Department of Computer Science

Partitive Techniques in Bayesian Data Analysis

Tuomas Sivula

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 18/2021

Partitive Techniques in Bayesian Data Analysis

Tuomas Sivula

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, Remote connection link https://aalto.zoom.us/j/63889388791, on 12 March 2021 at 12:00.

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

Supervising professor
Professor Aki Vehtari, Aalto University, Finland

Thesis advisor
Professor Aki Vehtari, Aalto University, Finland

Preliminary examiners
Professor Alp Kucukelbir, Columbia University, USA
Doctor Daniel Hernández-Lobato, Universidad Autónoma de Madrid, Spain

Opponent
Doctor Daniel Hernández-Lobato, Universidad Autónoma de Madrid, Spain

Aalto University publication series
DOCTORAL DISSERTATIONS 18/2021

ISBN 978-952-64-0268-0 (printed)
ISBN 978-952-64-0269-7 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0269-7

Unigrafia Oy
Helsinki 2021

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Abstract

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

Author
Tuomas Sivula

Name of the doctoral dissertation
Partitive Techniques in Bayesian Data Analysis

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 18/2021

Field of research Computational Science

Manuscript submitted 1 November 2020 **Date of the defence** 12 March 2021

Permission for public defence granted (date) 18 January 2021 **Language** English

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

Abstract

This dissertation analyses two popular methods used in Bayesian data analysis, that involve splitting the data set into disjoint sets. The analysed approximative methods include expectation propagation (EP) and leave-one-out cross-validation (LOO-CV), which are used in the context of distributed inference and model evaluation/comparison respectively. The main contribution of the dissertation is in analysing the applicability and behaviour of the methods under different situations.

The EP algorithm is a popular method for approximating a factorisable density. In the Bayesian context, for tractability, it has usually been applied pointwise. However, by including multiple observations in one approximated factor component, the method can be seen as a flexible framework for distributed inference. In addition, in hierarchical settings, it provides a convenient mean for dimension reduction by concentrating parameter inferences to separate units.

LOO-CV is a popular method used in model evaluation, comparison, and weighting for estimating the out-of-sample predictive performance of a model using the given observations. In some situations, obtaining the estimate is a computationally heavy operation. The dissertation addresses this issue in the context of Gaussian latent variable models (GLVM) by reviewing various more efficient methods for approximating the LOO-CV estimate. Based on the results, a suggestion of approaches with different levels of accuracy and computational complexity are proposed.

As the variability of the LOO-CV estimator can be high in some problems, it is important to take into account the related uncertainty when applying the LOO-CV method in practice. The current popular ways of estimating the uncertainty often leads to considerably underestimating the variability. The dissertation studies the behaviour of the uncertainty in a model comparison setting both theoretically and experimentally and identifies problematic cases, in which the estimated uncertainty is badly calibrated. The problematic cases include small data size, models making similar predictions, and model misspecification. In addition, the dissertation proposes an improved estimator for the variance of the LOO-CV estimator in the case of a Bayesian normal model. The proposed estimator serves as an example of the possibility of obtaining improved model-specific uncertainty estimates. This approach has not been discussed in the literature before.

Keywords Bayesian data analysis, model comparison, approximative distributed inference, Gaussian processes

ISBN (printed) 978-952-64-0268-0 **ISBN (pdf)** 978-952-64-0269-7

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

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

Pages 270 **urn** http://urn.fi/URN:ISBN:978-952-64-0269-7

Tekijä
Tuomas Sivula

Väitöskirjan nimi
Jaottelevat menetelmät bayesilaisessa data-analytiikassa

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 18/2021

Tutkimusala Laskennallinen tiede

Käsikirjoituksen pvm 01.11.2020 **Väitöspäivä** 12.03.2021

Väittelyluvan myöntämispäivä 18.01.2021 **Kieli** Englanti

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

Tiivistelmä

Tämä väitöskirja käsittelee kahta suosittua bayesilaisessa data-analytiikassa käytettyä menetelmää, joissa data jaetaan pistevieraisiin joukkoihin. Analysoitavat menetelmät ovat odotusarvon välittäminen (expectation propagation, EP), jota sovelletaan hajautettuun päättelyyn, ja jätä-yksi-pois ristiinvalidointi (leave-one-out cross-validation, LOO-CV), jota käytetään mallin evaluointiin ja vertailuun. Työn pääkontribuutio on menetelmien sovellettavuuden ja käyttäytymisen analysointi eri tilanteissa.

EP-algoritmi on suosittu menetelmä, jolla voidaan approksimoida osittuva tiheysfunktio. Bayesilaisessa kontekstissa menetelmää on algebrallisen mukautuvuuden vuoksi yleensä sovellettu pisteittäin. Osittamalla useita havaintoja yhteen approksimoitavaan tekijään, menetelmää voidaan soveltaa joustavasti myös hajautettuun laskentaan. Tämän lisäksi sen avulla voidaan pienentää laskennallista dimensionaalisuutta hierarkisissa ongelmissa keskittämällä eri parametrien päättelyt eri yksiköille.

LOO-CV on suosittu mallin evaluointi-, vertailu-, ja painotusmenetelmä, jolla voidaan estimoida havaintoaineiston ulkopuolista prediktiivistä suorituskykyä annetun havaintoaineiston perusteella. Joissain tilanteissa tämän estimaatin laskeminen on raskas operaatio. Väitöskirjassa tätä ongelmaa käsitellään Gaussinen latentti muuttuja -mallien (Gaussian latent variable models, GLVM) kontekstissa vertailemalla eri menetelmiä, joilla LOO-CV estimaattia voidaan approksimoida tehokkaammin. Tulosten perusteella esitetään suositellut lähestymistavat ongelman ratkaisemiseksi eri tarkkuuden ja laskennallisen vaativuuden tasoilla.

Joissain ongelmissa LOO-CV estimaattorin vaihtelevuus voi olla suuri. Tämän vuoksi on tärkeää arvioida ja huomioida siihen liittyvä epävarmuus sovellettaessa menetelmää käytännössä. Nykyiset suositut menetelmät tämän epävarmuuden estimoimiseksi usein aliarvioivat vaihtelevuuden huomattavasti. Tämä väitöskirja tutkii epävarmuuden käyttäytymistä mallinvertailutilanteessa sekä teoreettisesti että kokeellisesti ja identifioi ongelmallisia tilanteita, joissa estimoitu epävarmuus on huonosti kalibroitu. Havaitut ongelmat esiintyvät tilanteissa, missä havaintojoukko on pieni, mallien ennusteet ovat samankaltaisia, ja mallit kuvaavat ilmiötä huonosti. Lisäksi väitöskirja esittää paremman estimaattorin LOO-CV estimaattorin varianssille bayesilaisen normaalimallin tapauksessa. Esitetty estimaattori toimii esimerkkinä mahdollisuudesta rakentaa parempia mallikohtaisia estimaattoreita LOO-CV estimaattorin epävarmuudelle. Tätä näkökantaa ei ole esitetty kirjallisuudessa aiemmin.

Avainsanat bayesilainen data-analytiikka, mallinvertailu, hajautettu approksimatiivinen päättely, Gaussiset prosessit

ISBN (painettu) 978-952-64-0268-0 **ISBN (pdf)** 978-952-64-0269-7

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2021

Sivumäärä 270 **urn** http://urn.fi/URN:ISBN:978-952-64-0269-7

Preface

This dissertation reflects the studies and research that took place at Aalto University between the years 2014 and 2020 first as a master student and later as a doctoral candidate. I embarked on the journey in the Bayesian Methodology research group in the Department of Neuroscience and Biomedical Engineering, which was at the moment titled as Department of Biomedical Engineering and Computational Science. Later on, the group joined with the Statistical Machine Learning and Bioinformatics group to form a new group of Probabilistic Machine Learning in the Department of Computer Science. I thank Helsinki Institute for Information Technology, Academy of Finland, and Finnish Center for Artificial Intelligence for supporting this work. In addition, I thank Prof. Alp Kucukelbir for pre-examining this dissertation and Dr. Daniel Hernández-Lobato for pre-examining this dissertation and agreeing to serve as an opponent.

Most of all, I thank my supervisor and instructor Prof. Aki Vehtari for providing the possibility to work under his guidance. You have taught me a lot about the Bayesian data analysis field and the work in academia in general. In addition, you have presented me with numerous interesting problems, for which I have had the pleasure to explore answers. I thank you for all the encouragement and support for overcoming difficult situations and I appreciate all the work you have done to help me in the process.

I wish to thank all of you who have participated in joint work with me in various projects. In particular, I thank Prof. Måns Magnusson for close collaboration in the latest two publications included in this dissertation. Without your effort, this dissertation would not have been possible. You provided us with great support and your expertise and perception made a great impact on the outcome of the projects. In addition, I would like to thank Prof. Andrew Gelman, Dr. Pasi Jylänki, Dr. Dustin Tran, Dr. Swupnil Sahai, Dr. Paul Blomstedt, Prof. John P. Cunningham, Prof. David Schminovich, Prof. Christian P. Rober.

I am also grateful for all the talented coworkers I have had the privilege to work with during the years, including but not limited to Dr. Jarno Lintusaari, Dr. Juho Piironen, Dr. Olli-Pekka Koistinen, Eero Siivola,

Dr. Marko Järvenpää, Akash Kumar Dhaka, Kunal Ghosh, Topi Paananen, Markus Paasiniemi, and Ilkka Raiskinen. You have helped me in various situations covering many research, technical, and general matters. Above all, you have been a nice company to spend time with. In addition, I am grateful for all the staff at the Department of Computer Science, who has provided the research group with the environment needed for the research.

I am grateful to my parents Maritta and Eero for the supportive upbringing and for the encouragement to pursue a good education. In addition, I thank all my siblings and close relatives, with whom I have spent joyful time. I thank my friend and mentor Dr. Jouni Aalto, for all the supportive guidance and inspiring thoughts. Above all, I wish to express my gratitude to my closest ones, to my wife Emmi and my son Armas. I appreciate all that you are and do.

Espoo, 4th February 2021,

Tuomas Sivula

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 Aki Vehtari, Tommi Mononen, Ville Tolvanen, Tuomas Sivula, Ole Winther. Bayesian Leave-One-Out Cross-Validation Approximations for Gaussian Latent Variable Models. *Journal of Machine Learning Research*, 17(103), pp. 1–38, June 2016.

II Aki Vehtari, Andrew Gelman, Tuomas Sivula, Pasi Jylänki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P. Cunningham, David Schiminovich, Christian P. Robert. Expectation Propagation as a Way of Life: A Framework for Bayesian Inference on Partitioned Data. *Journal of Machine Learning Research*, 21(17), pp. 1–53, January 2020.

III Tuomas Sivula, Måns Magnusson, Aki Vehtari. Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison. Submitted to *a journal*, 88 pp., September 2020.

IV Tuomas Sivula, Måns Magnusson, Aki Vehtari. Unbiased estimator for the variance of the leave-one-out cross-validation estimator for a Bayesian normal model with fixed variance. Submitted to *a journal*, 22 pp., September 2020.

Author's Contribution

Publication I: "Bayesian Leave-One-Out Cross-Validation Approximations for Gaussian Latent Variable Models"

Aki Vehtari is the main author of the article. He carried out the theoretical analysis of the reviewed methods and did most of the writing. Tommi Mononen, Ville Tolvanen, Tuomas Sivula, and Ole Winther contributed to the main insights of the article and participated in the writing. Ole Winther derived the proof in Appendix A Under the guidance of Aki Vehtari, Tuomas Sivula implemented the experiment code and prepared the results.

Publication II: "Expectation Propagation as a Way of Life: A Framework for Bayesian Inference on Partitioned Data"

Andrew Gelman came up with the original idea of the article. Aki Vehtari further developed the idea and orchestrated the associated research. All the authors contributed to the theoretical analysis, writing, and participated in revising the article. Both Tuomas Sivula and Swupnil Sahai implemented around 50 % of the program code for the experiments and prepared the respective portion of the results in the article. During several rounds of revisions, Tuomas Sivula further analysed the properties of the proposed approach considerably. Considering the length of the paper (53 pages), Tuomas Sivula had a major role in the article. The three first authors—Aki Vehtari, Andrew Gelman, and Tuomas Sivula—had equal contribution.

Publication III: "Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison"

Aki Vehtari developed the original idea of the paper. All the authors contributed to the theoretical analysis and participated in revising the article. Tuomas Sivula did most of the writing, and guided by Måns Magnusson, derived the mathematical proofs and implemented the experiments.

Publication IV: "Unbiased estimator for the variance of the leave-one-out cross-validation estimator for a Bayesian normal model with fixed variance"

Tuomas Sivula is the main author of the article. He came up with the original idea and carried out the analysis. In addition he did most of the writing and implemented the experiments. Måns Magnusson and Aki Vehtari contributed to the original idea, to the analysis, and participated in revising the work.

Abbreviations

BB	Bayesian bootstrap
CCD	central composite design
CV	cross-validation
e-elpd	expecation of expected log pointwise predictive density over possible data sets
elpd	expected log pointwise predictive density
EP	expectation propagation
GLVM	Gaussian latent variable models
GP	Gaussian process
HMC	Hamiltonian Monte Carlo
LOO	leave-one-out
LOO-CV	leave-one-out cross-validation
lpd	log pointwise predictive density
MAP	maximum a posteriori
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
PIT	probability integral transform
SNEP	stochastic natural gradient expectation propagation
WAIC	widely applicable information criterion (also known as Watanabe-Akaike information criterion)

1. Introduction

Various approximative techniques in Bayesian data analysis involve data partitioning. Distributed inference and predictive performance estimation are examples of fields, for which such methods have been applied (see for example Gelman et al., 2013). These methods address important steps in the workflow of computational Bayesian data analysis and are useful tools for adapting to various intractable or unfeasible problems. Different properties affect the applicability and interpretability of these methods. In particular, accuracy and efficiency plays a great role and should be taken into consideration when applying these methods.

