Department of Computer Science

Bayesian Predictive Inference and Feature Selection for High-Dimensional Data

Juho Piironen

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 84/2019

Bayesian Predictive Inference and Feature Selection for High-Dimensional Data

Juho Piironen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 24 May 2019 at 12.

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

Supervising professor
Prof. Aki Vehtari, Aalto University, Finland

Thesis advisor
Prof. Aki Vehtari, Aalto University, Finland

Preliminary examiners
Prof. Jim Griffin, University College London, UK
Prof. James G. Scott, University of Texas at Austin, USA

Opponent
Dr. José Miguel Hernández-Lobato, University of Cambridge, UK

Aalto University publication series
DOCTORAL DISSERTATIONS 84/2019

ISBN 978-952-60-8538-8 (printed)
ISBN 978-952-60-8539-5 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-8539-5

Unigrafia Oy
Helsinki 2019

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Abstract

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

Author
Juho Piironen

Name of the doctoral dissertation
Bayesian Predictive Inference and Feature Selection for High-Dimensional Data

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 84/2019

Field of research Computational Science

Manuscript submitted 26 February 2019 **Date of the defence** 24 May 2019

Permission for public defence granted (date) 18 April 2019 **Language** English

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

Abstract

This thesis discusses Bayesian statistical inference in supervised learning problems where the data are scarce but the number of features large. The focus is on two important tasks. The first one is the prediction of some target variable of interest. The other task is feature selection, where the goal is to identify a small subset of features which are relevant for the prediction. A good predictive accuracy is often intrinsically valuable and a means to understanding the data. Feature selection can further help to make the model easier to interpret and reduce future costs if there is a price associated with predicting with many features.

Most traditional approaches try to solve both problems at once by formulating an estimation procedure that performs automatic or semiautomatic feature selection as a by-product of the predictive model fitting. This thesis argues that in many cases one can benefit from a decision theoretically justified two-stage approach. In this approach, one first constructs a model that predicts well but possibly uses many features. In the second stage, one then finds a minimal subset of features that can characterize the predictions of this model. The basic idea of this so called projective framework has been around for a long time but it has largely been overlooked in the statistics and machine learning community. This approach offers plenty of freedom for building an accurate prediction model as one does not need to care about feature selection at this point, and it turns out solving the feature selection problem often becomes substantially easier given an accurate prediction model that can be used as a reference.

The thesis focuses mostly on generalized linear models. To solve the problem of predictive model construction, the thesis introduces novel methods for encoding prior information about sparsity and regularization into the model. These methods can in some cases help to improve the prediction accuracy and robustify the posterior inference, but they also advance the current theoretical understanding of the fundamental characteristics of some commonly used prior distributions. The thesis explores also computationally efficient dimension reduction techniques that can be used as shortcuts for predictive model construction when the number of features is very large. Furthermore, the thesis develops the existing projective feature selection method further so as to make the computation fast and accurate for large number of features. Finally, the thesis takes the initial steps towards extending this framework to nonlinear and nonparametric Gaussian process models. The contributions of this thesis are solely methodological, but the benefits of the proposed methods are illustrated using example datasets from various fields, in particular from computational genetics.

Keywords Bayesian generalized linear models, feature selection, dimension reduction

ISBN (printed) 978-952-60-8538-8 **ISBN (pdf)** 978-952-60-8539-5

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2019

Pages 190 **urn** http://urn.fi/URN:ISBN:978-952-60-8539-5

Aalto-yliopisto

Tiivistelmä

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

Tekijä
Juho Piironen

Väitöskirjan nimi
Bayesilainen ennustava päättely ja piirrevalinta korkeaulotteisille aineistoille

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 84/2019

Tutkimusala Laskennallinen tiede

Käsikirjoituksen pvm 26.02.2019 **Väitöspäivä** 24.05.2019

Väittelyluvan myöntämispäivä 18.04.2019 **Kieli** Englanti

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

Tiivistelmä

Tämä väitöskirja käsittelee bayesilaista tilastollista päättelyä ohjatuissa oppimistehtävissä, joissa havaintoja on niukasti, mutta piirteiden määrä on suuri. Työssä keskitytään kahteen osaongelmaan. Ensimmäinen näistä on jonkin mielenkiinnon kohteena olevan muuttujan ennustaminen. Toinen ongelma on piirrevalinta, jossa tarkoituksena on löytää vain pieni joukko piirteitä, jotka ovat merkityksellisiä ennusteiden kannalta. Monissa tapauksissa hyvä ennustetarkkuus voi olla arvokasta sinällään ja usein auttaa ymmärtämään havaintoaineistoa. Piirrevalinta voi edelleen parantaa mallin tulkittavuutta ja selitettävyyttä, mutta sillä voidaan saavuttaa myös säästöjä, mikäli suuren piirremäärän käyttöön liittyy kustannuksia.

Valtaosa aiemmin ehdotetuista menetelmistä pyrkii ratkaisemaan molemmat ongelmat samanaikaisesti käyttäen estimointimenetelmää, jossa piirrevalinta saadaan varsinaisen ennustemallin sovittamisen sivutuotteena täysin tai lähes automaattisesti. Tässä työssä esitetään, että monissa tapauksissa voidaan päästä parempaan lopputulokseen, mikäli noudatetaan päätösteoreettisesti perusteltua kaksivaiheista lähestymistapaa. Tässä lähestymistavassa muodostetaan ensin malli, joka ennustaa hyvin, mutta joka mahdollisesti käyttää isoa määrää piirteitä. Piirrevalinta suoritetaan tämän jälkeen etsimällä pienin mahdollinen joukko piirteitä, joilla saavutetaan olennaisesti samanlaiset ennusteet kuin alkuperäisellä mallilla. Tätä niin kutsuttua projektiivista lähetysmistapaa on ehdotettu kirjallisuudessa jo kauan sitten, mutta menetelmä ei ole saanut ansaitsemaansa huomiota. Tämä menetelmä antaa paljon vapauksia ennustemallin rakentamiseen, koska mallintajan ei tässä vaiheessa tarvitse välittää piirrevalinnasta. Toisaalta piirrevalinta usein helpottuu huomattavasti, mikäli tässä vaiheessa voidaan hyödyntää aiemmin sovitettua tarkkaa ennustemallia ja käyttää tätä referenssinä.

Työssä keskitytään pääasiassa yleistettyihin lineaarimalleihin. Ennusteongelman ratkaisemiseksi työssä esitetään uusia menetelmiä harvuutta ja regularisointia koskevan priori-informaation sisällyttämiseksi ennustemalliin. Näillä menetelmillä voidaan joissakin tapauksissa parantaa mallin ennustekykyä ja tehdä mallin posteriori-laskennasta robustimpaa. Nämä tekniikat tuovat myös lisää teoreettista ymmärrystä eräiden usein käytettyjen priorijakaumien ominaisuuksista. Työssä tutkitaan myös laskennallisesti tehokkaita dimension redusointitekniikoita nopeuttamaan ennustemallin sovitusta havaintoaineistoissa, joissa piirteitä on hyvin paljon. Lisäksi työssä ehdotetaan alkuperäiseen projektiiviseen piirrevalintamenetelmään useita metodologisia parannuksia, joilla laskenta saadaan nopeaksi ja tarkaksi aineistoille, joissa piirteiden määrä on hyvin suuri. Työssä tutkitaan alustavasti myös, kuinka projektiivinen muuttujavalinta voidaan toteuttaa epälineaarisille ja ei-parametrisille malleille kuten gaussisille prosesseille. Väitöskirjan kontribuutiot ovat täysin metodologisia, mutta esitettyjen tekniikoiden etuja havainnollistetaan esimerkkiaineistoilla useilta sovellusaloilta, erityisesti laskennallisesta genetiikasta.

Avainsanat Bayesilaiset yleistetyt lineaarimallit, piirrevalinta, dimension redusointi

ISBN (painettu) 978-952-60-8538-8 **ISBN (pdf)** 978-952-60-8539-5

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2019

Sivumäärä 190 **urn** http://urn.fi/URN:ISBN:978-952-60-8539-5

Preface

This thesis is the culmination of a learning process and work that has taken place in Aalto University during years 2014–2018. The research was carried out first in the Bayesian Methodology group in the Department of Biomedical Engineering and Computational Science (BECS), and later at the Computer Science Department (CS) after the former group moved and joined together with the Statistical Machine Learning and Bioinformatics group to form the new Probabilistic Machine Learning (PML) group. I am grateful and want to acknowledge BECS Graduate School for partially funding my studies and making this research possible. I also wish to thank Prof. Jim Griffin and Prof. James G. Scott for pre-examining this thesis, and Dr. José Miguel Hernández-Lobato for agreeing to serve as an opponent.

Above all, I wish to thank my supervisor and instructor Prof. Aki Vehtari for the support and guidance during these years. I still remember the time when I was unsure whether to start the doctoral studies in the first place, and without your encouragement at the beginning, this thesis might never have come into existence. You have taught me many things about Bayesian statistical inference, but also about research and science in general. In addition to giving important instruction, I am very grateful that you have always given me plenty of freedom to explore my own ideas, and many of the papers in this thesis were born (at least partly) as a consequence of these explorations. I also appreciate the fact that the extra duties you have given me have always been very moderate, and I have always been able to focus on the research without too much disturbance.

I am also grateful to many colleagues whom I have worked with during these years and many of whom have provided me advice related to research and all sorts of things one might encounter in everyday work. In particular, I would like to thank (in a rough order I have gotten to know you) Prof. Arno Solin, Dr. Tomi Peltola, Janne Ojanen, Ville Tolvanen, Dr. Juho Kokkala, Olli-Pekka Koistinen, Tuomas Sivula, Dr. Jarno Lintusaari, Eero Siivola, Marko Järvenpää, Markus Paasiniemi, Topi Paananen, Akash Dhaka, Kunal Ghosh, Dr. Michael Riis Andersen, Gabriel Riutort Mayol, Dr. Måns Magnusson and Federico Pavone. I have enjoyed the numerous chats we have had; those which have been quite intellectual and taught me a lot, but also those perhaps not so intellectual ones

but which have been very entertaining nevertheless. I wish to thank also the other people in the PML group and the former Bayes group for a relaxed and enjoyable atmosphere. Special thanks go to the IT people at the BECS and CS departments, who have helped me in numerous computer related issues during the past few years.

I would also like to express my sincere gratitude to my family and in particular to my parents who have always supported and encouraged me in those endeavours I have decided to commit myself to. Last but definitely not least, thank you, Aura, for your support during these years.

Helsinki, April 29, 2019,

Juho Piironen

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 Juho Piironen and Aki Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017.

II Juho Piironen and Aki Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of Proceedings of Machine Learning Research, pages 905–913. PMLR. Fort Lauderdale, Florida, USA, 2017.

III Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

IV Juho Piironen and Aki Vehtari. Iterative supervised principal components. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of Proceedings of Machine Learning Research, pages 106–114. PMLR. Lanzarote, Spain, 2018.

V Juho Piironen, Markus Paasiniemi and Aki Vehtari. Projective inference in high-dimensional problems: prediction and feature selection. *Submitted*, 2018.

VI Juho Piironen and Aki Vehtari. Projection predictive model selection for Gaussian processes. In *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Salerno, Italy, 2016.

Author's Contribution

Publication I: "Comparison of Bayesian predictive methods for model selection"

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

Publication II: "On the hyperprior choice for the global shrinkage parameter in the horseshoe prior"

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

Publication III: "Sparsity information and regularization in the horseshoe and other shrinkage priors"

Both authors contributed in designing the content of the paper. The new methodological innovations are due to Piironen who also derived the new theoretical results, carried out the experiments and had the main responsibility of writing the article. Vehtari reviewed and proposed some additions to the manuscript.

Publication IV: "Iterative supervised principal components"

The topic was proposed by Piironen who also derived the new method and carried out the experiments. Piironen also had the main responsibility of writing

the article while Vehtari reviewed and proposed some modifications to the manuscript.

Publication V: "Projective inference in high-dimensional problems: prediction and feature selection"

Piironen and Vehtari both contributed in designing the topic of the study, but Piironen designed the content of the paper, derived the new methods and theoretical results, and carried out all the experiments. Piironen and Paasiniemi had about equal contributions in writing the software package. Piironen had the main responsibility of writing the article while Vehtari reviewed and proposed some additions to the manuscript.

Publication VI: "Projection predictive model selection for Gaussian processes"

The topic was proposed by Vehtari who also provided some ideas, but Piironen did most of the work in deriving the new method. Piironen also implemented the method, carried out all the experiments and had the main responsibility of writing the article. Vehtari reviewed and provided several small additions to the manuscript.

1. Introduction

This thesis deals mainly with Bayesian generalized linear models in setups where the data are scarce and the dimensionality of the feature space high—a regime which in the statistical jargon is often referred to as "small n, large d".[1] Linear models are often adopted as a default tool for these problems due to their interpretabiliy and ease of analysis but also because of their computational efficiency (at least relative to some of the more complex models). Another reason is statistical: often with very high-dimensional feature space and scarce data it can be difficult to learn nonlinear functions without overfitting, and in many cases linear models tend to be rich enough. Furthermore, some statistical relationships in the real-world mechanisms from which these datasets are collected can inherently be approximately linear or at least monotonic. A typical example could be the gene expression datasets where expressions of certain genes are either high or low for cancer samples and vice versa for controls—not for example so that *both* extremes (high and low) would be related to cancer and intermediate values to normal samples.

Due to the small sample sizes, these problems are often characterized by high uncertainties. Bayesian inference (e.g., O'Hagan and Forster, 2004; Gelman et al., 2013) provides a systematic framework for dealing with uncertainty using the rules of probability and by expressing the uncertainties using probability distributions. Accounting for uncertainty in the model parameters can result in better calibrated uncertainties also in predictions in comparison to using only point estimates for the parameters. Another benefit of Bayesian modeling is that it allows a natural way of incorporating prior information into the model. This can be useful, for example, for expressing that some parameter values are unlikely a priori which may improve the performance of the model. All this comes with a price, though; fully Bayesian inference can be computationally intensive. To alleviate this computational cost, one of the goals in this thesis is to study and develop techniques that can scale to large number of features.

The thesis is concerned with a setup which in machine learning is known as supervised learning: given a set of observations $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ infer the statisti-

[1] In the classical literature the number of features is often denoted by p, but we reserve this symbol for the probability density functions.

cal relationship between the features $\mathbf{x} = (x_1, \ldots, x_d)$ and the target variable y. The focus will be on two important problems. The first one is prediction: build a model which given new feature values $\tilde{\mathbf{x}}$ can predict the associated target variable $\tilde{y}$ as accurately as possible. In Bayesian formalism, this means learning a conditional probability distribution $p(\tilde{y} \mid \tilde{\mathbf{x}}, \mathcal{D})$. The second problem is feature selection: in many cases not all of the features are likely to play a crucial role in making the predictions, and we might want to identify a small subset of the features that can characterize the predictions. Feature selection can potentially have many benefits and we shall discuss these in some more detail in Chapter 3.

The most common approach of handling these two problems is to formulate an estimation procedure that performs (semi-)automatic feature selection at model fitting time. This means that both problems are attempted to be solved simultaneously, so that the feature selection is obtained as a "by-product" of the predictive model construction. One of the goals of this thesis is to challenge this traditional approach. We argue that in many cases one can gain if the two problems are solved in two stages using a decision theoretically justified approach: first construct a model that gives as good predictions as possible (not caring about feature selection), and then find a small subset of features that can characterize the predictions. As we shall see later on, perhaps rather surprisingly, this approach can both be computationally efficient and at the same time improve feature selection without sacrificing predictive accuracy.

The thesis consists of six publications and this introductory part. The contributions are solely methodological and covered in the original publications which can be found at the end of the thesis. The role of this introductory part is to provide a brief recap on the essential statistical methodology and summarize the overall philosophy behind the aforementioned two-stage approach for prediction and feature selection.

The remainder of this introductory part is structured as follows. Chapter 2 shortly reviews the used models. The focus is on Bayesian generalized linear models and their prior specification, but also Gaussian processes which are used in Publication VI are briefly discussed. Chapter 3 discusses the peculiarities encountered in problems with high-dimensional feature spaces. This chapter discusses the predictive inference and feature selection, and introduces the idea of the two-stage inference. Finally, the contributions of the thesis are briefly summarized in Chapter 4 followed by some concluding remarks in Chapter 5.

2. Bayesian linear models

This chapter briefly reviews the essential parts of Bayesian linear models that are relevant for the six publications. Section 2.1 discusses linear regression and Section 2.2 the generalized linear models which are used in Publications I–V. The prior choices are reviewed in Section 2.3 which also summarizes the methodological innovations of Publications II and III. Finally, Section 2.5 provides a brief introduction to Gaussian processes which are used in Publication VI.

2.1 Linear regression

Linear regression is one of the cornerstones of statistical analysis. In the basic setup, the goal is to model a real-valued target (or outcome) variable $y \in \mathbb{R}$ with features $\mathbf{x} = (x_1, \ldots, x_d)$ by assuming that the expected value of y given $\mathbf{x}$ is given by a linear combination of the features (often also referred to as covariates or predictors). With the customary assumption of normally distributed errors, the model can be written as

$$y_i = \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right), \quad i = 1, \ldots, n, \tag{2.1}$$

where i denotes the observation index and $\boldsymbol{\beta} \in \mathbb{R}^d$ the regression coefficients that determine how strongly each of the features is weighted in explaining variation of the target.[1] The other sources of variation not captured by the features $\mathbf{x}$ are modeled by the error terms ε_i that are assumed to be i.i.d. zero mean Gaussian random numbers with variance σ^2.

Assume we are given n measurements $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ and $\mathbf{X} = (\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_n^\mathsf{T}) \in \mathbb{R}^{n \times d}$. We can write (2.1) in a matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$. It is easy to show that the maximum likelihood estimates

[1] Typically an intercept term β_0 is added on the right hand side of model (2.1), which is equivalent to having an additional constant predictor $x_0 = 1$. We drop the intercept here for simplicity.

for the regression coefficients are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}, \tag{2.3}$$

provided that $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ exists, which is usually true when $d < n$.

Let us now assume a zero mean Gaussian prior for the regression coefficients

$$\boldsymbol{\beta} \,|\, \boldsymbol{\Lambda} \sim \mathrm{N}(0, \boldsymbol{\Lambda}), \tag{2.4}$$

where the covariance $\boldsymbol{\Lambda}$ is assumed to be given for now. With this choice, it is straightforward to show that the posterior for $\boldsymbol{\beta}$ given σ^2 and the data $\mathscr{D} = \{\mathbf{X}, \mathbf{y}\}$ is a Gaussian

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{\Lambda}, \sigma^2, \mathscr{D}) = \mathrm{N}\big(\boldsymbol{\beta} \,|\, \bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}\big), \tag{2.5}$$

where the mean and covariance are given by

$$\bar{\boldsymbol{\beta}} = \frac{1}{\sigma^2}\left(\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}, \tag{2.6}$$

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}. \tag{2.7}$$

It is straightforward to show that also the predictive distribution will be a Gaussian (see, e.g., O'Hagan and Forster, 2004, ch. 9). In case the noise variance σ^2 is also unknown (as it usually is), both the posterior for the regression coefficients and the predictive distribution become t-distributions (see O'Hagan and Forster, 2004, for more details).

A common choice is to use a diagonal prior covariance with a common variance, $\boldsymbol{\Lambda} = \tau^2\mathbf{I}$, which is referred to as (Bayesian) *ridge regression*. The prior variance τ^2 can be fixed to some constant value but a more flexible and adaptive approach is to give it a weakly informative prior and infer it from data along with other parameters (see Section 2.4 for more details about the inference). If we let the prior to approach uniform distribution $\tau \to \infty$, the posterior mean (2.6) will approach the maximum likelihood solution (2.3) as expected. Ridge regression is a reasonable choice when most of the features are assumed to have a regression coefficient β_j clearly distinguished from zero, but a variety of choices for $\boldsymbol{\Lambda}$ can be considered (see Section 2.3).

2.2 Generalized linear models

As the name suggests, generalized linear models (GLMs) (McCullagh and Nelder, 1989) generalize the setup discussed in Section 2.1. The limitation of model (2.1) is that it seems unreasonable for other than a real-valued, such as a discrete or positively constrained outcome. The GLM approach is to force the real-valued latent variable $f = \boldsymbol{\beta}^\mathsf{T}\mathbf{x}$ through a function that maps it to the target domain

and let this denote the expected value of y in an appropriate observation model. More specifically, we have

$$\mu = \mathrm{E}(y \mid x) = g^{-1}(f), \quad \text{or equivalently} \quad f = g(\mu),$$

where g is a monotonic function called the *link function* and g^{-1} its inverse, also known as the *response function*.

A common and important example of GLMs is the logistic regression model where the target value is either a Bernoulli distributed binary variable $y \in \{0, 1\}$ or a binomial distributed non-negative integer $y \in \{0, 1, 2, \ldots\}$. As the former is a special case of the latter, we can write the logistic regression model as

$$y_i \mid \mathbf{x}_i \sim \mathrm{Bin}\big(n_i, \mu_i\big), \quad \mu_i = \frac{1}{1 + \exp(-\boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i)}, \tag{2.8}$$

where n_i is the number of trials ($n_i = 1$ denoting the Bernoulli case) at feature values $\mathbf{x}_i$. The "success" probability $\mu \in (0, 1)$ for a given trial is given by the logistic response function $\mu = \frac{1}{1+\exp(-f)}$ that maps $f = \boldsymbol{\beta}^\mathsf{T} \mathbf{x} \in \mathbb{R}$ to the interval $(0, 1)$. Some common alternatives to the logistic response function are the probit and cauchit functions, which are the cumulative density functions of standard Gaussian and Cauchy distributions, respectively.

Other common examples of GLMs include the Poisson regression (e.g., Gelman et al., 2013, ch. 16) and survival models (e.g., Ibrahim et al., 2001) but we shall not discuss them further. All these observation models and link functions widen the applicability of the linear model but the downside is that due to the non-Gaussian likelihood, the posterior inference is no longer analytically available and some approximate inference technique must be adopted (see Section 2.4).

2.3 Prior choices

This section reviews some of the most common prior choices for the regression coefficients in GLMs. All priors discussed here can be formulated as scale mixtures of Gaussians (Section 2.3.1) which allows for convenient theoretical analysis. The priors imposed on the effective model complexity for different hyperprior choices are briefly discussed in Section 2.3.2.

2.3.1 Gaussian scale mixture priors

Many common priors can be obtained by placing a hyperprior on the prior covariance $\boldsymbol{\Lambda}$ in (2.4). By employing a few simplifying assumptions, one can also gain insights about how different choices affect the resulting posterior fit. Here we follow the analysis presented in Publications II and III.

Consider the posterior distribution for linear regression coefficients, given by Equation (2.5) (see Publication II or III for discussion about non-Gaussian likelihoods). Assuming the inverse $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ exists, with a little bit of manipulation

Table 2.1. Example prior distributions for the regression coefficients β_j that can be expressed as scale mixtures of Gaussians. The middle column gives the conditional prior for β_j given the hyperparameters, and the last column gives the hyperprior. All hyperparameters for which prior is not specified (τ, ν, π and c) are assumed to be given, although in practice these can be given hyperpriors as well. Symbol c is purposely used both in regularized horseshoe and spike-and-slab as it serves for the same purpose in both cases. For the inverse-gamma distribution, parameters a and b denote the shape and scale, respectively, and also for the exponential distribution, b denotes the scale.

Name	Prior	Hyperprior
Gaussian	$\mathrm{N}\left(0, \tau^2\lambda_j^2\right)$	$\lambda_j = 1$
Student-t_ν	”	$\lambda_j^2 \sim \text{Inv-Gamma}\left(a = b = \frac{\nu}{2}\right)$
Laplace	”	$\lambda_j^2 \sim \text{Exp}(b = 2)$
Horseshoe	”	$\lambda_j \sim \mathrm{C}^+(0,1)$
Regularized horseshoe	$\mathrm{N}\left(0, \tau^2\tilde{\xi}_j^2\right)$	$\tilde{\xi}_j^2 = \frac{c^2\lambda_j^2}{c^2+\tau^2\lambda_j^2}$, $\lambda_j \sim \mathrm{C}^+(0,1)$
Spike-and-slab	$\mathrm{N}\left(0, c^2\lambda_j^2\right)$	$\lambda_j \sim \text{Ber}(\pi)$

the posterior mean (2.6) can be rewritten as

$$\bar{\boldsymbol{\beta}} = \boldsymbol{\Lambda}\left(\boldsymbol{\Lambda} + \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\right)^{-1}\hat{\boldsymbol{\beta}}, \tag{2.9}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum-likelihood solution (2.3). Now, assuming further that the features $\mathbf{x}$ are uncorrelated with zero mean and unit variance, then $\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} \to \mathbf{I}$ as $n \to \infty$ (see Publication III for discussion on unequal feature variances). Using approximation $\mathbf{X}^\mathsf{T}\mathbf{X} \approx n\mathbf{I}$, with a diagonal prior covariance $\boldsymbol{\Lambda} = \tau^2\text{diag}\left(\lambda_1^2, \ldots, \lambda_d^2\right)$ the elements of $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_d)$ have a simple form

$$\bar{\beta}_j = (1 - \kappa_j)\hat{\beta}_j, \tag{2.10}$$

where

$$\kappa_j = \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2}. \tag{2.11}$$

We call the terms $\kappa_j \in (0,1)$ the *shrinkage factors* (or coefficients) following the terminology of Carvalho et al. (2009, 2010). Shrinkage factors describe how much the posterior mean is shrunk towards zero from the maximum likelihood solution. In particular, at one extreme we have $\kappa_j \to 1$ and $\bar{\beta}_j \to 0$, and on the other hand when $\kappa_j \to 0$ then $\bar{\beta}_j \to \hat{\beta}_j$. Allowing different features to have different local hyperparameter λ_j allows them also to have different shrinkage factors.

The shrinkage factors are a useful concept since they provide a tool for understanding the behaviour of some commonly used priors as well as a means for designing new ones. Since the shrinkage factors are determined by the hyperparameters λ_j and τ together with the noise variance σ^2, different (hyperprior) choices for λ_j and τ can be understood based on their effect on the *shrinkage profile*, that is, the prior imposed on κ_j.

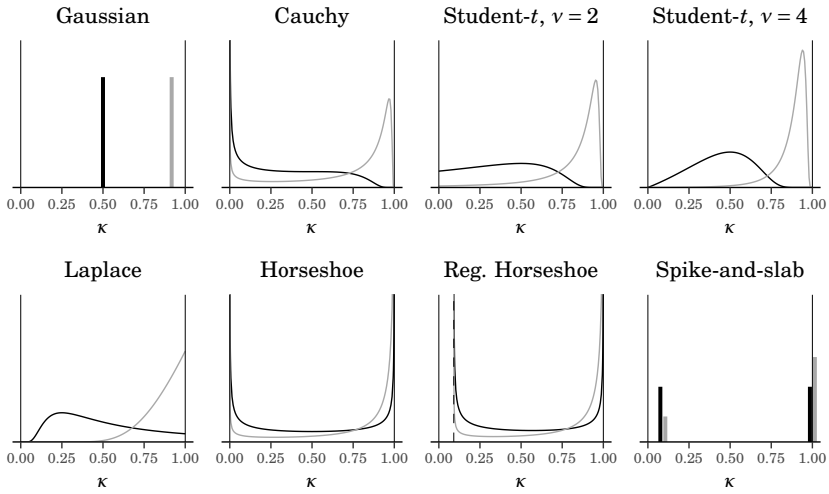

Figure 2.1. Priors densities imposed on the shrinkage factor (2.11) for different prior choices $p(\beta_j)$ (see Table 2.1). For Gaussian and spike-and-slab, the prior contains mass only at some discrete values depicted by the thick vertical bars. For all priors except spike-and-slab, black denotes the density when $\sqrt{n}\sigma^{-1}\tau = 1$ and grey denotes $\sqrt{n}\sigma^{-1}\tau = 0.3$. For spike-and-slab, black and grey denote cases $\frac{\pi}{1-\pi} = 1$ and $\frac{\pi}{1-\pi} = 0.3$, respectively (the bar locations are the same in both cases but are drawn here with a small horizontal shift to avoid overlap). For the regularized horseshoe and spike-and-slab, the left mode is located at $\kappa = \frac{1}{1+n\sigma^{-2}c^2}$, and for visualization we have selected slab scale $c = 1$ and $n\sigma^{-2} = 10$.

Table 2.1 lists some example priors that are Gaussian with diagonal covariance $\mathbf{\Lambda}$ when the hyperparameters are given. The corresponding hyperpriors are given in the last column. The imposed priors on a single shrinkage coefficient κ_j are shown in Figure 2.1. Gaussian prior (ridge regression) fixes the local variance parameters λ_j^2 to unity for each feature which results in a constant shrinkage for each coefficient β_j, and the magnitude of the shrinkage depends on the value for the global hyperparameter τ. Allowing the local hyperparameters λ_j to vary leads to a more flexible prior that allows the regression coefficients to adapt better to the observed data. In particular, the horseshoe prior (Carvalho et al., 2009, 2010; Polson and Scott, 2011) favors values both close to $\kappa_j = 0$ and $\kappa_j = 1$ which is useful for encoding prior information that some coefficients β_j are likely to be large and some close to zero. While both the Student-t family (Tipping, 2001; Gelman et al., 2008) and the Laplace prior (Park and Casella, 2008) can accommodate a wide range of values for κ_j, neither of them encourages both large and small values simultaneously.[2]

The horseshoe prior originally got its inspiration from the popular spike-and-

[2]It should be noted, though, that the relevance vector machine (RVM) of Tipping (2001) employs improper Student-t priors with $\nu = 0$ which *does* induce truly sparse solutions when the hyperparameters are optimized to the marginal maximum a posterior solution. This is because the hyperprior then becomes $p(\lambda_j) \propto \frac{1}{\lambda_j}$ which does encourage some of the λ_j to collapse to zero.

slab prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1993) that—in the form presented in Table 2.1—allows only two discrete values, $\kappa_j = 1$ or $\kappa_j = \frac{1}{1+n\sigma^{-2}c^2}$, where c denotes the slab scale. In other words, each coefficient β_j is either set exactly to zero ("spike") or given a Gaussian prior ("slab") with variance c^2. Several variants of the spike-and-slab prior have been proposed. For example, instead of a delta spike, one can use a distribution with small but nonzero variance, place a hyperprior on c to obtain a more heavy-tailed slab, and consider different slab widths c_j for each feature (George and McCulloch, 1993, 1997; Johnstone and Silverman, 2004; Ishwaran and Rao, 2005; Peltola et al., 2012). When the spike is taken to be a delta spike at the origin $\beta_j = 0$, integrating over the posterior uncertainty about the regression coefficients corresponds to Bayesian model averaging (BMA) where each feature combination is considered as a separate model (Raftery et al., 1997; Hoeting et al., 1999).

The original horseshoe and the spike-and-slab are not fully analogous, however. The difference is that while in spike-and-slab even the largest β_j will experience regularization by a Gaussian slab with scale c (that is, $\kappa_j = \frac{1}{1+n\sigma^{-2}c^2}$), the horseshoe encourages zero regularization (that is, $\kappa_j = 0$), see Figure 2.1. The regularized horseshoe introduced in Publication III bridges this gap by introducing a modified prior

$$\beta_j \,|\, \lambda_j, \tau, c \sim \mathrm{N}\big(0, \tau^2 \tilde{\xi}_j^2\big), \quad \tilde{\xi}_j^2 = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2}, \quad \lambda_j \sim \mathrm{C}^+(0,1). \tag{2.12}$$

Here the local parameters λ_j are given half-Cauchy priors as in the original horseshoe, but they enter the prior for β_j through the transformation $\tilde{\xi}_j$ which introduces a slab scale parameter c as in spike-and-slab. The idea is quite simple. For those β_j for which λ_j is small (that is, $\tau^2\lambda_j^2 \ll c^2$) we have $\tilde{\xi}_j^2 \approx \lambda_j^2$ and the prior is approximately the same as the original horseshoe (see Table 2.1). However, for those β_j for which λ_j is large (that is, $\tau^2\lambda_j^2 \gg c^2$) we have $\tilde{\xi}_j^2 \approx \frac{c^2}{\tau^2}$ and the prior for β_j approaches $\mathrm{N}\big(0, c^2\big)$, that is, a Gaussian slab. Figure 2.1 confirms that the new prior indeed mimics the shrinkage profile of the spike-and-slab with a finite slab width. Setting $c \to \infty$, we recover the original horseshoe which resembles the spike-and-slab with an infinitely wide slab.

Having a way to control the regularization for the largest regression coefficients can be useful with weakly identified parameters. An example are the logistic regression coefficients when the classes are perfectly separable because the likelihood becomes then flat. It is known that in such a situation the posterior moments may vanish for independent Cauchy-priors (Ghosh et al., 2018), and since also horseshoe has Cauchy-tails, it is vulnerable to the same phenomenon. Even if vanishing moments were not an issue, the regularized horseshoe is often empirically observed to robustify and speed-up the inference (see Publication III).

Similar extensions as for the spike-and-slab can be applied also to the regularized horseshoe. For example, instead of fixing c it can be given a hyperprior to allow a more flexible model. A reasonable hyperprior recommended in Publication III is $c^2 \sim \text{Inv-Gamma}\left(a = \frac{\nu}{2}, b = \frac{\nu s^2}{2}\right)$ which results in a Student-t slab

with scale s and ν degrees of freedom.

As a final remark, it should be noted that here we have discussed only some of the most commonly used and fairly well established priors. Indeed, several priors have been proposed many of which can be represented as scale mixtures of Gaussians as those mentioned above. These include the normal-gamma (Griffin and Brown, 2010), Bayesian hyper-Lasso (Griffin and Brown, 2011), three parameter beta normal scale mixture (Armagan et al., 2011), generalized double Pareto (Armagan et al., 2013), Dirichlet-Laplace (Bhattacharya et al., 2015), horseshoe+ (Bhadra et al., 2017) and R2-D2 (Zhang et al., 2017). Some empirical comparisons indicate that some of these perform quite similarly (Zhang et al., 2017; Tang et al., 2018) but much more research would be needed in order to get a good idea of the differences between all these priors.

2.3.2 Role of the global shrinkage

Figure 2.1 illustrates that the shape of the shrinkage profile changes when the value of the global parameter τ is changed (in spike-and-slab this role is played by the prior inclusion probability π). In particular, by decreasing the value of τ (or reducing π in spike-and-slab) one can place more mass near $\kappa_j = 1$ and encourage therefore more shrinkage. Given that τ has a notable effect on the shrinkage profile, how should one then decide the value or a hyperprior for it?

To address this issue, Publications II and III introduce the concept of *effective number of nonzero coefficients* which is defined as

$$m_{\text{eff}} = \sum_{j=1}^{d} (1 - \kappa_j). \tag{2.13}$$

This quantity measures the effective complexity of the model. In other words, those regression coefficients that are penalized very little (that is, $\kappa_j = 0$) contribute one to the sum, and those that are shrunk heavily towards zero (that is, $\kappa_j = 1$) contribute nothing. By studying the imposed prior on m_{eff} one can get an idea about how different prior choices for τ affect the effective model complexity.

Figure 2.2 illustrates this idea. The subplots show the histograms of prior draws for m_{eff} with two different choices for τ and π in the horseshoe and spike-and-slab priors, respectively. Decreasing the value of τ or π favors models with smaller effective complexity. For both priors, fixing the sparsity hyperparameter (τ or π) leads to a prior which is fairly informative about m_{eff}. A more flexible choice is to specify a hyperprior which leads to a less informative prior for m_{eff} (see Publications II and III for an illustration on this point for the horseshoe).

The results shown in Figure 2.2 are generated simply by drawing the hyperparameters $\lambda_1, \ldots, \lambda_d$ from their priors and then computing the shrinkage factors (2.11) and finally the effective number of nonzero coefficients (2.13). Clearly for the spike-and-slab this distribution is easily characterized also analytically since for a given π, m_{eff} is binomial distributed with success probability π. Publications II and III show that, even though the analytic form of the prior for m_{eff}

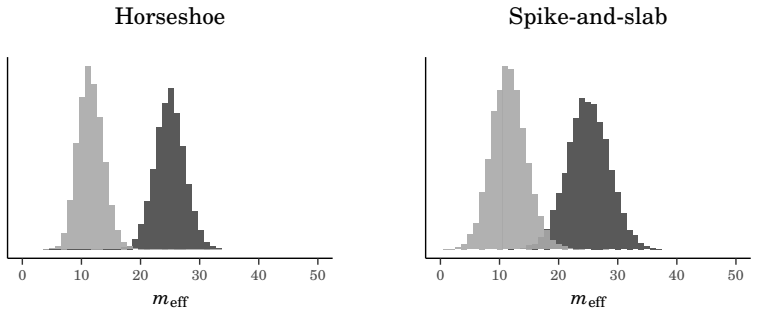

Figure 2.2. Illustration of the imposed priors on the model complexity for different choices of sparsity hyperparameters when $p = 50$. Left graph shows the histograms of prior draws for m_{eff} (Eq. (2.13)) for the horseshoe prior with $\sqrt{n}\sigma^{-1}\tau = 1$ (black) and with $\sqrt{n}\sigma^{-1}\tau = 0.3$ (gray). The right graph shows the same for the spike-and-slab (infinite slab width) with $\frac{\pi}{1-\pi} = 1$ (black) and $\frac{\pi}{1-\pi} = 0.3$ (gray). Notice that with spike-and-slab, m_{eff} obtains only integer values whereas with horseshoe it is real-valued.

is intractable for the horseshoe, the mean and variance of this distribution are analytically available for a given τ. This framework can also be used to design weakly informative default hyperpriors for τ based on the prior information about the sparsity (see Publications II and III for the procedure).

2.4 Inference

As discussed in Section 2.1, for the linear regression model (2.1) the posterior inference for $\boldsymbol{\beta}$ and σ^2 is analytically available when the prior covariance $\boldsymbol{\Lambda}$ is given, and when the noise variance is also given, the posterior has a simple Gaussian form. However, when the prior covariance $\boldsymbol{\Lambda}$ also has unknown hyperparameters, the posterior inference for these is no longer analytically possible. Still, as the marginal likelihood for a given $\boldsymbol{\Lambda}$ can be computed analytically, for simple prior covariance choices such as $\boldsymbol{\Lambda} = \tau^2\mathbf{I}$ the integration over τ can easily be approximated using numerical quadratures. This strategy is used in Publication I.

When $\boldsymbol{\Lambda}$ has many unknown hyperparameters (Section 2.3.1) or when the observation model is non-Gaussian (Section 2.2), one has to resort to more sophisticated inference algorithms. The most generic choice is to use Markov chain Monte Carlo (MCMC) algorithms (e.g., Robert and Casella, 2004) which can handle both of these issues. The advent of modern generic sampling tools such as Stan (Stan Development Team, 2018) has made MCMC inference efficient and easily available for a very wide class of models. Stan implements static Hamiltonian Monte Carlo (HMC) (e.g., Neal, 2011) and dynamic HMC (Hoffman and Gelman, 2014; Betancourt, 2017) which provide efficient inference in many cases even for high-dimensional parameter spaces. Stan is used for inference in Publications II–V.

The big advantage in MCMC is its generality, but this can come at the price of a large computation time. The analytical approximations based on Laplace approximation (e.g., Gelman et al., 2013, ch. 4), expectation propagation (EP) (Minka, 2001) or variational inference (e.g., Jordan et al., 1999; Bishop, 2006; Blei et al., 2017) can in many cases provide faster alternatives, but especially for EP and VI this typically means a substantial increase in the amount of analytical work and time required for the implementation. The automated variational inference algorithms that require minimal input from the user (Ranganath et al., 2014; Kucukelbir et al., 2017) hold promises, but these techniques are still too often either too inaccurate or fragile in order to really compete with MCMC as reliable black box inference algorithms.

2.5 Gaussian processes

Gaussian processes (GPs) (Rasmussen and Williams, 2006) are a rich and flexible class of models which contain GLMs as a special case. For this reason, they would deserve a chapter of their own but because of their limited role in this thesis (Publication VI) we shall discuss them only briefly here.

As were discussed in Section 2.2, in GLMs the expected value of the target variable y is obtained by transforming the latent function value f—which is obtained as a linear combination of the features $f = f(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{x}$—through the response function g^{-1}. Giving a prior distribution on the regression coefficients $\boldsymbol{\beta}$ induces a prior distribution on the latent function f. For instance, if $\boldsymbol{\beta} \sim \mathrm{N}(0, \boldsymbol{\Lambda})$, then the latent function values $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) = \mathbf{X}\boldsymbol{\beta}$ in an arbitrary collection of feature values $\mathbf{X} = (\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_n^\mathsf{T})$ will have a joint Gaussian distribution

$$\mathbf{f} \sim \mathrm{N}\big(0, \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^\mathsf{T}\big). \tag{2.14}$$

This is an example of a Gaussian process; a collection of random variables, any finite subset of which have a joint Gaussian distribution.

Instead of using a parametric model for f and then placing a prior on its parameters, the core idea of GP models is to place the prior directly on the latent function. That is, we assume that all the function values have a joint Gaussian distribution for which we specify mean and covariance functions, $m(\mathbf{x}) = \mathrm{E}(f(\mathbf{x}))$ and $k(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}\big(f(\mathbf{x}), f(\mathbf{x}')\big)$, that encode our prior assumptions about the function. In the previous GLM example, the mean function is simply $m(\mathbf{x}) = 0$ and the covariance (or kernel) between any two points $\mathbf{x}$ and $\mathbf{x}'$ is given by $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\mathsf{T}\boldsymbol{\Lambda}\mathbf{x}'$. Another common covariance function (and the one used in Publication VI) is the squared exponential (or exponentiated quadratic, or Gaussian) $k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\sum_{j=1}^d \frac{(x_j - x_j')^2}{\ell_j^2}\right)$ which produces smooth nonlinear functions. This essentially says that the function values in nearby points have a high covariance and the covariance decays to zero when the two points are far from each other. Here σ_f^2 and $\{\ell_j\}_{j=1}^d$ are hyperparameters that describe the overall magnitude of variation in the function values and how fast the covariance

decays in different input directions.

The actual "parameters" of a GP model are the latent function values $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$ for which the inference is in principle quite straightforward. Assuming a standard zero mean GP, the prior is Gaussian $\mathbf{f} \sim \mathrm{N}(0, \mathbf{K})$, where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and given an observation model (likelihood) $\mathbf{y} \sim p(\mathbf{y}|\mathbf{f})$, we can combine these to get the posterior distribution. If the observation model is Gaussian $\mathbf{y}|\mathbf{f} \sim \mathrm{N}(\mathbf{f}, \sigma^2\mathbf{I})$, the posterior for $\mathbf{f}$ given σ^2 remains Gaussian. It is then easy to show that the predictive distribution for the latent value at a given test point $\tilde{\mathbf{x}}$ will also be a Gaussian with mean and variance given by

$$\mathrm{E}(f(\tilde{\mathbf{x}})|\mathscr{D}) = \tilde{\mathbf{k}}^\mathsf{T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \tag{2.15}$$

$$\mathrm{Var}(f(\tilde{\mathbf{x}})|\mathscr{D}) = k(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}) + \tilde{\mathbf{k}}^\mathsf{T}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\tilde{\mathbf{k}}, \tag{2.16}$$

where $\tilde{\mathbf{k}} = (k(\tilde{\mathbf{x}}, \mathbf{x}_1), \ldots, k(\tilde{\mathbf{x}}, \mathbf{x}_n))$. The hyperparameters (noise variance and kernel parameters) are typically unknown in practice but since the marginal likelihood given the hyperparameters is analytically available (see Rasmussen and Williams, 2006, for details), these parameters can be estimated from the data. The most common strategy is to optimize them to the maximum marginal likelihood solution, but an alternative and more Bayesian approach is to integrate over them by using some deterministic or Monte Carlo based algorithm (e.g., Rue et al., 2009; Vanhatalo et al., 2010).

GPs provide an elegant and flexible way of encoding the prior assumptions of the underlying function into the model. The main drawback, however, is the computational cost; in a general case the exact inference scales cubicly $O(n^3)$ with the number of data points n due to the matrix inversion which becomes quickly prohibitive. Another complication is that the inference is analytically intractable for non-Gaussian likelihoods. Consequently, much of the GP literature has been focusing on making the inference feasible for large n and non-Gaussian likelihoods (e.g., Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006; Titsias, 2009; Hensman et al., 2013, 2015; Matthews et al., 2016; Hernández-Lobato and Hernández-Lobato, 2016; Salimbeni and Deisenroth, 2017; Hensman et al., 2018). Since this thesis focuses on problems with small n, we shall not discuss these techniques further. For a recent useful overview of many of these approaches, see Bui et al. (2017) and references therein.

3. High-dimensional problems

Inference and analysis for models discussed in Chapter 2 is typically straightforward when the model contains only a few predictors. However, problems with a large number of features—especially those where the number of features exceeds the number of observations—pose extra challenges. These problems usually arise from data collection processes where several features are measured but so that it is likely that not all of them are statistically predictive about the target variable.

In these problems the goals of the statistical analysis are often twofold: one would like to construct a model that predicts unseen data well but also to identify which of the features are relevant for prediction. The latter problem is typically referred to as feature or variable selection. Although sometimes overlooked, it is important to distinguish between two different problems that could both be considered as feature selection:

1. Identify a *minimal subset* of features so that adding more will not substantially improve the predictive performance.

2. Identify *all* those features (or as many as possible) that are statistically related to the target variable.

In machine learning literature where prediction is typically the most important concern, most authors refer to the first problem when talking about feature selection (see, for example, Guyon and Elisseeff, 2003). However, the latter problem—often called *multiple (hypothesis) testing*—is a much studied and still actively pursued topic in the statistics literature (e.g., Johnstone and Silverman, 2004; Scott and Berger, 2006, 2010; Efron, 2010). There the prediction typically plays a much smaller role and the main interest is to study the statistical relationships between the features and the target in order to better understand the real world process that generated the data.

This distinction is useful as it turns out that very often with large number of features many of them carry similar information, that is, there is a lot of redundancy. A simple example is the case of two features x_1 and x_2 which are

strongly correlated with each other and about equally correlated with the target variable y. Since $x_1 \approx x_2$, consequently $\beta_1 x_1 + \beta_2 x_2 \approx (\beta_1 + \beta_2)x_1 \approx (\beta_1 + \beta_2)x_2$ which shows why it can be possible to reduce the number of features without seriously affecting the predictions although the left-out feature(s) could not be considered as irrelevant. In other words, if we were interested in solving Problem 1, we would select either x_1 or x_2, whereas in Problem 2, we would want to identify both x_1 and x_2. Due to the different nature of the two problems, it is natural to expect that a single approach cannot be ideal for solving both problems. It should be emphasized here that this thesis focuses solely on the first problem and this is also what is meant by "feature selection" in what follows.

An important point argued in this chapter is that feature selection may not be necessary for obtaining good predictions even when the number of features is very large. This point is clearly illustrated in the papers of this thesis, see for example Publications I and II. For very high-dimensional problems it might be necessary to use some shortcuts in order to reduce the computation time (especially for Bayesian methods), but in many cases computation time can be reduced by dimension reduction techniques that do not necessarily perform any feature selection or at least use a relatively large number of features (see Section 3.2).

We shall briefly review some of the most commonly used classical techniques for feature selection in Section 3.1. Section 3.2 then discusses approaches that are useful for predictive model construction but that do not perform feature selection in the sense that they would attempt to produce a very sparse model. Section 3.3 then reviews some Bayesian approaches and discusses the projective framework that can be used to simplify non-sparse models if a truly sparse model is desired.

3.1 Classical approaches

Classical approaches for sparse estimation typically formulate the feature selection as a maximum likelihood estimation problem with an additional penalty that enforces sparsity in the solution. Probably the most well-known such method is the Lasso (Tibshirani, 1996), which for the GLMs can be written as

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ -\log p(\mathbf{y} \mid \boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_1 \right\}. \tag{3.1}$$

In Lasso the sparsity stems from the L_1-penalty on the regression coefficient vector. Solving (3.1) for a number of values for the regularization parameter λ yields a sequence of models with different number of nonzero regression coefficients. An appropriate value for λ is then typically selected based on the estimated predictive performance using cross-validation. Lasso has several advantages that have made it extremely popular: the method is very simple, the optimization problem is convex facilitating efficient computation (Friedman et al., 2010), and in most problems it yields reasonably good results (in terms of

predictive accuracy) while performing automatic feature selection.

On the other hand, one of the drawbacks is that for large values of λ (that is, for the sparsest models) the L_1-penalty tends to overshrink the nonzero coefficients and produce substantial bias in the estimation. To reduce this excessive bias, one must reduce λ which can cause many extra features to enter the model. This phenomenon is well-known and illustrated also in Publication V. Another consideration is that estimation of additional parameters such as the noise variance σ^2 in regression is non-trivial (Reid et al., 2016).

Another well-known penalization is the elastic net of Zou and Hastie (2005) which can be considered as a bridge between Lasso and ridge regression. The elastic net for GLMs is given by

$$\hat{\boldsymbol{\beta}}_{\alpha,\lambda} = \arg\min_{\boldsymbol{\beta}} \left\{ -\log p(\mathbf{y} \mid \boldsymbol{\beta}) + \lambda \left(\frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1 \right) \right\}, \tag{3.2}$$

which introduces a new parameter $\alpha \in [0,1]$. Lasso is obtained when $\alpha = 1$ and ridge when $\alpha = 0$. Intermediate values $\alpha \in (0,1)$ yield models with more nonzero coefficients than in Lasso, but the benefit is that the strongly correlated predictors tend to get selected in groups. Smaller values of α can also lead to somewhat better predictive accuracy even for very high-dimensional problems if there are plenty of relevant features (see Publication V).

There are also several other extensions or otherwise closely related approaches (see Hastie et al., 2015, for a useful overview). These include the group Lasso (Yuan and Lin, 2006) which can be used to select features in groups. The nonnegative garrote (Breiman, 1995) and adaptive Lasso (Zou, 2006) are closely related to each other and to the Lasso, and the former was actually the inspiration to the original Lasso paper (Hastie et al., 2015). These techniques can undo to some extent the undesirable excessive shrinkage to the largest coefficients inherent for Lasso while maintaining the convexity of the optimization problem. Thus they can also recover the true model (assuming such exists) under more general conditions than does the Lasso (see Zou, 2006, for more details). There are also some well developed non-convex penalties that can overcome the bias due to the excessive shrinkage. These include the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and the minimax concave penalty (MC+) (Zhang, 2010). Due to nonconvexity, finding the globally optimal solutions is difficult but some efficient heuristics that can find good locally optimal solutions have been developed (for example, Mazumder et al., 2011)

3.2 Predictive inference

The classical sparsity enforcing methods discussed in Section 3.1 attempt to perform the predictive model construction and feature selection simultaneously. However, even for very high-dimensional datasets it is often possible to come up with a model with high predictive accuracy even without doing much or any feature selection. This point is often overlooked but especially with Bayesian

methods and reasonable prior choices (see Sec. 2.3), it is usually possible to avoid overfitting without any feature selection. This is illustrated for example in Publications II and III where good results were obtained with a logistic regression model with (regularized) horseshoe prior in some example microarray datasets with very scarce data, $n < 100$, but the number of features going up to about $d \approx 7000$, so that $n \ll d$. Perhaps surprisingly, as demonstrated in Publication V, even ridge regression which encourages all regression coefficients to be away from zero can yield very accurate predictions in some of these problems. Naturally it depends on the problem which approach performs best, but it is safe to say that there are very high-dimensional datasets where no feature selection is required to achieve accurate predictions.

This does not mean, however, that the recommended or the most practical way of constructing a predictive model would be to use all features as they are. With high-dimensional feature spaces, especially fully Bayesian inference using MCMC can be computationally costly even for simple models such as logistic regression. In order to reduce the computation time one might want to use some shortcuts. Still, the computation time can often be reduced with dimension reduction techniques that do only little if any feature selection.

This was the key idea behind Neal and Zhang (2006) who were the overall winners of the NIPS 2003 feature selection challenge where the goal was to construct a model that optimizes prediction accuracy on a test set for five datasets with large number of features (Guyon et al., 2006). As classifiers in their challenge submissions, Neal and Zhang used Bayesian neural networks and Dirichlet diffusion trees. To reduce the computation time to something that could be handled by fully Bayesian inference for these models, they used feature screening (also called filtering) based on univariate feature relevance assessment (such as correlations with the class label) together with dimension reduction using principal component analysis (PCA). Depending on the dataset and submission—each contestant was allowed several submissions—they used either only screening or PCA, or a combination of the two. This way they reduced the dimensionality to about a few hundred features (the exact number being selected based on the results on the validation data) and then used sparsity promoting priors—or automatic relevance determination (ARD) as they called it—that could further adapt to features with different relevances. Many other contestants who achieved good results in the challenge used very similar ideas, and rather remarkably, most good results were obtained with very simple techniques. Another interesting take-home message from the overall results was that eliminating all or even most of the irrelevant features was not critical for obtaining a good classification accuracy (Guyon et al., 2006).

Dimensionality reduction using PCA can be an effective way of cutting down the computations, and this still typically corresponds to using all the original features (that is, no feature selection) since the principal components are almost always non-sparse. However, if the features contain a lot of variation unrelated to the variation in the target variable y, the first principal components may not

be very predictive about y and large number of principal components might be required in order to capture all the relevant variation. In these cases a more effective approach might be combining feature screening and PCA, an approach that is called supervised PCA (SPCA) (Bair et al., 2006). This approach works as follows:

1. Compute the univariate relevance scores $r_j = r(x_j, y)$ for each feature x_j.

2. Select some screening threshold γ, and retain only those features that have their score above this value, that is $r_j > \gamma$, and compute principal components from the reduced feature matrix $\mathbf{X}_\gamma$.

The scoring function r is typically taken to be the absolute sample correlation between x_j and y, but also other choices could be considered. The screening threshold γ can be either selected using cross-validation for the model constructed using the extracted features, or one could simply discard features with the score not statistically significantly different from zero (see Publication IV). The screening step attempts to discard variables which are irrelevant for predicting y which usually causes the predictive power to be more heavily loaded on the first few components facilitating more effective dimension reduction for predictive model construction. A probabilistic version of the above idea has also been proposed (Yu et al., 2006) but we shall not discuss it here.

Publication IV proposes a modification to the original SPCA by introducing an iterative screening process that could possibly discover also features that are not necessarily relevant alone but become relevant after some other features are included (the original SPCA would miss these). As one might expect, based on the comparisons on several benchmark datasets, no single method appears to perform better than the others in all cases, and the optimal method is dataset dependent (Publication IV).

3.3 Feature selection

As argued in Section 3.2, aggressive feature selection may not be necessary for obtaining a good predictive model. Still, even if we had a model that predicts well, in many cases some form of feature selection is beneficial, since it can aid data understanding by making the model easier to explain and interpret. Feature selection can also make the model much more convenient and faster to use at prediction time, and it can also help reducing future costs if there is a price associated with predicting with many features. In the following two sections, we shall briefly review the traditional Bayesian approaches with some caveats (Section 3.3.1) and then introduce the projective framework (Section 3.3.2) that is argued to be superior.

3.3.1 Selection based on posterior information

Some of the most common classical feature selection techniques were discussed in Section 3.1. In the Bayesian literature, the dominant approach by far is to formulate a prior that favors sparse solutions for the regression coefficients, and the most common choice is undoubtedly the spike-and-slab (e.g., Lee et al., 2003; Zhou et al., 2004). Also many of the other continuous shrinkage priors discussed in Section 2.3 could be considered. Unlike the classical methods, these approaches do not automatically produce a truly sparse model, since regardless of the prior, there will always be a nonzero posterior probability for each feature being included in the model. Sparse models could be produced for example by thresholding, so that those features with estimated posterior effect below some threshold are removed (Barbieri and Berger, 2004; Ishwaran and Rao, 2005; Narisetty and He, 2014). In case of spike-and-slab prior, an alternative strategy is to select the maximum a posteriori (MAP) model (e.g., Johnson and Rossell, 2012) which utilizes Bayes factors (Kass and Raftery, 1995; Han and Carlin, 2001) together with prior probabilities for different feature combinations. Spike-and-slab prior has also been used in combination with L_1-penalty for penalized maximum likelihood estimation (Ročková and George, 2018).

Unfortunately, the approach of inferring a good feature combination directly based on the posterior for the regression coefficients has many difficulties which are discussed in detail in Publication V. Firstly, as discussed in Section 3.2, the posterior inference for a sparsifying prior (or any prior for that matter) with a large number of features can be a great computational challenge if MCMC is used for inference. Analytical approximations based on EP or VI have been proposed to speed up the computation but these require a substantial amount of analytical work and can be complex to implement (Hernández-Lobato et al., 2010, 2013, 2015; Titsias and Lázaro-Gredilla, 2011; Carbonetto and Stephens, 2012). Secondly, for spike-and-slab the relative marginal likelihoods of different feature combinations can be sensitive to the hyperprior or hyperparameter choices (e.g., Kass and Raftery, 1995). The sensitivity of the posterior distribution to the hyperprior on the global shrinkage parameter in the horseshoe prior has been demonstrated in Publications II and III.

Another serious issue is that pretty much regardless of the prior, the marginal posteriors for the regression coefficients can be challenging to interpret when some of the features are correlated. As discussed in the introduction of this chapter, the reason is that when two features x_1 and x_2 are highly correlated (that is $x_1 \approx x_2$) then $\beta_1 x_1 + \beta_2 x_2 \approx (\beta_1 + \beta_2)x_1 \approx (\beta_1 + \beta_2)x_2$. In practice this means that the likelihood is relatively flat in the direction where $\beta_1 + \beta_2$ is constant, and therefore it provides little information whether both or only one of these coefficients should be nonzero.

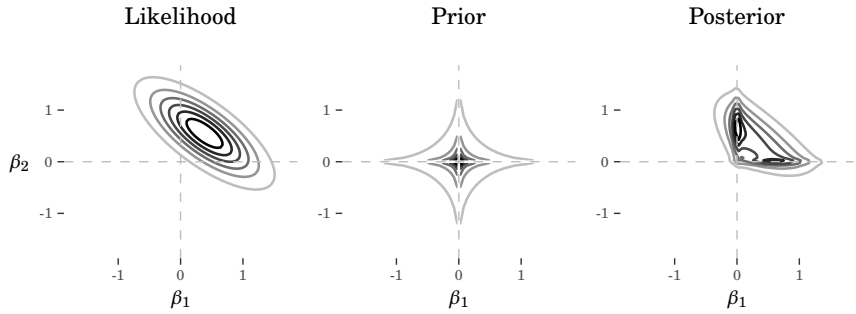

Figure 3.1. Illustration of a typical difficulty encountered with correlated features. The model is the simple linear regression (2.1) without intercept and assuming the noise variance σ^2 is known. Visualized are the likelihood, prior (horseshoe with $\tau = 1$) and posterior densities for the regression coefficients β_1 and β_2 for a random data realization with $n = 50$ observations when the features x_1 and x_2 have a correlation of $\rho = 0.8$ (see the text for more details). The likelihood for both coefficients being zero is small, but the data provides little evidence whether both or only one of them is nonzero. A sparsifying prior such as the horseshoe results in a multimodal posterior but does not help in solving the feature selection problem.

To illustrate this point, consider the following data generating mechanism:

$$\begin{aligned} f &\sim \mathrm{N}(0,1), \\ y \,|\, f &\sim \mathrm{N}\big(f, \sigma^2\big) \\ x_j \,|\, f &\sim \mathrm{N}\big(\sqrt{\rho} f, 1-\rho\big), \qquad j = 1,\ldots,d. \end{aligned} \tag{3.3}$$

The target variable values y are noisy observations from latent function values f which are drawn randomly from the standard Gaussian. The features x_j are also noisy observations from f which makes them correlated and on average equally predictive about y (each feature has unit variance and all pairwise feature correlations are equal to ρ). Figure 3.1 shows the likelihood, horseshoe prior and posterior densities for the regression coefficients with $d = 2$ features for a randomly generated dataset with $\rho = 0.8$, $\sigma^2 = 2^2$ and $n = 50$ observations (assuming the noise variance and the intercept $\beta_0 = 0$ are known for simplicity). The likelihood that both β_1 and β_2 are zero is small, but there are solutions with high likelihood where only one of the coefficients is nonzero. In other words, the likelihood is relatively uninformative about whether both or only one of these features should be included in the model. A sparsifying prior is not much of a help; it simply forces the posterior to become multimodal with modes at where one of the coefficients is close to zero, but it is still challenging to decide which of the modes should be selected and what would be the effect on the predictive performance. Publication V demonstrates that the problem becomes even more difficult when the number of correlating features increases, since for each feature the posterior mass starts to focus more near zero as then most of the features can be removed as long as some are retained. For a recent work on generalizations and limitations of marginal posterior based selection with correlated features, see Barbieri et al. (2018).

Finally, it is worth keeping in mind that even when none of the things discussed above were an issue, the selection based on marginal posterior probabilities attempts to identify the "true feature combination", that is, solve Problem 2 as framed in the introduction of this chapter. This means that the marginal posterior probabilities may not be good indicators about *how* relevant each of the features is. For example, we might have two features so that the first one explained 50 percent of the variation in y whereas the other one explained only 5 percent, but they both would end up having posterior probability close to one given enough data. For finding a sparse feature combination where features that are either redundant or have a negligible effect are removed, the projective framework (Sec. 3.3.2) offers a more natural answer. In addition to solving many of the issues described above, this framework offers a natural solution to how to make predictions with the selected feature combination—a question to which the traditional approaches do not give a clear answer.

3.3.2 Projection predictive framework

The projective feature selection discussed in Publications I, V and VI solves many problems inherent for traditional selection based on marginal posterior relevance assessment discussed at the beginning of Section 3.3. The key idea behind the projective philosophy is to separate the feature selection from the predictive model construction. More precisely, the projective framework (for feature selection) consists of the following two-step procedure:

1. Construct the best possible predictive model you can, which might be complex and potentially uses a lot of features.

2. If the model is too complex, find a simpler model (with acceptable complexity) that gives as similar predictions as possible compared to the original model. For a given model complexity (number of features), the model with the smallest predictive discrepancy compared to the original model should be selected.

The model constructed in the first stage is called the *reference model* and the simplification step a *projection*. The simplified models are usually referred to as a *submodels*. As we shall discuss in a moment, an important aspect of the projection is that it depends solely on the predictive properties of the reference model. In other words, the projection does not care how many features the reference model uses, or whether it employs a sparsifying prior, for instance. Another important aspect is that the simplification step is carried out *only* if the reference model is too complex or cannot be used for some other reason; if one was satisfied with the original model, there is no need for further feature selection or other simplification.

Projection for GLMs

There are a few different ways of formulating the projection, and these are reviewed in detail in Publication V, so we shall introduce them here only briefly. These techniques are generic in the sense that they do not assume any particular model family but they are still best suited for GLMs.

Suppose we have a reference model and a submodel which are parametrized by $\boldsymbol{\theta}_*$ and $\boldsymbol{\theta}$, respectively. If the reference model parameters are given, a natural way of selecting the submodel parameters would be to minimize the discrepancy between the predictive distributions $p(\tilde{y}|\tilde{\mathbf{x}},\boldsymbol{\theta}_*)$ and $p(\tilde{y}|\tilde{\mathbf{x}},\boldsymbol{\theta})$. In the original formulation of Goutis and Robert (1998) and Dupuis and Robert (2003), the discrepancy is measured as average Kullback–Leibler (KL) divergence between the two distributions over the empirical distribution of the features

$$\boldsymbol{\theta}_\perp = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}(p(\tilde{y}|\mathbf{x}_i,\boldsymbol{\theta}_*) \,\|\, p(\tilde{y}|\mathbf{x}_i,\boldsymbol{\theta})). \tag{3.4}$$

As discussed in Publication V, as long as the observation model of the submodel is in the exponential family, projection (3.4) is equivalent to finding the maximum likelihood parameters for $\boldsymbol{\theta}$ with the observed targets y_i replaced by their expected values $\mathrm{E}(\tilde{y}|\mathbf{x}_i,\boldsymbol{\theta}_*)$ as predicted by the reference model.[1] For this reason the above projection is fairly easy to compute for many models.

Since in fully Bayesian inference one accounts for the uncertainty in the reference model parameters, Goutis, Dupuis and Robert proposed taking a set of posterior draws $\{\boldsymbol{\theta}_*^s\}_{s=1}^S$ and projecting these individually to obtain a set of projected parameter values $\{\boldsymbol{\theta}_\perp^s\}_{s=1}^S$ for the submodel. The projection discrepancy is then defined as the average discrepancy over the draws. We refer to this as the *draw-by-draw* projection following terminology in Publication V.

An alternative strategy proposed by Tran et al. (2012) is to integrate over the uncertainty in the parameters of the reference model and form the full posterior predictive distribution $p(\tilde{y}|\tilde{\mathbf{x}},\mathscr{D}) = \int p(\tilde{y}|\tilde{\mathbf{x}},\boldsymbol{\theta}_*)p(\boldsymbol{\theta}_*|\mathscr{D})\mathrm{d}\boldsymbol{\theta}_* \approx \frac{1}{S}\sum_{s=1}^S p(\tilde{y}|\tilde{\mathbf{x}},\boldsymbol{\theta}_*^s)$ and then for the submodel find parameter point estimates that minimize the discrepancy to this distribution

$$\boldsymbol{\theta}_\perp = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}(p(\tilde{y}|\mathbf{x}_i,\mathscr{D}) \,\|\, p(\tilde{y}|\mathbf{x}_i,\boldsymbol{\theta})). \tag{3.5}$$

We refer to this as the *single point* projection as only point estimates for the submodel are computed. This has the advantage that it is much faster to compute since the computational complexity is the same as for projecting a single draw using (3.4) and therefore the computation time is cut down by a factor of S. In practice for GLMs the draw-by-draw projection (3.4) can sometimes yield a slightly more accurate predictive distribution for the submodel, but the difference to the single point projection is typically small and not worth the greatly increased computation time.

[1] Notice though, that this does not hold for the *dispersion* parameters, such as the noise variance σ^2 in regression. See Publication V for more detailed discussion.

Publication V proposes a *clustered* projection that attempts to maintain the accuracy of the draw-by-draw projection but with a greatly reduced computational cost. The idea is to cluster the posterior draws $\{\boldsymbol{\theta}_*^s\}_{s=1}^S$ of the reference model into M clusters $\{\boldsymbol{\theta}_*^s : s \in I_m\}$, $m = 1, \ldots, M$, and then perform a single point projection for each cluster. Here $I_1, \ldots, I_M$ denote the index sets that indicate which draw belongs to which cluster. To make the approach effective, the goal is to assign draws that result in similar predictive fit into the same cluster. Such clustering is easily obtained using, for example, k-means algorithm (see Section 3.3 in Publication V for more details). When the number of clusters approaches the number of draws, the clustered projection approaches the draw-by-draw projection. However, as illustrated in Publication V, often a small number of clusters such as $M = 5$ or $M = 10$ is enough for obtaining predictive distribution close to that of the draw-by-draw projection.

In GLMs where the projected parameters are the regression coefficients $\boldsymbol{\beta}$ (and potentially some dispersion parameter such as the noise variance), the search for sparse submodels is most conveniently done by using a single point projection with some sparsity enforcing penalty (see Section 3.1). For example, Publication V uses Lasso-type L_1-penalization

$$\boldsymbol{\beta}_\perp = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}\big(p(\tilde{y} \,|\, \mathbf{x}_i, \mathscr{D}) \,\|\, p(\tilde{y} \,|\, \mathbf{x}_i, \boldsymbol{\beta})\big) + \lambda ||\boldsymbol{\beta}||_1 \right\}, \tag{3.6}$$

which yields a sequence of models with varying number of nonzeros in $\boldsymbol{\beta}$ when λ is varied. In Publication V it is argued, however, that the penalization should only be used to sort the features and that the predictive accuracy of the sparsest submodels improves if the final projection is done without any penalty. This approach is similar in spirit to the Lasso-OLS hybrid (Efron et al., 2004) and the relaxed Lasso (Meinshausen, 2007). A fairly similar approach was also used by Tran et al. (2012) but with the difference that they used different penalties for different features which resembles more the adaptive Lasso. An alternative to L_1-penalization (or other penalties) is to use generic search heuristics such as forward stepwise excursion. Forward search is computationally more costly but can yield even better results. Furthermore, it has the benefit that it can be used also for draw-by-draw and clustered projections and it does not assume the model to be parametrized by a set of regression coefficients. Regardless of the search strategy, model size selection can be done by selecting the least number of features after which the predictive performance does not markedly improve. The predictive performance can be estimated using cross-validation (see Publication V for details).

One might wonder why the projection with L_1-penalized search would be better than simply using Lasso. This point has been discussed in detail in Publication V, but here we show a simple motivating example. Consider the toy data from the previous subsection given by Equation (3.3). We generated a random dataset of $n = 100$ observations with $\sigma^2 = 1$, together with $d = 100$ features with correlation $\rho = 0.5$. In addition, 900 irrelevant noise features

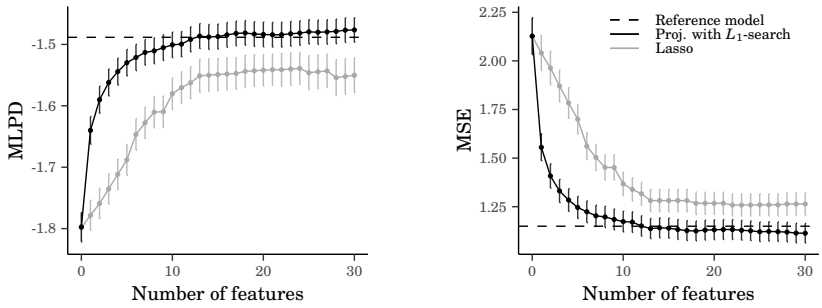

Figure 3.2. Illustration of projective selection. The training data has $n = 100$ observations with 1000 features out of which 100 are relevant but correlated with each other and therefore carry similar information (the rest are completely irrelevant). Left plot shows the mean log predictive density (MLPD) and right plot the predictive mean squared error (MSE) as a function of features selected, both evaluated on an independent test set of 1000 observations (vertical lines denote one standard error bars). The reference model (dashed horizontal) is obtained from Bayesian linear regression using the first 5 principal components. The projection (black) is the single point projection with L_1-search (Eq. (3.6)) but the predictions are computed without any penalization. Results for Lasso (gray) are shown for comparison.

generated from standard Gaussian were added so that the total number of features was 1000. The reference model was constructed simply by fitting the Bayesian linear regression model to the first 5 principal components of the features. The sparse submodels using the original features were found by the single point projection with L_1-search (Eq. (3.6)) but the predictions for the submodels were computed without any penalization as explained above.

Figure 3.2 shows the results. The Bayesian principal component regression gives more accurate results than the Lasso and the projected submodels eventually converge towards the reference model when more features are added. The submodels found by projection strictly dominate the ones found by Lasso in terms of accuracy for a given model size.[2] This example is a manifestation to the principle argued throughout this chapter; no feature selection is always needed for obtaining good predictions (principal component regression), and solving the prediction problem first may lead to improved feature selection (projection). This example illustrates also that the reference model construction can be computationally very efficient: in this case the principal component regression using MCMC takes only a few seconds on a standard laptop.

Projection for GPs

The above formulation of the projection (Equations (3.4) and (3.5)) is not ideally suited for the nonparametric GP models (Section 2.5). Due to their flexibility, it is nontrivial to design a projection that would be computationally feasible but would guarantee a small discrepancy between the reference model and the

[2]To compute the log predictive density for the Lasso, the noise variance σ^2 for each model size d' was estimated using the method proposed by Reid et al. (2016), that is, by dividing the squared residuals by $n - d'$.

submodel not only at the training points but also everywhere else in the feature space. This topic is pursued tentatively in Publication VI.

The approach proposed in the paper is to estimate the submodel hyperparameters by minimizing the KL-divergence between the posterior distributions for the latent values $\mathbf{f}$ in the full model and in the submodel with fewer features

$$\boldsymbol{\theta}_\perp = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}\big(\mathrm{N}\big(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*\big) \,\|\, \mathrm{N}\big(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\big)\big). \tag{3.7}$$

Here $\boldsymbol{\mu}_*$ and $\boldsymbol{\Sigma}_*$ are the posterior mean and covariance of $\mathbf{f}$ in the full model and $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ correspondingly for the submodel with hyperparameters $\boldsymbol{\theta}$. After having learned the hyperparameters of the submodel, the predictions can be made in a standard GP fashion. The results in Publication VI indicate that when combined with a forward search, the minimization of the above KL-criterion tends to find better tradeoff between submodel accuracy and sparsity than selecting features using the learned length-scale values (ARD). The drawback is that the KL-minimization is computationally hugely more expensive as in the forward excursion up to d' features the projection needs to be computed for $O(dd')$ models, and each projection has a computational complexity of $O(n^3)$. For more discussion on the method, see the original paper.

Related approaches

The approach of using a reference model with projective selection has received relatively little attention in the linear model literature but some closely related methods exist. The most closely related approach is the frequentist "preconditioning" for feature selection proposed by Paul et al. (2008). A related Bayesian method is the "posterior summary selection" by Hahn and Carvalho (2015) which is essentially merely a different formulation of the projection (a different loss function).

Another interesting idea is the "model compression" of Bucilă et al. (2006) where a complex neural network (or an ensemble of them) is replaced by a simpler one with the motivation to achieve faster out-of-sample predictions at test time. In this approach the smaller network is trained on a large (artificial) dataset where the labels are determined by the ensemble network, so the smaller network learns to mimic the larger model. A very similar method is the "knowledge distillation" of Hinton et al. (2015) which is greatly inspired by the model compression idea of Bucilă et al. It appears that since the paper by Hinton et al., the compression of neural networks has drawn notable attention; at the time of writing this, the knowledge distillation paper has received more than 1600 citations according to Google Scholar.

Yet another related method is the local interpretable model-agnostic explanation (LIME) framework (Ribeiro et al., 2016). In this method a simpler model (linear) is fitted to the predictions of a complex model (neural network) in the vicinity of a given test point which allows one to get insights of which features (such as pixels in an image) have high weights in the classification. This has been explored also from a more Bayesian viewpoint by Peltola (2018).

4. Summary of the contributions

This chapter briefly summarizes and discusses the main contributions of the six publications of the thesis.

4.1 Predictive inference (Publications I–IV)

Publication I compares several Bayesian methods for selecting features in linear regression and classification models over a wide range of simulated and real world datasets. Some of the methods are general purpose methods for estimating the predictive performance of any Bayesian model (such as cross-validation and information criteria) and some are designed for feature selection (for example marginal posterior relevance assessment). An important take-home message of the paper is that regardless of the used technique, feature selection rarely improves the predictive performance compared to accounting for model uncertainty with a reasonable prior over the competing models. Although it might appear surprising, this result is in perfect accordance to what has been advocated earlier by some other authors, see the discussion in Section 3.2. In the paper the full Bayesian solution is taken to be the Bayesian model averaging over the different feature combinations which—as pointed out in Section 2.3—is the same as using spike-and-slab prior for the model with all features. Given that empirically the horseshoe prior has been reported to give very comparable results to spike-and-slab on a variety of problems (Carvalho et al., 2009, 2010; Polson and Scott, 2011; Hernández-Lobato and Hernández-Lobato, 2013; Hernández-Lobato et al., 2015) it could be expected that the conclusions of Publication I are not sensitive to adopting the spike-and-slab instead horseshoe or some other sparsity promoting prior. The paper also gives recommendations about the preferred approaches for feature selection when simplification of the model is desirable (see Section 4.2).

Publications II and III discuss the horseshoe prior and advance the theoretical understanding of the role played by the global shrinkage parameter τ (see Section 2.3). Furthermore, the latter paper introduces the regularized horseshoe prior that can be used to control the shrinkage for the parameters that are far from zero. The connection to the spike-and-slab is discussed in detail. The

benefits of these methodological advances are twofold. Firstly, they aid the understanding about how the sparsity and regularization effects are encoded in the horseshoe prior and help formulating the prior information about these characteristics. Secondly, as demonstrated through practical examples, even weakly informative choices both for the sparsity and regularization can typically help to robustify and speed up the MCMC sampling and in some cases also improve the predictive accuracy simultaneously. It is worth emphasizing that neither of the proposed ideas—that is, how to incorporate sparsity and regularization information to the prior—is restricted to the horseshoe prior but can also be used with other priors that can be expressed as scale mixtures of Gaussians. This broadens their applicability and can potentially make them useful for other priors as well.

Also the empirical results of Publications II and III strongly support the idea that no feature selection is necessary for obtaining good predictions, but this can come with a high computational price. Motivated by this, Publication IV studies some computational shortcuts for datasets with a high-dimensional feature space. In addition to PCA and supervised PCA, the paper proposes an iterative version of the latter algorithm and compares the performance of these three methods when used to reduce the dimensionality to something that is conveniently handled by Bayesian methods. The paper concludes that in many cases the dimension reduction can be very effective; the model fitted using the reduced set of features can obtain a high accuracy while the computation time might be only a small fraction of what would be needed when fitting a Bayesian model using the original set of features.[1] The paper also concludes that none of the three dimension reduction algorithms performs better than the other two over all datasets, but in almost all experiments at least one of them gave very good results. Dimension reduction is something not routinely used in the Bayesian literature—perhaps because people have tendency to either use fully Bayesian methods or not Bayesian methods at all—but from a pragmatic point of view, this approach can certainly be very useful.

4.2 Feature selection (Publications I, V and VI)

As discussed in Section 4.1, Publication I compares several Bayesian model selection techniques for feature selection but recommends to avoid the selection completely if the predictive inference is the only concern. However, when a smaller subset of features need to be selected, the paper advocates the projective framework (Section 3.3.2) which demonstrates overall superior performance among the methods under comparison. The results indicate that the projection tends to find an excellent tradeoff between the number of features and predictive

[1]The paper does not actually compare to fully Bayesian approach using the original features, but this pattern applies to those datasets that were also used in Publication III where fully Bayesian inference (without dimension reduction) was used.

accuracy, which is obviously hugely beneficial if the goal is to find the simplest possible model that does not sacrifice much predictive accuracy compared to using all the features.

The paper argues that the superior performance of the projection over the other methods—especially the generic model selection techniques such as cross-validation and information criteria—is due to better tradeoff between bias and variance in the performance estimates for the submodels. For example leave-one-out cross-validation (LOO) is known to give a nearly unbiased accuracy estimate for any candidate model (Watanabe, 2010), but for small n the estimator has a high variance. Model selection based on optimizing LOO therefore results in high variability in the model selected. Furthermore, when many models are being compared (such as in feature selection), the selection process tends to (over)fit to the random noise in the LOO estimates. This can lead to a selection of a suboptimal model and considerable selection induced bias in the performance evaluation of the selected model (see Section 3 in Publication I). This phenomenon has been known for a long time (Stone, 1974; Rencher and Pun, 1980; Ambroise and McLachlan, 2002; Reunanen, 2003; Cawley and Talbot, 2010) but often tends to get overlooked. The projective selection on the other hand is likely to have some bias since the reference model is never perfect in practice but this is more than compensated by substantially reduced variability in the feature combination that gets selected.

Publication V studies the projection further by reviewing the different projection techniques and by proposing a new projection method that unifies the existing techniques and gives a good tradeoff between accuracy and speed (see Section 3.3.2). The paper makes also some other methodological contributions such as showing a fast way of validating the selection process and selecting an appropriate model size using approximate LOO. The paper provides also an extensive comparison to popular non-Bayesian techniques, in particular to Lasso and elastic net (Section 3.1), and shows the superiority of the projection in the "small n, large d" settings. Furthermore, the paper proves a theorem that gives a theoretical argument of why learning the parameters in a linear model via projection from a reference model can improve predictive accuracy compared to standard fitting to the observed data regardless of the method used to select the feature combination—a phenomenon that was empirically observed and demonstrated also in Publication I. The methods discussed in the paper are implemented in a freely available R-package `projpred` which makes them easily available to the community.[2]

Finally, Publication VI explores the implementation of the projective selection to GP models. The paper demonstrates empirically the difficulty of the input relevance assessment based on the length-scale values (automatic relevance determination, ARD) by showing that this tends to favor features along which the latent function is nonlinear. The paper proposes a way of implementing the projection to GPs together with empirical results that show the superiority

[2]The software is hosted at `https://github.com/stan-dev/projpred`.

compared to ranking the features using ARD. This technique could still be considered tentative and immature due to the high computational cost which limits its practical applicability (see Section 3.3.2). Another issue is that due the flexibility of the GPs the minimization of the KL-divergence can sometimes behave in an undesirable manner, namely so that a small divergence locally at the training data points does not guarantee small discrepancy elsewhere.

5. Conclusion

This thesis has focused on predictive inference and feature selection for problems with scarce data but high-dimensional feature space. The main argument has been that in many cases it can be beneficial to solve these problems in two stages, by first solving the prediction problem and then reducing the number of features using the projective framework if needed. This strategy can both be computationally efficient, give an excellent tradeoff between predictive accuracy and model complexity, and avoid several difficulties characteristic to the traditional Bayesian approaches. As discussed in some of the papers and this introduction, the advocated conceptual idea is not new but has been largely overlooked by the statistical community.

To make this approach easily accessible, the thesis has proposed and discussed practical and computationally efficient tools both for predictive model construction and the subsequent projective feature selection. The emphasis has been on Bayesian generalized linear models but preliminary exploration of these ideas to the Gaussian processes (Publication VI) has also been made with promising results. In addition, this work has advanced the theory of the continuous shrinkage priors by introducing new tools for formulating the sparsity and regularization information to the prior. These theoretical advances are useful for better understanding some of the priors that are currently in common use, but they can also help in the construction of a predictive model.

Currently the proposed framework is fairly mature for the GLMs but more work is needed for extending this to a wider range of models. Another aspect is that while the framework makes perfect sense conceptually and is observed to give good results in many cases empirically, the current theoretical understanding of the technique is very limited. There is no formal understanding of when the proposed approach could be expected to work better than the alternative strategies. This point has been touched upon in Publication V, but more research would be needed in this area.

References

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566.

Armagan, A., Clyde, M., and Dunson, D. B. (2011). Generalized beta mixtures of Gaussians. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 523–531.

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, 23:119–143.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897.

Barbieri, M. M., Berger, J. O., George, E. I., and Ročková, V. (2018). The median probability model and correlated variables. *arXiv:1807.08336*.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Bucilă, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 535–541. ACM.

Bui, T. D., Yan, J., and Turner, R. E. (2017). A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research*, 18(104):1–72.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In van Dyk, D. and Welling, M., editors, *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Cawley, G. C. and Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107.

Dupuis, J. A. and Robert, C. P. (2003). Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, 111(1-2):77–94.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman & Hall, third edition.

Gelman, A., Jakulin, A., Pittau, M. G., and Yu-Sung, S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–374.

Ghosh, J., Li, Y., and Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383.

Goutis, C. and Robert, C. P. (1998). Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika*, 85(1):29–37.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Griffin, J. E. and Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, 53(4):423–442.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Guyon, I., Gunn, S., Ben Hur, A., and Dror, G. (2006). Design and analysis of the NIPS2003 challenge. In Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A., editors, *Feature Extraction, Foundations and Applications*, pages 237 – 263. Springer.

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448.

Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association*, 96(455):1122–1132.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the Lasso and generalizations*. Chapman & Hall.

Hensman, J., Durrande, N., and Solin, A. (2018). Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. In Nicholson, A. and Smyth, P., editors, *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational Gaussian process classification. In Lebanon, G. and Vishwanathan, S. V. N., editors, *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360.

Hernández-Lobato, D. and Hernández-Lobato, J. M. (2013). Learning feature selection dependencies in multi-task learning. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 746–754.

Hernández-Lobato, D. and Hernández-Lobato, J. M. (2016). Scalable Gaussian process classification via expectation propagation. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 168–176.

Hernández-Lobato, D., Hernández-Lobato, J. M., and Dupont, P. (2013). Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14:1891–1945.

Hernández-Lobato, D., Hernández-Lobato, J. M., and Suárez, A. (2010). Expectation propagation for microarray data classification. *Pattern Recognition Letters*, 31(12):1618–1626.

Hernández-Lobato, J. M., Hernández-Lobato, D., and Suárez, A. (2015). Expectation propagation in linear regresssion models with spike-and-slab priors. *Machine Learning*, 99:437–487.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv:1503.02531*.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003). Gene selection: a Bayesian variable selection approach. *Bioinformatics*, 19(1):90–97.

Matthews, A. G. d. G., Hensman, J., Turner, R., and Ghahramani, Z. (2016). On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 231–239.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). *SparseNet*: coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, second edition.

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis*, 52(1):374–393.

Minka, T. P. (2001). Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1036.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Neal, R. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*, chapter 5. CRC Press.

Neal, R. and Zhang, J. (2006). High dimensional classification with Bayesian neural networks and Dirichlet diffusion trees. In Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A., editors, *Feature Extraction, Foundations and Applications*, pages 265–296. Springer.

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. Arnold, second edition.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Paul, D., Bair, E., Hastie, T., and Tibshirani, R. (2008). "Preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 36(4):1595–1618.

Peltola, T. (2018). Local interpretable model-agnostic explanations of Bayesian predictive models via Kullback-Leibler projections. In Aha, D. W., Darrell, T., Doherty, P., and Magazzeni, D., editors, *Proceedings of the 2nd Workshop on Explainable Artificial Intelligence*, pages 114–118.

Peltola, T., Marttinen, P., and Vehtari, A. (2012). Finite adaptation and multistep moves in the Metropolis–Hastings algorithm for variable selection in genome-wide association analysis. *PLoS ONE*, 7(11).

Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian statistics 9*, pages 501–538. Oxford University Press, Oxford.

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

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In Kaski, S. and Corander, J., editors, *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in Lasso regression. *Statistica Sinica*, 26(1):35–67.

Rencher, A. C. and Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics*, 22:49–53.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3:1371–1382.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. ACM.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer, second edition.

Ročková, V. and George, E. I. (2018). The spike-and-slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(2):319–392.

Salimbeni, H. and Deisenroth, M. (2017). Doubly stochastic variational inference for deep Gaussian processes. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4588–4599.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In Weiss, Y., Schölkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 1257–1264.

Stan Development Team (2018). Stan modeling language users guide and reference manual, version 2.18.0.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147.

Tang, X., Xu, X., Ghosh, M., and Ghosh, P. (2018). Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya A*, 80(2):215–246.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In van Dyk, D. and Welling, M., editors, *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574.

Titsias, M. K. and Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2339–2347.

Tran, M.-N., Nott, D. J., and Leng, C. (2012). The predictive Lasso. *Statistics and Computing*, 22(5):1069–1084.

Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29:1580–1607.

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

Yu, S., Yu, K., Tresp, V., Kriegel, H.-P., and Wu, M. (2006). Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 464–473, New York, NY, USA. ACM.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 68(1):49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, Y., Reich, B. J., and Bondell, H. D. (2017). High dimensional linear regression via the R2-D2 shrinkage prior. *arXiv:1609.00046*.

Zhou, X., Liu, K.-Y., and Wong, S. T. (2004). Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics*, 37:249–259.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67(2):301–320.

Aalto-DD 84/2019

ISBN 978-952-60-8538-8 (printed)
ISBN 978-952-60-8539-5 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

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

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS