Department of Biomedical Engineering and Computational Science

Sparse Bayesian Linear Models

Computational Advances and Applications in Epidemiology

Tomi Peltola

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 206/2014

Sparse Bayesian Linear Models

Computational Advances and Applications in Epidemiology

Tomi Peltola

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 auditorium E of the school on 12 January 2015 at 12.

Aalto University
School of Science
Department of Biomedical Engineering and Computational Science

Supervising professor
Jouko Lampinen

Thesis advisors
Prof. Aki Vehtari
Dr. Pekka Marttinen

Preliminary examiners
Prof. Matthew Stephens, The University of Chicago, USA
Dr. José Miguel Hernández-Lobato, Harvard University, USA

Opponent
Prof. Tom Heskes, Radboud University, The Netherlands

Aalto University publication series
DOCTORAL DISSERTATIONS 206/2014

ISBN 978-952-60-6011-8 (printed)
ISBN 978-952-60-6012-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-6012-5

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

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

Author
Tomi Peltola

Name of the doctoral dissertation
Sparse Bayesian Linear Models: Computational Advances and Applications in Epidemiology

Publisher School of Science

Unit Department of Biomedical Engineering and Computational Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 206/2014

Field of research Computational Engineering

Manuscript submitted 7 October 2014 **Date of the defence** 12 January 2015

Permission to publish granted (date) 10 November 2014 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Recent advances in measurement technologies have transformed the landscape of studies in the genetic and metabolic determinants of diseases and other complex traits. DNA and blood samples can be cost- and time-efficiently interrogated for millions of genetic markers and hundreds of circulating metabolites. While the scale and unbiased nature of the characterization of the individual samples creates opportunities for new discoveries, they also pose a challenge for the statistical analysis of the data. One approach for tackling the issues, and a focus of much recent research in statistical methodology, is searching for linear relationships with a sparsity assumption, that is, the presence of only a limited number of practically relevant relationships among the vast number of possibilities.

This thesis studies aspects of the statistical modelling and computation with the linearity and sparsity assumptions in the framework of Bayesian data analysis. First, a hierarchical extension of the spike and slab prior distribution for sparse linear regression modelling, to allow additive and dominant effects in genome-wide association analysis, is presented. The model is applied to search for genetic markers related to blood cholesterol levels. A tailored, finitely adaptive Markov chain Monte Carlo algorithm is studied for the computation. Second, an approach for constructing deterministic Gaussian approximations for Bayesian linear latent variable models using the expectation propagation method is described. The main advance is an efficient numerical solution to the moment integrals for bilinear probability factors. Third, a model for the prediction of the risk of adverse cardiovascular events in diabetic individuals using candidate biomarkers is presented. The model is extended hierarchically to include data from non-diabetic individuals. Shrinkage priors and projective covariate selection are applied to identify biomarkers with predictive value.

The results of the studies demonstrate benefits from the hierarchical Bayesian modelling. Despite the advances here and generally in the literature, the computation in sparse models and large datasets remains challenging. On the other hand, given the fast pace in the development of deterministic approximation methods, assessing their role in predictive covariate selection would seem timely.

Keywords Bayesian linear modelling, sparsity, Markov chain Monte Carlo, approximate inference

ISBN (printed) 978-952-60-6011-8 **ISBN (pdf)** 978-952-60-6012-5

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 130 **urn** http://urn.fi/URN:ISBN:978-952-60-6012-5

Tekijä
Tomi Peltola

Väitöskirjan nimi
Harvuutta suosivat bayesilaiset lineaarimallit: laskennallisia menetelmiä ja sovelluksia epidemiologiassa

Julkaisija Perustieteiden korkeakoulu

Yksikkö Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 206/2014

Tutkimusala Laskennallinen tekniikka

Käsikirjoituksen pvm 07.10.2014 **Väitöspäivä** 12.01.2015

Julkaisuluvan myöntämispäivä 10.11.2014 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Viimeaikaiset edistysaskeleet mittausteknologioissa ovat mahdollistaneet uudenlaisten tutkimusmenetelmien soveltamisen sairauksien ja muiden monitekijäisten piirteiden perinnöllisen ja aineenvaihdunnallisen taustan selvittämiseen. DNA- ja verinäytteistä pystytään mittaamaan verrattain nopeasti ja kustannustehokkaasti miljoonia geenitekijöitä ja satoja aineenvaihdunnan tuotteita. Vaikka yksittäisten näytteiden karakterisoinnin laajuus ja harhaton luonne johtanee uusiin löydöksiin, se myös asettaa haasteita aineistojen tilastolliselle analyysille. Lineaarisuus- ja harvuusoletukset ovat mahdollisia lähtökohtia näihin haasteisiin vastaamiseen ja ne ovatkin olleet viime aikoina tilastollisten menetelmien tutkimuksen keskiössä. Harvuus viittaa käsitykseen, jonka mukaan vain pieni osa kaikista mahdollisista tilastollisista yhteyksistä aineistossa ovat oleellisia.

Tässä väitöskirjassa tutkitaan tiettyjä näkökulmia bayesilaiseen tilastolliseen mallinnukseen ja laskentaan harvoissa lineaarimalleissa. Ensimmäisessä mallinnusongelmassa esitetään harvoja ratkaisuja suosivan spike and slab -priorijakauman hierarkkinen laajennus additiivisen ja dominantin perinnöllisen vaihtelun tutkimiseen genominlaajuisissa aineistoissa. Mallia sovelletaan veren kolesterolipitoisuuksiin vaikuttavien perinnöllisten tekijöiden etsimiseen. Työssä tutkitaan myös räätälöityä Markov-ketju Monte Carlo -algoritmia mallin laskennassa. Toisessa mallinnusongelmassa käsitellään determinististen gaussisten approksimaatioiden sovittamista lineaarisille latentti-muuttujamalleille expectation propagation -algoritmilla. Pääkontribuutio on tehokas numeerinen ratkaisu bilineaaristen todennäköisyystekijöiden momenteille. Kolmannessa mallinnusongelmassa esitellään kardiovaskulaaritapahtumien riskiennustemalli diabeetikoille pohjautuen vähän tutkittuihin aineenvaihdunnan tekijöihin. Mallille esitetään hierarkkinen laajennus ei-diabeetikoiden aineiston sisällyttämiseksi. Työssä sovelletaan harvoja ratkaisuja suosivia prioreja ja projektiivista muuttujanvalintaa ennustekykyisten tekijöiden tunnistamiseksi.

Tutkimusten tulokset heijastavat hierarkkisen bayesilaisen mallinnuksen hyötyjä. Tässä työssä ja yleisesti kirjallisuudessa esitetyistä edistyksistä huolimatta harvuutta suosivien mallien laskenta suurissa malliavaruuksissa on edelleen haastavaa. Toisaalta deterministisen approksimaatiomenetelmien nopean kehityksen avaamien mahdollisuuksien selvittäminen prediktiiviseen muuttujanvalintaan liittyen voisi olla ajankohtaista.

Avainsanat bayesilainen lineaarinen mallintaminen, harvuus, Markov-ketju Monte Carlo, likimääräinen päättely

ISBN (painettu) 978-952-60-6011-8 **ISBN (pdf)** 978-952-60-6012-5

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2014

Sivumäärä 130 **urn** http://urn.fi/URN:ISBN:978-952-60-6012-5

Preface

This thesis is a result of research conducted at the Department of Biomedical Engineering and Computational Science at Aalto University during 2010–2014. I am grateful for the financial support provided by the Finnish Doctoral Programme in Computational Sciences FICS and Tekniikan edistämissäätiö (Mirjam Helena Paloheimon erikoisrahasto).

Rather than this final product, it is certainly the path to it that leaves a stronger impression on the author. I would like to thank Prof. Aki Vehtari for keeping me strolling mostly in the right direction, but also allowing some space and time to wander about. During my stay in the Bayesian Methodology research group, I have learned a lot about research in general, and about probabilistic modelling and data analysis in particular. I want to thank Dr. Pekka Marttinen for the invaluable guidance and close collaboration during the first half of my doctoral studies, and for providing a model for analyticity in research and a keen eye for the English language, which are something to strive for. I would also like to thank Prof. Jouko Lampinen for providing an excellent environment for research, allowing uninterrupted focus on the work.

I wish to express my gratitude to Prof. Matthew Stephens and Dr. José Miguel Hernández-Lobato, both of whose research work I greatly admire, for their time and work put into pre-examining my thesis.

I am grateful to all my collaborators, Prof. Antti Jula, Prof. Markus Perola, Prof. Veikko Salomaa, Dr. Aki Havulinna, and Dr. Pasi Jylänki, for their time, expertise, and resources, without which this work would not have been possible. My gratitude also goes to the excellent and friendly IT support and administrative personnel at BECS. I would like to thank Prof. Kimmo Kaski, Prof. Mika Ala-Korpela, Prof. Per-Henrik Groop, and Dr. Ville-Petteri Mäkinen for introducing me to scientific research before my doctoral studies, and the director of FICS, Prof. Samuel Kaski, and

the coordinator of FICS, Dr. Ella Bingham, for the resources provided by the doctoral programme.

I have been lucky to have many inspiring colleagues during the years I have worked towards this thesis. Special thanks to Arno Solin, Juho Piironen, Ville Tolvanen, Dr. Jaakko Riihimäki, Janne Ojanen, Dr. Jouni Hartikainen, Dr. Simo Särkkä, Dr. Mari Myllymäki, Juho Kokkala, Dr. Jarno Vanhatalo, Olli-Pekka Koistinen, Dr. Tommi Mononen, Ville Väänänen, Henri Seijo, Tuomas Sivula, Dr. Eli Parviainen, Dr. Niina Sandholm, Dr. Taru Tukiainen, Antti Kangas, Jaakko Niemi, Niko Lankinen, Dr. Linda Kumpula, Aino Salminen, Sanna Kuusisto, Dr. Riku Linna, and the FinnDiane group at the Folkhälsan Research Center. Big thanks also to my fellow Bio03&04 students.

Thank you, mother and father, brother, grandparents, for your unconditional support. Emilia, you have my deepest gratitude.

Espoo, November 24, 2014,

Tomi Peltola

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 Tomi Peltola, Pekka Marttinen, Antti Jula, Veikko Salomaa, Markus Perola, and Aki Vehtari. Bayesian Variable Selection in Searching for Additive and Dominant Effects in Genome-Wide Data. *PLoS ONE*, 7, 1, e29115, January 2012.

II Tomi Peltola, Pekka Marttinen, and Aki Vehtari. Finite Adaptation and Multistep Moves in the Metropolis-Hastings Algorithm for Variable Selection in Genome-Wide Association Analysis. *PLoS ONE*, 7, 11, e49445, November 2012.

III Tomi Peltola, Pasi Jylänki, and Aki Vehtari. Expectation Propagation for Likelihoods Depending on an Inner Product of Two Multivariate Random Variables. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Journal of Machine Learning Research: Workshop and Conference Proceedings, 33, 769–777, Reykjavik, Iceland, April 2014.

IV Tomi Peltola, Aki S. Havulinna, Veikko Salomaa, and Aki Vehtari. Hierarchical Bayesian Survival Analysis and Projective Covariate Selection in Cardiovascular Event Risk Prediction. In *Proceedings of the Eleventh UAI Bayesian Modeling Applications Workshop*, CEUR Workshop Proceedings, Vol-1218, 79–88, Quebec, Canada, July 2014.

Author's Contribution

Publication I: "Bayesian Variable Selection in Searching for Additive and Dominant Effects in Genome-Wide Data"

Peltola participated in the design of the statistical analysis approach, implemented and ran the analysis, and wrote the first version of the manuscript. Initial analysis approach was designed by Vehtari and Marttinen in a collaborative project headed by Perola. The dataset, including imputations and preliminary quality control, was provided by Jula, Salomaa, and Perola. All co-authors reviewed the manuscript.

Publication II: "Finite Adaptation and Multistep Moves in the Metropolis-Hastings Algorithm for Variable Selection in Genome-Wide Association Analysis"

Peltola designed the study with Marttinen and Vehtari, implemented and ran the analysis, and wrote the main part of the manuscript. Marttinen wrote the examples 1 and 2. All co-authors participated in revising the manuscript.

Publication III: "Expectation Propagation for Likelihoods Depending on an Inner Product of Two Multivariate Random Variables"

Peltola derived the main contribution of the work, implemented and run the analysis, and wrote the main part of the manuscript. Jylänki contributed to the writing of the introduction and provided valuable knowledge on the intricacies of the expectation propagation algorithm. All co-authors commented on and provided revisions to the manuscript.

Publication IV: "Hierarchical Bayesian Survival Analysis and Projective Covariate Selection in Cardiovascular Event Risk Prediction"

Peltola participated in designing the statistical analysis approach, implemented and ran the analysis, and wrote the manuscript. Vehtari provided initial ideas for the analysis approach. Havulinna and Salomaa provided the dataset and expertise with regard to the application. Havulinna performed the imputation of the dataset. All co-authors participated in revising the manuscript.

1. Introduction

The unifying theme in this thesis is sparse Bayesian linear modelling. The work in the four publications consists of studying certain aspects of the computation in sparse Bayesian linear models (all publications to some extent; Publications II and III in particular), of applied modelling for identifying genetic markers in metabolic traits (Publications I and II), and of predictive modelling of the risk of adverse cardiovascular events with new candidate biomarkers (Publication IV). The introductory part of the thesis provides background and further motivation for the work presented in the publications. In particular, Chapter 2 reviews the specification of sparse Bayesian linear models and Chapter 3 discusses the computation. A summary of the publications is given in Chapter 4. A general motivation for the work and the contributions in the thesis are presented below.

Chronic, slowly-developing non-communicable diseases are a large problem for the ageing population of Finland (and worldwide) and a burden for the health-care system. Such diseases develop over lifetime as an interplay of lifestyle (diet, physical activity, etc.) and environmental and genetic components. Knowledge of the risk factors, diagnosis in early stages of development, and possibilities for targeted intervention would be essential for the prevention and effective treatment of the diseases. With advances in measurement technologies that have made the determination of genetic markers and snapshots of the metabolic state of the body possible in large scale, the epidemiological research has also partly shifted into *hypothesis-free* exploration of associations to identify promising candidate factors for further research. This, however, presents a challenge for the statistical analysis of the data as many of the interesting associations are individually expected to be weak and the number of samples is often relatively small compared to the number of measured characteristics. While

common statistical methods especially in genetics test each characteristic individually for association, the simultaneous modelling of all characteristics, for example by sparse linear models, would better account for the hypothesized multifactorial basis of many common complex diseases and traits [Plomin et al., 2009, Gibson, 2010].

From a statistical point of view, the application of linear models for identifying associations in these kinds of datasets can be motivated by their interpretability, stability, and computational efficiency. Incorporating the assumption of sparsity, that only few of the measured features will be associated to or predictive of the disease or trait studied, can further increase the interpretability (and possibly statistical and computational efficiency) of the model. The bias-variance decomposition provides a way to formalize this discussion with regard to stability (see, e.g., Bishop [2006, Chapter 3.2] for more details). Consider predicting values of an outcome y with a vector of covariates $\boldsymbol{x}$, which follow a joint distribution $p(y, \boldsymbol{x})$. The predictor function $f(\boldsymbol{x})$ is trained with a dataset $\mathcal{D} = \{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$. Let $h(\boldsymbol{x}) = \int yp(y|\boldsymbol{x})dy$, which is the optimal predictor under the squared error loss. The bias-variance decomposition divides the expected squared error of the predictions into three parts:

$$\mathrm{E}\big[(f(\boldsymbol{x}) - y)^2\big] = \mathrm{E}\big[(h(\boldsymbol{x}) - y)^2 + (\mathrm{E}_{\mathcal{D}}[f(\boldsymbol{x})] - h(\boldsymbol{x}))^2 + (f(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D}}[f(\boldsymbol{x})])^2\big]\,,$$

where E denotes expectation over $p(y, \boldsymbol{x})p(\mathcal{D})$ and $\mathrm{E}_{\mathcal{D}}$ over the distribution of training datasets $p(\mathcal{D})$. The first term on the right hand side is the intrinsic noise in predicting y using $\boldsymbol{x}$ and does not depend on f. The second term is the (squared) bias of f, and the third term is the variance of f (w.r.t. different training sets). A flexible model for f can make the bias small as it will be able to mimic h well in expectation, but it will also need a lot of training data to satisfy the large amount of freedom in its fitting. A linearity assumption for f, for example $f(\boldsymbol{x}) = \sum_j \beta_j x_j$ where the β_j are parameters fitted using the training data, is strong, and it is often expected to lead to bias in real applications. However, when the amount of training data is scarce, the variance term may be the dominating factor, and strong assumptions can alleviate this. The sparsity assumption can be seen as a further trade-off between bias and variance (e.g., when the number of covariates is larger than n, the coefficients in multiple linear regression models cannot be uniquely found with least squares fitting without regularization; see Chapter 2 for more discussion).

The linear models presented in this work are constructed in the framework of Bayesian modelling. The beginnings of the Bayesian approach

to statistical analysis dates back to the 18th century with the work of Thomas Bayes and Pierre-Simon Laplace [Stigler, 1986]. Bayesian modelling gained popularity in the second part of the 20th century, at least in part spurred by the availability of computational resources and methods required for its practical application (e.g., Beaumont and Rannala [2004], Wolpert [2004]). Modern introductions to Bayesian modelling are provided, for example, by O'Hagan and Forster [2004] and Gelman et al. [2014].

The core idea in Bayesian modelling is to represent uncertainties in the modelled system and in the observations made of the system using probability distributions. The inference is then conducted by the rules of probability theory, with the main result being the distribution of the quantities of interest conditional on the model assumptions and the observations. The conditional distribution can then be queried for model parameter estimates and related uncertainties, for predictions, and for answers to decision problems by means of decision theory. An important difference to frequentist statistics is that the model parameters are also modelled as random variables, even if the true parameters (to the extent that such can be thought to exist) are thought to have some fixed but unknown values and not governed by any random process. The uncertainties represented by the probability distributions then correspond to the lack of knowledge about the true values. A significant advantage of the all-encompassing probability formalism is that basic models can be extended and combined in hierarchies to form more complex ones. The related uncertainties are automatically propagated through the model structure following the rules of probability theory.

An essential tool to arrive at the conditional distribution of the quantities of interest is the Bayes' theorem:

$$p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}) = \frac{p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})},$$

where $p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}$. Here, $\boldsymbol{\theta}$ are the model parameters, $\mathcal{D}$ is the observed data, and $\mathcal{M}$ denotes the model assumptions. For notational convenience, $\mathcal{M}$ is often dropped from the equations (as will be done also here in the following chapters), and it is implicitly assumed understood that all inferences are conditional on the model assumptions. $p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M})$, as a probability distribution for $\mathcal{D}$, is the observation model. When thought of as a function of $\boldsymbol{\theta}$, with fixed observations $\mathcal{D}$, it is called the likelihood function. $p(\boldsymbol{\theta}|\mathcal{M})$ is the prior distribution, and contains the

knowledge about the model parameters $\boldsymbol{\theta}$ before the observations are accounted for. The conditional distribution of the model parameters after accounting for the observations, $p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$, is called the posterior distribution. $p(\mathcal{D}|\mathcal{M})$ is the prior predictive distribution of the observations. For fixed observations $\mathcal{D}$, it normalizes the posterior distribution and is called the marginal likelihood (or evidence) of the model.

For making predictions and quantifying their uncertainty, the posterior predictive distribution for an observable quantity $\tilde{y}$ is available as

$$p(\tilde{y}|\mathcal{D},\mathcal{M}) = \int p(\tilde{y}|\boldsymbol{\theta},\mathcal{D},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})d\boldsymbol{\theta}.$$

Often, all information in the data $\mathcal{D}$ related to the prediction of $\tilde{y}$ is contained in the posterior distribution, and $p(\tilde{y}|\boldsymbol{\theta},\mathcal{D},\mathcal{M}) = p(\tilde{y}|\boldsymbol{\theta},\mathcal{M})$. For example, a common assumption is that the observations $\mathcal{D} = \{y_i : i = 1,\ldots,n\}$ (and the future observable $\tilde{y}$), are exchangeable and modelled by conditionally independent probability distributions $p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M}) = \prod_i p(y_i|\boldsymbol{\theta},\mathcal{M})$.