The approach of distributing the inference can be viewed from multiple perspectives. In some situations, these methods can be used to improve the scalability of various computational inference algorithms by distributing the task at hand to smaller more manageable sub-problems. An alternative motivation for such a distributed setting could arise from a natural partitioning of the data or from the need to keep the separate parts private. Inference on partitioned data can also be considered as a way to handle different parts of the data differently—with different inference methods, sub-models, or precisions—taking into account part-specific objectives and properties of the data.

The predictive performance of a model is of interest in different situations. Various measures of the predictive performance can be useful as such for assessing the quality of the model's predictive capabilities or they can be used for comparing, selecting, and weighting multiple models (see for example Vehtari and Ojanen, 2012). In order to even out the inherent ability of models with different complexities to fit into the given data set, these measures reflect the performance of the model in the context of another data set from the respective data generating mechanism in general. Because the data generating mechanism is usually unknown, in practice such out-of-sample predictive performance measures need to be approximated based on the given data. A general approach for estimating these measures involves using a hold-out set of observations for evaluating the performance of a model trained on the other observations. By partitioning

the data into disjoint sets, this estimation can be carried out multiple times with different combinations of sets for the hold-out set. Cross-validation (CV) is a popular approach in which each partition is used once for the hold-out data set and in leave-one-out cross-validation (LOO-CV) each data point is partitioned into its own set (see for example Arlot and Celisse, 2010).

This dissertation analyses some popular methods involving data partition in the context of distributed inference, model evaluation, and model comparison. In particular, the applicability and behaviour of the methods in different situations are studied. The analysed properties include efficiency, approximation error, uncertainty, and estimation of the uncertainty. Based on the analysis, different approaches are compared and various recommendations and considerations are presented for the usage of the methods. In addition, aspects requiring future research are discussed.

While applied in different contexts, the analysed techniques are connected through the underlying approach of analysing one part of the data in the context of the left-out-data. This connection offers possibilities for applying concepts of one of the analysed method in the context of distributed inference for solving the problem in the context of predictive performance estimation; applying one particular method, expectation propagation (EP), for the former problem yields an approximation for the latter problem with only minimal additional computational cost.

1.1 Structure of the dissertation

This dissertation comprises this introductory part and four publications. The original publications, referred to as Publication I–IV, are included at the end of the dissertation. Each publication addresses a subset of the aforementioned concepts in the data partitive setting. All of the publications are methodological and analytical (Deming, 1942). Publications I, II and III are partly of review nature. The introductory part presents the problem settings discussed in the publications at a general level from a unified perspective. It also summarises the contributions from the publications while connecting them in the conceptual level.

The introductory part is organised as follows. First, reflecting the study in Publication II, Chapter 2 reviews the expectation propagation (EP) method in the context of distributed inference. Supporting the theme in publications I, III, and IV, Chapter 3 introduces the problem of estimating the predictive performance of a model with a focus on the LOO-CV estimate. Chapter 4 further studies the LOO-CV estimate by discussing the problem of estimating the uncertainty related to the LOO-CV predictive performance estimator in different problem situations while summarising the main findings and propositions from the associated publications III

and IV. As reviewed in Publication I, Chapter 5 combines concepts from all the other chapters to discuss various approaches for approximating the LOO-CV estimate in the context of Gaussian latent variable models (GLVM). Finally, Chapter 6 concludes the introductory part by discussing the main contributions of the dissertation.

2. Expectation propagation

Expectation propagation (EP) is a popular iterative algorithm for approximating a factorisable density with a density from a parametric family distribution. Opper and Winther (2000) first presented the initial idea of the method and shortly after Minka (2001b,a) presented it in a generalised form. It is often used in Bayesian inference for approximating an intractable posterior distribution.

Together in a more general group of algorithms called message-passing algorithms, EP provides convenient means for distributed inference on graph-structured models (see for example Pearl, 1986; Minka, 2001b; Chen and Wand, 2020). Publication II discusses and reviews this aspect and reflects it to a generalised setting of applying probabilistic programming to conveniently carry out the local inference on each partition. In addition, Publication II illustrates how applying EP for distributed inference in the context of hierarchical models can often be used to reduce the local dimensionality of the problem, which can offer drastic benefits in some high dimensional problems.

This chapter briefly presents the method and discusses its application in Bayesian inference. Following the discussion in Publication II, the application of the method for distributed inference is discussed, in particular in the context of hierarchical models. First, the algorithm is introduced and formulated in Section 2.1. Then Section 2.2 presents some general considerations related to the algorithm and Section 2.3 considers the possibilities EP offers in the context of distributed Bayesian inference. Finally, Section 2.4 discuss EP as a part of the more general framework of message-passing algorithms and review various extensions and modifications for it.

2.1 Algorithm

In EP, a target density $f(\theta)$ is approximated by a density $g(\theta)$ from some specified exponential family, such as the multivariate normal. The target

density is assumed to have some factorisation

$$f(\theta) \propto \prod_{k=1}^{K} f_k(\theta). \tag{2.1}$$

Each factor $f_k(\theta)$ is assigned a respective approximation $g_k(\theta)$ from the selected family of distributions and the target density is approximated with their product:

$$g(\theta) \propto \prod_{k=1}^{K} g_k(\theta). \tag{2.2}$$

Exponential family distributions have the following property: a product or division of two distributions from an exponential family is an unnormalised distribution from the same family. Thus $g(\theta)$ is also a distribution in the same selected family (see for example Minka, 2001b). The approximation $g(\theta)$ is referred to as the *global approximation* and each factor $f_k(\theta)$ in the target density together with the respective factor $g_k(\theta)$ in the approximation are referred to as the *sites*.

The algorithm iteratively updates each site approximation $g_k(\theta)$ by fixing other site approximations and considering the current target site factor $f_k(\theta)$. Accordingly, following the common terminology in EP literature (see for example Gelman et al., 2013, Section 13.8, p. 339), let

$$g_{-k}(\theta) \propto \prod_{t \neq k} g_t(\theta) = \frac{g(\theta)}{g_k(\theta)} \tag{2.3}$$

be referred to as the *cavity distribution*, and let

$$g_{\setminus k}(\theta) \propto f_k(\theta) g_{-k}(\theta) \tag{2.4}$$

be referred to as the *tilted distribution*. The cavity distribution in Equation (2.3) belongs to the selected approximating exponential family but the tilted distribution in Equation (2.4) depends on the target site factor $f(\theta)$ and is thus not restricted to be in this family. The algorithm is described as following:

1. Initialise each site approximation $g_k(\theta)$.
2. Repeat for each site $k \in \{1, 2, \ldots, K\}$ in any order, sequentially or asynchronously in parallel, until the global approximation $g(\theta)$ and the site approximations converge:
   (a) Compute the cavity distribution $g_{-k}(\theta)$.
   (b) Update the site approximation $g_k(\theta)$ so that the moments of the new global approximation $g(\theta) \propto g_k(\theta) g_{-k}(\theta)$ match the moments of the tilted distribution $g_{\setminus k}(\theta)$.

2.2 General considerations

Due to the exponential family restriction, the global approximation $g(\theta)$ and the cavity distribution $g_{-k}(\theta)$ in the EP algorithm, which comprises of products and divisions of the site approximations $g_k(\theta)$, are easy to obtain by summation or subtraction of the respective natural parameters of the distributions. This makes it easy to carry out step 2.a in the algorithm and to form the new global approximation after a site approximation has been updated in step 2.b.

The moment-matching step 2.b in the algorithm corresponds to minimising the KL-divergence $\mathrm{KL}(g_{\setminus k}(\theta)||g(\theta))$ (Minka, 2001b). Depending on the form of the target site factor $f_k(\theta)$ and the selected approximating distribution family, this step involves a potentially complex operation of inferring the moments of the tilted distribution.

In Bayesian context, EP is often applied to approximate the posterior distribution; the target density $f(\theta)$ corresponds to the posterior distribution $p(\theta \mid y)$, and each factor $f_k(\theta)$ corresponds to a likelihood component $p(y_k \mid \theta)$ or to the prior $p(\theta)$. The likelihood components assigned to each factor $f_k(\theta)$, $p(y_k \mid \theta)$ needs to be independent given the model parameters θ. Usually, one factor f_k is assigned to one observation, but as discussed in Publication II, multiple observations may be combined into one factor, which may provide various computational or structural advantages as later discussed in Section 2.3.

In some problems, the inference on the tilted distribution can be carried out analytically, but often approximative methods need to be used. In the context of Gaussian latent variable models (GLVM) in particular, discussed in more detail in Chapter 5, analytic solution is often available (see for example Opper and Winther, 2000; Minka, 2001b; Rasmussen and Williams, 2006, Section 3.6) or relatively efficient and quick numerical integration can be adopted (see for example Zoeter and Heskes, 2005). In some problems, however, one must resort to more complex approximative methods. Possible approaches include, but are not limited to, various mode-based approximation (for example the Laplace propagation Smola et al., 2004, shown to work well in many settings Rue et al., 2009), variational approximations (see for example Winn and Bishop, 2005), nested EP (Riihimäki et al., 2013; Hernandez-Lobato and Hernandez-Lobato, 2016), or simulation based approximations (such as Hamiltonian Monte Carlo (HMC) using the Stan probabilistic programming language (Carpenter et al., 2017)).

As indicated in the algorithm description, updating the site approximations in step 2 in the algorithm can be carried out sequentially or in parallel. The sites can also be updated asynchronously by initialising an update for a site using the latest approximations for the other sites. However, in this case, it is sensible to ensure that at least some other site approximation is updated between two concurrent updates for one site.

While in the EP algorithm, each update of the site approximation minimises the KL-divergence from the tilted distribution to the global approximation, it is not guaranteed that the KL-divergence from the target density to the global approximation is minimised. Furthermore, while being a fixed-point algorithm, is not guaranteed that the EP algorithm converges. Damping the updates of the site approximations can be used to mitigate these problems (Minka and Lafferty, 2002; Minka, 2005). In a damped site update, the site approximation in step 2.b of the algorithm is set to

$$g_k^{\text{new, damped}}(\theta) \propto g_k^{\text{old}}(\theta)^{1-\delta} g_k^{\text{new}}(\theta)^{\delta}, \tag{2.5}$$

where a damping factor $\delta \in (0,1]$ controls the step size of the update. In particular, damping is often necessary when the site updates are carried out in parallel; without damping, parallel EP updates often result in a global approximation that is deviated with respect to the sequential EP (Minka and Lafferty, 2002; Jylänki et al., 2011). Smaller damping can be used to avoid the error in the approximation while at the same time the convergence time is increased. Publication II discusses and demonstrates that it could be possible to use greater damping factor δ in the beginning for speedy start and reduce it during later iterations in order to avoid convergence problems and errors in the approximation.

2.3 Data partitioning

As discussed in Section 2.2, in Bayesian context EP has been usually applied to approximate the posterior distribution by allocating one EP site approximation for each observation likelihood factor $p(y_i \mid \theta)$. However, as discussed in Publication II, using one site to approximate the likelihood of multiple observations might offer various beneficial possibilities, namely speed and memory efficiency via distributed inference and dimension reduction in hierarchical settings. The following sections 2.3.1 and 2.3.2 respectively discuss these subjects in more detail.

2.3.1 Distributed setting

The EP algorithm conveniently work as a tool for distributed inference:

(a) The site updates in the EP algorithm in step 2 in the algorithm can be run in parallel.

(b) Because each site considers only the respective target factor $f_k(\theta)$, the data can be distributed to separate units.

(c) The information necessary to be shared among the sites between the iterations consists of distribution parameters of the selected approx-

imating family, for which the size depends only on the dimensionality of the parameter space.

(d) For convenience and generalisability, the inferences on the sites can be carried out approximately by utilising probabilistic programming.

Publication II discusses and shows experimentally, that increasing the number of sites offers possibilities for speed and memory efficiency while potentially increasing the error in the obtained global approximation. A desired partitioning of the data set can be applied to obtain a suitable trade-off between error in the approximation and the speed in the inference.

In some situations, the data and/or inference is distributed into conditionally independent sets by nature, for which distributed inference can be directly applied using EP. For example, as discussed in Publication II, EP can be used to combine multiple inferences of the same parameter carried out in different institutions. Motivation/reason for the separate inferences could be for example the complexity of the separate inferences as such or data privacy.

Various other generalisable distributed divide-and-conquer methods has been proposed for Bayesian inference (Ahn et al., 2012; Hoffman et al., 2013; Wang and Dunson, 2013; Neiswanger et al., 2014; Balan et al., 2014; Scott et al., 2016). As discussed in Publication II, one of the typical problems in the distributed Bayesian inference methods is the inappropriate conveying of the prior information to the separate inference units; either the prior $p(\theta)$ is distributed to each unit as such, in which case the combined inference may contain too much prior information, or the prior is reduced to $p(\theta)^{1/K}$, in which case the site-specific prior is too weak for good inference (see for example Gelman et al., 1996, 2008; Barthelmé and Chopin, 2014). In EP the prior information is naturally included as such in each of the site inferences without the aforementioned side-effects. Publication II compares the EP algorithm to the consensus Monte Carlo (MC) algorithm in a simulated hierarchical logistic regression problem setting. The results indicate that EP achieves better approximation accuracy with a comparable speed in all tested partitioning rates.

In addition to the speed and memory efficiency, the advantages of EP in a distributed setting come in the natural sharing of the information among the sites between iterations. Figure 2.1 demonstrates this computational benefit. A site receives information from the other sites via the cavity distribution, which operates as a prior for the current site update. Similar to a prior in conventional Bayesian inference, this information can indicate the area of importance for which the computation can be focused on.

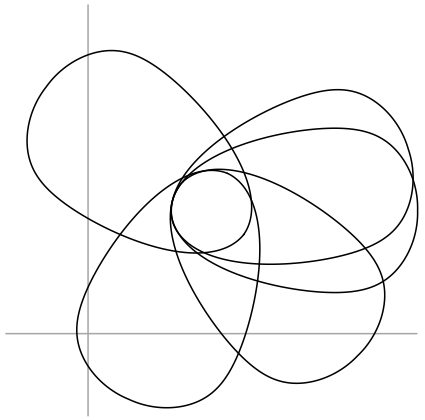

Figure 2.1. Sketch illustrating the benefits of EP in Bayesian computation. In this simple example, the parameter space θ has two dimensions, and the data have been split into five pieces. Each oval represents a contour of the likelihood $p(y_k|\theta)$ provided by a single partition of the data. Simple parallel computation of each piece separately would be inefficient because it would require the inference for each partition to cover its entire oval. By combining with the cavity distribution $g_{-k}(\theta)$, we can devote most of our computational effort to the area of overlap. Figure adopted from Publication II (Figure 2, p. 9).

2.3.2 Hierarchical setting

Using EP in hierarchical models can substantially reduce the parameter space on the site inferences and on the global approximation. By allocating all the likelihood components considering certain model parameters in one site, the corresponding parameter does not need to be shared to the other sites and to the global approximation. Only the parameters that affect multiple sites need to be included in the global approximation. If the inferences for the local parameters are needed after convergence, they can be obtained separately from the respective sites, or a joint sample can be generated as discussed in Publication II in Section 4.2.

The reduced dimensionality of the problem provides efficiency and allows modelling problems with big parameter spaces. Publication II features an example by Sahai (2018), in which a hierarchical problem in the field of astronomy comprising of 3258 parameters is approximated with EP using 30 sites each performing inference on 288 parameters: 18 parameters shared among the sites and 270 local parameters. In the concerned problem, by increasing the number of sites into the number of hierarchical groups, the local parameters could be further reduced to only 9 per site resulting in a total of 27 parameters per site.

2.4 Generalisations

As mentioned at the start of Chapter 2. EP belongs to a more general class of algorithms called message-passing algorithms (Minka, 2005). In this general form, the approximating distributions are not restricted to any family and, instead of minimising the KL-divergence from the tilted distribution to the global approximation by matching the moments, the step 2.b in the algorithm can more freely be defined to update the global approximation so that it approximates the tilted distribution, not necessarily by minimising any divergence measure.

The algorithmic definition of restricting the family of approximate distributions is useful in making the algorithm efficient due to the simple calculation of the tilted and cavity distributions and easy information sharing between the sites. The KL-divergence minimisation definition of step 2.b in the EP algorithm is also convenient and relatively easy to implement due to the simple moment-matching in the case of exponential approximating distribution family (Minka, 2001b). However, in particular for the latter algorithmic definition, various other implementations have been proposed. The following summarises some of the key approaches discussed in more detail in Publication II.

Minka (2004) propose an algorithm called power-EP, in which the KL-divergence minimisation in step 2.b in the EP algorithm is generalised, without losing the flexibility in the inference method, to a minimisation of the α-divergence with a tunable parameter α. In particular, when $\alpha = 1$, the divergence corresponds to $\mathrm{KL}(g_{\setminus k}(\theta)||g(\theta))$ used in EP, and when $\alpha = 0$, it corresponds to inverse KL-divergence $\mathrm{KL}(g(\theta)||g_{\setminus k}(\theta))$ also adopted in variational message-passing algorithms (Winn and Bishop, 2005). Greater values of α result in site updates that consider the whole parameter space of the tilted distribution in the approximation, whereas smaller values of α focus on good approximations near a mode of the tilted distribution (Minka, 2005). The mode focusing behaviour could be useful in situations when the tilted distribution is multimodal. Because of this flexibility, power-EP has shown to be more robust than EP in various situations, such as when the approximating family is ill-fitting (Minka, 2005) or when the prior is too vague (Seeger, 2008).

EP and the general message-passing algorithm framework does not ensure convergence. An alternative approach, which could also provide more reliable convergence properties, is to apply some energy optimisation method to the objective function with stationary points corresponding to the fixed point of some message-passing algorithm. Hasenclever et al. (2017) propose a method called stochastic natural gradient expectation propagation (SNEP), which is constructed by optimising the objective function of power-EP to get another message-passing algorithm with the corresponding optimum. Similar to the EP and power-EP, SNEP can

also be applied to use various inference methods in the sites, such as the conveniently generalisable MCMC. The SNEP method can be considered as a mean parameter space version of the power-EP, which updates the sites in the natural parameter space instead.

Because of the factorised nature of the message-passing algorithm, it is not necessary to define the algorithm consistently for each site. It is possible to use different methods of inference or approximation criteria for each site while using the same concept of cavity and tilted distribution to convey information between the sites. The method and criteria can also be changed during the iterations. For example, Publication II discuss the possibility of using EP with the moment matching in the early iterations for faster start and switching to use a compatible algorithm called stochastic natural gradient expectation propagation (SNEP, Hasenclever et al., 2017) for more reliable convergence.

3. Bayesian leave-one-out cross-validation

Evaluation of the predictive performance of a model is an important operation used in the Bayesian computational analysis workflow. The obtained accuracy or loss measure can be of interest by itself, for example when evaluating a forecast, or it may be used to compare two or more models in the process of selecting the applied model or exploring for possibilities improvements (see for example Vehtari and Lampinen, 2002; Vehtari and Ojanen, 2012; Gelman et al., 2013; Vehtari et al., 2017). In addition, it can be utilised in model averaging (see for example Geisser and Eddy, 1979; Gelfand, 1996; Madigan et al., 1996; Hoeting et al., 1999; Ando and Tsay, 2010; Yao et al., 2018).

In general, there are various approaches to setting up a model for a problem. One typical approach or involve starting off with a simple model and iteratively extend it while comparing the simpler one to the more complex one. In this nested model comparison setting, one ultimately decides on the level of complexity on the model while considering the advantages and disadvantages of the complexification (see for example Piironen et al., 2020). Another typical problem setting involves comparing non-nested models, where neither model is able to generalise into the other, as in the case of regression models using different sets of predictors (see for example Vehtari and Ojanen, 2012).

Ideally, the predictive performance of a model would be measured in the context of the mechanism that has generated the given data in general; instead of predicting the observed data set, the measure would reflect the performance of a model for a new unobserved data set. The motivation for this out-of-sample analysis is in putting models with different abilities to fit the given data, for example by the flexibility due to an increased number of parameters, to a common scale. As the true underlying data generating mechanism is not known in practice, in order to evaluate such an out-of-sample predictive performance measure, one needs to resort to various approximative methods. Leave-one-out cross-validation (LOO-CV) is one such a popular method (see for example Arlot and Celisse, 2010).

The LOO-CV estimation involves various problematic aspects, some of

which are addressed in publications I, III, and IV. First, the naive approach of obtaining this estimate requires fitting the model once for each observation in the data set, which can be overly time-consuming. Publication I discusses different fast approximations of the LOO-CV estimate in the specific context of Gaussian latent variable models. Second, estimating the uncertainty related to a LOO-CV estimate is a complex task and the currently popular approach often underestimates the uncertainty. Publication III study the behaviour of the LOO-CV estimate and the uncertainty in Bayesian model comparison and alleviates problematic situations. The problem is analysed both theoretically and experimentally in the case of linear regression models. Publication IV presents an alternative unbiased estimator for the variance of the sampling distribution for a simple Bayesian normal model.

This chapter introduces the LOO-CV problem setting and discusses the main challenges related to it while reflecting on the studies in the related publications I, III, and IV. The studied predictive performance measure and the associated LOO-CV estimator are introduced in sections 3.1 and 3.2 respectively. Section 3.3 discusses an alternative, related, and more easily obtainable approach for assessing the predictive performance, the widely applicable information criterion (WAIC), and contrasts it to LOO-CV. Section 3.4 further discusses how, being relatively time-consuming to obtain, LOO-CV can be estimated in practice.

3.1 Measuring predictive performance

Various measures of predictive accuracy exist, some of them being more application-specific, such as the classification accuracy, and others generally applicable, such as the log predictive density also known as the log-score. This work focuses on the latter measure. The log-score is a commonly used strictly proper and local scoring rule: the expected score is uniquely maximised by the true forecaster and the rule depends only on the forecaster at the realised predicted event (Gneiting and Raftery, 2007; Vehtari and Ojanen, 2012). The former property suggests honesty in the forecast in order to maximise the score and the latter property allows some bad forecasts to be penalised more than others.

Consider a data set of n observations $y = [y_1, y_2, \ldots, y_n]$, a stochastic variable with probability distribution $p_{\text{true}}(y)$ representing the true data generating mechanism, and its realisation y^{obs} corresponding to an observed data set. The within-sample measure for the predictive performance with log-score is the *log pointwise predictive density* (lpd):

$$\text{lpd}\left(\mathrm{M}_k \mid y^{\text{obs}}\right) = \sum_{i=1}^{n} \log p_k(y_i^{\text{obs}} \mid y^{\text{obs}}), \tag{3.1}$$

where $\log p_k(y_i \mid y^{\text{obs}})$ is the posterior predictive log density for the evaluated model M_k fitted for data set y^{obs}. For readability, the conditioning on possible covariates is omitted in the notation. Using the log-score, we define an out-of-sample predictive performance measure for evaluating the model M_k in the context of the observed data set and the underlying true data generating mechanism (Vehtari and Ojanen, 2012; Vehtari et al., 2017); given y^{obs} and $p_{\text{true}}(y)$, the *expected log pointwise predictive density* (elpd) is

$$\text{elpd}\left(\mathrm{M}_k \mid y^{\text{obs}}\right) = \sum_{i=1}^{n} \int p_{\text{true}}(y_i) \log p_k(y_i \mid y^{\text{obs}}) \mathrm{d}y_i \,. \tag{3.2}$$

The elpd measure defined in Equation (3.2) evaluates the predictive performance of a model for a new possible data set that could have been observed. It works in the context of a specific observed data set and respective true data generating mechanism. Thus this measure is of interest in the application-oriented setting, when evaluating a model or comparing models for a specific task with the one observed data set. In the algorithm oriented setting, one is interested in evaluating the predictive performance of a model when applied to any possible observed data set. In these cases one would instead use the expectation of elpd (e-elpd) over possible data sets y that could be observed as the predictive performance measure:

$$\text{e-elpd}\left(\mathrm{M}_k\right) = \mathrm{E}_y\left[{}^{\text{sv}}\text{elpd}\left(\mathrm{M}_k \mid y\right)\right]. \tag{3.3}$$

Here the notation ${}^{\text{sv}}\cdots$ is adopted from Publication III to remind that a term is a stochastic variable. This dissertation focuses on studying the elpd measure.

3.2 Leave-one-out cross-validation

The elpd and e-elpd predictive performance measures defined in equations (3.2) and (3.3) respectively involve the true data generating mechanism $p_{\text{true}}(y)$. As this distribution is usually not known in practice, elpd and e-elpd needs to be approximated (Bernardo and Smith, 1994; Vehtari and Ojanen, 2012). However, an estimator for elpd suffices also as an estimator for e-elpd as elpd itself is an estimator for e-elpd. Due to the double use of the data (constructing the posterior and predicting), the within-sample predictive accuracy measure lpd presented in Equation (3.1) is a biased estimator for elpd; typically a model tends to be more accurate in predicting the data used for training the model than some future data from the same data generating mechanism. If directly applied to select a model without any adjustment for the complexity, it leads to overfitting as more complex models tend to be more adaptive for the training data.

Cross-validation (CV) is a popular general approach for estimating elpd and other predictive performance measures. In CV, a subset of observations is left out of the data to use as an out-of-sample validation set while using the rest of the observations to train the model. The splitting of the data set is applied multiple times with different partitioning and the results are combined. K-fold CV involves evenly splitting the data set into K parts, where each part is used once as the validation set for the model trained using the rest. In leave-one-out CV (LOO-CV), $K = n$ so that each observation is in turn left out of the data.

For approximating $\mathrm{elpd}\left(\mathrm{M}_k \mid y^{\mathrm{obs}}\right)$, the LOO-CV estimate is formulated as

$$\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_k \mid y^{\mathrm{obs}}\right) = \sum_{i=1}^{n} \log p_k\left(y_i^{\mathrm{obs}} \mid y_{-i}^{\mathrm{obs}}\right), \tag{3.4}$$

where

$$\log p_k\left(y_i^{\mathrm{obs}} \middle| y_{-i}^{\mathrm{obs}}\right) = \log \int p_k\left(y_i^{\mathrm{obs}} \middle| \theta\right) p_k\left(\theta \middle| y_{-i}^{\mathrm{obs}}\right) \mathrm{d}\theta \tag{3.5}$$

is the leave-one-out (LOO) predictive log density for the ith observation y_i^{obs} with model M_k fitted for the data y_{-i}^{obs} consisting of all the other observations.

Similar to the within-sample measure lpd presented in Equation (3.1), the LOO-CV estimator ${}^{\mathrm{sv}}\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_k \mid y\right)$ presented in (3.4) uses the data twice. However, the cross-validation approach in LOO-CV addresses the problem of overfitting by never using an observation to predict itself. The LOO-CV estimator is consistent under mild assumptions, almost unbiased, and usually the bias decreases when the data size n grows (Arlot and Celisse, 2010, Section 5.1; Watanabe, 2010b).

When the objective of the predictive performance estimation is to compare models, the LOO-CV estimator can be directly applied to estimate the difference of the predictive performances of the models as the difference of the individual LOO-CV estimates; the difference of the predictive performances of models M_a and M_b fitted for the same data set y^{obs},

$$\mathrm{elpd}\left(\mathrm{M}_a, \mathrm{M}_b \mid y^{\mathrm{obs}}\right) = \mathrm{elpd}\left(\mathrm{M}_a \mid y^{\mathrm{obs}}\right) - \mathrm{elpd}\left(\mathrm{M}_b \mid y^{\mathrm{obs}}\right) \tag{3.6}$$

is estimated using LOO-CV as

$$\begin{aligned}
\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_a, \mathrm{M}_b \mid y^{\mathrm{obs}}\right) &= \widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_a \mid y^{\mathrm{obs}}\right) - \widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_b \mid y^{\mathrm{obs}}\right) \\
&= \sum_{i=1}^{n} \left(\log p_a\left(y_i^{\mathrm{obs}} \middle| y_{-i}^{\mathrm{obs}}\right) - \log p_b\left(y_i^{\mathrm{obs}} \middle| y_{-i}^{\mathrm{obs}}\right) \right) \\
&= \sum_{i=1}^{n} \left(\widehat{\mathrm{elpd}}_{\mathrm{LOO},i}\left(\mathrm{M}_a \mid y^{\mathrm{obs}}\right) - \widehat{\mathrm{elpd}}_{\mathrm{LOO},i}\left(\mathrm{M}_b \mid y^{\mathrm{obs}}\right) \right) \\
&= \sum_{i=1}^{n} \widehat{\mathrm{elpd}}_{\mathrm{LOO},i}\left(\mathrm{M}_a, \mathrm{M}_b \mid y^{\mathrm{obs}}\right).
\end{aligned} \tag{3.7}$$

Later on in the dissertation, together with functions accepting either a single model $f(\mathrm{M}_k \mid y)$ (model evaluation) or two models $f(\mathrm{M}_a, \mathrm{M}_b \mid y)$ (model comparison), notation $f(\cdot \mid y)$ is used to denote either one of these cases interchangeably but consistently so that always $\cdot = \mathrm{M}_k$ or $\cdot = \mathrm{M}_a, \mathrm{M}_b$ in one set of equations.

3.3 Widely applicable information criterion

An alternative approach for approximating elpd is to use the corresponding within-sample predictive accuracy measure lpd defined in Equation (3.1) either as such or with various adjustments. Widely applicable information criterion (also known as Watanabe-Akaike information criterion, WAIC) is one such a method, in which lpd is penalised with a term p_{WAIC}, often called the effective number of parameters, reflecting the complexity of the model being fit (Watanabe, 2010a,b):

$$\widehat{\mathrm{elpd}}_{\mathrm{WAIC}}\left(\mathrm{M}_k \mid y^{\mathrm{obs}}\right) = \mathrm{lpd}\left(\mathrm{M}_k \mid y^{\mathrm{obs}}\right) - p_{\mathrm{WAIC}}\,. \tag{3.8}$$

Two versions of the correction term p_{WAIC} has been proposed: one which is based on the difference between the training and the Gibbs utility,

$$p_{\mathrm{WAIC,G}} = 2\sum_{i=1}^{n}\Big(\log\big(\mathrm{E}_{\mathrm{post}}[p(y_i|\theta)]\big) - \mathrm{E}_{\mathrm{post}}[\log p(y_i|\theta)]\Big)\,, \tag{3.9}$$

and one which is based on the pointwise variance of the terms in the log predictive density,

$$p_{\mathrm{WAIC,V}} = \sum_{i=1}^{n}\mathrm{Var}_{\mathrm{post}}(p(y_i|\theta))\,, \tag{3.10}$$

where $\mathrm{E}_{\mathrm{post}}$ and $\mathrm{Var}_{\mathrm{post}}$ indicate the expectation and variance over the posterior distribution of θ respectively. Gelman et al. (2014) analyse popular information criteria, including WAIC, and compare their behaviour both theoretically and in practice. As discussed for example in Publication I and as experimentally demonstrated by Gelman et al. (2014), using $p_{\mathrm{WAIC,V}}$ as the correction term is preferable to $p_{\mathrm{WAIC,G}}$ due to closer connection to the LOO-CV via series expansion analysis.

Similar to the LOO-CV, WAIC is a consistent estimator for the elpd under mild assumptions and it is asymptotically equivalent to the Bayesian LOO-CV estimator (Watanabe, 2010b). However, as discussed in Publication I and by Gelman et al. (2014) and by Vehtari et al. (2017), LOO-CV has been found to be more robust than WAIC due to WAIC ignoring higher-order terms, which may be of significance in the finite data domain.

3.4 Fast approximations to LOO-CV

The naive way of evaluating the LOO-CV estimate involves fitting the model for n different data sets, which can be a costly process. In order to reduce the computational burden, various approximations for the LOO-CV estimate with smaller computational cost has been proposed. First, being asymptotically equivalent to LOO-CV, WAIC discussed in Section 3.3 can be considered as a fast approximation for LOO-CV. Other methods include, but are not limited to, various sample-based methods, methods based on EP or Laplace approximation, and numerical integration.

One popular approach is to use importance sampling for estimating the LOO-CV estimate, utilising the full posterior as the proposal distribution (Gelfand et al., 1992). Pareto smoothed importance sampling further stabilises the obtained weights while providing means for estimating the reliability of the estimate (Vehtari et al., 2019). Additionally, one can apply various adaptive importance sampling techniques, such as implicitly adaptive importance sampling (Paananen et al., 2020), or sub-sampling (Magnusson et al., 2019, 2020). The full posterior can be approximated using for example HMC using the Stan probabilistic programming language (Carpenter et al., 2017). Alternatively, Magnusson et al. (2019) propose to use Laplace or variational posterior approximations.

The EP method reviewed in Chapter 2 has a connection to the LOO-CV estimate via its cavity and tilted distribution. When applied to approximate the fully pointwise factorised likelihood, the cavity distribution presented in Equation (2.3), representing the contribution from all but one likelihood factor, can be considered as an approximation for the LOO posterior distribution:

$$p(\theta \mid y_{-i}) \approx g_{-i}(\theta). \tag{3.11}$$

Respectively, the LOO predictive density corresponds to the marginal likelihood of the tilted distribution:

$$p(y_i \mid y_{-i}) \approx \int p(y_i \mid \theta) g_{-i}(\theta) \mathrm{d}\theta. \tag{3.12}$$

In a converged EP algorithm, the cavity distribution has been determined using all the observations and thus $g_{-i}(\theta)$ is technically not independent of the observation y_i. Opper and Winther (2000) show that the EP method is LOO-CV consistent.

Although not used for approximating the LOO-CV estimator, an analogous approach involving Laplace approximation instead of EP was discussed by Cseke and Heskes (2011). Publication I utilises this for LOO-CV approximation while providing proof of LOO-CV consistency.

Another approach for approximating the LOO-CV estimate is to use numerical integration to approximate $p(y_i \mid y_{-i}) = \left(\int \frac{p(\theta|y)}{p(y_i|\theta)} \mathrm{d}\theta \right)^{-1}$ (Held et al.,

2010). Publication I adopts this method for Gaussian latent variable models (GLVM), in which the integral can be simplified to one-dimensional quadrature approximation, and proposes a method for further stabilising the errors in the tails.

Publication I discusses and compares various fast LOO-CV approximations in the context of GLVMs. Estimating the LOO-CV in this context is discussed in more detail in Chapter 5.

4. Uncertainty in LOO-CV

As discussed in Chapter 3, LOO-CV is a popular method for estimating the predictive performance (see for example Arlot and Celisse, 2010; Vehtari and Ojanen, 2012). The method is subject to the uncertainty arising from the estimation of the data generating mechanism $p_{\text{true}}(y_i)$ in Equation (3.2) using the finite observed data set y^{obs}. In order to rigorously interpret the obtained LOO-CV estimate, the associated uncertainty should be taken into account. Reflecting the analysis presented in Publication III, this chapter discusses this uncertainty and its estimation. In particular, various problematic cases, in which the LOO-CV estimator behaves problematically and/or the uncertainty is hard to estimate, are identified and reflected to the usage of the LOO-CV method.

It is known that, while the LOO-CV estimator is (under reasonable assumptions) asymptotically unbiased, its variance can be big (see for example Breiman, 1996; Arlot and Celisse, 2010, Section 5.2.1). The variance is affected by the problem setting and, as discussed by Arlot and Celisse, in particular by the stability of the learning algorithm. In this consideration, LOO-CV is well suited for Bayesian methods as they tend to be stable due to integrating over the uncertainty in the posterior distribution. However, model misspecification may cause instability also for Bayesian methods. In addition, as demonstrated for example by Vehtari et al. (2017), due to the smoothness of the log score, the variance in LOO-CV is usually lower than in K-fold CV in a Bayesian context.

Considering the possible high variability, in order to draw rigorous conclusions about a LOO-CV estimate, the uncertainty of the obtained estimate should be taken into account. In the case of evaluating the performance of a single model, the uncertainty can be indicated for example by error bars on the estimated predictive performance, and in the case of model comparison, the uncertainty might be used to estimate the probability of one model having better predictive performance than the other. As demonstrated in Publication III both theoretically and experimentally, by analysing the frequency properties of the LOO-CV estimator over possible data sets that could have been observed, the uncertainty of the LOO-CV estimate can

be significant especially in small sample sizes, when comparing models which produce similar predictions, and when the model(s) are misspecified with regards to the data generating process. Publication III also discusses that, in addition to the high LOO-CV variance, in these situations the estimation of the variance is also problematic.

Despite being asymptotically unbiased, when applied in a model selection problem setting by selecting the model with best LOO-CV estimate, the selection process induces bias into the resulting LOO-CV estimate of the predictive performance of the selected model. Due to this selection-induced bias, considering the uncertainty is particularly important in the typical nested model selection problem setting, where neglecting the uncertainty leads to overfitting (Piironen and Vehtari, 2017).

This chapter discusses the uncertainty of the LOO-CV estimator and indicates various problematic aspects, which makes it hard to estimate this uncertainty. First, Section 4.1 introduces an applied approach for representing the true uncertainty of the LOO-CV estimator. Then, Section 4.2 presents the current popular ways of estimating the uncertainty and highlights the problems associated with them later investigated in Sections 4.3–4.6. Finally, Section 4.7 draws some conclusions on the consequences of these behaviours on using LOO-CV in practice and interpreting its results.

4.1 Formulating the uncertainty

The uncertainty of the LOO-CV estimator can be formulated in different ways. This dissertation assumes the following generally applied interpretation, which is also studied in Publication III; given a data set y^{obs}, the stochastic variable representing the uncertainty about $\text{elpd}\left(\cdot \mid y^{\text{obs}}\right)$, either for model evaluation or for model comparison, when estimated with $\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right)$ is

$$^{\text{sv}}\widetilde{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right) = \widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right) - {}^{\text{sv}}\text{err}_{\text{LOO}}\left(\cdot \mid y\right), \tag{4.1}$$

where

$$^{\text{sv}}\text{err}_{\text{LOO}}\left(\cdot \mid y\right) = {}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y\right) - {}^{\text{sv}}\text{elpd}\left(\cdot \mid y\right) \tag{4.2}$$

is the distribution of the approximation error over possible data sets from the respective data generating mechanism. Appendix A in Publication III discusses the differences in the uncertainty when LOO-CV is used to estimate e-elpd instead of elpd and Appendix B in Publication III discusses various other approaches for formulating the uncertainty.

4.2 Estimating the uncertainty

Currently, there are two popular ways of estimating the uncertainty of a LOO-CV estimate presented in Equation (4.1): the normal approximation and the Bayesian bootstrap approximation (Vehtari and Lampinen, 2002; Vehtari and Ojanen, 2012; Vehtari et al., 2017; Yao et al., 2018). In the normal approximation, $^{\text{sv}}\widetilde{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right)$ is approximated with

$$^{\text{sv}}\widehat{\widetilde{\text{elpd}}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right) \sim \text{N}\left(\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right), \widehat{\text{SE}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right)\right), \tag{4.3}$$

where

$$\widehat{\text{SE}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right)^2 = \frac{n}{n-1}\sum_{i=1}^{n}\left(\widehat{\text{elpd}}_{\text{LOO},i}\left(\cdot \mid y^{\text{obs}}\right) - \frac{1}{n}\sum_{j=1}^{n}\widehat{\text{elpd}}_{\text{LOO},j}\left(\cdot \mid y^{\text{obs}}\right)\right)^2, \tag{4.4}$$

the sample variance of the individual LOO-CV terms multiplied by n, is an estimate for the variance of $^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y\right)$. The Bayesian bootstrap approximation applies a Dirichlet distribution to the terms $\widehat{\text{elpd}}_{\text{LOO},i}\left(\cdot \mid y^{\text{obs}}\right)$, $i = 1, 2, \ldots, n$ to model their sum and uses that as an approximation for the uncertainty (Rubin, 1981; Vehtari and Lampinen, 2002). For a good estimator for the uncertainty, the distribution of $^{\text{sv}}\widehat{\widetilde{\text{elpd}}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right)$ over possible data sets is a good representation of $\text{elpd}\left(\cdot \mid y^{\text{obs}}\right)$:

$$p\left(^{\text{sv}}\widehat{\widetilde{\text{elpd}}}_{\text{LOO}}\left(\cdot \mid y\right)\right) \approx p\left(^{\text{sv}}\text{elpd}\left(\cdot \mid y\right)\right). \tag{4.5}$$

For this, the calibration of the estimator can be analysed by the probability integral transform (PIT) method (see for example Gneiting et al., 2007).

As discussed in Publication III, estimating the uncertainty of a LOO-CV estimate, presented in Equation (4.1), is problematic due to multiple reasons:

1. No unbiased estimator for the variance of the sampling distribution $^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y\right)$ exist in general (Bengio and Grandvalet, 2004).
2. The sampling distribution $^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y\right)$ may be highly skewed.
3. The connection between the sampling distribution $^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y\right)$ and the uncertainty $^{\text{sv}}\widetilde{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right)$ may be weak.

While the normal and Bayesian bootstrap approximations are easy to obtain, the aforementioned aspects affect the behaviour of these approximations so that they may be badly calibrated in some situations. In the model comparison setting, the problematic cases include, but are possibly not limited to,

1. small data size,
2. models that make similar predictions, and

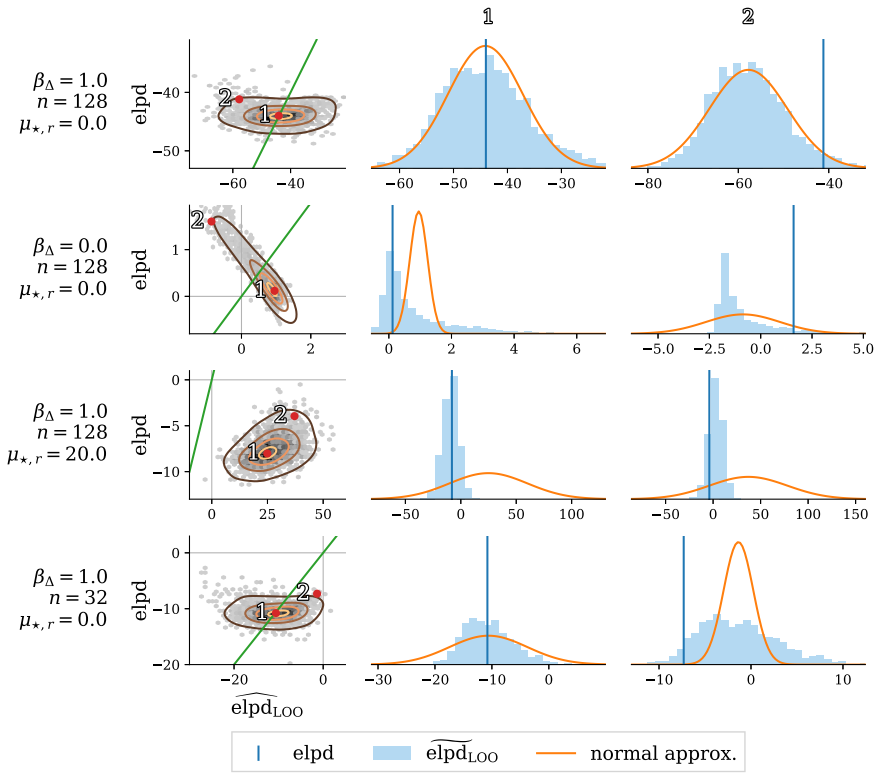

Figure 4.1. Demonstration of the estimated uncertainty in different problem settings in two cases: near the mode (labelled with 1) and at the tail area (labelled with 2) of the distribution of the predictive performance and its estimate. Parameter β_Δ controls the difference in the predictive performance of the models, n corresponds to the size of the data set, and $\mu_{\star,r}$ to the magnitude of an outlier observation. The experiments are described in more detail in Section 4 in Publication III. In the plots in the first column, the green diagonal line indicates where ${}^{\text{sv}}\text{elpd}\big(\text{M}_a, \text{M}_b \mid y\big) = {}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\big(\text{M}_a, \text{M}_b \mid y\big)$ and the brown-yellow lines illustrate density isocontours estimated with Gaussian kernel method with bandwidth 0.5. Bayesian bootstrap approximation to the uncertainty resembles the normal approximation in all the illustrated cases. Figure adopted from Publication III (Figure 2, p. 10).

3. misspecified models.

Apart from the case of models that make similar predictions, in addition to the model comparison problem setting discussed in Publication III, these problematic cases are likely to appear also in the case of model evaluation.

Figure 4.1 demonstrates the behaviour of the estimand ${}^{\text{sv}}\text{elpd}\big(\text{M}_a, \text{M}_b \mid y\big)$ and the LOO-CV estimate ${}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\big(\text{M}_a, \text{M}_b \mid y\big)$ over possible data sets from a known data generating mechanism in four different model comparison problem settings: well-behaving setting, models that make similar predictions, misspecified models, and small data size. Furthermore, the normal approximation ${}^{\text{sv}}\widehat{\widetilde{\text{elpd}}}_{\text{LOO}}\big(\text{M}_a, \text{M}_b \mid y^{\text{obs}}\big)$ is compared to the true uncertainty ${}^{\text{sv}}\widetilde{\text{elpd}}_{\text{LOO}}\big(\text{M}_a, \text{M}_b \mid y^{\text{obs}}\big)$ in two cases in each setting. Various points of interests can be seen from the figure. The following summarises the main

points:

1. The normal approximation can perform well in clear problem settings.
2. The estimand elpd and the estimator $\widehat{\text{elpd}}_{\text{LOO}}$ can be strongly negatively correlated when the models make similar predictions.
3. Outliers can make the LOO-CV estimator biased.
4. The normal approximation can under or overestimate the uncertainty in problematic settings.

The following sections 4.3–4.6 further discuss the reasons for the problematic cases.

4.3 Variance of the sampling distribution

The estimator of the variance of the sampling distribution presented in Equation (4.4) uses the sample variance of the individual LOO-CV terms $\widehat{\text{elpd}}_{\text{LOO},i}\left(\cdot \mid y^{\text{obs}}\right)$, $i = 1,2,\ldots,n$ multiplied by n to estimate the variance of the sum of these terms. This estimate relies on the assumption that the individual terms ${}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO},i}\left(\cdot \mid y\right)$ are independent and have the same variance so that

$$\text{Var}\left(\sum_{i=1}^{n} {}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO},i}\left(\cdot \mid y\right)\right) = \sum_{i=1}^{n} \text{Var}\left({}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO},i}\left(\cdot \mid y\right)\right) = n\sigma^2_{\text{LOO}}, \tag{4.6}$$

where

$$\sigma^2_{\text{LOO}} = \text{Var}\left({}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO},i}\left(\cdot \mid y\right)\right), \ \forall i = 1,2,\ldots,n. \tag{4.7}$$

However, while the assumption of equal variance is reasonable, the assumption of independence of the terms is not; each term ${}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO},i}\left(\cdot \mid y\right)$ depends on every observation y_i either as a part of the training set or as the predicted observation. As shown in Proposition 1 in Appendix C.1 in Publication III, the real variance of the LOO-CV estimator depends also on the covariance between individual terms:

$$\text{Var}\left({}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y\right)\right) = n\sigma^2_{\text{LOO}} + n(n-1)\gamma_{\text{LOO}}, \tag{4.8}$$

where σ^2_{LOO} is the variance of a LOO-CV term as defined in Equation (4.7) and

$$\gamma_{\text{LOO}} = \text{Cov}\left({}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO},i}\left(\cdot \mid y\right), {}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO},j}\left(\cdot \mid y\right)\right), \ \forall i,j = 1,2,\ldots,n, \ i \neq j. \tag{4.9}$$

On the other hand, as shown in Proposition 2 in Appendix C.1 in Publication III, the expectation of the estimator ${}^{\text{sv}}\widehat{\text{SE}}_{\text{LOO}}\left(\cdot \mid y\right)^2$ is $n\sigma^2_{\text{LOO}} - n\gamma_{\text{LOO}}$. Consequently, the bias of this estimator is $-n^2\gamma_{\text{LOO}}$. Being simple to obtain but limited in behaviour, the estimator ${}^{\text{sv}}\widehat{\text{SE}}_{\text{LOO}}\left(\cdot \mid y\right)^2$ is referred to as the naive variance estimator of the sampling distribution.

4.3.1 No unbiased sample variance estimator in general

A discouraging result by Bengio and Grandvalet (2004) states that, given the pointwise LOO-CV estimates $\widehat{\mathrm{elpd}}_{\mathrm{LOO},,}\left(\cdot \mid y^{\mathrm{obs}}\right)$ $i = 1,2,\ldots,n$, for any single model evaluation $\cdot = \mathrm{M}_k$ or model comparison $\cdot = \mathrm{M}_a, \mathrm{M}_b$, there is no unbiased estimate for the variance $\mathrm{Var}\left(^{\mathrm{sv}}\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\cdot \mid y\right)\right)$. They show experimentally that the contribution of γ_{LOO} can be of the same order of the contribution of σ^2_{LOO} in the variance presented in Equation (4.8) and thus it should be taken into account. Furthermore, it can be seen from the theoretically possible values of these terms (Bengio and Grandvalet, 2004, Lemma 8), that the expectation of the naive variance estimator can be infinitely too small or big compared to the true variance.

4.3.2 Improved problem specific variance estimators

Although Bengio and Grandvalet (2004) shows that there are no unbiased estimators for the sampling distribution of the LOO-CV variance in general, it could be possible to derive such estimators for specific models. Publication IV presents an unbiased variance estimator in the case of LOO-CV model evaluation of a simple Bayesian normal model with fixed variance. While the theorem by Bengio and Grandvalet (2004) deals with estimators based on the individual LOO-CV terms $\widehat{\mathrm{elpd}}_{\mathrm{LOO},,}\left(\cdot \mid y^{\mathrm{obs}}\right)$ $i = 1,2,\ldots,n$, the estimator in Publication IV utilises the specific model to form the estimate directly based on the observations y_i^{obs}, $i = 1,2,\ldots,n$. The same approach could be extended to other model settings and LOO-CV model comparison. However, analytic derivations required for an unbiased or improved estimator might not be possible in more complex problem settings, unlike in the case of the evaluation of the simple model in Publication IV. In these cases, some numerical approximation methods could be utilised instead. To our knowledge, the possibility of problem-specific LOO-CV variance estimators has not been extensively discussed before in literature.

As shown experimentally for example by Varoquaux et al. (2017) and Varoquaux (2018), the variance of the sampling distribution is often considerably underestimated in the model evaluation case using the naive variance estimate in the normal approximation. Furthermore, publication IV studies the bias theoretically under an example problem setting. Figure 4.2 illustrates the relative bias of the naive estimator and the improved estimator proposed in Publication IV in the scale of standard deviation. In this example problem setting, the theoretical expectation of the naive estimator is around 0.95, 1.15, and 0.8 times the true standard deviation when applied with well-matching, under-dispersed, and under-dispersed skewed heavy-tailed data respectively (0.9, 1.3, and 0.6 in the scale of the variance). The proposed improved estimator is unbiased in all these cases. Furthermore, Publication III identifies similar behaviour also

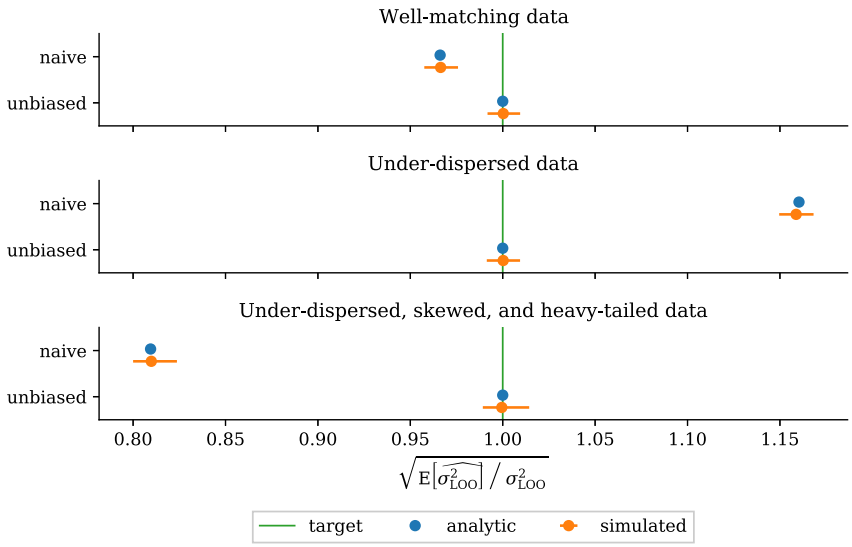

Figure 4.2. The expectation of the naive and unbiased LOO-CV variance estimators $\widehat{\sigma^2_{\mathrm{LOO}}}$ estimated using Bayesian bootstrap (BB, Rubin, 1981) in a simulated experiment under three different data generating mechanisms: well matching, under-dispersed, and under-dispersed skewed heavy-tailed data respectively. More information on the problem setting can be found from Section 3 in Publication IV. The x-axis is transformed to the square root of the ratio to the LOO-CV estimator's true variance σ^2_{LOO}. The analytic expectations (blue) match the simulated results (yellow) in all cases. The BB uncertainty is illustrated using a dot and a line corresponding to the mean and 95 % credible interval respectively. The naive estimator underestimates or overestimates the variance while the proposed improved estimator is unbiased. Figure adopted from Publication IV (Figure 1, p. 5).

for the model comparison case and for the Bayesian bootstrap approximation.

While exact unbiasedness itself is not necessary for a well-calibrated uncertainty estimator, it often connects to badly represented variability and bad calibrations. Figure 4.3 illustrates the connection of the bias to the other problematic properties affecting the calibration in a simulated Bayesian linear regression model comparison experiment adapted from Publication III. However, the calibration is also affected by other properties of the estimator and thus small bias—relative to the true variance of the sampling distribution—might not be problematic. Deriving improved problem-specific estimators of the variance, not necessarily unbiased ones, would likely also improve the calibration of the estimated uncertainty as a whole when applied to the normal approximation instead of the naive variance estimator. More research is needed for exploring the possibilities for such estimators and their effect on the calibration.

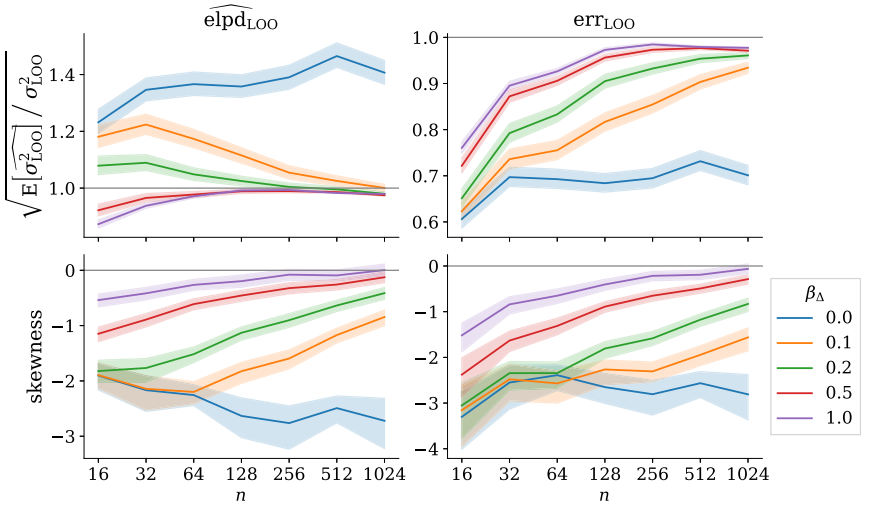

Figure 4.3. The relative expectation of the naive LOO-CV variance estimator $\widehat{\sigma^2_{\mathrm{LOO}}}$ in the scale of standard deviation as a function of the data size n in a model comparison problem setting for different magnitude of non-shared covariate effects β_Δ. The models make more similar predictions when β_Δ is small. The expectation is illustrated relative to the true variance σ^2_{LOO} of the sampling distribution ${}^{\mathrm{sv}}\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_a, \mathrm{M}_b \mid y\right)$ and the error distribution ${}^{\mathrm{sv}}\mathrm{err}_{\mathrm{LOO}}\left(\mathrm{M}_a, \mathrm{M}_b \mid y\right)$ The figure is adapted from the results of the experiments in Section 4 in Publication III. More details of the problem setting are given in the original work. Large bias in the variance estimator for the sampling distribution occurs together with a larger magnitude of skewness and underestimated variability in the error distribution. The problematic cases occur when β_Δ is small or with small sample sizes n.

4.4 Skewness of the sampling distribution

When estimating the uncertainty of a LOO-CV estimate, presented in Equation (4.3), estimating the variability, as discussed in Section 4.3, plays a great role. However, using only the variance in describing the uncertainty is a limited approach. In particular, when the estimated distribution is skewed and possibly heavy-tailed, variance as such is insufficient at representing the uncertainty. Even if the variance could be accurately estimated or known, the calibration of the estimated uncertainty can be very bad.

Publication III studies the possible skewness of the LOO-CV estimate and the error in a model comparison case, ${}^{\mathrm{sv}}\widehat{\mathrm{elpd}}_{\mathrm{LOO}}\left(\mathrm{M}_a, \mathrm{M}_b \mid y\right)$ and ${}^{\mathrm{sv}}\mathrm{err}_{\mathrm{LOO}}\left(\mathrm{M}_a, \mathrm{M}_b \mid y\right)$ respectively, both theoretically and experimentally in a simulated case study with known data generating mechanism. The results of the simulated experiment show that these distributions can be considerably skewed in the problematic cases of small data and models with similar predictions and situationally in the case of model misspecification. Figure 4.4 illustrates the skewness for a model comparison problem setting as a function of n with varying degree of difference in the predictions of

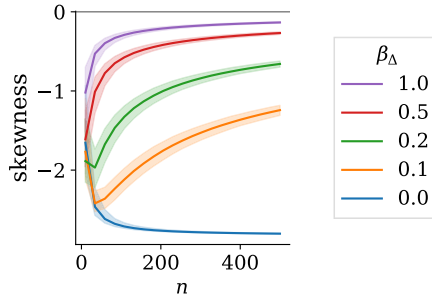

Figure 4.4. Illustration of the skewness conditional on the design matrix X for $^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\text{M}_a, \text{M}_b \mid y\right)$ as a function of the data size n. The data consist of an intercept and two covariates. One of the covariates with true effect β_Δ is considered only in the model M_b. The models make more similar predictions when β_Δ is small. The solid lines correspond to the median and the shaded area illustrates the 95 % confidence interval based on 2000 Xs independently simulated from the standard normal distribution. The problematic skewness of the error occurs with small n and β_Δ. It can also be seen that, when $\beta_\Delta = 0$, the magnitude of skewness does not fade away when n grows. Figure adapted from Publication III (Figure 4, p. 14).

the compared models. Furthermore, the theoretical analysis in the Publication III shows that in some cases, the problematic skewness does not fade away when n grows but converges into a constant nonzero value.

The popular normal approximation to the uncertainty of a LOO-CV estimate, presented in Equation (4.3), can not model the skewness of the estimated distribution. The other popular method, the Bayesian bootstrap approximation, is able to represent the skewness but has some problems in modelling skewed and possible heavy-tailed distributions (Rubin, 1981). Thus these methods may be badly calibrated in some situations. Publication III illustrates this unwanted behaviour in the problematic cases of skewed distribution. In order to have a robust, well-calibrated estimator for the uncertainty, the possible skewness should be taken into account in the estimator.

4.5 Bad connection between the sampling and the error distribution

Sections 4.3 and 4.4 discuss the difficulty of estimating the variability of the sampling distribution $^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y\right)$. Naturally, these problems also affect estimating the uncertainty $^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y^{\text{obs}}\right)$ presented in Equation (4.1). However, in addition to the sampling distribution, the true uncertainty is also affected by the distribution of the estimand $^{\text{sv}}\text{elpd}\left(\cdot \mid y\right)$. Instead of the sampling distribution $^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\cdot \mid y\right)$, the error distribution $^{\text{sv}}\text{err}_{\text{LOO}}\left(\cdot \mid y\right)$ is the distribution of interest in estimating the uncertainty.

Publication III analyses the sampling and error distributions in an ex-

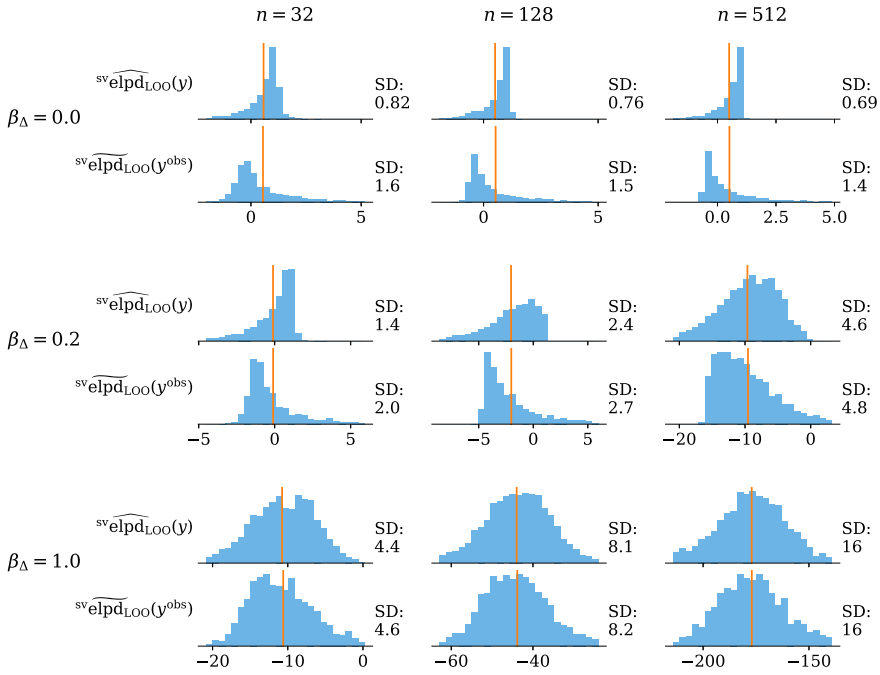

Figure 4.5. Illustration of the distributions of $\widehat{\text{elpd}}_{\text{LOO}}^{\text{sv}}(\text{M}_a, \text{M}_b \mid y)$ and $\widehat{\text{elpd}}_{\text{LOO}}^{\text{sv}}(\text{M}_a, \text{M}_b \mid y^{\text{obs}})$, where y^{obs} is such that $\widehat{\text{elpd}}_{\text{LOO}}(\text{M}_a, \text{M}_b \mid y^{\text{obs}}) = \text{E}\left[\widehat{\text{elpd}}_{\text{LOO}}^{\text{sv}}(\text{M}_a, \text{M}_b \mid y)\right]$, for various data sizes n and non-shared covariate effects β_Δ. The estimated standard deviation (SD) is indicated next to each histogram. The models make more similar predictions when β_Δ is small. The yellow lines show the means of the distributions and the corresponding sample standard deviation is displayed next to each histogram. In the problematic cases with small n and β_Δ, there is a weak connection in the skewness of the sampling and the error distributions. For brevity, model labels are omitted in the notation in the figure. Figure adopted from Publication III (Figure 16, p. 86).

ample Bayesian linear regression model comparison problem setting and shows that the connection between them can be weak in problematic situations. Figure 4.5 illustrates these distributions in a few different settings. It can be seen from the figure, that when the models make similar predictions, the distributions are considerably different. Namely, the standard deviation of the sampling distribution is smaller than of the error distribution and they are skewed to the opposite directions. Publication III further shows that increasing the number of observations decreases the difference in the standard deviation and more so when there is a larger difference in the predictions of the models. Because of the possible weak connection between the sampling and error distribution, even if the sampling distribution $\widehat{\text{elpd}}_{\text{LOO}}^{\text{sv}}(\cdot \mid y)$ would be known explicitly, it would not yield an optimally calibrated estimate for the uncertainty.

4.6 Model misspecification

Model misspecification may affect the behaviour of the LOO-CV estimate and the estimated uncertainty considerably. Publication III studies this behaviour both theoretically and experimentally in an example Bayesian linear regression problem setting involving one outlier observation with varying magnitude. While LOO-CV is generally asymptotically unbiased, the introduced outlier affects the convergence rate so that considerable bias is introduced even with large data sizes. The bias inevitably affects also the estimated uncertainty so that the calibration is bad; one would need to learn this bias in order to have good calibration.

In addition to introducing bias, the outlier also affects the skewness of the sampling distribution ${}^{\text{sv}}\widehat{\text{elpd}}_{\text{LOO}}\left(\text{M}_a, \text{M}_b \mid y\right)$ and the error distribution ${}^{\text{sv}}\text{err}_{\text{LOO}}\left(\text{M}_a, \text{M}_b \mid y\right)$; increase in the magnitude in the outlier decreases the magnitude of the skewness. This may situationally work in favour of better calibration of the estimated uncertainty as, at some level of the magnitude of the outlier, the positive effect of decreasing the skewness may be more effective than the negative effect of increasing the bias. Section 4.4 in Publication III illustrates one such case.

4.7 Consequences of the uncertainty in the estimated uncertainty

As discussed in sections 4.3–4.6, various reasons affect the behaviour of the LOO-CV sampling and error distributions and consequently the accuracy, calibration, and usability of the normal and Bayesian bootstrap approximation. Figure 4.6 illustrates the calibration of these approximations under various model comparison problem settings. It can be seen from the figure, that the calibration can be bad in the problematic cases: the models make similar predictions, there are outliers in the data, or the sample size is small.

In order to have an improved, robust estimator for the uncertainty with good calibration, the aforementioned problematic aspects should be taken into consideration in the estimator. Alternatively, the problematic cases should be acknowledged and diagnosed so that the uncertainty in the estimated uncertainty can be dealt with appropriately.

The actions taken upon diagnosed problematic cases vary by the application. Nevertheless, as discussed in the Publication III, proper model checking and expansion should be applied before using LOO-CV in order to avoid model misspecification in general. In addition, both in the case of model evaluation and comparison, the possibility for inaccurately estimated uncertainty of the LOO-CV estimate should be taken into account when the number of observations is small. In the case of model comparison, when the models make similar predictions, the difference in the predictive

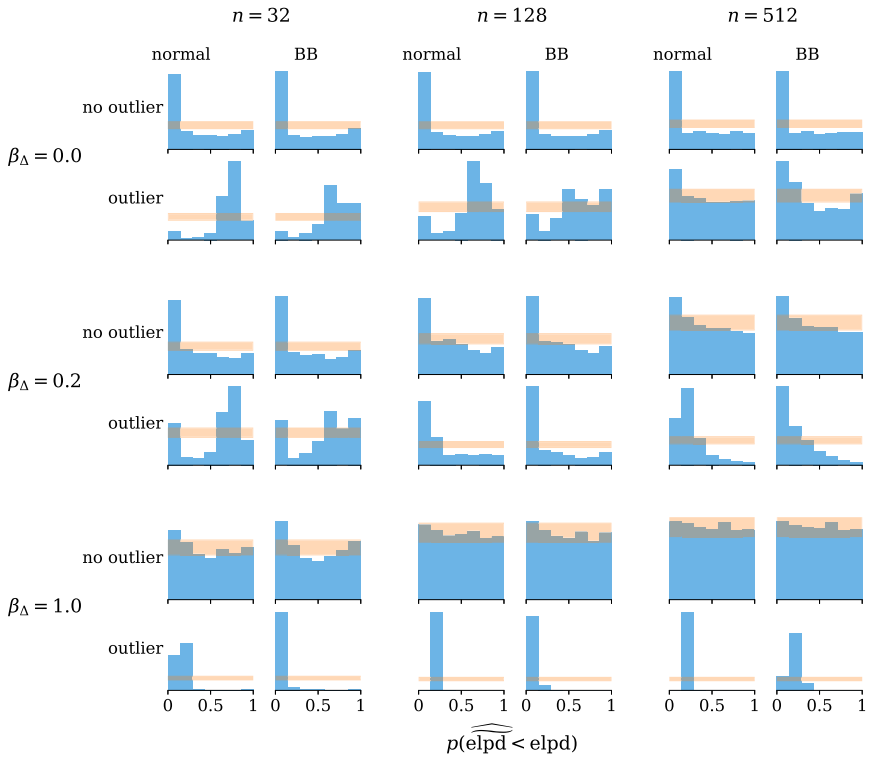

Figure 4.6. Calibration of the estimated uncertainty $^{\mathrm{sv}}\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(\mathrm{M}_a, \mathrm{M}_b \mid y)$ discussed in Equation (4.5) for various data sizes n and non-shared covariate effects β_Δ. The models make more similar predictions when β_Δ is small. The histograms illustrate the PIT values $p\left(^{\mathrm{sv}}\widehat{\mathrm{elpd}}_{\mathrm{LOO}}(\mathrm{M}_a, \mathrm{M}_b \mid y^{\mathrm{obs}}) < \mathrm{elpd}(\mathrm{M}_a, \mathrm{M}_b \mid y^{\mathrm{obs}})\right)$ over simulated data sets y, which would be uniform in a case of optimal calibration (see for example Gneiting et al., 2007; Talts et al., 2018). The yellow shading indicates the range of 99 % of the variation expected from uniformity. Two uncertainty estimators are presented: normal approximation and Bayesian bootstrap (BB). Cases with and without an outlier in the data are presented. The outlier observation has a deviated mean of 20 times the standard deviation of y_i. The calibration is better when β_Δ is large or n is big. The outlier makes the calibration worse, although with large n and small β_Δ, the calibration can be better. The latter behaviour is, however, situational to the selected magnitude of the outlier. Figure adopted from Publication III (Figure 9, p. 21).

performance is small and the uncertainty is badly calibrated. Because of this, it is hard to differentiate between small and zero effect sizes using LOO-CV.

5. Estimating the predictive performance of Gaussian latent variable models

As discussed in Chapter 3, assessing the predictive performance of a model is an important step in Bayesian data analysis, for which LOO-CV is a popular approximative approach. However, as discussed in Section 3.4, the naive brute force evaluation of the exact LOO-CV estimate involves fitting the model and evaluating the predictive density of an observation n times. This can be a costly operation for models with intractable posterior predictive distribution, such as occur with Gaussian latent variable models (GLVM). Publication I compares different approximations for the LOO-CV estimate together with various approaches for obtaining them and handling the hyperparameter inference in the context of GLVMs.

Reflecting the work in Publication I, this chapter discusses the process of estimating the predictive performance of GLVMs. First, the GLVMs are introduced in Section 5.1. Then, Section 5.2 discusses various approaches for estimating the LOO-CV predictive performance estimate. Finally, a summary of the suggested approaches is discussed in Section 5.3.

5.1 Gaussian latent variable models

GLVMs are a class of models, in which a response variable $y = [y_1, y_2, \ldots, y_n]$ is modeled given an explanatory variable $X = [x_1, x_2, \ldots, x_n]$ via a latent variable $f = [f_1, f_2, \ldots, f_n]$. The latent values have a joint Gaussian prior distribution $p(f \mid X, \theta)$, which depend on the explanatory variable X and hyperparameters θ. Similar to the Publication I, this dissertation focuses on models with factorisable likelihood; Observation y_i depends on the local latent value f_i and possibly some global parameters ϕ via an observation model $p(y_i \mid f_i, \phi)$. The conditional posterior of the latent values f is

$$p(f \mid y, X, \theta, \phi) = \frac{1}{Z} p(f \mid X, \theta) \prod_{i=1}^{n} p(y_i \mid f_i, \phi), \tag{5.1}$$

where the normalising factor Z corresponds to the marginal likelihood

$$p(y \mid X,\theta,\phi) = \int p(f \mid X,\theta) \prod_{i=1}^{n} p(y_i \mid f_i,\phi) \mathrm{d}f \,. \tag{5.2}$$

One popular category of submodels, used in the experiments in the Publication I, is the Gaussian process (GP) models (reviewed for example by Rasmussen and Williams, 2006). In a GP model, the prior on the latent values is multivariate Gaussian, $p(f \mid X,\theta) = \mathrm{N}(\mu_0,K)$, where μ_0 is the prior mean (often assumed to be zero) and for the covariance matrix, $K_{i,j} = k(x_i,x_j;\theta), i,j = 1,2,\ldots,n$, where $k(x_i,x_j;\theta)$ is a covariance function with parameters θ. The selected covariance function characterises the correlation between two points in their latent values.

5.1.1 Computing conditional posterior distribution

When the observation model is Gaussian, $p(y_i \mid f_i,\sigma^2) = \mathrm{N}(y_i \mid f_i,\sigma^2)$, where the global hyperparameter $\phi = \sigma^2$ is the noise variance, the conditional posterior of the latent values $p(f \mid y,X,\theta,\phi)$ and the respective marginal likelihood can be obtained analytically (see for example Rasmussen and Williams, 2006, Section 2.2). With a non-Gaussian likelihood however, these distributions usually need to be approximated. Popular methods for this approximation include expectation propagation (EP), reviewed in Chapter 2, and Laplace method (see for example Rasmussen and Williams, 2006; Gelman et al., 2013). These methods apply local Gaussian approximations to each likelihood component to obtain a global multivariate Gaussian approximation to the conditional posterior:

$$g(f \mid y,X,\theta,\phi) = \propto p(f \mid X,\theta) \prod_{i=1}^{n} g_i(f_i). \tag{5.3}$$

The EP and Laplace method has been shown to perform well for various problems (see for example Nickisch and Rasmussen, 2008; Rue et al., 2009; Vanhatalo and Vehtari, 2010; Vanhatalo et al., 2010; Jylänki et al., 2011; Cseke and Heskes, 2011; Riihimäki et al., 2013; Vanhatalo et al., 2013; Martins et al., 2013; Riihimäki and Vehtari, 2014; Tolvanen et al., 2014).

In the EP method, each cavity and tilted distribution, presented in equations (2.3) and (2.4) respectively, consider only one latent variable. Consequently, the inference on the moments of the one-dimensional tilted distribution can often be carried out analytically, as in the case of probit-likelihood $p(y_i \mid f_i) = \Phi(f_i y_i)$, or relatively efficiently using numerical integration. The marginal likelihood $p(y \mid X,\theta,\phi)$ can be approximated at the convergence by estimating the marginal likelihood $\hat{Z}_i = \int p(y_i \mid f_i) g_{-i}(f_i) \mathrm{d}f_i$ at each site and combining them (see for example Rasmussen and Williams, 2006).

As discussed in Section 2.4 in PublicationI, the Laplace method involves finding an approximation $\hat{f}$ for the mode of the posterior to apply the second-order Taylor expansion at, yielding a Gaussian approximation

$$g(f \mid y, X, \theta, \phi) = \mathrm{N}(\hat{f}, \hat{\Sigma}) \tag{5.4}$$

for the conditional posterior and an approximation for the marginal likelihood, where $\hat{\Sigma} = (K^{-1} + \tilde{\Sigma}^{-1})^{-1}$ and $\tilde{\Sigma}$ is diagonal with elements $\tilde{\Sigma}_i = -(\nabla_i \nabla_i \log p(y_i|f_i,\phi)|_{f_i=\hat{f}_i})^{-1}$. This approximation can also be matched with the factorised form in Equation (5.3). For the Laplace approximation, the site approximations can be formulated as

$$g_i^{\text{Laplace}}(f_i) = \mathrm{N}(f_i \mid \tilde{\mu}_i, \tilde{\Sigma}_i), \tag{5.5}$$

where $\tilde{\mu}_i = \hat{f} + \tilde{\Sigma}_i \nabla_i \log p(y_i|f_i,\phi)|_{f_i=\hat{f}_i}$. For more detailed description of the Laplace method, see for example Rasmussen and Williams (2006) or Gelman et al. (2013). This form is useful for obtaining marginal posterior approximations discussed in Section 5.1.2.

5.1.2 Marginal posterior distribution

Many approaches for approximating the LOO-CV estimate involve the marginal posterior distribution $p(f_i \mid y, X, \theta, \phi)$. As reviewed for example by Cseke and Heskes (2011), various methods have been proposed for approximating this distribution. The simplest approach is to use the marginal $g(f_i \mid y, X, \theta, \phi)$ of the obtained EP or Laplace Gaussian approximations to the joint posterior.

Instead of simply approximating the marginal posterior using the joint approximation, further improvement can be made by considering (conditionalisation for y, X, θ, ϕ has been dropped for brevity)

$$p(f_i) \propto g(f_i)\epsilon_i(f_i) \underbrace{\int g(f_{-i} \mid f_i) \prod_{j \neq i} \epsilon_j(f_j) \mathrm{d}f_{-i}}_{c_i(f_i)}, \tag{5.6}$$

where $\epsilon_i(f_i) = p(y_i \mid f_i)/g_i(f_i)$ is the ratio of the true likelihood and the site approximation. Local correction methods consider only the effect of $\epsilon_i(f_i)$ while global correction methods also consider $c_i(f_i)$.

The tilted distribution inferred in the EP method corresponds to one such local correction method and is obtained without any additional computation while obtaining the joint posterior approximation. Analogous local correction for the Laplace method can be derived by defining respective Laplace based cavity and tilted distribution: $\frac{g^{\text{Laplace}}(f|y,X,\theta,\phi)}{g_i^{\text{Laplace}}(f_i)} p(y_i \mid f_i, \phi)$ (Cseke and Heskes, 2011). Publication I proposes an alternative way of computing an equivalent approximation where the Laplace cavity distribution is formulated using linear response theory.

Notable global correction methods include methods called EP-FACT and LA-CM2 based on the EP and Laplace approximations to the joint posterior distribution respectively (Cseke and Heskes, 2011). The global correction methods are, however, more computationally complex. In addition, based on the experimental results by Cseke and Heskes (2011), compared to the local corrections, the difference is often small.

5.1.3 Hyperparameters

Various methods can be utilised for marginalising out the hyperparameters θ and ϕ from the conditional posterior distribution. The marginal posterior for the parameters

$$p(\theta,\phi \mid y,X) \propto p(\theta,\phi)p(y \mid X,\theta,\phi) \tag{5.7}$$

can be formed using the true or approximated marginal likelihood $p(y \mid X,\theta,\phi)$. This can be utilised to numerically integrate over θ and ϕ using for example various Monte Carlo methods (see a list of references by Vanhatalo et al., 2013), importance sampling (see for example Vehtari, 2001; Vehtari and Lampinen, 2002; Held et al., 2010), deterministic central composite design (CCD) method (Rue et al., 2009), or importance weighted CCD (Held et al., 2010; Vanhatalo et al., 2013).

In cases, where the marginal posterior distribution $p(\theta,\phi \mid y,X)$ is narrow, it may be sufficient to use type II maximum a posteriori (MAP) approximation instead of integrating over the hyperparameters, that is select $(\hat{\theta},\hat{\phi}) = \arg\max_{\theta,\phi} p(\theta,\phi \mid y,X)$. Such cases often occur when the dimensionality of the hyperparameters is small and n is large.

5.2 Approximating the LOO-CV estimate for GLVMs

Obtaining the LOO-CV predictive performance estimate presented in Equation (3.4) can be a computationally intensive task for GLVMs. Various approaches have been proposed for approximating it. This section introduces the problem setting and reviews some of the approximative approaches.

Hyperparameters In the context of estimating the predictive performance, hyperparameters can be considered on different levels. Instead of directly approximating the LOO predictive density $p(y_i \mid x_i, y_{-i}, X_{-i})$, one can first obtain the conditional predictive density $p(y_i \mid x_i, y_{-i}, X_{-i}, \theta, \phi)$ and the posterior for the hyperparameters $p(\theta,\phi \mid y_{-i}, X_{-i})$ and integrate out (θ,ϕ):

$$p(y_i \mid x_i, y_{-i}, X_{-i}) = \int p(y_i \mid x_i, y_{-i}, X_{-i}, \theta, \phi)p(\theta,\phi \mid y_{-i}, X_{-i})\mathrm{d}\theta\mathrm{d}\phi. \tag{5.8}$$

Furthermore, one can approximate $p(\theta,\phi \mid y_{-i}, X_{-i}) \approx p(\theta,\phi \mid y,X)$ in Equation (5.8) (Marshall and Spiegelhalter, 2003). Finally, as discussed in

Section 5.1.3, in some situations it is reasonable to apply type II MAP point estimate $(\hat{\theta},\hat{\phi})$ to approximate $p(y_i \mid x_i, y_{-i}, X_{-i}) \approx p(y_i \mid x_i, y_{-i}, X_{-i}, \hat{\theta}, \hat{\phi})$.

Full posterior formulation As demonstrated in Publication I, the LOO predictive density can be formulated in a couple of ways (dropping θ and ϕ for brevity):

$$p(y_i \mid x_i, y_{-i}, X_{-i}) = \int p(y_i \mid f_i) p(f_i \mid x_i, y_{-i}, X_{-i}) \mathrm{d}f_i \tag{5.9}$$

$$= \left(\int \frac{p(f_i \mid y, X)}{p(y_i \mid f_i)} \mathrm{d}f_i \right)^{-1}. \tag{5.10}$$

In some situations, $p(f_i \mid x_i, y_{-i}, X_{-i})$ can be obtained analytically or it can be approximated efficiently. Otherwise, one often needs to approximate $p(f_i \mid y, X)$ and the integration over f_i.

Gaussian likelihood When the likelihood $p(y_i \mid f_i, \phi)$ is Gaussian, the LOO predictive density can be obtained analytically. In this case, $p(f_i \mid x_i, y_{-i}, X_{-i}, \theta, \phi)$ corresponds to the cavity distribution discussed in the context of EP approximation in Section 5.1.1. Sundararajan and Keerthi (2001) present an alternative formulation for the corresponding distribution and use this to maximise the LOO log predictive density with respect to the hyperparameters.

EP and Laplace approximations Various general approaches for approximating the LOO-CV estimate, as discussed in Section 3.4, can also be applied in the context of GLVMs. In particular, the EP and Laplace methods, discussed in Section5.1 for GLVMs, provide convenient means for approximating the LOO-CV estimate via their marginal cavity distribution approximations discussed in Section 5.1.2. The marginal likelihood of the tilted distribution, which corresponds to the LOO predictive density, is obtained for free as a by-product of the EP algorithm and with small additional computation for the Laplace method. More details on the implementation in this context are provided in Publication I.

Quadrature approximations The numerical integration method by Held et al. (2010), discussed in Section 3.4, can also be conveniently applied to Equation (5.10) to yield a convenient one-dimensional quadrature approximation for the LOO-CV estimate. If applied together with EP or Laplace approximation for the marginal posterior distribution of the latent value using the local corrections, this method corresponds to the respective EP or Laplace approximation discussed above. More complex methods can be obtained by considering various other global correction methods for approximating the marginal posterior distribution discussed in Section 5.1.2. Publication I presents a more stable version of this method, in which the contributions from difficult observations are biased towards the full posterior behaviour in order to be more robust to approximation errors in the tails.

5.3 Suggested workflow

Based on experimental results, Publication I suggest a workflow with incremental levels of computational complexity for assessing the predictive performance in the context of GLVMs (list adopted from Publication I):

1. Find the MAP estimate $(\hat{\phi},\hat{\theta})$ using the Laplace method to approximately integrate over the latent values f.
2. Using $(\hat{\phi},\hat{\theta})$ obtained in the previous step, use EP to integrate over the latent values and check whether the predictive performance improves substantially compared to using the Laplace method (we may also re-estimate $\hat{\phi}$ and $\hat{\theta}$).
3. Integrate over ϕ and θ and check whether integration over the parameters improves predictive performance.

For the LOO-CV estimate conditional on the hyperparameters, the EP and Laplace method with local marginal corrections, discussed in sections 3.4 and 5.2, provide the best trade-off between computational cost and accuracy and are thus recommended. For full Bayesian LOO-CV approximation, both methods can be used together with importance sampling or importance weighted CCD for hyperparameter inference discussed in Section 5.1.3. The more complex quadrature approximations with global marginal corrections give useful results but they are considerably slower and often the EP method with local correction is more accurate. The WAIC method does not provide any benefits.

6. Discussion

This dissertation analyses several techniques in Bayesian data analysis, that utilise data partitioning as an approximative approach for distributing the inference or estimating the predictive performance. The discussed methods are widely used to address various intractable or unfeasible problems. The work focuses on analysing the expectation propagation (EP) algorithm as a framework for inference on a partitioned data, and the leave-one-out cross-validation (LOO-CV) method for model evaluation and comparison.

The main contribution of the dissertation is in improving the understanding of the behaviour of the methods and their capabilities. The accuracy, efficiency, and applicability of different implementations and adaptations of the methods and alternative approaches are reviewed and compared. The behaviour of the methods is analysed under different problem settings. Consequently, based on the results, the dissertation suggests approaches for different situations and proposes points of consideration.

The EP algorithm is a popular general algorithm useful for approximative Bayesian inference. In the past, motivated by the ability to perform the sub-inferences analytically, the method has mainly been applied by factorising the data set pointwise. However, by partitioning the data into bigger sets and possibly utilising more complex approximative methods for the sub-inferences, the algorithm can be seen as a framework for distributed inference with a trade-off between accuracy and efficiency in the number of partitions. In hierarchical settings, the framework can be used to reduce the dimensionality of the problem by distributing conditionally independent likelihood contributions to separate partitions.

In the context of the LOO-CV, the dissertation analyses the applicability and uncertainty of the method in various situations. By studying the frequency properties of the LOO-CV estimator both theoretically and experimentally in the model comparison setting, the following problematic cases are identified:

1. small data size,

2. comparing models that make similar predictions, and
3. misspecified models.

In these cases, the obtained estimates of the uncertainty related to the LOO-CV estimate can be badly calibrated. This emphasises the importance of model checking and highlights the uncertainty of the analysis when identifying small effects sizes and analysing small data sets.

Being inefficient to obtain exactly in practice, various approaches for approximating the LOO-CV estimator are studied and reviewed in the context of Gaussian latent variable models (GLVM). Respectively, a workflow with levels increasing in accuracy and computational complexity is presented. In addition, addressing the uncertainty of the estimator, an improved estimator for the variance of the LOO-CV estimator is presented in the case of a Bayesian normal model. The possibility of obtaining such model-specific estimators has not been discussed in the literature before and the presented estimator serves as an example of improving on the current general way of estimating the uncertainty. However, more research is needed to inspect the possibilities of this approach in a wider range of models.

References

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Tomohiro Ando and Ruey Tsay. Predictive likelihood for bayesian model selection and averaging. *International Journal of Forecasting*, 26(4):744 – 763, 2010. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2009.08.001. URL http://www.sciencedirect.com/science/article/pii/S0169207009001290.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

Anoop Korattikara Balan, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 181–189, 2014.

Simon Barthelmé and Nicolas Chopin. Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association*, 109:315–333, 2014.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of K-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105, 2004.

José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, 1994.

Leo Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, December 1996. doi: 10.1214/aos/1032181158. URL https://doi.org/10.1214/aos/1032181158.

Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, (in press), 2017.

Wilson Ye Chen and Matt Wand. Factor graph fragmentization of expectation propagation. *Journal of the Korean Statistical Society*, 49(3):722–756, January 2020. doi: 10.1007/s42952-019-00033-9.

Botond Cseke and Tom Heskes. Approximate marginals in latent gaussian models. *Journal of Machine Learning Research*, 12:417–454, 02 2011.

W. Edwards Deming. On a classification of the problems of statistical inference. *Journal of the American Statistical Association*, 37(218):173–185, 1942. doi: 10.1080/01621459.1942.10500624. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1942.10500624.

Seymour Geisser and William F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160, 1979. ISSN 01621459. URL http://www.jstor.org/stable/2286745.

Alan E. Gelfand. Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 145–162. Chapman & Hall, 1996.

Alan E. Gelfand, Dipak K. Dey, Hong Chang, Alan E. Gelf, Acce. Ior, Hong Chang U, Alan E. Gelf, Dipak K. Dey, and Hong Chang. Model determination using predictive distributions with implementation via sampling-based-methods (with discussion). In *In Bayesian Statistics 4*. University Press, 1992.

Andrew Gelman, Frederic Bois, and Jiming Jiang. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, 91:1400–1412, 1996.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2:1360–1383, 2008.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Taylor and Francis, 3rd edition, 2013. ISBN 9781439840955.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6): 997–1016, 2014.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of American Statistical Association*, 102:359–379, 2007.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed Bayesian learning with stochastic natural gradient expectation propagation and the posterior server. *Journal of Machine Learning Research*, 18(106):1–37, 2017.

Leonhard Held, Birgit Schrödle, and Håvard Rue. *Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA*, pages 91–110. Springer, 01 2010. doi: 10.1007/978-3-7908-2413-1_6.

Daniel Hernandez-Lobato and Jose Miguel Hernandez-Lobato. Scalable Gaussian process classification via expectation propagation. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 168–176, Cadiz, Spain, 2016.

Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John William Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1): 1303–1347, 2013.

Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12: 3227–3257, 2011.

David Madigan, Adrian E Raftery, C Volinsky, and J Hoeting. Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*, pages 77–83, 1996.

Måns Magnusson, Michael Andersen, Johan Jonasson, and Aki Vehtari. Bayesian leave-one-out cross-validation for large data. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 4244–4253. PMLR, 2019.

Måns Magnusson, Aki Vehtari, Johan Jonasson, and Michael Andersen. Leave-one-out cross-validation for bayesian model comparison in large data. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 341–351, Online, August 2020. PMLR.

E. C. Marshall and D. Spiegelhalter. Approximate cross-validatory predictive checks in disease mapping models. *Statistics in medicine*, 22 10:1649–60, 2003.

Thiago G Martins, Daniel Simpson, Finn Lindgren, and Håvard Rue. Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, 67:68–83, 2013.

Thomas P. Minka. *A family of algorithms for approximate Bayesian inference.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2001a.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller, editors, *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI-2001)*, pages 362–369. Morgan Kaufmann, San Francisco, Clif., 2001b.

Thomas P. Minka. Power EP. Technical report, Microsoft Research, Cambridge, 2004.

Thomas P. Minka. Divergence measures and message passing. Technical report, Microsoft Research, Cambridge, 2005.

Thomas P. Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-2002)*, pages 352–359. Morgan Kaufmann, San Francisco, CA, 2002.

Willie Neiswanger, Chong Wang, and Eric P. Xing. Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, UAI 2014, Quebec City, Quebec, Canada, July 23-27, 2014*, pages 623–632, 2014.

Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, October 2008.

Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.

Topi Paananen, Juho Piironen, Paul-Christian Bürkner, and Aki Vehtari. Implicitly adaptive importance sampling, 2020.

Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288, 1986.

Juho Piironen and Aki Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27:711–735, 2017.

Juho Piironen, Markus Paasiniemi, and Aki Vehtari. Projective inference in high-dimensional problems: Prediction and feature selection. *Electron. J. Statist.*, 14 (1):2155–2197, 2020. doi: 10.1214/20-EJS1711. URL https://doi.org/10.1214/20-EJS1711.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

Jaakko Riihimäki and Aki Vehtari. Laplace approximation for logistic Gaussian process density estimation and regression. *Bayesian analysis*, 9(2):425–448, 2014.

Jaakko Riihimäki, Pasi Jylänki, and Aki Vehtari. Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning Research*, 14:75–109, 2013.

Donald B. Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9(1):130–134, 1981.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal statistical Society B*, 71(2):319–392, 2009.

Swupnil Sahai. *Topics in Computational Bayesian Statistics With Applications to Hierarchical Models in Astronomy and Sociology*. PhD thesis, Columbia University Academic Commons, 2018.

Steven L. Scott, Alexander W. Blocker, and Fernando V. Bonassi. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 2016. URL http://www.tandfonline.com/doi/full/10.1080/17509653.2016.1142191.

Matthias Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

Alexander J. Smola, Vishy Vishwanathan, and Eleazar Eskin. Laplace propagation. In S. Thrun, L.K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing 16*, 2004.

S. Sundararajan and S. Sathiya Keerthi. Predictive approaches for choosing hyperparameters in gaussian processes. *Neural Computation*, 13(5):1103–1118, 2001.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration, 2018.

Ville Tolvanen, Pasi Jylänki, and Aki Vehtari. Expectation propagation for nonstationary heteroscedastic Gaussian process regression. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014. DOI:10.1109/MLSP.2014.6958906.

Jarno Vanhatalo and Aki Vehtari. Speeding up the binary Gaussian process classification. In Peter Grünwald and Peter Spirtes, editors, *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 1–9, 2010.

Jarno Vanhatalo, Ville Pietiläinen, and Aki Vehtari. Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15): 1580–1607, 2010.

Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. Gpstuff: Bayesian modeling with gaussian processes. *Journal of Machine Learning Research*, 14:1175–1179, April 2013. ISSN 1532-4435.

Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68 – 77, 2018.

Gaël Varoquaux, Pradeep Reddy Raamana, Denis A. Engemann, Andrés Hoyos-Idrobo, Yannick Schwartz, and Bertrand Thirion. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, 145:166 – 179, 2017.

Aki Vehtari. *Bayesian Model Assessment and Selection Using Expected Utilities*. PhD thesis, Helsinki University of Technology, 2001.

Aki Vehtari and Jouko Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468, 2002.

Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2019.

Xiangyu Wang and David B. Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Sumio Watanabe. Equations of states in singular statistical estimation. *Neural Networks*, 23(1):20–34, January 2010a. ISSN 0893-6080. doi: 10.1016/j.neunet.2009.08.002. URL https://doi.org/10.1016/j.neunet.2009.08.002.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010b.

John Winn and Christopher M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1003, 2018.

Onno Zoeter and Tom Heskes. Gaussian quadrature based expectation propagation. In Robert Cowell and Zoubin Ghahramani, editors, *International Workshop on Artificial Intelligence and Statistics (AISTATS)*, volume 10, 2005.

Aalto-DD 18/2021

ISBN 978-952-64-0268-0 (printed)
ISBN 978-952-64-0269-7 (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 DISSERTATIONS