Department of Computer Science

Computational methods for Bayesian model assessment

Topi Paananen

Aalto University

DOCTORAL THESES

Aalto University publication series
DOCTORAL THESES 96/2022

Computational methods for Bayesian model assessment

Topi Paananen

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall A208d Jeti, Otakaari 5, on 15 September 2022 at 12.

Aalto University
School of Science
Department of Computer Science
Probabilistic Machine Learning group

Supervising professor
Professor Aki Vehtari, Aalto University, Finland

Thesis advisor
Professor Aki Vehtari, Aalto University, Finland

Preliminary examiners
Reader Víctor Elvira, University of Edinburgh, UK
Professor Lorin Crawford, Brown University, USA

Opponent
Reader Víctor Elvira, University of Edinburgh, UK

Aalto University publication series
DOCTORAL THESES 96/2022

ISBN 978-952-64-0869-9 (printed)
ISBN 978-952-64-0870-5 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0870-5

Unigrafia Oy
Helsinki 2022

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Abstract

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

Author
Topi Paananen

Name of the doctoral thesis
Computational methods for Bayesian model assessment

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL THESES 96/2022

Field of research Computer Science

Manuscript submitted 4 March 2022 **Date of the defence** 15 September 2022

Permission for public defence granted (date) 11 May 2022 **Language** English

☐ **Monograph** ☒ **Article thesis** ☐ **Essay thesis**

Abstract

This thesis studies computational tools for Bayesian modelling workflow. The focus is on two important areas. The first part of the thesis discusses the use of importance sampling for model assessment. Importance sampling is a generic computational tool that is used in many different applications. In model assessment, it is commonly used to speed up the computation of leave-one-out cross-validation. This thesis studies techniques for adapting the proposal distribution of importance sampling in order to improve the accuracy of leave-one-out cross-validation computations. To accomplish this, this thesis introduces a generic method for adapting a proposal distribution implicitly, which is thus applicable to a variety of complex distributions. The thesis also studies the special characteristics of adaptive importance sampling for self-normalized importance sampling, which is commonly used, for example, with probabilistic programming tools.

The thesis also studies importance sampling techniques for analyzing the sensitivity of Bayesian model posteriors to the choice of prior and likelihood. These methods can be useful for detecting various modelling issues, such as overly influential prior distributions. Importance sampling is beneficial for this purpose due to its simplicity, and these methods can be automated and used as part of Bayesian modelling workflow easily.

The second part of this thesis studies the assessment of the importance of variables in supervised learning. A common approach for generic variable importance assessment is to analyze the predictions of the model in real or transformed observations. This thesis presents methods for incorporating the predictive uncertainty of the model in such evaluation of variable importance. This thesis introduces the concept of uncertainty-aware sensitivity that generalizes sensitivity analysis from single predictions to probability distributions. The thesis develops an analytical framework for uncertainty-aware sensitivity as well as practical algorithms for its computation with different supervised learning models. The method is utilized to assess the importance of variables and variable interactions in Gaussian process models applied to different settings. The uncertainty-aware methods are especially useful for Gaussian processes and other flexible models where both variable importance and predictive uncertainty can vary significantly depending on the point of evaluation.

The contributions of this thesis are mostly methodological, and an integral part of the contribution is code written as part of the publications. While the thesis does not focus on any single application, the applicability of the studied methods is demonstrated with a variety of freely available data sets from different fields.

Keywords model assessment, importance sampling, variable importance

ISBN (printed) 978-952-64-0869-9 **ISBN (pdf)** 978-952-64-0870-5

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2022

Pages 126 **urn** http://urn.fi/URN:ISBN:978-952-64-0870-5

Tekijä
Topi Paananen

Väitöskirjan nimi
Laskennalliset menetelmät bayesilaisten mallien arvioinnissa

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL THESES 96/2022

Tutkimusala Tietotekniikka

Käsikirjoituksen pvm 04.03.2022 **Väitöspäivä** 15.09.2022

Väittelyluvan myöntämispäivä 11.05.2022 **Kieli** Englanti

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

Tiivistelmä

Tämä väitöskirja tutkii laskennallisia menetelmiä osana bayesilaista mallinnusta. Väitöskirja keskittyy kahteen pääaihealueeseen. Ensimmäinen näistä on painotusotannan käyttö bayesilaisten mallien arvioinnissa. Painotusotanta on laajasti käytetty ja yleispätevä laskennallinen menetelmä. Mallien arvioinnissa painotusotantaa käytetään yleisesti nopeuttamaan ristiinvalidoinnin laskentaa. Tämä väitöskirja tutkii painotusjakauman mukauttamismenetelmiä, joiden avulla ristiinvalidoinnin laskenta tarkentuu. Väitöskirja esittelee painotusjakaumien implisiittiseen mukauttamiseen soveltuvan menetelmän, jonka avulla monimutkaisia jakaumia voidaan hyödyntää tehokkaammin. Väitöskirja tutkii erityisesti painotusjakaumien mukauttamista itsenormalisoitavassa painotusotannassa, joka on yleisesti käytetty muun muassa probabilististen ohjelmointityökalujen yhteydessä.

Tämä väitöskirja tutkii myös painotusotannan käyttöä bayesilaisten mallien priorien ja uskottavuusfunktioiden herkkyyden analysoimiseen. Nämä menetelmät ovat hyödyllisiä havaitsemaan erilaisia mallinnusongelmia kuten liian rajoittavia priorijakaumia. Painotusotantamenetelmät ovat hyödyllisiä tähän tarkoitukseen niiden yksinkertaisuuden vuoksi, ja väitöskirjassa tutkitut menetelmät ovat helposti automatisoitavissa osaksi bayesilaista mallinnusprosessia.

Väitöskirjan jälkimmäinen osa tutkii muuttujien tärkeyden arviointia ohjatussa oppimisessa. Yleinen lähestymistapa tähän on tilastollisen mallin ennusteiden tarkastelu havaintopisteissä tai muunnelluissa havaintopisteissä. Tämä väitöskirja tutkii menetelmiä, joilla mallin ennusteiden epävarmuus voitaisiin hyödyntää tällaisessa tarkastelussa. Väitöskirja esittelee konseptin epävarmuustietoinen herkkyys, joka yleistää herkkyysanalyysin siten, että yksittäisten ennusteiden herkkyyden sijaan tarkastellaan ennustejakaumien herkkyyttä. Väitöskirja esittelee epävarmuustietoisen herkkyyden analyyttisen perustan sekä käytännöllisiä algoritmeja erilaisille tilastollisille malleille. Menetelmiä hyödynnetään muuttujien sekä niiden välisten vuorovaikutusten tärkeyden arviointiin gaussisissa prosesseissa useissa eri mallinnussovelluksissa. Epävarmuustietoista herkkyyttä hyödyntävät menetelmät ovat erityisen kiinnostavia gaussisissa prosesseissa sekä muissa joustavissa malleissa, sillä niissä sekä muuttujien tärkeys että ennusteiden epävarmuus voivat vaihdella merkittävästi tarkastelupisteestä riippuen.

Tämän väitöskirjan kontribuutiot ovat pääasiassa metodologisia, ja olennainen osa niitä on julkaisujen yhteydessä kirjoitettu lähdekoodi. Väitöskirja ei keskity mihinkään tiettyyn sovellukseen, mutta tutkittujen menetelmien soveltuvuutta havainnollistetaan käyttämällä useita julkisia datalähteitä.

Avainsanat mallien arviointi, painotusotanta, muuttujien tärkeys

ISBN (painettu) 978-952-64-0869-9 **ISBN (pdf)** 978-952-64-0870-5

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2022

Sivumäärä 126 **urn** http://urn.fi/URN:ISBN:978-952-64-0870-5

Preface

This doctoral thesis is the result of my work in the Probabilistic Machine Learning (PML) group in the Department of Computer Science in Aalto University between the years 2017 and 2022. The journey towards this thesis started with a cold email when I asked Prof. Aki Vehtari for guidance on a master's thesis. With a background in physics and mathematics, I took a leap of faith into the world of Bayesian modelling. What started as a master's thesis project quite naturally extended to a full doctoral thesis.

I am grateful to everyone who has been a part of my doctoral journey. Most of all, I thank my supervisor and instructor Prof. Aki Vehtari for all the guidance and help during these years. I thank you for trusting me by taking me under your supervision, and for giving me freedom to explore my own ideas and learn at my own pace. Your vast knowledge and support has inspired me and pushed me forward during the years. I also thank Prof. Lorin Crawford and Dr. Víctor Elvira for examining this thesis, and Dr. Elvira for agreeing to serve as an opponent in my public defence.

During my doctoral studies, I have been honoured to work with many inspiring and talented people. Prof. Michael Riis Andersen started as a postdoc in the PML group soon after I started my master's thesis work. From the start, you have been my role model of an innovative and methodical researcher. I thank you for your guidance and friendship, and I cherish the memories of our discussions both on and off work. I thank Dr. Sebastian Weber for all the guidance and mentorship during my internship at Novartis Pharma Ag. Even though the work did not end up in this thesis, the trip was a valuable experience in my doctoral journey. I also thank my other collaborators, who have inspired and helped me in numerous ways. Thank you Dr. Juho Piironen, Dr. Paul Bürkner, Noa Kallioinen, Alejandro Catalina Feliú, and Isaac Sebenius.

I am very grateful for having had the pleasure to share the joys and troubles of research with many colleagues during these years. Many of you have also become close friends of mine. I am also grateful to the service personnel at the department of Computer Science for their services and help in securing the conditions for my research work.

I express my thanks to my family and in particular to my parents who have always supported and encouraged me throughout my life. My most deepest gratitude I express to Tea Tähtinen for her unlimited support and patience, and for always believing in me.

Espoo, July 14, 2022,

Topi Paananen

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 Topi Paananen, Juho Piironen, Paul-Christian Bürkner and Aki Vehtari. Implicitly Adaptive Importance Sampling. *Statistics and Computing*, 31:16. 19 pages, February 2021.

II Noa Kallioinen, Topi Paananen, Paul-Christian Bürkner and Aki Vehtari. Detecting and diagnosing prior and likelihood sensitivity with power-scaling. Submitted to *a journal.* 21 pages, July 2021.

III Topi Paananen, Michael Riis Andersen and Aki Vehtari. Uncertainty-aware Sensitivity Analysis Using Rényi Divergences. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence,* volume 161 of Proceedings of Machine Learning Research, pages 1185-1194, July 2021.

IV Topi Paananen, Juho Piironen, Michael Riis Andersen and Aki Vehtari. Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS),* volume 89 of Proceedings of Machine Learning Research, pages 1743-1752, April 2019.

Author's Contribution

Publication I: "Implicitly Adaptive Importance Sampling"

The topic was proposed by Vehtari. Paananen designed and carried out all the experiments, and came up with methodological improvements to the initial idea. Paananen and Bürkner wrote the software together. Paananen had the main responsibility of writing the article while Piironen, Bürkner, and Vehtari reviewed and proposed suggestions to the manuscript.

Publication II: "Detecting and diagnosing prior and likelihood sensitivity with power-scaling"

The topic was proposed by Vehtari and Bürkner. Kallioinen improved the original idea, and Paananen came up with methodological additions. Kallioinen designed and carried out the experiments and case studies. Kallioinen implemented the majority of the software code. Paananen implemented the code for importance weighted moment matching. Kallioinen had the main responsibility of writing the article while Paananen, Bürkner and Vehtari reviewed and proposed suggestions to the manuscript.

Publication III: "Uncertainty-aware Sensitivity Analysis Using Rényi Divergences"

The methodological innovations are due to Paananen who also derived all the theoretical results and carried out the experiments. Paananen also had the main responsibility of writing the article while Andersen and Vehtari proposed suggestions to the manuscript.

Publication IV: "Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution"

The topic was proposed by Vehtari and Piironen. Paananen designed and carried out all the experiments. Paananen had the main responsibility of writing the article while Piironen, Andersen and Vehtari reviewed and proposed suggestions to the manuscript.

1. Introduction

1.1 Background and research environment

Model assessment is an integral part of statistical modelling and analysis workflow (Gelman et al., 2020). Workflow is a concept that encompasses not only statistical methods, but many aspects needed in the practical use of statistics, such as decision making, programming, and subject matter knowledge. Understanding good workflow practices is essential to be able to efficiently use statistical methods for solving real-world problems. Recognition of the importance of good workflow has increased in recent years in the statistical literature and other fields that rely on statistical methods (Turner & Lambert, 2015; Lee et al., 2019; Gabry et al., 2019; Gelman et al., 2020; Schad et al., 2021).

Importance sampling is a commonly used tool in Bayesian statistics. In model assessment, it is extensively used for computing out-of-sample predictive performance estimates of models with leave-one-out cross-validation (Gelfand et al., 1992; Vehtari et al., 2016; Vehtari et al., 2017; Bürkner et al., 2020, 2021). A notable recent advancement has been the ability to easily yet dependably diagnose the reliability of importance sampling leave-one-out cross-validation after its computation (Vehtari et al., 2017; Vehtari et al., 2020; Vehtari et al., 2021). If the diagnostics indicate inadequate accuracy, reliable and computationally efficient alternative methods to compute cross-validation have been lacking. For example, as alternative to importance sampling leave-one-out cross-validation, Vehtari et al. (2017) recommend sampling directly from the leave-one-out posterior distributions for unreliable cases, or resorting to K-fold cross-validation. However, both of these alternatives can be inaccurate similarly to importance sampling, and do not have diagnostics as readily available. Publication I addresses this research gap by studying adaptive importance sampling methods that can be used to improve results for those leave-one-out folds that are diagnosed as unreliable.

In Bayesian model checking, assessing the effect of the prior and likelihood on the posterior distribution of a model is a topic that has been studied for many decades (Canavos, 1975; Skene et al., 1986; Hill & Spall, 1994; Lopes & Tobias, 2011; Depaoli et al., 2020). However, recently, this model checking step has not been routinely seen in empirical studies that use Bayesian methods (van de Schoot et al., 2017). One of the possible reasons for its infrequent use is the lack of tools that can be used with modern modelling workflows. Many existing methods are specific to certain model types (Hunanyan et al., 2021; Roos et al., 2021), or require a substantial amount of manual tuning (Jacobi et al., 2018; McCartan, 2021). This thesis addresses the need for general purpose tools for prior and likelihood sensitivity analysis. The methods developed in Publication II are readily usable with modern probabilistic programming and Bayesian modelling tools such as the packages `posterior` (Bürkner et al., 2022), `rstan` (Stan Development Team, 2021), and `brms` (Bürkner, 2017) for the R language (R Core Team, 2021).

The latter part of this thesis studies model assessment in supervised learning, where the statistical relationship between predictors $\mathbf{x} = (x_1, ..., x_D)$ and a target variable y is inferred given a set of observation pairs $\mathscr{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$. In many cases, some of the predictor variables are unimportant for predicting the target variable, which means that ignoring some of the unimportant predictors may be beneficial. Even if none of the predictors are completely trivial, knowing their relative importance can improve understanding of the used model and data. Thus, variable importance assessment is an important part of the Bayesian statistical workflow for supervised learning models.

Variable importance can be assessed by analyzing the derivatives of a model's predictions (Ruck et al., 1990; Refenes & Zapranis, 1999; Leray & Gallinari, 1999; Simonyan et al., 2014; Sundararajan et al., 2017; Cui et al., 2020). This approach is attractive because of its flexibility, as it can be used with a variety of different supervised learning models, as opposed to methods that rely on certain properties of the model for evaluating variable importance. While variable selection has been studied extensively in the Bayesian context (see e.g. O'Hara & Sillanpää, 2009; Piironen & Vehtari, 2017), research into using derivatives of probabilistic predictions for variable importance has been lacking, and existing methods ordinarily rely on the derivatives of singular prediction values. Publications III and IV of this thesis focus on this research gap, and study methods for incorporating predictive uncertainty in the evaluation of variable importance with derivatives.

1.2 Objectives and scope

The statistical analysis workflow typically includes multiple steps, including model fitting, model checking, and model re-evaluation. Each step may depend on multiple algorithms or computational tools that perform computations to help the modeller. These tools exist on various levels of specificity and abstraction. The most basic tools, such as importance sampling, are simply mathematical operations that perform a specific computation. More advanced algorithms combine multiple low-level operations in order to perform the same task better. For example, adaptive importance sampling tools combine multiple importance sampling steps with other operations with the goal of computing an importance sampling task with better accuracy. Algorithms with different levels of complexity can be combined into modelling tools that hide some of the complexity by providing intuitive user interfaces and guidance for the user.

To make the complete statistical analysis workflow effective and enjoyable, the used tools need to be reliable, computationally efficient, and intuitive to use. With these objectives in mind, the high-level goal of this thesis is to study computational tools for different steps of the Bayesian modelling workflow. This thesis addresses three research questions, which are presented in this section. The section also describes how the research questions are connected to the publications that constitute this thesis. Next, the research questions are presented together with a summary of the extent to which this thesis addresses each research question. The contributions of the thesis are described in more detail in the subsequent chapters.

Research Question 1: *The posterior distribution of a Bayesian model is often a good proposal distribution for importance sampling leave-one-out cross-validation. For observations for which this is not the case, can accurate results be achieved without model refits by utilizing adaptive importance sampling?*

Research question 1 is addressed by Publication I, which studies adaptive importance sampling for leave-one-out cross-validation. The publication studies ways to make adaptive importance sampling easy and effective for Bayesian modelling where model posteriors are often approximated with Markov chain Monte Carlo methods. Publication I presents an importance weighted moment matching algorithm that uses an existing Monte Carlo sample as the starting point as opposed to the common approach of explicitly adapting parametric proposal distributions. The paper also compares these two approaches in multiple scenarios with different models. Publication I studies the properties of self-normalized importance sampling, and introduces a novel double adaptation procedure and a multiple importance sampling proposal distribution specifically suited for self-normalized

importance sampling. Self-normalized importance sampling is used often in Bayesian modelling, because it can be used without knowing the normalization constants of the distributions used.

Research Question 2: *Probabilistic programming tools offer an easy interface for Bayesian inference by generating Monte Carlo samples based on user-specified models. Can prior and likelihood sensitivity analysis be easily and reliably carried out without model refits, based on a probabilistic programming model and its result?*

Research question 2 is addressed by Publication II, which studies easily usable methods for prior and likelihood sensitivity analysis. In a similar fashion as in Publication I, Publication II uses importance sampling, and a Monte Carlo sample from the model posterior as the starting point for the analysis. The paper proposes a power-scaling sensitivity analysis method that requires only the ability to evaluate the densities of the model prior and likelihood, which are readily available with many probabilistic programming tools. The paper examines several alternative sensitivity measures for power-scaling, and compares their accuracy and interpretability. Publication II also considers adaptive importance sampling for improving the reliability of the sensitivity analysis results.

Research Question 3: *Derivatives of the predictions of a supervised learning model can be used as measures for the relative importance of predictor variables. Can variable importance be identified more accurately by taking into account the uncertainty of the predictions? Is such a measure useful for identifying important predictor variables from Gaussian process models?*

Research question 3 is addressed by Publications III and IV. Publication III introduces a novel framework for uncertainty-aware sensitivity analysis called R-sens. The framework generalizes derivative-based variable importance analysis to Bayesian models by differentiating the Rényi divergence of predictive distributions. Publication III demonstrates that this framework is extensible to discovering second-order interaction effects of variables. The uncertainty-aware sensitivity method is derived analytically for Gaussian process models with various observation models as well as some generalized linear models. Publication III compares the uncertainty-aware sensitivity to naive sensitivity methods in their ability to identify important variables and interactions in both simulated and real scenarios. Publication IV presents a simpler numerical approximation of the uncertainty-aware sensitivity method, which does not require differentiating predictive equations by hand. Publications III and IV both study the usefulness of uncertainty-aware sensitivity for ranking variables in Gaussian process models. Publication IV compares uncertainty-aware sensitivity to ranking based on the parameters of the Gaussian process

model, whereas Publication III compares it to several existing variable importance methods.

1.3 Thesis structure

This introductory part of the thesis is organized into the following chapters. Chapter 2 discusses the use of importance sampling for Bayesian model assessment, providing the theoretical background for research questions 1 & 2, and presenting the contributions of Publications I and II. In Chapter 3, the discussion shifts to variable importance assessment for supervised learning models. The chapter also discusses Gaussian process models, and summarizes the methodological and experimental contributions of Publications III and IV, both of which address research question 3. Finally, Chapter 4 concludes the introductory part of the thesis by discussing the significance of the results as well as limitations and directions for future research.

2. Importance sampling for model assessment

Importance sampling (IS) is a useful computational tool for many tasks in Bayesian modelling. This chapter discusses methods that utilize importance sampling for Bayesian model assessment. First, sections 2.1 and 2.2 discuss importance sampling and adaptive importance sampling to give the necessary background for Publications I and II. Section 2.3 introduces Bayesian leave-one-out cross-validation, and Section 2.4 discusses the results and contributions of Publication I towards research question 1. Finally, Section 2.5 discusses the use of importance sampling for prior and likelihood sensitivity analysis, which is the topic of Publication II. Section 2.6 then analyses research question 2, and presents the contributions of Publication II.

2.1 Importance sampling

After model inference, many quantities computed from a posterior distribution are expectations. The expectation of a function $h(\theta)$ over the probability density function $p(\theta)$ is

$$\mu = \mathbb{E}_p[h(\theta)] = \int h(\theta)p(\theta)\mathrm{d}\theta. \tag{2.1}$$

Here we assume that $h(\theta)$ is integrable with respect to $p(\theta)$. Using a sample of independent draws $\{\theta^{(1)},\theta^{(2)},...,\theta^{(S)}\}$ from $p(\theta)$, the *simple Monte Carlo* estimator of μ is

$$\hat{\mu}_{\mathrm{MC}} = \widehat{\mathbb{E}_p[h(\theta)]} = \frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)}), \text{ when } \theta^{(s)} \sim p(\theta).$$

After approximating the posterior distribution $p(\theta)$ by generating a sample of Monte Carlo draws, many model assessment steps require computation of expectations over multiple distributions $\{g_1(\theta), g_2(\theta), ...\}$ that are somehow similar to the posterior $p(\theta)$. For example, in leave-one-out cross-validation, one repeatedly computes expectations over posteriors with a

single observation left out (Gelfand et al., 1992; Peruggia, 1997; Epifani et al., 2008; Vehtari et al., 2017). A similar situation is encountered in both classical (Efron, 1979) and Bayesian Bootstrap (Rubin, 1981), and as we shall see in Section 2.5, also prior sensitivity analysis.

The computation of multiple expectations can be made faster with importance sampling by using the same sample of draws for multiple expectations (Owen, 2013; Elvira & Martino, 2021). There exist variations of importance sampling estimators. This thesis focuses mostly on self-normalized importance sampling, but Publication I briefly discusses the properties of different estimators. Self-normalized importance sampling is useful in most practical situations, because it can be used without knowing the normalization constants of the distributions used. Situations like this are commonplace in Bayesian modelling, for example when using probabilistic programming tools or Markov chain Monte Carlo sampling algorithms.

The self-normalized importance sampling (SNIS) estimator of the expectation of a function $h(\theta)$ over the distribution $g_k(\theta)$ is

$$\hat{\mu}_{\text{SNIS}} = \widehat{\mathbb{E}_{g_k}[h(\theta)]} = \frac{\sum_{s=1}^{S} \frac{g_k(\theta^{(s)})}{p(\theta^{(s)})} h(\theta^{(s)})}{\sum_{s=1}^{S} \frac{g_k(\theta^{(s)})}{p(\theta^{(s)})}}$$

$$= \frac{\sum_{s=1}^{S} w^{(s)} h(\theta^{(s)})}{\sum_{s=1}^{S} w^{(s)}}, \text{ when } \theta^{(s)} \sim p(\theta). \tag{2.2}$$

Here, $g_k(\theta)$ is often called the target distribution, and $p(\theta)$ is the proposal distribution. $w^{(s)}$ are importance weights, which represent the density ratio of the target $g_k(\theta)$ and the proposal $p(\theta)$. Importance sampling is widely used also outside Bayesian modelling, with applications ranging from rare event estimation (Rubino & Tuffin, 2009) to optimal control (Kappen & Ruiz, 2016).

The performance of importance sampling depends critically on the choice of the proposal distribution. In self-normalized importance sampling, there are two criteria that define performance. First, the proposal must be a good match to the target distribution. Second, the proposal must match the function whose expectation is being computed. While the former criterion only concerns importance sampling, the latter is important for any Monte Carlo estimator. This issue is sometimes overlooked, even though it can be a significant source of error, as discussed in Publication I.

Many strategies have been proposed to reduce the variance of importance sampling by performing transformations, such as truncation or smoothing, to the importance weights (Ionides, 2008; Koblents & Míguez, 2015; Miguez et al., 2018; Vehtari et al., 2021). In many situations, these techniques are beneficial even though they introduce bias in the estimate. In addition to stabilizing the estimate, Vehtari et al. (2021) propose a diagnostic for the practical pre-asymptotic convergence rate. By estimating

the pre-asymptotic convergence rate, the diagnostic can identify whether sampling can be made accurate with reasonable computational cost. The Pareto $\hat{k}$ diagnostic of Vehtari et al. (2021) is an essential tool for the methods discussed in Publications I and II. There are also a variety of other importance sampling diagnostics, such as the effective sample size and its variations (Martino et al., 2017; Agarwal et al., 2021).

2.2 Adaptive importance sampling

As discussed in Section 2.1, choosing a good proposal distribution is critical for the feasibility of importance sampling. In Bayesian model assessment, if a sample of Monte Carlo draws has been generated from the model posterior as part of inference, it is a natural first choice as the proposal. To make this proposal better for a specific computational task, it can be adapted for the task at hand. The procedure of improving the proposal distribution in importance sampling is called adaptive importance sampling (AIS) (Cappé et al., 2004; Cornuet et al., 2012; Martino et al., 2015; Bugallo et al., 2017). As discussed in Publication I, the principles of many AIS algorithms can be summarized to three steps:

1. generating draws from the proposal distribution(s)

2. computing importance weights for each draw

3. adapting the proposal distribution(s) based on the importance weights.

These three steps are typically repeated until a pre-defined convergence or stopping criterion is reached. Importance sampling diagnostics (Martino et al., 2017; Vehtari et al., 2021; Agarwal et al., 2021) can be valuable for determining the progress of AIS methods, which is also demonstrated in Publications I and II.

To enable resampling (step 1) in every iteration, many AIS methods use proposal distributions with an analytical representation that are easy to sample from, for example Gaussian or Student-t distributions. One of the conclusions of Publication I is that such standard distributions are not always good enough when posterior distributions are high-dimensional or have non-standard form. Publication I uses an implicitly adaptive importance sampling algorithm that does not resample between iterations, and can therefore be used with more generic proposal distributions. For example, draws from a posterior obtained with Markov chain Monte Carlo methods can be used as long as the posterior density is computable up to an unknown normalization constant. The implicit adaptation means only transforming the Monte Carlo draws, which is implicitly equal to

adapting the proposal distribution. Publication I uses the probabilistic programming tool Stan (Stan Development Team, 2022) to demonstrate that implicitly adaptive importance sampling can be easily coupled with probabilistic programming tools.

Publication I presents an implicitly adaptive importance sampling algorithm called importance weighted moment matching (IWMM). The algorithm uses a combination of affine transformations which match the standard moments of a proposal sample to its importance weighted moments, approximately matching the proposal distribution to the target distribution. Analogous implicit matching transformations have been discussed in different contexts, for example by Voter (1985) and Meng and Schilling (2002). The main benefit of the moment matching affine transformations is their simplicity, as the procedure can be automated and used similarly for a wide variety of models. On the other hand, as the target moments are computed using importance sampling, it is possible that they are not accurate enough for successful moment matching. Moreover, even if moment matching is successful, there is no guarantee that it will result in a better proposal distribution. The IWMM algorithm uses the Pareto $\hat{k}$ diagnostic (Vehtari et al., 2021) to measure the progress and determine a stopping criterion for the adaptation. For a more detailed description, see Publication I.

2.2.1 Double adaptation

Typically, the goal of adaptive importance sampling is to find a proposal which is close to the distribution over which the expectation is defined, hence also the name target distribution. Therefore, many AIS methods set this distribution as the target of adaptation by using the importance ratios (as defined in equation (2.2)) to guide the adaptation. Publication I discusses the optimal proposal distributions for different Monte Carlo estimators (Kahn & Marshall, 1953; Hesterberg, 1988), and emphasizes that the distribution over which the expectation is defined is not always a good target for AIS. Based on this remark, Publication I proposes to use the expectation-specific optimal proposal distribution as the target of adaptation.

For the expectation $\mathbb{E}_p[h(\theta)]$, the optimal proposal for self-normalized importance sampling is proportional to

$$g_{\text{SNIS}}^{\text{opt}}(\theta) \propto p(\theta)|h(\theta) - \mathbb{E}_p[h(\theta)]|. \tag{2.3}$$

This optimal density can be significantly more complex than for other estimators because of the requirement to simultaneously estimate both the numerator and denominator of equation (2.2). Publication I discusses this in detail and proposes a double adaptation procedure for SNIS that combines two implicitly adapted proposal distributions into a single proposal.

This is called the *split proposal*, because it is based on an approximation of the piecewise defined density of equation (2.3) that is split into two separate components:

$$g_{\text{SNIS}}^{\text{split}}(\theta) \propto p(\theta)|h(\theta)| + p(\theta)\mathbb{E}_p[h(\theta)]. \tag{2.4}$$

The split proposal is based on a technique called multiple importance sampling, which allows combining multiple distributions into a single proposal distribution (see e.g. Elvira et al., 2019). Publication I also shows that the double adaptation can be a valuable improvement to standard AIS methods that use parametric proposal distributions. Adaptive importance sampling specifically for self-normalized importance sampling has been recently studied also by Lamberti et al. (2018) and Rainforth et al. (2020).

2.3 Bayesian leave-one-out cross-validation

An important part of the Bayesian modelling workflow is the assessment of predictive performance. It is often useful by itself, but it can also be used for model comparison or improvement (see for example Vehtari and Lampinen (2002), Vehtari and Ojanen (2012), Gelman et al. (2013), Vehtari et al. (2017)) or model averaging (Geisser & Eddy, 1979; Ando & Tsay, 2010; Yao et al., 2018). Predictive performance is typically measured using a *utility* or *cost*, and there exist a wide range of possible measures. One generally applicable utility measure is the log predictive density (also known as the log-score) (Good, 1952). It is widely used because of its desirable properties such as strict propriety (Bernardo, 1979; Geisser & Eddy, 1979; Bernardo & Smith, 1994; Gneiting & Raftery, 2007; Vehtari & Ojanen, 2012; Krüger et al., 2021).

Consider we have an observed data set y_{obs} which is used to fit a model. Consider also a true data generating mechanism with probability distribution $p_{\text{true}}(y)$, and N observations $y = (y^{(1)}, ..., y^{(N)})$ from the true data generating mechanism. Using the log-score, the out-of-sample predictive performance measure for a model fitted on y_{obs} for a new possible data set can be evaluated with the expected log pointwise predictive density (ELPD)

$$\text{ELPD} = \sum_{i=1}^{N} \int p_{\text{true}}(y^{(i)}) \log p(y^{(i)} \mid y_{\text{obs}}) \mathrm{d}y^{(i)}, \tag{2.5}$$

where $p(y^{(i)} \mid y_{\text{obs}})$ is the posterior predictive density of the evaluated model that is fitted on the data y_{obs}.

The ELPD measure defined in equation (2.5) includes the true data generating mechanism which is generally unknown. In practice, ELPD is very often approximated with the observed data set y_{obs}. To avoid using

the same data for inference and evaluation, procedures such as cross-validation (CV) that partition the data are commonly used. In CV, the data is partitioned into subsets, and one set is left out of inference to be used as an out-of-sample validation set. The inference and validation is then repeated using the different subsets for evaluation in turn. When the data is split into K parts, the procedure is called K-fold CV. In leave-one-out cross-validation (LOO-CV), $K = N$, meaning that only one observation is left out for evaluation at a time, and model fitting is repeated N times. The ELPD estimator based on LOO-CV is

$$\widehat{\mathrm{ELPD}}_{\mathrm{LOO}} = \sum_{i=1}^{N} \log \int p(y_{\mathrm{obs}}^{(i)} \mid \theta) p(\theta \mid y_{\mathrm{obs}}^{(-i)}) \, \mathrm{d}\theta = \sum_{i=1}^{N} \log p(y_{\mathrm{obs}}^{(i)} \mid y_{\mathrm{obs}}^{(-i)}), \tag{2.6}$$

where $y_{\mathrm{obs}}^{(i)}$ refers to observation i, and $y_{\mathrm{obs}}^{(-i)}$ to all other observations.

2.3.1 Importance sampling leave-one-out cross-validation

Naively evaluating the LOO-CV estimator of equation (2.6) with Monte Carlo methods requires generating samples from each of the N LOO posteriors separately. This can be very costly when N is large and when the model fitting is expensive. Fortunately, this cost can be significantly reduced with importance sampling.

With observations that are independent conditionally on the model parameters θ, it is straightforward to compute leave-one-out cross-validation with importance sampling by using draws $\{\theta^{(1)}, \theta^{(2)}, ..., \theta^{(S)}\}$ from the model posterior with all observations included, $p(\theta \mid y_{\mathrm{obs}})$ (Gelfand et al., 1992). In this case, the self-normalized importance sampling estimator of the ELPD of observation i is

$$\widehat{\mathrm{ELPD}}_{\mathrm{LOO},i} = p(y_{\mathrm{obs}}^{(i)} \mid y_{\mathrm{obs}}^{(-i)}) \approx \frac{\frac{1}{S} \sum_{s=1}^{S} w_{\mathrm{loo},i}^{(s)} \, p(y_{\mathrm{obs}}^{(i)} \mid \theta^{(s)})}{\frac{1}{S} \sum_{s=1}^{S} w_{\mathrm{loo},i}^{(s)}} = \frac{1}{\frac{1}{S} \sum_{s=1}^{S} w_{\mathrm{loo},i}^{(s)}}. \tag{2.7}$$

Here, $p(y_{\mathrm{obs}}^{(i)} \mid \theta^{(s)})$ is the likelihood, and $w_{\mathrm{loo},i}^{(s)}$ are the unnormalized importance weights, which are simply

$$w_{\mathrm{loo},i}^{(s)} = \frac{1}{p(y_{\mathrm{obs}}^{(i)} \mid \theta^{(s)})} \left(\propto \frac{p(\theta^{(s)} \mid y_{\mathrm{obs}}^{(-i)})}{p(\theta^{(s)} \mid y_{\mathrm{obs}})} \right). \tag{2.8}$$

Typically the normalization constants of the posterior distributions are unknown, and the importance weights are thus the ratio of the target distribution (LOO posterior) and the proposal distribution (full data posterior) up to an unknown constant. IS LOO-CV can also be used in many situations where the likelihood does not fully factorize, such as Gaussian latent variable models (Vehtari et al., 2016), time-series models (Bürkner et al., 2020), and non-factorized normal or Student-t models (Bürkner et al., 2021).

2.4 Research question 1 results

While IS can make LOO-CV faster by re-using the full data posterior draws for multiple expectations, it is not always accurate. It is traditionally thought that instability of IS LOO-CV estimators is the result of the fact that the full data posterior likely has a smaller variance than the LOO posterior. Publication I studies the causes of the instability in detail and concludes that similar stability issues can arise also in naive LOO-CV when using draws from LOO posteriors, a phenomenon that often tends to get overlooked. Publication I presents multiple situations where both IS LOO-CV and naive LOO-CV are extremely unstable. A typical example is a situation where one observation is an outlier that is very unexpected from the perspective of the model. In these cases, the double adaptation scheme proposed in Publication I (see Section 2.2.1) is crucial for accurate model assessment.

Publication I discusses the special structure of the LOO-CV expectation when using the log-score utility. It is special because the posterior conditioned on all observations is an optimal proposal distribution for the numerator of the self-normalized importance sampling estimator, which is also seen in equation (2.7) where the numerator simplifies to 1. This property makes it especially favourable to use the full data posterior distribution as the initial proposal in adaptive importance sampling. Moreover, the double adaptation can be avoided and a single adaptation targeting the LOO posterior distribution is enough. However, using the split proposal discussed in Section 2.2.1 is still important, because the full data posterior adapted towards the LOO posterior is no longer optimal for the numerator.

Consider the following illustrative example data from Publication I. We generate 29 observations from a standard normal distribution, and set the value for a 30'th observation to represent an outlier observation. The data is modelled with a normal distribution with unknown mean and variance. We generate draws from the posterior distribution and evaluate the model with leave-one-out cross-validation and logarithmic score. We evaluate the ELPD score using both naive LOO-CV (sampling from the LOO posterior) and IS LOO-CV (sampling from the full data posterior). The experiment is repeated with outlier values from 3 to 14. The key question to consider is: can either of these proposals be relied on for accurately estimating the model's ability to predict the outlier, i.e. computing $\widehat{\mathrm{ELPD}}_{\mathrm{LOO},30}$?

Figure 2.1 shows the results of the example. The left plot shows the $\widehat{\mathrm{ELPD}}_{\mathrm{LOO},30}$ as a function of the value of the outlier observation. For this simple model, the $\mathrm{ELPD}_{\mathrm{LOO},30}$ can be computed analytically, and it is depicted with a cross symbol. PSIS refers to Pareto smoothed importance sampling using the full data posterior (Vehtari et al., 2017), and PSIS+IWMM refers to using the importance weighted moment matching algorithm from Publication I in addition to PSIS. Naive means sampling

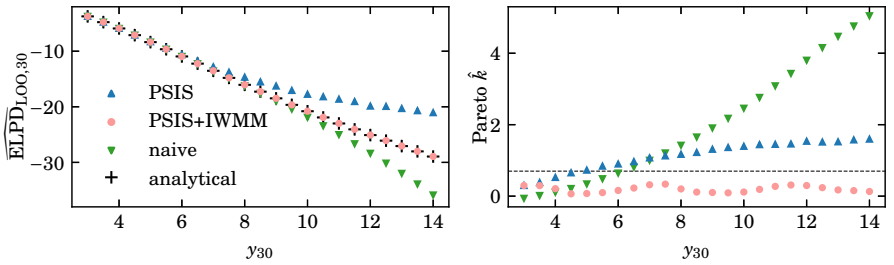

Figure 2.1. Illustration of the accuracy of LOO-CV computation for an outlier observation (adapted from Publication I). The data has 29 random observations generated from a standard normal distribution, and the value of y_{30} is varied. The data is modelled as normally distributed. The left plot shows the $\widehat{\mathrm{ELPD}}_{\mathrm{LOO},30}$ estimates and the right plot the Pareto $\hat{k}$ diagnostic value as a function of the outlier observation y_{30}. The black crosses depict the true analytical $\mathrm{ELPD}_{\mathrm{LOO},30}$ values. The sampling results are averaged from 50 independent repetitions. Error bars representing 95% intervals of the mean over the repetitions are almost indistinguishable. The dashed line at $\hat{k} = 0.7$ represents the diagnostic threshold, below which the results can be considered practically useful.

from the LOO posterior and computing ELPD with simple Monte Carlo sampling. The plot illustrates that as the outlier observation becomes more unexpected for the model, both the naive and PSIS estimates become biased in opposite directions. PSIS+IWMM estimate using the split proposal can correct the bias and give accurate results. Publication I shows that the same can be achieved with naive+IWMM, but PSIS+IWMM is typically cheaper to evaluate. The right plot of Figure 2.1 presents the Pareto $\hat{k}$ diagnostic (Vehtari et al., 2021) value for each estimator, which correctly indicates when the sampling is unreliable. The dashed line at $\hat{k} = 0.7$ represents the diagnostic threshold, below which the results can be considered practically useful (Vehtari et al., 2017). In this example, a more conservative value $\hat{k} = 0.5$ was used as the threshold for stopping IWMM.

To give a better illustration of the root cause of the biased results, we can compare the different proposal distributions to the theoretical optimal proposal given by equation (2.3). Figure 2.2 shows the discrepancy between the optimal proposal and three candidate proposal distributions: the full data posterior (used by PSIS), the LOO posterior (used by naive), and a split proposal constructed from the full data and LOO posteriors (approximated by PSIS+IWMM). In this figure, the model is simplified so that the data variance is fixed to 1, and the model posterior contains only a single parameter, the mean of the data. From the figure, it is evident that as the outlier becomes more exceptional to the model, neither the full data posterior nor the LOO posterior can match both tails of the optimal proposal, which would be essential for reliable sampling. On the contrary, the behaviour of the split proposal is the opposite, and it matches the optimal proposal better as the outlier becomes more exceptional.

Publication I shows with several additional data sets and models that

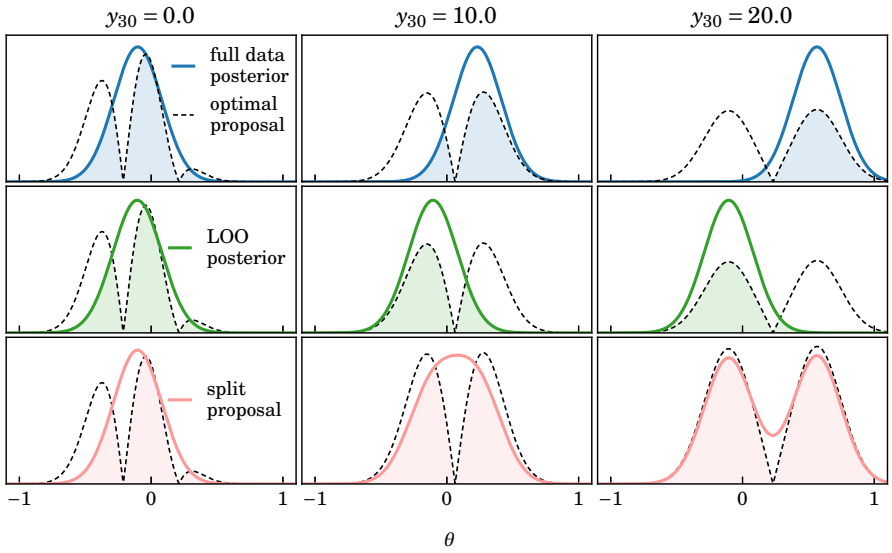

Figure 2.2. Illustration of the discrepancy between the theoretically optimal proposal distribution for self-normalized importance sampling, and three possible proposal distributions, when the data has an outlier observation (adapted from Publication I). The model is a Gaussian model with unknown mean and variance 1. The data has 29 observations generated from a standard normal distribution, and a 30'th observation y_{30} with three different values represented by the three columns. The top row shows the full data posterior distribution, the middle row the LOO posterior, and the bottom row the split proposal distribution constructed from the full data and LOO posteriors. The shaded areas represent the overlapping probability densities of the optimal proposal and each proposal candidate.

the sampling instability is an issue for both naive LOO-CV and IS LOO-CV. In many cases, more reliable cross-validation estimates are obtained with IWMM. Thus, adaptive importance sampling is not only a computationally cheap alternative to sampling directly from the leave-one-out posterior of each unreliable case, but also often more accurate. In order to be easily usable, the IWMM algorithm is implemented as part of the R-package `loo` (Vehtari et al., 2020).

Publication I also compares IWMM, that uses the full data posterior as the initial proposal distribution, to similar adaptive importance sampling with Gaussian or Student-t proposal distributions. Especially with high-dimensional posteriors, IWMM performs significantly better than the parametric proposals. Thus, one of the conclusions is that the full data posterior is a good initial proposal distribution for adaptive importance sampling in leave-one-out cross-validation.

2.5 Prior and likelihood sensitivity

Choosing the prior and likelihood for a model are some of the cornerstones of Bayesian inference. Analyzing the sensitivity of the posterior to changes in both of these components is thus an important part of model assessment (Lopes & Tobias, 2011; Depaoli et al., 2020). This type of sensitivity analysis can uncover issues such as prior-data conflict (Evans & Moshonov, 2006; Al Labadi & Evans, 2017; Reimherr et al., 2021) and likelihood noninformativity (Gelman et al., 2017).

The sensitivity of the posterior to changes in the likelihood and prior can be analyzed by perturbing the prior or likelihood, and quantifying the change inflicted on the posterior distribution. An abnormally high sensitivity can be further analyzed to evaluate its cause and determine whether it is acceptable or if a change in the model is required.

Prior-data conflict can be caused by informative priors that disagree with the observed data, but are not completely dominated by the likelihood. Such priors can be intentional, but if not, this can be a reason to reconsider the choice of priors. Prior-data conflict typically makes the posterior sensitive to changes in both the prior and likelihood. Likelihood noninformativity, on the other hand, can be caused by the likelihood exhibiting less information than the prior, for example because of imbalanced data. This is seen as the posterior being sensitive to changes in the prior, but insensitive to likelihood perturbations. Such difference in sensitivity can also be caused by the formulation of the model and not the data. For example, a model with a Student-t likelihood and Gaussian prior can exhibit likelihood noninformativity just because of their different shape. Publication II presents a comprehensive explanation of the different combinations of prior and likelihood sensitivity.

2.5.1 Power-scaling

Publication II presents a widely applicable procedure for assessing prior and likelihood sensitivity, which is based on perturbing both components separately by raising them to some power $\alpha > 0$. This can be interpreted as a distribution agnostic way to strengthen ($\alpha > 1$) or weaken ($\alpha < 1$) the power-scaled term relative to the other, while keeping the support of the scaled component unchanged.

The main benefit of power-scaling is its simplicity – it is quick to compute and easy to use with a large variety of different models without model-specific tuning. However, even though power-scaling is in principle distribution agnostic, the way it is used for the prior and likelihood may differ depending on the structure of the model. Publication II discusses this aspect in more detail especially related to hierarchical models. Power-scaling is also intuitive – for many standard probability distributions, it

simply alters the parameters of the distribution, but keeps the functional form the same. For examples of the results of power-scaling on some common distributions, see Publication II.

2.5.2 Importance sampling for sensitivity analysis

Like Publication I, Publication II also focuses on the setting where the model posterior is approximated using Monte Carlo methods. Typically, to understand the implications of power-scaling, a range of α values needs to be analyzed. For example, power-scaled priors ranging from $\alpha = 0.5$ to $\alpha = 2$ represent a range of alternative priors that could be considered instead of the specific prior ($\alpha = 1$) that the model uses, which is always more or less subjective. If the cost of sampling from the posterior is high, sampling from a range of power-scaled posteriors can be costly. Importance sampling is a suitable tool to avoid this cost by re-using the Monte Carlo sample generated from the original model.

With the posterior distribution of the original model as the proposal, and considering a posterior with power-scaled prior as the target distribution, the importance ratios are given as

$$w^{(s)}_{\alpha,\text{prior}} = \frac{p_{\text{prior}}(\theta^{(s)})^{\alpha} p(y \mid \theta^{(s)})}{p_{\text{prior}}(\theta^{(s)}) p(y \mid \theta^{(s)})} = p_{\text{prior}}(\theta^{(s)})^{\alpha-1}. \tag{2.9}$$

When scaling the likelihood, the importance weights are

$$w^{(s)}_{\alpha,\text{likelihood}} = p(y \mid \theta^{(s)})^{\alpha-1}. \tag{2.10}$$

Importance sampling enables quick computation of various sensitivity measures. For example, the mean and standard deviation of the power-scaled posteriors are easy-to-understand quantities for analysing the effect of power-scaling. More elaborate measures, such as the cumulative Jensen-Shannon divergence can be used as well. See Publication II for a more in-depth discussion on the benefits of different approaches. In addition to analyzing a range of power-scaled posteriors, Publication II also presents a method for assessing the general sensitivity of the current model ($\alpha = 1$) with importance sampling. This is achieved by computing the derivative of a sensitivity measure at $\alpha = 1$ for a specific summary of the posterior.

2.6 Research question 2 results

Publication II demonstrates that the power-scaling method is a viable solution to Research question 2. Power-scaling is utilized in multiple case studies and compared to alternative prior and likelihood sensitivity analysis methods. The conclusion of Publication II is that power-scaling can identify prior and likelihood sensitivity from different models and data

where sensitivity is known to exist. The ease of use is demonstrated by using the methods in combination with modern probabilistic programming and Bayesian modelling tools such as the packages `posterior` (Bürkner et al., 2022) and `brms` (Bürkner, 2017) for the R language (R Core Team, 2021). The method itself is also implemented in a freely available R-package `priorsense`.

Section 2.4 discussed the instability of importance sampling, which can be diagnosed for example with the Pareto $\hat{k}$ diagnostic (Vehtari et al., 2021). Instability can hinder the use of importance sampling also for power-scaling. Publication II utilizes the Pareto diagnostic, and uses the IWMM adaptive importance sampling method from Publication I if the diagnostic indicates inadequate accuracy. The IWMM framework is very generic, and thus the basic principles apply to power-scaled posteriors similarly as to leave-one-out cross-validation, although some nuances are different. In LOO-CV, the ELPD estimator can be very difficult to compute even naively, i.e. using the LOO posterior as the proposal distribution, as demonstrated in Section 2.4. This is because the function whose expectation is computed, the likelihood of the left-out observation, can attain very high values in the tails of the LOO posterior. In power-scaling, we typically compute expectations of simpler functions, such as the first few moments to summarize the power-scaled distributions. In power-scaling, the difficulties arise typically simply due to the mismatch of the proposal distribution (original model posterior) and the power-scaled posteriors.

To illustrate why this may happen, let us revisit the simulation example from Section 2.4 and Figure 2.2. Instead of setting a single observation as an outlier, let us simply generate all 30 observations randomly from a standard normal distribution. Let us again model the data with a Gaussian likelihood with unknown mean and variance of 1. The only unknown parameter is the data mean, for which we set a Normal(0,10) prior.

We could formulate the power-scaled model using e.g. probabilistic programming, and generate separately draws from power-scaled posteriors with different values of α. However, simply re-using the draws from the original posterior and using importance sampling is much faster to compute. Therefore, let us now consider the question: how good is the original model posterior distribution as a proposal distribution for power-scaling, specifically for computing the mean of the distribution whose likelihood is power-scaled? We then scale the likelihood by raising it to different powers of α, and attempt to compute the mean of the power-scaled posteriors.

Figure 2.3 shows the comparison of the optimal proposal distribution from equation (2.3) along with the original posterior and the power-scaled posterior for three different α values. The figure shows that when $\alpha < 1$, the power-scaled posterior becomes wider than the original posterior, making importance sampling less reliable. The power-scaled posterior in the

bottom row widens similarly to the optimal proposal, and remains a good proposal distribution with different values of α. This means that the use of double adaptation and the split proposal (see Section 2.2.1) is rarely necessary, and it is enough to use the power-scaled posterior as the target of adaptive importance sampling.

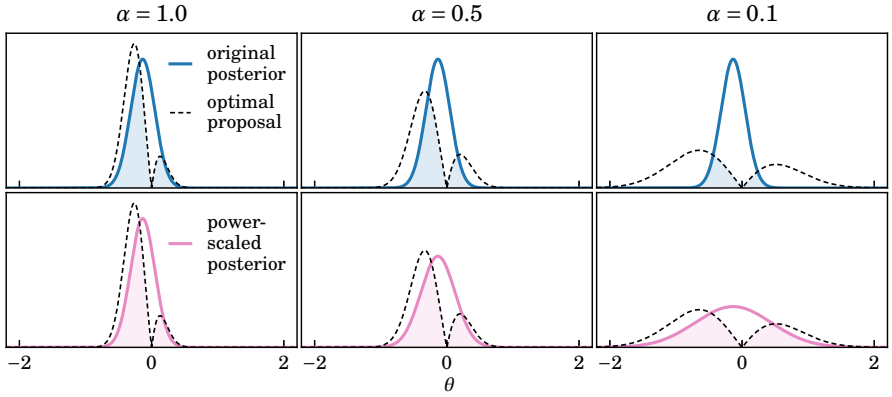

Figure 2.3. Illustration of the discrepancy between the theoretically optimal proposal distribution for SNIS, and two possible proposal distributions: the original posterior and the power-scaled posterior. The model is a Gaussian model with unknown mean and known variance. The data consists of 30 observations from a standard normal distribution. The columns correspond to three different values for the power-scaling parameter α, and the dashed line is the optimal proposal for computing the mean of the power-scaled posterior. The shaded areas represent the probability density difference of the optimal proposal and the proposal candidates.

3. Variable importance sensitivity analysis

In supervised learning, it is often interesting to study how important each predictor variable is for predicting the target variable. Such a task is called variable importance assessment, although it can also be framed as variable selection where the goal is to explicitly choose a subset of the most important variables. Variable importance assessment is often an important part of the Bayesian statistical workflow for supervised learning models. This chapter introduces the concepts of Publications III and IV, which study variable importance from a Bayesian perspective, such that the predictive uncertainty of the model is taken into account.

First, Section 3.1 gives a brief introduction to derivative based variable importance. Section 3.2 then introduces the methods presented in Publications III and IV. Section 3.3 introduces the basic concepts of Gaussian process models, and presents the implementation of uncertainty-aware sensitivity method for Gaussian processes. Finally, Section 3.4 discusses the contributions of Publications III and IV towards research question 3.

3.1 Variable importance

The term variable importance is often used to refer to generic methods that use the predictions of a model to assess the importance of predictors in supervised learning models. Consequently, they are also called model-agnostic methods. There exist a myriad different approaches to assessing variable importance (Wei et al., 2015; Molnar, 2019), but they can be loosely categorized into global and local methods. Global methods attempt to describe the average contribution of variables, whereas local methods describe the contribution in a specific observation or prediction. The methods of Publications III and IV fall in the category of local methods. However, the distinction of local and global methods is not restrictive, since local methods are often used to also evaluate the average contribution of variables, also in Publications III and IV.

One sub-category of local variable importance methods are derivative

based sensitivity analysis methods (Ruck et al., 1990; Refenes & Zapranis, 1999; Leray & Gallinari, 1999; Simonyan et al., 2014; Sundararajan et al., 2017; Cui et al., 2020). Consider a supervised learning model trained on a set of observation pairs $\mathscr{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$. Let us denote the prediction function of the model for the target variable $\tilde{y}$ at test point $\tilde{\mathbf{x}}$ as $f(\tilde{\mathbf{x}})$. Derivative based methods assess the sensitivity of f to changes in the predictors, which is quantified by the partial derivative with respect to a single predictor $\tilde{x}_d$, $\frac{\partial f(\tilde{\mathbf{x}})}{\partial \tilde{x}_d}$. The absolute value, or square, is typically used to give equal importance to positive and negative derivatives. The local importance values can also be summed over the observed $\mathbf{x}$ values to compute an estimate of the global importance of x_d. This is called the expected absolute derivative (EAD) (see for example Cui et al. (2020))

$$\mathrm{EAD}(x_d) = \mathrm{E}_{p(\mathbf{x})}\left[\left|\frac{\partial f(\mathbf{x})}{\partial x_d}\right|\right]. \tag{3.1}$$

3.2 Uncertainty-aware sensitivity

From a Bayesian perspective, the derivative-based methods have a major shortcoming as they only use a single prediction value, and ignore the uncertainty of the prediction. The aim of the uncertainty-aware sensitivity analysis discussed in Publications III and IV is to solve this shortcoming. This is achieved by selecting a statistical divergence measure, and differentiating this measure computed from the predictive distribution. An example of a spectrum of suitable divergences that measure the dissimilarity of two probability distributions are the Rényi divergences, which are studied in Publication III. Rényi divergences are parameterized by an order parameter α which defines the properties of the divergence. In practice, both Publications III and IV use the Kullback-Leibler divergence (KLD), which is the Rényi divergence obtained when the order parameter α approaches 1 (Kullback & Leibler, 1951).

To illustrate the uncertainty-aware sensitivity, let us define a predictive distribution $p(\tilde{y} \mid \boldsymbol{\lambda}(\tilde{\mathbf{x}}))$, where $\tilde{y}$ is the target variable, and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_M)$ are parameters that are conditioned on predictor values $\tilde{\mathbf{x}}$. The goal is to evaluate how sensitive the predictive distribution is to infinitesimal variations in $\tilde{\mathbf{x}}$, which is achieved by differentiating the Rényi divergence between two predictive distributions in the limit where they coincide:

$$\left.\frac{\partial^2 \mathscr{D}_\alpha^p[\tilde{\mathbf{x}}']}{(\partial \tilde{x}_d')^2}\right|_{\tilde{\mathbf{x}}'=\tilde{\mathbf{x}}} = \left(\frac{\partial \boldsymbol{\lambda}(\tilde{\mathbf{x}})}{\partial \tilde{x}_d}\right)^{\mathsf{T}} \mathbf{H}_{\boldsymbol{\lambda}(\tilde{\mathbf{x}}')}(\mathscr{D}_\alpha^p[\tilde{\mathbf{x}}']) \left(\frac{\partial \boldsymbol{\lambda}(\tilde{\mathbf{x}}')}{\partial \tilde{x}_d'}\right)\Bigg|_{\tilde{\mathbf{x}}'=\tilde{\mathbf{x}}}. \tag{3.2}$$

Here, $\mathscr{D}_\alpha^p[\tilde{\mathbf{x}}'] \equiv \mathscr{D}_\alpha[\, p(\tilde{y} \mid \boldsymbol{\lambda}(\tilde{\mathbf{x}})) || p(\tilde{y} \mid \boldsymbol{\lambda}(\tilde{\mathbf{x}}'))]$ is the Rényi divergence of order α from one predictive distribution to another, and $\mathbf{H}_{\boldsymbol{\lambda}(\tilde{\mathbf{x}}')}(\mathscr{D}_\alpha^p[\tilde{\mathbf{x}}'])$ is the Hessian of the Rényi divergence. The Hessian can be approximated with

the Fisher information matrix $\mathrm{I}(\boldsymbol{\lambda}(\tilde{\mathbf{x}}))$ of the predictive distribution $p(\tilde{y} \mid \boldsymbol{\lambda}(\tilde{\mathbf{x}}))$ (Kullback, 1959; Haussler & Opper, 1997; van Erven & Harremos, 2014). Based on this derivative, Publication III introduces the uncertainty-aware sensitivity measure called R-sens. Based on equation (3.2), the measure for predictor x_d evaluated at $\tilde{\mathbf{x}}$ is defined as

$$\text{R-sens}(\tilde{\mathbf{x}}, x_d, \alpha) \equiv \sqrt{\alpha \left(\frac{\partial \boldsymbol{\lambda}(\tilde{\mathbf{x}})}{\partial \tilde{x}_d} \right)^{\mathsf{T}} \mathrm{I}(\boldsymbol{\lambda}(\tilde{\mathbf{x}})) \left(\frac{\partial \boldsymbol{\lambda}(\tilde{\mathbf{x}})}{\partial \tilde{x}_d} \right)}. \tag{3.3}$$

Here α is the order of Rényi divergence used.

The connection of equation (3.3) to the standard derivative based variable importance $\left| \frac{\partial f(\tilde{\mathbf{x}})}{\partial \tilde{x}_d} \right|$ is not immediately obvious. However, for certain distribution types, equation (3.3) is easily interpretable and the two formulas are comparable. For example, consider a predictive distribution in the location-scale family (for example Gaussian or Student-t), where $\boldsymbol{\lambda}$ contains a location parameter λ_1 (comparable to a single prediction $f(\tilde{\mathbf{x}})$), a scale parameter λ_2 (representing predictive uncertainty), and possibly some auxiliary parameters. In this case, the product inside the square root of equation (3.3) contains two easily interpretable terms. First, it contains a term proportional to the square of $\frac{\partial \lambda_1}{\partial x_d}$ that is multiplied by a factor depending on the scale parameter λ_2. After taking the square root, this term is comparable to $\left| \frac{\partial f(\tilde{\mathbf{x}})}{\partial \tilde{x}_d} \right|$ that is scaled by the predictive uncertainty. Second, it contains a term proportional to the square of $\frac{\partial \lambda_2}{\partial x_d}$, which influences the sensitivity independently of $\frac{\partial \lambda_1}{\partial x_d}$, even if $\frac{\partial \lambda_1}{\partial x_d} = 0$.

Figure 3.1 illustrates the contributions of the two components for a model with location-scale predictive distribution. The top plot represents 20 observation pairs $\mathscr{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$, and the predictive distribution of a Gaussian process model with an exponentiated quadratic covariance function (equation (3.9)). The dashed line represents the mean prediction, and the shaded areas 1-3 standard deviations of the Gaussian predictive distribution. The middle part of Figure 3.1 shows the standard sensitivity, $\left| \frac{\partial f(\tilde{\mathbf{x}})}{\partial \tilde{x}_d} \right|$, and the R-sens uncertainty-aware sensitivity. In the domain with many observations, the two sensitivity measures are almost equal, but they begin to deviate when the predictive uncertainty increases. The bottom plot shows separately the two components of uncertainty-aware sensitivity (defined as R-sens with the other component set to zero). Component 1 is proportional to the standard sensitivity, but scaled by the inverse standard deviation (see equation (3.14)). Component 2, on the other hand, is proportional to the derivative of the predictive variance with respect to x.

Publication III illustrates the components of R-sens in more detail, using a linear model as a simple example. For the linear model, the uncertainty-aware sensitivities of the variables are interpretable, and approach values proportional to the absolute regression coefficients as the number of observations increases.

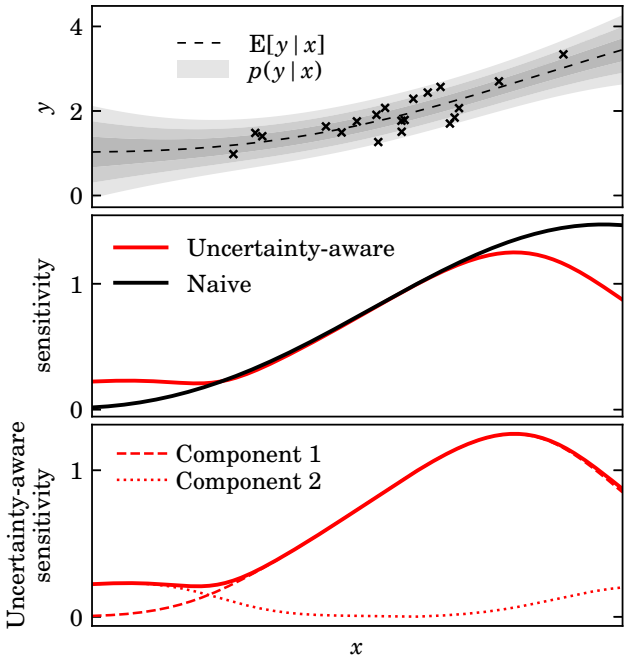

Figure 3.1. Top: Example of input and output data, and the predictive distribution $p(y \mid x)$ of a Gaussian process model. Middle: standard sensitivity $\left|\frac{\partial \mathrm{E}[y|x]}{\partial x}\right|$ of the model's mean prediction to perturbations in x (black), and uncertainty-aware sensitivity (red), which is adjusted based on uncertainty about y. Bottom: The uncertainty-aware sensitivity consists of multiple components – one for each parameter in the predictive distribution. Here, component 1 depends on $\frac{\partial \mathrm{E}[y|x]}{\partial x}$ but component 2 does not.

The uncertainty-aware sensitivity framework can be used equally well with different types of distributions, making the method applicable to both continuous and discrete target variables. The formulation of uncertainty-aware sensitivity in Publication III has two obvious restrictions. First, it assumes an analytical form for the predictive distribution. The exact distribution does not have to be available in closed form, but it can be an approximation or density estimate for example. However, because the methods rely on uncertainty, the results can be unexpected if the predictive distribution is a poor description of the true uncertainty. Second, the formulation requires the derivatives of the predictive distribution's parameters with respect to predictors. This restriction can be somewhat alleviated with automatic differentiation methods (see e.g. Baydin et al., 2018), which is touched on in Publication III.

Publication IV (published before Publication III) presents a finite difference based approximation, called KL, of the R-sens sensitivity measure. Instead of differentiating predictive distributions analytically, the KL method computes the divergence between fixed and perturbed predictive distributions. In variable importance assessment, a perturbed predictive

distribution is evaluated in a point with slightly different value for a single predictor variable. Publication III shows that the KL measure can be obtained as a Taylor approximation of R-sens. In practise, both KL and R-sens give practically equivalent results for single variable sensitivities. The idea of the KL method is very simple both computationally and conceptually as it only requires being able to evaluate a divergence based on the predictive distribution. It is easy to implement for arbitrary supervised learning models, and would also be straightforward to evaluate with arbitrary statistical divergences or distances.

Publication III develops the idea of the KL method further into the uncertainty-aware sensitivity framework, which has several benefits. First, it avoids possible errors resulting from too small or too large perturbation. Second, framing the problem as the derivative of a statistical divergence allows better interpretability and enables more theoretical understanding of the procedure. Publication III touches on interpretability, but more research is needed to understand the theoretical properties of uncertainty-aware sensitivity more deeply. Third, using analytical derivatives allows evaluating variable interaction effects by using higher order derivatives.

3.2.1 Variable interactions

In many supervised learning tasks, it is important to take into account the interaction effects of the predictor variables. A variable interaction simply means that the joint effect of two variables is different than the sum of their individual effects. The idea of derivative based importance is directly extendable to second-order variable interactions by introducing second derivatives (Wei et al., 2015; Cui et al., 2020). Analogously to $\left|\frac{\partial f(\tilde{\mathbf{x}})}{\partial \tilde{x}_d}\right|$, the absolute values of the derivatives with respect to both x_d and x_e, $\left|\frac{\partial^2 f(\tilde{\mathbf{x}})}{\partial \tilde{x}_d \partial \tilde{x}_e}\right|$, quantify the sensitivity of $f(\tilde{\mathbf{x}})$ to the second-order interaction effect of x_d and x_e.

The uncertainty-aware sensitivity idea can be extended to second-order interactions by differentiating a statistical divergence with respect to two variables. However, Publication III shows that at least when using Rényi divergences and Gaussian process models, the full fourth derivative contains cross-derivative terms which may not be useful for discovering variable interactions. Dropping some of these terms leads to a useful measure with formulation analogous to R-sens:

$$\text{R-sens}_2(\tilde{\mathbf{x}}, \{x_d, x_e\}, \alpha) \equiv \sqrt{\alpha \left(\frac{\partial^2 \boldsymbol{\lambda}(\tilde{\mathbf{x}})}{\partial \tilde{x}_d \partial \tilde{x}_e}\right)^{\mathsf{T}} \mathrm{I}(\boldsymbol{\lambda}(\tilde{\mathbf{x}})) \left(\frac{\partial^2 \boldsymbol{\lambda}(\tilde{\mathbf{x}})}{\partial \tilde{x}_d \partial \tilde{x}_e}\right)}. \tag{3.4}$$

3.3 Gaussian processes

Gaussian processes are a class of stochastic processes which are suitable for defining flexible prior distributions for functions in a Bayesian approach to supervised learning (Rasmussen & Williams, 2006). This section reviews the basics of Gaussian process modelling from a Bayesian viewpoint and how the importance of variables can be assessed for Gaussian process models, which is studied in Publications III and IV.

A Gaussian process (GP) defines a distribution over a function space, where a finite subset of evaluation points form a joint Gaussian distribution. A GP is fully defined by its mean and covariance functions:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{3.5}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[f(\mathbf{x}), f(\mathbf{x}')]. \tag{3.6}$$

These define the prior assumptions about the function to be inferred. Very often the mean function is set to zero, and only the covariance function is used to determine the a priori assumptions. The GP framework can be utilized to set a prior directly on functions mapping from the input variable(s) to the output variable(s). Inference is thus carried out without explicitly parameterizing the model. As the GP prior is infinite dimensional, Gaussian processes can be considered nonparametric models.

In supervised learning, observations are typically assumed to be noisy realisations of an underlying process. This is the standard approach also in Gaussian process inference, where the prior is combined with a likelihood to get a posterior for the values of the latent process. If the observation model is Gaussian, the posterior for a finite collection of latent values remains Gaussian. Moreover, at any new point, the latent values have a Gaussian predictive distribution with mean and variance given by

$$\mathbb{E}[f(\tilde{\mathbf{x}})] = \tilde{\mathbf{k}}^{\mathsf{T}} (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \tag{3.7}$$

$$\mathrm{Var}[f(\tilde{\mathbf{x}})] = k(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \tilde{\mathbf{k}}^{\mathsf{T}} (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \tilde{\mathbf{k}}. \tag{3.8}$$

Here, $\mathbf{K}_{i,j} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, $\tilde{\mathbf{k}} = \{k(\tilde{\mathbf{x}}, \mathbf{x}^{(1)}), \ldots, k(\tilde{\mathbf{x}}, \mathbf{x}^{(N)})\}$, $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ represent training data, and σ^2 is the noise variance.

Gaussian process models are very flexible due to the myriad possibilities for covariance functions. One common choice is the exponentiated quadratic (EQ) covariance function

$$k_{\mathrm{EQ}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{l_d^2} \right). \tag{3.9}$$

Here σ_f determines the overall magnitude of variation for the latent function, whereas the length-scale parameters l_d determine how rapidly the covariance decays in different input directions. Varying functions can

be constructed for example by adding and multiplying different covariance functions together.

The key property of the EQ covariance function is that it is infinitely differentiable, producing smooth nonlinear functions. For the uncertainty-aware sensitivity methods described in Section 3.3.1, infinite differentiability is not necessary, but the covariance function must be differentiable at least once. The EQ covariance function is used throughout Publications III and IV.

The parameters of the Gaussian process kernel and the used likelihood are typically called hyperparameters. In Publications III and IV, the hyperparameters are estimated from data via optimization by finding the maximum of the posterior distribution of hyperparameters. In order to make the optimization more stable, the publications use inverse-gamma prior distributions on the length-scale parameters, as described in Publication IV.

3.3.1 Uncertainty-aware sensitivity for Gaussian processes

One of the main attractions of Gaussian process models is their ability to represent uncertainty. For example, Wilson et al. (2020) describe them as the gold standard for many modelling scenarios when the trustworthy representation of uncertainty is important. Because of this property, Gaussian process models are a suitable use case for the uncertainty-aware sensitivity analysis. Both Publications III and IV use these methods for assessing variable importance of Gaussian process models.

If we take the Gaussian process with Gaussian observation model, presented in Section 3.3, as an example, the noisy target variable at a new test point has an equivalent predictive distribution as the latent values, but with added noise variance σ^2

$$\begin{aligned} p(\tilde{y} \mid \tilde{\mathbf{x}}, \mathbf{y}) &= \text{Normal}(\tilde{y} \mid \text{E}[\tilde{y}], \text{Var}[\tilde{y}]) \\ &= \text{Normal}(\tilde{y} \mid \text{E}[f(\tilde{\mathbf{x}})], \text{Var}[f(\tilde{\mathbf{x}})] + \sigma^2). \end{aligned} \tag{3.10}$$

The derivatives of these parameters with respect to the predictors are needed to compute the R-sens uncertainty-aware sensitivity measure. Because of the equivalence of equation (3.10), the derivatives of $\text{E}[\tilde{y}]$ and $\text{Var}[\tilde{y}]$ can be represented as

$$\frac{\partial \text{E}[\tilde{y}]}{\partial \tilde{x}_d} = \frac{\partial \text{E}[f(\tilde{\mathbf{x}})]}{\partial \tilde{x}_d} \tag{3.11}$$

$$\frac{\partial \text{Var}[\tilde{y}]}{\partial \tilde{x}_d} = \frac{\partial \text{Var}[f(\tilde{\mathbf{x}})]}{\partial \tilde{x}_d}. \tag{3.12}$$

These are obtained by differentiating equations (3.7) and (3.8), respectively.

The Fisher information matrix of the distribution $\text{Normal}(\tilde{y} \mid \text{E}[\tilde{y}], \text{Var}[\tilde{y}])$

is

$$\mathrm{I_N}(\mathrm{E}[\tilde{y}], \mathrm{Var}[\tilde{y}]) = \begin{pmatrix} \frac{1}{\mathrm{Var}[\tilde{y}]} & 0 \\ 0 & \frac{1}{2(\mathrm{Var}[\tilde{y}])^2} \end{pmatrix}. \tag{3.13}$$

Putting these components together, the R-sens measure (for $\alpha = 1$) becomes

$$\text{R-sens}(\tilde{\mathbf{x}}, x_d) = \sqrt{\frac{1}{\mathrm{Var}[\tilde{y}]} \left(\frac{\partial \mathrm{E}[f(\tilde{\mathbf{x}})]}{\partial \tilde{x}_d} \right)^2 + \frac{1}{2(\mathrm{Var}[\tilde{y}])^2} \left(\frac{\partial \mathrm{Var}[f(\tilde{\mathbf{x}})]}{\partial \tilde{x}_d} \right)^2}. \tag{3.14}$$

Publication III shows that similar closed form equations can be derived also for Bernoulli and Poisson observations models. The uncertainty-aware sensitivity is thus applicable to discrete distributions as well.

For Gaussian process models, the cost of computing the variance of the predictive distribution is of order $\mathcal{O}(N^2)$, whereas computing the mean is only $\mathcal{O}(N)$. These are reasonably cheap compared to the $\mathcal{O}(N^3)$ cost of inference, but if predictions are needed for many observations and predictor variables, the cost can exceed the inference cost. Publication III briefly discusses the costs if approximation methods are used for Gaussian process inference. For many of these methods, the prediction cost is scaled down equally to the cost of inference, which means that the uncertainty-aware sensitivity methods can scale to larger data sets to some extent.

Publication III discusses the use of automatic differentiation for implementing R-sens. While the implementation for individual variables is relatively straightforward, for variable interactions it is not. Publication III demonstrates that computing the fourth derivative directly with automatic differentiation is not very useful for identifying interactions, unless some cross-derivative terms are dropped. Despite this, automatic differentiation can be a useful tool for extending the uncertainty-aware methods to various models.

3.3.2 Evaluating variable importance

The separate length-scale parameters of the EQ covariance function in equation (3.9) describe the length-scale of variations for each variable. The parameter values can also be used as indicators of variable importance, which is called automatic relevance determination (ARD) (MacKay, 1994; Neal, 1996). ARD simply means using the inverses of the length-scale parameters as estimates of variable importance. Due to its simplicity, this is a relatively commonly used procedure for Gaussian processes and other models with analogous parameters (Williams & Rasmussen, 1996; Hensman et al., 2013).

Despite its popularity, using ARD to determine variable importance with Gaussian process models has two drawbacks. The length-scale parameters by themselves are poorly identifiable, meaning that the parameter values may vary significantly depending on the other parameters of the model (H.

Zhang, 2004). Moreover, ARD is biased towards variables with a nonlinear effect over equally important linear or near-linear variables (Piironen & Vehtari, 2016).

3.4 Research question 3 results

Publication III compares uncertainty-aware sensitivity to standard derivative-based sensitivity in both simulated and real scenarios. The paper concludes that in many practical situations, the difference of the two is small in terms of their ability to find important variables. However, when predictions are very uncertain, uncertainty-aware sensitivity can outperform standard derivative based sensitivity. This difference may be even greater if the predictions are evaluated outside the training observations, for example if some predictor values are permuted. Permutations are not studied in this thesis, but it would be an interesting topic for future research.

Publication IV compares the uncertainty-aware sensitivity to ARD in several different scenarios. One of the conclusions of Publication IV is that also R-sens is biased towards predictors with a nonlinear effect, but this bias is smaller than for ARD. R-sens also typically selects predictors with better predictive performance compared to ARD. While the benefit of ARD is that is requires no extra evaluation after fitting the Gaussian process model, the improved identification of important variables with R-sens may make it worth the extra cost.

Publication III compares the derivative based methods to multiple other variable importance methods. Overall, derivative based methods perform well for discovering both individual variables and second-order interactions. For interaction discovery, derivative based methods treat an interaction effect equally important regardless of whether the variables contain individual main effects or not, which is not the case for all variable importance methods. In many cases, this property is beneficial as it allows better separation of main effects and interaction effects.

4. Discussion

4.1 Scientific and practical impact of the work

One of the research gaps presented in Section 1.1 was the lack of efficient methods for improving the accuracy of leave-one-out cross-validation of difficult observations. If some cases are diagnosed as unreliable, sampling directly from the leave-one-out posteriors is a generally used alternative to importance sampling (Vehtari et al., 2017). In addition to being computationally costly, this thesis shows that it can suffer from the same inaccuracy as importance sampling. This thesis not only sheds light on the issues of this generally used alternative, but also is the first work to utilize adaptive importance sampling to provide a better alternative solution.

The thesis also demonstrates that using the model posterior as the initial proposal distribution is sometimes much better than using standard parametric proposal distributions (Bugallo et al., 2017). In a broader sense, this thesis has shown that adaptive importance sampling can be both accurate and computationally efficient when using a Bayesian model posterior as the initial proposal distribution. This may be found useful in other applications as well, such as Bootstrap.

The results of this thesis align with the long history of research about the importance of prior and likelihood sensitivity analysis (Canavos, 1975; Skene et al., 1986; Hill & Spall, 1994; Lopes & Tobias, 2011; Depaoli et al., 2020). What this thesis shows is that quick and simple procedures may be just as efficient as more complex or model-specific sensitivity analysis methods. The presented power-scaling method is novel, and may inspire more research in generic prior and likelihood sensitivity methods.

In variable importance assessment, this thesis introduces a concept of uncertainty-aware sensitivity. Derivatives of predictions have been used extensively as measures of variable importance (Ruck et al., 1990; Refenes & Zapranis, 1999; Leray & Gallinari, 1999; Simonyan et al., 2014; Sundararajan et al., 2017; Cui et al., 2020). To the best of our knowledge, this

work is the first to evaluate variable importance by differentiating statistical divergences of predictive distributions. This thesis has demonstrated that such an approach is possible analytically, laying the foundation for possible further research, both theoretical and empirical.

Overall, this thesis focuses on making Bayesian statistical modelling workflow easier and more effective. A significant factor in the relevance of this work lies in its applicability to practitioners of Bayesian statistics, especially those using the Stan statistical modelling platform (Stan Development Team, 2022). Bayesian statistics is used widely across different fields of science, ranging from ecology and social sciences to medicine and genetics. Model assessment is an important part of the Bayesian statistical workflow, and implementation of the methods of this thesis in openly available software packages has made them easy to use for all practitioners.

4.2 Limitations and recommendations for future research

This thesis studies importance sampling solutions for Bayesian cross-validation as well as prior and likelihood sensitivity analysis. While the methods were demonstrated with multiple different types of models, it is possible that there are some settings where the methods are not useful. For example, more detailed research might be needed to get the most out of these methods with hierarchical models. In the domain of cross-validation, an interesting area to study would be the use of adaptive importance sampling for other cross-validation settings. For example, leave-one-cluster-out cross-validation (Merkle et al., 2019) may be even more difficult for importance sampling, and more elaborate solutions may be needed. With power-scaling, understanding how the model's parameterization affects the power-scaling results is an interesting direction of future research. In addition, the practical interpretation of the results of power-scaling might need further study.

Publication I identified the need for special treatment of self-normalized importance sampling in leave-one-out cross-validation. While a viable solution was introduced in the form of the split proposal distribution, more research in its properties would be valuable, in addition to comparison to other related approaches (Lamberti et al., 2018; Rainforth et al., 2020).

This thesis demonstrates that in many practical situations, the uncertainty-aware sensitivity does not significantly differ from standard sensitivity analysis that uses singular prediction values. In simulated scenarios, it is seen that when the model has significant predictive uncertainty, taking it into account can lead to more accurate identification of variable importance. However, more research would be needed to identify practical situations where the uncertainty-aware sensitivity would be beneficial. While this

method has potential to be a useful variable importance method, at this point it can be considered a proof of concept. More research would be needed to also understand its theoretical properties more deeply, such as the effect of the choice of divergence, and its applicability to different probabilistic models. It is also possible that the uncertainty-aware sensitivity concept is found even more beneficial in other fields of research, such as deep learning (e.g. Maddox et al., 2019) or computer vision (e.g. J. Zhang et al., 2021).

References

Agarwal, M., Vats, D., & Elvira, V. (2021). A principled stopping rule for importance sampling. *arXiv preprint*. https://arxiv.org/abs/2108.13289

Al Labadi, L., & Evans, M. (2017). Optimal Robustness Results for Relative Belief Inferences and the Relationship to Prior–Data Conflict. *Bayesian Analysis*, *12*(3), 705–728. https://doi.org/10.1214/16-BA1024

Ando, T., & Tsay, R. (2010). Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting*, *26*(4), 744–763. https://doi.org/10.1016/j.ijforecast.2009.08.001

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2018). Automatic differentiation in machine learning: A survey. *Journal of machine learning research*, *18*(153), 1–43. http://jmlr.org/papers/v18/17-468.html

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics*, 686–690.

Bernardo, J. M., & Smith, A. F. (1994). *Bayesian theory*. Wiley.

Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J., & Djuric, P. M. (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, *34*(4), 60–79. https://doi.org/10.1109/MSP.2017.2699226

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., Gabry, J., Kay, M., & Vehtari, A. (2022). Posterior: Tools for working with posterior distributions [R package version 1.2.0]. https://mc-stan.org/posterior/

Bürkner, P.-C., Gabry, J., & Vehtari, A. (2020). Approximate leave-future-out cross-validation for Bayesian time series models. *Journal of Statisti-*

cal Computation and Simulation, *90*(14), 2499–2523. https://doi.org/10.1080/00949655.2020.1783262

Bürkner, P.-C., Gabry, J., & Vehtari, A. (2021). Efficient leave-one-out cross-validation for Bayesian non-factorized normal and Student-t models. *Computational Statistics*, *36*(2), 1243–1261. https://doi.org/10.1007/s00180-020-01045-4

Canavos, G. C. (1975). Bayesian estimation: A sensitivity analysis. *Naval Research Logistics Quarterly*, *22*(3), 543–552. https://doi.org/10.1002/nav.3800220310

Cappé, O., Guillin, A., Marin, J.-M., & Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, *13*(4), 907–929. https://doi.org/10.1198/106186004X12803

Cornuet, J.-M., Marin, J.-M., Mira, A., & Robert, C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, *39*(4), 798–812. https://doi.org/10.1111/j.1467-9469.2011.00756.x

Cui, T., Marttinen, P., & Kaski, S. (2020). Learning Global Pairwise Interactions with Bayesian Neural Networks. *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*.

Depaoli, S., Winter, S. D., & Visser, M. (2020). The Importance of Prior Sensitivity Analysis in Bayesian Statistics: Demonstrations Using an Interactive Shiny App. *Frontiers in Psychology*, *11*, 608045. https://doi.org/10.3389/fpsyg.2020.608045

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Elvira, V., & Martino, L. (2021). Advances in importance sampling. *arXiv preprint*. https://arxiv.org/abs/2102.05407

Elvira, V., Martino, L., Luengo, D., Bugallo, M. F., Et al. (2019). Generalized multiple importance sampling. *Statistical Science*, *34*(1), 129–155. https://doi.org/10.1214/18-STS668

Epifani, I., MacEachern, S. N., & Peruggia, M. (2008). Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics*, *2*, 774–806. https://doi.org/10.1214/08-EJS259

Evans, M., & Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, *1*(4), 893–914. https://doi.org/10.1214/06-BA129

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical*

Society: Series A (Statistics in Society), *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153–160.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion), In *Bayesian statistics 4*, Oxford University Press.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, *19*(10), 555. https://doi.org/10.3390/e19100555

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman; Hall/CRC.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian workflow. *arXiv preprint*. https://arxiv.org/abs/2011.01808

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Good, I. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *14*(1), 107–114.

Haussler, D., & Opper, M. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, *25*(6), 2451–2492. https://doi.org/10.1214/aos/1030741081

Hensman, J., Fusi, N., & Lawrence, N. D. (2013). Gaussian processes for big data, In *Proceedings of the 29th conference on Uncertainty in artificial intelligence*. http://auai.org/~w-auai/uai2013/prints/papers/244.pdf

Hesterberg, T. (1988). *Advances in importance sampling* (Doctoral dissertation). Stanford University.

Hill, S., & Spall, J. (1994). Sensitivity of a Bayesian analysis to the prior distribution. *IEEE Transactions on Systems, Man, and Cybernetics*, *24*(2), 216–221. https://doi.org/10.1109/21.281421

Hunanyan, S., Rue, H., Plummer, M., & Roos, M. (2021). Quantification of empirical determinacy: the impact of likelihood weighting on posterior location and spread in Bayesian meta-analysis estimated with JAGS and INLA. *arXiv preprint*. https://arxiv.org/abs/2109.11870

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, *17*(2), 295–311. https://doi.org/10.1198/106186008X320456

Jacobi, L., Joshi, M., & Zhu, D. (2018). Automated sensitivity analysis for Bayesian inference via Markov chain Monte Carlo: Applications to Gibbs sampling. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2984054

Kahn, H., & Marshall, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America*, *1*(5), 263–278.

Kappen, H. J., & Ruiz, H. C. (2016). Adaptive importance sampling for control and inference. *Journal of Statistical Physics*, *162*(5), 1244–1266. https://doi.org/10.1007/s10955-016-1446-7

Koblents, E., & Míguez, J. (2015). A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models. *Statistics and Computing*, *25*(2), 407–425. https://doi.org/10.1007/s11222-013-9440-2

Krüger, F., Lerch, S., Thorarinsdottir, T., & Gneiting, T. (2021). Predictive Inference Based on Markov Chain Monte Carlo Output. *International Statistical Review*, *89*(2), 274–301. https://doi.org/10.1111/insr.12405

Kullback, S. (1959). Statistics and information theory. *J. Wiley and Sons, New York*.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Lamberti, R., Petetin, Y., Septier, F., & Desbouvries, F. (2018). A Double Proposal Normalized Importance Sampling Estimator, In *2018 IEEE Statistical Signal Processing Workshop (SSP)*. https://doi.org/10.1109/SSP.2018.8450849

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., Et al. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*(3), 141–153. https://doi.org/10.1007/s42113-019-00029-y

Leray, P., & Gallinari, P. (1999). Feature selection with neural networks. *Behaviormetrika*, *26*(1), 145–166. https://doi.org/10.2333/bhmk.26.145

Lopes, H. F., & Tobias, J. L. (2011). Confronting Prior Convictions: On Issues of Prior Sensitivity and Likelihood Robustness in Bayesian Analysis. *Annual Review of Economics*, *3*(1), 107–131. https://doi.org/10.1146/annurev-economics-111809-125134

MacKay, D. J. (1994). Bayesian nonlinear modeling for the prediction competition. *ASHRAE transactions*, *100*(2), 1053–1062.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for Bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html

Martino, L., Elvira, V., & Louzada, F. (2017). Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, *131*, 386–401. https://doi.org/10.1016/j.sigpro.2016.08.025

Martino, L., Elvira, V., Luengo, D., & Corander, J. (2015). An Adaptive Population Importance Sampler: Learning From Uncertainty. *IEEE Transactions on Signal Processing*, *63*(16), 4422–4437. https://doi.org/10.1109/TSP.2015.2440215

McCartan, C. (2021). *adjustr: Stan model adjustments and sensitivity analyses using importance sampling*. https://corymccartan.github.io/adjustr/

Meng, X.-L., & Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, *11*(3), 552–586. https://doi.org/10.1198/106186002457

Merkle, E., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, *84*, 802–829. https://doi.org/10.1007/s11336-019-09679-0

Miguez, J., Mariño, I. P., & Vázquez, M. A. (2018). Analysis of a nonlinear importance sampling scheme for Bayesian parameter estimation in state-space models. *Signal Processing*, *142*, 281–291. https://doi.org/10.1016/j.sigpro.2017.07.030

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Neal, R. M. (1996). Bayesian learning for neural networks. *Lecture Notes in Statistics*, *118*.

O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian analysis*, *4*(1), 85–117. https://doi.org/10.1214/09-BA403

Owen, A. (2013). *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/

Peruggia, M. (1997). On the Variability of Case-Deletion Importance Sampling Weights in the Bayesian Linear Model. *Journal of the American Statistical Association*, *92*(437), 199–207. https://doi.org/10.1080/01621459.1997.10473617

Piironen, J., & Vehtari, A. (2016). Projection predictive model selection for Gaussian processes, In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. https://doi.org/10.1109/MLSP.2016.7738829

Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

Rainforth, T., Golinski, A., Wood, F., & Zaidi, S. (2020). Target–Aware Bayesian Inference: How to Beat Optimal Conventional Estimators. *Journal of Machine Learning Research*, *21*(88), 1–54. http://jmlr.org/papers/v21/19-102.html

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT press Cambridge.

Refenes, A.-P., & Zapranis, A. (1999). Neural model identification, variable selection and model adequacy. *Journal of Forecasting*, *18*(5), 299–332.

Reimherr, M., Meng, X.-L., & Nicolae, D. L. (2021). Prior sample size extensions for assessing prior impact and prior–likelihood discordance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*(3), 413–437. https://doi.org/10.1111/rssb.12414

Roos, M., Hunanyan, S., Bakka, H., & Rue, H. (2021). Sensitivity and identification quantification by a relative latent model complexity perturbation in bayesian meta-analysis. *Biometrical Journal*. https://doi.org/10.1002/bimj.202000193

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, *9*(1), 130–134. https://doi.org/10.1214/aos/1176345338

Rubino, G., & Tuffin, B. (2009). *Rare event simulation using Monte Carlo methods*. John Wiley & Sons.

Ruck, D. W., Rogers, S. K., & Kabrisky, M. (1990). Feature selection using a multilayer perceptron. *Journal of Neural Network Computing*, *2*(2), 40–48.

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological methods*, *26*(1), 103. https://doi.org/10.1037/met000027

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps, In *Workshop at international conference on learning representations*. https://www.robots.ox.ac.uk/~vgg/publications/2014/Simonyan14a

Skene, A. M., Shaw, J. E. H., & Lee, T. D. (1986). Bayesian modelling and sensitivity analysis. *The Statistician*, *35*(2), 281. https://doi.org/10.2307/2987533

Stan Development Team. (2021). RStan: The R interface to Stan [R package version 2.21.3]. https://mc-stan.org/

Stan Development Team. (2022). *Stan Modelling Language Users Guide and Reference Manual* (Version 2.28). https://mc-stan.org

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks, In *Proceedings of the 34th international conference on machine learning*, PMLR. https://proceedings.mlr.press/v70/sundararajan17a.html

Turner, K. J., & Lambert, P. S. (2015). Workflows for quantitative data analysis in the social sciences. *International Journal on Software Tools for Technology Transfer*, *17*(3), 321–338. https://doi.org/10.1007/s10009-014-0315-4

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239. https://doi.org/10.1037/met0000100

van Erven, T., & Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, *60*(7), 3797–3820. https://doi.org/10.1109/TIT.2014.2320500

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). Loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models [R package version 2.4.1]. https://mc-stan.org/loo/

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., & Lampinen, J. (2002). Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities. *Neural Computation*, *14*(10), 2439–2468. https://doi.org/10.1162/08997660260293292

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016). Bayesian Leave-One-Out Cross-Validation Approximations for Gaussian Latent Variable Models. *Journal of Machine Learning Research*, *17*(103), 1–38. http://jmlr.org/papers/v17/14-540.html

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. https://doi.org/10.1214/12-SS102

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2021). Pareto Smoothed Importance Sampling. *arXiv preprint*. https://arxiv.org/abs/1507.02646

Voter, A. F. (1985). A monte carlo method for determining free-energy differences and transition state theory rate constants. *The Journal of chemical physics*, *82*(4), 1890–1899.

Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, *142*, 399–432. https://doi.org/10.1016/j.ress.2015.05.018

Williams, C., & Rasmussen, C. (1996). Gaussian Processes for Regression, In *Advances in neural information processing systems*, MIT Press. https://proceedings.neurips.cc/paper/1995/hash/7cce53cf90577442771720a370c3c723-Abstract.html

Wilson, J., Borovitskiy, V., Terenin, A., Mostowsky, P., & Deisenroth, M. (2020). Efficiently sampling functions from Gaussian process posteriors, In *Proceedings of the 37th international conference on machine learning*, PMLR. https://proceedings.mlr.press/v119/wilson20a.html

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, *13*(3), 917–1007. https://doi.org/10.1214/17-BA1091

Zhang, H. (2004). Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of the American Statistical Association*, *99*(465), 250–261. https://doi.org/10.1198/016214504000000241

Zhang, J., Fan, D.-P., Dai, Y., Anwar, S., Saleh, F., Aliakbarian, S., & Barnes, N. (2021). Uncertainty inspired RGB-D saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2021.3073564

Aalto-DT 96/2022

ISBN 978-952-64-0869-9 (printed)
ISBN 978-952-64-0870-5 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL THESES