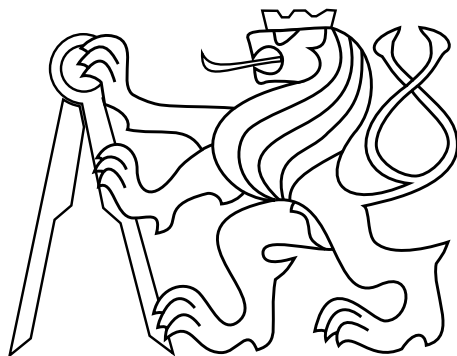Master Thesis

Bayesian Parameter Estimation of State-Space Models with Intractable Likelihood

Bc. Tomáš Kala

SUPERVISOR: ING. KAMIL DEDECIUS, PHD.

APRIL 2019

DEPARTMENT OF COMPUTER SCIENCE
FACULTY OF ELECTRICAL ENGINEERING
CZECH TECHNICAL UNIVERSITY IN PRAGUE

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Kala**   Jméno: **Tomáš**   Osobní číslo: **434690**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra počítačů**

Studijní program: **Otevřená informatika**

Studijní obor: **Bioinformatika**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Bayesovské odhadování parametrů stavových modelů při nedostupné věrohodnostní funkci

Název diplomové práce anglicky:

Bayesian parameter estimation of state-space models with intractable likelihood

Pokyny pro vypracování:

Stavové modely představují velmi populární formalismus vhodný pro popis celé řady různých náhodných procesů, od časových řad po aplikace v teorii řízení. Pokud tyto modely neobsahují statické parametry, lze pro jejich odhad použít např. Kalmanův filtr a jeho varianty, dále particle filtraci aj. Pokud ovšem statické parametry obsahují, tyto filtry zpravidla nekonvergují a nezbývá, než přikročit k optimalizačním technikám typu maximalizace věrohodnosti či particle Markov chain Monte Carlo. Další komplikace nastávají, pokud navíc není věrohodnostní funkce pro pozorovanou veličinu dostupná, nebo je nepřesná či příliš komplikovaná. Diplomová práce je specificky zaměřena poslední zmíněnou problematiku.
Specifické pokyny
1. Seznamte se s metodami pro odhadování stavových modelů pomocí kalmanovské filtrace a sekvenční Monte Carlo filtrace. Nastudujte problematiku statických parametrů a jejich odhadu.
2. Proveďte rešerši ohledně využití daných metod v oblasti bioinformatiky, například v modelování buněčných procesů.
3. Seznamte se s metodami ABC - Approximate Bayesian Computation a jejich využití ve filtraci stavových modelů.
4. Navrhněte vhodný způsob odhadování statických parametrů stavových modelů s využitím metod ABC.
5. Experimentálně (na vhodném problému z oblasti molekulární biologie) a případně teoreticky ověřte vlastnosti navržené metody, diskutujte její vlastnosti a navrhněte další možné směry výzkumu.

Seznam doporučené literatury:

1] C. C. Drovandi, A. N. Pettitt, and R. A. McCutchan, "Exact and approximate Bayesian inference for low integer-valued time series models with intractable likelihoods," Bayesian Anal., vol. 11, no. 2, pp. 325–352, 2016.
[2] S. Martin, A. Jasra, S. S. Singh, N. Whiteley, P. Del Moral, and E. McCoy, "Approximate Bayesian Computation for Smoothing," Stoch. Anal. Appl., vol. 32, no. 3, pp. 397–420, 2014.
[3] T. B. Schön, A. Svensson, L. Murray, and F. Lindsten, "Probabilistic learning of nonlinear dynamical systems using sequential Monte Carlo," Mech. Syst. Signal Process., vol. 104, pp. 866–883, May 2018.
[4] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," J. R. Stat. Soc. Ser. B (Statistical Methodol., vol. 72, no. 3, pp. 269–342, Jun. 2010.
[5] K. Dedecius, "Adaptive kernels in approximate filtering of state-space models," Int. J. Adapt. Control Signal Process., vol. 31, no. 6, pp. 938–952, Jun. 2017.

Jméno a pracoviště vedoucí(ho) diplomové práce:

Ing. Kamil Dedecius, Ph.D.,   ÚTIA AV ČR

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **04.02.2019**   Termín odevzdání diplomové práce: **24.05.2019**

Platnost zadání diplomové práce: **20.09.2020**

_____   _____   _____
Ing. Kamil Dedecius, Ph.D.   podpis vedoucí(ho) ústavu/katedry   prof. Ing. Pavel Ripka, CSc.
podpis vedoucí(ho) práce   podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací.
Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

_____   _____
Datum převzetí zadání   Podpis studenta

Abstract

Abstract in English

Abstrakt

Abstract in Czech

Author statement for graduate thesis:

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, date ....

<div align="right">signature</div>

Acknowledgements

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Probabilistic modelling arises in a wide variety of situations. Often, the measurements one uses to perform inference have been carried out with an unknown error. Frequently, one also does not have access to a correct model for the particular situation — the true model is either unknown, or such model is even impossible to formulate.

In the former case, one naturally assumes a random error associated with the observations, and attempts to infer something from the data while accounting for this randomness.

In the latter case, one has no choice but to work with a given, although possibly simplified model, purely because of insufficient domain knowledge. Connected with such a model is some degree of uncertainty about its parameters. It is often beneficial to think of these parameters as random variables themselves, in accordance with the Bayesian methodology (Robert, 2007). Such formulation allows to formulate one's prior beliefs about the parameter values, and then updating them upon receiving new observations.

In this thesis, we work with state-space models (SSMs) consisting of a sequence of observed random variables y_t indexed by discrete time $t = 1, \ldots, T$, which have been generated by a latent random process x_t, $t = 1, \ldots, T$. The distribution of x_t and y_t is assumed to be parameterized by a static parameter θ. Our goal is to perform posterior inference about this parameter, given a sequence of observations y_t. Furthermore, we assume that the likelihood function of the SSM is intractable and must be approximated. This assumption is well-grounded, as the likelihood is only available in severely restricted cases, to be discussed in Chapter 3, together with a formal definition of the SSM.

Our contribution is twofold. First, we show how to apply the Approximate Bayesian Computation (ABC) methodology (Rubin et al., 1984; Pritchard et al., 1999) to obtain an estimate of the likelihood even under a misspecified observation model. Our formulation allows for arbitrary kernel functions with automatically determined widths, unlike the simple accept-reject routine typically discussed in the literature. Second, we apply the resulting model to the genetic auto-regulation process in prokaryotes. Such situation is suitable for a state-space model with a possibly misspecified observation model, as all attempts to model such a complex system are necessarily simplified. To quote the famous statistician George E. P. Box, *"all models are wrong, but some are useful"* (Box, 1979). This statement is particularly true for such situations, and it is our hope that our model is indeed useful.

The rest of the thesis is organized as follows:

In Chapter 2, we discuss the related work. First, we review some classical works on Markov Chain Monte Carlo (MCMC) methods. We also discuss how these can be applied to state-space models with an intractable likelihood function. Next, we overview the literature on ABC methods and how these could be used to obtain a suitable likelihood estimate even when the observation model is incorrect. Finally, we discuss the application of related methods in bioinformatics and molecular biology.

In Chapter 3, we describe the assumed form of a state-space model. We show how one would implement a sampler to approximately infer the static parameters given a sequence of observations. We also show that in this basic form, such sampler is unusable, since it relies on the evaluation of the likelihood function of the observed sequence, which is intractable (up to certain special cases). We then show how this likelihood can be estimated using the particle filter (Doucet et al., 2001) without affecting the asymptotical properties of the sampler.

Chapter 4 provides a description of the ABC method, and also how it can be applied to estimate the likelihood even under a misspecified model. We discuss the pros and cons of such approach as well as potential issues and how to address them.

Chapter 5 provides numerical studies, where we apply the model developed in Chapter 4 to several examples and compare it with the model utilizing the particle filter. We perform stability tests and show how the models behave under outliers and model misspecification. This chapter also includes the prokaryotic auto-regulation study discussed earlier.

Finally, Chapter 6 concludes the thesis and discusses some possible directions to be investigated in the future.

Chapter 2

Related work

Chapter 3

Learning the parameters of a state-space model

Chapter 4

Approximate Bayesian Computation

Chapter 5

Applications

Chapter 6

Conclusion and future work

Bibliography

G. E. Box. Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier, 1979.

A. Doucet, A. Smith, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics. Springer New York, 2001. ISBN 9780387951461. URL `https://books.google.cz/books?id=uxX-koqKtMMC`.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

D. B. Rubin et al. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.