While, given the model and observations, deriving the posterior distribution is often conceptually easy, carrying out the computations in practice may pose a challenge. In general, the required integrals are analytically intractable except for simple models. In discrete models, the sums may contain too many terms for exact evaluation in a reasonable time. Consequently, the computation of Bayesian models is an active research area. Two main approaches are numerical integration, especially by stochastic sampling, which is asymptotically exact, and deterministic approximations, which approximate the posterior distribution with simpler, tractable distributions.

The contributions in the publications are as follows. Publication I studies an application of sparse Bayesian linear regression using a spike and slab prior to genome-wide association analysis, where the aim is to identify genetic variants associated to blood cholesterol levels. The prior is extended to allow additive and dominant genetic effects, and the elicitation of hyperparameters based on previous studies is discussed. A Markov chain Monte Carlo (MCMC) algorithm with an automatically tuned proposal distribution is described for computation. Performance gains relative to a popular one-variant-at-a-time hypothesis testing approach are demonstrated in simulations, and the approach is applied to a real genome-wide dataset comprising of 3895 Finnish individuals. Publication II studies the computational approach in this application further and shows that the efficiency of the MCMC sampling can be increased by non-standard extensions of the basic Metropolis–Hastings algorithm for the spike and

slab models. A possible problem with a common Gaussian effect size prior is also demonstrated in the real dataset, while an alternative, more flexible prior seems to avoid the troubling behaviour. Publication III extends the applicability of a deterministic posterior approximation algorithm, expectation propagation, to approximate likelihood factors that depend on an inner product of multivariate Gaussian (or approximately Gaussian) random variables. Such models include, for example, Bayesian principal component analysis, among other linear latent variable models. The main contribution is to show how the analytically intractable multidimensional integrals can be reduced to one-dimensional problems, to be evaluated numerically. Publication IV presents an application of Bayesian survival regression for adverse cardiovascular event risk prediction in diabetic individuals using a set of new candidate biomarkers. Increased predictive performance is demonstrated on extending the model hierarchically to include data of non-diabetic individuals. A comparison of the Gaussian, Laplace, and horseshoe priors for the biomarker regression coefficients is presented, and finally, projective covariate selection is applied within cross-validation to rank the candidate biomarkers and examine the predictive performance of submodels corresponding to biomarker subsets.

Source codes for the software implemented as a part of this work are available at `https://github.com/to-mi/` and on the website of the research group: `http://becs.aalto.fi/en/research/bayes/bmagwa/` (Publications I and II), `http://becs.aalto.fi/en/research/bayes/epwx/` (Publication III), `http://becs.aalto.fi/en/research/bayes/diabcvd/` (Publication IV).

2. Bayesian Linear Models

This chapter describes the Bayesian formulation of generalized linear models in regression and latent variable models (Section 2.1), reviews approaches to encode assumptions about sparsity and correlation in the prior distribution for the linear model coefficients (Section 2.2), and briefly discusses covariate selection (Section 2.3). The main aim is to present background and recent developments as pertaining to the applications in genome-wide association analysis and cardiovascular event risk prediction. The literature on Bayesian linear models is too large for exhaustive coverage. The chapter begins with a brief definition of a *linear model*, and a discussion of the classical least squares solution to the multiple linear regression problem.

The defining feature of a linear model, as applied in this work, is the dependence of a probabilistic model on a linear combination, say, $p(y|\sum_j \beta_j x_j)$, where the probability distribution of y depends on the linear combination $\sum_j \beta_j x_j$ in some model specific way. Here, β_j and x_j, $j = 1, \ldots, m$, are model parameters or observed quantities. The linear combination is often written as an inner product: $\sum_j \beta_j x_j = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}$, where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_m]^{\mathrm{T}}$ and $\boldsymbol{x} = [x_1, \ldots, x_m]^{\mathrm{T}}$ are column vectors. As an example, in linear regression $\boldsymbol{\beta}$ are the regression coefficients and $\boldsymbol{x}$ are the observed covariate values. However, $\boldsymbol{x}$ may also be allowed to stand for unobserved or uncertain quantities, giving a bilinear dependence on the model parameters. A fixed $\boldsymbol{x}$ can also contain known nonlinear or constant terms (the canonical example being the polynomial $\boldsymbol{x} = [1, z, z^2, \ldots, z^m]$ of some observed quantity z). Nevertheless, all of these cases will be referred to as linear, although they may not correspond to a linear mapping in precise mathematical terms. The essential distinction in probabilistic models will be which of the quantities are modelled as random variables and which are taken as fixed values.

It is instructive to consider here the classical least squares solution to the multiple linear regression problem, as it illustrates some of the possible challenges in the application of linear models and, as such, forms a basis for much of the following discussion. Let $\boldsymbol{y} = [y_1, \ldots, y_n]^\mathrm{T} \in \mathbb{R}^n$ be n observed outcome values and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\mathrm{T} \in \mathbb{R}^{n \times m}$ the $n \times m$ matrix of observed covariate values for the n samples. The linear regression model associates the covariates to the outcome by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{r},$$

where $\boldsymbol{r} = [r_1, \ldots, r_n]^\mathrm{T}$, $r_i = \boldsymbol{y}_i - \boldsymbol{\beta}^\mathrm{T}\boldsymbol{x}_i$, are the residuals. The least squares solution for $\boldsymbol{\beta}$ minimizes the squared sum of the residuals, that is, $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} \boldsymbol{r}^\mathrm{T}\boldsymbol{r}$. Finding the root of the derivative with respect to $\boldsymbol{\beta}$ gives the classical solution $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathrm{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathrm{T}\boldsymbol{y}$ provided that $\boldsymbol{X}^\mathrm{T}\boldsymbol{X}$ is invertible. With this solution, the residuals $\hat{\boldsymbol{r}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ are orthogonal to the column space of $\boldsymbol{X}$, $\hat{\boldsymbol{r}}^\mathrm{T}\boldsymbol{X} = \boldsymbol{0}$: in other words, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the projection of $\boldsymbol{y}$ to the space spanned by $\boldsymbol{X}$, such that the Euclidean distance to $\boldsymbol{y}$ is minimized. The least squares solution coincides with the maximum likelihood solution when the observation model for $\boldsymbol{y}$ is the Gaussian $\mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, where σ^2 is the residual variance and $\boldsymbol{I}$ the $n \times n$ identity matrix, and with the posterior mean of a Bayesian model with the same observation model and improper, flat prior distribution for $\boldsymbol{\beta}$.

The greater the number of samples n, the more confidently the coefficients $\boldsymbol{\beta}$ can be identified. When y_i, for $i = 1, \ldots, n$, are assumed independent with known σ^2, the sampling variance of the least squares estimator of $\boldsymbol{\beta}$ is $\sigma^2(\boldsymbol{X}^\mathrm{T}\boldsymbol{X})^{-1}$ (see, e.g., Milton and Arnold [1995, p. 485]). This coincides with the posterior variance of $\boldsymbol{\beta}$ in the aforementioned Bayesian model (given σ^2). If the covariates are approximately uncorrelated and scaled to zero-mean and unit variance, $(\boldsymbol{X}^\mathrm{T}\boldsymbol{X})^{-1} \approx n^{-1}\boldsymbol{I}$, and the variance for a single β_j estimator is $\frac{\sigma^2}{n}$. On the other hand, if $\boldsymbol{X}$ contains nearly collinear (linearly dependent) columns, the corresponding β_j will be difficult to identify: for example, for two covariates with correlation ρ, the expected diagonal elements of $(\boldsymbol{X}^\mathrm{T}\boldsymbol{X})^{-1}$ scale in the inverse of $1 - \rho^2$. For a more geometrical intuition (following the idea in Gelman et al. [2014, p. 365–366]), one may consider a case when the data points fall near a line in the (x_1, x_2, y)-space: the data then provides little information on how to orient the regression plane about this line (Figure 2.1).

Especially interesting is the case when $m > n$, that is, there are more covariates than samples, as it nowadays frequently arises in applications

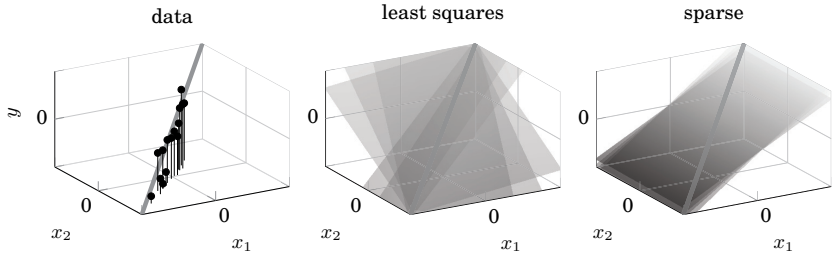

Figure 2.1. Left panel shows an example of 15 data points generated near the thick grey diagonal line in the (x_1, x_2, y)-space. The middle panel shows five regression planes fitted using least squares to five replicate datasets similar to the left panel. The right panel shows five posterior mean solutions when zero-mean Gaussian priors are placed on the two coefficients, with large variance for β_1 and small for β_2.

as mentioned in the Introduction. Then, the columns of $\boldsymbol{X}$ are necessarily linearly dependent, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is not invertible, and there is no unique solution for $\boldsymbol{\beta}$ (but an infinite number of solutions making the residuals zero are available when the rank of $\boldsymbol{X}$ is n). However, with further knowledge of the problem than immediately available in the observed $\boldsymbol{y}$ and $\boldsymbol{X}$, it may be possible to express preferences for or rule out certain kinds of solutions. A natural approach in the Bayesian framework is to encode the preferences, for example sparsity, in the prior distribution (Figure 2.1, right panel). Section 2.2 presents some approaches for constructing such prior distributions.

The minimization of the squared sum of residuals in regression analysis is a means to an end. The true aim is to make inferences that generalize beyond the available data, be they predictions for new samples or conclusions about the relevance of covariates. Even when $n > m$, one often has to deal with limitations in the informativeness of the data with regard to all the questions one would like to tackle with it: for example, covariate selection (Section 2.3) presents a combinatorial problem with the number of subsets of covariates being 2^m for m covariates. Bayesian hierarchical modelling provides a way to stabilize the inference process by sharing information across related model parameters via the prior structure.

2.1 Generalized Linear Models

The Gaussian observation model, $y_i \sim \mathrm{N}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2)$, is appropriate for continuous outcomes y_i taking positive and negative values, with normally distributed residuals around the mean $\mu_i = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i$. The framework of gen-

eralized linear models extends the family of observation models for other types of quantities $y_i \in \mathcal{Y}$ (see McCullagh and Nelder [1989] for an extensive coverage of the framework and, for example, Gelman et al. [2014, Chapter 16] for a Bayesian treatment). Below, the modelling of binary outcomes, where $y_i \in \{0, 1\}$, and event-free survival times, where $y_i \in (0, \infty)$, in addition to the continuous outcomes $y_i \in \mathbb{R}$, will be considered in particular as they are relevant for the publications.

The construction blocks of generalized linear models are

1. the linear combination $\eta_i = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i$,
2. the link function g connecting η_i and the mean μ_i of the outcome as $\eta_i = g(\mu_i)$,
3. the probability model $p(y_i|\mu_i, \phi)$ such that its mean $\mathrm{E}[y_i|x_i, \phi] = \mu_i$.

The link function is any function that is strictly monotonic, and, in particular, has the inverse mapping $g^{-1}(\eta_i) = \mu_i$. Often the probability model is chosen from the exponential family and the link function is differentiable, which makes finding the maximum likelihood estimate of the model parameters convenient (see McCullagh and Nelder [1989]). The optional parameter ϕ is often referred to as dispersion. An advantage of the general formulation from a Bayesian point of view is that the prior specification for $\boldsymbol{\beta}$ can follow the same guidelines for all models within the framework (although one needs to consider specific differences in the interpretation of η_i between models, e.g., scale; nevertheless, a monotonically increasing g guarantees that an increase in η_i always corresponds to an increase in μ_i).

2.1.1 Continuous Outcomes

For the Gaussian observation model, g is the identity function giving $\mu_i = \eta_i$, and the probability model is the namesake $p(y_i|\mu_i, \phi) = \mathrm{N}(y_i|\mu_i, \sigma^2)$ with $\phi = \sigma^2$. With conjugate priors for $\boldsymbol{\beta}$ and σ^2, a Gaussian linear regression model can be written as

$$
\begin{aligned}
2p(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2) &= \mathrm{N}(y_i|\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2), && \text{for } i = 1, \ldots, n, \\
p(\beta_j|\sigma^2) &= \mathrm{N}(\beta_j|0, \sigma^2\tau_j^2), && \text{for } j = 1, \ldots, m, \\
p(\sigma^2) &= \text{Inv-}\chi^2(\sigma^2|\nu, s^2), &&
\end{aligned}
$$

where $\tau_1^2, \ldots, \tau_m^2$, ν, and s^2 are hyperparameters that are either fixed or given prior distributions. The priors for $\boldsymbol{\beta}$ are further discussed in Section 2.2. The elicitation of ν and s^2 in a genome-wide association analysis application is considered in Publication I. The Gaussian observation model

is used in Publications I, II, and III. Another common observation model for continuous data is the t distribution, which has heavier tails than the normal distribution and can thus be more robust to outliers.

2.1.2 Binary Outcomes

Binary outcomes, for example disease status, can be modelled using the Bernoulli distribution: $p(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}$, where $y_i \in \{0, 1\}$ and μ_i is the probability that $y_i = 1$. There is no parameter ϕ. Now, the link function should be chosen such that $\mu_i = g^{-1}(\eta_i) \in (0, 1)$ where the linear combination η_i can generally take values along the whole real axis. Common link functions include the logistic (or logit) function, $g(\mu_i) = \log \frac{\mu_i}{1-\mu_i}$, and the probit function, $g(\mu_i) = \Phi^{-1}(\mu_i)$, where Φ is the cumulative distribution function of the standard normal distribution. The two functions are similar (when scaled appropriately) for $\mu_i \in (0.2, 0.8)$, while their tail behaviour differs. The models can be formulated using auxiliary variables u_i (with mean η_i), such that $y_i = 1$ if $u_i > 0$ and zero otherwise, and the distribution of u_i is normal for the probit model and the heavier tailed logistic for the logistic model. An advantage of the logistic model for interpretation is that changes in η_i correspond to changes in the log-odds. The probit model has convenient computational properties for Bayesian analysis, and it is used in Publication III. By reparametrizing y_i as $z_i = 2y_i - 1 \in \{-1, +1\}$, the probit regression model can be written as

$$\begin{aligned} p(z_i|\boldsymbol{x}_i, \boldsymbol{\beta}) &= \Phi(z_i \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i), && \text{for } i = 1, \ldots, n, \\ p(\beta_j) &= \mathrm{N}(\beta_j|0, \tau_j^2), && \text{for } j = 1, \ldots, m. \end{aligned}$$

The prior for $\boldsymbol{\beta}$ is conjugate for the auxiliary variable formulation.

2.1.3 Survival Time Outcomes

In survival analysis, the outcome is a survival time $y_i \in (0, \infty)$, where survival is understood loosely as the absence of some event of interest. For example in Publication IV, y_i is the time until a cardiovascular event from the beginning of the study. Survival analysis is often complicated by incomplete observation of the survival times (*censoring*). Perhaps the most common type of censoring is right-censoring: instead of observing y_i, one observes $t_i = \min(y_i, c_i)$, where c_i is the time of censoring. Observations may be censored, for example, because of not presenting the event before

the end of the follow-up in the study or because of death from an unrelated cause during the study. More generally, observations may also be left- or interval-censored. However, in the following only right-censoring is considered as it is the relevant type for the application in Publication IV. Ibrahim et al. [2001] provide for a book-length discussion of Bayesian survival analysis.

Common parametric probability models for survival analysis include the log-normal, log-logistic, and Weibull distributions. These can be formulated in the generalized linear model framework by taking g to be the logarithm function (possibly with an additive term independent of η_i) or by modelling the logarithm of the survival time with normal, logistic, or Weibull distribution with the identity link (and possible additive term). Alternatively, the models can be formulated as accelerated failure time models by considering density functions of form $p(\exp(\eta_i)y_i|\phi)\exp(\eta_i)$, where the effect of η_i is to lengthen or contract the time axis [Kalbfleisch and Prentice, 2002, Section 2.3.3].

However, it is perhaps more natural, at least in applications where the accurate prediction of the event time is difficult from the outset, to compare the implied *hazard* functions of the probability models. The hazard function is defined as $h(y_i|\eta_i,\alpha) = \frac{p(y_i|\eta_i,\alpha)}{S(y_i|\eta_i,\alpha)}$, where the model is parametrized using η_i and the shape parameter α instead of μ_i and ϕ. The *survivor* function $S(y_i)$ gives the probability of survival until y_i, and it is equal to the complementary cumulative distribution function of the observation model $p(y_i|\eta_i,\alpha)$. The hazard function obtains then an interpretation as the instantaneous risk of the event at y_i, having survived until that time. For the log-normal model, the hazard first increases until a maximum value and then decreases towards zero. For the log-logistic model, the hazard is decreasing if $\alpha \leq 1$ and behaves qualitatively similarly to the log-normal otherwise. The Weibull hazard function is decreasing if $\alpha < 1$, constant for $\alpha = 1$ (corresponding to exponential probability model), and increasing if $\alpha > 1$.

The Weibull model is used in the cardiovascular event risk prediction application in Publication IV. The formula for the hazard function is

$$h(y_i|\boldsymbol{x}_i,\boldsymbol{\beta},\alpha) = \alpha y_i^{\alpha-1}\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i),$$

which shows that it is a proportional hazards model: the ratio of hazards at any time y for samples i and j is $\frac{h(y|\boldsymbol{x}_i,\boldsymbol{\beta},\alpha)}{h(y|\boldsymbol{x}_j,\boldsymbol{\beta},\alpha)} = \exp\left(\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{x}_j)\right)$, which does not depend on y, but only on the effects of $\boldsymbol{x}$. The Weibull model is the only accelerated failure time model with this property [Kalbfleisch

and Prentice, 2002, Section 2.3.4]. The popular Cox proportional hazards model [Cox, 1972] has the hazard $h(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = h_0(y_i)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$, differing from the Weibull model in that the *baseline* hazard function h_0 is estimated non-parametrically (see Ibrahim et al. [2001, Chapter 3] and Joensuu et al. [2012] for Bayesian approaches). A Weibull regression model can be written as

$$\begin{aligned} p(t_i|\boldsymbol{x}_i, v_i, \boldsymbol{\beta}, \alpha) &= \alpha^{v_i} t_i^{v_i(\alpha-1)} \exp\left(v_i\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i - t_i^{\alpha}\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)\right), \quad \text{for } i = 1, \ldots, n, \\ p(\beta_j) &= \mathrm{N}(\beta_j|0, \tau_j^2), \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{for } j = 1, \ldots, m, \\ p(\log(\alpha)) &= \mathrm{N}(\log(\alpha)|0, s^2), \end{aligned}$$

where the indicator $v_i = 1$ if t_i corresponds to an observed event time, and $v_i = 0$ if t_i is a censoring time (for right-censored data), and a Gaussian prior is used for the shape parameter α on log-scale. For the censored observations ($v_i = 0$), it is only known that they survived until t_i. The likelihood function $p(t_i|\boldsymbol{x}_i, v_i = 0, \boldsymbol{\beta}, \alpha)$ then reduces to the survivor function, while $p(t_i|\boldsymbol{x}_i, v_i = 1, \boldsymbol{\beta}, \alpha)$ is the actual Weibull observation model.

2.1.4 Regression Models and Latent Variable Models

The modelling of $\boldsymbol{x}_i$ was not discussed above. In regression models, $\boldsymbol{x}_i$ are observed covariate values and modelling them as random variables is often not necessary. On the other hand, $\boldsymbol{x}_i$ are unobserved in linear latent variable models, and a prior distribution $p(\boldsymbol{x}_i|\boldsymbol{\psi})$ is required.

Linear regression models are fitted using a dataset of observed pairs of outcomes and covariates, $\mathcal{D} = \{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$, with the aim of explaining variation in the outcome using the covariates. When covariates are accurately and fully observed, regression modelling does not require specifying the observation model for the covariates, $p(\boldsymbol{x}_i|\boldsymbol{\psi})$, on the condition that the parameter $\boldsymbol{\psi}$ is a priori independent of $\boldsymbol{\beta}$ and ϕ (the outcome observation model parameters; following Gelman et al. [2014, p. 354]): when $p(\boldsymbol{\beta}, \phi, \boldsymbol{\psi}) = p(\boldsymbol{\beta}, \phi)p(\boldsymbol{\psi})$ and the observation models are as above, the Bayes' theorem gives the posterior distribution $p(\boldsymbol{\beta}, \phi, \boldsymbol{\psi}|\mathcal{D}) = p(\boldsymbol{\beta}, \phi|\mathcal{D})p(\boldsymbol{\psi}|\mathcal{D})$. A model for the covariates is thus not needed for studying the association of the outcome and the covariates or for making predictions for new samples. In particular, the models described in the previous sections are as such sufficient for regression. However, this holds only if there is no uncertainty in the covariate values. Otherwise the uncertainty should be accounted for in the model (see Gustafson [2004] for accounting for measurement uncertainty in covariates and Gelman et al.

[2014, Chapter 18] for missingness).

Latent variable models aim to model variation in multivariate observations $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{im}]^{\mathrm{T}}$ using unobserved latent variables $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{iK}]^{\mathrm{T}}$: the observed dataset is $\mathcal{D} = \{\boldsymbol{y}_i : i = 1, \ldots, n\}$ and the model seeks to explain dependencies between the observed variables. Consider the following linear latent variable model:

$$
\begin{aligned}
p(y_{ij}|\boldsymbol{x}_i, \boldsymbol{\beta}_j) &= \mathrm{N}(y_{ij}|\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i, \sigma_j^2), \qquad &\text{for } i = 1, \ldots, n \text{ and } j = 1, \ldots, m,\\
p(\boldsymbol{\beta}_j) &= \mathrm{N}(\boldsymbol{\beta}_j|0, \tau_j^2\boldsymbol{I}), \qquad &\text{for } j = 1, \ldots, m,\\
p(\boldsymbol{x}_i) &= \mathrm{N}(\boldsymbol{x}_i|0, \boldsymbol{I}), \qquad &\text{for } i = 1, \ldots, n,
\end{aligned}
$$

which corresponds to a probabilistic formulation of principal component analysis (when $\sigma_j^2 = \sigma^2$ for all j) or factor analysis (when σ_j^2 are allowed to differ) (see Bishop [1999b] and Bishop [2006, p. 580–586]). Integrating over $\boldsymbol{x}_i$ gives $p(\boldsymbol{y}_i|\boldsymbol{B}, \boldsymbol{\Sigma}) = \mathrm{N}(\boldsymbol{y}_i|\boldsymbol{0}, \boldsymbol{\Sigma} + \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}})$, where $\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m]^{\mathrm{T}}$ and $\boldsymbol{\Sigma}$ is a diagonal matrix with elements σ_j^2. When K is small compared to m, as is usually the case, this is essentially a low-rank model for the covariance of $\boldsymbol{y}_i$. Depending on the application, the latent variables $\boldsymbol{x}_i$ may have an intuitive interpretation, or they may only function as an abstract, low-dimensional, smoothed representation of $\boldsymbol{y}_i$. The generalized linear model framework can be applied to extend the observation models to other than the Gaussian. Moreover, latent variable models can also be extended, for example, to model covariance between multiple *views* of a dataset (canonical correlation analysis, group factor analysis; see Klami et al. [2013]), to perform two-way clustering [Hochreiter et al., 2010], or, as a predictive model, to impute missing values. The model structure of the errors-in-variables regression is also similar, with the covariates $\boldsymbol{x}_i$ inaccurately observed. Approximate computation in linear latent variable models of this form is considered in Section 3.2 and Publication III.

2.2 Prior Distributions for the Linear Model Coefficients

This section elaborates on the specification of prior distributions for the linear model coefficients $\boldsymbol{\beta}$. First, approaches to encoding the assumption of sparsity or shrinkage in the prior distribution are reviewed. Second, joint modelling of subgroups in the data, such that each group has their own $\boldsymbol{\beta}$ parameter, but the parameters are tied in the prior distribution, is considered.

2.2.1 Sparsity and Shrinkage

There are two prevalent Bayesian approaches to encoding sparsity or shrinkage in the prior. The first uses a mixture prior, which allocates the coefficients into two components: in-model (non-zero coefficients) and out-of-model (zero or near zero coefficients). The second approach does not explicitly try to divide the coefficients into relevant and irrelevant, but uses a prior distribution with substantial amount of mass near zero to shrink apparently irrelevant coefficients. The approaches are sometimes distinguished by the prior being absolutely continuous (latter) or not (former, although not all formulations).

Spike and Slab Priors
The two-component mixture prior can be formulated with latent binary indicator variables γ_j as

$$p(\beta_j|\gamma_j = 1) = \mathrm{N}(\beta_j|0, \tau_j^2),$$
$$p(\beta_j|\gamma_j = 0) = \delta_0(\beta_j),$$

where δ_0 is the Dirac delta function. A priori the probability that $\beta_j = 0$ is equal to the prior probability of $\gamma_j = 0$. Following Mitchell and Beauchamp [1988] the prior is referred to as the spike and slab prior, although their formulation used a uniform prior on the slab and directly specified the probability of the spike at $\beta_j = 0$ using a parameter which was not given a prior distribution. George and McCulloch [1993] introduced the latent variables γ_j and put independent Bernoulli priors on them. Instead of the Dirac delta, George and McCulloch [1993] use a normal distribution with small variance, such that coefficients with $\gamma_j = 0$ are deemed practically irrelevant, although they are not constrained to equal zero exactly.

Some essential features of spike and slab priors can be illustrated by considering the Bayes factor (the ratio of marginal likelihoods) or posterior odds between submodels, which differ in a single γ_j. Assuming the normal linear regression model with the conjugate prior,

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \mathrm{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}),$$
$$p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma}) = \mathrm{N}(\boldsymbol{\beta}|\boldsymbol{0}, \sigma^2\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}),$$
$$p(\sigma^2) = \text{Inv-}\chi^2(\sigma^2|\nu, s^2),$$

makes computing the marginal likelihoods analytically tractable. Here, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ is assumed diagonal with elements $\tau_0^2 + \tau_1^2$ for all β_j with $\gamma_j = 1$ and

τ_0^2 for β_j with $\gamma_j = 0$ (if $\tau_0^2 = 0$, $\mathrm{N}(\beta_j|0,0)$ is taken to denote $\delta_0(\beta_j)$). The marginal likelihood $p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\gamma})$ is given by the multivariate t distribution density function with $\nu+n$ degrees of freedom, zero mean and scale matrix $s^2(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{X}^{\mathrm{T}})$. The Bayes factor for the larger model against the smaller is

$$\mathrm{BF} = \frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\gamma}_{-k},\gamma_k=1)}{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\gamma}_{-k},\gamma_k=0)} = \left(\frac{S_1^2+\nu s^2}{S_0^2+\nu s^2}\right)^{-\frac{\nu+n}{2}} \left(\frac{\det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_1\boldsymbol{X}^{\mathrm{T}})}{\det(\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^{\mathrm{T}})}\right)^{-\frac{1}{2}},$$

where $\boldsymbol{\gamma}_{-k}$ refers to the configuration of $\boldsymbol{\gamma}$ excluding γ_k, $S_1^2 = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{I} + \boldsymbol{X}\boldsymbol{\Sigma}_1\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{y}$ with $\boldsymbol{\Sigma}_1$ being the $\boldsymbol{\beta}$ prior scale matrix for $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{-j}, \gamma_k = 1)$. S_0^2 and $\boldsymbol{\Sigma}_0$ are similarly defined for $\gamma_k = 0$. The posterior odds between the models is $\frac{p(\boldsymbol{\gamma}_{-k},\gamma_k=1|\boldsymbol{y},\boldsymbol{X})}{p(\boldsymbol{\gamma}_{-k},\gamma_k=0|\boldsymbol{y},\boldsymbol{X})} = \frac{p(\boldsymbol{\gamma}_{-k},\gamma_k=1)}{p(\boldsymbol{\gamma}_{-k},\gamma_k=0)} \times \mathrm{BF}$. Noting that $\boldsymbol{X}\boldsymbol{\Sigma}_1\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^{\mathrm{T}} + \boldsymbol{x}_k\tau_1^2\boldsymbol{x}_k^{\mathrm{T}}$, where $\boldsymbol{x}_k$ is vector of values of the kth covariate for the n samples, and using the matrix determinant lemma and the Sherman–Morrison formula [Golub and Van Loan, 2013, p. 65], the contribution of the kth covariate can be teased out:

$$\mathrm{BF} = \left(1 - \frac{\frac{\tau_1^2}{1+\tau_1^2 c}\left((\boldsymbol{y}-\bar{\boldsymbol{y}}_0)^{\mathrm{T}}\boldsymbol{x}_k\right)^2}{S_0^2+\nu s^2}\right)^{-\frac{\nu+n}{2}} (1+\tau_1^2 c)^{-\frac{1}{2}}, \tag{2.1}$$

where $c = \boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{x}_k - \boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{x}_k$ measures independent variation in $\boldsymbol{x}_k$ and $\bar{\boldsymbol{y}}_0 = \boldsymbol{X}(\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ is the posterior mean of $\boldsymbol{y}$ under the $(\boldsymbol{\gamma}_{-k}, \gamma_k = 0)$ model. For $\tau_0^2 = 0$, the formula holds with the kth covariate (and all other covariates with $\gamma_j = 0$) excluded from $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_0$.

Now, consider the case with $\tau_0^2 = 0$, $c = nc^*$ with $c^* > 0$, and $r = (\boldsymbol{y}-\bar{\boldsymbol{y}}_0)^{\mathrm{T}}\boldsymbol{x}_k = 0$ as $n \to \infty$, that is, the kth covariate does not have any explanatory power: then the first factor in BF is 1 and the second scales approximately as $n^{-\frac{1}{2}}$, and $\mathrm{BF} \to 0$. When $r^2 > 0$, the first factor scales as $a^{-\frac{\nu+n}{2}}$, $a < 1$, and consequently $\mathrm{BF} \to \infty$. Referring to this kind of behaviour, George and McCulloch [1997] interprets the Dirac spike formulation of the spike and slab prior as excluding only covariates whose coefficients cannot be distinguished from zero (as opposed to considering the practical significance of covariates). For another illustration, consider BF as a function of only τ_1^2 (ignoring the possible dependence of $\boldsymbol{\Sigma}_0$ on τ_1^2 for simplicity). If $\tau_1^2 \to 0$, then $\mathrm{BF} \to 1$ as the two models become the same. On the other hand, when $\tau_1^2 \to \infty$ (and $c > 0$), $\mathrm{BF} \to 0$ (Lindley's paradox; see also, e.g., Geweke [1996], Smith and Kohn [1996] for this observation in spike and slab models). The behaviour of BF is, of course, not monotone in $\tau_1^2 \in (0,\infty)$ as shown in Figure 2.2. Nevertheless, increasing τ_1^2 will decrease the posterior odds in the tail. This highlights that the

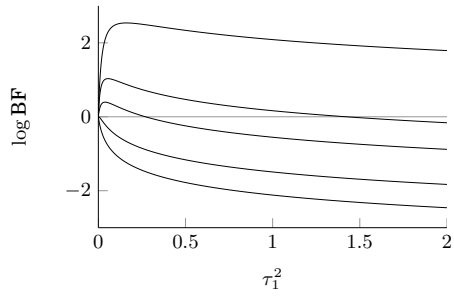

Figure 2.2. Example of the effect of τ_1^2 on the Bayes factor. The lines show the $\log \mathrm{BF}$ against the null model as a function of τ_1^2 for five covariates in the conjugate Gaussian model ($\nu = 1$, $s^2 = 1$, $n = 50$). The data were generated by drawing five covariate vectors $\boldsymbol{x}_j$ and their coefficients β_j from the standard normal distribution and adding normal-distributed noise with variance 20 in the linear model to get the y_i.

prior on τ_1^2, in addition to the prior on $\boldsymbol{\gamma}$, will be important in determining what kind of covariate effects the model will distinguish from zero. A striking empirical demonstration of this is seen in Publication II.

When the 2^m different configurations of the indicator vector $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_m]$ are interpreted as submodels, posterior expectations in the spike and slab model correspond to Bayesian model averaging over this model space (see, e.g., Raftery et al. [1996], Clyde [1999]). Following Raftery et al. [1996], the variance of the model averaged posterior predictive distribution for $\tilde{y}$ (given covariate values $\tilde{\boldsymbol{x}}$) can be decomposed as

$$\mathrm{var}_{\tilde{y}|\mathcal{D}}[\tilde{y}] = \mathrm{E}_{\boldsymbol{\gamma},\boldsymbol{\theta}|\mathcal{D}}[\sigma^2] + \mathrm{E}_{\boldsymbol{\gamma}|\mathcal{D}}[\mathrm{var}_{\boldsymbol{\theta}|\boldsymbol{\gamma},\mathcal{D}}\,[\bar{y}_{\boldsymbol{\theta},\boldsymbol{\gamma}}]] + \mathrm{var}_{\boldsymbol{\gamma}|\mathcal{D}}[\bar{y}_{\boldsymbol{\gamma}}],$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, $\bar{y}_{\boldsymbol{\theta},\boldsymbol{\gamma}} = \mathrm{E}_{\tilde{y}|\boldsymbol{\theta},\boldsymbol{\gamma},\mathcal{D}}[\tilde{y}]$ is the mean prediction given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and $\bar{y}_{\boldsymbol{\gamma}} = \mathrm{E}_{\tilde{y},\boldsymbol{\theta}|\boldsymbol{\gamma},\mathcal{D}}[\tilde{y}]$ is the mean prediction given $\boldsymbol{\gamma}$. The first term on the right hand side represents sampling uncertainty, the second represents within-model parameter uncertainty, and the last represents model uncertainty. If a single model (corresponding to some configuration of $\boldsymbol{\gamma}$), such as the highest posterior probability model, would be used for the prediction instead of the model average, the last term in the decomposition would be ignored completely. Raftery et al. [1996] give examples of the importance of accounting for model uncertainty in survival regression (see Draper [1995] for a more general discussion of uncertainty related to model structure).

A common approach to model space prior specification is to take $p(\boldsymbol{\gamma}|\boldsymbol{\pi}) = \prod_j \mathrm{Bernoulli}(\gamma_j|\pi_j)$. When there are no a priori dependencies in $\boldsymbol{\gamma}$, the prior can be simplified by setting $\pi_j = \pi$ for all j. This implies that the cardinality of $\boldsymbol{\gamma}$ (or *model size*), $|\boldsymbol{\gamma}| = \sum_{j=1}^m \gamma_j$, has a binomial prior dis-

tribution: $|\gamma| \sim \text{Bin}(m, \pi)$. Prior knowledge on the number of relevant covariates (in regression context) can then be used to guide the specification of π. Kohn et al. [2001] provide equations for setting a beta distribution prior on π (leading to a beta-binomial marginal prior distribution for $|\gamma|$) based on the prior expected mean and variance of the cardinality (see Publication I for application in genome-wide association analysis). Scott and Berger [2010] show that the beta-binomial prior performs multiplicity adjustment in the sense that the larger the number of covariates m, the larger the prior penalty from adding a covariate (the penalty also depends on the size of the base model). On the other hand, any fixed choice of π independent of m cannot adjust for the multiplicity. A simple choice would be to set $\pi = 0.5$, which gives all configurations of γ equal prior probabilities. However, because there are a large number of models with size around $\frac{m}{2}$ relative to the number of models with small or large cardinalities, the prior distribution is concentrated around models of size $\frac{m}{2}$ as can be verified from the binomial distribution. The prior expected model size with the beta-binomial prior also scales linearly in m, unless the parameters of the beta prior depend on m (in particular, with unit parameters the expected model size is $\frac{m}{2}$, but with a larger variance than with fixed $\pi = 0.5$; see Wilson et al. [2010]).

There are multiple ways to extend the product Bernoulli prior to account for structure in γ. For example, Publication I considers a case with multiple possible functional forms per covariate, Chipman [1996] considers polynomial and interaction terms (among others), and Li and Zhang [2010] extends the prior to a Markov random field, which allows specifying dependencies between covariate inclusion using a graph structure. One issue with the product Bernoulli distribution is that having multiple *proxies* for the same effect, in extreme, copies of a single covariate or more generally collinear covariates, will increase the prior probability of including at least one of the proxies in the model [George, 1999]. George [2010] proposes solving the issue with model space priors that dilute the prior mass among similar models.

As discussed above, the prior distribution of the slab, $p(\beta_j|\gamma_j = 1)$, is important in determining what kind of effects the model will consider as non-zero. A common approach is to set it to a normal distribution or a scale mixture of normal distributions (see Gianola et al. [2009], Guan and Stephens [2011], Kärkkäinen and Sillanpää [2012], Zhou et al. [2013], Publications I and II for some examples and discussion in genetics appli-

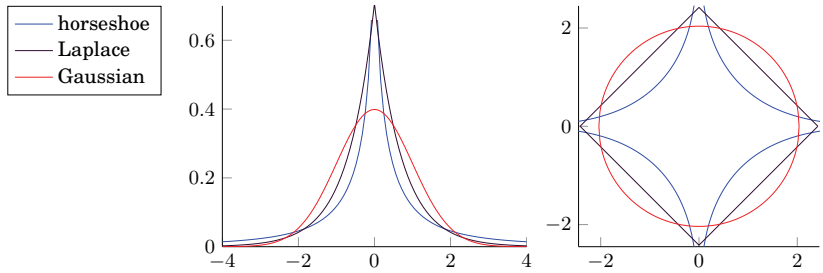

Figure 2.3. One-dimensional densities (left) and contours in two-dimensions (right) for the horseshoe, Laplace and Gaussian priors (marginalized over λ_1^2, λ_2^2 with $\tau^2 = 1$).

cations; an excellent discussion of the interplay of the model size prior and the normal distribution slab, with illustrations in economics datasets, is provided by Ley and Steel [2009]). Normal scale mixture distributions are discussed in the next section.

Normal Scale Mixture Priors

Many shrinkage priors that do not rely on discrete mixture components can be represented as normal scale mixtures:

$$p(\beta_j|\lambda_j^2, \tau^2) = \mathrm{N}(\beta_j|0, \lambda_j^2\tau^2) \quad \text{and} \quad \lambda_j^2 \sim p(\lambda_j^2), \text{ for } j = 1, \ldots, m,$$

where $p(\lambda_j^2)$ is some distribution for the local (specific to each j) variance parameters. τ^2 is a global variance parameter, which is shared for all j. Common examples include

$$\begin{aligned} \text{Laplace (e.g., Park and Casella [2008])} \quad & \lambda_j^2 \sim \text{Exponential}, \\ \text{Student's } t \text{ [Tipping, 2001]} \quad & \lambda_j^2 \sim \text{Inv-}\chi^2, \\ \text{normal-gamma [Griffin and Brown, 2010]} \quad & \lambda_j^2 \sim \text{Gamma}, \\ \text{horseshoe [Carvalho et al., 2010]} \quad & \lambda_j \sim \text{Half-Cauchy}. \end{aligned}$$

For more examples, see references in Polson and Scott [2011]. The Gaussian prior is also a special case with $\lambda_j^2 = 1$. One- and two-dimensional densities are illustrated in Figure 2.3 for the horseshoe, Laplace and Gaussian priors, which span a representative range of sparsity-inducing behaviours.

Assuming the least squares estimate is well-defined, the shrinkage effect of the priors can be contrasted to it (see, e.g., Carvalho et al. [2010], Griffin and Brown [2010], Polson and Scott [2011]). Let $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, and consider a regression model with the Gaussian likelihood and the observations transformed to $\hat{\boldsymbol{\beta}}$, that is, $p(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \sigma^2) = \mathrm{N}(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$. If

the prior is $p(\boldsymbol{\beta}|\boldsymbol{\lambda}^2, \tau^2, \sigma^2) = \mathrm{N}(\boldsymbol{\beta}|\mathbf{0}, \sigma^2\tau^2\boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with elements λ_j^2, the posterior distribution of $\boldsymbol{\beta}$ is

$$
\begin{aligned}
p(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda^2}, \tau^2, \sigma^2) &= \mathrm{N}(\boldsymbol{\beta}|\boldsymbol{m}, \boldsymbol{\Sigma}), \\
\boldsymbol{m} &= (\boldsymbol{I} - (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}(\tau^2\boldsymbol{\Lambda} + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})^{-1})\hat{\boldsymbol{\beta}}, \\
\boldsymbol{\Sigma} &= \sigma^2(\tau^{-2}\boldsymbol{\Lambda}^{-1} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}.
\end{aligned}
$$

Simplifying by taking $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \boldsymbol{I}$, $m_j = (1 - \frac{1}{1+\tau^2\lambda_j^2})\hat{\beta}_j$ and $\Sigma_{jj} = \sigma^2\frac{\tau^2\lambda_j^2}{1+\tau^2\lambda_j^2}$. Studying the shrinkage factor $\frac{1}{1+\tau^2\lambda_j^2}$, Polson and Scott [2011] suggest choosing a prior for τ^2 with a substantial amount of mass near zero to facilitate overall shrinkage, and choosing the λ_j^2 prior to have heavy tails so that some β_j can escape the shrinkage. The formula for the posterior variance shows that small τ^2 will also contract the posterior for near zero coefficients (i.e., for those with λ_j^2 also small).

Carvalho et al. [2010] plot the implied prior distributions of the shrinkage factors for multiple priors (with $\tau^2 = 1$), which shows that the advocated horseshoe prior has a considerable amount of mass near the extremes (no shrinkage and full shrinkage), while, for example, the Laplace prior density decays to zero at the no shrinkage extreme. More formally, they also show that the amount of shrinkage effected by the horseshoe prior decays to zero when the observation $\hat{\beta}_j \to \infty$. On the other hand, such result does not hold for the Laplace prior, suggesting that it may overshrink large coefficients (see also Griffin and Brown [2010], Polson and Scott [2011]). The comparison of Gaussian, Laplace and horseshoe priors in a survival regression application presented in Publication IV shows that the Gaussian and Laplace priors shrink the strongest coefficient more than the horseshoe, while the posterior uncertainty in near-zero coefficients is larger for the Gaussian and Laplace priors, as expected from above discussion for a sparse situation. The amount of shrinkage in a spike and slab prior with a Gaussian slab also does not vanish for a large observation $\hat{\beta}_j$ [Griffin and Brown, 2010], suggesting that heavier tailed priors may be more appropriate also for the slab (although the global variance parameter τ^2 does not need to be responsible for inducing overall sparsity there; see also Publication II).

The half-t family of distributions has been suggested as appropriate weakly informative priors for scale parameters in hierarchical models [Gelman, 2006]. Following this, Polson and Scott [2012] advocate the special case of half-Cauchy prior for the global scale parameter τ as a good default choice also in shrinkage models. In particular, the common con-

jugate inverse-χ^2 (or inverse-gamma) prior on τ^2 is not appropriate for inducing shrinkage as its density approaches zero for $\tau^2 \to 0$.

Shrinkage and sparsity penalties are also an active research topic in non-Bayesian regularized regression. Many of the methods have a connection to the Bayesian approaches as maximum a posteriori solutions to the regression problem, and the regularization terms can be represented as $-\sum_j \log \int p(\beta_j|\lambda_j^2, \tau^2)p(\lambda_j^2)d\lambda_j^2$. For example, the Gaussian prior leads to ridge regression and the Laplace prior to lasso regression [Tibshirani, 1996]. Both are very popular approaches and have efficient algorithms for computation as the optimization problems are convex. The lasso regression can perform variable selection by producing exact zeroes as solutions of some coefficients. However, the properties and theoretical results for the regularization approaches do not directly extend to fully Bayesian analysis.

2.2.2 Hierarchical Modelling of Subgroups

Consider having observations for multiple subgroups in regression, or having multiple datasets $\mathcal{D}^{(l)} = \{(y_i^{(l)}, \boldsymbol{x}_i^{(l)}) : i = 1, \ldots, n^{(l)}\}$, for $l = 1, \ldots, L$, where the same type of outcome and same covariates are observed in each dataset (i.e., $(y_i^{(l)}, \boldsymbol{x}_i^{(l)}) \in \mathcal{Y} \times \mathcal{X}$ for all i and l), but the regression functions are not necessarily expected to be identical. When the L subproblems are related, it can make sense to model them jointly. Especially subgroups with a small number of observations $n^{(l)}$, having limited amount of information to fit the regression, could benefit from sharing information across related subproblems.

Hierarchical modelling provides a natural approach to formulate such joint models (see Gelman and Hill [2007] and Gelman et al. [2014, Chapter 15] for broader coverage). Let the subproblems, at the first level of the hierarchy, follow a generalized linear model $p(y_i^{(l)}|\boldsymbol{x}_i^{(l)}, \boldsymbol{\beta}^{(l)}, \phi^{(l)})$. The parameters of the model are tied at the second level using a joint prior distribution $p(\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(L)})$ (more generally, the $\phi^{(l)}$ could also be included in the joint prior). A common approach is to take $(\boldsymbol{\beta}^{(l)}|\boldsymbol{w}) \sim \mathrm{N}(\boldsymbol{Z}_l\boldsymbol{w}, \boldsymbol{\Sigma}_l)$, where $\boldsymbol{Z}_l$ are group-level characteristics for the lth group, $\boldsymbol{w}$ are group-level coefficients, and $\boldsymbol{\Sigma}_l$ is a covariance matrix (see, e.g., Seltzer et al. [1996] and Gelman and Hill [2007]). If $\boldsymbol{w} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_w)$, the marginal prior for the $\boldsymbol{\beta}^{(l)}$, integrated over $\boldsymbol{w}$, is $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma} + \boldsymbol{Z}\boldsymbol{\Sigma}_w\boldsymbol{Z}^{\mathrm{T}})$, where $\boldsymbol{\beta}$ is a stacked vector of the $\boldsymbol{\beta}^{(l)}$, $l = 1, \ldots, L$, $\boldsymbol{Z}$ is a stacked matrix of the $\boldsymbol{Z}_l$, and $\boldsymbol{\Sigma}$ is a block diagonal matrix with blocks $\boldsymbol{\Sigma}_l$. The group-level characteris-

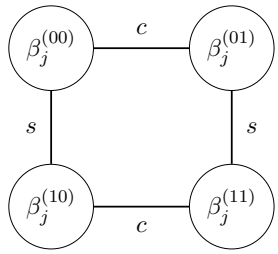

Figure 2.4. Graphical structure of the prior for the jth regression coefficient. Modified from Figure 1 in Publication IV.

tics induce correlations between the $\boldsymbol{\beta}^{(l)}$ in the prior.

An alternative approach, in some cases more convenient and economical in parameters, is to directly specify a structure for the covariance matrix $\boldsymbol{\Sigma}_\beta$ of the marginal prior $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$. Focusing on a single coefficient in all of the L submodels, $\boldsymbol{\beta}_j = [\beta_j^{(1)}, \ldots, \beta_j^{(L)}]^{\mathrm{T}}$, Publication IV proposes setting $\boldsymbol{\Sigma}_{\beta,j} = r_j^2 \lambda \boldsymbol{\Lambda}^{-1}$, where $\lambda \boldsymbol{\Lambda}^{-1}$ is a correlation matrix and r_j^2 is a variance parameter, which can be used to define shrinkage (as in Section 2.2.1) for the jth regression coefficient jointly across the subgroups. In the setting of Publication IV, where diabetes status and sex of the individuals are used to define four subgroups, the precision matrix $\boldsymbol{\Lambda}$ is chosen to encode the Markov random field structure illustrated in Figure 2.4 by taking

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1+c+s & -c & -s & 0 \\ -c & 1+c+s & 0 & -s \\ -s & 0 & 1+c+s & -c \\ 0 & -s & -c & 1+c+s \end{bmatrix},$$

where c and s are non-negative parameters that control the correlation between adjacent submodels (λ, as a function of c and s, is used to scale $\boldsymbol{\Lambda}^{-1}$ to a correlation matrix). The precision matrix $\boldsymbol{\Lambda}$ can also be thought of as the Laplacian matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ of the graph in Figure 2.4, where $\boldsymbol{D}$ is a diagonal matrix of node degrees, defined[1] as $\mathrm{degree}(l) = 1 + c + s$, and $\boldsymbol{W}$ is a weighted adjacency matrix with weights as given in the figure (following Liu et al. [2014], who use a more general form of a graph Laplacian inspired prior to learn the regularization dependence structure in covariates). For more details and the full prior specification, including priors on c and s, see Publication IV.

Although the approach is above arrived at in terms of hierarchical mod-

[1]This is a slightly unorthodox definition, but it is similar to that by Liu et al. [2014]. More commonly the degree would be only the sum of edge weights $c + s$ (see, for example, Luxburg [2007]).

elling of subgroups, it could also be formulated as a multi-task or transfer learning problem [Pan and Yang, 2010]. In particular, the approach can be seen as a specific Bayesian version of the multi-task graph regularization proposed by Evgeniou et al. [2005] and further studied by Sheldon [2008]. The negative logarithm of the prior on $\boldsymbol{\beta}_j$, dropping terms independent of $\boldsymbol{\beta}_j$, can be written as $\frac{1}{2r_j^2\lambda}(S_2 + cS_c + sS_s)$, where $S_2 = (\beta_j^{(00)})^2+(\beta_j^{(01)})^2+(\beta_j^{(10)})^2+(\beta_j^{(11)})^2$, $S_c = (\beta_j^{(00)}-\beta_j^{(01)})^2+(\beta_j^{(10)}-\beta_j^{(11)})^2$ and $S_s = (\beta_j^{(00)} - \beta_j^{(10)})^2 + (\beta_j^{(01)} - \beta_j^{(11)})^2$, which is similar to the multi-task regularization penalty. This form also emphasizes the role of c and s as controlling the prior similarity of the subgroup models. The approach can also be formulated as a version of the sparse Bayesian multi-task model proposed by Archambeau et al. [2011], where a (zero-mean) matrix-variate Gaussian prior is placed on $\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m]$ with row covariance $\boldsymbol{\Omega}$ (over the m covariates) and column covariance $\boldsymbol{\Sigma}$ (over the L tasks). Taking $\boldsymbol{\Omega}$ diagonal with diagonal elements r_j^2 and $\boldsymbol{\Sigma} = \lambda\boldsymbol{\Lambda}^{-1}$ gives the prior above.

2.3 Covariate Selection

Covariate selection refers to the problem of choosing a subset of the available covariates according to some criterion. The motivation for the selection may be, for example, simplicity, identification of relevant covariates, or reduction of costs related to obtaining covariate values in the future. The sparsity and shrinkage priors do not address this problem fully as they will not lead to exact point mass posteriors at zero for any of the covariate effects. Such priors, however, may be practically enough to rule out or select some covariates. For example, the posterior inclusion probabilities in spike and slab models, $p(\gamma_j = 1|\mathcal{D})$ for $j = 1, \ldots, m$, can be used to rank the covariates according to their posterior relevance, and this strategy might be enough to choose candidates for further study (e.g., replication or molecular biology experiments in genetics applications). Nevertheless, a more formal approach to covariate selection can be beneficial.

Covariate selection can be formulated as a model selection problem within decision theory (see Kadane and Lazar [2004] and Vehtari and Ojanen [2012] for reviews of Bayesian model selection). The formulation entails defining a space of possible decisions (here, covariate subsets) and a utility function, the expected value of which is estimated to assess the decisions.

In practice, the definitions are often given only loosely as formulating utility functions that carefully balance the benefits and costs of the decisions can be difficult and not worth the effort in many studies.

Two goals can be distinguished in the analysis[2]: 1) covariate selection for structure and 2) covariate selection for prediction. In the first case, the aim is to find covariates associated to the outcome variable (such that the associations generalize beyond the observed dataset), which provides information about the structure of the system under study. Here, the decision does not need to correspond to a selection of a model (with a subset of the covariates and their parameters), but can also be a selection of a subset of the covariates (without any reference to model parameters). The decision theoretic formulation can be complicated as the *true* structure may be unobservable (at least at the level of the available data) or even difficult to properly define (and, in particular, often almost certainly not included within the candidate models). Nevertheless, statistics assessing model fit, such as Bayes factors, posterior model probabilities, or posterior inclusion probabilities of covariates in spike and slab models, can be motivated from decision theoretic perspective (see, e.g., Kadane and Lazar [2004], Barbieri and Berger [2004], Vehtari and Ojanen [2012]) and used to perform covariate selection. Furthermore, simulation studies, which posit a certain data generating mechanism (and, in particular, structure) that hopefully captures some features of how the actual observed data arises, can be used to evaluate the utility of using a certain model and to compare utilities between models in different imaginary scenarios. A common approach in spike and slab linear models is to select covariates whose posterior inclusion probabilities reach a certain threshold, and evaluate the utility of the selection based on the true positive and false positive rates for the identification of true associations in simulations (see, e.g., Publication I, Guan and Stephens [2011]). Guan and Stephens [2011] also study the calibration of the posterior inclusion probabilities, which is important when choosing a suitable threshold to select covariates, for example, for follow-up studies. On the other hand, the sensitivity of the inferences to and the behaviour of the covariate selection under random perturbations of the data can be evaluated without reference to a ground-truth (Publication I and Guan and Stephens [2011]).

[2]Consideration of formal causal analysis is omitted, although one often hopes that the inferences can be used to speculate or generate hypotheses of causal associations. See Shmueli [2010] for discussion contrasting predictive and causal statistical modelling.

In the second case, the main aim is the prediction of unobserved (but observable) outcome values given the covariates. Vehtari and Ojanen [2012] provide an extensive review of methods for model selection from the Bayesian predictive perspective. The utility function is in this case chosen to measure the predictive performance (e.g., squared error or logarithm of predictive density) and related costs. Another main ingredient for practical application is a method for estimating the expected utility. Here, the possibilities include using the training dataset, an independent test dataset, cross-validation, or a reference model for the outcome distribution. The first gives biased estimates of the generalization performance as the same data is used to train and test the models. An independent dataset (w.r.t. the training data) does not have this problem, but it may often not be available unless a part of the available dataset is omitted from the training data. Cross-validation follows this idea, while aiming for more efficient use of the available data, by partitioning the dataset into multiple parts and testing on one part at a time while training on the others. The reference model approach uses some model $p(y|\boldsymbol{x})$ as a reference measure over which the expected utility of another model can be computed. As the model comparisons are based on the reference model, there is no further fitting to the training dataset in the selection phase [Vehtari and Lampinen, 2004, Vehtari and Ojanen, 2012]. On the other hand, the reference model needs to be a good model for the outcomes for it to make sense to evaluate the expected utilities over it. The encompassing model or the model average are candidates for the reference model in linear regression as they include the uncertainties related to all the covariates (see Dupuis and Robert [2003], Vehtari and Lampinen [2004], and Publication IV for examples). Finally, the model space is often too large, with size 2^m for m covariates if all subsets are allowed, for exhaustive enumeration of the expected utilities for all models, and some search strategy (e.g., stochastic sampling, stepwise selection methods) or theoretical optimality results are needed. Related to spike and slab linear models, Barbieri and Berger [2004] show that the median probability model (i.e., the model where covariates with posterior inclusion probabilities ≥ 0.5 are included) is the optimal predictive model under the squared error loss when using the model average as reference (with strong conditions; essentially requiring orthogonality of the covariates).

Comparing models based on predictive performance does not require specifying prior probabilities on the model space. The projection method

of Goutis and Robert [1998], Dupuis and Robert [2003] for generalized linear models takes a further step and does not require prior specifications for model parameters in the submodels but only on an encompassing reference model. The parameters of a submodel are defined by a Kullback–Leibler (KL) divergence[3] based projection of the reference model parameters to the restricted parameter space of the submodel:

$$\boldsymbol{\theta}_{\perp}^{*} = \operatorname*{argmin}_{\boldsymbol{\theta}_{\perp}} \int \mathrm{KL}\left[p(y|\boldsymbol{\theta}, \boldsymbol{x}) \parallel p(y|\boldsymbol{\theta}_{\perp}, \boldsymbol{x}_{\perp})\right] p(\boldsymbol{x}) dx,$$

where $p(y|\boldsymbol{\theta}, \boldsymbol{x})$ is the observation model of the reference model with parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi) \in \mathbb{R}^m \times \mathbb{R}^+$ and $\boldsymbol{x} = (\boldsymbol{x}_{\perp}, \boldsymbol{x}_{\top}) \in \mathbb{R}^m$, and $p(y|\boldsymbol{\theta}_{\perp}, \boldsymbol{x}_{\perp})$ is the observation model in the restricted space $\boldsymbol{\theta}_{\perp} = (\boldsymbol{\beta}_{\perp}, \phi_{\perp}) \in \mathbb{R}^k \times \mathbb{R}^+$, $k < m$, where the covariates $\boldsymbol{x}_{\top}$ have been omitted. $p(\boldsymbol{x})$ is the distribution of the covariates, the integral over which is in practice approximated by summing over the available dataset. Dupuis and Robert [2003] propose performing the projection for each Markov chain Monte Carlo sample of the reference model posterior distribution and using the posterior mean of the KL divergence to guide the covariate selection.

In Publication IV, the projection approach is applied in the Weibull survival model. A forward selection search is used to traverse the model space and provides a ranking of the candidate biomarkers. Using a threshold for the KL divergence to choose a single submodel seems difficult, as the divergence slowly decreases towards zero as more biomarkers are added and the selection would be sensitive to the arbitrary threshold. Publication IV examines applying the projective search within cross-validation to estimate and compare the predictive performance of the submodels along the forward selection path.

A proper Bayesian approach to covariate selection should account for the uncertainty in dropping covariates. The projection approaches may be able to account for this in a limited fashion as the reference model posterior distribution includes the uncertainties about the regression coefficients of all covariates, and the submodels are defined with respect to it. However, this feature seems currently under-studied. Lindley [1968] considers covariate selection in linear regression with a model $p(\boldsymbol{x})$ for the covariates. The uncertainty in leaving out some covariates is then accounted for by integrating over the conditional distribution of the left-out covariates given the in-model covariates. However, building good models for covariates is often difficult especially in high-dimensional cases.

[3]The KL divergence from p to q is defined as $\mathrm{KL}[p \parallel q] = \int p(y) \log \frac{p(y)}{q(y)} dy$.

3. Computational Approaches

Computing normalized posterior distributions, predictive distributions, and summaries, such as expectations, of the posterior distributions requires calculating integrals that are analytically tractable only in limited special cases, which, in particular, generally exclude sparse and shrinkage models. In discrete parameter spaces, even if the required sums would be composed of tractable terms, there can be too many terms for direct evaluation in a reasonable time. To apply sparse Bayesian modelling in complex, real-world problems, efficient methods to deal with the integrals are needed.

Well-behaved, low-dimensional analytically intractable integrals can often be efficiently calculated using quadrature or cubature rules, such as the composite Simpson's rule or Gaussian quadrature, which evaluate the integrand at a finite, deterministic grid of points. The error in the estimate is of order $S^{-\frac{c}{d}}$, where S is the number of function evaluations, d the dimensionality of the integral, and c a constant [Haber, 1970]. For a fixed error tolerance, the required number of evaluation points grows exponentially with d, making quadrature impractical for high-dimensional problems.

Remarkably, using stochastic evaluation points, Monte Carlo integration, to estimate the integral can bypass the curse of dimensionality. As direct sampling from the posterior distribution is often not possible, there exists various Monte Carlo methods that use surrogate distributions to facilitate sampling (see, e.g., Robert and Casella [2004]). The (probabilistic) error in the estimate is independent of the dimension d and is proportional to $\frac{1}{\sqrt{S}}$ (with certain conditions). However, methods that use a global surrogate sampling distribution can still be practically limited by the curse of dimensionality as constructing good surrogate distributions, necessary for efficient implementation, can be difficult for high-dimensional prob-

lems. Markov chain Monte Carlo (MCMC) methods alleviate the problem by constructing the samples in a Markov chain, such that the samples are generated from a surrogate distribution that is allowed to depend on the previous sample. This obviates the need for a globally good surrogate distribution, albeit at the cost of inducing dependencies along the sample chain. Section 3.1.1 describes MCMC sampling for spike and slab models, the computation of which is studied in Publications I and II. Section 3.1.2 briefly discusses MCMC approaches for hierarchical models with continuous parameters.

An alternative to numerical integration is to approximate the posterior distribution within a simpler, tractable family of distributions, and compute posterior summaries over the approximating distribution. Advantages over MCMC methods include often faster computation and easier convergence diagnostics, albeit there is generally no guarantee that the fitted approximation will be good, whereas MCMC can, in principle, made arbitrarily accurate by running longer chains. An early and popular approximation method is the Laplace's method, which expands the logarithm of the posterior distribution at its mode using a second order Taylor series, giving a Gaussian approximation centred at the posterior mode (e.g., O'Hagan and Forster [2004, Section 9.10]). The Gaussian approximating form can also be motivated based on the asymptotic normality of posterior distributions (on some conditions; see, e.g., O'Hagan and Forster [2004, Section 3.20]) and on the tractability of the multivariate normal distribution. However, with limited data, centring the approximation at the mode may not capture well the bulk of the posterior distribution (see, e.g., Bishop [2006, Chapter 10]), and alternative methods, such as variational Bayes and expectation propagation (EP), that are motivated by minimizing the Kullback–Leibler divergence between the approximation and the true posterior distribution may fare better. These methods can also use more general families of approximating distributions than Gaussian (especially other exponential family distributions). EP for inner product factors and spike and slab priors is reviewed in Section 3.2. Extending the applicability of EP to the former is the main contribution in Publication III.

3.1 Markov Chain Monte Carlo

MCMC methods work by producing a chain of samples $\boldsymbol{\theta}^{(s)}, s = 1, \ldots, S$, with the Markov structure $p(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(s)}) = p(\boldsymbol{\theta}^{(s+1)}|\boldsymbol{\theta}^{(s)})$, such that the stationary distribution of the chain is the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ and the average of some function f computed over the chain, $\frac{1}{S}\sum_s f(\boldsymbol{\theta}^{(s)})$, converges to the expectation of f, $\mathrm{E}_{p(\boldsymbol{\theta}|\mathcal{D})}[f(\boldsymbol{\theta})]$ (as $S \to \infty$ assuming the expectation exists). The convergence relies on the theory of Markov chains and appropriately designed algorithms to generate the sample chain. Extensive and more formal treatments of MCMC methods are provided, for example, by Tierney [1994], O'Hagan and Forster [2004, Chapter 10], and Robert and Casella [2004]. A brief introduction is given below.

The Metropolis–Hastings (MH) algorithm [Metropolis et al., 1953, Hastings, 1970] is the basis of many MCMC methods. A new sample to the chain is generated in two steps:

1. Propose a new state $\boldsymbol{\theta}^*$ from a conditional distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(s)})$.
2. Set $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^*$ with the *acceptance* probability

$$a(\boldsymbol{\theta}; \boldsymbol{\theta}^*) = \min\left(1, \frac{p(\boldsymbol{\theta}^*|\mathcal{D})q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}|\mathcal{D})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right).$$

   Otherwise, set $\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)}$.

A simple example of a MH algorithm is the Gaussian random walk algorithm for $\boldsymbol{\theta} \in \mathbb{R}^m$ with the proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \sigma_q^2\boldsymbol{I})$. The Gibbs sampler [Geman and Geman, 1984] is another popular algorithm, especially for conditionally conjugate Bayesian models. There, the model parameters are divided into m subcomponents $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$, and the new sample is generated by drawing each subcomponent in turn from its conditional posterior distribution $p(\boldsymbol{\theta}_j^{(s+1)}|\boldsymbol{\theta}_{-j}^{(s')}, \mathcal{D})$, where $\boldsymbol{\theta}_{-j}^{(s')}$ denotes the other subcomponents than jth at their $(s+1)$th or sth state depending on if they have or have not been sampled at the current cycle. Each draw can be seen as a special case of a MH proposal, where the acceptance probability turns out to be 1.

Following the results in Robert and Casella [2004], the MH algorithm produces a Markov chain with the transition kernel (or transition probabilities if the state space is finite)

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}'|\boldsymbol{\theta})a(\boldsymbol{\theta}, \boldsymbol{\theta}') + (1 - r(\boldsymbol{\theta}))\delta_{\boldsymbol{\theta}}(\boldsymbol{\theta}'),$$

where $r(\boldsymbol{\theta}) = \int q(\boldsymbol{\theta}'|\boldsymbol{\theta})a(\boldsymbol{\theta}, \boldsymbol{\theta}')d\boldsymbol{\theta}'$ and δ_x is point mass at x. The MH kernel K satisfies the detailed balance condition with the posterior distribution $p(\cdot|\mathcal{D})$ (or equivalently reversibility) $K(\boldsymbol{\theta}, \boldsymbol{\theta}')p(\boldsymbol{\theta}|\mathcal{D}) = K(\boldsymbol{\theta}', \boldsymbol{\theta})p(\boldsymbol{\theta}'|\mathcal{D})$,

which ensures that $p(\cdot|\mathcal{D})$ is the invariant distribution of K. Convergence of the MH algorithm requires further that K is irreducible and aperiodic. Irreducibility refers to the ability of the Markov chain to reach with positive probability any set of positive posterior probability from any starting point in the state space. Aperiodicity denies the possibility of the chain from deterministically cycling between disjoint subsets of the state space, which is usually trivially satisfied by MH algorithms that reject some proposals. The Gibbs sampler is not necessarily reversible when the updates are performed in a deterministic order, but can nevertheless be proven to lead to a valid procedure.

Apart from choosing an efficient algorithm, a practical difficulty in the application of MCMC is to know whether the sampling has converged to a stationary phase and how many samples should be collected for estimating the posterior summaries of interest. In Publications I, II, and IV, the convergence was assessed by visually inspecting MCMC parameter traces and by computing potential scale reduction factors, which compare the within-chain variances of parameters to pooled-chains variance to assess mixing between multiple MCMC chains started from different initial points [Gelman and Rubin, 1992]. The diagnostics can only find problems with convergence, but cannot ensure it. The asymptotic variance of the sample average estimate over the MCMC chain (assuming the central-limit theorem holds) is $\frac{\sigma^2}{S}(1 + 2\sum_{s=1}^{\infty} \rho_s)$, where σ^2 is the posterior variance of the estimated parameter and ρ_s is the lag-s autocorrelation [Geyer, 1992]. Comparing to independent sampling motivates the definition of the effective sample size for S dependent MCMC samples as $\text{ESS} = \frac{S}{1+2\sum_{s=1}^{\infty} \rho_s}$. Geyer [1992] provides methods to estimate the autocorrelation time in the denominator. ESS can also be used to compare the efficiency of MCMC methods, although the computational costs of the methods should also be accounted for (see, e.g., Lamnisos et al. [2009] and Publication II).

3.1.1 Spike and Slab Linear Models

In spike and slab models (Section 2.2.1), the computation of the normalizing constant of the posterior distribution $Z = \sum_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}) p(\mathcal{D}|\boldsymbol{\gamma})$, where the sum runs over all configurations of the binary variables in $\boldsymbol{\gamma}$, and the computation of posterior expectations $\mathrm{E}[f(\boldsymbol{\gamma})] = \sum_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma}) p(\boldsymbol{\gamma}|\mathcal{D})$ form a bottleneck: the number of terms in the sum is 2^m, where m is the number of covariates, and grows infeasible for exhaustive computation for m around

20–30. Sampling with MCMC can, however, be feasible for much larger m.

Let the parameters of the model be $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is a division of other parameters than $\boldsymbol{\gamma}$ such that the marginal $p(\boldsymbol{\gamma}|\boldsymbol{\theta}_1, \mathcal{D}) = \int p(\boldsymbol{\gamma}, \boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathcal{D}) d\boldsymbol{\theta}_2$ is conveniently computable. For example, in normal linear regression models with the conjugate prior, one might take $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}, \sigma^2)$. The following overall Gibbs-like sampling scheme can then be used to generate $\boldsymbol{\theta}^{(s+1)}$ given $\boldsymbol{\theta}^{(s)}$ and is used in Publications I and II:

1. Sample $\boldsymbol{\theta}_1^{(s+1)}$ from the conditional distribution $p(\boldsymbol{\theta}_1|\boldsymbol{\gamma}^{(s)}, \boldsymbol{\theta}_2^{(s)}, \mathcal{D})$.
2. Sample $\boldsymbol{\gamma}^{(s+1)}$ from the conditional distribution $p(\boldsymbol{\gamma}|\boldsymbol{\theta}_1^{(s+1)}, \mathcal{D})$, where $\boldsymbol{\theta}_2$ has been marginalized.
3. Sample $\boldsymbol{\theta}_2^{(s+1)}$ from the conditional distribution $p(\boldsymbol{\theta}_2|\boldsymbol{\gamma}^{(s+1)}, \boldsymbol{\theta}_1^{(s+1)}, \mathcal{D})$.

The last two steps can be seen as a factorized draw from the joint conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(s+1)}, \mathcal{D})$. Sampling of $\boldsymbol{\theta}_2$ may not be needed at all, if the conditional distribution in step 1 is independent of it (and posterior inference on $\boldsymbol{\theta}_2$ is not of interest). Such *collapsed* schemes can increase the sampling efficiency, and are discussed in more generality by van Dyk and Park [2008]. Moreover, separating the sampling of the discrete parameters $\boldsymbol{\gamma}$ and the continuous parameters, for example in normal linear regression by taking $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}, \sigma^2)$, is convenient in avoiding trans-dimensional proposals [Green, 1995] when the Dirac spike is used. Non-collapsed Gibbs sampling for non-conjugate normal spike model was introduced by George and McCulloch [1993]. Later, George and McCulloch [1997] compared it to collapsed sampling in the conjugate model and found the latter advantageous for the computation.

The following discussion focuses on sampling with the Dirac spike and a conjugate slab, such that at least $\boldsymbol{\beta}$ can be integrated out in step 2 (although to avoid trans-dimensional updates, it is enough to be able to marginalize a single β_j when updating a (γ_j, β_j)-pair; see Geweke [1996] for a Gibbs sampling algorithm). It is assumed that standard sampling methods can be applied in steps 1 and 3, and these will not be discussed.

Step 2 is often implemented by proposing changes to a single γ_j variable or a block of γ_j variables at a time using Gibbs or Metropolis–Hastings sampling. In both cases the main computational burden is usually evaluating the Bayes factor (BF) between the current and proposed states (e.g., equation (2.1) in Section 2.2.1). Gibbs sampling proceeds either in random or fixed order over $\boldsymbol{\gamma}$, sampling each new γ_j from its conditional Bernoulli distribution [Smith and Kohn, 1996, George and McCulloch, 1997]. The

basic Metropolis–Hastings approach proceeds by selecting one γ_j at uniformly random and proposing a change to its state (the original formulation of Madigan and York [1995], Raftery et al. [1997] also allows proposing the current model). A second strategy, often better in large model spaces and sparse settings, is to generate the proposals by first selecting whether to add, remove, or, possibly, swap states of two variables before selecting which variable(s) (see especially Lamnisos et al. [2009] and also Guan and Stephens [2011], Publications I and II). Otherwise the sampling process will mainly evaluate additions that are rarely accepted. Similarly, Kohn et al. [2001] study metropolized versions of Gibbs sampling, where the draw from the conditional distribution of γ_j is replaced with a MH step that proposes the new state based on the prior. When the proposed state is equal to the old, it can avoid the expensive evaluation of the marginal likelihood $p(\mathcal{D}|\boldsymbol{\gamma}^*)$ (or BF). Publications I and II use the second MH approach and consider some extensions of it, which are described below.

Adaptive Proposal Distributions

The efficiency of the Metropolis–Hastings algorithm depends on the proposal distribution. Adaptive methods aim to tune the proposals during the sampling to maximize the efficiency. This, however, destroys the Markov structure of the sample chain, and requires non-standard theory for proving the convergence to correct target distribution. Although the theory for adaptive MCMC has been developing in the recent years (see, e.g., Rosenthal [2011] for a review), a simple and safe approach is to adapt the proposal distribution only during an initial sampling period and fix the distribution before collecting samples for posterior inference (automatic tuning or finite adaptation; see, e.g., Pasarica and Gelman [2010]). Then, as long as the final transition kernel is valid, convergence results follow from the standard theory.

In spike and slab models with large model spaces, it can make sense to try to focus the sampling effort on variables with greatest uncertainty of their inclusion: if some variable, say jth, should be almost certainly included (or excluded), it should almost always stay at $\gamma_j = 1$ ($\gamma_j = 0$) in the sample chain, whereas a variable with greater uncertainty should flip between inclusion and exclusion. Nott and Kohn [2005] provide conditions for convergence of adaptive samplers on finite state space, and propose a metropolized Gibbs sampler, where the proposal for changing the state of γ_j is formed based on the mean and covariance of $\boldsymbol{\gamma}$ in the past sam-

ple chain. Publications I and II use finite adaptation, with selection of which variable to consider for update selected according to the marginal inclusion probabilities that are updated during an initial phase of the sampling. A simplified, random-scan version of the Nott–Kohn adaptive sampler (without estimation of covariance which can be difficult in large model space) can be seen to be similar to the sampler in Publication II, as discussed in the supplementary material of the publication. Sampling based on estimated marginal inclusion probabilities is used also in the Bayesian Adaptive Sampling algorithm of Clyde et al. [2011], which samples the model space without replacement (and, consequently, is not an MCMC algorithm).

The algorithms discussed above have not been proven optimal in any rigorous fashion. The Nott–Kohn algorithm is motivated by modelling the conditional density in Gibbs sampling by a linear approximation. The algorithm in Publication I (and consequently Publication II) was partly inspired by the non-adaptive use of single variable association statistics of Guan and Stephens [2011] to form their proposal distribution. Clyde et al. [2011] also show that the product of the marginal inclusion probabilities form the closest product model to the true posterior distribution in Kullback–Leibler divergence sense.

Block Proposals

Simultaneously updating a block of $\boldsymbol{\gamma}$, say $(\gamma_{j_1}, \ldots, \gamma_{j_k})$ for some block size k, in Gibbs sampling can improve the mixing of the Markov chain especially with correlated variables. However, the block update requires computation of the marginal likelihoods of 2^k models, which can be considerably more expensive than k single updates, although George and McCulloch [1997] note that speed-ups are possible using matrix decomposition update algorithms and careful consideration of the order in which the marginal likelihoods of the models are evaluated.

The Metropolis–Hastings algorithm can also be made to propose changing multiple variables in one proposal (e.g., George and McCulloch [1997]). Proposing swaps of the states of two variables $(\gamma_{j_1}, \gamma_{j_2})$ in areas of correlated covariates is a useful special case, for example, in genetics applications, where the variants along the genome are known to be correlated [Guan and Stephens, 2011], and is used in Publications I and II. More generally, Lamnisos et al. [2009] demonstrate empirically that block proposals can increase sampling efficiency (with uniform proposal distribution

for selecting which variables to update). Guan and Stephens [2011] also propose using occasional multistep proposals, but provide little details or analysis of the gains.

Publication II constructs block proposals with (finitely) adaptive proposal distribution in two steps: 1) sample block size k from geometric distribution, 2) sample sequentially which variables to update. The latter step can be further divided into an iteration of two steps: 1) uniformly in random choose to add or remove a variable, 2) sample which variable to add (or remove) among those out of the model (in the model) with probabilities proportional to the adaptively estimated marginal inclusion probabilities (marginal exclusion probabilities). The sequential sampling is formalized in the auxiliary variable framework of Storvik [2011] to demonstrate its validity. For more details, see Publication II. The results there show that using the adaptive proposal distribution may increase the efficiency of block proposals over those with uniform sampling. This seems sensible as using large k tends to decrease the MH acceptance rate, but the adaptive proposal distributions may be able to counteract this to some extent by generating better proposals.

As the acceptance rate of the proposals can be sensitive to the block size k, Publication II uses finite adaptation to automatically tune the parameter of the geometric distribution by optimizing the expected squared jumping distance $\mathrm{E}[\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|^2]$ over the transition kernel [Pasarica and Gelman, 2010]. Lamnisos et al. [2013] propose to adapt the block size proposals by acceptance rate coercion. This is simpler than jump distance optimization, but requires specifying a target acceptance rate, theoretical optimal values of which are only available for certain simple problems.

Advanced alternative ways to generate large block updates in spike and slab models include the Evolutionary Stochastic Search [Bottolo and Richardson, 2010], Parallel Hierarchical Sampling [Rigat and Mira, 2012] and Swendsen–Wang [Nott and Green, 2004] algorithms. These can be especially beneficial to tackle problems with multimodality. The first two run multiple interacting chains and can be seen as complementary approaches to the sampling methods described above as they also use local within-chain updates.

Delayed rejection

Delayed rejection is an extension of the Metropolis–Hastings algorithm, which continues the proposal process on rejections [Tierney and Mira,

1999, Mira, 2001]. If a proposal to move from $\boldsymbol{\gamma}^{(s)}$ to $\boldsymbol{\gamma}^*$ is rejected, instead of setting $\boldsymbol{\gamma}^{(s+1)} = \boldsymbol{\gamma}^{(s)}$, a new proposal is made from a proposal distribution $q(\boldsymbol{\gamma}'^*|\boldsymbol{\gamma}^{(s)}, \boldsymbol{\gamma}^*)$ that is allowed to depend on the first proposal. The MH acceptance probability of the second proposal is constructed to satisfy the detailed balance. The rationale for delaying the rejection comes from the decreased probability of staying put in combination with the result of Peskun [1973], which states that increasing the off-diagonal elements of the transition probability matrix leads to lower asymptotic variance of expectations. The increase in computation in the continued proposals is a possible caveat.

Delayed rejection is used for rejected block proposals in Publication II. The idea is to take advantage of the computations required in evaluating the marginal likelihood of the k-block update in order to compute the full set of 2^k marginal likelihoods for the models involved in the block, similar to discussed by George and McCulloch [1997] for block Gibbs sampling. A second stage proposal is then performed with this knowledge at hand. Details are given in the publication and its supplementary material.

3.1.2 Continuous-parameter Hierarchical Models

Gibbs sampling and random walk Metropolis–Hastings algorithms are popular approaches for Bayesian hierarchical models in continuous parameter spaces, such as the generalized linear models with normal scale mixture priors discussed in Chapter 2. Yet, they can be severely inefficient in cases with strongly correlated parameters or for densities with a peaked high probability area and an extensive low probability tail. Such pathologies are illustrated and discussed, for example, by Neal [2011] and Betancourt and Girolami [2013]. Reparametrizations or a well-chosen random walk proposal covariance matrix may help, but there may not be any reparametrization or covariance, which would work well in the whole parameter space.

The Hamiltonian Monte Carlo (HMC) algorithm has been suggested as an alternative approach (see Neal [2011] for a review). HMC is an auxiliary variable MH algorithm, where the proposals are generated by simulating Hamiltonian dynamics in an augmented state space, where the (unnormalized) posterior distribution is interpreted as inducing a potential energy function $U(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta}|\mathcal{D})$ and the distribution of the auxiliary variables $\boldsymbol{\xi}$ a kinetic energy function $K(\boldsymbol{\xi})$. It is usually assumed that $\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{M})$, where $\boldsymbol{M}$ is some covariance matrix. The Hamiltonian

of the system is $H(\boldsymbol{\theta}, \boldsymbol{\xi}) = U(\boldsymbol{\theta}) + K(\boldsymbol{\xi})$, and the Hamilton's equations are

$$\frac{d\theta_j}{dt} = \frac{\partial H}{\partial \xi_j},$$
$$\frac{d\xi_j}{dt} = -\frac{\partial H}{\partial \theta_j},$$

for $j = 1, \ldots, m$ when $\boldsymbol{\theta} \in \mathbb{R}^m$. The simulation of the dynamics keeps H constant. The HMC proposal is generated by first sampling a new $\boldsymbol{\xi}$ from its distribution and then simulating the dynamics for some time t. In practice, the simulation needs to be done numerically (except for simple special cases) and the arising error in H corrected for by a MH acceptance step.

The auxiliary variables function as momentum parameters to suppress the random walk behaviour and lead to better exploration of the posterior distribution. Theoretically, HMC has better scaling with dimensionality than the random walk MH algorithm, and it has much better behaviour for many cases, where random-walk MH fails (see Neal [2011] and Betancourt and Girolami [2013]). Parametrization and selection of the covariance matrix $\boldsymbol{M}$ can have considerable effect also on the efficiency of HMC. A recent further generalization of HMC, Riemannian manifold Hamiltonian Monte Carlo [Girolami and Calderhead, 2011], promises to help in this regard by exploiting local curvature information.

A hindrance to the use of HMC has been that the computational and statistical efficiency of the numerical simulation of the Hamiltonian dynamics can be sensitive to the discretization step size and trajectory length parameters. Moreover, the gradient of the logarithm of the posterior distribution needs to be available and can be laborious to calculate for complex hierarchical models. However, recently these problems have been mitigated by the No-U-Turn sampling (NUTS) algorithm [Hoffman and Gelman, 2014] and the Stan software [Stan Development Team, 2014] which implements the NUTS algorithm. NUTS can automatically choose the simulation trajectory length based on heuristics, while maintaining the detailed balance, and provides a finite adaptation approach to optimize the step size parameter. Stan uses an automatic differentiation algorithm to relieve the user from calculating the gradient and requires only the model description be given in its modelling language.

Stan is applied in Publication IV to sample from the posterior distribution of hierarchical survival models. The publication also gives the eigendecomposition of the joint prior discussed in Section 2.2.2, which can be useful in reparametrizing the model for HMC or other MCMC sampling

methods.

3.2 Expectation Propagation

Expectation propagation (EP) [Minka, 2001a,b] is an algorithm for constructing deterministic, exponential family approximations to posterior distributions $p(\boldsymbol{\theta}|\mathcal{D})$. The approximating distribution $q(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta}|\mathcal{D})$ should be close to the true posterior in some sense, and choosing in which sense different approximations with different properties can be obtained. Minimizing the Kullback–Leibler divergence $\mathrm{KL}[p(\boldsymbol{\theta}|\mathcal{D}) \parallel q(\boldsymbol{\theta})]$ with q in the exponential family would provide an approximation, which preserves the moments of the true posterior (up to those needed for characterizing q) [Herbrich, 2005]. However, the minimization is usually intractable when approximations are sought, as it would entail being able to compute the true posterior moments. EP makes a simplification leading to an algorithm for iteratively refining q by tractable KL projections. It has been found to provide useful approximations for many Bayesian models, including linear models with spike and slab [Hernández-Lobato et al., 2008, 2010b], Laplace [Seeger, 2008b], and horseshoe priors [Hernández-Lobato and Hernández-Lobato, 2013].

Consider a model, for which the posterior distribution is written as a product of terms $t_j(\boldsymbol{\theta})$[1],

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z}\prod_j t_j(\boldsymbol{\theta}). \tag{3.1}$$

For example, for the generalized linear models with conditionally independent observations $\mathcal{D} = \{y_i : i = 1, \ldots, n\}$ considered in Chapter 2, $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z}p(\boldsymbol{\theta})\prod_i^n p(y_i|\boldsymbol{\theta})$, and one could take $t_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ and $t_j(\boldsymbol{\theta}) = p(y_i|\boldsymbol{\theta})$ for all $i = j$. EP approximates the posterior with

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx q(\boldsymbol{\theta}) = \frac{1}{\tilde{Z}}\prod_j \tilde{t}_j(\boldsymbol{\theta}), \tag{3.2}$$

where the $\tilde{t}_j(\boldsymbol{\theta})$ are suitable exponential family form functions, which approximate the corresponding $t_j(\boldsymbol{\theta})$ terms in (3.1). In general, individual $\tilde{t}_j(\boldsymbol{\theta})$ terms need not be proper probability densities (the likelihood terms $t_j(\boldsymbol{\theta}) = p(y_i|\boldsymbol{\theta})$ are not necessarily either), as long as $q(\boldsymbol{\theta})$ is. For example, taking $\tilde{t}_j(\boldsymbol{\theta}) = \tilde{Z}_j \exp(-\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\tilde{\boldsymbol{\Gamma}}_j\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\tilde{\boldsymbol{\mu}}_j)$ gives a Gaussian form with mean

[1] Publication III, unfortunately, uses t_j to denote an approximate term, while using t_j for a term in the true posterior and $\tilde{t}_j$ for its approximation would be more standard notation and follow Minka [2001a].

$\tilde{\boldsymbol{m}}_j = \tilde{\boldsymbol{\Gamma}}_j^{-1}\tilde{\boldsymbol{\mu}}_j$ and covariance $\tilde{\boldsymbol{\Gamma}}_j^{-1}$, where $\tilde{\boldsymbol{\Gamma}}_j$ is not necessarily required to be positive definite. Putting the Gaussian term approximations $\tilde{t}_j(\boldsymbol{\theta})$ into (3.2) gives a multivariate normal approximation $q(\boldsymbol{\theta})$ with mean and precision matrix

$$\boldsymbol{m} = \tilde{\boldsymbol{\Gamma}}^{-1}\sum_j \tilde{\boldsymbol{\mu}}_j \quad \text{and} \quad \boldsymbol{\Gamma} = \sum_j \tilde{\boldsymbol{\Gamma}}_j. \tag{3.3}$$

It is a general property of exponential family densities that their functional form is closed under products and quotients. Seeger [2008a] describes properties of the exponential family distributions relevant for EP. A good overview of EP with Gaussian posterior approximation is given by Cseke and Heskes [2011, Appendix C]. The Gaussian case is considered in the following.

The EP algorithm is used to find the parameters $(\tilde{\boldsymbol{\mu}}_j, \tilde{\boldsymbol{\Gamma}}_j)$ characterizing the posterior approximation. The parallel EP algorithm [van Gerven et al., 2009] consists of the following steps:

1. Initialize the parameters $(\tilde{\boldsymbol{\mu}}_j, \tilde{\boldsymbol{\Gamma}}_j)$ of the term approximations $\tilde{t}_j(\boldsymbol{\theta})$ and calculate the full approximation $q(\boldsymbol{\theta})$ using (3.3).
2. For each site j:
   (a) Calculate the *cavity distribution* $q^{\setminus j}(\boldsymbol{\theta}) \propto \frac{q(\boldsymbol{\theta})}{\tilde{t}_j(\boldsymbol{\theta})}$. Gaussian approximation implies a Gaussian cavity distribution with the mean and precision matrix

   $$\boldsymbol{m}^{\setminus j} = (\boldsymbol{\Gamma}^{\setminus j})^{-1}(\boldsymbol{\Gamma}\boldsymbol{m} - \tilde{\boldsymbol{\mu}}_j) \quad \text{and} \quad \boldsymbol{\Gamma}^{\setminus j} = \boldsymbol{\Gamma} - \tilde{\boldsymbol{\Gamma}}_j. \tag{3.4}$$

   (b) Form the *tilted distribution* $\hat{p}(\boldsymbol{\theta}) \propto t_j(\boldsymbol{\theta})q^{\setminus j}(\boldsymbol{\theta})$, and find an updated approximation as $\hat{q}(\boldsymbol{\theta}) = \operatorname{argmin}_{\tilde{q}} \mathrm{KL}[\hat{p}(\boldsymbol{\theta}) \parallel \tilde{q}(\boldsymbol{\theta})]$, where $\tilde{q}$ is restricted to the approximating family. For the Gaussian approximation, $\hat{q}(\boldsymbol{\theta})$ is a Gaussian distribution with mean and covariance

   $$\hat{\boldsymbol{m}} = \mathrm{E}_{\hat{p}}[\boldsymbol{\theta}] \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \mathrm{E}_{\hat{p}}[(\boldsymbol{\theta} - \hat{\boldsymbol{m}})(\boldsymbol{\theta} - \hat{\boldsymbol{m}})^{\mathrm{T}}]. \tag{3.5}$$

   (c) Update the parameters of the site approximation $\tilde{t}_j(\boldsymbol{\theta})$ such that $\tilde{t}_j(\boldsymbol{\theta})q^{\setminus j}(\boldsymbol{\theta}) = \hat{q}(\boldsymbol{\theta})$. Here,

   $$\tilde{\boldsymbol{\mu}}_j = \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{m}} - \boldsymbol{\Gamma}^{\setminus j}\boldsymbol{m}^{\setminus j} \quad \text{and} \quad \tilde{\boldsymbol{\Gamma}}_j = \hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Gamma}^{\setminus j}. \tag{3.6}$$

3. Update the full approximation $q(\boldsymbol{\theta})$ using (3.3).
4. Repeat steps 2 and 3 until convergence.

The applicability of EP depends on the tractability of the above computations, where especially the computation of the moments of the tilted

distribution in step 2b is crucial. When a site $t_j(\boldsymbol{\theta})$ depends only on a part of the parameter vector $\boldsymbol{\theta}$ or on a linear combination (as in linear regression $t_j(\boldsymbol{\beta}) = p(y_i|\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)$), the representation of $\tilde{t}_j(\boldsymbol{\theta})$ and the moment computations can be cast into lower dimensionality (see, e.g., Cseke and Heskes [2011, Appendix C] and Seeger [2008b] for Gaussian EP). Posterior terms $t_j(\boldsymbol{\theta})$ that are directly of the same form as their approximation need not be processed when initialized as $\tilde{t}_j(\boldsymbol{\theta}) = t_j(\boldsymbol{\theta})$. EP can also provide an approximation of the marginal likelihood as $\tilde{Z} = \int \prod_j \tilde{t}_j(\boldsymbol{\theta}) d\boldsymbol{\theta}$, which requires also matching the zeroth moment to find $\tilde{Z}_j$ in step 2b.

The sequential EP algorithm [Minka, 2001a,b] differs from the above in that the full approximation is updated after every site update (i.e., step 3 above is performed as a fourth task in step 2). The parallel algorithm has been found to provide faster convergence in some problems (e.g., Seeger and Nickisch [2011]). Neither algorithm is guaranteed to convergence in general. The EP solution can be shown to correspond to a stationary point of an objective function [Minka, 2001b,c], and alternative optimization algorithms with provable convergence are available [Minka, 2001c, Heskes and Zoeter, 2002, Opper and Winther, 2005, Seeger and Nickisch, 2011]. Damping is a simple stabilizing modification, where the parameters of the site approximation (step 2c) are updated to a convex combination of the new and old values [Minka and Lafferty, 2002]. A more drastic modification is to force the variance parameters of the site approximations to positive, although it is expected to give a worse approximation than a converging unrestricted one [Minka, 2001a].

Publication III describes how to construct a Gaussian approximation with EP to the posterior distribution of linear latent variables models (Section 2.1.4). The main issue is approximating each likelihood term $p(y_{ij}|\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i, \phi)$ with a Gaussian $\tilde{t}(\boldsymbol{\beta}_j, \boldsymbol{x}_i)$ (assuming fixed ϕ), where $\boldsymbol{x}_i \in \mathbb{R}^K$ is the latent variable and $\boldsymbol{\beta}_j \in \mathbb{R}^K$ its coefficient vector. A sparse model can be achieved by placing a spike and slab prior on $\boldsymbol{\beta}_j$. The required EP updates (step 2) are described below. The likelihood and sparsity prior updates are decoupled, so the following likelihood updates are independent of the form of the prior on $\boldsymbol{\beta}_j$ and $\boldsymbol{x}_i$ as long as the approximation is Gaussian (similarly, the prior term update is independent of the likelihood). For a latent variable model with $p(\boldsymbol{x}_i) = \mathrm{N}(\boldsymbol{x}_i|\mathbf{0}, \boldsymbol{I})$, the corresponding term $\tilde{t}(\boldsymbol{x}_i)$ can be fixed to $p(\boldsymbol{x}_i)$.

3.2.1 Inner Product Terms

Consider the $(n+m)K$-dimensional Gaussian EP approximation $q(\boldsymbol{\beta}, \boldsymbol{x})$ for a model with nm likelihood terms $p(y_{ij}|\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i, \phi)$ that depend on $\boldsymbol{\beta}_j$ and $\boldsymbol{x}_i$ through their inner product. The approximate factors $\tilde{t}_{ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i)$ can be taken as $2K$-dimensional Gaussian terms, with the understanding that on applying equations (3.3) and (3.4) the site parameters need to be appropriately aligned with the parameters of the full approximation (see Cseke and Heskes [2011, Appendix C]). Publication III assumes a factorizing approximation $q(\boldsymbol{\beta}, \boldsymbol{x}) = \prod_j q(\boldsymbol{\beta}_j) \prod_i q(\boldsymbol{x}_i)$, with $\tilde{t}_{ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i) = \tilde{t}_{ij}(\boldsymbol{\beta}_j)\tilde{t}_{ij}(\boldsymbol{x}_i)$, which is a special case of the more general treatment below.

Step 2a

As a single likelihood term depends on only a single $\boldsymbol{\beta}_j$ and $\boldsymbol{x}_i$, the moment matching reduces to considering only their marginal distribution. The marginal $q^{\backslash ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i) = \mathrm{N}([\boldsymbol{\beta}_j^{\mathrm{T}}\ \boldsymbol{x}_i^{\mathrm{T}}]^{\mathrm{T}}|\boldsymbol{m}^{\backslash ij}, (\boldsymbol{\Gamma}^{\backslash ij})^{-1})$ of the cavity distribution $q^{\backslash ij}(\boldsymbol{\beta}, \boldsymbol{x})$, found by applying (3.4), is then needed. For the factorizing approximation, the marginal is conveniently available already in the precision matrix parametrization, as the full precision matrix is block diagonal and the diagonal blocks are equal to the precision matrices of the marginals.

Step 2b

Unfortunately, computing the moments (3.5) of the tilted distribution,

$$\hat{p}(\boldsymbol{\beta}_j, \boldsymbol{x}_i) \propto p(y_{ij}|\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i, \phi) q^{\backslash ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i),$$

seems analytically intractable already for a Gaussian likelihood, and involves $2K$-dimensional integrals ruling out naive quadrature in general.

Publication III proposes to use an integral transform of the Dirac delta function $\delta(\xi) = \frac{1}{2\pi}\int \exp(\imath t\xi) dt$ [Olver et al., 2010, p. 37–38] to rewrite the integrals. For example, for the normalization of the tilted distribution:

$$\begin{aligned}
\hat{Z} &= \int p(y_{ij}|\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i, \phi) q^{\backslash ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i) d(\boldsymbol{\beta}_j, \boldsymbol{x}_i) \\
&= \iint p(y_{ij}|f, \phi)\delta(f - \boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i) df q^{\backslash ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i) d(\boldsymbol{\beta}_j, \boldsymbol{x}_i) \\
&= \iint p(y_{ij}|f, \phi)\frac{1}{2\pi}\int \exp(\imath t(f - \boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i)) dt df q^{\backslash ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i) d(\boldsymbol{\beta}_j, \boldsymbol{x}_i).
\end{aligned}$$

At first sight, this formal representation of the integral using the Dirac delta transform seems to complicate rather than simplify the problem. To advance, the integration orders of f and t, and of t and $(\boldsymbol{\beta}_j, \boldsymbol{x}_i)$ are

changed:

$$\begin{aligned}\hat{Z} &= \frac{1}{2\pi}\iint p(y_{ij}|f,\phi)\exp(\imath tf)df\int \exp(-\imath t\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i))q^{\backslash ij}(\boldsymbol{\beta}_j,\boldsymbol{x}_i)d(\boldsymbol{\beta}_j,\boldsymbol{x}_i)dt \\ &= \frac{1}{2\pi}\iint L(t,f)df\int C(t,\boldsymbol{\beta}_j,\boldsymbol{x}_i)d(\boldsymbol{\beta}_j,\boldsymbol{x}_i)dt,\end{aligned}$$

where $L(t,f)$ is the integrand over f and $C(t,\boldsymbol{\beta}_j,\boldsymbol{x}_i)$ is the integrand over $(\boldsymbol{\beta}_j,\boldsymbol{x}_i)$. Conditions for the validity of the above manipulations are discussed in the supplementary material of Publication III in particular for the Gaussian and probit likelihoods (the latter is seen to require special care).

Now, the integrand $C(t,\boldsymbol{\beta}_j,\boldsymbol{x}_i)$ can be seen to be of unnormalized Gaussian form in $(\boldsymbol{\beta}_j,\boldsymbol{x}_i)$, with complex-valued mean and covariance. In particular, the mean and covariance for the concatenated variable $[\boldsymbol{\beta}_j^{\mathrm{T}}\ \boldsymbol{x}_i^{\mathrm{T}}]^{\mathrm{T}}$ are

$$\bar{\boldsymbol{m}}(t) = \bar{\boldsymbol{V}}(t)\boldsymbol{\Gamma}^{\backslash ij}\boldsymbol{m}^{\backslash ij} \quad \text{and} \quad \bar{\boldsymbol{V}}(t) = \left(\boldsymbol{\Gamma}^{\backslash ij} + \imath t\begin{bmatrix}\boldsymbol{0} & \boldsymbol{I}\\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\right)^{-1},$$

where $\boldsymbol{\Gamma}^{\backslash ij}$ is required to be positive definite for the complex generalization of the Gaussian distribution (see, e.g., Neretin [2011, p. 10–11]), and $\bar{\boldsymbol{V}}$ and, hence, $\bar{\boldsymbol{m}}$ are functions of t. The tractability of the Gaussian distribution allows analytical solutions to the integrals of $C(t,\boldsymbol{\beta}_j,\boldsymbol{x}_i)$ over $(\boldsymbol{\beta}_j,\boldsymbol{x}_i)$. Then, assuming $L(t) = \int L(t,f)df$ is analytically (or conveniently numerically) tractable, the tilted distribution moments are available as

$$\begin{aligned}\hat{Z} &= \frac{1}{2\pi}\int L(t)D(t)dt, \\ \hat{\boldsymbol{m}} &= \frac{1}{\hat{Z}}\frac{1}{2\pi}\int L(t)D(t)\bar{\boldsymbol{m}}(t)dt, \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{\hat{Z}}\frac{1}{2\pi}\int L(t)D(t)[\bar{\boldsymbol{V}}(t) + \bar{\boldsymbol{m}}(t)\bar{\boldsymbol{m}}(t)^{\mathrm{T}}]dt - \hat{\boldsymbol{m}}\hat{\boldsymbol{m}}^{\mathrm{T}},\end{aligned}$$

where

$$D(t) = |\boldsymbol{\Gamma}^{\backslash ij}|^{\frac{1}{2}}|\bar{\boldsymbol{V}}(t)|^{\frac{1}{2}}\exp(-\frac{1}{2}[(\boldsymbol{m}^{\backslash ij})^{\mathrm{T}}\boldsymbol{\Gamma}^{\backslash ij}\boldsymbol{m}^{\backslash ij} - (\boldsymbol{m}^{\backslash ij})^{\mathrm{T}}\boldsymbol{\Gamma}^{\backslash ij}\bar{\boldsymbol{V}}(t)\boldsymbol{\Gamma}^{\backslash ij}\boldsymbol{m}^{\backslash ij}]).$$

The integrals over the auxiliary variable t seem analytically intractable, but can often be solved efficiently numerically (with the caveat of possible problematic oscillatory behaviour of the integrands).

Efficient evaluation of the integrands as a function of t require efficient means of evaluating the factors $\bar{\boldsymbol{V}}(t)$, $|\bar{\boldsymbol{V}}(t)|$, $\boldsymbol{a}^{\mathrm{T}}\bar{\boldsymbol{V}}(t)\boldsymbol{a}$, and $\bar{\boldsymbol{V}}(t)\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\bar{\boldsymbol{V}}(t)$, where $\boldsymbol{a}$ is some vector independent of t. To this end, a convenient representation for $\bar{\boldsymbol{V}}(t)$ is $\bar{\boldsymbol{V}}(t) = \boldsymbol{S}(\boldsymbol{I}+\imath t\boldsymbol{\Lambda})^{-1}\boldsymbol{S}^{\mathrm{T}}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix and $\boldsymbol{S}$ is such a matrix that $\boldsymbol{S}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{I}\\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\boldsymbol{S} = \boldsymbol{\Lambda}$ and $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{\Gamma}^{\backslash ij}\boldsymbol{S} = \boldsymbol{I}$. Here,

$\mathbf{\Gamma}^{\setminus ij}$ is required to be positive definite, and $\boldsymbol{S}$ and $\mathbf{\Lambda}$ can be solved from a generalized eigenvalue problem [Golub and Van Loan, 2013, p. 497–500]. The dependency on t is then restricted to the $2K$ terms of form $1 + \imath t\lambda_k$, and, in particular, $\boldsymbol{S}$ and $\mathbf{\Lambda}$ are independent of t.

Step 2c

The new parameters of the site approximation follow directly from (3.6) by using the parameters of the cavity distribution and the moments of the tilted distribution.

3.2.2 Spike and Slab Prior Terms

EP for spike and slab priors has been studied in generalized linear single- and multi-task regression models [Hernández-Lobato et al., 2008, 2010b,a], in group- and network-sparse generalizations of the prior [Hernández-Lobato et al., 2011, 2013], and in latent variable models [Rattray et al., 2009].

Consider the spike and slab prior

$$p(\beta_{jk}|\gamma_{jk}, \tau_0^2, \tau_1^2) = \gamma_{jk}\mathrm{N}(\beta_{jk}|0, \tau_1^2) + (1 - \gamma_{jk})\mathrm{N}(\beta_{jk}|0, \tau_0^2),$$
$$p(\gamma_{jk}|\pi_{jk}) = \text{Bernoulli}(\gamma_{jk}|\pi_{jk}),$$

where $\tau_1^2 > \tau_0^2 \geq 0$ (if $\tau_0^2 = 0$ then $\mathrm{N}(\beta_{jk}|0, 0)$ is interpreted as the Dirac spike $\delta_0(\beta_{jk})$) and π_{jk} are assumed given. The approximation is assumed to factorize as $q(\boldsymbol{\beta_j}, \boldsymbol{\gamma_j}) = q(\boldsymbol{\beta}_j)q(\boldsymbol{\gamma}_j)$, where $\boldsymbol{\beta}_j \in \mathbb{R}^K$ has a Gaussian approximation and $\boldsymbol{\gamma}_j \in \{0, 1\}^K$ is taken to have a product-Bernoulli approximation. Let

$$\tilde{t}_{jk}(\beta_{jk}, \gamma_{jk}) = \tilde{t}_{jk}(\beta_{jk})\tilde{t}_{jk}(\gamma_{jk}) = \tilde{Z}_{jk} \exp(-\frac{1}{2}\tilde{\Gamma}_{jk}\beta_{jk}^2 + \beta_{jk}\tilde{\mu}_{jk})\tilde{\rho}_{jk}^{\gamma_{jk}}(1-\tilde{\rho}_{jk})^{1-\gamma_{jk}}$$

correspond to the approximate term for $p(\beta_{jk}|\gamma_{jk}, \tau_0^2, \tau_1^2)$. The approximate term corresponding to $p(\gamma_{jk}|\pi_{jk})$ can be fixed to the prior as it is already of Bernoulli form.

Step 2a

The cavity distribution $q^{\setminus jk}(\beta_{jk}, \gamma_{jk}) = q^{\setminus jk}(\beta_{jk})q^{\setminus jk}(\gamma_{jk})$ is computed as follows. As $q(\gamma_{jk}) \propto \tilde{t}_{jk}(\gamma_{jk})p(\gamma_{jk}|\pi_{jk})$, equation (3.4) gives $q^{\setminus jk}(\gamma_{jk}) = p(\gamma_{jk}|\pi_{jk})$ always and updating $\tilde{t}_{jk}(\gamma_{jk})$ during the EP iterations is not necessary. $q^{\setminus jk}(\beta_{jk}) = \mathrm{N}(\beta_{jk}|m^{\setminus jk}, \Sigma^{\setminus jk})$ is the marginal of the K-dimensional $q^{\setminus jk}(\boldsymbol{\beta}_j)$. Cseke and Heskes [2011, Appendix C.2] gives the parameters in terms of the kth marginal mean m_{jk} and variance Σ_{jk} of $q(\boldsymbol{\beta}_j)$ as

$$m^{\setminus jk} = \frac{m_{jk} - \Sigma_{jk}\tilde{\mu}_{jk}}{1 - \tilde{\Gamma}_{jk}\Sigma_{jk}} \quad \text{and} \quad \Sigma^{\setminus jk} = \frac{\Sigma_{jk}}{1 - \tilde{\Gamma}_{jk}\Sigma_{jk}}.$$

Step 2b

The tilted distribution is

$$\hat{p}(\beta_{jk}, \gamma_{jk}) \propto \left[\gamma_{jk}\pi_{jk}\mathbf{N}(\beta_{jk}|0, \tau_1^2) + (1-\gamma_{jk})(1-\pi_{jk})\mathbf{N}(\beta_{jk}|0, \tau_0^2)\right] \mathbf{N}(\beta_{jk}|m^{\backslash jk}, \Sigma^{\backslash jk}).$$

Computing the mean and variance with regard to β_{jk} involves only analytically tractable Gaussian integrals. Let $\hat{Z} = \hat{Z}_1 + \hat{Z}_0$ be the normalization constant, where

$$\begin{aligned}\hat{Z}_1 &= \pi_{jk}\mathbf{N}(0|m^{\backslash jk}, \tau_1^2 + \Sigma^{\backslash jk}),\\ \hat{Z}_0 &= (1-\pi_{jk})\mathbf{N}(0|m^{\backslash jk}, \tau_0^2 + \Sigma^{\backslash jk}).\end{aligned}$$

The tilted distribution moments are then

$$\begin{aligned}\hat{m} &= \frac{1}{\hat{Z}}\left[\hat{Z}_1 s_1 + \hat{Z}_0 s_0\right] m^{\backslash jk},\\ \hat{\Sigma} &= \frac{1}{\hat{Z}}\left[\hat{Z}_1 s_1 \left(1 + s_1 \frac{(m^{\backslash jk})^2}{\Sigma^{\backslash jk}}\right) + \hat{Z}_0 s_0 \left(1 + s_0 \frac{(m^{\backslash jk})^2}{\Sigma^{\backslash jk}}\right)\right] \Sigma^{\backslash jk} - \hat{m}^2,\end{aligned}$$

where $s_1 = \frac{\tau_1^2}{\tau_1^2 + \Sigma^{\backslash jk}}$ and $s_0 = \frac{\tau_0^2}{\tau_0^2 + \Sigma^{\backslash jk}}$. The approximate posterior inclusion probability can be computed once the EP algorithm has converged as $q(\gamma_{jk} = 1) = \mathbf{E}_{\hat{p}}[\gamma_{jk}] = \frac{\hat{Z}_1}{\hat{Z}}$.

Step 2c

The parameters of the site approximation are updated using (3.6) with the cavity parameters and the tilted distribution moments.

3.3 Alternative Variational Approaches

Many approximation schemes consider minimizing the global Kullback–Leibler divergence from the approximation to the true posterior distribution $\mathrm{KL}[q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathcal{D})]$. The reversal of the direction of the divergence compared to the local KL projections in EP gives the approximation different properties [Bishop, 2006, p. 467–470]: $\min \mathrm{KL}[q \parallel p]$ avoids putting mass in q to regions with low p, whereas $\min \mathrm{KL}[p \parallel q]$ needs to put mass in q anywhere where p has mass. In practice, when the approximation is forced to compromise, the former tends to underestimate variance, while the latter overestimate. As a drastic case, on approximating a multimodal p with a unimodal q, the former can fit to one mode, while the latter needs to span all. Other types of compromises are available with more general divergence measures [Minka, 2005].

Computational tractability in the minimization of $\mathrm{KL}[q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathcal{D})]$ is achieved with suitable restrictions to the approximation and with the design of efficient optimization algorithms (this is an active research area; for some recent general approaches see, e.g., Salimans and Knowles [2013], Ranganath et al. [2014]). A computationally convenient approach in conditionally conjugate models is to assume a factorizing (mean-field) approximation $q(\boldsymbol{\theta}) = \prod_j q(\boldsymbol{\theta}_j)$, which allows an iterative minimization algorithm with closed-form updates [Bishop, 2006, Chapter 10].

The mean-field approach has been applied to approximate linear latent variable models (e.g., Bishop [1999a], Rattray et al. [2009], Stern et al. [2009], Klami et al. [2013], Hernández-Lobato et al. [2014]). Rattray et al. [2009] study sparse factor models with the spike and slab prior. They also propose a hybrid algorithm, where EP is used for the sparsity-inducing prior and the reversed KL for the inner product dependent likelihood terms. Similar hybrid approaches are also used by Stern et al. [2009] and Hernández-Lobato et al. [2014], who motivate the reverse KL update by its symmetry breaking property.

In regression models with the spike and slab prior, the mean-field approximation has been applied by Logsdon et al. [2010] and Carbonetto and Stephens [2012] to genome-wide association analysis. Carbonetto and Stephens [2012] present a hybrid approach, where the integration over the hyperparameters of the model is performed numerically with the marginal likelihood (integration over $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) replaced by the variational lower bound estimate. In line with the above discussion about multimodal posteriors, the results show that the approximation tends to give a single representative variable a high posterior inclusion probability among a set of correlated variables harbouring an association (while the true posterior dilutes the mass among the set). Nevertheless, the approach is found to provide a useful and fast alternative to sampling based inference. Mean-field approximation methods have also been developed for continuous shrinkage priors, including the Laplace and horseshoe priors [Seeger, 2008b, Neville et al., 2014].

4. Summary of the Publications

This chapter provides a brief summary of Publications I–IV. The models are presented here in a unified notation of the thesis, which contains some differences to the attached original work.

4.1 Spike and Slab Linear Model with Additive and Dominant Effects for Genome-wide Association Analysis (I)

During the last decade or so, three key advances in genetics, beginning with the sequencing of a reference human genome and followed by the cataloguing of common human genetic variation and the availability of cost-effective genotyping, have changed the landscape of genetic epidemiology from candidate gene association studies to genome-wide approaches [Lander, 2011]. Genome-wide association studies (GWAS) of common diseases and traits mainly focus on interrogating two-allele single nucleotide polymorphisms (SNPs), which are genomic locations where the common A, T, C, or G nucleotide (major allele) has mutated at some point (or points) in the evolutionary history to another (minor allele) that has been inherited by a part of the study population in one or both copies of the locus.

Primary statistical analysis in GWAS is often conducted by testing each of the hundreds of thousands or millions of SNPs genotyped in the study for association one at a time. This requires stringent multiple hypothesis testing correction and does not allow sharing any information across the tests. Publication I studies the application of the spike and slab linear regression (Section 2.2.1) in GWAS with regard to an extension to additive and dominant genetic effects. An excellent general discussion of what one might expect from an application of spike and slab models in GWAS is provided by Guan and Stephens [2011].

The observation model in Publication I is assumed Gaussian,

$$p(y_i|\boldsymbol{e}_i, \boldsymbol{x}_i, \boldsymbol{b}, \boldsymbol{\beta}, \sigma^2) = \mathrm{N}(y_i|\boldsymbol{b}^{\mathrm{T}}\boldsymbol{e}_i + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2),$$
$$p(\sigma^2) = \text{Inv-}\chi^2(\sigma^2|\nu, s^2),$$

where $\boldsymbol{e}_i$ is a vector of demographic and other fixed covariates (with a Gaussian prior on $\boldsymbol{b}$ omitted here), and $\boldsymbol{x}_i$ is a vector of the SNP genotypes. Common additive coding specifies $x_{ij} = 0, 1, 2$ for the three genotypes $MM, Mm/mM, mm$, where M is the major allele and m the minor. Publication I extends the spike and slab approach to allow other genetic effect types. In particular, categorical variables $t_j \in \{A, AH\}$ are introduced, such that A corresponds to the additive coding, and AH to a linear combination of the additive and heterozygous (i.e., $0, 1, 0$) codings that encompasses purely dominant $(0, 1, 1)$ and recessive $(0, 0, 1)$ effects as special cases. Extending each SNP to contribute two terms in $\boldsymbol{x}_i$, $\boldsymbol{x}_{ij} = [x_{ij,A}, x_{ij,H}]$ with the different codings, the following prior structure for the coefficients $\boldsymbol{\beta}$ is specified:

$$p([\beta_{j,A}\ \beta_{j,H}]^{\mathrm{T}}|\sigma^2, \boldsymbol{\tau^2}, t_j, \gamma_j) = \begin{cases} \delta_0(\beta_{j,A})\delta_0(\beta_{j,H}), & \text{if } \gamma_j = 0, \\ \mathrm{N}(\beta_{j,A}|0, \sigma^2\tau_A^2)\delta_0(\beta_{j,H}), & \text{if } \gamma_j = 1, t_j = A, \\ \mathrm{N}(\beta_{j,A}|0, \sigma^2\tau_A^2)\mathrm{N}(\beta_{j,H}|0, \sigma^2\tau_H^2), & \text{if } \gamma_j = 1, t_j = AH, \end{cases}$$
$$p(\tau_l^2) = \text{Inv-}\chi^2(\tau_l^2|\nu_l, s_l^2), \quad \text{for } l = A, H,$$
$$p(\gamma_j|\pi) = \text{Bernoulli}(\gamma_j|\pi),$$
$$p(\pi) = \text{Beta}(\pi|a_1, a_0),$$
$$p(t_j|\phi, \gamma_j = 1) = \text{Bernoulli}(t_j|\phi), \quad \text{with } t_j \in \{A, AH\},$$
$$p(\phi) = \text{Beta}(\phi|b_A, b_{AH}).$$

When $\gamma_j = 0$, the value of t_j is irrelevant. The redundant parametrization with γ_j and t_j helps to separate the prior specification with regard to the number of expected associations and the types of the genetic effects, but is not strictly necessary. More generally, t_j could also be allowed to specify other codings of the genotypes. The specification of the hyperparameters is discussed in the publication, and a finitely adaptive Markov chain Monte Carlo algorithm is described for the computation (Section 3.1.1).

The behaviour of the model is studied using simulated observations y_i based on real genotype data. The results show that the extended spike and slab approach has better causal variant identification performance compared to a popular single-SNP approach and handles better the extension to allow multiple types of genetic effects. In particular, the exten-

sion of the model space does not markedly decrease the performance when only additive effects are simulated. An application to the association analysis of high-density and low-density lipoprotein cholesterol blood levels in 3895 individuals with over one million SNPs is presented. The genetic regions with the highest posterior inclusion probabilities correspond mostly to previously implicated regions.

4.2 Tuning the Metropolis–Hastings Algorithm for High-dimensional Spike and Slab Linear Models (II)

Publication II considers the Markov chain Monte Carlo computation for the spike and slab model in GWAS in more detail. The basic Metropolis–Hastings approach for sampling $\boldsymbol{\gamma}$, which proposes a change to the state of a single γ_j variable at a time with the specific variable chosen uniformly in random, can suffer from long autocorrelations and poor mixing between correlated variables.

Three extensions of the basic approach are studied: 1) finite adaptation of the proposal distribution of which variable to update, 2) multistep (or block) proposals that propose to change the states of multiple γ_j variables in one go, and 3) delayed rejection for the block proposals (see Section 3.1.1). A method to tune the block size for the multistep proposals is also described. The extensions, together with two metropolized Gibbs sampling approaches [Kohn et al., 2001, Nott and Kohn, 2005], are compared with regard to the effective sample size per time unit (Section 3.1), and are seen to provide some gain in the sampling efficiency. Better mixing of the multistep algorithm with delayed rejection compared to single-step sampling is also demonstrated in the real GWAS data.

In addition, the Gaussian slab prior controlled by a single shared variance parameter used in Publication I is demonstrated to exhibit arguably undesirable multimodal behaviour in the real data. Using a heavier tailed prior distribution, capable of accommodating a wider range of non-null coefficient sizes, is advocated instead. The modified part of the prior structure (ignoring the possibility of multiple effect types t_j, as done in Publication II) can be written as

$$\begin{aligned} p(\beta_j|\lambda_j^2, \tau, \gamma_j = 1) &= \mathrm{N}(\beta_j|0, \sigma^2\tau^2\lambda_j^2), \\ p(\lambda_j^2) &= \text{Inv-}\chi^2(\lambda_j^2|\nu_\lambda, s_\lambda^2), \\ p(\tau) &= \mathrm{N}(\tau|\mu, 1), \end{aligned}$$

where, in particular, the local variance parameters λ_j^2 are introduced to increase the flexibility of the prior (Section 2.2.1).

4.3 Expectation Propagation for Inner Product Factors (III)

Publication III studies the application of expectation propagation (EP) to approximate probabilistic factors that depend on an inner product of random variables (Section 3.2). Examples of such factors include the likelihood terms in generalized linear models when accounting for uncertainty in both the coefficients $\boldsymbol{\beta}$ and the latent variables $\boldsymbol{x}$ is needed (Section 2.1.4). The structure of these kinds of models can be challenging for efficient sampling based inference.

The main issue in applying EP here is the computation of the tilted distribution moments, which are analytically intractable $2K$-dimensional integrals over $\boldsymbol{\beta} \in \mathbb{R}^K$ and $\boldsymbol{x} \in \mathbb{R}^K$. Publication III shows how an integral transform of the Dirac delta function can be used to cast the problem into $O(K)$ one-dimensional integrals, which are then evaluated numerically (Section 3.2.1). The main requirement is that the posterior distribution of the coefficients and latent variables are approximated as Gaussian. Conditions on the approximated probability factor with regard to the integral transform are discussed in the supplementary material of the publication in particular for Gaussian and probit likelihood terms.

The accuracy of the EP posterior approximation is studied in sparse principal component analysis models of the following form:

$$
\begin{aligned}
p(y_{ij}|\boldsymbol{x}_i, \boldsymbol{\beta}_j) &= \mathrm{N}(y_{ij}|\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i, \sigma^2) \text{ or } \mathrm{Bernoulli}(y_{ij}|\Phi(\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{x}_i)),\\
p(x_{ik}) &= \mathrm{N}(x_{ik}|0,1),\\
p(\beta_{jk}|\gamma_{jk}=1) &= \mathrm{N}(\beta_{jk}|0,\tau^2),\\
p(\beta_{jk}|\gamma_{jk}=0) &= \delta_0(\beta_{jk}),\\
p(\gamma_{jk}) &= \mathrm{Bernoulli}(\gamma_{jk}|\pi),
\end{aligned}
$$

where σ^2, τ^2 and π are assumed fixed for simplicity, and the observations are $\mathcal{D} = \{\boldsymbol{y}_i \in \mathbb{R}^m : i = 1, \ldots, n\}$ for the Gaussian case ($\boldsymbol{y}_i \in \{0,1\}^m$ for the probit case). Comparisons to the hybrid algorithm that uses variational Bayes to approximate the likelihood terms and EP to approximate the spike and slab terms [Rattray et al., 2009] are presented in simulated datasets with Gibbs sampling as a reference. The results show that the proposed EP approach can in some cases be notably more accurate.

4.4 Hierarchical Survival Modelling and Covariate Selection for Cardiovascular Event Risk Prediction (IV)

As opposed to genetic markers, measurements of metabolic biomarkers provide a temporal, molecular-level snapshot of the state of the dynamic interplay of genetic, metabolic, environmental, and lifestyle factors. Identifying metabolic biomarkers related to disease processes is essential for accurate risk assessment, informed and early decisions for preventive treatment, and for the generation of hypotheses about treatment targets.

Publication IV presents a modelling approach with the goal of finding biomarkers predictive of adverse cardiovascular events in diabetic individuals. The available data consists of a cohort of 7932 Finnish individuals with measurements of 55 candidate biomarker blood levels and 15 years of follow-up for the cardiovascular events. At the beginning of the study, 401 of the participants had diabetes. Given this relatively limited number of diabetic individuals, a joint model for the diabetic and non-diabetic individuals is formulated based on the assumption that, although diabetes confers an increased risk of cardiovascular disease, the risk factors are likely to be similar to some extent. Separate Weibull observation models (Section 2.1.3) are assumed for the four subgroups formed based on the diabetes status (Non/Diabetic) and sex (Men/Women),

$$p(t_i|\nu_i, l_i, \boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \text{Weibull}(t_i|\mu^{(l_i)} + \boldsymbol{b}^{(l_i)\text{T}}\boldsymbol{e}_i + \boldsymbol{\beta}^{(l_i)\text{T}}\boldsymbol{x}_i, \alpha^{(l_i)}, \nu_i),$$

where t_i is an observed event or censoring time (with $\nu_i = 1$ or 0, respectively), $l_i \in \{NM, NW, DM, DW\}$ is a subgroup indicator, and $\boldsymbol{e}_i$ and $\boldsymbol{x}_i$ are vectors of established risk factor and candidate biomarker levels. The models are tied together using joint prior distributions. For the jth biomarker, with $\boldsymbol{\beta}_j = [\beta_j^{(NM)}, \ldots, \beta_j^{(DW)}]^\text{T}$,

$$p(\boldsymbol{\beta}_j|r_j^2, c, s) = \text{N}(\boldsymbol{\beta}_j|\boldsymbol{0}, r_j^2\boldsymbol{\Sigma}),$$

where r_j^2 is used to specify a sparsity promoting scale mixture prior (Section 2.2.1) on the whole coefficient vector $\boldsymbol{\beta}_j$, and the 4×4 correlation matrix $\boldsymbol{\Sigma}$ is parametrized by c and s, where the former controls the similarity of the models between men and women and the latter between non-diabetic and diabetic individuals (see Section 2.2.2). Similar priors are used for the $\boldsymbol{b}_j$ and the shape parameters $\boldsymbol{\alpha}$, and the three prior correlation structures are tied at a higher level. Posterior samples are obtained using Markov chain Monte Carlo sampling (Section 3.1.2).

Predictive performance estimates from ten-fold cross-validation are used to compare different models, including a comparison of the horseshoe, Laplace and Gaussian shrinkage priors (Section 2.2.1). The results imply three main conclusions: 1) the candidate biomarkers contain additional predictive power over the established risk factors, 2) the joint model predicts better for the diabetic individuals than using a separate model, 3) the horseshoe prior provides the best shrinkage behaviour among the three alternatives for this data.

The projective covariate selection approach of Dupuis and Robert [2003] with a forward selection search strategy is then used to find biomarker subsets with predictive value (Section 2.3) using the joint model with horseshoe shrinkage as a reference. The projections are performed only on the part of the posterior samples pertaining to modelling the diabetic subgroups (DM and DW). Ten-fold cross-validation is used to estimate the predictive performance of the submodels along the forward selection path. The results imply that there are two to ten predictive biomarkers. The uncertainty in the cross-validation comparisons is relatively large, making a definite choice of a single submodel difficult.

5. Discussion

This thesis has focused on the computation and application of Bayesian models in settings, where the amount of data is scarce compared to the goals of the statistical analysis, and relatively strong constraints on the hypothesis space seem necessary for uncovering meaningful patterns. Such analysis risks being biased, but avoids drowning in the variance of trying to satisfy the large number of degrees of freedom in flexible models. The phenomena studied are complex at a deep level (e.g., the physiological interplay of genetic, metabolic, lifestyle, and environmental factors in the development of disease) and often lacking in knowledge and data especially at the metabolic level and with respect to time. It is then difficult to formulate useful physical models of the underlying processes for addressing questions of interest, for example, in epidemiology. The statistical models applied here are superficial in the sense of ignoring a lot of the true complexity. The aim is usefulness at the appropriate level rather than exactness.

The main ingredients in the models studied in this thesis were linearity and sparsity. Together they enable formulating models that provide interpretable summaries of statistical relationships between outcomes and covariates in high-dimensional datasets. In addition, Publications I and IV demonstrated the (well-known) benefits of hierarchical modelling. Publication I extended the spike and slab prior to include different genetic effect types, with robust performance in identifying true associations despite the increase in the size of the model space. Publication IV showed gains in the performance of predicting the risk of adverse cardiovascular events in diabetic individuals by tying in a model of non-diabetic individuals, whose data were more plentiful.

The sparsity-inducing prior distributions present computational challenges as analytic computation is not possible in general. Non-log-concave

priors that can strongly favour sparsity, including the spike and slab and horseshoe priors, but excluding the Laplace and Gaussian priors, result in multimodal posterior distributions when the data are not informative enough to rule out the multiple competing hypotheses. In the worst case, accurately representing the posterior distribution may have combinatorial complexity [Seeger, 2008b]. Tailored Markov chain Monte Carlo sampling methods, such as those studied in Publication II for genome-wide association analysis, can be used to improve mixing especially when there is application-specific knowledge about the possible difficulties. On the other hand, the current deterministic approximation algorithms are mainly based on fitting a multivariate Gaussian distribution for the linear model coefficients. Apart from being inappropriate for the multimodal cases, the symmetric Gaussian form cannot capture the skewness of the posterior distribution often arising from shrinkage priors. Yet, the variational approaches have been found to provide useful and fast, if not in all respects accurate, inference methods even for spike and slab models (e.g., Hernández-Lobato et al. [2010b], Carbonetto and Stephens [2012], Publication III; however, a comprehensive study of expectation propagation for approximating the posterior quantities of interest in genome-wide association analysis, such as posterior inclusion probabilities, seems lacking). In summary, given the possible brittleness of the computation, it seems wise at present to recommend the sparse multiple linear regression as a complementary tool rather than as a replacement to the common single-covariate analysis, for example, in high-throughput genetics applications.

Publication III described an approach for applying the expectation propagation algorithm in models with bilinear probabilistic factors, and studied the approximation in sparse principal component analysis as a proof of concept. While the initial results are encouraging, some issues remain to fully characterize the potential of the approach and its robust application in real-world datasets. First, the Simpson's composite rule was applied in Publication III for the numerical integration over the auxiliary variable t in the moment computations. While simple, it may not be the most efficient and robust method with the possibly oscillating integrand. Studying in detail when the integrand is oscillatory would be necessary. Second, linear latent variable models often have symmetries, such as sign or rotation ambiguities between the latent variables and the coefficients, which can present as multiple modes in the tilted distribution. The mean of the tilted distribution can then be a poor choice to

centre the approximation at. Moreover, expectation propagation is known to exhibit convergence problems in multimodal cases (e.g., Seeger [2008b], Jylänki et al. [2011]). Erasing the symmetries already in the model specification would then be desirable. It is also unclear which of the sequential, parallel, or more elaborate implementations of the algorithm would provide here the best tradeoff between robustness and computational efficiency. Last, the covariance structure of the Gaussian approximation can have a large effect on the accuracy and the computational requirements of the inference. Publication III used a factorizing approximation between the latent variables and their coefficients, rendering the covariance matrix block-diagonal with the block size determined by the dimensionality of the latent space. Moment computations for the general case of unrestricted covariance were described in Section 3.2.1. For the sparse principal component analysis model without the factorization assumption, the precision matrix will exhibit a sparse structure ($\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m$ are conditionally independent, as are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$). However, the covariance matrix will in general be dense. The computational and memory requirements may be prohibitive unless the number of observations n, their dimensionality m, and the dimensionality of the latent space K are small. In the other extreme, a fully diagonal or other suitable very sparse covariance approximation could provide efficient inference, but their characterization in practice remains a future work.

The sparsity-promoting priors do not lead to truly sparse posterior distributions. Yet, there are multiple reasons for selecting subsets of representative covariates (cost in future tasks, ease of exposition, etc.). Publication IV studied projective covariate selection in an epidemiological disease risk prediction setting. The results, for one thing, highlight that appreciating the uncertainty in the process is important, and pinpointing a single subset of covariates was found difficult. On the other hand, the Kullback–Leibler projection, $\min_{\boldsymbol{\theta}_\perp} \mathrm{KL}[p(y|\boldsymbol{x}, \boldsymbol{\theta}) \parallel p(y|\boldsymbol{x}_\perp, \boldsymbol{\theta}_\perp)]$, in the method of Dupuis and Robert [2003] is perhaps not the preferred one. Defining the projection from the true predictive distribution of the reference model to some distribution q, $\min_q \mathrm{KL}[p(y|\boldsymbol{x}) \parallel q(y|\boldsymbol{x}_\perp)]$, seems more desirable, but does not immediately give a tractable solution. Vehtari and Lampinen [2004] use this KL for covariate selection, but not in a projective approach (projection should be able to give a smaller KL divergence than a $q(y|\boldsymbol{x}_\perp)$ which is fitted independently of the reference model). Suitable fixed-form restrictions on q could render the KL projec-

tion tractable, similar to variational inference (see also the discussion in Section 9 of Minka [2005]). Another interesting research direction would be to explicitly model the conditional distribution of the left-out covariates $\boldsymbol{x}_\top$, $p(\boldsymbol{x}_\top|\boldsymbol{x}_\perp)$. The computation of the predictive distribution $q(y|\boldsymbol{x}_\perp)$ by marginalization of $\boldsymbol{x}_\top$, with appropriate approximations, would then bear similarity to the developments in Publication III. While correlations among the covariates are usually thought an impediment to covariate selection, the modelling of the covariate distribution would in fact capitalize on them (see also Lindley [1968]).

Bibliography

Cédric Archambeau, Shengbo Guo, and Onno Zoeter. Sparse Bayesian multi-task learning. In *Advances in Neural Information Processing Systems 24*, pages 1755–1763, 2011.

Maria Maddalena Barbieri and James O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897, 2004.

Mark A. Beaumont and Bruce Rannala. The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–261, 2004.

Michael Betancourt and Mark Girolami. Hamiltonian Monte Carlo for hierarchical models. *arXiv preprint arXiv:1312.0906v1*, 2013.

Christopher M. Bishop. Variational principal components. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, volume 1, pages 509–514, 1999a.

Christopher M. Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems 11*, pages 382–388, 1999b.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Leonard Bottolo and Sylvia Richardson. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3):583–618, 2010.

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

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

Hugh Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996.

Merlise A. Clyde. Bayesian model averaging and model search strategies. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 157–172. Oxford University Press, 1999.

Merlise A. Clyde, Joyee Ghosh, and Michael L. Littman. Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101, 2011.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220, 1972.

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

David Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):45–97, 1995.

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

Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, third edition, 2014.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

E. I. George. Discussion to "Bayesian model averaging and model search strategies" by Merlise A. Clyde. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 175–177. Oxford University Press, 1999.

Edward I. George. Dilution priors: Compensating for model space redundancy. In James O. Berger, T. Tony Cai, and Iain M. Johnstone, editors, *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, pages 158–165. Institute of Mathematical Statistics, 2010.

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

Edward I. George and Robert E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997.

J. Geweke. Variable selection and model comparison in regression. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 5*, pages 609–620. Oxford University Press, 1996.

Charles J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4): 473–482, 1992.

Daniel Gianola, Gustavo de Los Campos, William G. Hill, Eduardo Manfredi, and Rohan Fernando. Additive genetic variability and the Bayesian alphabet. *Genetics*, 183(1):347–363, 2009.

Greg Gibson. Hints of hidden heritability in GWAS. *Nature Genetics*, 42(7): 558–560, 2010.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The John Hopkins University Press, fourth edition, 2013.

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

Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

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

Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815, 2011.

Paul Gustafson. *Measurement Error and Misclassification in Statistical Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall/CRC, 2004.

Seymour Haber. Numerical evaluation of multiple integrals. *SIAM Review*, 12 (4):481–526, 1970.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Ralf Herbrich. Minimising the Kullback–Leibler divergence. Technical report, Microsoft Research, 2005.

Daniel Hernández-Lobato and José Miguel Hernández-Lobato. Learning feature selection dependencies in multi-task learning. In *Advances in Neural Information Processing Systems 26*, pages 746–754, 2013.

Daniel Hernández-Lobato, José Miguel Hernández-Lobato, Thibault Helleputte, and Pierre Dupont. Expectation propagation for Bayesian multi-task feature selection. In *Machine Learning and Knowledge Discovery in Databases*, pages 522–537. Springer, 2010a.

Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Alberto Suárez. Expectation propagation for microarray data classification. *Pattern Recognition Letters*, 31(12):1618–1626, 2010b.

Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Pierre Dupont. Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14:1891–1945, 2013.

José Miguel Hernández-Lobato, Tjeerd Dijkstra, and Tom Heskes. Regulator discovery from gene expression time series of malaria parasites: a hierachical approach. In *Advances in Neural Information Processing Systems 20*, pages 649–656, 2008.

Jose Miguel Hernández-Lobato, Daniel Hernández-Lobato, and Alberto Suárez. Network-based sparse Bayesian classification. *Pattern Recognition*, 44(4):886–900, 2011.

José Miguel Hernández-Lobato, Neil Houlsby, and Zoubin Ghahramani. Probabilistic matrix factorization with non-random missing data. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1512–1520, 2014.

Tom Heskes and Onno Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 216–223, 2002.

Sepp Hochreiter, Ulrich Bodenhofer, Martin Heusel, Andreas Mayr, Andreas Mitterecker, Adetayo Kasim, Tatsiana Khamiakova, Suzy Van Sanden, Dan Lin, Willem Talloen, Luc Bijnens, Hinrich W. H. Göhlmann, Ziv Shkedy, and Djork-Arné Clevert. FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527, 2010.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

Joseph G. Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Springer, 2001.

Heikki Joensuu, Aki Vehtari, Jaakko Riihimäki, Toshirou Nishida, Sonja E. Steigen, Peter Brabec, Lukas Plank, Bengt Nilsson, Claudia Cirilli, Chiara Braconi, Andrea Bordoni, Magnus K. Magnusson, Zdenek Linke, Jozef Sufliarsky, Massimo Federico, Jon G. Jonasson, Angelo Paolo Dei Tos, and Piotr Rutkowski. Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *The Lancet Oncology*, 13(3):265–274, 2012.

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

Joseph B. Kadane and Nicole A. Lazar. Methods and criteria for model selection. *Journal of the American Statistical Association*, 99(465):279–290, 2004.

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, second edition, 2002.

Hanni P. Kärkkäinen and Mikko J. Sillanpää. Back to basics for Bayesian model building in genomic selection. *Genetics*, 191(3):969–987, 2012.

Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013.

Robert Kohn, Michael Smith, and David Chan. Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11(4):313–322, 2001.

Demetris Lamnisos, Jim E. Griffin, and Mark F. J. Steel. Transdimensional sampling algorithms for Bayesian variable selection in classification problems with many more variables than observations. *Journal of Computational and Graphical Statistics*, 18(3):592–612, 2009.

Demetris Lamnisos, Jim E. Griffin, and Mark F. J. Steel. Adaptive Monte Carlo for Bayesian variable selection in regression models. *Journal of Computational and Graphical Statistics*, 22(3):729–748, 2013.

Eric S. Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197, 2011.

Eduardo Ley and Mark F. J. Steel. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674, 2009.

Fan Li and Nancy R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214, 2010.

D. V. Lindley. The choice of variables in multiple regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 31–66, 1968.

Fei Liu, Sounak Chakraborty, Fan Li, Yan Liu, and Aurelie C. Lozano. Bayesian regularization via graph Laplacian. *Bayesian Analysis*, 9(2):449–474, 2014.

Benjamin A. Logsdon, Gabriel E. Hoffman, and Jason G. Mezey. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, 11:58, 2010.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

David Madigan and Jeremy York. Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232, 1995.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, second edition, 1989.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

J. S. Milton and Jesse C. Arnold. *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*. McGraw-Hill, third edition, 1995.

Thomas Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, 2005.

Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

Thomas P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001a.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369, 2001b.

Thomas P. Minka. The EP energy function and minimization schemes. Technical report, 2001c.

Antonietta Mira. On Metropolis–Hastings algorithms with delayed rejection. *Metron*, 59(3–4):231–241, 2001.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman & Hall/CRC, 2011.

Yurii A. Neretin. *Lectures on Gaussian Integral Operators and Classical Groups*. European Mathematical Society, 2011.

Sarah E. Neville, John T. Ormerod, and M. P. Wand. Mean field variational Bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electronic Journal of Statistics*, 8(1):1113–1151, 2014.

David J. Nott and Peter J. Green. Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics*, 13(1), 2004.

David J. Nott and Robert Kohn. Adaptive sampling for Bayesian variable selection. *Biometrika*, 92(4):747–763, 2005.

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

Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

Manfred Opper and Ole Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177–2204, 2005.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Cristian Pasarica and Andrew Gelman. Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, 20(1):343–364, 2010.

P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

Robert Plomin, Claire M. A. Haworth, and Oliver S. P. Davis. Common disorders are quantitative traits. *Nature Reviews Genetics*, 10(12):872–878, 2009.

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

Nicholas G. Polson and James G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

Adrian E. Raftery, David Madigan, and Chris T. Volinsky. Accounting for model uncertainty in survival analysis improves predictive performance. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 5*, pages 323–349. Oxford University Press, 1996.

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

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

Magnus Rattray, Oliver Stegle, Kevin Sharp, and John Winn. Inference algorithms and learning theory for Bayesian sparse factor analysis. *Journal of Physics: Conference Series*, 197(1):012002, 2009.

F. Rigat and A. Mira. Parallel hierarchical sampling: A general-purpose interacting Markov chains Monte Carlo algorithm. *Computational Statistics and Data Analysis*, 56(6):1450–1467, 2012.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, second edition, 2004.

Jeffrey S. Rosenthal. Optimal proposal distributions and adaptive MCMC. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 93–111. Chapman & Hall/CRC, 2011.

Tim Salimans and David A. Knowles. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

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

Matthias Seeger. Expectation propagation for exponential families. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2008a.

Matthias Seeger and Hannes Nickisch. Fast convergent algorithms for expectation propagation approximate Bayesian inference. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 652–660, 2011.

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

Michael H. Seltzer, Wing Hung Wong, and Anthony S. Bryk. Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics*, 21(2):131–167, 1996.

Daniel Sheldon. Graphical multi-task learning. In *NIPS 2008 Workshop: "Structured Input – Structured Output"*, 2008.

Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

Michael Smith and Robert Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

Stan Development Team. Stan: A C++ library for probability and sampling, 2014. URL `http://mc-stan.org/`.

David Stern, Ralf Herbrich, and Thore Graepel. Matchbox: Large scale online Bayesian recommendations. In *18th International World Wide Web Conference*, pages 111–120, 2009.

Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press, 1986.

Geir Storvik. On the flexibility of Metropolis–Hastings acceptance probabilities in auxiliary variable proposal generation. *Scandinavian Journal of Statistics*, 38(2):342–358, 2011.

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

Luke Tierney. Markov chains for exploring posterior distributions (with discussion). *The Annals of Statistics*, 4:1701–1762, 1994.

Luke Tierney and Antonietta Mira. Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine*, 18(17–18):2507–2515, 1999.

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

David A. van Dyk and Taeyoung Park. Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482): 790–796, 2008.

Marcel van Gerven, Botond Cseke, Robert Oostenveld, and Tom Heskes. Bayesian source localization with the multivariate Laplace prior. In *Advances in Neural Information Processing Systems 22*, pages 1901–1909, 2009.

Aki Vehtari and Jouko Lampinen. Model selection via predictive explanatory power. Technical Report B38, Laboratory of Computational Engineering, Helsinki University of Technology, 2004.

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

Melanie A. Wilson, Edwin S. Iversen, Merlise A. Clyde, Scott C. Schmidler, and Joellen M. Schildkraut. Bayesian model search and multilevel inference for SNP association studies. *The Annals of Applied Statistics*, 4(3):1342–1364, 2010.

Robert L. Wolpert. A conversation with James O. Berger. *Statistical Science*, 19 (1):205–218, 2004.

Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, 9(2):e1003264, 2013.

Aalto-DD 206/2014

ISBN 978-952-60-6011-8 (printed)
ISBN 978-952-60-6012-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Biomedical Engineering and Computational Science

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS