_1^2(t)$ and $U_1^3(t)$. The length scale and magnitude of $U_0^1(t)$ are given by $\ell_1(t) = g\left(U_1^2(t)\right)$ and $\sigma_1 = g\left(U_1^3(t)\right)$, respectively, for $g\colon \mathbb{R}^2 \to \mathbb{R}_{>0}$. The solution $V(t) = \begin{bmatrix} U_0^1(t) & U_1^2(t) & U_1^3(t) \end{bmatrix}^\mathsf{T}$ is said to be a Matérn $\nu = 3/2$ SS-DGP.

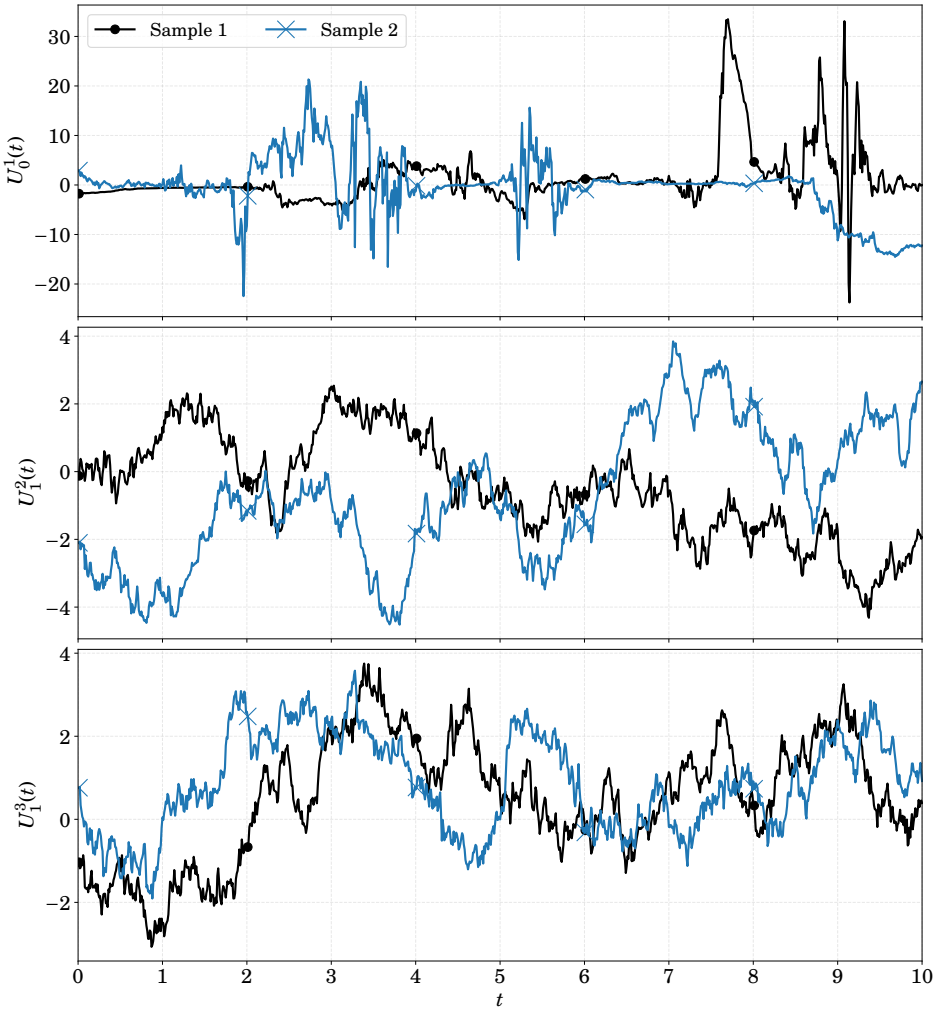


Figure 4.4. This figure shows two samples (plotted in different colours and markers) drawn from the Matérn $\nu = 1/2$ SS-DGP defined in Example 4.16.

It is worth noting that for this model the Euler–Maruyama scheme gives a singular discretisation covariance.

The Examples 4.16 and 4.17 feature Matérn SS-DGPs with only three GP elements. This hierarchy/depth can be continued further to represent higher degrees of non-stationarity.

Figures 4.4 and 4.5 illustrate a few samples drawn from the Matérn SS-DGPs defined in Examples 4.16 and 4.17, respectively. More specifically, for the Matérn $\nu = 1/2$ SS-DGP in Example 4.16 we use $\ell_2 = \ell_3 = \sigma_2 = \sigma_3 = 2$ and $g(u) = \exp(u)$, while for the Matérn $\nu = 3/2$ SS-DGP in Example 4.17 we use $\ell_2 = \ell_3 = 0.5$, $\sigma_2 = \sigma_3 = 2$ and $g(u) = \exp\left(\begin{bmatrix}1 & 0\end{bmatrix} u\right)$. The initial states are standard Gaussian random vectors with unit covariances in both cases.

From Figures 4.4 and 4.5, we can observe non-stationary in the behaviour of U_0^1. This results from its length scale and magnitude being driven by its parent GPs U_1^2 and U_1^3. As an example, Sample 2 (blue line) of U_0^1 in Figure 4.5 exhibits low-magnitude high-frequency jittering around

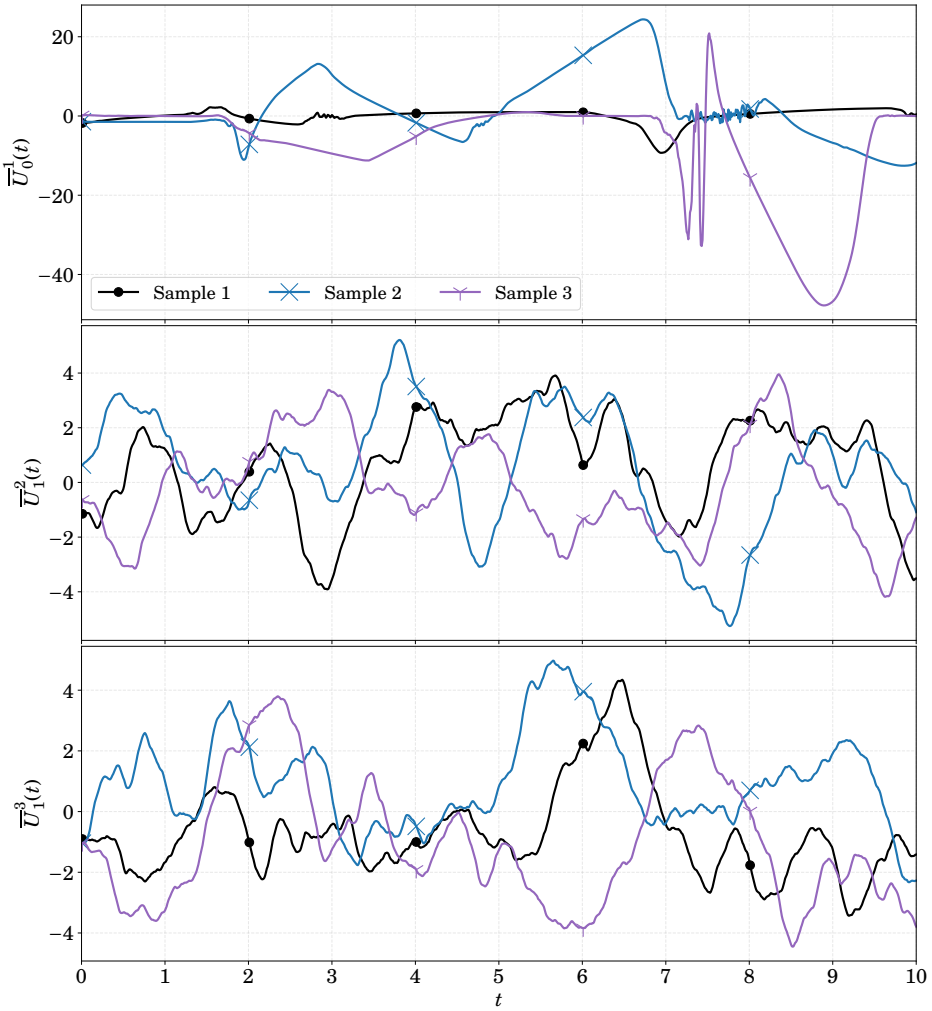


Figure 4.5. This figure shows three samples (plotted in different colours and markers) drawn from the Matérn $\nu = 3/2$ SS-DGP defined in Example 4.17.

$t \in [7,8]$ because the length scale $g(\overline{U}_1^2)$ and magnitude $g(\overline{U}_1^3)$ are relatively small on $t \in [7,8]$. On the other hand Sample 3 (magenta line) of $\overline{U}_0^1$ in Figure 4.5 exhibits high-magnitude medium-frequency jittering around $t \in [7,8]$ because the length scale $g(\overline{U}_1^2)$ and magnitude $g(\overline{U}_1^3)$ are relatively average and high, respectively, on $t \in [7,8]$.

The main usefulness of this Matérn construction is that the resulting Matérn SS-DGPs can provide generic priors for modelling a wide class of continuous functions (with smoothness parameter $\gamma - 1$). These priors are flexible in the sense that they have non-stationary characteristics which can be learnt from data. In Chapter 5, we will show some real applications of Matérn SS-DGPs.

Apart from the Matérn construction, it is also possible to build SS-DGPs by formulating SDE coefficients in some other meaningful ways. For example, Solin and Särkkä (2014) construct SDEs that represent quasi-periodic oscillators, and Rangapuram et al. (2018) parametrise SDEs with

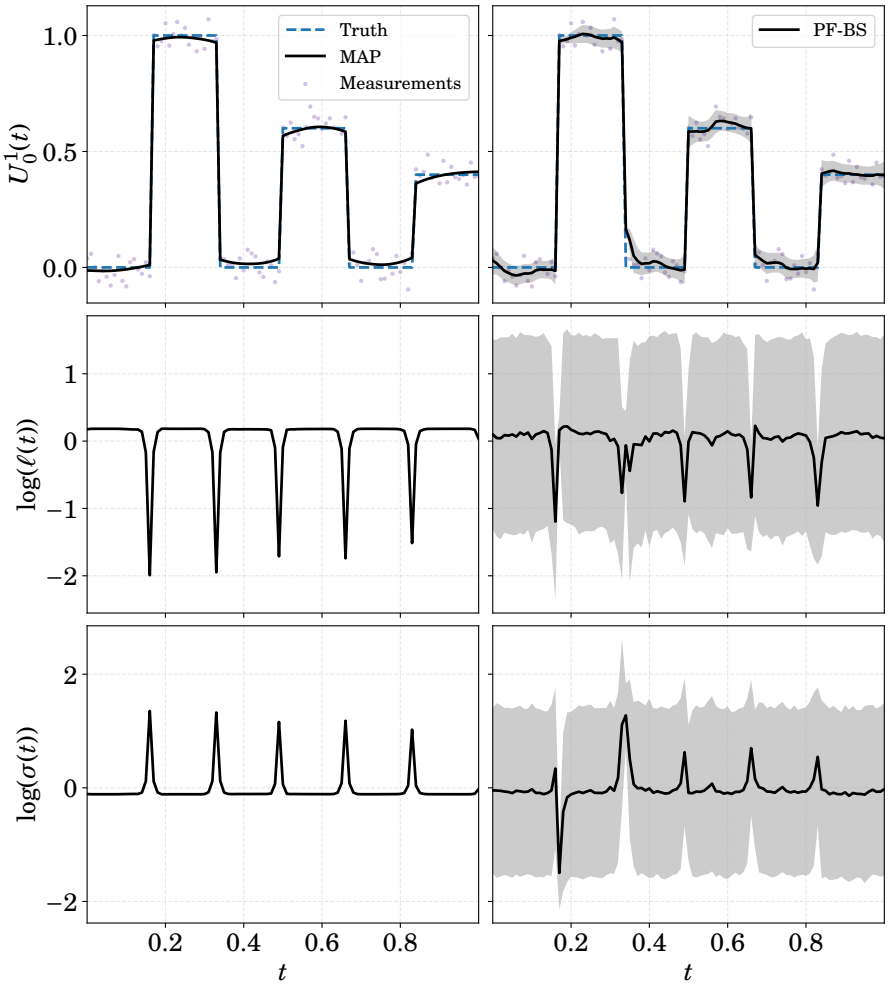


Figure 4.6. Matérn SS-DGP regression on a rectangular signal. The first column corresponds to the MAP estimate of the SDE state, while the second column corresponds to a full posterior estimate using PF-BS (particle filter and backward simulation smoother).

neural networks for time series forecasting.

4.7 SS-DGP Regression

In this section, we show how to solve SS-DGP regression problems for discrete measurement data. Since SS-DGPs are characterised by SDEs, we view these problems as continuous-discrete Bayesian smoothing problems (see, Section 2.2).

Let $V : \mathbb{T} \to \mathbb{R}^{\sum_{i=1}^{L} d_i}$ be an SS-DGP defined as per Equation (4.21). Suppose that we measure V at $t_1, t_2, \ldots, t_T \in \mathbb{T}$, by a (non-linear) function $h : \mathbb{R}^{\sum_{i=1}^{L} d_i} \to \mathbb{R}^{d_y}$ and additive Gaussian noises $\xi_k \sim \mathrm{N}(0, \Xi_k)$ for $k = 1, 2, \ldots, T$. We consider the SS-DGP regression problem in its continuous-discrete state-space form

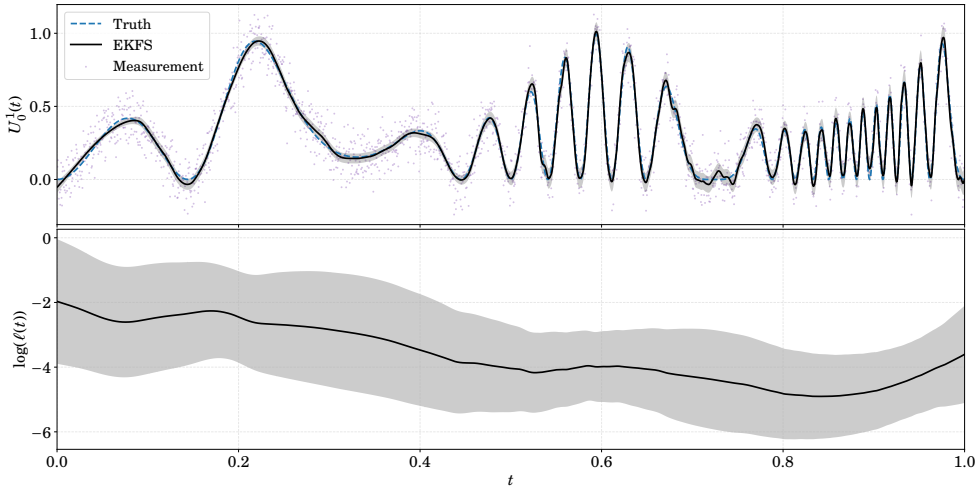


Figure 4.7. Matérn SS-DGP regression on a composite sinusoidal signal by using EKFS (extended Kalman filter and smoother).

$$\begin{aligned} \mathrm{d}V(t) &= a(V(t))\,\mathrm{d}t + b(V(t))\,\mathrm{d}W(t), \quad V(t_0) = V_0, \\ Y_k &= h(V_k) + \xi_k, \end{aligned} \tag{4.38}$$

where we abbreviate $V_k \coloneqq V(t_k)$. Suppose we have a set of measurements $y_{1:T} = \{y_1, y_2, \ldots, y_T\}$, we want to learn the (smoothing) posterior density $p_{V_k \mid Y_{1:T}}(v_k \mid y_{1:T})$ for $k = 1, 2, \ldots, T$, or more generally $p_{V(t) \mid Y_{1:T}}(v, t \mid y_{1:T})$, for any $t \in \mathbb{T}$. We can then use the methods presented in Section 2.2 to solve the regression/continuous-discrete smoothing problem above.

Thanks to the Markov property of the SS-DGP prior, we can solve this regression problem in linear computational time with respect to T by leveraging Bayesian filtering and smoothing methods. This is in contrast with batch DGPs, where one often needs to solve matrix inversions of dimension $T \times T$.

In Figures 4.6 and 4.7, we plot some SS-DGP regression examples taken from Zhao et al. (2021a). Compared to the GP regression shown in Figure 1.1, we can see the advantages of SS-DGPs for fitting irregular data. Also, the estimated length scale and magnitude parameters can explain well the changes of regime in the data generating process.

It would be also possible to extend SS-DGP regression to classification by modifying the measurement model in Equation (4.38) accordingly (Neal, 1999; Rasmussen and Williams, 2006; Ángel F. García-Fernández et al., 2019). For example, we can assume that the measurement follows a categorical distribution, with parameters determined by the SS-DGP states (Rasmussen and Williams, 2006).

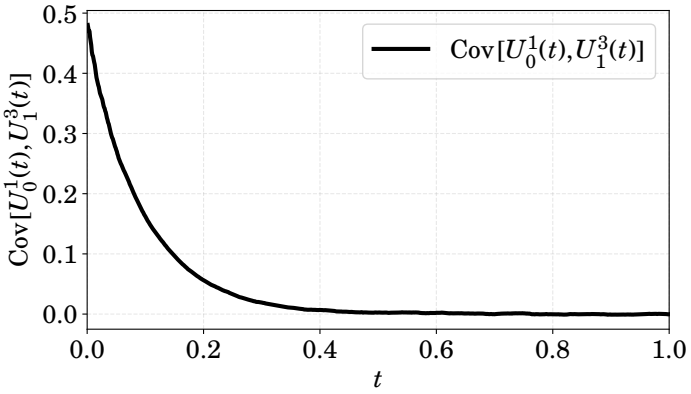


Figure 4.8. Evolution of $\mathrm{Cov}\left[U_0^1(t), U_1^3(t)\right]$ in Example 4.16, estimated by 20,000 independent Monte Carlo runs. Parameters are $\ell_2 = \ell_3 = \sigma_2 = \sigma_3 = 0.1$, and the initial $\mathrm{Cov}\left[U_0^1(0), U_1^3(0)\right] = 0.5$. We observe that the covariance (numerically) converges to zero monotonically as t grows.

4.8 Identifiability analysis of Gaussian approximated SS-DGP regression

Gaussian filters and smoothers (GFSs, see, Section 2.2.3) are widely used classes of Bayesian filters and smoothers. Moreover, Zhao et al. (2021a) show that GFSs are particularly efficient for solving SS-DGP regression problems. However, for a certain class of SS-DGPs, GFSs cannot identify (i.e., estimate the posterior density of) their state components as $t \to \infty$. More specifically, in this section, we show how – under some weak assumptions on the SS-DGP regression model coefficients – the posterior (cross-)covariance estimates of the regression problem solutions at the measurements times t_k collapse to 0 as $k \to \infty$.

To explain the problem in short, let us suppose that we have an SS-DGP regression model with the SDE given by Example 4.16, and that we measure the first GP element U_0^1 with additive Gaussian noises. Further suppose that we apply GFSs (see, Algorithm 2.11) to solve the regression problem. It turns out that this SDE has a vanishing $\mathrm{Cov}\left[U_0^1(t), U_1^3(t)\right] \to 0$ as $t \to \infty$ (see, Figure 4.8 for a numerical illustration), and that the GFS estimated posterior $\mathrm{Cov}\left[U_0^1(t_k), U_1^3(t_k) \mid y_{1:k}\right]$ vanishes to zero as $k \to \infty$ too. Consequently, the Kalman gain for the component U_1^3 converges to zero as $k \to \infty$. This means that the posterior distribution of U_1^3 estimated by GFSs will use no information from measurements as $k \to \infty$.

In order to formulate the problem, we limit ourselves to a class of SS-DGP regression models for which the dispersion term of the observed GP element is parametrised by another GP element. Formally, we consider $U_0^1 \colon \mathbb{T} \to \mathbb{R}$ and $U_1^2 \colon \mathbb{T} \to \mathbb{R}$ that are the solutions of the pair of SDEs

$$\begin{aligned} \mathrm{d}U_0^1(t) &= A^1(\psi(t)) U_0^1(t)\,\mathrm{d}t + B^1\left(\psi(t), U_1^2(t)\right)\mathrm{d}W^1(t), \\ \mathrm{d}U_1^2(t) &= A^2(\varphi(t)) U_1^2(t)\,\mathrm{d}t + B^2(\varphi(t))\,\mathrm{d}W^2(t), \end{aligned} \tag{4.39}$$

on a filtered probability space $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P})$, where the initial conditions, the processes $\psi\colon \mathbb{T} \to \mathbb{R}$ and $\varphi\colon \mathbb{T} \to \mathbb{R}$, and the Wiener processes $W^1\colon \mathbb{T} \to \mathbb{R}$ and $W^2\colon \mathbb{T} \to \mathbb{R}$ are mutually independent.

Remark 4.18. *Note that the index 2 on U_1^2 is arbitrary as the definition of a DGP is invariant of reindexing of its components (see, Definition 4.6).*

Remark 4.19. *The SDEs given by Equation* (4.39) *represent a class of SS-DGPs for which an inner GP element U_1^2 parametrises the dispersion term of the measured GP element U_0^1. Since the parents of U_0^1 and U_1^2 are not necessarily Gaussian (but are instead conditional Gaussian), we generically name their parents ψ and φ which can be any well-defined processes. As an example, in the left figure of Figure 4.2, one can imagine ψ as the representation of U_1^3, U_3^6, and U_3^7, while φ as the representation of U_2^4 and U_2^5.*

Let the random variables

$$Y_k = U_0^1(t_k) + \xi_k, \quad \xi_k \sim \mathrm{N}(0, \Xi_k), \tag{4.40}$$

for $k = 1, 2, \ldots$ stand for the measurements at time $t_1, t_2, \ldots$, and assume that $\inf_k \{t_k - t_{k-1} \colon k = 1, 2, \ldots\} > 0$.

We use the following assumptions.

Assumption 4.20. *The coefficients $A^1\colon \mathbb{R} \to \mathbb{R}_{<0}$, $B^1\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $A^2\colon \mathbb{R} \to \mathbb{R}_{<0}$, $B^2\colon \mathbb{R} \to \mathbb{R}$, and the initial conditions are chosen regular enough so that the weak uniqueness (see, Definition 2.6) holds for the SDE in Equation* (4.39) .

Assumption 4.21. *The components U_0^1 and U_1^2 at the initial time t_0 satisfy $\mathbb{E}\left[\left|U_0^1(t_0)\right|\right] < \infty$, $\mathbb{E}\left[\left|U_1^2(t_0)\right|\right] < \infty$, and $\mathbb{E}\left[\left|U_0^1(t_0) U_1^i(t_0)\right|\right] < \infty$.*

Assumption 4.22. *There exists constants $c_{A^1} < 0$ and $c_{A^2} < 0$ such that $\mathbb{P}$-almost surely the processes $(A^1 \circ \psi)(t) \leq c_{A^1}$ and $(A^2 \circ \varphi)(t) \leq c_{A^2}$ for all $t \in \mathbb{T}$.*

Assumption 4.23. *There exist constants $c_A > 0$ and $c_B > 0$ such that, for all $t \in \mathbb{T}$, $\mathbb{E}\left[\left(A^1(\psi(t)) U_0^1(t)\right)^2\right] \leq c_A$ and $\mathbb{E}\left[\left(B^1\left(\psi(t), U_1^2(t)\right)\right)^2\right] \geq c_B$.*

Assumption 4.24. *There exists $\Xi_{\inf} > 0$ such that for every $k = 1, 2, \ldots$, either $\Xi_k > \Xi_{\inf}$ or $\Xi_k = 0$.*

Assumption 4.20 ensures SDEs (4.39) be well-defined. Assumption 4.21 postulates absolute integrability of the initial conditions which is used in the proof of Lemma 4.25. Assumption 4.23 aims to yield a positive lower bound for $\mathrm{Var}\left[U_0^1(t)\right]$ as used in Corollary 4.29.

Assumption 4.22 is the key assumption to have the prior covariance vanishing. This assumption is pragmatic because it ensures that the mean of U_0^1 and U_1^2 shrinks to zero over time. Also, if one considers $-1/A^1$ and $-1/A^2$ as length scales, then this assumption guarantees their positivity.

Lemma 4.25. *Under Assumptions 4.20 to 4.22,*

$$\lim_{t\to\infty} \operatorname{Cov}\left[U_0^1(t), U_1^2(t)\right] = 0. \tag{4.41}$$

Proof. By Itô's formula and the law of total expectation, one can find that

$$\begin{aligned}\operatorname{Cov}\left[U_0^1(t), U_1^2(t)\right] = \mathbb{E}\left[\mathbb{E}\left[U_0^1(t_0)U_1^2(t_0) \mid \psi(t_0), \varphi(t_0)\right] e^{\int_{t_0}^t A^1(\psi(s))+A^2(\varphi(s))\,\mathrm{d}s}\right] \\ - \mathbb{E}\left[\mathbb{E}\left[U_0^1(t_0) \mid \psi(t_0)\right] e^{\int_{t_0}^t A^1(\psi(s))\,\mathrm{d}s}\right] \\ \times \mathbb{E}\left[\mathbb{E}\left[U_1^2(t_0) \mid \varphi(t_0)\right] e^{\int_{t_0}^t A^2(\varphi(s))\,\mathrm{d}s}\right].\end{aligned} \tag{4.42}$$

Next, knowing that $A^1 \circ \psi$ and $A^2 \circ \varphi$ are upper bounded by Assumption 4.22, we can then apply the triangle inequality and conditional Jensen's inequality (Klenke, 2014, Theorem 8.20) to bound the three terms in the equation above. As an example, the first term admits

$$\begin{aligned}&\left|\mathbb{E}\left[\mathbb{E}\left[U_0^1(t_0)U_1^2(t_0) \mid \psi(t_0), \varphi(t_0)\right] e^{\int_{t_0}^t A^1(\psi(s))+A^2(\varphi(s))\,\mathrm{d}s}\right]\right| \\ &\leq \mathbb{E}\left[\left|\mathbb{E}\left[U_0^1(t_0)U_1^2(t_0) \mid \psi(t_0), \varphi(t_0)\right]\right| \left|e^{\int_{t_0}^t A^1(\psi(s))+A^2(\varphi(s))\,\mathrm{d}s}\right|\right] \\ &\leq \mathbb{E}\left[\left|U_0^1(t_0)U_1^2(t_0)\right|\right] e^{(c_{A^1}+c_{A^2})(t-t_0)}.\end{aligned} \tag{4.43}$$

For the rest two expectation terms in Equation (4.42), *mutatis mutandis*. Finally, by taking limits on both side of Equation (4.42) and using the bounds above, one arrives at Equation (4.41). □

Remark 4.26. *The proof above slightly deviates from the proof given in Zhao et al. (2021a, Lemma 1) as the original proof uses Assumption 4.22 in a different order, which yields an unnecessarily stricter bound than that of Equation* (4.43)*.*

Lemma 4.27. *Under Assumption 4.21, for every $\epsilon > 0$, there exists $\zeta_\epsilon > 0$ such that*

$$\begin{aligned}&\operatorname{Var}\left[U_0^1(t)\right] \\ &\geq \frac{1}{z(t)} \int_{t_0}^t z(s) \left(\mathbb{E}\left[\left(B^1\left(\psi(s), U_1^2(s)\right)\right)^2\right] - 2\epsilon\sqrt{\mathbb{E}\left[\left(A^1(\psi(s))U_0^1(s)\right)^2\right]}\right) \mathrm{d}s,\end{aligned} \tag{4.44}$$

where

$$z(t) = \exp\left\{\int_{t_0}^t 2\zeta_\epsilon \sqrt{\mathbb{E}\left[\left(A^1(\psi(s))U_0^1(s)\right)^2\right]}\,\mathrm{d}s\right\}.$$

Proof. We give the idea of the proof, for details, see, Zhao et al. (2021a). The first step is to express $\operatorname{Var}\left[U_0^1(t)\right]$ as the solution of an integral/differential equation. Then by using Hölder's inequality one can obtain an integral/differential inequality. Finally, by using the integrating factor method on $\left(z(t)\operatorname{Var}\left[U_0^1(t)\right]\right)$, one can recover the desired bound. □

Remark 4.28. *We can also use Theorem 2.15 to obtain alternative positive lower bounds by letting* $v(x) = \sqrt{x}$ *or* $v(x) = \epsilon + \zeta_\epsilon x$ *in Theorem 2.15. The resulting bounds do not involve the dispersion term* $\mathbb{E}\left[\left(B^1\left(\psi(s), U_1^2(s)\right)\right)^2\right]$ *but includes the initial variance* $\operatorname{Var}\left[U_0^1(t_0)\right]$ *instead.*

Corollary 4.29. *Under Assumptions 4.20 and 4.23, there exists an* $\epsilon > 0$ *such that*

$$\operatorname{Var}\left[U_0^1(t)\right] \geq \frac{(c_B - 2\epsilon\sqrt{c_A})(t - t_0)}{\exp\left(2\zeta_\epsilon\sqrt{c_A}(t - t_0)\right)} > 0, \tag{4.45}$$

for all $t \in \mathbb{T}$.

Proof. The bound follows from Lemma 4.27 and Assumption 4.23. In order to make the bound positive, one then needs to choose $\epsilon < c_B/(2\sqrt{c_A})$. □

We can now analyse the limit of the posterior covariance

$$\operatorname{Cov}\left[U_0^1(t_k), U_1^2(t_k) \mid y_{1:k}\right] \approx P_k^{1,2} \tag{4.46}$$

as approximated by Gaussian filters as $k \to \infty$. In order to do so, in Algorithm 4.30, we consider an abstract general form of Gaussian filters that suppose perfect integration in the prediction step. We use the notations $\operatorname{Cov}\left[U_0^1(t_k), U_1^2(t_k)\right]_{z_s}$ and $\operatorname{Var}\left[U_0^1(t_k)\right]_{z_s}$ to represent the values of $\operatorname{Cov}\left[U_0^1(t_k), U_1^2(t_k)\right]$ and $\operatorname{Var}\left[U_0^1(t_k)\right]$, respectively, at time t_k starting from any initial value z_s at time $t_s < t_k \in \mathbb{T}$.

Algorithm 4.30 (Abstract Gaussian filter for $P_k^{1,2}$)**.** *Suppose that we have initial conditions* $P_0^{1,2} = \operatorname{Cov}\left[U_0^1(t_0), U_1^2(t_0)\right]$ *and* $P_0^{1,1} = \operatorname{Var}\left[U_0^1(t_0)\right]$. *Starting from* $k = 1$ *the abstract Gaussian filter predicts*

$$\begin{aligned} \overline{P}_k^{1,2} &= \operatorname{Cov}\left[U_0^1(t_k), U_1^2(t_k)\right]_{P_{k-1}^{1,2}}, \\ \overline{P}_k^{1,1} &= \operatorname{Var}\left[U_0^1(t_k)\right]_{P_{k-1}^{1,1}}, \end{aligned} \tag{4.47}$$

and updates

$$\begin{aligned} P_k^{1,2} &= \overline{P}_k^{1,2} - \frac{\overline{P}_k^{1,1}\overline{P}_k^{1,2}}{\overline{P}_k^{1,1} + \Xi_k}, \\ K_k^{1,2} &= \frac{\overline{P}_k^{1,2}}{\overline{P}_k^{1,1} + \Xi_k}, \end{aligned} \tag{4.48}$$

for $k = 1, 2, \ldots$

Remark 4.31. *Algorithm 4.30 is a skeleton of Algorithm 2.11 that is only concerned with the covariance estimates* $P_k^{1,2}$ *for* $k = 1, 2, \ldots$. *However, this algorithm assumes that the predictions through the SDE are done exactly as per Equation* (4.47)*, which is usually unrealistic in practice. Zhao et al. (2021a) explain how this abstraction is derived.*

We can finally state the main result of this section.

Theorem 4.32. *Suppose that Assumptions 4.20 to 4.24 hold. Further assume that* $\left|\operatorname{Cov}\left[U_0^1(t_k), U_1^2(t_k)\right]_{z_{k-1}}\right| \leq |z_{k-1}|$ *for all initial* $z_{k-1} \in \mathbb{R}$ *and* $k = 1,2,\ldots$*, then Algorithm 4.30 gives*

$$\lim_{k\to\infty} P_k^{1,2} = 0. \tag{4.49}$$

Proof. The basic idea is to expand the recursion in Algorithm 4.30 for $k = 1,2,\ldots$, and by mathematical induction one can prove that

$$\left|P_k^{1,2}\right| \leq \left|P_0^{1,2}\right| \prod_{j=1}^{k} D_j,$$

where

$$D_j = \frac{R_j}{\overline{P}_j^{1,1} + R_j}.$$

Hence, the limit of $P_k^{1,2}$ depends on the limit of $\prod_{j=1}^{k} D_j$. Although D_j is always less than 1, the infinite product $\prod_{j=1}^{\infty} D_j$ does not necessarily converge to zero (e.g., Viète's formula). However, Lemma 4.27 and Assumption 4.24 ensure that $\overline{P}_j^{1,1}$ is lower bounded uniformly by some positive P_{inf}, so that $D_j < \frac{R_j}{R_j + P_{\text{inf}}}$, for all $j > 0$. Assumption 4.24 then allows to conclude. For details, see Zhao et al. (2021a). □

The consequence of Theorem 4.32 is that the Kalman gain $K_k^{1,2}$ in Algorithm 4.30 will also converge to zero as $k \to \infty$. It means that the Kalman update for the state $U_1^2(t_k)$ will not use information from measurements in the limit $k \to \infty$.

4.9 L^1-regularised batch and state-space DGP regression

Constrained/regularised regression, for example, the sparsity-inducing least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996) method is an important topic in statistics, machine learning, and inverse problems (Kaipio and Somersalo, 2005; Hastie et al., 2015). Heuristically, sparsity in DGPs may also yield several benefits, in particular for modelling discontinuous signals for which the length scale around the discontinuities should jump from a high value to almost zero (see, e.g., Figure 4.6). In this section, we show how L^1-regularisation can be interpreted and implemented in the context of batch and state-space DGP regressions.

Regularised batch DGP regression

For the sake of exposition and to keep notations simple, we will restrict ourselves to a shallow DGP, with only one observed GP element depending

on two latent GPs:

$$\begin{aligned} U_0^1(t) \,\big|\, \mathscr{U}^1 &\sim \mathrm{GP}\big(0, C^1(t,t';\mathscr{U}^1)\big), \\ U_1^2(t) &\sim \mathrm{GP}\big(0, C^2(t,t')\big), \\ U_1^3(t) &\sim \mathrm{GP}\big(0, C^3(t,t')\big), \\ Y_k &= U_0^1(t_k) + \xi_k, \quad \xi_k \sim \mathrm{N}(0, \Xi_k), \end{aligned} \tag{4.50}$$

where $\mathscr{U}^1 = \{U_1^2, U_1^3\}$, and we let the DGP $V(t) = \begin{bmatrix} U_0^1(t) & U_1^2(t) & U_1^3(t) \end{bmatrix}^\mathsf{T}$. Suppose that at times $\{t_k \in \mathbb{T} : k = 1,2,\ldots,T\}$ we have measurements $y_{1:T} = \{y_k : k = 1,2,\ldots,T\}$, the DGP regression aims to learn the posterior density

$$\begin{aligned} & p_{V_{1:T}\,|\,Y_{1:T}}(v_{1:T} \mid y_{1:T}) \\ & \propto p_{Y_{1:T}\,|\,V_{1:T}}(y_{1:T} \mid v_{1:T})\, p_{V_{1:T}}(v_{1:T}) \\ & = p_{Y_{1:T}\,|\,U^1_{0,1:T}}\big(y_{1:T} \mid u^1_{0,1:T}\big)\, p_{U^1_{0,1:T}\,|\,U^2_{1,1:T},U^3_{1,1:T}}\big(u^1_{0,1:T} \mid u^2_{1,1:T}, u^3_{1,1:T}\big) \\ & \quad \times p_{U^2_{1,1:T}}\big(u^2_{1,1:T}\big)\, p_{U^3_{1,1:T}}\big(u^3_{1,1:T}\big), \end{aligned} \tag{4.51}$$

where we define $V_{1:T} \coloneqq \{V(t_k) : k = 1,2,\ldots,T\}$, $U^1_{0,1:T} \coloneqq \big\{U_0^1(t_k) : k = 1,2,\ldots,T\big\}$, and similarly for $U^2_{1,1:T}$ and $U^3_{1,1:T}$. Now let us introduce three regularisation-inducing matrices $\Phi^1 \in \mathbb{R}^{T\times T}$, $\Phi^2 \in \mathbb{R}^{T\times T}$, and $\Phi^3 \in \mathbb{R}^{T\times T}$. We are interested in learning the posterior density (4.51) under an L^1-regularisation of the GP elements, that is by introducing the penalty terms

$$\big\|\Phi^1 u^1_{0,1:T}\big\|_1, \quad \big\|\Phi^2 u^2_{1,1:T}\big\|_1, \quad \text{and} \quad \big\|\Phi^3 u^3_{1,1:T}\big\|_1. \tag{4.52}$$

In other words, we encourage the Φ-transformed variables to be sparse in the L^1 norm sense. For example, if we let Φ to be the identity matrix (respectively, a finite difference matrix), then the resulting penalty will correspond to increasing elementwise sparsity (respectively, reducing the total variation of the function).

We consider a maximum a posterior (MAP) approach for solving Equation (4.51), and we express the regularised DGP regression problem as a penalised optimisation problem. Namely, by taking the negative log of Equation (4.51), we get an objective function

$$\begin{aligned} \mathscr{L}^{\mathrm{B}} &\coloneqq \mathscr{L}^{\mathrm{B}}(v_{1:T}) \coloneqq \mathscr{L}^{\mathrm{B}}(u^1_{0,1:T}, u^2_{1,1:T}, u^3_{1,1:T}) \\ &= \big\|u^1_{0,1:T} - y_{1:T}\big\|^2_{\Xi_{1:T}} + \big\|u^1_{0,1:T}\big\|^2_{C^1_{1:T}} + \log\det\big(2\pi C^1_{1:T}\big) \\ &\quad + \big\|u^2_{1,1:T}\big\|^2_{C^2_{1:T}} + \log\det\big(2\pi C^2_{1:T}\big) + \big\|u^3_{1,1:T}\big\|^2_{C^3_{1:T}} + \log\det\big(2\pi C^3_{1:T}\big), \end{aligned} \tag{4.53}$$

where we omit the factor $1/2$ and let $v_{1:T} \coloneqq \big\{u^1_{0,1:T}, u^2_{1,1:T}, u^3_{1,1:T}\big\}$ for simplicity. In the above Equation (4.53), notation $\|x\|_G = (xG^{-1}x)^{1/2}$ stands for the G-weighted Euclidean norm given a non-singular matrix G. We write

$C^1_{1:T} \in \mathbb{R}^{T\times T}$ for the matrix obtained by evaluating the covariance function C^1 on the Cartesian grid $(t_1,\ldots,t_T)\times(t_1,\ldots,t_T)$, and similarly for $C^2_{1:T}$ and $C^3_{1:T}$. The noise covariance $\Xi_{1:T}$ is the diagonal matrix of $\Xi_1,\ldots,\Xi_T$. We now introduce the regularisation term

$$\begin{aligned}\mathscr{L}^{\mathrm{B-REG}} &:= \mathscr{L}^{\mathrm{B-REG}}(v_{1:T}) \\ &= \lambda_1 \left\|\Phi^1 u^1_{0,1:T}\right\|_1 + \lambda_2 \left\|\Phi^2 u^2_{1,1:T}\right\|_1 + \lambda_3 \left\|\Phi^3 u^3_{1,1:T}\right\|_1,\end{aligned} \tag{4.54}$$

where the positive parameters λ_1, λ_2, and λ_3 stand for the strength of regularisation. The regularised batch DGP (R-DGP) regression aims at solving

$$v_{1:T} = \underset{v_{1:T}}{\arg\min}\ \mathscr{L}^{\mathrm{B}} + \mathscr{L}^{\mathrm{B-REG}}. \tag{4.55}$$

Remark 4.33. *It is important to recall that the covariance matrix $C^1_{1:T}$ in Equation* (4.53) *depends on the objective variables $u^2_{1,1:T}$ and $u^3_{1,1:T}$, hence, $\mathscr{L}^{\mathrm{B}}$ can be non-convex.*

ADMM solution of regularised batch DGP regression

There are many approaches to solving penalised optimisation problems of the form given in Equation (4.55) (Ruszczyński, 2006; Nocedal and Wright, 2006), however $\mathscr{L}^{\mathrm{B}}$ is usually non-convex, and $\mathscr{L}^{\mathrm{B-REG}}$ is not differentiable-everywhere. Heuristically, we can interpret the gradients of $\mathscr{L}^{\mathrm{B-REG}}$ at non-differential points by subgradients and then use gradient descent (GD) methods to find the minima. These GD-based approaches, however, can suffer from a slow convergence rate (Hastie et al., 2015) which limits their applicability in practice.

Zhao et al. (2021c) propose using the alternating direction method of multipliers (ADMM, Boyd and Vandenberghe, 2004) for solving the optimisation problem in Equation (4.55). The idea is to split the complicated optimisation problem in Equation (4.55) into simpler subproblems. To do so, let us introduce auxiliary variables $\theta_{1:T} := \left\{\theta^1_{1:T}, \theta^2_{1:T}, \theta^3_{1:T}\right\}$, where $\theta^1_{1:T} \in \mathbb{R}^T$, $\theta^2_{1:T} \in \mathbb{R}^T$, and $\theta^3_{1:T} \in \mathbb{R}^T$. We rewrite Equation (4.55) as an equality constrained problem

$$\begin{aligned} v_{1:T} = \underset{v_{1:T}}{\arg\min}\ & \left\|u^1_{0,1:T} - y_{1:T}\right\|^2_{\Xi_{1:T}} + \left\|u^1_{0,1:T}\right\|^2_{C^1_{1:T}} + \log\det\left(2\pi C^1_{1:T}\right) \\ & + \left\|u^2_{1,1:T}\right\|^2_{C^2_{1:T}} + \log\det\left(2\pi C^2_{1:T}\right) + \left\|u^3_{1,1:T}\right\|^2_{C^3_{1:T}} + \log\det\left(2\pi C^3_{1:T}\right) \\ & + \lambda_1 \left\|\theta^1_{1:T}\right\|_1 + \lambda_2 \left\|\theta^2_{1:T}\right\|_1 + \lambda_3 \left\|\theta^3_{1:T}\right\|_1 \\ & \text{subject to} \\ & \theta^1_{1:T} = \Phi^1 u^1_{0,1:T}, \quad \theta^2_{1:T} = \Phi^2 u^2_{1,1:T}, \quad \theta^3_{1:T} = \Phi^3 u^3_{1,1:T}. \end{aligned} \tag{4.56}$$

Then let us introduce multiplier variables $\eta_{1:T} \coloneqq \{\eta^1_{1:T}, \eta^2_{1:T}, \eta^3_{1:T}\}$, where $\eta^1_{1:T} \in \mathbb{R}^T$, $\eta^2_{1:T} \in \mathbb{R}^T$, and $\eta^3_{1:T} \in \mathbb{R}^T$. We can construct the augmented Lagrangian function $\mathscr{L}(v_{1:T}, \theta_{1:T}, \eta_{1:T})$ associated with problem (4.56) as

$$
\begin{aligned}
&\mathscr{L}^{\mathrm{A}}(v_{1:T}, \theta_{1:T}, \eta_{1:T}) \\
&= \left\| u^1_{0,1:T} - y_{1:T} \right\|^2_{\Xi_{1:T}} + \left\| u^1_{0,1:T} \right\|^2_{C^1_{1:T}} + \log\det\left(2\pi C^1_{1:T}\right) \\
&\quad + \left\| u^2_{1,1:T} \right\|^2_{C^2_{1:T}} + \log\det\left(2\pi C^2_{1:T}\right) + \left\| u^3_{1,1:T} \right\|^2_{C^3_{1:T}} + \log\det\left(2\pi C^3_{1:T}\right) \\
&\quad + \lambda_1 \left\| \theta^1_{1:T} \right\|_1 + \left(\eta^1_{1:T}\right)^{\mathsf{T}} \left(\Phi^1 u^1_{0,1:T} - \theta^1_{1:T}\right) \\
&\quad + \lambda_2 \left\| \theta^2_{1:T} \right\|_1 + \left(\eta^2_{1:T}\right)^{\mathsf{T}} \left(\Phi^2 u^2_{1,1:T} - \theta^2_{1:T}\right) \\
&\quad + \lambda_3 \left\| \theta^3_{1:T} \right\|_1 + \left(\eta^3_{1:T}\right)^{\mathsf{T}} \left(\Phi^3 u^3_{1,1:T} - \theta^3_{1:T}\right) \\
&\quad + \frac{\rho_1}{2} \left\| \Phi^1 u^1_{0,1:T} - \theta^1_{1:T} \right\|^2_2 + \frac{\rho_2}{2} \left\| \Phi^2 u^2_{1,1:T} - \theta^2_{1:T} \right\|^2_2 + \frac{\rho_3}{2} \left\| \Phi^3 u^3_{1,1:T} - \theta^3_{1:T} \right\|^2_2,
\end{aligned}
\tag{4.57}
$$

where $\rho_1 > 0$, $\rho_2 > 0$, and $\rho_3 > 0$ are penalty parameters. The ADMM method works by generating a sequence of estimates $\{v^{(i)}_{1:T}, \theta^{(i)}_{1:T}, \eta^{(i)}_{1:T} : i = 0, 1, \ldots\}$ to iteratively approximate the optimal $\{v_{1:T}, \theta_{1:T}, \eta_{1:T}\}$ of Equation (4.57), as shown in the following algorithm.

Algorithm 4.34 (ADMM for R-DGP regression)**.** *Let $\{v^{(0)}_{1:T}, \theta^{(0)}_{1:T}, \eta^{(0)}_{1:T}\}$ be a given initial estimate. Then for $i = 0, 1, \ldots$, the ADMM algorithm updates the estimate by solving the following subproblems iteratively*

$$
\begin{aligned}
v^{(i+1)}_{1:T} = \operatorname*{arg\,min}_{v_{1:T}} &\left\| u^1_{0,1:T} - y_{1:T} \right\|^2_{\Xi_{1:T}} + \left\| u^1_{0,1:T} \right\|^2_{C^1_{1:T}} + \log\det\left(2\pi C^1_{1:T}\right) \\
&+ \left\| u^2_{1,1:T} \right\|^2_{C^2_{1:T}} + \log\det\left(2\pi C^2_{1:T}\right) + \left\| u^3_{1,1:T} \right\|^2_{C^3_{1:T}} + \log\det\left(2\pi C^3_{1:T}\right) \\
&+ \left(\eta^{1,(i)}_{1:T}\right)^{\mathsf{T}} \left(\Phi^1 u^1_{0,1:T} - \theta^{1,(i)}_{1:T}\right) + \frac{\rho_1}{2} \left\| \Phi^1 u^1_{0,1:T} - \theta^{1,(i)}_{1:T} \right\|^2_2 \\
&+ \left(\eta^{2,(i)}_{1:T}\right)^{\mathsf{T}} \left(\Phi^2 u^2_{1,1:T} - \theta^{2,(i)}_{1:T}\right) + \frac{\rho_2}{2} \left\| \Phi^2 u^2_{1,1:T} - \theta^{2,(i)}_{1:T} \right\|^2_2 \\
&+ \left(\eta^{3,(i)}_{1:T}\right)^{\mathsf{T}} \left(\Phi^3 u^3_{1,1:T} - \theta^{3,(i)}_{1:T}\right) + \frac{\rho_3}{2} \left\| \Phi^3 u^3_{1,1:T} - \theta^{3,(i)}_{1:T} \right\|^2_2,
\end{aligned}
\tag{4.58}
$$

$$
\begin{aligned}
\theta^{(i+1)}_{1:T} = \operatorname*{arg\,min}_{\theta_{1:T}} &\lambda_1 \left\| \theta^1_{1:T} \right\|_1 + \frac{\rho_1}{2} \left\| \Phi^1 u^{1,(i+1)}_{0,1:T} - \theta^1_{1:T} + \frac{1}{\rho_1} \eta^{1,(i)} \right\|^2_2 \\
&+ \lambda_2 \left\| \theta^2_{1:T} \right\|_1 + \frac{\rho_2}{2} \left\| \Phi^2 u^{2,(i+1)}_{1,1:T} - \theta^2_{1:T} + \frac{1}{\rho_2} \eta^{2,(i)} \right\|^2_2 \\
&+ \lambda_3 \left\| \theta^3_{1:T} \right\|_1 + \frac{\rho_3}{2} \left\| \Phi^3 u^{3,(i+1)}_{1,1:T} - \theta^3_{1:T} + \frac{1}{\rho_3} \eta^{3,(i)} \right\|^2_2,
\end{aligned}
\tag{4.59}
$$

and

$$\begin{aligned}
\eta_{1:T}^{1,(i+1)} &= \eta_{1:T}^{1,(i)} + \rho_1 \left(\Phi^1 u_{0,1:T}^{1,(i+1)} - \theta_{1:T}^{1,(i+1)}\right), \\
\eta_{1:T}^{2,(i+1)} &= \eta_{1:T}^{2,(i)} + \rho_2 \left(\Phi^2 u_{1,1:T}^{2,(i+1)} - \theta_{1:T}^{2,(i+1)}\right), \\
\eta_{1:T}^{3,(i+1)} &= \eta_{1:T}^{3,(i)} + \rho_3 \left(\Phi^3 u_{1,1:T}^{3,(i+1)} - \theta_{1:T}^{3,(i+1)}\right).
\end{aligned} \tag{4.60}$$

The subproblem in Equation (4.58) is a standard unconstrained optimisation problem which can be solved numerically by a vast number of non-linear optimisers. For a review of such optimisers, we refer the reader to Nocedal and Wright (2006). As for the subproblem in Equation (4.59), one can use the soft thresholding scheme in order to obtain a closed-form solution (Hastie et al., 2015; Boyd et al., 2011).

Convergence analysis of Algorithm 4.34

The goal is now to analyse whether the sequence generated by Algorithm 4.34 converges to a local minimum. For notational convenience we concatenate the objective variables in $v_{1:T}$ in a vector $\overline{v}_{1:T} \in \mathbb{R}^{3T}$ defined by $\overline{v}_{1:T} \coloneqq \begin{bmatrix} u_{0,1:T}^1 & u_{1,1:T}^2 & u_{1,1:T}^3 \end{bmatrix}^\mathsf{T}$.

We consider the following assumptions on the DGP and the minimisation problem parameters.

Assumption 4.35. *The covariance matrix $C_{1:T}^1$ is strictly positive definite. That is, $\lambda_{\min}(C_{1:T}^1)$ has a positive lower bound uniformly for all $u_{1,1:T}^2 \in \mathbb{R}^T$ and $u_{1,1:T}^3 \in \mathbb{R}^T$.*

Assumption 4.36. *The penalty parameters ρ_i and sparsity parameters Φ_i for $i = 1,2,3$ satisfy*

$$\frac{\rho_i}{2}\left(\lambda_{\min}(\Phi_i)\right)^2 - \frac{c_v}{2} \geq 0, \tag{4.61}$$

where the constant c_v is defined in Lemma 4.38.

Assumption 4.35 is an important prerequisite for Theorem 4.39 because the proof in Zhao et al. (2021c) requires that $(C_{1:T}^1)^{-1}$ be bounded so that the Lagrangian function in Equation (4.57) admits a lower bound independent of its arguments.

The constant c_v in Assumption 4.36 is a fixed number determined by $\mathscr{L}^{\mathrm{B}}$ (and therefore by the DGP model itself) and is independent of data $y_{1:T}$. This gives a lower bound on the free penalty parameters ρ_1, ρ_2, and ρ_3 for the problem to be well defined.

Remark 4.37. *Zhao et al. (2021c) considers the specific case of a non-stationary Matérn covariance function C_{NS} defined as per Equation* (4.19)*. Due to this choice, additional assumptions need to be introduced so as to proceed with the convergence analysis of the problem in Equation* (4.55)*.*

Lemma 4.38 (Lipschitz condition)**.** *Suppose that there exists a constant $c_v > 0$ such that the norm $\|\mathrm{H}_{\overline{v}_{1:T}} \mathscr{L}^B\|_2 \leq c_v$ for all $\overline{v}_{1:T} \in \mathbb{R}^{3T}$. Then for every two vectors $\overline{v}^1_{1:T}, \overline{v}^2_{1:T} \in \mathbb{R}^{3T}$,*

$$\begin{aligned} &\left| \mathscr{L}^{\mathrm{B}}\left(\overline{v}^1_{1:T}\right) - \mathscr{L}^{\mathrm{B}}\left(\overline{v}^2_{1:T}\right) - \left(\overline{v}^1_{1:T} - \overline{v}^2_{1:T}\right)^{\mathsf{T}} \nabla_{\overline{v}_{1:T}} \mathscr{L}^{\mathrm{B}}\left(\overline{v}^2_{1:T}\right) \right| \\ &\quad \leq \frac{c_v}{2} \left\| \overline{v}^1_{1:T} - \overline{v}^2_{1:T} \right\|_2^2. \end{aligned} \tag{4.62}$$

Proof. See, Lemma 1.2.2 and 1.2.3 in Nesterov (2004). □

Theorem 4.39. *Suppose that Assumptions 4.35 and 4.36 hold. Further assume that the subproblem in Equation* (4.58) *has a stationary point. Then the sequence $\left\{ v^{(i)}_{1:T}, \theta^{(i)}_{1:T}, \eta^{(i)}_{1:T} : i = 0, 1, \dots \right\}$ generated by Algorithm 4.34 converges to a local minimum.*

Proof. The key is to prove that the sequence $\left\{ \mathscr{L}^{\mathrm{A}}\left(v^{(i)}_{1:T}, \theta^{(i)}_{1:T}, \eta^{(i)}_{1:T}\right) : i = 0, 1, \dots \right\}$ is non-increasing and lower bounded over $i = 0, 1, \dots$ Using the convexity of subproblem (4.59) we can then prove the convergence of Algorithm 4.34 (see, e.g., Boyd and Vandenberghe, 2004; Nesterov, 2018). We refer the reader to Zhao et al. (2021c) for details. □

Regularised SS-DGP regression

We can also derive the state-space versions of regularised DGPs. However, the resulting method turns out to be very similar to that of batch DGPs. We therefore only sketch out the basic idea in this section, and refer to Zhao et al. (2021c) for details.

Consider the state-space representation of the DGP defined in Equation (4.50). The first step is to derive the state-space version of the MAP objective function in Equation (4.53). One approach is to discretise the SDE as in Equation (4.28), then we can factorise the SDE prior density over $t_1, \dots, t_T$. As shown in Zhao et al. (2021c), the state-space (approximate) MAP objective function reads

$$\begin{aligned} \mathscr{L}^{\mathrm{S}}(v_{1:T}) = v_0^{\mathsf{T}} P_0^{-1} v_0 + \sum_{k=1}^{T} \Big[&\| y_k - H v_k \|^2_{\Xi_k} + \| v_k - f_{k-1}(v_{k-1}) \|^2_{Q_{k-1}(v_{k-1})} \\ &+ \log \det \left(2\pi Q_{k-1}(v_{k-1}) \right) \Big], \end{aligned} \tag{4.63}$$

where the matrix H selects U^1_0 from V. Now, let Ψ^1, Ψ^2, and Ψ^3 be three suitable matrices that, respectively, select the components U^1_0, U^2_1, and U^3_1 from V. We can now introduce the regularisation term

$$\mathscr{L}^{\mathrm{S-REG}}(v_{1:T}) = \sum_{k=0}^{T} \sum_{i=1}^{3} \lambda_i \| \Psi^i v_k \|_1. \tag{4.64}$$

One can then analogously derive the augmented Lagrangian function and the corresponding ADMM algorithm in their state-space versions.

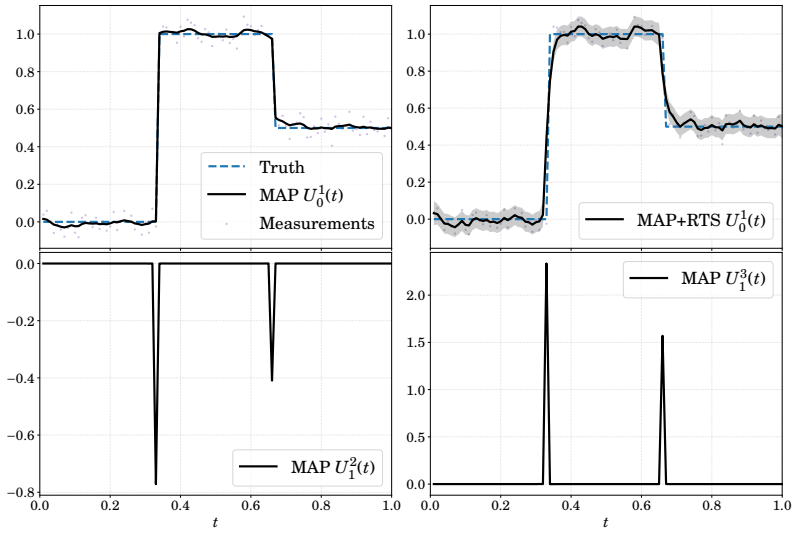


Figure 4.9. Regularised SS-DGP regression on a rectangular signal. The uncertainty is quantified by using Equation (4.65) and an RTS smoother.

Remark 4.40. *The computational complexities of $\mathscr{L}^{\mathrm{S}}$ and $\mathscr{L}^{\mathrm{B}}$ are linear and cubic with respect to T, respectively. Hence, the state-space version is significantly computationally cheaper than the batch version when the number of measurements is large.*

Remark 4.41. *$\mathscr{L}^{\mathrm{B}}$ and $\mathscr{L}^{\mathrm{S}}$ are generally not equal, since discretisations of SS-DGPs often involve approximations.*

Uncertainty quantification

The regularised DGP regression method presented in this section is rooted in the MAP framework, which provides point estimates of the quantities at hand, and ignores the uncertainty in the solution. This can be partially remedied by using, for example, Laplace's method to approximate $p_{V_{1:T} \mid Y_{1:T}}(v_{1:T} \mid y_{1:T})$ around the MAP estimate with a Gaussian density (Bishop, 2006). However, computing the Hessian (of dimension $T \times T$) can be computationally intensive and limits the applicability for high dimensional problems such as ours.

Another solution is to solve the subproblems of ADMM by using Bayesian solvers instead of deterministic optimisers, resulting in an estimate of the uncertainty in the form of a posterior distribution on the solution. For instance, it is known that the iterated extended Kalman smoother is in some sense equivalent to the Gauss–Newton method (Bell, 1994; Särkkä and Svensson, 2020), and Gao et al. (2019a) showed that this connection could be extended to the ADMM method. For a review of these, we refer the reader to the discussion in Gao (2020).

However, if we are only interested in the marginal posterior density $p_{U_{0,1:T}^1 \mid Y_{1:T}}\left(u_{0,1:T}^1 \mid y_{1:T}\right)$ instead of the full density $p_{V_{1:T} \mid Y_{1:T}}(v_{1:T} \mid y_{1:T})$, then

we can leverage the hierarchical nature of DGPs to approximate the marginal density efficiently. In order to do so, we can write the approximation

$$\begin{aligned} & p_{U^1_{0,1:T} \mid U^2_{1,1:T}, U^3_{1,1:T}, Y_{1:T}}\left(u^1_{0,1:T} \mid u^2_{1,1:T}, u^3_{1,1:T}, y_{1:T}\right) \\ & \approx p_{U^1_{0,1:T} \mid Y_{1:T}}\left(u^1_{0,1:T} \mid u^{2,\star}_{1,1:T}, u^{3,\star}_{1,1:T}, y_{1:T}\right), \end{aligned} \tag{4.65}$$

where $u^{2,\star}_{1,1:T}, u^{3,\star}_{1,1:T}$ stand for the MAP estimates of $U^2_{1,1:T} \mid y_{1:T}$ and $U^3_{1,1:T} \mid y_{1:T}$. Afterwards, computing $p_{U^1_{0,1:T} \mid Y_{1:T}}\left(u^1_{0,1:T} \mid y_{1:T}\right)$ simply consists in solving a standard GP regression problem, which can be obtained in closed form (Zhao et al., 2021c).

Figure 4.9 illustrates such an example of regularised SS-DGP, where we set the sparsity inducing matrices to be identity matrices (Zhao et al., 2021c). The latent states U^2_1 and U^3_1 exhibit spiking behaviours, being almost zero except at the two discontinuities.

5. Applications

In this chapter, we present the experimental results in Publications IV, III, V, II, and VI. These works are mainly concerned with the applications of state-space (deep) GPs. Specifically, in Section 5.1 we show how to use the SS-GP regression method to estimate unknown drift functions in SDEs. Similarly, under that same state-space framework, in Section 5.2 we show how to estimate the posterior distributions of the Fourier coefficients of signals. Sections 5.3 and 5.4 illustrate how SS-DGPs can be used to model real-world signals, such as gravitational waves, accelerometer recordings of human motion, and maritime vessel trajectories.

5.1 Drift estimation in stochastic differential equations

Consider a scalar-valued stochastic process $X : \mathbb{T} \to \mathbb{R}$ governed by a stochastic differential equation

$$\mathrm{d}X(t) = a(X(t))\,\mathrm{d}t + b\,\mathrm{d}W(t), \quad X(t_0) = X_0, \tag{5.1}$$

where $b \in \mathbb{R}$ is a constant, $W : \mathbb{T} \to \mathbb{R}$ is a Wiener process, and $a : \mathbb{R} \to \mathbb{R}$ is an *unknown* drift function. Suppose that we have measurement random variables $X(t_1), X(t_2), \ldots, X(t_T)$ of X at time instances $t_1, t_2, \ldots, t_T \in \mathbb{T}$, the goal is to estimate the drift function a from these measurements.

One way to proceed is to assume a parametric form of function $a = a_\vartheta(\cdot)$ and estimate its parameters ϑ by using, for example, maximum likelihood estimation (Dacunha-Castelle and Florens-Zmirou, 1986; Yoshida, 1992; Kessler, 1997; Aït-Sahalia, 2003) or Monte Carlo methods (Roberts and Stramer, 2001; Beskos et al., 2006).

In this chapter, we are mainly concerned with the GP regression approach for estimating the unknown a (Papaspiliopoulos et al., 2012; Ruttor et al., 2013; Garcìa et al., 2017; Batz et al., 2018; Opper, 2019). The key idea of this approach is to assume that the unknown drift function is distributed according to a GP, that is

$$a(x) \sim \mathrm{GP}(0, C(x, x')). \tag{5.2}$$

Having at our disposal measurements $X(t_1), X(t_2), \ldots, X(t_T)$ observed directly from SDE (5.1), we can formulate the problem of estimating a as a GP regression problem. In order to do so, we discretise the SDE in Equation (5.1) and thereupon define the measurement model as

$$Y_k := X(t_k) - X(t_{k-1}) = \check{f}_{k-1}(X(t_{k-1})) + \check{q}_{k-1}(X(t_{k-1})) \tag{5.3}$$

for $k = 1, 2, \ldots, T$, where the function $\check{f}_{k-1} \colon \mathbb{R} \to \mathbb{R}$ and the random variable $\check{q}_{k-1}$ represent the exact discretisation of X at t_k from t_{k-1}. We write the GP regression model for estimating the drift function by

$$\begin{aligned} a(x) &\sim \mathrm{GP}(0, C(x, x')), \\ Y_k &= \check{f}_{k-1}(X_{k-1}) + \check{q}_{k-1}(X_{k-1}). \end{aligned} \tag{5.4}$$

The goal now is to estimate the posterior density of $a(x)$ for all $x \in \mathbb{R}$ from a set of data $y_{1:T} = \{x_k - x_{k-1} \colon k = 1, 2, \ldots, T\}$.

However, the exact discretisation of non-linear SDEs is rarely possible. In practice, we often have to approximate $\check{f}_{k-1}$ and $\check{q}_{k-1}$ by using, for instance, Euler–Maruyama scheme, Milstein's method, or more generally Itô–Taylor expansions (Kloeden and Platen, 1992). As an example, application of the Euler–Maruyama method to Equation (5.1) gives

$$\begin{aligned} \check{f}_{k-1} &\approx a(x)\,\Delta t_k, \\ \check{q}_{k-1} &\approx b\,\delta_k, \end{aligned} \tag{5.5}$$

where $\Delta t_k := t_k - t_{k-1}$ and $\delta_k \sim \mathrm{N}(0, \Delta t_k)$.

However, the discretisation by the Euler–Maruyama scheme can sometimes be crude, especially when the discretisation step is relatively large, making the measurement representation obtained from it inaccurate. Zhao et al. (2020b) show that if the prior of a is chosen of certain regularities, it is possible to leverage high-order Itô–Taylor expansions in order to discretise the SDE with higher accuracy. As an example, suppose that the GP prior a is twice-differentiable almost surely. Then, the Itô–Taylor strong order 1.5 (Itô-1.5) method (Kloeden and Platen, 1992) gives

$$\begin{aligned} \check{f}_{k-1}(x) &\approx a(x)\,\Delta t_k + \frac{1}{2}\left(\frac{\mathrm{d}a}{\mathrm{d}x}(x)\,a(x) + \frac{1}{2}\frac{\mathrm{d}^2 a}{\mathrm{d}x^2}(x)\,b^2\right)\Delta t_k^2, \\ \check{q}_{k-1}(x) &\approx b\,\delta_{1,k} + \frac{\mathrm{d}a}{\mathrm{d}x}(x)\,b\,\delta_{2,k}, \end{aligned} \tag{5.6}$$

where

$$\begin{bmatrix} \delta_{1,k} \\ \delta_{2,k} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{(\Delta t_k)^3}{3} & \frac{(\Delta t_k)^2}{2} \\ \frac{(\Delta t_k)^2}{2} & \Delta t_k \end{bmatrix}\right). \tag{5.7}$$

Indeed, using a higher order Itô–Taylor expansion can lead to a better measurement representation, however, this in turn requires more computations and limits the choice of the prior model. It is also worth mentioning

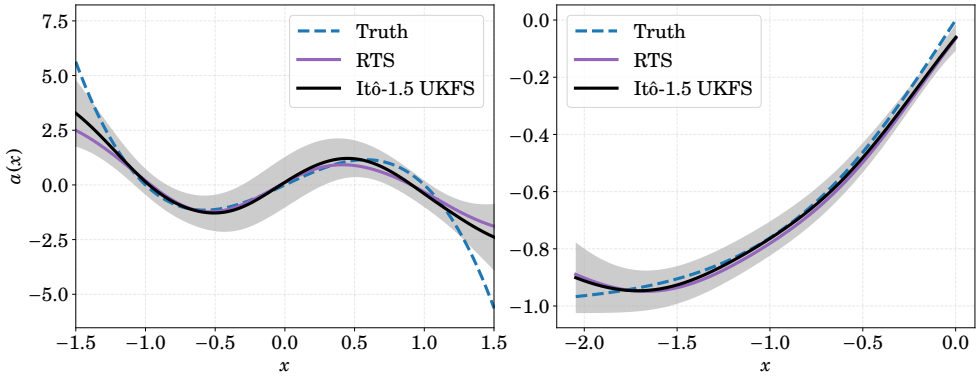


Figure 5.1. Estimation of drift functions $a(x) = 3(x - x^3)$ (left) and $a(x) = \tanh(x)$ (right) by Zhao et al. (2020b). UKFS stands for unscented Kalman filter and RTS smoother (UKFS). Shaded area stands for 0.95 confidence interval associated with the UKFS estimation.

that if one uses the approximations of high order Itô–Taylor expansions – such as the one in Equation (5.6) – the resulting measurement representation in the GP regression model (5.4) is no longer linear with respect to a. Consequently, the GP regression solution may not admit a closed-form solution.

One problem of this GP regression-based drift estimation approach is that the computation can be demanding if the number of measurements T is large. Moreover, if the measurements are densely located then the covariance matrices used in GP regression may be numerically close to singular. These two issues are already discussed in Introduction and Section 4.2. In addition, the GP regression model is not amenable to high order Itô–Taylor expansions, as these expansions result in non-linear measurement representations and require to compute the derivatives of a up to a certain order.

Zhao et al. (2020b) address the problems above by considering solving the GP regression problem in Equation (5.4) under the state-space framework. More precisely, they put an SS-GP prior over the unknown a instead of a standard batch GP. The main benefit of doing so for this application is that the SS-GP regression solvers are computationally more efficient for large-scale measurements compared to the standard batch GP regression (see, Introduction and Section 4.2). Moreover, in order to use high order Itô–Taylor expansions, Zhao et al. (2020b) consider putting SS-GP priors over a of the Matérn family, so that the derivatives of a naturally appear as the state components of a (see, Section 4.6). In this way, computing the covariance matrices of the derivatives of a is no longer needed.

Remark 5.1. *Note that the SS-GP approach requires to treat* $X(t_1), X(t_2), \ldots, X(t_T)$ *as time variables and sort their data* $x_{1:T} = \{x_1, x_2, \ldots, x_T\}$ *in temporal order.*

In Figure 5.1, we show a representative result from Zhao et al. (2020b),

where the SS-GP approach is employed to approximate the drift functions of two SDEs. In particular, the solutions are obtained by using the Itô-1.5 discretisation, and an unscented Kalman filter and an RTS smoother. For more details regarding the experiments the reader is referred to Zhao et al. (2020b).

5.2 Probabilistic spectro-temporal signal analysis

Let $z\colon \mathbb{T} \to \mathbb{R}$ be a periodic signal. In signal processing, it is often of interest to approximate the signal by Fourier expansions of the form

$$z(t) \approx \alpha_0 + \sum_{n=1}^{N} \left[\alpha_n \cos(2\pi \mathring{f}_n t) + \beta_n \sin(2\pi \mathring{f}_n t)\right], \tag{5.8}$$

where $\{\mathring{f}_n : n = 1,2,\ldots,N\}$ stand for the frequency components, and N is a given expansion order. When z satisfies certain conditions (Katznelson, 2004), the representation in the equation above converges as $N \to \infty$ (in various modes).

Let use denote $y_k := y(t_k)$ and suppose that we have a set of measurement data $y_{1:T} = \{y_k : k = 1,2,\ldots,T\}$ of the signal at time instances $t_1, t_2, \ldots, t_T \in \mathbb{T}$. In order to quantify the truncation and measurement errors, we introduce Gaussian random variables $\xi_k \sim \mathrm{N}(0, \Xi_k)$ for $k = 1,2,\ldots,T$ and let

$$Y_k = \alpha_0 + \sum_{n=1}^{N} \left[\alpha_n \cos(2\pi \mathring{f}_n t_k) + \beta_n \sin(2\pi \mathring{f}_n t_k)\right] + \xi_k \tag{5.9}$$

represent the random measurements of z at t_k. The goal now is to estimate the coefficients $\{\alpha_0, \alpha_n, \beta_n : n = 1,2,\ldots,N\}$ from the data $y_{1:T}$. We call this problem the *spectro-temporal estimation* problem.

One way to proceed is by using the MLE method (Bretthorst, 1988), but Qi et al. (2002); Zhao et al. (2018, 2020a) show that we can also consider this spectro-temporal estimation problem as a GP regression problem. More precisely, the modelling assumption is that

$$\begin{aligned} \alpha_0(t) &\sim \mathrm{GP}\big(0, C_\alpha^0(t,t')\big), \\ \alpha_n(t) &\sim \mathrm{GP}\big(0, C_\alpha^n(t,t')\big), \\ \beta_n(t) &\sim \mathrm{GP}\big(0, C_\beta^n(t,t')\big), \end{aligned} \tag{5.10}$$

for $n = 1,2,\ldots,N$, and that the measurements follow

$$Y_k = \alpha_0(t_k) + \sum_{n=1}^{N} \left[\alpha_n(t_k) \cos(2\pi \mathring{f}_n t_k) + \beta_n(t_k) \sin(2\pi \mathring{f}_n t_k)\right] + \xi_k,$$

for $k = 1,2,\ldots,T$. This results in a standard GP regression problem therefore, the posterior distribution of coefficients $\{\alpha_0, \alpha_n, \beta_n : n = 1,2,\ldots,N\}$ have

a close-form solution. However, solving this GP regression problem is, in practice, infeasible when the expansion order N and the number of measurements T are large. This is due to the fact that one needs to compute $2N+1$ covariance matrices of dimension $T \times T$ and compute their inverse.

Zhao et al. (2018) propose to solve this spectro-temporal GP regression problem under the state-space framework, that is, by replacing the GP priors in Equation (5.10) with their SDE representations. Since SS-GPs have already been extensively discussed in previous sections, we omit the resulting state-space spectro-temporal estimation formulations. However, the details can be found in Section 4.2 and in Zhao et al. (2018).

The computational cost of the state-space spectro-temporal estimation method is substantially cheaper than that of standard batch GP methods. Indeed, Kalman filters and smoothers only need to compute one E-dimensional covariance matrix at each time step (see, Algorithm 2.10) instead of those required by batch GP methods. The dimension E is equal to the sum of all the state dimensions of the SS-GPs $\{\alpha_0, \alpha_n, \beta_n : n = 1,2,\ldots,N\}$.

Zhao et al. (2020a) further extend the state-space spectro-temporal estimation method by putting quasi-periodic SDE priors (Solin and Särkkä, 2014) over the Fourier coefficients instead of the Ornstein–Uhlenbeck SDE priors used in Zhao et al. (2018). This consideration generates a time-invariant version of the measurement model in Equation (5.9), thus, one can apply steady-state Kalman filters and smoothers (SS-KFSs) in order to achieve lower computational costs. The computational cost is further reduced because SS-KFSs do not need to compute the E-dimensional covariances of the state in their filtering and smoothing loops. Instead, the state covariances in SS-KFSs are replaced by a pre-computed steady covariance matrix obtained as the solution of its discrete algebraic Riccati equation (DARE). Moreover, solving the DARE is independent of data/measurements, which is especially useful when the model is known or fixed. However, SS-KFSs may not always be computationally efficient when $N \gg T$, since solving an E-dimensional DARE can be demanding when E is large.

Zhao et al. (2018, 2020a) show that the state-space spectro-temporal estimation method can be a useful feature extraction mechanism for detecting atrial fibrillation from electrocardiogram signals. More specifically, the spectro-temporal method estimates the spectrogram images of atrial fibrillation signals. These images are then fed to a deep convolutional neural network classifier which is tasked with recognising atrial fibrillation manifestations.

Since the measurement noises $\{\xi_k : k = 1,2,\ldots,T\}$ in Equation (5.9) encode the truncation and measurement errors, it is also of interest to estimate them. This is done in Gao et al. (2019b), where the variances Ξ_k of ξ_k for $k = 1,2,\ldots,T$ are estimated under the alternating direction method of multipliers.

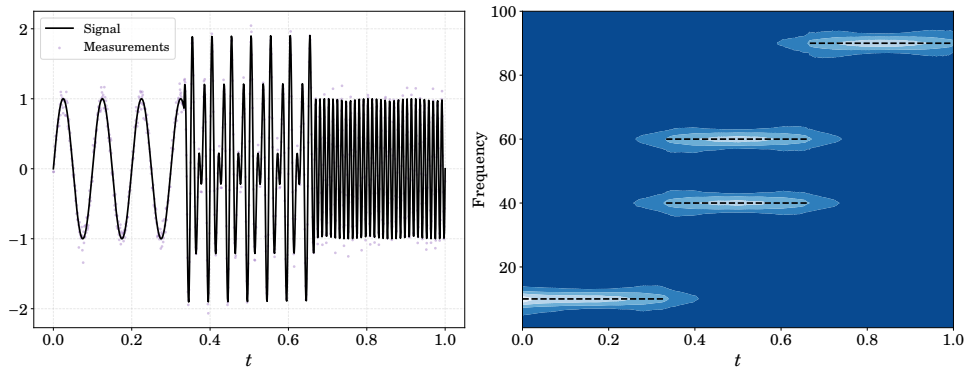


Figure 5.2. Spectrogram (right, contour plot) of a sinusoidal signal (left) generated by Kalman filtering and RTS smoothing using the method in Section 5.2. Dashed black lines (right) stand for the ground truth frequencies.

Figure 5.2 illustrates an example of using the state-space spectro-temporal estimation method to estimate the spectrogram of a sinusoidal signal with multiple frequency bands.

5.3 Signal modelling with SS-DGPs

In this section, we apply SS-DGPs for modelling gravitational waves and human motion (i.e., acceleration). We consider these as SS-DGP regression problems, where the measurement models are assumed to be linear with respect to the SS-DGPs with additive Gaussian noises. As for their priors, we chose the Matérn $\nu = 3/2$ SS-DGP in Example 4.17, except that the parent GPs U_1^2 and U_1^3 use the Matérn $\nu = 1/2$ representation.

Modelling gravitational waves

Gravitational waves are curvatures of spacetime caused by the movement of objects with mass (Maggiore, 2008). Since the time Albert Einstein predicted the existence of gravitational waves theoretically from a linearised field equation in 1916 (Einstein and Rosen, 1937; Hill et al., 2017), much effort has been done to observe their presence (Blair, 1991). In 2015, the laser interferometer gravitational-wave observatory (LIGO) team first observed a gravitational wave from the merging of a black hole binary (event GW150914, Abbott et al., 2016). This wave/signal is challenging for standard GPs to fit because the frequency of the signal changes over time. It is then of our interest to see if SS-DGPs can fit this gravitational wave signal.

Figure 5.3 plots the SS-DGP fit for the gravitational wave observed in the event GW150914. In the same figure, we also show the fit from a Matérn $\nu = 3/2$ GP as well as a waveform (which is regarded as the ground truth) computed from the numerical relativity (purple dashed lines) for

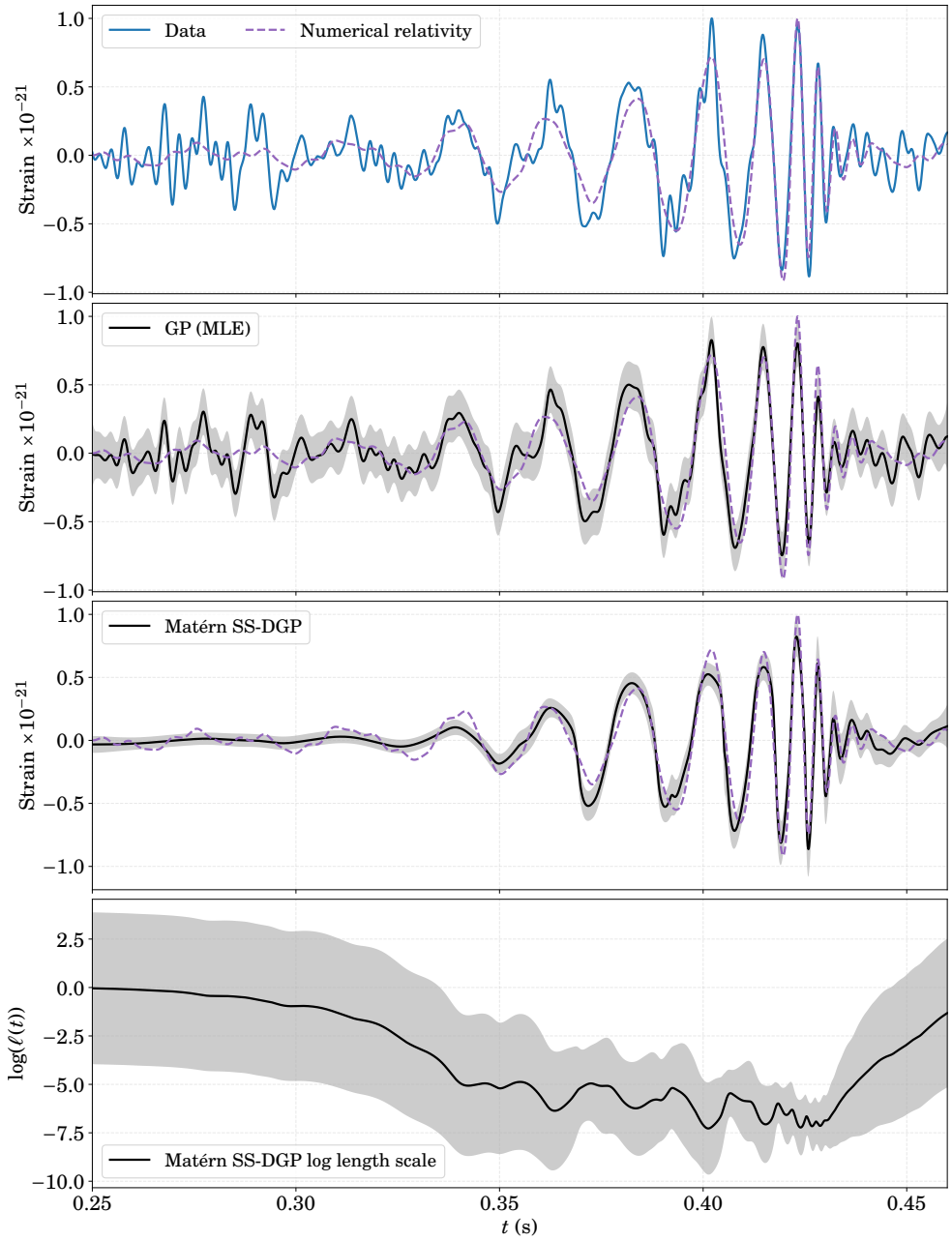


Figure 5.3. Matérn $\nu = 3/2$ SS-DGP regression (solved by cubature Kalman filter and smoother) for the gravitational wave in event GW150914 (Hanford, Washington). The shaded area stands for 0.95 confidence interval. Details about the data can be found in Zhao et al. (2021a).

comparison. Details about the experiment and data are found in Zhao et al. (2021a).

Figure 5.3 shows that the GP fails to give a reasonable fit to the gravitational wave because the GP over-adapts the high-frequency section of the signal around 0.4 s. On the contrary, the SS-DGP does not have such a problem, and the fit is closer to the numerical relativity waveform compared that of the GP. Moreover, the estimated length scale (in log transformation)

can interpret the data in the sense that the length scale value decreases as the signal frequency increases.

Modelling human motion

We apply the regularised SS-DGP (R-SS-DGP) presented in Section 4.9 to fit an accelerometer recording of human motion. The reason for using R-SS-DGP here is that the recording (see, the first row of Figure 5.4) is found to have some sharp changes and artefacts. Hence, we aim at testing if we can use sparse length scale and magnitude to describe such data. The collection of accelerometer recordings and the experiment settings are detailed in Hostettler et al. (2018) and Zhao et al. (2021c), respectively.

A demonstrative result is shown in Figure 5.4. We see that the fit of R-SS-DGP is smoother than that of GP. Moreover, the posterior variance of R-SS-DGP is also found to be reasonably smaller than GP. It is also evidenced from the figure that the GP does not handle the artefacts well, for example, around times $t = 55$ s and 62 s. Finally, we find that the learnt length scale and magnitude (in log transformation) are sparse, and that they can respond sharply to the abrupt signal changes and artefacts.

5.4 Maritime situational awareness

Another area of applications of (deep) GPs is autonomous maritime navigation. In Thombre et al. (2020), we present a literature review on the sensor technology and machine learning methods for autonomous vessel navigation. In particular, we show that GP-based methods are able to analyse ship trajectories (Rong et al., 2019), detect navigation abnormality (Kowalska and Peel, 2012; Smith et al., 2014), and detect/classify vessels (Xiao et al., 2017).

In Figure 5.5, we present an example for fitting an automatic identification system (AIS) recording by using an R-SS-DGP. The recording is taken from MS Finlandia (Helsinki–Tallinn) by Fleetrange Oy on December 10, 2020. We see from the figure that the fit of R-SS-DGP is smoother than that of GP. Moreover, the learnt length scale and magnitude parameters are flat and jump at the acceleration/deceleration points.

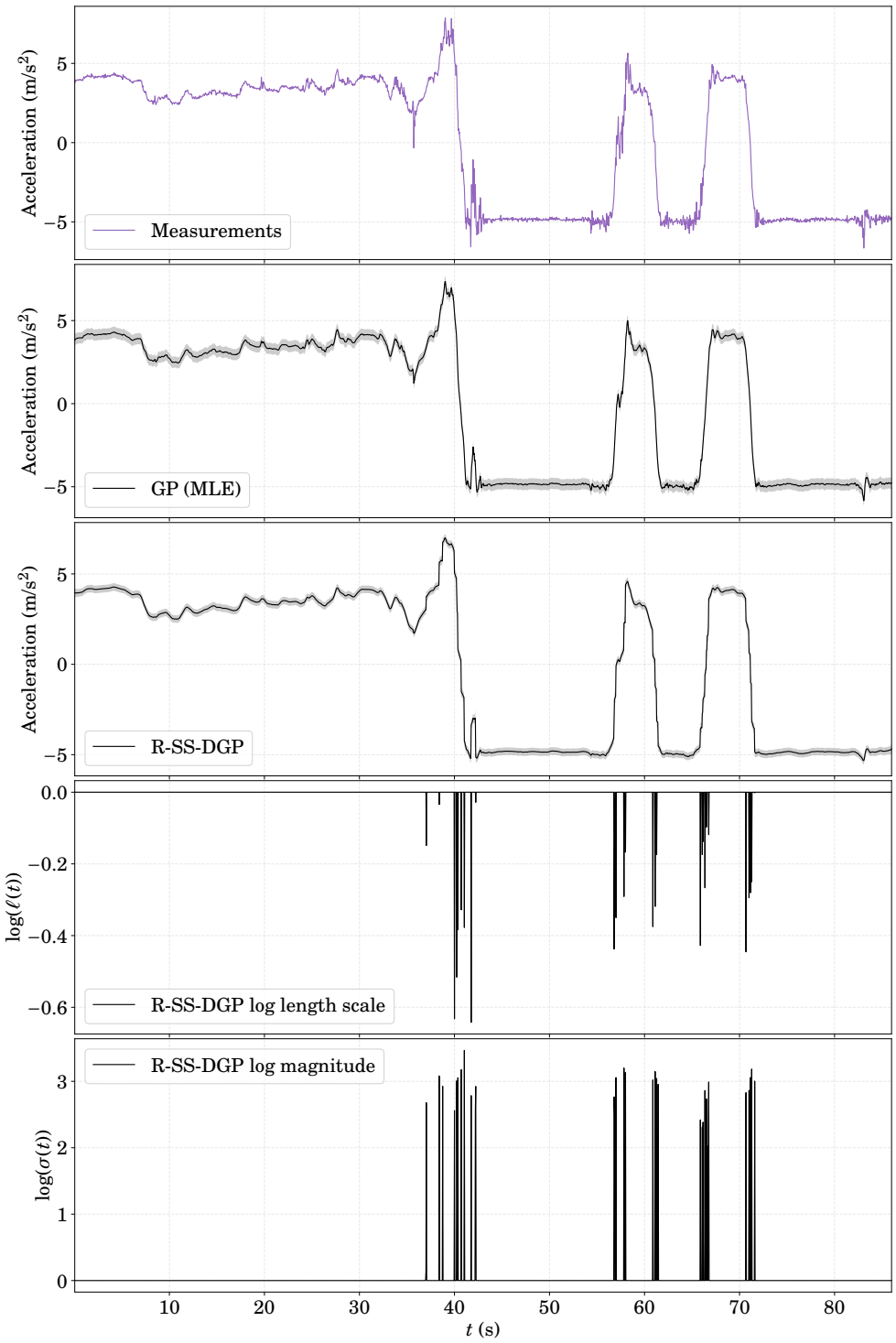


Figure 5.4. Human motion modelling with an R-SS-DGP. The GP here uses a Matérn $\nu = 3/2$ covariance function. Shaded area stands for 0.95 confidence interval.

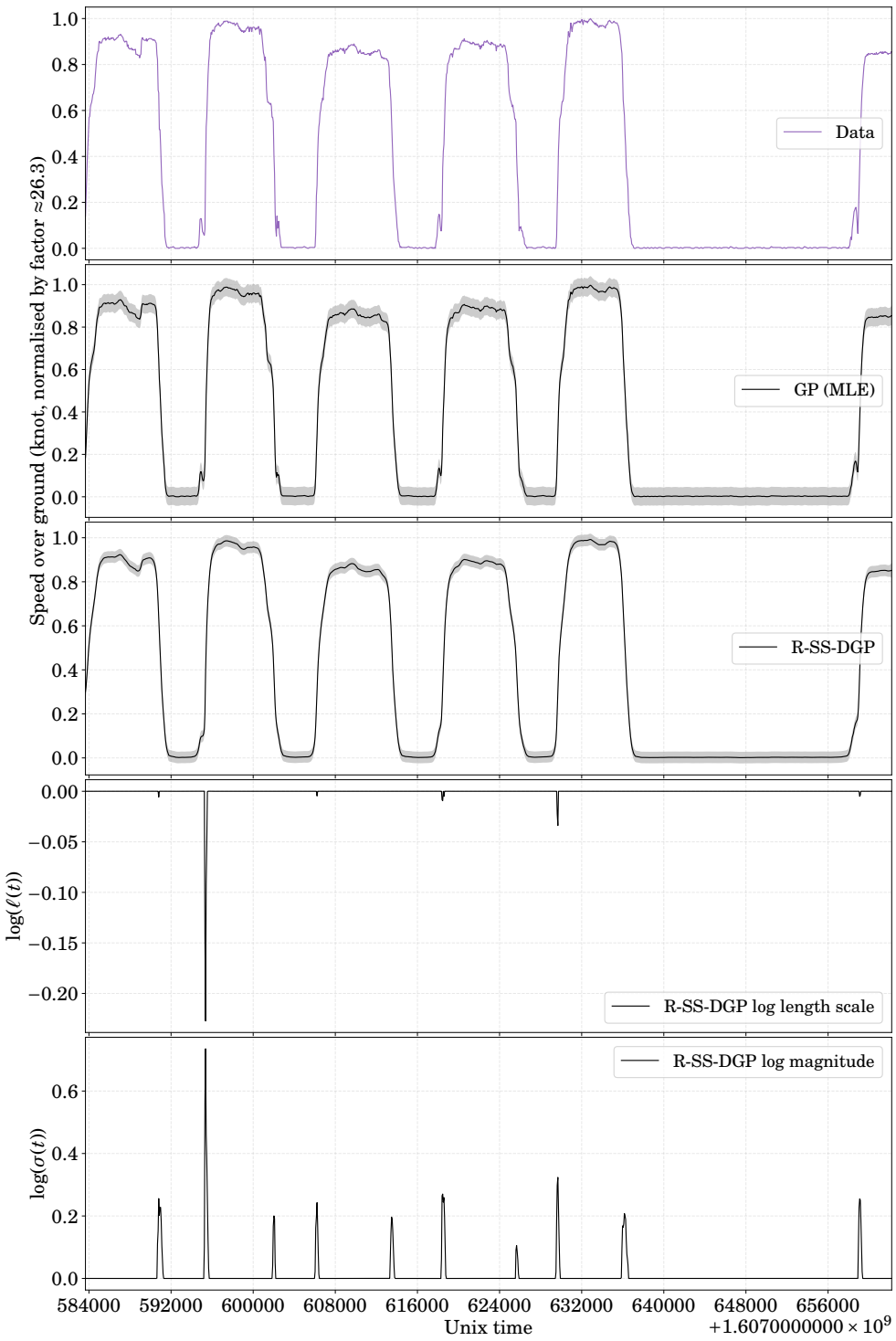


Figure 5.5. Modelling AIS recording (speed over ground) of MS Finlandia with an R-SS-DGP. The GP here uses a Matérn $\nu = 3/2$ covariance function. Shaded area stands for 0.95 confidence interval.

6. Summary and discussion

In this chapter we present a concise summary of Publications I–VII as well as discussion on a few unsolved problems and possible future extensions.

6.1 Summary of publications

This section briefly summaries the contributions of Publications I–VII and highlights their significances.

Publication I (Chapter 3)

This paper proposes a new class of non-linear continuous-discrete Gaussian filters and smoothers by using the Taylor moment expansion (TME) scheme to predict the means and covariances from SDEs. The main significance of this paper is that the TME method can provide asymptotically exact solutions of the predictive mean and covariances required in the Gaussian filtering and smoothing steps. Secondly, the paper analyses the positive definiteness of TME covariance approximations and thereupon presents a few sufficient conditions to guarantee the positive definiteness. Lastly, the paper analyses the stability of TME Gaussian filters.

Publication II (Chapter 4)

This paper introduces state-space representations of a class of deep Gaussian processes (DGPs). More specifically, the paper defines DGPs as vector-valued stochastic processes over collections of conditional GPs, thereupon, the paper represents DGPs in hierarchical systems of the SDE representations of their conditional GPs. The main significance of this paper is that the resulting state-space DGPs (SS-DGPs) are Markov processes, so that the SS-DGP regression problem is computationally cheap (i.e., linear with respect to the number of measurements) by using continuous-discrete filtering and smoothing methods. Secondly, the paper identifies that for a

certain class of SS-DGPs the Gaussian filtering and smoothing methods fail to learn the posterior distributions of their state components. Finally, the paper features a real application of SS-DGPs in modelling a gravitational wave signal.

Publication III (Section 5.2)

This paper is an extension of Publication V. In particular, the quasi-periodic SDEs are used to model the Fourier coefficients instead of the Ornstein–Uhlenbeck ones used in Publication V. This consideration leads to state-space models for which the measurement representations are time-invariant therefore, one can use steady-state Kalman filters and smoothers to solve the spectro-temporal estimation problem with lower computational cost compared to Publication V. This paper also expands the experiments for atrial fibrillation detection by taking into account more classifiers.

Publication IV (Section 5.1)

This paper is concerned with the state-space GP approach for estimating unknown drift functions of SDEs from partially observed trajectories. This approach is significant mainly in terms of computation, as the computational complexity scales linearly in the number of measurements. In addition, the state-space GP approach allows for using high-order Itô–Taylor expansions in order to give accurate SDE discretisations without the necessity to compute the covariance matrices of the derivatives of the GP prior.

Publication V (Section 5.2)

This paper introduces a state-space probabilistic spectro-temporal estimation method and thereupon applies the method for detecting atrial fibrillation from electrocardiogram signals. The so-called probabilistic spectro-temporal estimation is a GP regression-based model for estimating the coefficients of Fourier expansions. The main significance of this paper is that the state-space framework allows for dealing with large sets of measurements and high-order Fourier expansions. Also, the combination of the spectro-temporal estimation method and deep convolutional neural networks shows efficacy for classifying a class of electrocardiogram signals.

Publication VI (Section 5.4)

This paper reviews sensor technologies and machine learning methods for autonomous maritime vessel navigation. In particular, the paper lists and reviews a number of studies that use deep learning and GP methods

for vessel trajectory analysis, ship detection and classification, and ship tracking. The paper also features a ship detection example by using a deep convolutional neural network.

Publication VII (Section 4.9)

This paper solves L^1-regularised DGP regression problems under the alternating direction method of multipliers (ADMM) framework. The significance of this paper is that one can introduce regularisation (e.g., sparseness or total variation) at any level of the DGP component hierarchy. Secondly, the paper provides a general framework that allows for regularising both batch and state-space DGPs. Finally, the paper presents a convergence analysis for the proposed ADMM solution of L^1-regularised DGP regression problems.

6.2 Discussion

Finally, we end this thesis with discussion on some unsolved problems and possible future extensions.

Positive definiteness analysis for high-order and high-dimensional TME covariance approximation

Theorem 3.5 provides a sufficient condition to guarantee the positive definiteness of TME covariance approximations. However, the use of Theorem 3.5 soon becomes infeasible as the expansion order M and the state dimension d grow large. In practice, it can be easier to check the positive definiteness numerically when d is small.

Practical implementation of TME

A practical challenge with implementing TME consists in the presence of derivative terms in $\mathscr{A}$ (see, Equation (3.5)). This in turn implies that the iterated generator $\mathscr{A}^M$ further requires the computation of derivatives of the SDE coefficients up to order M. While the derivatives of $\mathscr{A}$ are easily computed by hand, the derivatives in $\mathscr{A}^M$ require more consideration as they involve numerous applications of the chain rule, not to mention the multidimensional operator $\overline{\mathscr{A}}$ in Remark 3.3.

While in our current implementation we chose to use symbolic differentiation (for ease of implementation as well as portability across languages), several things can be said against using it. Symbolic differentiation explicitly computes full Jacobians, where only vector-Jacobian/Jacobian-vector products would be necessary. This induces an unnecessary overhead that grows with the dimension of the problem. Also, symbolic differentiation is

usually independent of the philosophy of modern differentiable programming frameworks and the optimisation for parallelisable hardware (e.g., GPUs), hence they may incur a loss of performance on these.

Automatic differentiation tools, for instance, TensorFlow and JaX are amenable to computing the derivatives in $\overline{\mathscr{A}}$. Furthermore, they provide efficient computations for Jacobian-vector/vector-Jacobian products. We hence argue that these tools are worthwhile for performance improvement in the future[1].

Generalisation of the identifiability analysis

The identifiability analysis in Section 4.8 is limited to SS-DGPs for which the GP elements are one-dimensional. This dimension assumption is used in order to derive Equation (4.42) in closed-form. However, it is of interest to see whether we can generalise Lemma 4.25 for SS-DGPs that have multidimensional GP elements.

The abstract Gaussian filter in Algorithm 4.30 assumes that the prediction steps are done exactly. However, this assumption may not always be realistic because Gaussian filters often involve numerical integrations to predict through SDEs, for example, by using sigma-point methods. Hence, it is important to verify if Lemma 4.25 still holds when one computes the filtering predictions by some numerical means.

Spatio-temporal SS-DGPs

SS-DGPs are stochastic processes defined on temporal domains. In order to model spatio-temporal data, it is necessary to generalise SS-DGPs to take values in infinite-dimensional spaces (Prato and Zabczyk, 2014). A path for this generalisation is to leverage the stochastic partial differential equation (SPDE) representations of spatio-temporal GPs. To see this, let us consider an $\mathbb{H}$-valued stochastic process $U\colon \mathbb{T} \to \mathbb{H}$ governed by a well-defined SPDE

$$\mathrm{d}U(t) = A\,U(t)\,\mathrm{d}t + B\,\mathrm{d}W(t)$$

with some boundary and initial conditions, where $A\colon \mathbb{H} \to \mathbb{H}$ and $B\colon \mathbb{W} \to \mathbb{H}$ are linear operators, and $W\colon \mathbb{T} \to \mathbb{W}$ is a $\mathbb{W}$-valued Wiener process. Then we can borrow the idea presented in Section 4.3 to form a spatio-temporal SS-DGP by hierarchically composing such SPDEs of the form above.

A different path for generalising SS-DGPs is shown by Emzir et al. (2020). Specifically, they build deep Gaussian fields based on the SPDE representations of Matérn fields (Whittle, 1954; Lindgren et al., 2011). However, we should note that this approach gives random fields instead of spatio-temporal processes.

[1]By the time of the pre-examination of this thesis, the TME method is now implemented in JaX as an open source library (see, Section 1.2).

References

Rachid Ababou, Amvrossios C. Bagtzoglou, and Eric F. Wood. On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology*, 26(1):99–133, 1994. Cited on page 65.

Benjamin P. Abbott et al. Observation of gravitational waves from a binary black hole merger. *Physical Review Letters*, 116(6):061102, 2016. Cited on page 102.

Yacine Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, 70(1):223–262, 2003. Cited on pages 54 and 97.

Brian D. O. Anderson. Fixed interval smoothing for nonlinear continuous time systems. *Information and Control*, 20(3):294–300, 1972. Cited on page 37.

Brian D. O. Anderson and John B. Moore. Detectability and stabilizability of time-varying discrete-time linear systems. *SIAM Journal on Control and Optimization*, 19(1):20–32, 1981. Cited on page 55.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010. Cited on page 36.

Ángel F. García-Fernández, Filip Tronarp, and Simo Särkkä. Gaussian process classification using posterior linearization. *IEEE Signal Processing Letters*, 26 (5):735–739, 2019. Cited on page 84.

Ienkaran Arasaratnam and Simon Haykin. Cubature Kalman filters. *IEEE Transactions on Automatic Control*, 54(6):1254–1269, 2009. Cited on page 36.

Ienkaran Arasaratnam, Simon Haykin, and Robert J. Elliott. Discrete-time nonlinear filtering algorithms using Gauss–Hermite quadrature. *Proceedings of the IEEE*, 95(5):953–977, 2007. Cited on page 36.

Ienkaran Arasaratnam, Simon Haykin, and Thomas R. Hurd. Cubature Kalman filtering for continuous-discrete systems: Theory and simulations. *IEEE Transactions on Automatic Control*, 58(10):4977–4993, 2010. Cited on page 61.

Cédric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice*, volume 1, pages 1–16. PMLR, 2007. Cited on page 37.

Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-taylor. Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems 20*, pages 1–8. Curran Associates, Inc., 2008. Cited on page 37.

Patrik Axelsson and Fredrik Gustafsson. Discrete-time solutions to the continuous-time differential Lyapunov equation with applications to Kalman filtering. *IEEE Transactions on Automatic Control*, 60(3):632–643, 2015. Cited on page 33.

Michael Baake and Ulrike Schlägel. The Peano–Baker series. In *Proceedings of the Steklov Institute of Mathematics*, volume 275, pages 155–159, 2011. Cited on pages 39 and 76.

Alan Bain and Dan Crisan. *Fundamentals of Stochastic Filtering*. Springer-Verlag New York, 2009. Cited on pages 37 and 41.

Yaakov Bar-Shalom, Xiao-Rong Li, and Thiagalingam Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*. John Wiley & Sons, 2002. Cited on page 61.

Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry*. Springer-Verlag Berlin Heidelberg, 2006. Cited on page 49.

Philipp Batz, Andreas Ruttor, and Manfred Opper. Approximate Bayes learning of stochastic differential equations. *Physical Review E*, 98(2):022109, 2018. Cited on page 97.

Randal Beard, John Kenney, Jacob Gunther, Jonathan Lawton, and Wynn Stirling. Nonlinear projection filter based on Galerkin approximation. *Journal of Guidance, Control, and Dynamics*, 22(2):258–266, 1999. Cited on page 37.

Bradley M. Bell. The iterated Kalman smoother as a Gauss–Newton method. *SIAM Journal on Optimization*, 4(3):626–636, 1994. Cited on page 95.

Dennis S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009. Cited on page 38.

Alexandros Beskos and Gareth O. Roberts. Exact simulation of diffusions. *The Annals of Applied Probability*, 15(4):2422–2444, 2005. Cited on page 78.

Alexandros Beskos, Omiros Papaspiliopoulos, Gareth O. Roberts, and Paul Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):333–382, 2006. Cited on page 97.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. Cited on page 95.

David G. Blair, editor. *The Detection of Gravitational Waves*. Cambridge University Press, 1991. Cited on page 102.

Jose Blanchet and Fan Zhang. Exact simulation for multivariate Itô diffusions. *Advances in Applied Probability*, 52(4):1003–1034, 2020. Cited on page 78.

Dirk Blömker, Kody J. H. Law, Andrew M. Stuart, and Konstantinos C. Zygalakis. Accuracy and stability of the continuous-time 3DVAR filter for the Navier–Stokes equation. *Nonlinearity*, 26(8):2193–2219, 2013. Cited on page 56.

Vladimir I. Bogachev. *Gaussian Measures*. American Mathematical Society, 1998. Cited on page 26.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. Cited on pages 91 and 94.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1): 1–122, 2011. Cited on page 93.

Carlos A. Braumann. *Introduction to Stochastic Differential Equations with Applications to Modelling in Biology and Finance*. John Wiley & Sons, 2019. Cited on page 25.

George L. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag Berlin Heidelberg, 1988. Cited on page 100.

Damiano Brigo, Bernard Hanzon, and François LeGland. A differential geometric approach to nonlinear filtering: the projection filter. *IEEE Transactions on Automatic Control*, 43(2):247–252, 1998. Cited on page 37.

William L. Brogan. *Modern Control Theory*. Pearson, 3rd edition, 2011. Cited on pages 39 and 76.

Thang Bui, José M. Hernández-Lobato, Daniel Hernández-Lobato, Yingzhen Li, and Richard Turner. Deep Gaussian processes for regression using approximate expectation propagation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 1472–1481, New York, USA, 2016. PMLR. Cited on page 22.

Roberto Calandra, Jan Peters, Carl E. Rasmussen, and Marc P. Deisenroth. Manifold Gaussian processes for regression. In *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3338–3345, Vancouver, Canada, 2016. Cited on page 21.

Claudio Canuto and Anita Tabacco. *Mathematical Analysis II*. Springer International Publishing, 2nd edition, 2014. Cited on page 38.

Subhash Challa and Yaakov Bar-Shalom. Nonlinear filter design using Fokker–Planck–Kolmogorov probability density evolutions. *IEEE Transactions on Aerospace and Electronic Systems*, 36(1):309–315, 2000. Cited on page 37.

Krzysztof Chalupka, Christopher K. I. Williams, and Iain Murray. A framework for evaluating approximation methods for Gaussian process regression. *Journal of Machine Learning Research*, 14:333–350, 2013. Cited on page 19.

Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer International Publishing, 2020. Cited on page 36.

Kai Lai Chung and Ruth J. Williams. *Introduction to Stochastic Integration*. Probability and Its Applications. Birkhäuser Boston, 2nd edition, 1990. Cited on pages 25, 28, and 29.

Adrien Corenflos, James Thornton, George Deligiannidis, and Arnaud Doucet. Differentiable particle filtering via entropy-regularized optimal transport. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 2100–2111. PMLR, 2021a. Cited on page 36.

Adrien Corenflos, Zheng Zhao, and Simo Särkkä. Gaussian process regression in logarithmic time. *arXiv preprint arXiv:2102.09964*, 2021b. Cited on pages 20 and 67.

Jarrad Courts, Adrian Wills, and Thomas B. Schön. Gaussian variational state estimation for nonlinear state-space models. *IEEE Transactions on Signal Processing*, 2021. In press. Cited on page 37.

Lehel Csató and Manfred Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–668, 2002. Cited on page 19.

Jeffery J. DaCunha. Transition matrix and generalized matrix exponential via the Peano–Baker series. *Journal of Difference Equations and Applications*, 11 (15):1245–1264, 2005. Cited on page 76.

Didier Dacunha-Castelle and Danielle Florens-Zmirou. Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, 19(4):263–284, 1986. Cited on pages 42, 45, 54, and 97.

Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, volume 31, pages 207–215, Scottsdale, Arizona, USA, 2013. PMLR. Cited on page 22.

Philip J. Davis and Philip Rabinowitz. *Methods of Numerical Integration*. Academic Press, 1984. Cited on page 35.

Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10 (3):197–208, 2000. Cited on page 36.

Arnaud Doucet, Nando de Freitas, and Neil Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. Springer-Verlag New York, 2001. Cited on page 37.

Matthew M. Dunlop, Mark A. Girolami, Andrew M. Stuart, and Aretha L. Teckentrup. How deep are deep Gaussian processes? *Journal of Machine Learning Research*, 19(54):1–46, 2018. Cited on page 22.

David K. Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014. Cited on page 22.

David K. Duvenaud, Oren Rippel, Ryan Adams, and Zoubin Ghahramani. Avoiding pathologies in very deep networks. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, volume 33, pages 202–210, Reykjavík, Iceland, 2014. PMLR. Cited on page 22.

Eugene B. Dynkin. *Markov Processes: Volume 1*. Springer-Verlag Berlin Heidelberg, 1965. Cited on page 43.

Albert Einstein and Nathan Rosen. On gravitational waves. *Journal of the Franklin Institute*, 223(1):43–54, 1937. Cited on page 102.

Muhammad Emzir, Sari Lasanen, Zenith Purisha, Lassi Roininen, and Simo Särkkä. Non-stationary multi-layered Gaussian priors for Bayesian inversion. *Inverse Problems*, 37(1):015002, 2020. Cited on pages 22, 69, and 110.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 1986. Cited on page 29.

Geir Evensen. *Data Assimulation: The Ensemble Kalman Filter*. Springer-Verlag Berlin Heidelberg, 2nd edition, 2009. Cited on page 37.

Danielle Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4):547–557, 1989. Cited on pages 42, 44, and 45.

Avner Friedman. *Stochastic Differential Equations and Applications: Volume 1*. Academic Press, 1975. Cited on page 75.

Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274 of *Graduate Texts in Mathematics*. Springer International Publishing Switzerland, 2016. Cited on pages 29 and 30.

Rui Gao. *Rescursive Smoother Type Variable Splitting Methods for State Estimation*. PhD thesis, Aalto University, 2020. Cited on page 95.

Rui Gao, Filip Tronarp, and Simo Särkkä. Iterated extended Kalman smoother-based variable splitting for L_1-regularized state estimation. *IEEE Transactions on Signal Processing*, 97(19):5078–5092, 2019a. Cited on page 95.

Rui Gao, Filip Tronarp, Zheng Zhao, and Simo Särkä. Regularized state estimation and parameter learning via augmented Lagrangian Kalman smoother method. In *Proceedings of the 29th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Pittsburgh, PA, USA, 2019b. Cited on page 101.

Constantino A. Garcìa, Abraham Otero, Paulo Félix, Jesús Presedo, and David G. Márquez. Nonparametric estimation of stochastic differential equations with sparse Gaussian processes. *Physical Review E*, 96(2):022104, 2017. Cited on page 97.

Jacob Gardner, Geoff Pleiss, Kilian Q. Weinberger, David Bindel, and Andrew G. Wilson. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems 31*, pages 1–11. Curran Associates, Inc., 2018. Cited on page 20.

Mark N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, University of Cambridge, 1997. Cited on pages 20 and 72.

Simon J. Godsill, Arnaud Doucet, and Mike West. Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association*, 99(465): 156–168, 2004. Cited on page 36.

Alexander Grigorievskiy, Neil Lawrence, and Simo Särkkä. Parallelizable sparse inverse formulation Gaussian processes (SpInGP). In *Proceedings of the 27th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Tokyo, Japan, 2017. Cited on page 20.

Jonathan L. Gross, Jay Yellen, and Mark Anderson. *Graph Theory and Its Applications*. Chapman & Hall/CRC, 3rd edition, 2018. Cited on page 69.

Jouni Hartikainen and Simo Särkkä. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Proceedings of the 20th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 379–384, 2010. Cited on pages 66 and 76.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015. Cited on pages 89, 91, and 93.

Markus Heinonen, Henrik Mannerström, Juho Rousu, Samuel Kaski, and Harri Lähdesmäki. Non-stationary Gaussian process regression with Hamiltonian Monte Carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 732–740, Cadiz, Spain, 2016. PMLR. Cited on page 22.

Uwe Helmke and Joachim Rosenthal. Eigenvalue inequalities and Schubert calculus. *Mathematische Nachrichten*, 171(1):207–225, 1995. Cited on page 38.

Philipp Hennig, Michael A. Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179):20150142, 2015. Cited on page 17.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 282–290. AUAI Press, 2013. Cited on page 19.

Dave Higdon, Jenise Swall, and J. Kern. Non-stationary spatial modeling. *Bayesian Statistics*, 6(1):761–768, 1999. Cited on pages 20 and 72.

Clyde D. Hill, Paweł Nuroski, Lydia Bieri, David Garfinkle, and Nicolás Yunes. The mathematics of gravitational waves. *Notice of the AMS*, 64(7):686–707, 2017. Cited on page 102.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991. Cited on page 38.

Roland Hostettler, Tuomas Lumikari, Lauri Palva, Tuomo Nieminen, and Simo Särkkä. Motion artifact reduction in ambulatory electrocardiography using inertial measurement units and Kalman filtering. In *Proceedings of the 21st International Conference on Information Fusion (FUSION)*, pages 780–787, Cambridge, UK, 2018. Cited on page 104.

Stefano M. Iacus. *Simulation and Inference for Stochastic Differential Equations: With R Examples*. Springer-Verlag New York, 2008. Cited on page 47.

Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic Differential Equations and Diffusion Processes*. North Holland, 2nd edition, 1992. Cited on pages 25, 43, and 80.

Kazufumi Itô and Kaiqi Xiong. Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 45(5):910–927, 2000. Cited on pages 34 and 56.

Kiyosi Itô. Stochastic integral. In *Proceedings of the Imperial Academy*, volume 20, pages 519–524, 1944. Cited on page 27.

Kiyosi Itô. *Stochastic Processes: Lectures given at Aarhus University*. Springer-Verlag Berlin Heidelberg, 2004. Cited on page 43.

Andrew H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970. Cited on pages 31, 35, 37, 41, and 55.

Bin Jia, Ming Xin, and Yang Cheng. Sparse-grid quadrature nonlinear filtering. *Automatica*, 48(2):327–341, 2012. Cited on page 36.

Simo J. Julier and Jeffrey K. Uhlmann. Unscented filtering and nonlinear estimation. In *Proceedings of the IEEE*, volume 92, pages 401–422, 2004. Cited on page 36.

Jari P. Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. Springer-Verlag New York, 2005. Cited on page 89.

Rudolf E. Kálmán and Richard S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83(1):95–108, 1961. Cited on page 36.

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag New York, 2nd edition, 1991. Cited on pages 25, 26, 27, 28, 40, 63, 64, 66, 67, and 75.

Toni Karvonen, Silvère Bonnabel, Eric Moulines, and Simo Särkkä. On stability of a class of filters for nonlinear stochastic systems. *SIAM Journal on Control and Optimization*, 58(4):2023–2049, 2020. Cited on pages 56 and 57.

Yitzhak Katznelson. *An Introduction to Harmonic Analysis*. Cambridge University Press, 3rd edition, 2004. Cited on page 100.

Mathieu Kessler. Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics*, 24(2):211–229, 1997. Cited on pages 42, 54, and 97.

Mathieu Kessler, Alexander Lindner, and Michael Sørensen. *Statistical Methods for Stochastic Differential Equations*. Chapman & Hall/CRC, 2012. Cited on page 78.

Hassan K. Khalil. *Nonlinear Systems*. Pearson, 3rd edition, 2002. Cited on page 79.

Rafail Khasminskii. *Stochastic Stability of Differential Equations*. Springer-Verlag Berlin Heidelberg, 2012. Cited on page 42.

Genshiro Kitagawa. Non-Gaussian state–space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032–1041, 1987. Cited on page 32.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer-Verlag London, 2nd edition, 2014. Cited on page 87.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag Berlin Heidelberg, 1992. Cited on pages 25, 42, 51, 77, and 98.

Juš Kocijan. *Modelling and Control of Dynamic Systems Using Gaussian Process Models*. Springer International Publishing, 2016. Cited on page 17.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. Cited on page 69.

Leonid B. Koralov and Yakov G. Sinai. *Theory of Probability and Random Processes*. Lecture Notes in Mathematics. Springer-Verlag Berlin Heidelberg, 2nd edition, 2007. Cited on page 26.

Kira Kowalska and Leto Peel. Maritime anomaly detection using Gaussian process active learning. In *Proceedings of the 15th International Conference on Information Fusion (FUSION)*, pages 1164–1171, Singapore, 2012. Cited on page 104.

Shinsuke Koyama. Projection smoothing for continuous and continuous-discrete stochastic dynamic systems. *Signal Processing*, 144:333–340, 2018. Cited on page 37.

Gennady Yu. Kulikov and Maria V. Kulikova. Accurate numerical implementation of the continuous-discrete extended Kalman filter. *IEEE Transactions on Automatic Control*, 59(1):273–279, 2014. Cited on page 42.

Hui-Hsiung Kuo. *Gaussian Measures in Banach Spaces*, volume 463 of *Lecture Notes in Mathematics*. Springer-Verlag New York, 1975. Cited on page 26.

Hui-Hsiung Kuo. *Introduction to Stochastic Integration*. Universitext. Springer-Verlag New York, 2006. Cited on pages 26, 27, 29, and 43.

Harold J. Kushner. On the differential equations satisfied by conditional probablitity densities of Markov processes, with applications. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 2(1):106–119, 1964. Cited on page 37.

Carl E. Langenhop. Bounds on the norm of a solution of a general differential equation. In *Proceedings of the American Mathematical Society*, volume 11, pages 795–799, 1960. Cited on page 39.

Kody J. H. Law, Abhishek Shukla, and Andrew M. Stuart. Analysis of the 3DVAR filter for the partially observed Lorenz'63 model. *Discrete and Continuous Dynamical Systems*, 34(3):1061–1078, 2014. Cited on page 56.

Kody J. H. Law, Andrew M. Stuart, and Konstantinos C. Zygalakis. *Data Assimilation: A Mathematical Introduction*. Springer International Publishing Switzerland, 2015. Cited on page 37.

Miguel Lázaro-Gredilla. Bayesian warped Gaussian processes. In *Advances in Neural Information Processing Systems 25*, pages 1–9. Curran Associates, Inc., 2012. Cited on page 21.

Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl E. Rasmussen, and Figueiras-Vidal R. Aníbal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11(63):1865–1881, 2010. Cited on page 20.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 3870–3882. PMLR, 2020. Cited on page 37.

Yaowei Li. Non-stationary State Space Gaussian Processes. Master thesis, Aalto University, 2020. Cited on page 77.

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. Cited on pages 20 and 110.

Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When Gaussian process meets big data: a review of scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4405–4423, 2020. Cited on page 19.

Gabriel J. Lord, Catherine E. Powell, and Tony Shardlow. *An Introduction to Computational Stochastic PDEs*, volume 50 of *Cambridge Texts in Applied Mathematics*. Cambridge University Press, 2014. Cited on pages 26 and 60.

Michele Maggiore. *Gravitational Waves: Volume 1: Theory and Experiments*. Oxford University Press, 2008. Cited on page 102.

Xuerong Mao. *Stochastic Differential Equations and Applications*. Woodhead Publishing, 2nd edition, 2008. Cited on page 75.

Bertil Matérn. *Spatial Variation: Stochastic models and their applications to some problems in forest surveys and other sampling investigations*. PhD thesis, Stockholm University, 1960. Cited on page 64.

Peter S. Maybeck. *Stochastic Models, Estimation, and Control: Volume 2*. Academic Press, 1982. Cited on pages 35 and 37.

Edward Meeds and Simon Osindero. An alternative infinite mixture of Gaussian process experts. In *Advances in Neural Information Processing Systems 18*, pages 883–890. MIT press, 2006. Cited on page 20.

Per C. Moan and Jitse Niesen. Convergence of the Magnus series. *Foundations of Computational Mathematics*, 8(3):291–301, 2008. Cited on pages 39 and 76.

Karla Monterrubio-Gómez, Lassi Roininen, Sara Wade, Theodoros Damoulas, and Mark Girolami. Posterior inference for sparse hierarchical non-stationary models. *Computational Statistics and Data Analysis*, 148:106954, 2020. Cited on page 22.

Peter Mörters and Yuval Peres. *Brownian Motion*, volume 30 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2010. Cited on page 26.

Radford M. Neal. Regression and classification using Gaussian process priors. In *Proceedings of the Sixth Valencia International Meeting*, volume 6 of *Bayesian Statistics*, pages 475–501. Oxford University Press, 1999. Cited on page 84.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004. Cited on page 94.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer International Publishing, 2nd edition, 2018. Cited on page 94.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer-Verlag New York, 2nd edition, 2006. Cited on pages 91 and 93.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer-Verlag Berlin Heidelberg, 6th edition, 2007. Cited on pages 25, 27, 28, 29, 43, and 80.

Manfred Opper. Variational inference for stochastic differential equations. *Annalen der Physik*, 531(3):1800233, 2019. Cited on page 97.

E. E. Osborne. On pre-conditioning of matrices. *Journal of the ACM*, 7(4):338–345, 1960. Cited on page 79.

Tohru Ozaki. A local linearization approach to nonlinear filtering. *International Journal of Control*, 57(1):75–96, 1993. Cited on page 77.

Baburao G. Pachpatte. *Inequalities for Differential and Integral Equations*, volume 197 of *Mathematics in Science and Engineering*. Academic Press, 1998. Cited on page 39.

Christopher J. Paciorek and Mark J. Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in Neural Information Processing Systems 16*, pages 273–280. MIT Press, 2004. Cited on pages 21, 22, and 72.

Christopher J. Paciorek and Mark J. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506, 2006. Cited on pages 21 and 72.

Raymond E. A. C. Paley and Nobert Wiener. *Fourier Transform in the Complex Domain*, volume 19 of *Colloquium Publications*. American Mathematical Society, 1934. Cited on page 26.

Omiros Papaspiliopoulos, Yvo Pokern, Gareth O. Roberts, and Andrew M. Stuart. Nonparametric estimation of diffusions: a differential equations approach. *Biometrika*, 99(3):511–531, 2012. Cited on page 97.

Beresford N. Parlett and Christian Reinsch. Balancing a matrix for calculation of eigenvalues and eigenvectors. In *Handbook for Automatic Computation*, pages 315–326. Springer-Verlag Berlin, 1971. Cited on page 79.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer-Verlag New York, 2014. Cited on page 67.

Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 152 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, 2nd edition, 2014. Cited on pages 26 and 110.

Yuan Qi, Thomas P. Minka, and Rosalind W. Picara. Bayesian spectrum estimation of unevenly sampled nonstationary data. In *Proceedings of the 37th IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1473–1476, Orlando, FL, USA, 2002. Cited on page 100.

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. Cited on page 19.

Rahul Radhakrishnan, Abhinoy Kumar Singh, Shovan Bhaumik, and Nutan Kumar Tomar. Multiple sparse-grid Gauss–Hermite filtering. *Applied Mathematical Modelling*, 40(7–8):4441–4450, 2016. Cited on page 36.

Syama S. Rangapuram, Matthias W. Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. In *Advances in Neural Information Processing Systems 31*, pages 7785–7794. Curran Associates, Inc., 2018. Cited on page 82.

Pritam Ranjan, Ronald Haynes, and Richard Karsten. A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data. *Technometrics*, 53(4):366–378, 2011. Cited on pages 65 and 66.

Carl E. Rasmussen and Zoubin Ghahramani. Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems 14*, pages 881–888. MIT press, 2002. Cited on page 20.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. Cited on pages 17, 21, 63, 65, 68, and 84.

Konrad Reif, Stefan Günther, Engin Yaz, and Rolf Unbehauen. Stochastic stability of the discrete-time extended Kalman filter. *IEEE Transactions on Automatic Control*, 44(4):714–728, 1999. Cited on page 55.

Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017. Cited on page 72.

Gonzalo Rios and Felipe Tobar. Compositionally-warped Gaussian processes. *Neural Networks*, 118:235–246, 2019. Cited on page 21.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer New York, 2nd edition, 2004. Cited on page 78.

Gareth O. Roberts and Osnat Stramer. On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika*, 88 (3):603–421, 2001. Cited on page 97.

Stephen Roberts, Michael A. Osborne, Mark Ebden, S. Reece, N. Gibson, and S. Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013. Cited on page 17.

Chris Rogers and David Williams. *Diffusions, Markov Processes, and Martingales*. Cambridge University Press, 2nd edition, 2000. Cited on pages 28 and 29.

Lassi Roininen, Mark Girolami, Sari Lasanen, and Makku Markkanen. Hyperpriors for Matérn fields with applications in Bayesian inversion. *Inverse Problems & Imaging*, 13(1):1–29, 2019. Cited on pages 22 and 69.

H. Rong, A. P. Teixeira, and C. G. Soares. Ship trajectory uncertainty prediction based on a Gaussian process model. *Ocean Engineering*, pages 499–511, 2019. Cited on page 104.

Iurii A. Rozanov. Markov random fields and stochastic partial differential equations. *Mathematics of the USSR-Sbornik*, 32(4):515–534, 1977. Cited on page 67.

Iurii A. Rozanov. *Markov Random Fields*. Springer-Verlag New York, 1982. Cited on page 67.

Håvard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, 2005. Cited on page 20.

Håvard Rue and Sara Martino. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, 137(10):3177–3192, 2007. Cited on page 20.

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. Cited on page 20.

Andrzej P. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006. Cited on page 91.

Andreas Ruttor, Philipp Batz, and Manfred Opper. Approximate Gaussian process inference for the drift function in stochastic differential equations. In *Advances in Neural Information Processing Systems 26*, pages 1–9. Curran Associates, Inc., 2013. Cited on page 97.

Hugh Salimbeni and Marc P. Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems 30*, pages 1–12. Curran Associates, Inc., 2017a. Cited on page 22.

Hugh Salimbeni and Marc P. Deisenroth. Deeply non-stationary Gaussian processes. In *NIPS Workshop on Bayesian Deep Learning*, 2017b. Cited on pages 22 and 69.

Paul D. Sampson and Peter Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992. Cited on page 21.

Neville Sancho. On the approximate moment equations of a nonlinear stochastic differential equation. *Journal of Mathematical Analysis and Applications*, 29 (2):384–391, 1970. Cited on page 35.

Simo Särkkä. On unscented Kalman filtering for state estimation of continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, 52(9):1631–1641, 2007. Cited on page 42.

Simo Särkkä. Continuous-time and continuous-discrete-time unscented Rauch–Tung–Striebel smoothers. *Signal Processing*, 90(1):225–235, 2010. Cited on page 42.

Simo Särkkä. *Bayesian Filtering and Smoothing*, volume 3 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, 2013. Cited on pages 29, 31, 32, 33, 36, 37, 41, and 56.

Simo Särkkä and Ángel F. García-Fernández. Temporal parallelization of Bayesian smoothers. *IEEE Transactions on Automatic Control*, 366(1):299–306, 2021. Cited on page 67.

Simo Särkkä and Juha Sarmavuori. Gaussian filtering and smoothing for continuous-discrete dynamic systems. *Signal Processing*, 93(2):500–510, 2013. Cited on pages 34 and 35.

Simo Särkkä and Arno Solin. On continuous-discrete Cubature Kalman filtering. In *Proceedings of 16th IFAC Symposium on System Identification*, volume 45, pages 1221–1226, 2012. Cited on page 36.

Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*, volume 10 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, 2019. Cited on pages 20, 31, 32, 33, 34, 37, 42, 43, 52, 60, 67, and 77.

Simo Särkkä and Lennart Svensson. Levenberg–Marquardt and line-search extended Kalman smoothers. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5875–5879, 2020. Cited on page 95.

Simo Särkkä, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: a look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013. Cited on pages 20, 66, 67, 76, and 79.

René L. Schilling. *Measures, Integrals and Martingales*. Cambridge University Press, 2nd edition, 2017. Cited on page 74.

René L. Schilling and Lothar Partzsch. *Brownian Motion: An Introduction to Stochastic Processes*. De Gruyter, 2012. Cited on pages 26, 29, 30, and 43.

Alexandra M. Schmidt and Anthony O'Hagan. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758, 2003. Cited on page 21.

Yi Shen, Qi Luo, and Xuerong Mao. The improved LaSalle-type theorems for stochastic functional differential equations. *Journal of Mathematical Analysis and Applications*, 318(1):134–154, 2006. Cited on page 75.

Mark Smith, Steven Reece, Stephen Roberts, and Iead Rezek. Maritime abnormality detection using Gaussian processes. *Knowledge and Information Systems*, 38(3):717–740, 2014. Cited on page 104.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press, 2006. Cited on page 19.

Edward Snelson and Zoubin Ghahramani. Local and global sparse Gaussian process approximations. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, volume 2, pages 524–531. PMLR, 2007. Cited on page 19.

Edward Snelson, Zoubin Ghahramani, and Carl E. Rasmussen. Warped Gaussian processes. In *Advances in Neural Information Processing Systems 16*, pages 1–8. MIT Press, 2004. Cited on page 21.

Jasper Snoek, Kevin Swersky, Rich Zemel, and Ryan Adams. Input warping for Bayesian optimization of non-stationary functions. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1674–1682. PMLR, 2014. Cited on page 72.

Arno Solin. *Stochastic Differential Equation Methods for Spatio-Temporal Gaussian Process Regression*. PhD thesis, Aalto University, 2016. Cited on pages 66 and 79.

Arno Solin and Simo Särkkä. Explicit link between periodic covariance functions and state space models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 904–912, Reykjavík, Iceland, 2014. PMLR. Cited on pages 82 and 101.

Ruslan Leont'evich Stratonovich. A new representation for stochastic integrals and equations. *SIAM Journal on Control*, 4(2):362–371, 1966. Cited on page 27.

Daniel W. Stroock and Sathamangalam R. S. Varadhan. Diffusion processes with continuous coefficients, I and II. *Communications on Pure and Applied Mathematics*, 22(3 and 4):345–400 and 478–530, 1969. Cited on pages 28 and 29.

Daniel W. Stroock and Sathamangalam R. S. Varadhan. *Multidimensional Diffusion Processes*. Springer-Verlag Berlin Heidelberg, 1979. Cited on page 28.

Sarang Thombre, Zheng Zhao, Henrik Ramm-Schmidt, José M. Vallet García, Tuomo Malkamäki, Sergey Nikolskiy, Toni Hammarberg, Hiski Nuortie, M. Z. H. Bhuiyan, Simo Särkkä, and Ville V. Lehtola. Sensors and AI techniques for situational awareness in autonomous ships: a review. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–20, 2020. In press. Cited on page 104.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996. Cited on page 89.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5, pages 567–574. PMLR, 2009. Cited on page 19.

Ke A. Wang, Geoff Pleiss, Jacob R. Gardner, Stephen Tyree, Kilian Q. Weinberger, and Andrew G. Wilson. Exact Gaussian processes on a million data points. In *Advances in Neural Information Processing Systems 32*, pages 14648–14659. Curran Associates, Inc., 2019. Cited on page 20.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen. *Mathematische Annalen*, 71(4):441–479, 1912. Cited on page 38.

Peter Whittle. On statioanry process in the plane. *Biometrika*, 41(3–4):434–449, 1954. Cited on page 110.

Nobert Wiener. Differential-space. *Journal of Mathematics and Physics*, 2(1-4): 131–174, 1923. Cited on page 26.

Christopher K. I. Williams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, 1998. Cited on page 21.

Andrew G. Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 370–378, Cadiz, Spain, 2016. PMLR. Cited on page 21.

Zhipeng Xiao, Bin Dai, Hongdong Li, Tao Wu, Xin Xu, Yujun Zeng, and Tongtong Chen. Gaussian process regression-based robust free space detection for autonomous vehicle by 3-D point cloud and 2-D appearance information fusion. *International Journal of Advanced Robotic Systems*, 14(4):1–20, 2017. Cited on page 104.

Kaiqi Xiong, H. Y. Zhang, and C. W. Chan. Performance evaluation of UKF-based nonlinear filtering. *Automatica*, 42(2):261–270, 2006. Cited on page 56.

Dongbin Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010. Cited on pages 41 and 42.

Toshio Yamada and Shinzo Watanabe. On the uniqueness of solutions of stochastic differential equations. *Journal of Mathematics of Kyoto University*, 11(1):155–167, 1971. Cited on page 29.

Nakahiro Yoshida. Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis*, 41(2):220–242, 1992. Cited on page 97.

Moshe Zakai. On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 11(3):230–243, 1969. Cited on page 37.

Michael Minyi Zhang and Sinead A. Williamson. Embarrassingly parallel inference for Gaussian processes. *Journal of Machine Learning Research*, 20(169): 1–26, 2019. Cited on page 20.

Zheng Zhao and Simo Särkkä. Non-linear Gaussian smoothing with Taylor moment expansion. *IEEE Signal Processing Letters*, 2021. In press. Cited on pages 54 and 58.

Zheng Zhao, Simo Särkkä, and Ali Bahrami Rad. Spectro-temporal ECG analysis for atrial fibrillation detection. In *Proceedings of the 28th IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Aalborg, Denmark, 2018. Cited on pages 100 and 101.

Zheng Zhao, Simo Särkkä, and Ali Bahrami Rad. Kalman-based spectro-temporal ECG analysis using deep convolutional networks for atrial fibrillation detection. *Journal of Signal Processing Systems*, 92(7):621–636, 2020a. Cited on pages 100 and 101.

Zheng Zhao, Filip Tronarp, Roland Hostettler, and Simo Särkkä. State-space Gaussian process for drift estimation in stochastic differential equations. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5295–5299, Barcelona, Spain, 2020b. Cited on pages 98, 99, and 100.

Zheng Zhao, Muhammad Emzir, and Simo Särkkä. Deep state-space Gaussian processes. *Statistics and Computing*, 31(6):75, 2021a. Cited on pages 18, 22, 36, 80, 84, 85, 87, 88, 89, and 103.

Zheng Zhao, Toni Karvonen, Roland Hostettler, and Simo Särkkä. Taylor moment expansion for continuous-discrete Gaussian filtering. *IEEE Transactions on Automatic Control*, 66(9):4460–4467, 2021b. Cited on pages 42, 47, 49, 54, 57, and 61.

Zheng Zhao, Gao Rui, and Simo Särkkä. Hierarchical non-stationary temporal Gaussian processes with L^1-regularization. *arXiv preprint arXiv:2105.09695*, 2021c. Cited on pages 75, 76, 80, 91, 93, 94, 96, and 104.

Errata

Publication I

In Example 7, the coefficient $\Phi_{x,2}$ should multiply with a factor 2.

Publication I

Zheng Zhao, Toni Karvonen, Roland Hostettler, and Simo Särkkä. Taylor moment expansion for continuous-discrete Gaussian filtering. *IEEE Transactions on Automatic Control*, Volume 66, Issue 9, Pages 4460–4467, December 2020.

Publication II

Zheng Zhao, Muhammad Emzir, and Simo Särkkä. Deep state-space Gaussian processes. *Statistics and Computing*, Volume 31, Issue 6, Article number 75, Pages 1–26, September 2021.

Publication III

Zheng Zhao, Simo Särkkä, and Ali Bahrami Rad. Kalman-based spectro-temporal ECG analysis using deep convolutional networks for atrial fibrillation detection. *Journal of Signal Processing Systems*, Volume 92, Issue 7, Pages 621–636, April 2020.

Publication IV

Zheng Zhao, Filip Tronarp, Roland Hostettler, and Simo Särkkä. State-space Gaussian process for drift estimation in stochastic differential equations. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, Pages 5295–5299, May 2020.

Publication V

Zheng Zhao, Simo Särkkä, and Ali Bahrami Rad. Spectro-temporal ECG analysis for atrial fibrillation detection. In *Proceedings of the IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark, 6 pages, September 2018.

Publication VI

Sarang Thombre, Zheng Zhao, Henrik Ramm-Schmidt, José M. Vallet García, Tuomo Malkamäki, Sergey Nikolskiy, Toni Hammarberg, Hiski Nuortie, M. Zahidul H. Bhuiyan, Simo Särkkä, and Ville V. Lehtola. Sensors and AI techniques for situational awareness in autonomous ships: a review. Accepted for publication in *IEEE Transactions on Intelligent Transportation Systems*, 20 pages, September 2020.

Publication VII

Zheng Zhao, Rui Gao, and Simo Särkkä. Hierarchical Non-stationary temporal Gaussian processes with L^1-regularization. Submitted to *Statistics and Computing*, May 2021.

Aalto-DD 163/2021

ISBN 978-952-64-0602-2 (printed)
ISBN 978-952-64-0603-9 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Electrical Engineering and Automation
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS