MSc Mathematics

Track: Stochastics

Master Thesis

Sparse Variational Inference and Bayesian High-Dimensional Regression

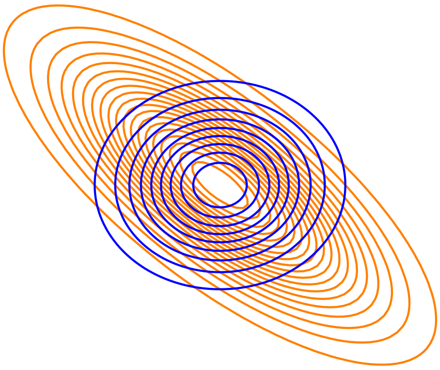
by

Gabriel Clara

August 11, 2021

Supervisor: Dr. Botond Szabó

Second examiner: Dr. Paulo J. de Andrade Serra



Department of Mathematics

Faculty of Sciences

VU UNIVERSITY AMSTERDAM

Abstract

A mean-field variational algorithm for Bayesian sparse logistic regression models with spike-and-slab priors is developed. The novelty of the approach lies in the use of Laplace densities for the continuous parts of the prior, as opposed to the commonplace Gaussian densities. Several issues unique to the Laplace approach are discussed: derivation of the objective function, numerical optimization, and efficient implementation. To complement the theoretical properties of the algorithm, a simulation study featuring a variety of Bayesian methods for sparse logistic regression is conducted. Additionally, some time is spent on exposition of issues relating to variational inference and contraction of the resulting approximate posterior distributions.

Title: Sparse Variational Inference and Bayesian High-Dimensional Regression
Author: Gabriel Clara,
Supervisor: Dr. Botond Szabó
Second examiner: Dr. Paulo J. de Andrade Serra
Date: August 11, 2021

Department of Mathematics
VU University Amsterdam
de Boelelaan 1081, 1081 HV Amsterdam
http://www.math.vu.nl/

For JT

Preface

The following thesis contains the collected efforts from my collaboration with Dr. Botond Szabó and Dr. Kolyan Ray on spike-and-slab variational inference during my time as a student at the Vrije Universiteit Amsterdam. In the arduous process of turning my notes into a sound development of the subject matter, I added exposition of general issues in Bayesian high-dimensional inference, variational methods, computer programming, and others. Whenever I thought it to be a useful point of view, I chose to emphasize abstract definitions before treating the concrete setting of sparse logistic regression, a style that I find under-represented in the related literature.

The thesis is divided into two parts. The first part focuses on the mathematical derivation of the variational algorithm for the sparse logistic model. The second part illustrates the implementation of the algorithm and includes a simulation study with a variety of Bayesian methods for sparse logistic regression. Additionally, a short chapter before the first part collects the notation used throughout the thesis, though the notation is usually introduced whenever relevant in the main text as well.[1]

Throughout the project, I have learned immensely from many meetings with Dr. Botond Szabó and Dr. Kolyan Ray, both in terms of the subject matter as well as mathematical research itself. Due to a variety of factors, meeting in-person was impossible and so most of the work was done remotely. I am greatly indebted to them, both for their mentorship and their willingness to accommodate my remote situation. Moreover, I would like to thank Dr. Paulo Serra for grading the thesis.

On the personal side I owe much to my family. Without them this thesis — let alone my pursuit of mathematical research — would certainly be impossible. Due to the same factors cited above, I have not been able to meet them for most of the time that I have worked on this thesis, yet this has not prevented them from being ever-supportive.

Lastly, I could not have completed this thesis without Thi. There were many times throughout the past year when life seemed too difficult to make progress, but between that and finishing a thesis of her own she still managed to give me strength. I hope both her and my own perseverance and support are reflected in each other's works.

Gabriel Clara
Boston, Massachusetts

[1] In the preface of the famous algebra treatise [Lan02], the author refers to the practice of repeating definitions as *lèse Bourbaki*, which is perhaps one of the funnier mathematical jokes.

Contents

Notation and Conventions

This short, preliminary chapter details most of the mathematical notation and conventions used throughout part I. As may be expected in the subject of statistics, most of the conventions will broadly relate to either measure theory, linear algebra, or optimization, focusing on the real number field $\mathbb{R}$.

Given $x \in \mathbb{R}$, the segments $\{y \in \mathbb{R} \mid y > x\}$ and $\{y \in \mathbb{R} \mid y < x\}$ will be abbreviated as $\mathbb{R}_{>x}$ and $\mathbb{R}_{<x}$. It will sometimes be useful to consider functions taking infinite values, for example in the extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. The n-fold Cartesian product $\mathbb{R}^n$ will always be viewed as a topological vector space with point-wise addition and scalar multiplication. A linear operator $\tau : \mathbb{R}^n \to \mathbb{R}^m$ may in this case be taken as synonymous with the matrix $X \in \mathbb{R}^{m \times n}$ representing it with respect to the standard basis $\{e_i\}_{i \in n}$.

$\mathbb{R}_{>x}$, $\mathbb{R}_{<x}$

$\overline{\mathbb{R}}$

$\mathbb{R}^n$

e_i

There are various norms and inner products defined on $\mathbb{R}^n$, with the most relevant being the ℓ^p-norms $\|v\|_p^p = \sum_{i=1}^n |v_i|^p$ with $p \in (1, \infty)$. For $p = 2$, the norm is induced by the inner product $\langle v, w \rangle = v^\mathrm{t} w$. If X_i, $i \in \{1, \dots p\}$, denotes the i^th column of X, then a semi-norm on $\mathbb{R}^{m \times n}$ is given by $\|X\| = \max_{j \in \{1,\dots,n\}} \|X_j\| = \max_{j \in \{1,\dots n\}} (X^\mathrm{t} X)_{jj}^{1/2}$.

ℓ^p, $\|\cdot\|_p$

$v^\mathrm{t} w$

$\|X\|$

A discussion of measure theory often starts by defining a σ-algebra. Here, this will usually mean either the Borel σ-algebra $\mathscr{B}(\mathbb{R})$, generated by the open intervals, or the collection of Lebesgue measurable sets. Taking the product topology on $\mathbb{R}^n$, for any $n \in \mathbb{N}$, yields straightforward generalizations to higher dimensions. In this context, the functions $\pi_j : \bigotimes_{i=1}^n S_n \to S_j$, $j \in \{1, \dots, n\}$, will always denote the respective coordinate projections.

$\mathscr{B}(\mathbb{R})$

π_j

In any measurable space $(\Omega, \mathfrak{F})$, a reference probability measure will be denoted by $\mathbb{P} : \mathfrak{F} \to [0, 1]$. For any $p \in \mathbb{R}_{>0}$, write

$\mathbb{P} : \mathfrak{F} \to [0, 1]$

$$\mathcal{L}^p(\mathbb{P}) = \left\{ f : \Omega \to \overline{R} \;\middle|\; \|f\|_p^p = \int_\Omega |f|^p \,\mathrm{d}\mathbb{P} < \infty \right\},$$

$\mathcal{L}^p(\mathbb{P})$, $\|\cdot\|_p$

then $\|\cdot\|_p$ defines a semi-norm on the space of functions with finite p^th moment. Its restriction to the quotient space $L^p(\mathbb{P}) = \mathcal{L}^p(\mathbb{P}) / \overset{\text{a.s.}}{=}$ yields a genuine norm and Banach space structure.

$L^p(\mathbb{P})$

The collection of finite signed measures on a measurable space $(\Omega, \mathfrak{F})$ and its subset, the probability measures, will be signified by $\mathfrak{M}(\Omega)$ and $\mathscr{P}(\Omega)$. The former may be given the structure of a Banach space via the total variation norm

$\mathfrak{M}(\Omega)$, $\mathscr{P}(\Omega)$

$$\|\mu\|_\text{TV} = \sup_{A \in \mathfrak{F}} \mu(A) + \left| \inf_{A \in \mathfrak{F}} \mu(A) \right|.$$

$\|\cdot\|_\text{TV}$

A particularly important probability measure for modeling sparsity is the Dirac δ-distribution $\delta_0 : 2^\mathbb{R} \to \{0, 1\}$, defined via $\delta_0(A) = \mathbf{1}_{\{0 \in A\}}$. As an operator on a space of suitable functions, δ_0 may also be identified with the evaluation functional $\delta_0(f) = f(0)$.

$\delta_0 : 2^\mathbb{R} \to \{0, 1\}$

$\delta_0(f)$

Given a measure Q on $(\Omega, \mathfrak{F})$ and $P \in \mathfrak{M}(\Omega)$, the relation $P \ll Q$ of absolute continuity signifies that $Q(A) = 0$ implies $P(A) = 0$ for all $A \in \mathfrak{F}$. If additionally Q is σ-finite and P finite then there exists unique $\mathrm{d}P/\mathrm{d}Q \in L^1(Q)$, satisfying

$P \ll Q$

$$P(A) = \int_A \frac{\mathrm{d}P}{\mathrm{d}Q}\, \mathrm{d}Q$$

$\frac{\mathrm{d}P}{\mathrm{d}Q}$

for all $A \in \mathfrak{F}$. This so-called Radon-Nikodym derivative gives rise to the crucial object

$$\mathrm{KL}(P, Q) = \int_\Omega \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P,$$

$\mathrm{KL}(P, Q)$

the Kullback-Leibler divergence, or relative entropy, between probability measures.

A real-valued random variable, or vector, will refer to a measurable mapping $X : \Omega \to \mathbb{R}^d$ on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. Any such mapping induces a probability measure $\mathbb{P}^X = \mathbb{P} \circ X^{-1}$ on $\mathbb{R}^d$. If $\mathbb{P}^X \ll \mu$, for the Lebesgue measure μ of appropriate dimension, then $\mathrm{d}\mathbb{P}^X/\mathrm{d}\mu$ will be called the density of X. Some common random variables in $\mathbb{R}$ and their densities are

rand. variable

$\mathbb{P}^X$

density

$$X \sim \mathcal{N}(\mu, \sigma^2), \qquad \mu \in \mathbb{R},\ \sigma \in \mathbb{R} \setminus \{0\} \qquad \frac{\mathrm{d}\mathbb{P}^X}{\mathrm{d}\mu}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

$\mathcal{N}(\mu, \sigma^2)$

$$X \sim \mathrm{Lap}(\nu, \lambda), \qquad \nu \in \mathbb{R},\ \lambda \in \mathbb{R}_{>0} \qquad \frac{\mathrm{d}\mathbb{P}^X}{\mathrm{d}\mu}(x) = \frac{1}{2\lambda} e^{-|x-\nu|/\lambda}$$

$\mathrm{Lap}(\lambda)$

$$X \sim \mathrm{Cauchy}(\mu, \tau), \qquad \mu \in \mathbb{R},\ \tau \in \mathbb{R}_{>0}, \qquad \frac{\mathrm{d}\mathbb{P}^X}{\mathrm{d}\mu}(x) = \frac{\tau}{\pi} \frac{1}{(x-\mu)^2 + \tau^2}$$

$\mathrm{Cauchy}(\mu, \tau)$

$$X \sim \mathrm{Unif}[a, b], \qquad a \in \mathbb{R},\ b \in \mathbb{R}_{>a} \qquad \frac{\mathrm{d}\mathbb{P}^X}{\mathrm{d}\mu}(x) = \frac{\mathbf{1}_{[a,b]}(x)}{b-a}$$

$\mathrm{Unif}[a, b]$

$$X \sim \mathrm{Gamma}(\alpha, \beta), \qquad \alpha, \beta \in \mathbb{R}_{>0}, \qquad \frac{\mathrm{d}\mathbb{P}^X}{\mathrm{d}\mu}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

$\mathrm{Gamma}(\alpha, \beta)$

$$X \sim \mathrm{Beta}(\alpha, \beta), \qquad \alpha, \beta \in \mathbb{R}_{>0} \qquad \frac{\mathrm{d}\mathbb{P}^X}{\mathrm{d}\mu}(x) \propto \mathbf{1}_{[0,1]}(x) \cdot x^{\alpha-1}(1-x)^{\beta-1}.$$

$\mathrm{Beta}(\alpha, \beta)$

Additionally, if $X \sim \mathrm{Unif}[0, 1]$ and $p \in [0, 1]$, then $Y \sim \mathrm{Binom}(p)$ shall signify $Y = \mathbf{1}_{\{X \leq p\}}$.

$\mathrm{Bin}(p)$

Frequently, statistics texts employ some variation on the phrase *"Let $X_1, X_2, \ldots, X_n$ be an independent, identically distributed (i.i.d.) sample"*, meaning that the measure $\mathbb{P}^{(X_1, \ldots, X_n)}$ factorizes into a product of identical measures. If each X_i is a random vector in $\mathbb{R}^d$ and has density $f : \mathbb{R}^d \times \Theta \to \mathbb{R}$, parameterized over some set Θ, then

i.i.d.

$$\mathcal{L}(\theta \mid X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n f(x_i, \theta)$$

$\mathcal{L}(\theta \mid X = x)$

will denote the likelihood of $\theta \in \Theta$. Additionally, $\ell(\theta \mid X = x)$ shall denote $\log \mathcal{L}(\theta \mid X = x)$.

$\ell(\theta \mid X = x)$

Given $f, g : \mathbb{N} \to \mathbb{R}$, the order relation $f = O(g)$ will signify the existence of $M \in \mathbb{R}$ and $m \in \mathbb{N}$ satisfying $f(n) \leq Mg(n)$, for all $n \geq m$. Similarly, the order relation $f = o(g)$ shall denote that for every $\varepsilon \in \mathbb{R}_{>0}$, there exists $m \in \mathbb{N}$ such that $f(n) \leq \varepsilon g(n)$, whenever $n \geq m$.

Lastly, a common theme in the parts relating to optimization will be convexity. A subset C of a real vector space V will be called convex if $x, y \in C$ and $t \in [0,1]$ imply $tx+(1-t)y \in C$. convex set A function $f : V \to \overline{\mathbb{R}}$ will be labeled convex if the sets $\{(x,y) \in V \times \mathbb{R} \mid y \geq f(x)\}$ and convex function $\{x \in V \mid f(x) < \infty\}$ are both convex. For sufficiently smooth functions with domain $\mathbb{R}$, this turns out to be equivalent to $f'' \geq 0$. Given $x, y \in \mathbb{R}$ and a differentiable function f, convexity enables the lower-bound $f(y) \geq f'(x)(y-x)$.

Part I.

Theory and Methodology

1. A Bayesian Approach to Sparse Logistic Regression

The first chapter introduces the logistic regression problem and discusses some of the common modeling choices dealing with sparsity, focusing especially on the Bayesian setting. These choices are mainly taken as motivation for the mathematical problems studied in subsequent chapters. For a thorough treatment of modeling aspects in high-dimensional regression see chapter 18 of [HTF09].

1.1. The Logistic Model

Binary regression problems are a ubiquitous topic in modern machine learning and statistics. The prototypical example motivating the study of these problems is perhaps the prediction and classification of medical problems according to patients' characteristics. Suppose a medical researcher collects various information from hospital patients such as demographic data, medical records, or maybe even DNA samples. Further, suppose some subset of the patients has been diagnosed with a specific medical condition. The researcher may then wonder if there is a way to predict the diagnosis of a newly admitted patient regarding this condition, based solely on how the data collected from this patient compares to the previously observed patients. This is known as a binary regression problem, which may be formalized mathematically as follows.

Definition 1.1 (Binary Regression). *Given $n \in \mathbb{N}$, consider a sequence $\{X_i\}_{i=1}^n$ of observations in a measurable space $(\mathcal{X}, \mathscr{X})$ with corresponding binary labels $\{Y_i\}_{i=1}^n \in \{0,1\}^n$. The binary regression problem concerns the modeling and subsequent estimation of the unknown regression function $p : \mathcal{X} \times \{0,1\} \to [0,1]$, given by*

$$p(x,y) = \mathbb{P}(Y = y \mid x)$$

where $Y : \Omega \to \{0,1\}$ is a binary random variable, independent for every $x \in \mathcal{X}$.

In principle, the model could be any non-, semi-, or fully parametric measurable map $p : \mathcal{X} \times \{0,1\} \to [0,1]$. Only the parametric case will be considered here, the most popular model of which uses the so-called logistic map.

Definition 1.2 (Logistic Regression). *In the binary regression setting, suppose $\mathcal{X} = \mathbb{R}^p$. The logistic regression model postulates a linear dependence of the logarithmic odds ratio on X,*

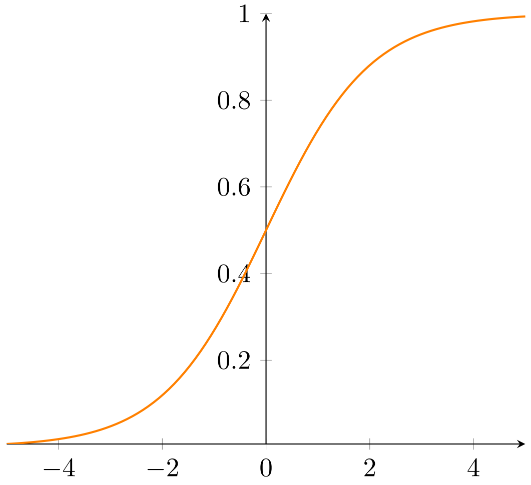


Figure 1.1.: Plot of the logistic function $\psi(x) = (1 + e^{-x})^{-1}$.

parameterized by $\theta \in \mathbb{R}^p$. In particular,

$$\begin{aligned} X_i^{\mathsf{t}}\theta &= \log \frac{p(X_i, 1)}{p(X_i, 0)} \\ &= \log \frac{\mathbb{P}(Y_i = 1 \mid X_i)}{1 - \mathbb{P}(Y_i = 1 \mid X_i)} \end{aligned}$$

which may be rearranged to get $\mathbb{P}(Y_i = 1 \mid X_i) = \left(1 + e^{-X_i^{\mathsf{t}}\theta}\right)^{-1}$.

The function $\psi : \mathbb{R} \to [0, 1]$ given by $\psi(x) = (1 + e^{-x})^{-1}$ is the titular logistic function. From the perspective of generalized linear modeling, ψ plays the role of a link function. A popular alternative is given by $p(x, 1) = \Phi(x^{\mathsf{t}}\theta)$, where $\Phi : \mathbb{R} \to [0, 1]$ is the cumulative distribution function of a standard normal variable. In the linear modeling literature this is also known as the probit link function.

In a heuristic sense, the logistic function “squeezes” the estimated log-odds vector $X\theta \in \mathbb{R}^n$ into the probability simplex $[0, 1]^n$. This incurs an unfortunate loss of resolution, as illustrated in figure 1.1. While ψ separates probabilities close to $\frac{1}{2}$ well, its inverse image of $[0, \varepsilon] \cup [0, 1 - \varepsilon]$ is actually most[1] of $\mathbb{R}$ for arbitrarily small $\varepsilon \in \mathbb{R}_{>0}$. The practical implications of this phenomenon will be discussed in part II.

1.2. High-Dimensional Inference

A feature — as well as a problem — of many data sets encountered in modern statistical practice is their complexity. A particular way in which complexity can manifest itself is high dimensionality. In the hospital patient example given earlier, the number of genes contained

[1] In the sense that the inverse image is the complement of a set with compact closure.

in a DNA sample may be several orders of magnitude larger than the number of observed patients.

In the context of generalized linear models, this creates the problem of solving an underdetermined linear system $X\theta$, where $X \in \mathbb{R}^{n \times p}$ and $p \gg n$. In this case, the operator X cannot be injective by the rank-nullity theorem, making the underlying parameter θ of any generalized linear model based on $X\theta$ unidentifiable.

A common theme in the modern statistical literature dealing with this problem is the restriction to cases where the parameter θ admits a hidden lower-dimensional structure that enables a principled analysis of the problem. This may be a reasonable assumption in many practical settings, for example only few gene mutations may cause a hereditary illness while most mutations go unnoticed. Similarly, a noisy radio signal may consist of few frequencies that contain a message and many more frequencies that do not. In the linear setting, this requires a suitably large number of coordinates of the parameter θ to be zero, a concept known as sparsity.

Definition 1.3 (s-Sparsity). *Given $p \in \mathbb{N}$ and a field F, suppose $\Theta = \bigoplus_{i\in\{1,\dots,p\}}^{\text{ext}} \Theta_i$ is the external direct sum of F-vector spaces. For $s \in \{1,\dots,p\}$, an element $\theta \in \Theta$ is said to be s-sparse if*

$$\sum_{i=1}^{n} \mathbf{1}_{\{\theta_i=0\}} = s.$$

A technical discussion of how sparse θ should be in the logistic model 1.2 will be postponed until chapter 4. For now, the sparse high-dimensional setting mainly serves as a motivator for further modeling choices.

One of the most well-known high-dimensional inference techniques is the so-called LASSO[2] estimator first introduced in [Tib96]. In mathematical terms, it may also be called the ℓ_1-regularized regression estimator.

Definition 1.4 (LASSO Regression). *Consider the linear regression problem $Y = X\theta + \varepsilon$, where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $\varepsilon \sim \mathcal{N}(0, I_n)$. Given a regularization parameter $\lambda \in \mathbb{R}_{>0}$, the LASSO estimator $\widetilde{\theta}$ of the parameter $\theta \in \mathbb{R}^p$ is the minimizer of*

$$\theta \longmapsto \|Y - X\theta\|_2^2 + \lambda\|\theta\|_1.$$

Likewise, a LASSO estimator can be defined for regression. The regularization term $\lambda\|\theta\|_1$ induces sparsity in $\widetilde{\theta}$ by penalizing models that include too many coordinates. The magnitude of λ controls the severity of the penalty, allowing for flexibility. In contrast, the ℓ_2-regularized estimator, also known as ridge regression, shrinks unimportant coordinates to zero, but does not induce sparsity. See [Bv11] and [HTF09] for mathematical, as well as practical, treatments of regularized regression estimators.

[2] **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

1.3. Model Selection Priors

Bayesian methods for sparse high-dimensional problems have received considerable attention in recent years, owing to advances in both computational efficiency and theoretical understanding. A contemporary survey of progress in this area may be found in [BCG21].

The Bayesian approach can model sparsity directly via the chosen prior, offering an elegant and easily interpretable setup. In particular, most of the attention here will be directed towards a specific kind of model selection prior.

Definition 1.5 (Model Selection Prior). *Given a finite sequence $(\Theta_1, \mathscr{G}_1), \ldots, (\Theta_p, \mathscr{G}_p)$ of measurable spaces, denote by $(\Theta, \mathscr{G})$ the Cartesian product $\Theta = \prod_{i=1}^n \Theta_i$, endowed with the product σ-algebra $\mathscr{G} = \bigotimes_{i=1}^n \mathscr{G}_i$. A measure $\Pi : \mathscr{G} \to [0,1]$ is said to be a model selection prior if it admits a hierarchical representation*

$$\begin{aligned} s &\sim \pi_p \\ S \mid s &\sim \mathrm{Unif}_{p,s} \\ \theta_S &\sim \Pi_S \\ \theta_{S^c} &\sim \Pi_{S^c} \end{aligned}$$

where π_p is a prior on $\{1, \ldots, p\}$, $\mathrm{Unif}_{p,s}$ denotes the uniform measure on $\big\{S \in 2^p \mid |S| = s\big\}$, and $\Pi_S : \bigotimes_{i \in S} \mathscr{G}_i \to [0,1]$ and $\Pi_{S^c} : \bigotimes_{i \notin S} \mathscr{G}_i \to [0,1]$ are probability measures.

In case each factor Θ_i is a real vector space, sparsity may be induced by taking $\Pi_{S^c} = \bigotimes_{i \notin S} \delta_0$. More concretely, if $\Theta = \mathbb{R}^p$ and Π_S has Lebesgue density $g_S : \mathbb{R}^S \to \mathbb{R}$, then the prior density of (S, θ) satisfies

$$(S, \theta) \longmapsto \pi_p(|S|) \tfrac{1}{\binom{p}{s}} g_S(\theta_S) \delta_0(\theta_{S^c})$$

Taking π_p binomially distributed and $g_S = \bigotimes_{i=1}^{|S|} g$ for a density $g : \mathbb{R} \to \mathbb{R}$ yields a so-called spike-and-slab prior[3], which will be used in chapter 3.

The main computational drawback of model selection priors is that evaluating the posterior requires summing over all 2^p possible models. This renders computation via Markov chain Monte Carlos methods rather inefficient. Over the course of the next few chapters, an alternative based on optimization techniques will be explored, starting with a review in chapter 2.

[3] Counterintuitively, the continuous part g is referred to as the slab, despite being a “spike in the signal”.

2. A Short Review of Variational Inference

The following chapter presents a short overview of variational inference. In the context of Bayesian statistics, it represents a viable but less understood alternative to traditional sampling based methods. To circumvent the difficult integrals arising from complex prior distributions, the posterior is approximated by simpler distributions, framing the computation as an optimization problem. With some judicious choices of objective function and approximating class, the resulting algorithms can perform multiple orders of magnitude faster than Markov chain Monte Carlo methods, but their theoretical properties are a somewhat subtle issue that has not been studied conclusively in full generality.

The presentation chosen here is slightly more abstract than perhaps usual and phrased in the language of measure theory. For a summary with a more statistical flavor see [BKM17], which also contains a large bibliography of applications in bio-informatics, robotics, natural language processing, and other fields.

2.1. The Abstract Problem

In its most general form, variational inference can be cast as an optimization problem in the space of measures on an arbitrary measurable space $(\Omega, \mathscr{F})$. Recall that $\mathfrak{M}(\Omega)$ denotes the vector space of finite signed measures on $(\Omega, \mathscr{F})$, which may be endowed with the structure of a Banach space via the total variation norm $\|\mu\|_{\mathrm{TV}} = \sup_{A \in \mathscr{F}} \mu(A) + |\inf_{A \in \mathscr{F}} \mu(A)|$.

Definition 2.1 (Abstract Variational Problem). *Let a functional $f : \mathfrak{M}(\Omega) \to \mathbb{R}$ and a set $\mathscr{Q} \subset \mathfrak{M}(\Omega)$ of admissible measures be given, then the abstract variational problem consists of computing the value $\inf_{\mu \in \mathscr{Q}} f(\mu)$ as well as a minimizer $\mu_0 \in \mathscr{Q}$ attaining the infimum, provided the latter exists.*

Without additional assumptions on f and $\mathscr{Q}$ the problem may very well be too general to analyze, but it represents a unified framework for the sequel. Statistical inference is perhaps most naturally framed by considering $\mathscr{Q}$ contained within $\mathscr{P}(\Omega) \subset \mathfrak{M}(\Omega)$, the convex subcollection of probability measures. For example, suppose $\mathbb{P} : \mathscr{F} \to [0,1]$ is a reference probability measure and let a strongly measurable map $X : \Omega \to S$ into a metric space (S, d) be given. Generalized notions of mean and variance can be cast as the minimizer and value of the optimization problem

$$\mathbb{V}(X) := \inf_{\mu \in \mathscr{Q}} \mathbb{E}\left[\int_S d(X,y)^2 \,\mathrm{d}\mu(y)\right]$$

where $\mathscr{Q} = \left\{\delta_x \in \mathscr{P}(S) \mid x \in S\right\}$. Solutions to variational problems of this type are also known as barycenters, see [ALP20; Stu03] for recent accounts of this subject.

A slightly different problem — closer in nature to variational inference as encountered in the machine learning literature — concerns the approximation of problematic probability distributions. In particular, if the functional f captures in some sense the dis-similarity between $\mu \in \mathscr{P}(\Omega)$ and a target measure ν, then computing the minimizer over a suitably chosen class $\mathscr{Q}$ of simpler distributions approximates ν. In the Bayesian setting, the measure ν will typically represent an intractable posterior distribution, for example arising from complex non-conjugate and/or high-dimensional priors. Approximate inference in this sense suggests an alternative to the usual sampling based methods used in such situations.

The advantages and disadvantages of this approach are heavily dependent upon the specific choice of dis-similarity functional f and approximating class $\mathscr{Q}$ of of probability measures. This requires a careful balance between computational tractability and quality of the approximation. Perhaps somewhat surprisingly, the most practical choices for f are statistical divergences, as opposed to genuine metrics on $\mathscr{P}(\Omega)$.

Definition 2.2 (Statistical Divergences)**.** *Fix a measurable space* $(\Omega, \mathscr{F})$ *and* $\mathscr{Q} \subset \mathscr{P}(\Omega)$*. A function* $D : \mathscr{Q} \times \mathscr{Q} \to \mathbb{R}$ *is said to be a statistical divergence if the following hold:*

(i) For all $\mu, \nu \in \mathscr{Q}$*,* $D(\mu, \nu) \geq 0$.

(ii) For all $\mu, \nu \in \mathscr{Q}$*,* $D(\mu, \nu) = 0$ *if, and only if,* $\mu = \nu$.

Though the rigorous study of parametric statistical models via divergences could be said to have begun with [KL51], it is a still emerging field at the intersection of differential geometry, information theory, and probability theory. See [Ay+17] for a current account of the mathematical fundamentals underpinning the subject.

2.2. The Kullback-Leibler Divergence

The most popular divergence by far in the variational inference literature is the so-called Kullback-Leibler divergence, first introduced in [KL51]. The monograph [Kul97] by one of the namesake inventors gives a (for its time) complete account of the classical theory of frequentist statistics through the lens of this divergence.

An accessible introduction covering basic examples of variational inference with respect to the Kullback-Leibler divergence may be found in chapter 10 of [Bis06]. In recent years, some works have explored the use of more general classes of divergences in variational inference, see for example [Amb+18; Ran+16; SBK20], but the Kullback-Leibler divergence remains the gold standard in practice.

Definition 2.3 (Kullback-Leibler Divergence)**.** *Let* $\mu, \nu \in \mathscr{P}(\Omega)$ *be given. Provided that* $\nu(A) = 0$ *implies* $\mu(A) = 0$ *for all* $A \in \mathscr{F}$*, the Kullback-Leibler divergence* $\mathrm{KL}(\mu, \nu)$ *between* μ *and* ν *is given by*

$$\mathrm{KL}(\mu, \nu) = \int_\Omega \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu.$$

It follows immediately that $\mathrm{KL}(\cdot,\cdot)$ is a divergence in the sense of definition 2.2. As opposed to a genuine metric, it is asymmetric and does not satisfy the triangle inequality. If both μ and ν are dominated by a third measure, say the Lebesgue measure $\mathrm{d}x$ when $\Omega = \mathbb{R}^d$, then the Kullback-Leibler divergence between their respective densities $p, q : \mathbb{R}^d \to \mathbb{R}$ reduces to

$$\mathrm{KL}(p,q) = \int_{\mathbb{R}^d} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x.$$

A first convergence result is given by the following well-known inequality, see section 4.11 of [BLM16] for a modern probabilistic proof.

Lemma 2.4 (Kullback-Csiszár-Pinsker Inequality). *Suppose μ, $\nu \in \mathscr{P}(\Omega)$ satisfy the conditions mentioned in the previous definition, then*

$$\|\mu - \nu\|_{\mathrm{TV}}^2 \leq \tfrac{1}{2}\mathrm{KL}(\mu, \nu).$$

The result seems deceptively powerful, implying that convergence in Kullback-Leibler divergence entails convergence in total variation. The required choice of variational class $\mathscr{Q}$ makes this matter more subtle. Taking $\mathscr{Q}$ too simple, every $\mu \in \mathscr{Q}$ may be "far away" from the target measure ν in Kullback-Leibler divergence. Conversely, should $\mathscr{Q}$ be too rich, then $\mathrm{KL}(\mu_0, \nu)$ may be small for some $\mu_0 \in \mathscr{Q}$, but actually computing the minimizer could become intractable.

2.3. The Mean-Field Family

Just as the Kullback-Leibler divergence is the usual choice of objective function, so-called mean-field families are the most accessible choice of admissible measures $\mathscr{Q}$. There are of course works exploring the use of more generic variational families, see for example [HB15; JRH20].

Definition 2.5 (Mean-Field Families). *Suppose $\Omega = \prod_{n=1}^{\infty} \Omega_n$ is endowed with the product σ-algebra $\mathscr{F}$, induced by the coordinate projections $\pi_n : \Omega \to (\Omega_n, \mathscr{F}_n)$. A collection $\mathscr{Q} \subset \mathscr{P}(\Omega)$ is said to be a mean-field family if every $\mu \in \mathscr{Q}$ is of the form $\mu = \bigotimes_{n=1}^{\infty} \mu_n$, where $\mu_n \in \mathscr{P}(\Omega_n)$.*

Such collections may be taken non-empty, provided that each $\mathscr{P}(\Omega_n)$ is non-empty, see theorem 8.2.2 of [Dud02]. When $\Omega = \mathbb{R}^d$, a particularly common example is given by taking each μ_n as the measure induced by a Gaussian distribution. In this case, the family can be parametrized by a vector $(\mu, \sigma) \in \mathbb{R}^d \times \mathbb{R}^d_{>0}$.

The power of the mean-field approach lies in the independence of the coordinates. For example, the posterior arising from the variable selection prior 1.5 contains a practically intractable sum over the 2^p possible models, rendering sampling methods inefficient. A mean-field distribution models the inclusion of each dimension separately, bypassing the problem. Chapter 3 concerns the rigorous derivation of such an algorithm for the sparse logistic model.

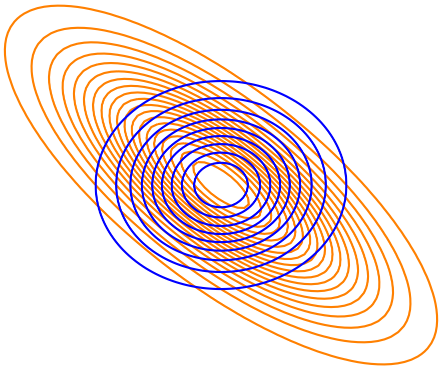


Figure 2.1.: Contours of a multivariate Gaussian density with diagonal covariance matrix (blue, small circular contours) and non-diagonal covariance matrix (orange, large ellipsoidal contours).

The obvious drawback of mean-field measures is their inability to account for correlation between the variables. Consider again the variational class of multivariate Gaussian distributions with diagonal covariance matrix. If the target distribution also has a multivariate Gaussian distribution, but with non-diagonal covariance matrix, then the variational approximation may look similar to figure 2.1. By definition, the Kullback-Leibler objective heavily penalizes putting mass onto sets where the target measure has small density because the term $\log \frac{p(x)}{q(x)}$ grows larger in this case. The reverse is not true, integrating against $p(x)\,\mathrm{d}x$ mitigates the penalty incurred by putting little mass onto sets where the target measure has large density. Accordingly, mean-field variational inference is prone to underestimating the variance of the target distribution. More principled discussions of this phenomenon may be found in appendix B of [LT16], or section 1.3 of [TS11]. As demonstrated empirically in chapter 6, reliable uncertainty quantification may still be possible in some cases.

2.4. The Emergence of Theory

Results concerning the theoretical properties of variational inference have started appearing only in recent years. A particularly recent development concerns the rate of contraction of the approximate posterior, contextualizing the variational approach within the now classical framework of posterior contraction by [GGv00].

Definition 2.6 (Posterior Contraction Rates). *Denote by $\theta \sim \Pi_n$ the posterior arising from $n \in \mathbb{N}$ observations $X_1, \dots, X_n$ in a measurable space $(X, \mathscr{X})$ and some prior Π on the semi-metric parameter set (Θ, d) of a parameterized statistical model $\theta \mapsto \mathbb{P}_\theta$. A sequence $\{\alpha_n\}_{n\in\mathbb{N}} \subset \mathbb{R}_{\geq 0}$ is said to be a posterior contraction rate for Π_n at $\theta_0 \in \Theta$ if*

$$\Pi_n\big(\theta \in \Theta \mid d(\theta, \theta_0) \geq L_n \alpha_n\big) \longrightarrow 0$$

in $\mathbb{P}_{\theta_0}$-probability as $n \to \infty$, for any sequence $\{L_n\}_{n\in\mathbb{N}} \subset \mathbb{R}_{\geq 0}$ diverging to infinity.

The general theory of contraction rates in parametric, as well as non-parametric, settings has matured to a reasonable state of completion for models dominated by a reference measure. The expansive textbook [Gv17] collects most of the relevant developments. The most successful approach has been dubbed "prior mass and testing", as illustrated by the following general theorem:

Theorem 2.7 (Theorem 8.11 of [Gv17]). *Denote by $\mathscr{P}$ the collection of probability measures on $\mathcal{X}$ dominated by some reference measure μ on a sample space $(\mathcal{X}, \mathscr{X})$. Consider a sequence $\{\Pi_n\}_{n\in\mathbb{N}}$ of prior distributions on $\mathscr{P}$ and suppose that $p_0 \in \mathscr{P}$ is the true density of given i.i.d. observations $X_1, \ldots, X_n$. Let $d : \mathscr{P} \times \mathscr{P} \to \mathbb{R}_{\geq 0}$ be a metric such that for every $n \in \mathbb{N}$, $\varepsilon \in \mathbb{R}_{>0}$, and $p_1 \in \mathscr{P}$ with $d(p_1, p_0) > \varepsilon$, there exist constants $\xi, K \in \mathbb{R}_{>0}$ and a measurable mapping $\psi_n : \mathcal{X}^n \to [0,1]$ satisfying*

$$\int \psi_n \,\mathrm{d}(p_0\mu)^n \leq e^{-Kn\varepsilon^2}$$

$$\sup_{p:d(p,p_1)>\xi\varepsilon} \int 1 - \psi_n \,\mathrm{d}(p\mu)^n \leq e^{-Kn\varepsilon^2}.$$

Given a sequence of partitions $\mathscr{P} = \mathscr{P}_{n,1} \cup \mathscr{P}_{n,2}$ and $\varepsilon_n, \overline{\varepsilon}_n \geq \sqrt{n}$, the posterior contraction rate at p_0 is ε_n, provided that the following are satisfied for sufficiently large j:

(i) $\dfrac{\Pi_n\big(p \in \mathscr{P} \mid j\varepsilon_n < d(p,p_0) \leq 2j\varepsilon_n\big)}{\Pi_n\big(B_0(p_0,\varepsilon_n)\big)} \leq e^{Kn\varepsilon_n^2 j^2/2}$

(ii) $\sup_{\varepsilon\geq\varepsilon_n} \log \mathscr{N}\Big(\xi\varepsilon, \big\{p \in \mathscr{P}_{n,1} \mid d(p,p_0) \leq 2\varepsilon\big\}, d\Big) \leq n\varepsilon_n^2$

(iii) $\dfrac{\Pi_n(\mathscr{P}_{n,2})}{\Pi_n\big(B_0(p_0,\overline{\varepsilon}_n)\big)} = o\big(e^{-D_n n\overline{\varepsilon}_n^2}\big)$ *for some $D_n \to \infty$,*

where $B_0(p_0,\varepsilon) = \big\{p \in \mathscr{P} \mid \mathrm{KL}(p_0,p) < \varepsilon^2\big\}$ and $\mathscr{N}(\varepsilon, A, d)$ is the minimal number of d-balls with radius ε covering A.

A general result quantifying the contraction rate of the variational posterior under very similar conditions was recently obtained by [ZG20]. Similar contraction results inspired by fractional posteriors appeared independently around the same time in [AR20; YPB20]. In the parametric setting, a first variational Bernstein-von Mises type result has been obtained as well, see [WB19].

A more particular approach, analyzing the contractive properties of variational posteriors arising from spike-and-slab priors in high-dimensional linear and logistic regression can be found in [RS21] and [RSC20]. The latter presents the results most relevant to the sparse logistic model introduced in chapter 1. A summary of the pertinent conditions and theorems will be given in chapter 4.

3. A Variational Algorithm for the Sparse Logistic Model

The main goal of this chapter is to derive a variational algorithm for the Bayesian logistic regression model with spike-and-slab prior. The novelty of the approach lies in choosing heavy-tailed Laplace densities for the prior, as opposed to the commonplace Gaussian densities. This will also be the source of numerous implementational issues, to be discussed in part II. Some of this material has been published as part of [RSC20], though the presentation here will be much more complete.

3.1. Defining the Variational Problem

The model selection prior as given in definition 1.5 is not entirely suitable for some of the computations in this chapter, warranting a reformulation involving auxiliary binary variables.

Definition 3.1 (Hierarchical Spike-and-Slab Prior). *Let a density function $g : \mathbb{R} \to \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}_{>0}$ be given. The spike-and-slab prior Π for $\theta \in \mathbb{R}^p$ with slab-density g may be defined via the hierarchical scheme*

$$\begin{aligned} w_j \mid \alpha, \beta &\overset{i.i.d.}{\sim} \mathrm{Beta}(\alpha, \beta) \\ z_j \mid w_j &\overset{i.i.d.}{\sim} \mathrm{Binom}(w_j) \\ \theta_j \mid z_j &\overset{ind.}{\sim} z_j g + (1 - z_j)\delta_0. \end{aligned}$$

Intuitively, the $w_j \in [0,1]$ represent the expected fraction of non-zero coefficients in θ. Any prior knowledge of the sparsity level may be included in the model by adjusting α and β. As is customary, the ultimate object of interest is the posterior distribution for θ arising from Π.

Proposition 3.2 (Spike-and-Slab Posterior). *Given data $X \in \mathbb{R}^{n \times p}$ and $Y \in \{0,1\}^n$, let Π be the hierarchical spike and slab prior, as in definition 3.1, for the parameter $\theta \in \mathbb{R}^p$ of the logistic regression model 1.2. The posterior probability of Lebesgue measurable $\Theta \subset \mathbb{R}^p$ may be evaluated as*

$$\Pi(\theta \in \Theta \mid X, Y) = \frac{\int_\Theta \prod_{i=1}^n p_\theta(X_i)^{Y_i} \big(1 - p_\theta(X_i)\big)^{1-Y_i} \, \mathrm{d}\Pi(\theta)}{\int_{\mathbb{R}^p} \prod_{i=1}^n p_\theta(X_i)^{Y_i} \big(1 - p_\theta(X_i)\big)^{1-Y_i} \, \mathrm{d}\Pi(\theta)}.$$

Proof. Writing $p_i = p_\theta(X_i)$ for each $i \in \{1, \dots, n\}$, the likelihood of the logistic model is given by $\mathcal{L}(\theta \mid X, Y) = \prod_{i=1}^n p_i^{Y_i}(1 - p_i)^{1-Y_i}$. A simple application of Bayes' rule concludes the proof. □

Notice that integrating against $\mathrm{d}\Pi$ requires summing over all 2^p possible sparse configurations of θ since

$$\Pi(\theta \in \Theta) = \sum_{S \subset \{1,\ldots,p\}} \prod_{j=1}^{p} \mathbb{P}\big(z_j = \mathbf{1}_{\{j \in S\}}\big) \cdot \left(z_j \int_{\pi_j(\Theta)} g \ \mathrm{d}\mu + (1 - z_j) \cdot \delta_0\big(\pi_j(\Theta)\big) \right)$$

for any Lebesgue measurable $\Theta \subset \mathbb{R}^p$. This renders both the analytical evaluation of $\Pi(\cdot \mid X, Y)$ and the simulation of posterior draws via Markov chain Monte Carlo methods futile. Applying the variational techniques discussed in chapter 2 must begin with a suitable class of approximation measures circumventing the combinatorial explosion.

Definition 3.3 (Spike-and-Slab Mean-Field Family). *For fixed $p \in \mathbb{N}$, the mean-field family for the hierarchical spike-and-slab prior 3.1 is defined as*

$$\mathscr{P}_{\mathrm{MF}} = \left\{ \left(\bigotimes_{j=1}^{p} \gamma_j \mathcal{N}(\mu_j, \sigma_j^2) + (1 - \gamma_j)\delta_0 \right) \in \mathscr{P}\big(\mathbb{R}^p\big) \;\middle|\; \mu_j \in \mathbb{R}, \sigma_j^2 \in \mathbb{R}_{>0}, \gamma_j \in [0,1] \right\}$$

The element of $\mathscr{P}_{\mathrm{MF}}$ corresponding to a particular choice of $\mu \in \mathbb{R}^p$, $\sigma^2 \in \mathbb{R}^p_{>0}$, and $\gamma \in [0,1]^p$ will be denoted by $P_{\mu,\sigma,\gamma}$ and its associated expectation operator by $\mathbb{E}_{\mu,\sigma,\gamma}$.

Allowing individual inclusion probabilities γ_j for each coordinate of θ breaks their interdependence through the z_j. Accordingly, the Kullback-Leibler divergence between $\mathbb{P}_{\mu,\sigma,\gamma}$ and $\Pi(\cdot \mid X, Y)$ should factorize over $\{1, \ldots, p\}$, as long as the expectation is taken with respect to a measure in the mean-field family.

Definition 3.4 (Variational Approximate Posterior). *Let $\mathrm{d}\Pi(\cdot \mid X, Y)$ be the posterior of the logistic regression model, as given in proposition 3.2, and suppose $\mathscr{P}_{\mathrm{MF}}$ is the mean-field family of definition 3.3. The mean-field variational approximation Π^* of $\Pi(\cdot \mid X, Y)$ is defined via*

$$\begin{aligned} \Pi^* &= \arg\min\nolimits_{P_{\mu,\sigma,\gamma} \in \mathscr{P}_{\mathrm{MF}}} \mathrm{KL}\big(P_{\mu,\sigma,\gamma}, \Pi(\cdot \mid X, Y)\big) \\ &= \arg\min\nolimits_{P_{\mu,\sigma,\gamma} \in \mathscr{P}_{\mathrm{MF}}} \int \log\left(\frac{\mathrm{d}P_{\mu,\sigma,\gamma}}{\mathrm{d}\Pi(\cdot \mid X, Y)} \right) \mathrm{d}P_{\mu,\sigma,\gamma}. \end{aligned}$$

As desired, the integral defining Π^* factorizes over the coordinates of the model, which will eventually allow for tractable optimization along each dimension $j = 1, \ldots, p$. Indeed,

$$\begin{aligned} \int \log\left(\frac{\mathrm{d}P_{\mu,\sigma,\gamma}}{\mathrm{d}\Pi(\cdot \mid X, Y)} \right) \mathrm{d}P_{\mu,\sigma,\gamma} &= \int \log\left(\frac{\mathrm{d}P_{\mu,\sigma,\gamma}(\theta)}{\mathcal{L}(\theta \mid X, Y)\ \mathrm{d}\Pi(\theta)} \right) \mathrm{d}P_{\mu,\sigma,\gamma}(\theta) + C \\ &= \int \log\left(\frac{\mathrm{d}P_{\mu,\sigma,\gamma}}{\mathrm{d}\Pi}(\theta) \right) - \ell(\theta \mid X, Y)\ \mathrm{d}P_{\mu,\sigma,\gamma}(\theta) + C \\ &= \mathrm{KL}(P_{\mu.\sigma,\gamma}, \Pi) - \int \ell(\theta \mid X, Y)\ \mathrm{d}P_{\mu,\sigma,\gamma}(\theta) + C \end{aligned}$$

for some real constant C, which normalizes the posterior. The Radon-Nikodym derivative $\mathrm{d}P_{\mu,\sigma,\gamma}/\mathrm{d}\Pi$ may be written as a product over $\{1, \ldots, p\}$ due to the factorized structure of $P_{\mu,\sigma,\gamma}$. Unfortunately, the term involving $\ell(\theta \mid X, Y)$ still defies explicit computation, requiring a somewhat technical detour before revisiting the Kullback-Leibler divergence.

3.2. Bounding the Likelihood

Recall that the likelihood $\mathcal{L}(\theta \mid X, Y)$, arising from observed data $X \in \mathbb{R}^{n\times p}$ and $Y \in \{0,1\}^p$, for the parameter $\theta \in \mathbb{R}^p$ of the logistic regression model 1.2 is given by

$$\begin{aligned}\mathcal{L}(\theta \mid X, Y) &= \prod_{i=1}^{n} p_\theta(X_i)^{Y_i}\big(1-p_\theta(X_i)\big)^{1-Y_i} \\ &= \prod_{i=1}^{n}\left(\frac{1}{1+e^{-X_i^{\mathrm{t}}\theta}}\right)^{Y_i}\left(1-\frac{1}{1+e^{-X_i^{\mathrm{t}}\theta}}\right)^{1-Y_i}.\end{aligned}$$

Subsequently, terms of the problematic form $\log\left(1+e^{-X_i^{\mathrm{t}}\theta}\right)$ emerge in the log-likelihood $\ell(\theta \mid X, Y)$. These terms admit no closed-form solution when integrated against $\mathrm{d}P_{\mu,\sigma,\gamma}$, demanding a different approach. To start, rearranging the log-likelihood yields

$$\begin{aligned}\ell(\theta \mid X, Y) &= \sum_{i=1}^{n} Y_i \log\left(\frac{1}{1+e^{-X_i^{\mathrm{t}}\theta}}\right) + (1-Y_i)\log\left(1-\frac{1}{1+e^{-X_i^{\mathrm{t}}\theta}}\right) \\ &= \sum_{i=1}^{n} Y_i \log\left(\frac{1}{1+e^{-X_i^{\mathrm{t}}\theta}}\right) + (1-Y_i)\log\left(\frac{e^{-X_i^{\mathrm{t}}\theta}}{1+e^{-X_i^{\mathrm{t}}\theta}}\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{e^{-X_i^{\mathrm{t}}\theta}}{1+e^{-X_i^{\mathrm{t}}\theta}}\right) + Y_i X_i^{\mathrm{t}}\theta + Y_i\log\left(1+e^{-X_i^{\mathrm{t}}\theta}\right) - Y_i\log\left(1+e^{-X_i^{\mathrm{t}}\theta}\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{1+e^{X_i^{\mathrm{t}}\theta}}\right) + Y_i X_i^{\mathrm{t}}\theta.\end{aligned}$$

With some ingenuity and basic convex analysis, the first term in the sum may be lower-bounded. This technique, based on convex conjugation, seems to have first appeared in [JJ00], who investigated variational inference for classical logistic regression models. For completeness, a proof of the following lemma will be given:

Lemma 3.5 (Jaakkola-Jordan [JJ00]). *Let $\psi : \mathbb{R} \to [0,1]$ be the standard logistic function $\psi(x) = (1+e^{-x})^{-1}$, then*

$$\log\psi(x) \geq \frac{x-\eta}{2} + \log\psi(\eta) - \frac{1}{4\eta}\tanh(\eta/2)(x^2-\eta^2)$$

for all $\eta \in \mathbb{R}$.

Proof. First, rearrange some terms to get

$$\begin{aligned}\log\psi(x) &= -\log(1+e^{-x}) \\ &= -\log\left(\frac{e^{x/2}}{e^{x/2}} + \frac{e^{-x/2}}{e^{x/2}}\right) \\ &= \frac{x}{2} - \log\left(e^{x/2}+e^{-x/2}\right) \\ &= \frac{x}{2} - \log\left(\sqrt{e^x}+\sqrt{e^{-x}}\right).\end{aligned}$$

This observation may seem arbitrary, but the second term happens to be convex in the variable x^2. Indeed, the first two derivatives with respect to x^2 are given by

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x^2}\log\left(\sqrt{e^x}+\sqrt{e^{-x}}\right) &= \frac{1}{2x}\frac{\mathrm{d}}{\mathrm{d}x}\log\left(\sqrt{e^x}+\sqrt{e^{-x}}\right) \\
&= \frac{1}{4x}\frac{1}{\sqrt{e^x}+\sqrt{e^{-x}}}\left(\frac{e^x}{\sqrt{e^x}}-\frac{e^{-x}}{\sqrt{e^{-x}}}\right) \\
&= \frac{1}{4x}\frac{\sqrt{e^x}-\sqrt{e^{-x}}}{\sqrt{e^x}+\sqrt{e^{-x}}} \\
&= \frac{1}{4x}\tanh x/2 \\
\frac{\mathrm{d}}{\mathrm{d}x^2}\left(\frac{1}{4x}\tanh x/2\right) &= \frac{1}{2x}\frac{1}{(4x)^2}\left(2x\cosh^{-2}(x)-4\tanh(x/2)\right) \\
&= \frac{x-2\sinh(x/2)\cosh(x/2)}{16x^3\cosh^2(x/2)}
\end{aligned}$$

Observing that the numerator in the very last term is a continuous odd function with 0 as a fixed point, it suffices to show that

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(x-2\sinh(x/2)\cosh(x/2)\big) = 1-\sinh^2(x/2)-\cosh^2(x/2)$$

is negative for all $x\in\mathbb{R}$. This follows immediately from the identity $\sinh^2(x)+\cosh^2(x)=\cosh(2x)$ and the fact that $\cosh(x)\geq 1$ for all $x\in\mathbb{R}$. Subsequently, a first order expansion concedes the lower bound

$$\begin{aligned}
\log\psi(x) &\geq \frac{x}{2}-\log\left(e^{\eta/2}+e^{-\eta/2}\right)-\frac{1}{4\eta}\tanh(\eta/2) \\
&= \frac{x-\eta}{2}+\log\psi(\eta)-\frac{1}{4\eta}\tanh(\eta/2)\big(x^2-\eta^2\big)
\end{aligned}$$

where $\eta\in\mathbb{R}$ is arbitrary. □

The problematic terms of the log-likelihood may be written in the form $\psi\big(-X_i^{\mathrm{t}}\theta\big)$. Introducing a free parameter $\eta_i\in\mathbb{R}$ for each $i=1,\dots,n$ and substituting the lower bound of the proposition results in

$$\begin{aligned}
\ell(\theta\mid X,Y) &\geq \sum_{i=1}^{n}\frac{-X_i^{\mathrm{t}}\theta-\eta}{2}+\log\psi(\eta_i)-\frac{1}{4\eta_i}\tanh(\eta_i/2)\big(X_i^{\mathrm{t}}\theta^2-\eta_i^2\big)+Y_iX_i^{\mathrm{t}}\theta \\
&= \sum_{i=1}^{n}\log\psi(\eta_i)-\frac{\eta_i}{2}+(Y_i-1/2)X_i^{\mathrm{t}}\theta-\frac{1}{4\eta_i}\tanh(\eta_i/2)\big(X_i^{\mathrm{t}}\theta^2-\eta_i^2\big)
\end{aligned}$$

Letting $\eta\in\mathbb{R}^n$ be the vector of free parameters, the expression on the last line will be denoted by $f(\theta,\eta)$ from here on. The expected value of $f(\theta,\eta)$ depends on the variational parameters through the measure $P_{\mu,\sigma,\gamma}$ so it may be desirable to optimize η with the aim of tightening the lower bound. Fortunately, this optimization problem separates over the coordinates of η, each of which admits a maximizer.

Proposition 3.6. *For each non-zero $\alpha \in \mathbb{R}$, define the function $g_\alpha : \mathbb{R} \to \mathbb{R}$ via*

$$g_\alpha(x) = \log \psi(x) - \frac{x}{2} - \frac{1}{4x} \tanh(x/2)\big(\alpha^2 - x^2\big),$$

then g_α is symmetric around 0 *and admits maximizers at $x = \pm\alpha$.*

Proof. Recalling that $\tanh(-x) = -\tanh(x)$ and $\psi(-x) = 1 - \psi(x)$, symmetry follows by direct computation of

$$\begin{aligned}
\log \psi(-x) + \frac{x}{2} &= \log\Big(e^{x/2}\big(1 - \psi(x)\big)\Big) \\
&= \log\left(e^{x/2}\left(1 - \frac{1}{1+e^{-x}}\right)\right) \\
&= \log\left(e^{x/2}\frac{e^{-x}}{1+e^{-x}}\right) \\
&= \log\left(e^{-x/2}\frac{1}{1+e^{-x}}\right) \\
&= \log \psi(x) - \frac{x}{2}.
\end{aligned}$$

Since $\mathrm{d}\tanh(x)/\mathrm{d}x = \cosh(x)^{-2}$ and $\psi(x) = \tanh(x/2)/2 + 1/2$, the derivative of g_α with respect to x takes the form

$$\begin{aligned}
\frac{\partial g_\alpha(x)}{\partial x} &= -\frac{\mathrm{d}\log(1+e^{-x})}{\mathrm{d}x} - \frac{1}{2} + \frac{1}{4}\left(\frac{\mathrm{d}\tanh(x/2)x}{\mathrm{d}x} + \alpha^2 \frac{\mathrm{d}\tanh(x/2)/x}{\mathrm{d}x}\right) \\
&= \frac{e^{-x}}{1+e^{-x}} - \frac{1}{2} + \frac{1}{4}\left(\tanh(x/2) + \frac{x}{2\cosh(x/2)^2} + \alpha^2 \frac{2\tanh(x/2) - \cosh(x/2)^2 x}{2x^2}\right) \\
&= \frac{1}{2}\tanh(-x/2) + (x^2 + \alpha^2)\frac{\tanh(x/2)}{4x^2} + \frac{x^2 - \alpha^2}{8x\cosh(x/2)^2} \\
&= \frac{1}{2}\tanh(-x/2) + \frac{1}{2}\tanh(x/2) + (\alpha^2 - x^2)\frac{\tanh(x/2)}{4x^2} + \frac{x^2 - \alpha^2}{8x\cosh(x/2)^2} \\
&= (\alpha^2 - x^2)\left(\frac{\tanh(x/2)}{4x^2} - \frac{1}{8x\cosh(x/2)^2}\right)
\end{aligned}$$

Clearly, $x^2 = \alpha^2$ is a critical point of g_α and to conclude the proof it suffices to show that this indeed constitutes a maximum. To this end, observe that the second partial derivative with respect to x is given by

$$\begin{aligned}
\frac{\partial^2 g_\alpha(x)}{\partial^2 x} &= (\alpha - x^2)\tfrac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\tanh(x/2)}{4x^2} - \frac{1}{8x\cosh(x/2)^2}\right) - \frac{\tanh(x/2)}{2x} + \frac{1}{4\cosh(x/2)^2} \\
&= (\alpha - x^2)\tfrac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\tanh(x/2)}{4x^2} - \frac{1}{8x\cosh(x/2)^2}\right) - \frac{2\sinh(x/2)\cosh(x/2) - x}{4x\cosh(x/2)^2} \\
&= (\alpha - x^2)\tfrac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\tanh(x/2)}{4x^2} - \frac{1}{8x\cosh(x/2)^2}\right) - \frac{\sinh(x) - x}{4x\cosh(x/2)^2},
\end{aligned}$$

where the last equality follows from the half-argument formula $\sinh(x)\cosh(x) = \sinh(2x)/2$. The first term vanishes for non-zero α whenever $x = \pm\alpha$ and $\sinh(x)/x > 1$ for all non-zero $x \in \mathbb{R}$, which concludes the argument. □

In view of this result, the free parameters η_i may be restricted to $\mathbb{R}_{>0}$. Accordingly, the tightest lower bound on the expected value of $f(\theta,\eta)$, achieved at

$$\widetilde{\eta} = \arg\min_{\eta \in \mathbb{R}^n_{>0}} \mathbb{E}_{\mu,\sigma,\gamma}\big[f(\theta,\eta)\big],$$

satisfies $\widetilde{\eta}_i^2 = \mathbb{E}_{\mu,\sigma,\gamma}\big[(X_i^{\mathrm{t}}\theta)^2\big]$ in each coordinate. Computing this expectation forms the last piece of the lower bound puzzle.

Proposition 3.7. *Suppose $\theta \sim P_{\mu,\sigma,\gamma}$ for some fixed $P_{\mu,\sigma,\gamma} \in \mathscr{P}_{\mathrm{MF}}$, as defined in 3.3, then*

$$\begin{aligned}\mathbb{E}_{\mu,\sigma,\gamma}\big[X_i^{\mathrm{t}}\theta\big] &= \sum_{j=1}^{p} \gamma_j \mu_j X_{ij}\\ \mathbb{E}_{\mu,\sigma,\gamma}\big[(X_i^{\mathrm{t}}\theta)^2\big] &= \sum_{j=1}^{p} \gamma_j X_{ij}^2(\mu_j^2 + \sigma_j^2) + \sum_{j=1}^{p}\sum_{k\neq j}(\gamma_j\mu_j X_{ij})(\gamma_k\mu_k X_{ik}).\end{aligned}$$

Proof. Recall that $\theta_j \sim \gamma_j \mathcal{N}(\mu_j,\sigma_j^2) + (1-\gamma_j)\delta_0$, independently for each coordinate, with respect to $P_{\mu,\gamma,\sigma}$. Accordingly, $\mathbb{E}_{\mu_\sigma,\gamma}[\theta_j] = \gamma_j\mu_j$ from which the first claim follows by linearity.

Next, note that $\theta_j \mathbf{1}_{\{z_j=1\}} \neq 0$ and $\theta_j \mathbf{1}_{\{z_j=0\}} = 0$, both $P_{\mu,\sigma,\gamma}$-almost surely. Expanding the square, the fully factorized structure of $P_{\mu,\sigma,\gamma}$ yields

$$\begin{aligned}\mathbb{E}_{\mu,\sigma,\gamma}\big[(X_i^{\mathrm{t}}\theta)^2\big] &= \mathbb{E}_{\mu,\sigma,\gamma}\left[\left(\sum_{j=1}^{p} X_{ij}\theta_j \mathbf{1}_{\{z_j=1\}} + \sum_{j=1}^{p} X_{ij}\theta_j \mathbf{1}_{\{z_j=0\}}\right)^2\right]\\ &= \mathbb{E}_{\mu,\sigma,\gamma}\left[\left(\sum_{j=1}^{p} X_{ij}\theta_j \mathbf{1}_{\{z_j=1\}}\right)^2\right] + \mathbb{E}_{\mu,\sigma,\gamma}\left[\underbrace{\left(\sum_{j=1}^{p} X_{ij}\theta_j \mathbf{1}_{\{z_j=0\}}\right)^2}_{\overset{a.s.}{=}0}\right]\\ &\quad + 2\mathbb{E}_{\mu,\sigma,\gamma}\left[\left(\sum_{j=1}^{p} X_{ij}\theta_j \mathbf{1}_{\{z_j=1\}}\right)\underbrace{\left(\sum_{j=1}^{p} X_{ij}\theta_j \mathbf{1}_{\{z_j=0\}}\right)}_{\overset{a.s.}{=}0}\right]\\ &= \mathbb{E}_{\mu,\sigma,\gamma}\left[\left(\sum_{j=1}^{p} X_{ij}\theta_j \mathbf{1}_{\{z_j=1\}}\right)^2\right]\end{aligned}$$

By definition, the expectation of both $\mathbf{1}_{\{z_j=1\}}$ and its square are equal to $P_{\mu,\sigma,\gamma}(z_j = 1) = \gamma_j$. Expanding the squared sum in the previous display, independence of the z_j and θ_j results in

$$\begin{aligned}\mathbb{E}_{\mu,\sigma,\gamma}\big[(X_i^{\mathrm{t}}\theta)^2\big] &= \sum_{j=1}^p \mathbb{E}_{\mu,\sigma,\gamma}\Big[\big(X_{ij}\theta_j\mathbf{1}_{\{z_j=1\}}\big)^2\Big] + \sum_{j=1}^p\sum_{k\neq j}\mathbb{E}_{\mu,\sigma,\gamma}\Big[X_{ij}\theta_j\mathbf{1}_{\{z_j=1\}}X_{ik}\theta_k\mathbf{1}_{\{z_k=1\}}\Big]\\ &= \sum_{j=1}^p \gamma_j X_{ij}^2(\mu_j^2+\sigma_j^2) + \sum_{j=1}^p\sum_{k\neq j}(\gamma_j\mu_jX_{ij})(\gamma_k\mu_kX_{ik}),\end{aligned}$$

which concludes the proof. □

3.3. Coordinate Ascent Equations

To proceed, a slab-density g will have to be chosen first. An algorithm with Gaussian slabs seems to have first appeared in [CS12] and features surprisingly simple update equations. From a theoretical standpoint, it may be desirable to choose a density with heavier tails, for example a Laplace density, a phenomenon first discussed in [Cv12]. This choice enables bounds on the contraction rate of the approximate posterior towards a sufficiently sparse truth in the frequentist sense, as shown in [RSC20]. The main goal of this section is to rigorously derive the update equations for the Laplace spike-and-slab prior. With some small adjustments, the Gaussian case follows by a similar argument and will be discussed very briefly for completeness. While the Gaussian prior results in closed-form updates, the Laplace prior requires numerical optimization of the update equations, which will be treated in part II.

Now that all preliminaries are in place, the mean-field parameters μ, σ, and γ may be optimized. Throughout this section, suppose that a free parameter $\eta \in \mathbb{R}^n_{>0}$, optimal or not, has been fixed. Substituting the bound on the log-likelihood, the objective is to minimize the right-hand side of

$$\mathrm{KL}\big(P_{\mu,\sigma,\gamma}, \Pi(\cdot \mid X, Y)\big) \leq \mathrm{KL}(P_{\mu,\sigma,\gamma}, \Pi) - \mathbb{E}_{\mu,\sigma,\gamma}\big[f(\theta,\eta)\big] + C$$

with respect to the mean-field parameters. The constant $C \in \mathbb{R}$ arises from the normalizing factor in Bayes' rule, but in the sequel it will denote any constant independent from the current objective parameters.

Recall that a Laplace density with location $\nu \in \mathbb{R}$ and scale $\lambda \in \mathbb{R}_{>0}$ takes the form $f_{\nu,\lambda}(x) = \frac{1}{2\lambda}e^{-|x-\nu|/\lambda}$. Similarly, $\phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$ will denote the Gaussian density with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_{>0}$. Due to the factorized structure of the mean-field class,

$$\begin{aligned}\mathrm{KL}(P_{\mu,\sigma,\gamma}, \Pi) &= \int \log\left(\frac{\mathrm{d}P_{\mu,\sigma,\gamma}}{\mathrm{d}\Pi}\right)\,\mathrm{d}P_{\mu,\sigma,\gamma}\\ &= \int\cdots\int \sum_{j=1}^p \log\left(\frac{\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j}}{\mathrm{d}\Pi_j}\right)\,\mathrm{d}P_{\mu_1,\sigma_1,\gamma_1}\cdots\mathrm{d}P_{\mu_p,\sigma_p,\gamma_p}\end{aligned}$$

where $P_{\mu_j,\sigma_j,\gamma_j}$ and Π_j are the projections onto the j^{th} coordinate of the respective measures. For fixed $j \in \{1,\dots,p\}$ only the j^{th} term in the sum contains μ_j, σ_j, and γ_j, so all other terms can be absorbed into the additive constant C when optimizing those parameters. Moreover, the conditional measures $P_{\mu_j,\sigma_j|z_j=1}$ and $\Pi_{j|z_j=1}$ are non-atomic and hence respectively singular with $P_{\mu_j,\sigma_j|z_j=0} = \delta_0$ and $\Pi_{j|z_j=0} = \delta_0$. Since $\Pi_j(z_j = 1) = \frac{\alpha}{\alpha+\beta}$, the integrals appearing in the Kullback-Leibler divergence take the form

$$\begin{aligned}\int \log\left(\frac{\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j}}{\mathrm{d}\Pi_j}\right)\,\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j} &= \int \log\left(\frac{\gamma_j \mathrm{d}P_{\mu_j,\sigma_j|z_j=1} + (1-\gamma_j)\delta_0}{\frac{\alpha}{\alpha+\beta}\mathrm{d}\Pi_{j|z_j=1} + \frac{\beta}{\alpha+\beta}\delta_0}\right)\,\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j}\\ &= \int \log\left(\frac{\gamma_j \phi_{\mu_j,\sigma_j}(\theta_j) + (1-\gamma_j)\delta_0}{\frac{\alpha}{\alpha+\beta}f_{\nu,\lambda}(\theta_j) + \frac{\beta}{\alpha+\beta}\delta_0}\right)\,\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j}(\theta_j).\end{aligned}$$

The next step will be to optimize the difference between the latter expression and the variational likelihood bound $\mathbb{E}_{\mu,\sigma,\gamma}\big[f(\theta,\eta)\big]$ with respect to μ_j and σ_j. Collecting terms independent of the objective parameters, observe that

$$\begin{aligned}\mathbb{E}_{\mu_j,\sigma_j,\gamma_j}\left[\log\frac{\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j}}{\mathrm{d}\Pi_j}\right] &= \mathbb{E}_{\mu_j,\sigma_j,\gamma_j}\left[\big(\mathbf{1}_{\{z_j=1\}} + \mathbf{1}_{\{z_j=0\}}\big)\log\frac{\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j}}{\mathrm{d}\Pi_j}\right]\\ &= \gamma_j \mathbb{E}_{\mu_j,\sigma_j,\gamma_j|z_j=1}\left[\log\frac{\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j}}{\mathrm{d}\Pi_j}\right]\\ &\qquad + (1-\gamma_j)\mathbb{E}_{\mu_j,\sigma_j,\gamma_j|z_j=0}\left[\log\frac{\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j}}{\mathrm{d}\Pi_j}\right]\\ &= \gamma_j \int \log\left(\frac{\phi_{\mu_j,\sigma_j}}{f_{\nu,\lambda}}\right)\,\mathrm{d}P_{\mu_j,\sigma_j|z_j=1} + C,\end{aligned}$$

so it suffices to optimize the latter integral as a function of μ_j and σ_j. Similarly,

$$\begin{aligned}\mathbb{E}_{\mu,\sigma,\gamma}\big[f(\theta,\eta)\big] &= \gamma_j \mathbb{E}_{\mu,\sigma,\gamma|z_j=1}\big[f(\theta,\eta)\big] + (1-\gamma_j)\mathbb{E}_{\mu,\sigma,\gamma|z_j=0}\big[f(\theta,\eta)\big]\\ &= \gamma_j \int f(\theta,\eta)\,\mathrm{d}P_{\mu,\sigma,\gamma|z_j=1} + C\end{aligned}$$

further simplifies the objective function.

Proposition 3.8. *For any free parameter $\eta \in \mathbb{R}^n_{>0}$, data $X \in \mathbb{R}^{n\times p}$ and $Y \in \{0,1\}^n$, and prior parameters $\nu \in \mathbb{R}$ and $\lambda \in \mathbb{R}_{>0}$, the coordinates of the minimizer of*

$$(\mu_j,\sigma_j) \longmapsto \int \log\left(\frac{\phi_{\mu_j,\sigma_j}}{f_{\nu,\lambda}}\right)\,\mathrm{d}P_{\mu_j,\sigma_j|z_j=1} - \int f(\theta,\eta)\,\mathrm{d}P_{\mu,\sigma,\gamma|z_j=1}$$

are the respective minimizers of

$$\begin{aligned}\mu_j \longmapsto & \tfrac{|\sigma_j|}{\lambda}\sqrt{\tfrac{2}{\pi}}e^{-\frac{1}{2\sigma_j^2}(\mu_j-\nu)^2} - \tfrac{\nu}{\lambda}\mathrm{erf}\left(\tfrac{\mu_j-\nu}{\sqrt{2}|\sigma_j|}\right)\\ & + \mu_j\left(\sum_{i=1}^n \tfrac{1}{2\eta_i}X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik} - \sum_{i=1}^n (Y_i - 1/2)X_{ij}\right)\\ & + \mu_j^2\sum_{i=1}^n \tfrac{1}{4\eta_i}\tanh(\eta_i/2)X_{ij}^2\\ \sigma_j \longmapsto & \tfrac{|\sigma_j|}{\lambda}\sqrt{\tfrac{2}{\pi}}e^{-\frac{1}{2\sigma_j^2}(\mu_j-\nu)^2} + \tfrac{\mu_j-\nu}{\lambda}\mathrm{erf}\left(\tfrac{\mu_j-\nu}{\sqrt{2}|\sigma_j|}\right) - \log|\sigma_j|\\ & + \sigma_j^2\sum_{i=1}^n \tfrac{1}{4\eta_i}\tanh(\eta_i/2)X_{ij}^2.\end{aligned}$$

Proof. The first step will be to evaluate the Kullback-Leibler term. Expanding the Radon-Nikodym derivative as a function of θ_j yields

$$\begin{aligned}\log\left(\frac{\phi_{\mu_j,\sigma_j}}{f_{\nu,\lambda}}(\theta_j)\right) &= \log\left(\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-\frac{1}{2\sigma_j^2}(\theta_j-\mu_j)^2}\right) - \log\left(\frac{1}{2\lambda}e^{-\frac{1}{\lambda}|\theta_j-\nu|}\right)\\ &= -\log|\sigma_j| - \frac{1}{2}\left(\frac{\theta_j-\mu_j}{\sigma_j}\right)^2 + \frac{1}{\lambda}\left|\theta_j - \nu\right| + C.\end{aligned}$$

By definition, $(\theta_j - \mu_j)^2/\sigma_j^2 \sim \chi^2$ with respect to the variational measure $P_{\mu_j,\sigma_j|z_j=1}$, so its expectation may be absorbed into the additive constant C. Similarly,

$$\frac{\theta_j - \nu}{\lambda} \sim \mathcal{N}\left(\tfrac{\mu_j-\nu}{\lambda}, \tfrac{\sigma_j^2}{\lambda^2}\right)$$

under $P_{\mu_j,\sigma_j|z_j=1}$ and its absolute value has the distribution of a so-called folded Gaussian random variable. Substituting the well-known expression for the mean of such a random variable yields

$$\mathbb{E}_{\mu_j,\sigma_j|z_j=1}\left[\frac{\left|\theta_j-\nu\right|}{\lambda}\right] = \tfrac{|\sigma_j|}{\lambda}\sqrt{\tfrac{2}{\pi}}e^{-\frac{1}{2\sigma_j^2}(\mu_j-\nu)^2} + \tfrac{\mu_j-\nu}{\lambda}\mathrm{erf}\left(\tfrac{\mu_j-\nu}{\sqrt{2}|\sigma_j|}\right)$$

where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-s^2}\,\mathrm{d}s$ denotes the error function. This concludes the computation of the Kullback-Leibler divergence between $P_{\mu_j,\sigma_j|z_j=1}$ and the slab prior.

The expectation of $f(\theta, \eta)$ with respect to $P_{\mu,\sigma,\gamma}$ has already been evaluated in the previous section. Recall that for fixed $\eta \in \mathbb{R}^n_{\geq 0}$ the only terms depending on either μ_j or σ_j are of the form $\sum_{i=1}^n (Y_i - 1/2)\mathbb{E}_{\mu,\sigma,\gamma}\left[X_i^{\mathrm{t}}\theta\right]$ or $\sum_{i=1}^n \frac{1}{4\eta_i}\tanh(\eta_i/2)\mathbb{E}_{\mu,\sigma,\gamma}\left[X_i^{\mathrm{t}}\theta^2\right]$. Denote by $\widetilde{\gamma}$ the vector satisfying $\widetilde{\gamma}_k = \mathbf{1}_{\{k\neq j\}}\gamma_k + \mathbf{1}_{\{k=j\}}$ for every $k = 1, \dots, p$, then $P_{\mu,\sigma,\widetilde{\gamma}}$ and $P_{\mu,\sigma,\gamma|z_j=1}$ agree on all Lebesgue measurable subsets of $\mathbb{R}^p$. Accordingly,

$$\begin{aligned}\mathbb{E}_{\mu,\sigma,\gamma|z_j=1}\big[f(\theta,\eta)\big] &= \mathbb{E}_{\mu,\sigma,\widetilde{\gamma}}\big[f(\theta,\eta)\big] \\ &= \sum_{i=1}^{n}(Y_i - 1/2)\mathbb{E}_{\mu,\sigma,\widetilde{\gamma}}\big[X_i^{\mathrm{t}}\theta\big] - \sum_{i=1}^{n}\tfrac{1}{4\eta_i}\tanh(\eta_i/2)\mathbb{E}_{\mu,\sigma,\widetilde{\gamma}}\big[X_i^{\mathrm{t}}\theta^2\big] + C.\end{aligned}$$

Substituting the closed-form expressions of the expectations, exchanging the sums, and collecting the relevant terms shows that

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - 1/2)\mathbb{E}_{\mu,\sigma,\widetilde{\gamma}}\big[X_i^{\mathrm{t}}\theta\big] &= \sum_{i=1}^{n}(Y_i - 1/2)\sum_{k=1}^{p}\widetilde{\gamma}_k\mu_k X_{ik} \\ &= \sum_{k=1}^{p}\sum_{i=1}^{n}(Y_i - 1/2)\widetilde{\gamma}_k\mu_k X_{ik} \\ &= \sum_{i=1}^{n}(Y_i - 1/2)\mu_j X_{ij} + C.\end{aligned}$$

Similarly, the second term may be expressed in closed form as

$$\begin{aligned}\sum_{i=1}^{n}\tfrac{1}{4\eta_i}\tanh(\eta_i/2)\mathbb{E}_{\mu,\sigma,\widetilde{\gamma}}\big[X_i^{\mathrm{t}}\theta^2\big] &= \sum_{i=1}^{n}\tfrac{1}{4\eta_i}\tanh(\eta_i/2)\bigg(\sum_{k=1}^{p}\widetilde{\gamma}_k X_{ik}^2\big(\mu_k^2 + \sigma_k^2\big) \\ &\qquad + \sum_{k=1}^{p}\sum_{l\neq k}(\widetilde{\gamma}_k\mu_k X_{ik})(\widetilde{\gamma}_l\mu_l X_{il})\bigg) \\ &= \sum_{i=1}^{n}\tfrac{1}{4\eta_i}\tanh(\eta_i/2)\bigg(\big(\mu_j^2 + \sigma_j^2\big)X_{ij}^2 \\ &\qquad + 2\mu_j X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik}\bigg) + C\end{aligned}$$

which concludes the evaluation of the conditional lower bound $\mathbb{E}_{\mu,\sigma,\gamma|z_j=1}\big[f(\theta,\eta)\big]$.

Combining both computations, the optimal mean and standard deviation are given by any joint minimizer of

$$\begin{aligned}(\mu_j,\sigma_j) \longmapsto{} & \tfrac{|\sigma_j|}{\lambda}\sqrt{\tfrac{2}{\pi}}e^{-\frac{1}{2\sigma_j^2}(\mu_j-\nu)^2} + \tfrac{\mu_j-\nu}{\lambda}\operatorname{erf}\Big(\tfrac{\mu_j-\nu}{\sqrt{2}|\sigma_j|}\Big) - \log|\sigma_j| \\ & + \sum_{i=1}^{n}\tfrac{1}{4\eta_i}\tanh(\eta_i/2)\bigg(\big(\mu_j^2 + \sigma_j^2\big)X_{ij}^2 + 2\mu_j X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik}\bigg) \\ & - \sum_{i=1}^{n}(Y_i - 1/2)\mu_j X_{ij}.\end{aligned}$$

Collecting the terms relevant to either μ_j or σ_j yields the update equations in their desired form. □

As mentioned, the update equations as derived in the previous proposition do not admit closed form minimizers and require numerical optimization. Fortunately, the opposite is true for the inclusion probabilities γ.

Proposition 3.9. *For any free parameter $\eta \in \mathbb{R}^n_{>0}$, data $X \in \mathbb{R}^{n\times p}$ and $Y \in \{0,1\}^n$, and prior parameters $\nu \in \mathbb{R}$ and $\lambda \in \mathbb{R}_{>0}$, the minimizer of*

$$\gamma_j \longmapsto \mathrm{KL}\big(P_{\mu_j,\sigma_j,\gamma_j}, \Pi\big) - \mathbb{E}_{\mu,\sigma,\gamma}\big[f(\theta,\eta)\big]$$

satisfies

$$\begin{aligned} -\log \tfrac{\gamma_j}{1-\gamma_j} = \tfrac{|\sigma_j|}{\lambda}\sqrt{\tfrac{2}{\pi}}e^{-\frac{1}{2\sigma_j^2}(\mu_j-\nu)^2} + \tfrac{\mu_j-\nu}{\lambda}\mathrm{erf}\Big(\tfrac{\mu_j-\nu}{\sqrt{2}|\sigma_j|}\Big) + \log\tfrac{\lambda}{|\sigma_j|} + \log\tfrac{\beta}{\alpha} + \log\tfrac{\sqrt{2}}{\sqrt{\pi}} - \tfrac{1}{2} \\ -\sum_{i=1}^n (Y_i - 1/2)\mu_j X_{ij} + \sum_{i=1}^n \tfrac{1}{4\eta_i}\tanh(\eta_i/2)\Bigg(\big(\mu_j^2+\sigma_j^2\big)X_{ij}^2 + 2\mu_j X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik}\Bigg). \end{aligned}$$

Proof. To evaluate the Kullback-Leibler term, recall that $\mathrm{KL}(P_{\mu_j,\sigma_j,\gamma_j}, \Pi)$ may be decomposed into $\gamma_j\mathrm{KL}(P_{\mu_j,\sigma_j,\gamma_j|z_j=1}, \Pi) + (1-\gamma_j)\mathrm{KL}(P_{\mu_j,\sigma_j,\gamma_j|z_j=0}, \Pi)$ where

$$\begin{aligned} \gamma_j\mathrm{KL}\big(P_{\mu_j,\sigma_j,\gamma_j|z_j=1}, \Pi\big) &= \gamma_j\Big(\log\gamma_j - \log\tfrac{\alpha}{\alpha+\beta}\Big) + \gamma_j\int \log\left(\frac{\phi_{\mu_j,\sigma_j}}{f_{\nu,\lambda}}\right)\mathrm{d}P_{\mu_j,\sigma_j|z_j=1} \\ (1-\gamma_j)\mathrm{KL}\big(P_{\mu_j,\sigma_j,\gamma_j|z_j=0}, \Pi\big) &= (1-\gamma_j)\Big(\log(1-\gamma_j) - \log\tfrac{\beta}{\alpha+\beta}\Big). \end{aligned}$$

Substituting the already computed expressions for $\int \log\phi_{\mu_j,\sigma_j}/f_{\nu,\lambda}\,\mathrm{d}P_{\mu_j,\sigma_j|z_j=1}$ and the expectation of $f(\theta,\eta)$, the optimal γ_j minimizes

$$\begin{aligned} \gamma_j \longmapsto & \gamma_j\log\gamma_j + (1-\gamma_j)\log(1-\gamma_j) \\ & +\gamma_j\Bigg(\tfrac{|\sigma_j|}{\lambda}\sqrt{\tfrac{2}{\pi}}e^{-\frac{1}{2\sigma_j^2}(\mu_j-\nu)^2} + \tfrac{\mu_j-\nu}{\lambda}\mathrm{erf}\Big(\tfrac{\mu_j-\nu}{\sqrt{2}|\sigma_j|}\Big) + \log\tfrac{\sqrt{2}\lambda}{\sqrt{\pi}|\sigma_j|} - \tfrac{1}{2} - \log\tfrac{\alpha}{\alpha+\beta} + \log\tfrac{\beta}{\alpha+\beta} \\ & -\sum_{i=1}^n (Y_i - 1/2)\mu_j X_{ij} + \sum_{i=1}^n \tfrac{1}{4\eta_i}\tanh(\eta_i/2)\Big(\big(\mu_j^2+\sigma_j^2\big)X_{ij}^2 + 2\mu_j X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik}\Big)\Bigg). \end{aligned}$$

Observe that this latter function is of the form $g(x) = x\log x + (1-x)\log(1-x) + cx$ for some $c \in \mathbb{R}$. Clearly,

$$\begin{aligned} f'(x) &= \log x - \log(1-x) + c \\ f''(x) &= \tfrac{1}{x} + \tfrac{1}{1-x} \end{aligned}$$

and $f'(x) = 0$ whenever $\log\frac{x}{1-x} = -c$. Moreover, f is convex since $1 \geq \gamma_j$, thereby concluding the proof. □

To conclude this chapter, it will be shown that replacing the Laplace slabs with a Gaussian density ϕ_{μ_0,σ_0} yields closed form update equations. Later on, in part II, it will be demonstrated that despite this advantage the algorithm with Gaussian slabs does not terminate substantially faster in test cases.

Proposition 3.10. *For any free parameter $\eta \in \mathbb{R}^n_{>0}$, data $X \in \mathbb{R}^{n\times p}$ and $Y \in \{0,1\}^n$, and prior parameters $\mu_0 \in \mathbb{R}$ and $\sigma_0^2 \in \mathbb{R}_{>0}$, the variational update equations for the Gaussian spike-and-slab prior are given by*

$$\begin{aligned}
\mu_j &= \frac{\frac{\mu_0}{\sigma_0^2} - \sum_{i=1}^n \frac{1}{2\eta_i}\tanh(\eta_i/2)X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik} + \sum_{i=1}^n (Y_i - 1/2)X_{ij}}{\frac{1}{\sigma_0^2} + \sum_{i=1}^n \frac{1}{2\eta_i}\tanh(\eta_i/2)X_{ij}^2}\\
\sigma_j^2 &= \frac{1}{\frac{1}{\sigma_0^2} + \sum_{i=1}^n \frac{1}{2\eta_i}\tanh(\eta_i/2)X_{ij}^2}\\
-\log\frac{\gamma_j}{1-\gamma_j} &= \left(\frac{1}{2}\left(\frac{\sigma_j^2}{\sigma_0^2} + \frac{(\mu_j-\mu_0)^2}{\sigma_0^2}\right) + \log\left|\frac{\sigma_0}{\sigma_j}\right| - \frac{1}{2} - \log\frac{\alpha}{\alpha+\beta} + \log\frac{\beta}{\alpha+\beta} - \sum_{i=1}^n (Y_i - 1/2)\mu_j X_{ij}\right.\\
&\quad\left. + \sum_{i=1}^n \frac{1}{4\eta_i}\tanh(\eta_i/2)\Big((\mu_j^2 + \sigma_j^2)X_{ij}^2 + 2\mu_j X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik}\Big)\right).
\end{aligned}$$

Proof. Observe that in the objective function $\mathrm{KL}(P_{\mu,\sigma,\gamma}) + \mathbb{E}_{\mu,\sigma,\gamma}\big[f(\theta,\eta)\big]$ only the Radon-Nikodym terms $\mathrm{d}P_{\mu_j,\sigma_j,\gamma_j}/\mathrm{d}\Pi_j$ depend on the choice of prior slabs, allowing most of the computations to carry over from the previous propositions.

As, before, minimizing the objective with respect to μ_j and σ_j is done while conditioning on $z_j = 1$. In particular,

$$\begin{aligned}
\int \log\left(\frac{\phi_{\mu_j,\sigma_j}}{\phi_{\mu_0,\sigma_0}}\right)\,\mathrm{d}P_{\mu_j,\sigma_j|z_j=1} &= \int \log\left|\frac{\sigma_0}{\sigma_j}\right| - \frac{\big(\theta_j-\mu_j\big)^2}{2\sigma_j^2} + \frac{\big(\theta_j-\mu_0\big)^2}{2\sigma_0^2}\,\mathrm{d}P_{\mu_j,\sigma_j|z_j=1}\\
&= \log\left|\frac{\sigma_0}{\sigma_j}\right| - \frac{1}{2} + \frac{1}{2}\left(\frac{\sigma_j^2}{\sigma_0^2} + \frac{(\mu_j-\mu_0)^2}{\sigma_0^2}\right).
\end{aligned}$$

Combined with the results of the previous propositions, this yields the objective functions

$$\begin{aligned}
\mu_j &\longmapsto \frac{(\mu_j-\mu_0)^2}{2\sigma_0^2} + \mu_j^2\sum_{i=1}^n \frac{1}{4\eta_i}\tanh(\eta_i/2)X_{ij}^2\\
&\qquad + \mu_j\left(\sum_{i=1}^n \frac{1}{2\eta_i}\tanh(\eta_i/2)X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik} - \sum_{i=1}^n (Y_i - 1/2)X_{ij}\right)\\
\sigma_j &\longmapsto \frac{\sigma_j^2}{2\sigma_0^2} - \log\big|\sigma_j\big| + \sigma_j^2\sum_{i=1}^n \frac{1}{4\eta_i}\tanh(\eta_i/2)X_{ij}^2.
\end{aligned}$$

Note that the first function is convex, being quadratic in μ_j with positive coefficient. Differentiating, the optimal μ_j must satisfy

$$\frac{\mu_j-\mu_0}{\sigma_0^2} + \mu_j\sum_{i=1}^n \frac{1}{2\eta_i}\tanh(\eta_i/2)X_{ij}^2 + \sum_{i=1}^n \frac{1}{2\eta_i}X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik} - \sum_{i=1}^n (Y_i - 1/2)X_{ij} = 0$$

which results in the claimed update equation. Similarly, the objective function for σ_j is convex when restricted to $\sigma_j \in \mathbb{R}_{>0}$, being the sum of a quadratic with positive coefficient and the convex function $-\log \sigma_j$. Differentiating, the minimizer must satisfy

$$\sigma_j \left(\tfrac{1}{\sigma_0^2} + \sum_{i=1}^{n} \tfrac{1}{2\eta_i} \tanh(\eta_i/2) X_{ij}^2 \right) - \tfrac{1}{\sigma_j} = 0$$

which may be rearranged into the claimed form. Lastly, following the same steps as in the previous proposition shows that the optimal γ_j minimizes

$$\begin{aligned}\gamma_j \longmapsto & \gamma_j \log \gamma_j + (1-\gamma_j)\log(1-\gamma_j) \\ & + \gamma_j \left(\tfrac{1}{2}\left(\tfrac{\sigma_j^2}{\sigma_0^2} + \tfrac{(\mu_j-\mu_0)^2}{\sigma_0^2} \right) + \log\left| \tfrac{\sigma_0}{\sigma_j} \right| - \tfrac{1}{2} - \log \tfrac{\alpha}{\alpha+\beta} + \log \tfrac{\beta}{\alpha+\beta} - \sum_{i=1}^{n} (Y_i - 1/2)\mu_j X_{ij} \right. \\ & \left. + \sum_{i=1}^{n} \tfrac{1}{4\eta_i} \tanh(\eta_i/2) \Big((\mu_j^2 + \sigma_j^2) X_{ij}^2 + 2\mu_j X_{ij} \sum_{k \neq j} \gamma_k \mu_k X_{ik} \Big) \right).\end{aligned}$$

which concludes the proof. □

4. A Glimpse of Asymptotic Consistency

The following chapter contributes a short overview of the main technical conditions and theorems regarding the asymptotic behavior of variational inference in sparse logistic regression. The presentation mainly follows [RSC20], which combines techniques from the linear case [RS21] with the mathematical machinery of [Atc17; CSv15; NR20]. Throughout this chapter, denote by Π^* the variational approximate posterior 3.4 of the sparse logistic model with Laplace spike-and-slab prior 3.1. The results and definitions are given in the hybrid frequentist/Bayesian setting, pioneered in [Sch65] and synthesized into its modern form starting with the seminal article [GGv00].

Assumption 4.1. *There exists a true parameter $\theta_0 \in \mathbb{R}^p$ such that the observed sample $Y \in \{0,1\}^n$ of the logistic regression model 1.2 is a random sample from the probability measure $\mathbb{P}_{\theta_0}$ on $\{0,1\}^n$ induced by θ_0.*

4.1. Compatibility and Coherence

One of the central problems in the mathematical theory of high-dimensional regression lies in the underspecified nature of the relevant linear equations. More precisely, given a design matrix $X \in \mathbb{R}^{n\times p}$ with $p \gg n$, the existence of a unique $\widetilde{\theta} \in \mathbb{R}^p$ satisfying $v = X\widetilde{\theta}$ for some given $v \in \mathbb{R}^n$ cannot be guaranteed without additional assumptions. In the logistic regression model 1.2, the vector v represents the logarithmic odds ratios of the binary outcomes $Y \in \{0,1\}^n$.

As mentioned in chapter 1, one such assumption is sparsity, meaning that the number of non-zero coordinates of $\widetilde{\theta}$ is suitably small compared to p. Given $\theta \in \mathbb{R}^p$, denote by s_θ the sparsity level $\sum_{j=1}^p \mathbf{1}_{\{\theta_j \neq 0\}}$ of θ. Of particular interest is the unknown true sparsity level s_{θ_0}, which will be shortened to s_0 from here on. Moreover, write S_0 for the subset of $\{1,\dots,p\}$ that contains the non-zero indexes of θ_0. Recall that $\theta_{S_0} \in \mathbb{R}^{s_0}$ is the projection of θ onto the subspace spanned by these dimensions.

The key requirement on the design matrix is a so-called local invertibility condition for $X^{\mathrm{t}}X$. Intuitively, it ensures that the Gram matrix is approximately injective when restricted to sufficiently sparse vectors. The main goal of the remainder of this section is to make this notion more precise. Recall that $\psi : \mathbb{R} \to [0,1]$ denotes the logistic function $\psi(x) = (1+e^{-x})^{-1}$, the second derivative of which is given by

$$\psi''(x) = \psi(x)\big(1-\psi(x)\big).$$

Definition 4.2 (Compatibility Numbers). *Given $X \in \mathbb{R}^{n\times p}$ and $\theta_0 \in \mathbb{R}^p$, write $W \in \mathbb{R}^{n\times n}$ for the diagonal matrix with entries $W_{ii} = \psi''(X_i^{\mathrm{t}}\theta_0)$. There are three compatibility type constants* *associated with X, given by*

$$\begin{aligned}\underline{\kappa} &= \inf_{\theta\neq 0}\left\{\frac{\left\|W^{1/2}X\theta\right\|_2^2}{\|X\|^2\|\theta\|_2^2}\;\middle|\;\left\|\theta_{S_0^c}\right\|_1 \leq 7\left\|\theta_{S_0}\right\|_1\right\},\\ \underline{\kappa}(s) &= \inf_{\theta\neq 0}\left\{\frac{\left\|W^{1/2}X\theta\right\|_2^2}{\|X\|^2\|\theta\|_2^2}\;\middle|\;s_\theta \leq s\right\},\\ \overline{\kappa}(s) &= \inf_{\theta\neq 0}\left\{\frac{\|X\theta\|_2^2}{\|X\|^2\|\theta\|_2^2}\;\middle|\;s_\theta \leq s\right\}.\end{aligned}$$

The first compatibility number $\underline{\kappa}$ involves vectors that are only approximately sparse, in the sense that the ℓ^1-norm of their projections onto S_0^c is not too large. The constant 7 is somewhat arbitrary and was used by [Atc17; CSv15]. In contrast, the other two compatibility numbers are defined with respect to truly sparse vectors. The model is said to be compatible if $\underline{\kappa} > 0$ in which case $\left\|W^{1/2}X\theta\right\|_2^2$ is lower-bounded by $\underline{\kappa}\|X\|^2\|\theta\|_2^2$. Scaling X appropriately, $\|X\| = 1$ and $\sqrt{\overline{\kappa}(s)}$ gives an upper bound on the operator norm of X restricted to s-sparse vectors. Similarly, $\sqrt{\underline{\kappa}(s)}$ gives a lower bound on the scaled singular values of s-dimensional sub-matrices of $W^{1/2}X$. For additional reference on compatibility see [Bv11], in particular section 6.13.

Assumption 4.3. *There exists a constant $\alpha \in \mathbb{R}_{>0}$, to be specified later, such that for some given $L \in \mathbb{R}_{>0}$ it holds that*

$$\|X\| \geq \alpha s_0\sqrt{\log p}\max\left\{\frac{50(L+2)\|X\|_\infty}{\underline{\kappa}\big((L+1)s_0\big)}, \frac{64}{\underline{\kappa}}\right\}.$$

Normalizing the design matrix, $\|X\|$ is of order $\sqrt{n}$, transforming the assumption into a minimum sample size requirement. In particular, n should asymptotically be of larger order than $s_0 \log p$, matching [Atc17].

A slightly different approach, termed mutual coherence, emphasizes the role played by correlation between the design matrix columns. Intuitively, its goal is to measure how far the matrix is from orthogonal and its advantage lies in the fact that correlations are somewhat easier to interpret than compatibility constants. See [BTW07; DET06] for some examples of approaching sparsity via mutual coherence bounds.

Definition 4.4 (Mutual Coherence). *The mutual coherence* $\mathrm{mc}(X)$ *of a matrix $X \in \mathbb{R}^{n\times p}$ is given by the maximum correlation between its columns, that is*

$$\mathrm{mc}(X) = \max_{i\neq j}\frac{\left|\langle X_i, X_j\rangle\right|}{\left\|X_i\right\|_2\left\|X_j\right\|_2}.$$

In fact, the two conditions are related, as recounted below. For a proof of the following lemma see [RSC20], which slightly modifies the proof of the related lemma 1 in [CSv15].

Lemma 4.5 (Lemma 6 of [RSC20]). *Suppose $\|X\theta_0\|_\infty$ is bounded and $\min_{i\neq j} \frac{\|X_i\|_2}{\|X_j\|_2} \geq \eta$, then*

$$\begin{aligned}\overline{\kappa}(s) &\leq 1 + s\mathrm{mc}(X)\\ \underline{\kappa}(s) &\geq C\big(\eta^2 - s\mathrm{mc}(X)\big)\\ \underline{\kappa} &\geq C\big(\eta^2 - 64s_0\mathrm{mc}(X)\big).\end{aligned}$$

Provided that the conditions of the lemma are satisfied with $\eta > 0$ and $s_0 = o\big(\mathrm{mc}(X)^{-1}\big)$, the compatibility constants are bounded away from zero and infinity in which case the theorems of the next section take on particularly attractive forms.

Perhaps the simplest compatible setup is the normal means model with observations $Y_i = \theta_i + \varepsilon$, where $n = p$ and $\varepsilon \sim \mathcal{N}(0,1)$. The design matrix is simply the identity matrix $X = I_p$, meaning $\mathrm{mc}(X) = 0$. The compatibility numbers all equal one in this case. Using basic facts of orthogonal transformations, these results are easily extended to arbitrary orthogonal design matrices. Next, suppose $X_{ij} = M_{ij}/\|M_j\|_2$, where $M_j = (M_{1j}, \dots, M_{nj})$, for a random matrix $M \in \mathbb{R}^{n\times p}$ with independent and identically distributed entries. Provided that

$$\begin{aligned}\mathbb{E}\left[e^{t|M_{ij}|^\alpha}\right] &< \alpha\\ \log p &= o\left(n^{\frac{\alpha}{4+\alpha}}\right)\end{aligned}$$

for some $\alpha, t \in \mathbb{R}_{>0}$, theorem 2 of [CJ11] shows that the compatibility numbers are bounded away from zero with probability tending to one as $n \to \infty$, with sparsity levels $s_n = o\big(\sqrt{n/\log p}\big)$. In particular, this covers the case $M_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, whenever $\log p = o(\sqrt[3]{n})$. See appendix D of [RS21] for further details and examples of suitable design matrices.

4.2. Consistency and Contraction

Recall that in its general form a model selection prior on $\theta \in \mathbb{R}^p$ first selects a sparsity level $s \in \{1, \dots, p\}$, according to some prior $s \sim \pi_p$, and then draws $S \subset \{1, \dots p\}$ with $|S| = s$ uniformly at random. The non-zero values θ_S are then drawn from some prior on $\mathbb{R}^{|S|}$, in this case a product of Laplace distributions. The penultimate assumption concerns the weight given by π_p to dense models.

Assumption 4.6. *There exist constants $C_1, C_2, C_3, C_4 \in \mathbb{R}_{>0}$ such that*

$$C_1 p^{-C_3}\pi_p(s-1) \leq \pi_p(s) \leq C_2 p^{-C_4}\pi_p(s-1)$$

is satisfied for all $s \in \{1, \dots, p\}$.

See [CSv15; Cv12] for further discussion of assumptions of this kind. In particular, taking π_p binomially distributed as in 3.1 satisfies the assumption. A characterization of spike-and-slab priors as a sub-class of model selection priors is given in theorem 2.3 of [ES21].

Using the assumptions stated so far, the general results of [Atc17] imply contraction of the true posterior 3.2 at the min-max rate if the parameter λ of the Laplace slabs is taken proportional to $\|X\|_\infty \sqrt{n \log p}$. Under a similar assumption on λ, the main theorems of [RSC20] show that the same holds true for the variational approximate posterior Π^*.

Assumption 4.7. *The constant α of assumption 4.3 can be taken so that the slab-parameter $\lambda \in \mathbb{R}_{>0}$ is bounded by*

$$2\|X\|\sqrt{\log p} \leq \lambda \leq \alpha\|X\|\sqrt{\log p}.$$

Now, the first theorem of [RSC20] can be stated, concerning the asymptotic consistency of Π^* in ℓ^2-norm and mean-squared prediction error $\|p_\theta - p_0\|_n^2 = \frac{1}{n}\sum_{i=1} n\big(\psi(X_i^{\mathrm{t}}\theta) - \psi(X_i^{\mathrm{t}}\theta_0)\big)^2$. To this end, define the constants

$$\begin{aligned} L_0 &= 2\max\left\{C_4/5, \frac{1.1 + 4\alpha^2/\underline{\kappa} + 2C_4 + \log\big(4 + \overline{\kappa}(s_0)\big)}{C_4}\right\} \\ C_\kappa &= L_0\left(\frac{\overline{\kappa}(L_0 s_0)}{\underline{\kappa}(s_0 + 4L_0 s_0/C_4)^2} + \frac{1}{\underline{\kappa}(L_0 s_0)}\right). \end{aligned}$$

Theorem 4.8 (Theorem 1 of [RSC20]). *Suppose all numbered assumptions of this chapter hold and in particular assumption 4.3 for every L_n in a sequence $\{L_n\}_{n\in\mathbb{N}}$, then the variational approximate posterior Π^* satisfies*

$$\begin{aligned} \mathbb{E}_{\theta_0}\left[\Pi^*\left(\theta \in \mathbb{R}^p \;\middle|\; \|\theta - \theta_0\|_2 \geq \frac{\sqrt{L_n}}{\underline{\kappa}(L_n s_0)}\frac{\sqrt{s_0\log p}}{\|X\|}\right)\right] &= O\left(\tfrac{C_\kappa}{L_n}\right) + o(1) \\ \mathbb{E}_{\theta_0}\left[\Pi^*\left(\theta \in \mathbb{R}^p \;\middle|\; \|p_\theta - p_0\|_n \geq \frac{\sqrt{L_n\overline{\kappa}(L_n s_0)}}{\underline{\kappa}(L_n s_0)}\frac{\sqrt{s_0\log p}}{\sqrt{n}}\right)\right] &= O\left(\tfrac{C_\kappa}{L_n}\right) + o(1) \end{aligned}$$

for all $n \in \mathbb{N}$.

If C_κ/L_n vanishes as $n \to \infty$, then the theorem implies contraction of Π^* at the min-max rate. Provided that the compatibility constants are bounded away from zero and infinity, the right-hand side expressions vanish at the respective rates $\sqrt{L_n s_0 \log p}/\|X\|$ and $L_n\sqrt{s_0\log p/n}$, whenever $L_n \to \infty$ arbitrarily slowly. Under the same assumptions, one can show that the variational posterior concentrates on models of sparsity level at most a constant multiple of s_0.

Theorem 4.9 (Theorem 3 of [RSC20]). *In the setting of the previous theorem, it holds for all $n \in \mathbb{N}$ that*

$$\mathbb{E}_{\theta_0}\Big[\Pi^*\big(\theta \in \mathbb{R}^p \;\big|\; s_\theta \geq L_n s_0\big)\Big] = O\left(\tfrac{C_\kappa}{L_n}\right) + o(1).$$

The last theorem describes the asymptotic behavior of the variational posterior predictive mean $\widehat{p}(x) = \int \psi(x^{\mathrm{t}}\theta)\,\mathrm{d}\Pi^*(\theta)$.

Theorem 4.10 (Theorem 2 of [RSC20]). *In the setting of the previous theorems, suppose that $n \geq 1.1 + 4\alpha^2/\underline{\kappa} + 2C_4 + \log\big(4 + \overline{\kappa}(s_0)\big)$, then*

$$\mathbb{P}_{\theta_0}\left(\theta \in \mathbb{R}^p \;\middle|\; \|\widehat{p} - p_0\|_n \geq \frac{\sqrt{L_n\overline{\kappa}\left(\frac{2n}{C_4\log p} + s_0\right)}}{\underline{\kappa}\left(\frac{2n}{C_4\log p} + (1 - 2/C_4)s_0\right)}\frac{\sqrt{s_0\log p}}{\sqrt{n}}\right) = O\left(\tfrac{C_\kappa}{L_n}\right) + o(1)$$

for every $n \in \mathbb{N}$.

Note that the sample size condition of the previous theorem is implied by assumption 4.3 if $\|X\|$ is of order $\sqrt{n}$. Again, provided that the compatibility constants are bounded away from zero and infinity, the right-hand side vanishes at the rate $\sqrt{L_n s_0 \log p/n}$, whenever $L_n \to \infty$ arbitrarily slowly.

The theorems given here are stated in asymptotic form for simplicity. The non-asymptotic counterpart of theorem 4.8 can be found in section 10 of [RSC20]. The proof relies on bounding $\mathrm{KL}\big(\Pi^*, \Pi(\cdot, \mid X, Y)\big)$ on a suitably chosen sequence $\{A_n\}_{n\in\mathbb{N}}$ of events with $\mathbb{P}_{\theta_0}$-probability tending to one. Provided that the true posterior puts exponentially small mass outside of a given set $\Theta_n \subset \mathbb{R}^p$ on A_n, a technical lemma of [RS21] bounds the $\mathbb{P}_{\theta_0}$-expectation of $\Pi^*(\Theta_n^c)$ on A_n via the expected Kullback-Leibler divergence. Choosing Θ_n accordingly then yields the theorems above. The sequence of events A_n is constructed by a localization argument common in non-parametric Bayesian statistics, see for example [NR20]. Proving that the true posterior indeed concentrates on the desired sets Θ_n requires adapting the techniques of [CSv15] and their generalizations in [Atc17].

Part II.

Implementation and Simulation

5. The sparsevb R-Package

The following chapter details the inner workings of the sparsevb software package [CSR21], available for download in the *Comprehensive R Archive Network* (CRAN). The package contains variational algorithms for both linear and logistic regression models with spike-and-slab priors, featuring Laplace as well as Gaussian slab densities. The main goal of the chapter is to detail the implementation of the logistic algorithm with Laplace slabs, as derived in chapter 3. Its linear counterpart, developed in [RS21], may be implemented in analogous fashion.

5.1. Introduction

The first step perhaps considered a proper part of "implementation" — as opposed to "mathematical description" — is the choice of tools, for example among the many available programming languages. Important aspects of this choice include the target audience, available libraries, and package development ecosystem. First and foremost, the sparsevb package exists in an academic context, both as proof-of-concept and as an evolving reference implementation to facilitate further development and testing. The articles [RS21] and [RSC20] are perhaps most relevant for [1] the methodological and theoretical statistics research communities, making an implementation in the R programming language [R C21a] natural.

The expansive R package ecosystem includes many options for numerical linear algebra and optimization routines. Scalability is often proclaimed to be the main practical argument for variational inference, making speed and efficiency a key consideration in this choice. To leverage the relevant advantages of compiled code, the R language may be interfaced with C/C++ or Fortran. The low-level mechanics of the interface are presented in chapter 5 of the excellent manual [R C21b]. Passing R objects back and forth between the R environment and other languages via internal SEXP and SEXPREC objects is quite complicated in its own right, compared to the usual R script development process. Thankfully, the expansive and well-maintained Rcpp package [EF11; Edd13; EB18] eases the low-level burden on R package developers, implementing a high(er)-level wrapper of the C++ interface.

Since its initial CRAN release in 2008, Rcpp has spawned its own community and extended set of developer tools. Regarding numerical linear algebra, two popular C++ libraries are included via RcppArmadillo [ES14] and RcppEigen [BE13]. The Armadillo library [SC16; SC18] functions as a templated C++ wrapper for the low-level *Basic Linear Algebra Subprograms* (BLAS) and *Linear Algebra Package* (LAPACK) reference implementations of linear algebra routines. In practice, its speeds depends upon the specific BLAS and LAPACK implementations, inviting the use of multi-threaded high-performance replacements such as OpenBLAS [XQY12; Wan+13]. In contrast, the Eigen library [GJ+20] implements its own

[1] In the sense of being understood and contextualized by.

low-level linear algebra subroutines with high-level C++ interface. This perhaps allows for slightly more low-level options regarding storage orders and typing, but the decision was made to implement the `sparsevb` package using `RcppArmadillo`.

A major factor influencing this choice is access to the `Armadillo` based numerical optimization library `ensmallen` [Bha+18] with corresponding R-bindings available via the `RcppEnsmallen` package [BE18]. Many recent as well as classic stochastic and/or derivative-free optimization algorithms are included in the library. As of version 0.1.0, the `sparsevb` package supports the limited memory Broyden-Fletcher-Goldfarb-Shanno[2] (L-BFGS) algorithm for all numerical optimization tasks. Belonging to the quasi-Newton class of optimization routines, the L-BFGS algorithm iteratively approximates the inverse Hessian matrix of the objective function through evaluations of the gradient. For a detailed explanation of the algorithm see chapters 6 and 7 of [NW06]. Here, the discussion will now turn towards implementation of the objective functions derived in chapter 3.

5.2. Variational Update Loop

Recall from chapter 3 that for fixed $j \in \{1, \dots, p\}$ the optimal variational mean and standard deviation are the joint minimizers of

$$\begin{aligned}(\mu_j, \sigma_j) \longmapsto & \tfrac{|\sigma_j|}{\lambda}\sqrt{\tfrac{2}{\pi}}e^{-\frac{1}{2\sigma_j^2}(\mu_j-\nu)^2} + \tfrac{\mu_j-\nu}{\lambda}\operatorname{erf}\Big(\tfrac{\mu_j-\nu}{\sqrt{2}|\sigma_j|}\Big) - \log|\sigma_j| \\ & + \sum_{i=1}^n \tfrac{1}{4\eta_i}\tanh(\eta_i/2)\left(\big(\mu_j^2+\sigma_j^2\big)X_{ij}^2 + 2\mu_j X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik}\right) \\ & - \sum_{i=1}^n (Y_i - 1/2)\mu_j X_{ij}.\end{aligned}$$

The first goal of this section is to derive an efficient order of computation for the expressions appearing in the objective function. For the sake of brevity, set $\nu = 0$ and define

$$\begin{aligned}\Gamma_j(\mu,\sigma) &= \tfrac{|\sigma_j|}{\lambda}\sqrt{\tfrac{2}{\pi}}e^{-\frac{1}{2\sigma_j^2}\mu_j^2} \\ \Lambda_j(\mu,\sigma) &= \tfrac{\mu_j}{\lambda}\operatorname{erf}\Big(\tfrac{\mu_j}{\sqrt{2}|\sigma_j|}\Big) \\ \Xi_j(\mu,\sigma) &= \big(\mu_j^2+\sigma_j^2\big)\sum_{i=1}^n \tfrac{1}{4\eta_i}\tanh(\eta_i/2)X_{ij}^2 \\ \Upsilon_j(\mu,\sigma) &= \mu_j\left(\sum_{i=1}^n \tfrac{1}{2\eta_i}\tanh(\eta_i/2)X_{ij}\sum_{k\neq j}\gamma_k\mu_k X_{ik} - \sum_{i=1}^n (Y_i-1/2)X_{ij}\right)\end{aligned}$$

then the joint objective function for μ_j and σ_j is given by $h_j(\mu,\sigma) = \big(\Gamma_j + \Lambda_j + \Xi_j + \Upsilon_j\big)(\mu,\sigma) - \log|\sigma_j|$. Notice that Γ_j, Λ_j, and Ξ_j depend only on λ, X, and η. The former two are known

[2] Remarkably, the original unlimited memory algorithm was discovered independently by each of the four authors, see [Bro70a; Bro70b; Fle70; Gol70; Sha71].

constants throughout execution of the algorithm and η is held fixed when updating the variational parameters. Accordingly, the relevant coefficients may be computed simultaneously for each $j \in \{1, \dots, p\}$ to take advantage of efficient vectorized sub-routines. The same holds true for the coefficients $z_j := \sum_{i=1}^n (Y_i - 1/2) X_{ij}$ appearing in Υ_j.

Given $d \in \mathbb{N}$, denote by $\circledcirc : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ the Hadamard product, computed via element-wise multiplication

$$(v \circledcirc w)_j = v_j w_j$$

for all $j \in \{1, \dots, d\}$. The bottleneck in computing Υ_j is then the matrix/vector multiplication $\sum_{k \neq j} \gamma_k \mu_k X_{ik} = X(I_p - e_j e_j^{\mathrm{t}})(\gamma \circledcirc \mu)$, where $e_j \in \mathbb{R}^p$ is the j^{th} standard basic vector. The coefficient clearly depends on the values of $\gamma \circledcirc \mu$ at indexes other than j, forcing it to be recomputed after every coordinate ascent update. A naive implementation of this computation would thereby incur a complexity of order $O(np^2)$ per iteration.

Write $X_j = X e_j$ for the j^{th} column of X. Observe that $(X e_j)\big(e_j^{\mathrm{t}}(\gamma \circledcirc \mu)\big) = \gamma_j \mu_j X_j$ only depends on the values γ_j and μ_j. Accordingly, one may opt to compute $w = X(\gamma \circledcirc \mu)$ a single time during initialization and then recompute

$$\begin{aligned} w &\leftarrow w - \gamma_j \mu_j X_j \\ &\vdots \\ w &\leftarrow w + \gamma_j^{\text{new}} \mu_j^{\text{new}} X_j \end{aligned}$$

during each coordinate ascent update. This results in the improved complexity $O\big(p \max(n, p)\big)$ per iteration since the other components of Υ_j can be precomputed in vectorized fashion.

With all four pieces of the joint objective function in hand, the values of μ_j and σ_j may now be updated via a suitable numerical optimization subroutine. As mentioned in the previous section, the `sparsevb` package currently supports the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm for this task.

To complete the coordinate ascent update along dimension $j \in \{1, \dots, p\}$, a new inclusion probability γ_j will have to be computed next. Thankfully, most of the update equation derived in proposition 3.9 is actually given by the optimal value $\inf_{\mu_j, \sigma_j} h_j(\mu, \sigma)$, as computed in the previous step. Indeed, the equality

$$\begin{aligned} -\log \tfrac{\gamma_j}{1-\gamma_j} = &\underbrace{\tfrac{|\sigma_j|}{\lambda} \sqrt{\tfrac{2}{\pi}} e^{-\frac{1}{2\sigma_j^2}\mu_j^2}}_{=\Gamma_j} + \underbrace{\tfrac{\mu_j}{\lambda} \operatorname{erf}\Big(\tfrac{\mu_j}{\sqrt{2}|\sigma_j|}\Big)}_{=\Lambda_j} + \log \tfrac{\lambda}{|\sigma_j|} + \log \tfrac{\beta}{\alpha} + \log \tfrac{\sqrt{2}}{\sqrt{\pi}} - \tfrac{1}{2} \\ &\underbrace{- \sum_{i=1}^n (Y_i - 1/2) \mu_j X_{ij} + \sum_{i=1}^n \tfrac{1}{4\eta_i} \tanh(\eta_i/2) \Bigg(\big(\mu_j^2 + \sigma_j^2\big) X_{ij}^2 + 2\mu_j X_{ij} \sum_{k \neq j} \gamma_k \mu_k X_{ik} \Bigg)}_{=\Xi_j + \Upsilon_j} \\ &= \log \tfrac{\beta}{\alpha} + \log \tfrac{\lambda}{|\sigma_j|} + h_j(\mu, \sigma) + \log \tfrac{\sqrt{2}}{\sqrt{\Pi}} - \tfrac{1}{2} \end{aligned}$$

is valid immediately after updating (μ_j, σ_j) and may be solved for γ_j in straightforward manner. As mentioned in definition 1.2, the inverse of $x \mapsto \log x - \log(1 - x)$ is the logistic function $x \mapsto (1 + e^{-x})^{-1}$, ensuring that $\gamma_j \in [0, 1]$.

After completing a full loop over $j \in \{1, \dots, p\}$ of recomputing μ_j, σ_j, γ_j, the lower-bound parameter η is adjusted to provide for the tightest possible bound. Following propositions 3.6 and 3.7, the optimal value is given element-wise by

$$\eta_i^2 = \sum_{j=1}^{p} \gamma_j X_{ij}^2 (\mu_j^2 + \sigma_j^2) + \sum_{j=1}^{p} \sum_{k \neq j} (\gamma_j \mu_j X_{ij})(\gamma_k \mu_k X_{ik}).$$

The first term may be computed simultaneously for all $i \in \{1, \dots, n\}$ via the vectorized expression $(X \circledcirc X)\big(\gamma \circledcirc (\mu^2 + \sigma^2)\big)$. In contrast, the second term seems problematic. A naive implementation may use the matrix $A \in \mathbb{R}^{p \times p}$ with identical columns $A = (\gamma \circledcirc \mu, \dots, \gamma \circledcirc \mu)$ to compute

$$\eta^2 = (X \circledcirc X)\big(\gamma \circledcirc (\mu^2 + \sigma^2)\big) + \Big(\big(X\big(A - \operatorname{diag}(A)\big)\big) \circledcirc X\Big)(\gamma \circledcirc \mu).$$

The latter suffers from a complexity of order $O(np^2)$ due to the dense matrix-matrix multiplication XA. To circumvent this bottleneck, observe that

$$\left(\sum_{j=1}^{p} \gamma_j \mu_j X_{ij}\right)^2 = \sum_{j=1}^{p} \big(\gamma_j \mu_j X_{ij}\big)^2 + \sum_{j=1}^{p} \sum_{k \neq j} (\gamma_j \mu_j X_{ij})(\gamma_k \mu_k X_{ik})$$

for every $i \in \{1, \dots, n\}$. Accordingly, the second sum may be computed simultaneously via the vectorized expression $\big(X(\gamma \circledcirc \mu)\big)^2 - (X \circledcirc X)(\gamma \circledcirc \mu \circledcirc \gamma \circledcirc \mu)$, reducing the complexity of computing η to $O(np)$ per iteration.

Finally, the algorithm performs a convergence check for the binary entropy of γ, up to a specified tolerance, before entering the next iteration. A summary of the implementation discussed up to this point is given in algorithm 1, detailing the exact order of computations.

5.3. Empirical Hyper-Parameter Initialization

Discussions regarding the philosophy, practice, and theory of choosing priors and their parameters in Bayesian models are a mainstay of classroom education as well as active research. A conclusive treatment of the subject may very well remain ever-elusive. The interested reader shall be referred to chapter 3 of [Rob07] for a starting point treating foundational aspects.

Free from philosophical or theoretical burdens, this section will take a heuristic approach. The Kullback-Leibler objective function minimized by the variational method is well-known to be non-convex and hence potentially quite sensitive to initialization as well as hyper-parameter choices. In practice, it may not be guaranteed that every user of the `sparsevb` package takes a principled approach in this matter, necessitating a robust initialization routine for the algorithm.

Given $X \in \mathbb{R}^{n \times p}$, $Y \in \{0, 1\}^n$, and a regularization strength $\lambda \in \mathbb{R}_{>0}$, denote by $\widehat{\theta}$ and $\widetilde{\theta}$ the ℓ_2- and ℓ_1-regularized estimators for the parameter $\theta \in \mathbb{R}^p$ of the logistic regression

Algorithm 1: Overview of the variational update loop

```
  Data: obs $= \{X \in \mathbb{R}^{n\times p}, Y \in \{0,1\}^n\}$, prior $= \{\alpha, \beta, \lambda \in \mathbb{R}_{>0}\}$;
      init $= \{\mu \in \mathbb{R}^p, \sigma^2 \in \mathbb{R}^p_{>0}, \gamma \in [0,1]^p\}$;
      term $= \{N \in \mathbb{N}, \varepsilon \in \mathbb{R}_{>0}\}$ ;          // $N$ max. iterations, $\varepsilon$ tolerance
  Result: post $= \{\mu \in \mathbb{R}^p, \sigma^2 \in \mathbb{R}^p_{>0}, \gamma \in [0,1]^p\}$;
$\eta \leftarrow (1,\dots,1)$ ;                                          // init. free param.
$H_0 \leftarrow H(\gamma)$ ;                        // $H(\gamma) := \gamma\log_2(\gamma) + (1-\gamma)\log_2(1-\gamma)$
$v \leftarrow \gamma \circledcirc \mu$ ;                                        // element-wise mult.
$w \leftarrow Xv$;
$z \leftarrow X^{\mathrm{t}}\big(Y - \frac{1}{2}\mathbf{1}\big)$ ;                                             // $\mathbf{1} := (1,\dots,1)$
  for $i = 1$ to $N$ do
      for $k = 1$ to $n$ do
          $\zeta_k \leftarrow \frac{1}{4\eta_k}\tanh(\eta_k/2)$ ;                     // compute hyperbolic term
      end
      $\xi \leftarrow (X \circledcirc X)^{\mathrm{t}}\zeta$ ;                      // compute squared coefficients
      for $j = 1$ to $p$ do
          $\Gamma_j \leftarrow \left((\mu_j, \sigma_j) \mapsto \frac{|\sigma_j|}{\lambda}\sqrt{\frac{2}{\pi}}e^{-\mu_j^2/2\sigma_j^2}\right)$;
          $\Lambda_j \leftarrow \left((\mu_j, \sigma_j) \mapsto \frac{\mu_j}{\lambda}\mathrm{erf}\left(\frac{\mu_j}{\sqrt{2}|\sigma_j|}\right)\right)$;
          $\Xi_j \leftarrow \big((\mu_j, \sigma_j) \mapsto \xi_j\big(\mu_j^2 + \sigma_j^2\big)\big)$;
          $w \leftarrow w - v_jX_j$ ;                           // delete $j^{\mathrm{th}}$ term from sum
          $\Upsilon_j \leftarrow \big((\mu_j, \sigma_j) \mapsto \mu_j\big(2(\zeta \circledcirc X_j)^{\mathrm{t}}w - z_j\big)\big)$;
          $h_j \leftarrow \Gamma_j + \Lambda_j + \Xi_j + \Upsilon_j - \big(\sigma_j \mapsto \log|\sigma_j|\big)$;
          $(\mu_j, \sigma_j) \leftarrow \arg\min_{\mu_j,\sigma_j} h_j(\mu, \sigma)$ ;   // joint minimizer of prop. 3.8
          $\gamma_j \leftarrow \big(1 + e^{\log\beta/\alpha + h_j(\mu,\sigma) + \log\lambda + \log\sqrt{2/\pi} - 1/2}\big)^{-1}$ ;     // solves prop. 3.9
          $v_j \leftarrow \gamma_j\mu_j$ ;                                // update with new val.
          $w \leftarrow w + v_jX_j$;
      end
      $\eta^2 \leftarrow (X \circledcirc X)\big(\gamma \circledcirc (\mu^2 + \sigma^2)\big) + w \circledcirc w - (X \circledcirc X)(v \circledcirc v)$ ;  // prop. 3.6-3.7
      $H_i \leftarrow H(\gamma)$ ;                           // check entropy convergence
      if $\|H_i - H_{i-1}\|_\infty \le \varepsilon$ then
          return post $\leftarrow \{\mu, \sigma^2, \gamma\}$;
      end
  end
```

model 1.2. Exact definitions of the estimators are given in section 1.2. As suggested in section 4.2 of [RS21], sensitivity of the variational algorithm with respect to the updating order over $j \in \{1, \ldots, p\}$ may be alleviated via a preliminary estimator like $\widehat{\theta}$. Indeed, if the true parameter θ_0 has small magnitude in its first few coordinates, then a lexicographical order may result in false positives since the KL-objective can still decrease by attempting to fit the data with those coordinates. Accordingly, [RS21] propose to update the coordinates in decreasing order of $|\widehat{\theta}|$, aiming to circumvent highly sub-optimal local minima in this way. The ordering may be formalized via the unique function $\varsigma : \mathbb{R}^p \to \mathfrak{S}_p$ that satisfies

$$\left|v_{\sigma(i)}\right| \leq \left|v_{\sigma(j)}\right|$$

whenever $i \geq j$ and $\varsigma(v) = \sigma$. Uniqueness of the function formally follows from the recursion principle, see section 1.3 of [Dud02]. The choice of $\widehat{\mu}$ as opposed to $\widetilde{\mu}$ is due to the ℓ_1-penalty inducing many zeros in $\widetilde{\theta}$, making the latter unsuitable for generation of a robust updating order.

On the other hand, the LASSO estimator $\widetilde{\theta}$ may be used to empirically initialize the parameters of the Beta(α, β)-prior on inclusion probabilities. The expected prior inclusion probability of any particular coordinate $j \in \{1, \ldots, p\}$ is given by $\alpha/(\alpha + \beta)$. In an intuitive sense, $\alpha > \beta$ captures the case where more than half of coordinates are expected to be included and vice-versa when $\alpha < \beta$. This motivates the initialization

$$\begin{aligned} \alpha &= \left\|\tilde{\theta}\right\|_0 \\ \beta &= p - \left\|\tilde{\theta}\right\|_0 \end{aligned}$$

where $\|v\|_0 = \sum_{j=1}^p \mathbf{1}_{\{v_j \neq 0\}}$. Accordingly, if $\|\widetilde{\theta}\|_0 = q \in \{1, \ldots, p\}$, then $\alpha/(\alpha + \beta) = q/p$ captures the proportion of non-zero coordinates of the sparse vector $\widetilde{\theta}$.

Algorithm 2: Overview of the initialization routine

```
Data: $X \in \mathbb{R}^{n \times p}, Y \in \{0,1\}^n$;
Result: $\alpha, \beta \in \mathbb{R}_{>0}, \mu \in \mathbb{R}^p$, order $\in \mathfrak{S}_p$;

$\widehat{\theta} \leftarrow$ cv.ridge($X, Y$);                         // via cv.glmnet
order $\leftarrow \varsigma(\widehat{\theta})$;

$\widetilde{\theta} \leftarrow$ cv.lasso($X, Y$);                         // via cv.glmnet
$\alpha \leftarrow \|\tilde{\theta}\|_0$;
$\beta \leftarrow p - \|\tilde{\theta}\|_0$;

return $\mu \leftarrow \widehat{\theta}, \alpha, \beta$, order;
```

Computation of $\widehat{\mu}$ and $\widetilde{\mu}$ is handled via the popular `glmnet` package [FHT10].[3] Formally, the package computes regularized regression estimators with so-called elastic net penalties,

[3] Interestingly, the numerical parts of the `glmnet` package are written in Mortran 2.0, a very rarely seen dialect of Fortran 77.

meaning a convex combination of ℓ_1- and ℓ_2-penalties. Of course, only the edge cases of pure ridge/LASSO regression are needed here. As an added benefit, a regularization parameter λ may be chosen via cross validation. Naturally, all initialization routines presented in this section, summarized in algorithm 2, may be turned off by manually passing initial values to the algorithm.

6. Simulation Study

The last chapter will start off with a summary of some Bayesian variable selection methods available via the *Comprehensive R Archive Network* (CRAN), focusing on packages supporting logistic regression models. Having introduced the "competitors", simulation results assessing their qualities compared to the variational algorithm developed here will be presented.

6.1. Bayesian Methods for Sparse Logistic Regression

Certainly, many more algorithms claim to perform variable selection or estimation for sparse logistic regression than could ever be introduced and tested here. In turn, a judicious, as well as representative, choice of popularly available inference methods is required for an expressive simulation study. Most algorithms may fall into either a "sampling" or "optimizing" category. The former refers chiefly to Markov chain Monte Carlo (MCMC) methods, which aim to sample from a target probability measure — for example an intractable posterior distribution — by constructing a Markov chain with desired limit. Assuming that the chain has mixed well, quantities such as posterior mean, median, or variance may be approximated by integrating along the chain of samples. For reference or a basic introduction to MCMC methodology see the standard work [RC04].

Among MCMC implementations supporting logistic regression available on CRAN, the most mature option may very well be the `rstanarm` package [Goo+20]. It provides a high-level interface similar to the `glm()` function included with the `stats` library of R. The underlying computations are carried out via the probabilistic programming language Stan [Sta21], using the `RStan` interface [Sta20]. Shrinkage in sparse settings may be performed using the regularized horshoe prior introduced in [PV17]. Given $\tau, c \in \mathbb{R}_{>0}$, the prior takes the hierarchical form

$$\begin{aligned} \lambda_j &\overset{i.i.d.}{\sim} \text{Cauchy}^+(0,1) \\ \sigma_j^2 &= \frac{(c\tau\lambda_j)^2}{c^2 + (\tau\lambda_j)^2} \\ \theta_j \mid \sigma_j^2 &\sim \mathcal{N}\big(0, \sigma_j^2\big), \end{aligned}$$

where $X \sim \text{Cauchy}^+(0,1)$ if $X = |Y|$ for $Y \sim \text{Cauchy}(0,1)$. This is termed the half-Cauchy distribution, in analogy with the half-normal distribution encountered in chapter 3. The samples generated by the algorithm are then used to estimate the posterior mean of θ, yielding a fully Bayesian point estimator. Unlike the spike-and-slab prior 3.1, unimportant coefficients are shrunk directly, without modeling their inclusion probabilities.

Valued for their theoretical elegance, MCMC methods have become the most popular general purpose method of computing posterior distributions. The asymptotic guarantees of sampling algorithms may come at a hefty price in terms of computation time, however. To guarantee sufficient mixing of the chain, large parts of the posterior support need to be explored, requiring many samples. Several more scalable alternatives exist, based either on optimization algorithms or on modified sampling methods. A recent representative of the latter category, the skinny Gibbs sampler introduced in [NSH19] is based on spike-and-slab priors of the form

$$\begin{aligned} z_j \mid w_j &\overset{ind.}{\sim} \mathrm{Binom}(w_j) \\ \theta_j \mid z_j, \sigma_0^2, \sigma_1^2 &\sim \mathcal{N}\big(0, \sigma_{z_j}^2\big) \end{aligned}$$

with given $\sigma_0^2 \ll \sigma_1^2$. During each iteration, the usual Gibbs sampler for the model would require sampling a vector from a p-variate normal distribution, derived in section 2.2 of [NSH19]. This step incurs a large computational complexity, demanding inversion of a certain quadratic form in X to calculate the covariance matrix. To alleviate this bottleneck, the skinny Gibbs sampler only samples coefficients with $z_j = 1$ during the current iteration from a full-rank normal distribution and generates other coefficients from a product of independent normal distributions. As explained in section 2.3 of [NSH19], a modified update of the indicators z_j is needed to guarantee nice theoretical properties, but the computational complexity is improved nevertheless.

The remaining methods considered here are in the optimizing category, including of course the variational algorithms with Gaussian or Laplace slabs, as derived in chapter 3. The most direct competitor is perhaps the `varbvs` package, based on [CS12], which introduced the usage of variational inference for spike-and-slab priors with Gaussian slabs. Additionally, the package includes an importance sampling routine to average multiple variational posteriors arising from different hyper-parameter setting. The empirical initialization of hyper-parameters described in chapter 5 presents a heuristic, but scalable, alternative to the averaging sub-routine.

Another popular optimization-based alternative for latent variable models is the expectation maximization (EM) framework. In the context of spike-and-slab priors, the unobserved variables are of course the binomially distributed indicators $z_1, \dots, z_p$, preventing tractable computation of the posterior $\Pi(\cdot \mid X, Y)$ due to a summation over $\{0,1\}^p$, see the paragraphs surrounding proposition 3.2. The EM algorithm is a general purpose technique for iterative maximization of log-likelihoods/densities with such intractable latent components. Accordingly, it enables maximum likelihood estimation in a frequentist context, or maximum a posteriori (MAP) estimation in Bayesian settings. The first representative of this class is the `BhGLM` package [Yi+18], which applies an iteratively re-weighted EM algorithm to maximize the posterior log-density arising from the spike-and-slab prior

$$\begin{aligned} w &\sim \mathrm{Unif}[0,1] \\ z_j &\overset{i.i.d.}{\sim} \mathrm{Binom}(w) \\ \theta_j \mid z_j, \lambda_0, \lambda_1 &\sim \mathrm{Lap}\big(0, \lambda_{z_j}\big) \end{aligned}$$

with given $\lambda_0 \ll \lambda_1$. A slightly different approach is utilized by the `BinaryEMVS` package based on [MSW16], which adapts the expectation maximization variable selection (EMVS) algorithm introduced in [RG14] to the binary setting. The stochastic dual coordinate ascent algorithm of [SZ13] is employed to solve the underlying optimization problems, arising from the hierarchical prior

$$\begin{aligned} w \mid \alpha, \beta &\sim \mathrm{Beta}(\alpha, \beta) \\ z_j \mid w &\overset{i.i.d.}{\sim} \mathrm{Binom}(w) \\ \sigma^2 \mid \kappa, \tau &\sim \mathrm{InvGamma}(\kappa, \tau) \\ \theta_j \mid z_j, \sigma^2, \nu_0, \nu_1 &\sim \mathcal{N}\big(0, \sigma^2 \nu_{z_j}\big) \end{aligned}$$

with given $\alpha, \beta, \kappa, \tau, \nu_0, \nu_1 \in \mathbb{R}_{>0}$. As expected, a random variable is InvGamma(κ, τ)-distributed if its reciprocal is Gamma(κ, τ)-distributed, constituting the conjugate prior for a normal family.

6.2. Simulation Results

A comparison of algorithms may only ever be as good as its metrics. All algorithms described in the previous section perform either estimation of the unknown parameter θ, selection of a model $S \subset \{1, \ldots, p\}$, or both. Each task will be assessed via two metrics.

In the simulated setting, both the true parameter θ_0 as well as the true probabilities $\mathbb{P}_{\theta_0}(Y_i = 1) = \psi(X_i \theta_0)$ are known, inviting a direct comparison with the estimates $\widetilde{\theta}$ and $\mathbb{P}_{\widetilde{\theta}}(Y_i = 1) = \psi(X_i \widetilde{\theta})$. The quality of the parameter estimate will be assessed via the ℓ_2-distance $\|\widetilde{\theta} - \theta_0\|_2$. On the other hand, the probability $\mathbb{P}_{\widetilde{\theta}}(Y_i = 1)$ may be seen as a prediction for the value of Y_i, tempting a comparison via the mean squared prediction error (MSPE). In particular, it may be estimated via its sample variant, given by $\sqrt{\frac{1}{n}\sum_{i=1}^n \big(\psi(X_i\widetilde{\theta}) - \psi(X_i\theta_0)\big)^2}$.

Similarly, the true model $S_0 \subset \{1, \ldots, p\}$ of non-zero coordinates is known. The algorithms performing variable selection do so by returning a vector $\widetilde{\gamma} \in [0,1]^p$ of inclusion probabilities, which may be converted to an estimated model via the customary threshold of 0.5. The estimated model will be assessed by computing the true positive rate (TPR) $s_0^{-1} \sum_{j=1}^p \mathbf{1}_{\{\theta_{0,j} \neq 0 \text{ and } \widetilde{\gamma}_j > 1/2\}}$ among included variables and the false discovery rate (FDR) $s^{-1} \sum_{j=1}^p \mathbf{1}_{\{\theta_{0,j} = 0 \text{ and } \widetilde{\gamma}_j > 1/2\}}$ among included variables. Here, s_0 denotes the true sparsity level and s the sparsity level of the estimated model.

The results of eight simulated tests are shown in tables 6.2 and 6.3, split into estimation and model selection metrics. Each value represents an average over 200 seeded runs, ensuring that every algorithm is given exactly the same design matrices and binary labels. For the first eight tests, relatively small design matrices $X \in \mathbb{R}^{100 \times 200}$ were chosen to allow all algorithms to compete. Table 6.5 shows metrics for larger simulated tests, including only algorithms that are scalable enough to complete the test within a reasonable time. Typically, a sampling algorithm like Stan may require on the order of 8000 samples, making comparisons using larger matrices infeasible. On the other hand, the variational algorithms included in the `sparsevb` package will usually converge within less than 100 iterations, a difference clearly reflected in the average running times. All times are given for an Intel `i7-8550u` mobile processor, using version 4.0.5 of R built against `openBLAS`. The simulations show that the `sparsevb` package performs on par with its direct competitor `varbvs`.

Figure 6.1 illustrates some typical outcomes for the `sparsevb` and `varbvs` algorithms in tests 1, 3, and 4. The sequence of plots also exhibits some consequences of the inverse logistic function "losing resolution" away from one half, as pointed out in chapter 1. In the first test, featuring a single non-zero coordinate, the algorithms have no trouble finding said coordinate. In tests 3 and 4, with larger numbers of non-zero coordinates, the algorithms miss some of those coordinates since the combined signal strength is large. As the inverse logistic function is barely injective close to the boundary of the unit interval, the algorithms can fit the data reasonably well[1] without finding every single non-zero coordinate.

Recall from chapter 2 that mean-field variational algorithms tend to underestimate the variance, making uncertainty quantification difficult. Table 6.4 shows the average coverage and length of credible intervals for the Laplace algorithm in the first eight tests. As illustrated, reliable uncertainty quantification may still be possible in some cases, but perhaps further theoretical investigation is necessary to understand this phenomenon.

Finally, table 6.1 shows results for the Laplace algorithm with a various values of the prior scale parameter λ. As already mentioned in [RSC20], large values of λ indicate harsher shrinkage, causing the algorithm to perform worse. In contrast, simulations in linear regression, carried out by [RS21], show a more varied outcome with larger values performing sometimes better and sometimes worse.

Table 6.1.: Comparison of Laplace algorithm with different prior scale

λ	TPR	FDR	ℓ_2	MSPE	Time
20	0.19 ± 0.12	$\mathbf{0.01 \pm 0.10}$	4.40 ± 1.09	0.31 ± 0.06	$\mathbf{0.63 \pm 0.15}$
5	0.31 ± 0.13	0.02 ± 0.08	3.82 ± 1.03	0.26 ± 0.05	0.70 ± 0.09
2	0.36 ± 0.13	0.03 ± 0.09	$\mathbf{3.61 \pm 0.94}$	0.24 ± 0.05	0.72 ± 0.11
1	$\mathbf{0.37 \pm 0.12}$	0.03 ± 0.09	3.66 ± 0.89	$\mathbf{0.23 \pm 0.05}$	0.72 ± 0.29
$\frac{1}{2}$	0.33 ± 0.13	0.30 ± 0.34	9.86 ± 36.69	0.25 ± 0.04	0.68 ± 0.12

$X \in \mathbb{R}^{100\times 200}$

$X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$

$s_0 = 10$

$\theta_{0,S_0} \overset{i.i.d.}{\sim} \mathrm{Unif}[-3,3]$

[1] In the sense of converging to a local minimum.

Table 6.2.: Comparison of Bayesian estimators for sparse logistic regression

	Laplace	Gauss	VarBVS	BhGLM	Stan	BinEMVS
ℓ_2	1.36 ± 1.18	2.64 ± 2.04	1.80 ± 1.47	2.29 ± 1.55	**1.23 ± 1.00**	2.48 ± 1.30
	1.31 ± 0.54	**1.23 ± 0.49**	1.28 ± 0.51	1.35 ± 0.42	1.72 ± 0.70	3.71 ± 1.10
	3.67 ± 0.98	3.48 ± 0.88	3.74 ± 1.09	3.78 ± 0.85	**2.95 ± 0.76**	3.32 ± 0.91
	11.91 ± 1.57	11.66 ± 1.61	12.16 ± 1.60	11.79 ± 1.48	10.21 ± 1.68	**9.78 ± 1.82**
	0.90 ± 0.52	1.30 ± 0.70	0.99 ± 0.54	1.35 ± 0.59	1.28 ± 0.85	2.58 ± 1.35
	2.00 ± 0.62	3.61 ± 0.19	2.43 ± 0.59	3.41 ± 0.33	1.67 ± 1.26	**1.52 ± 0.55**
	3.43 ± 0.51	5.04 ± 0.16	3.72 ± 0.55	5.16 ± 0.28	**1.57 ± 0.94**	2.56 ± 0.45
	5.00 ± 0.65	6.33 ± 0.18	5.05 ± 0.53	6.79 ± 0.39	**1.73 ± 0.75**	3.75 ± 0.45
MSPE	**0.05 ± 0.04**	0.08 ± 0.04	**0.05 ± 0.04**	0.09 ± 0.06	0.08 ± 0.05	0.22 ± 0.09
	0.17 ± 0.06	**0.16 ± 0.05**	0.16 ± 0.06	0.17 ± 0.05	0.20 ± 0.05	0.29 ± 0.04
	0.23 ± 0.05	0.21 ± 0.05	0.24 ± 0.06	0.22 ± 0.04	**0.21 ± 0.03**	0.21 ± 0.04
	0.32 ± 0.06	0.30 ± 0.06	0.35 ± 0.05	0.32 ± 0.04	0.17 ± 0.04	**0.16 ± 0.03**
	0.07 ± 0.05	0.09 ± 0.04	**0.07 ± 0.05**	0.11 ± 0.05	0.12 ± 0.04	0.24 ± 0.07
	0.06 ± 0.03	0.09 ± 0.01	**0.06 ± 0.02**	0.09 ± 0.02	0.09 ± 0.02	0.16 ± 0.03
	0.07 ± 0.02	0.10 ± 0.01	**0.07 ± 0.02**	0.11 ± 0.02	0.11 ± 0.02	0.15 ± 0.03
	0.09 ± 0.03	0.12 ± 0.02	**0.09 ± 0.02**	0.14 ± 0.03	0.12 ± 0.03	0.14 ± 0.03
Time	0.64 ± 0.10	0.46 ± 0.05	1.23 ± 0.60	**0.14 ± 0.05**	48.18 ± 4.35	6.64 ± 3.03
	0.59 ± 0.10	0.50 ± 0.09	1.13 ± 0.46	**0.14 ± 0.05**	55.24 ± 5.37	9.83 ± 1.05
	0.63 ± 0.09	0.47 ± 0.04	1.45 ± 0.42	**0.13 ± 0.04**	57.81 ± 6.35	9.86 ± 0.41
	0.65 ± 0.11	0.49 ± 0.04	1.42 ± 0.47	**0.14 ± 0.05**	64.43 ± 7.79	9.96 ± 0.55
	0.61 ± 0.08	0.47 ± 0.04	1.13 ± 0.44	**0.13 ± 0.05**	54.71 ± 6.01	7.92 ± 2.93
	0.74 ± 0.19	0.44 ± 0.04	1.99 ± 0.76	**0.12 ± 0.03**	59.41 ± 6.79	5.30 ± 2.64
	0.76 ± 0.07	0.46 ± 0.04	2.14 ± 0.46	**0.13 ± 0.04**	61.50 ± 6.68	6.67 ± 3.12
	0.74 ± 0.08	0.47 ± 0.04	2.27 ± 0.48	**0.13 ± 0.03**	61.35 ± 8.02	7.98 ± 2.93

All tests: $X \in \mathbb{R}^{100\times 200}$ with $X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.
(1) $s_0 = 1$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-10,10]$
(2) $s_0 = 5$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-2,2]$
(3) $s_0 = 10$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-3,3]$
(4) $s_0 = 20$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-5,5]$
(5) $s_0 = 2$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-5,5]$
(6) $s_0 = 2$, $\theta_{0,1:2} = 5$
(7) $s_0 = 3$, $\theta_{0,1:3} = 5$
(8) $s_0 = 4$, $\theta_{0,1:4} = 5$

Table 6.3.: Comparison of Bayesian model selectors for sparse logistic regression

	Laplace	Gauss	VarBVS	Skinny	BinEMVS
TPR	0.90 ± 0.30	0.91 ± 0.29	0.92 ± 0.27	0.91 ± 0.27	**0.96 ± 0.20**
	0.44 ± 0.21	0.47 ± 0.21	0.44 ± 0.21	0.42 ± 0.19	**0.63 ± 0.20**
	0.36 ± 0.13	0.41 ± 0.14	0.33 ± 0.14	0.32 ± 0.13	**0.56 ± 0.13**
	0.15 ± 0.09	0.19 ± 0.09	0.11 ± 0.08	0.09 ± 0.06	**0.36 ± 0.10**
	0.75 ± 0.28	0.76 ± 0.28	0.75 ± 0.28	0.77 ± 0.27	**0.83 ± 0.25**
	1.00 ± 0.00	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
	1.00 ± 0.00	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
	1.00 ± 0.04	1.00 ± 0.02	1.00 ± 0.02	**1.00 ± 0.01**	1.00 ± 0.02
FDR	0.03 ± 0.14	0.04 ± 0.15	**0.02 ± 0.11**	0.34 ± 0.10	0.44 ± 0.36
	0.04 ± 0.12	0.06 ± 0.15	0.04 ± 0.13	0.24 ± 0.11	0.57 ± 0.18
	0.05 ± 0.13	0.05 ± 0.11	**0.02 ± 0.10**	0.19 ± 0.10	0.31 ± 0.16
	0.08 ± 0.16	0.10 ± 0.15	**0.04 ± 0.12**	0.24 ± 0.10	0.29 ± 0.16
	0.03 ± 0.12	0.03 ± 0.12	**0.01 ± 0.08**	0.28 ± 0.10	0.51 ± 0.29
	0.01 ± 0.07	**0.00 ± 0.00**	**0.00 ± 0.00**	0.20 ± 0.04	0.18 ± 0.21
	0.01 ± 0.05	**0.00 ± 0.00**	**0.00 ± 0.00**	0.14 ± 0.03	0.14 ± 0.17
	0.01 ± 0.04	0.00 ± 0.02	**0.00 ± 0.01**	0.11 ± 0.03	0.12 ± 0.14
Time	0.64 ± 0.10	**0.46 ± 0.05**	1.23 ± 0.60	1.84 ± 0.13	6.64 ± 3.03
	0.59 ± 0.10	**0.50 ± 0.09**	1.13 ± 0.46	2.12 ± 0.18	9.83 ± 1.05
	0.63 ± 0.09	**0.47 ± 0.04**	1.45 ± 0.42	2.31 ± 0.23	9.86 ± 0.41
	0.65 ± 0.11	**0.49 ± 0.04**	1.42 ± 0.47	2.07 ± 0.21	9.96 ± 0.55
	0.61 ± 0.08	**0.47 ± 0.04**	1.13 ± 0.44	2.02 ± 0.10	7.92 ± 2.93
	0.74 ± 0.19	**0.44 ± 0.04**	1.99 ± 0.76	2.06 ± 0.02	5.30 ± 2.64
	0.76 ± 0.07	**0.46 ± 0.04**	2.14 ± 0.46	2.35 ± 0.16	6.67 ± 3.12
	0.74 ± 0.08	**0.47 ± 0.04**	2.27 ± 0.48	2.64 ± 0.04	7.98 ± 2.93

All tests: $X \in \mathbb{R}^{100\times 200}$ with $X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

(1) $s_0 = 1$, $\theta_{0,S_0} \overset{i.i.d.}{\sim} \mathrm{Unif}[-10,10]$

(2) $s_0 = 5$, $\theta_{0,S_0} \overset{i.i.d.}{\sim} \mathrm{Unif}[-2,2]$

(3) $s_0 = 10$, $\theta_{0,S_0} \overset{i.i.d.}{\sim} \mathrm{Unif}[-3,3]$

(4) $s_0 = 20$, $\theta_{0,S_0} \overset{i.i.d.}{\sim} \mathrm{Unif}[-5,5]$

(5) $s_0 = 2$, $\theta_{0,S_0} \overset{i.i.d.}{\sim} \mathrm{Unif}[-5,5]$

(6) $s_0 = 2$, $\theta_{0,1:2} = 5$

(7) $s_0 = 3$, $\theta_{0,1:3} = 5$

(8) $s_0 = 4$, $\theta_{0,1:4} = 5$

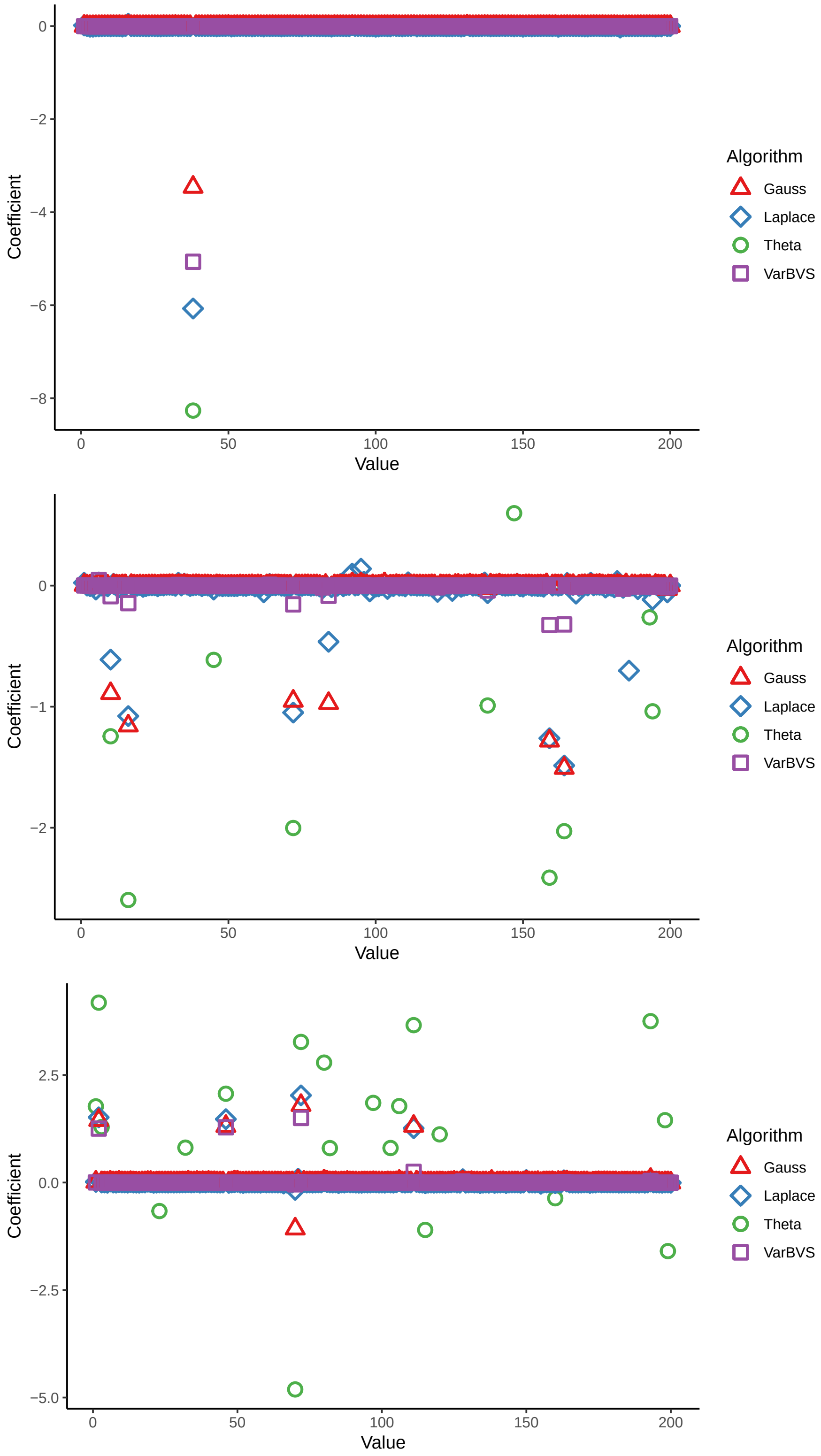


Figure 6.1.: Single outcomes for the `sparsevb` and `varbvs` algorithms in tests 1, 3, and 4.

Table 6.4.: Uncertainty quantification via the Laplace algorithm

	Non-Zero Coverage	Zero Coverage	Non-Zero Length	Zero Length
Test 1	0.98 ± 0.16	1.00 ± 0.00	0.04 ± 0.21	0.01 ± 0.01
Test 2	0.98 ± 0.06	0.98 ± 0.01	0.02 ± 0.06	0.02 ± 0.02
Test 3	0.95 ± 0.07	0.95 ± 0.01	0.04 ± 0.06	0.04 ± 0.02
Test 4	0.89 ± 0.07	0.90 ± 0.01	0.05 ± 0.04	0.05 ± 0.03
Test 5	0.98 ± 0.09	0.99 ± 0.00	0.03 ± 0.13	0.01 ± 0.01
Test 6	0.11 ± 0.26	1.00 ± 0.00	1.21 ± 0.08	0.01 ± 0.01
Test 7	0.00 ± 0.00	1.00 ± 0.00	1.17 ± 0.06	0.01 ± 0.01
Test 8	0.00 ± 0.00	1.00 ± 0.00	1.13 ± 0.08	0.01 ± 0.01

All tests: $X \in \mathbb{R}^{100 \times 200}$ with $X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.
(1) $s_0 = 1$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-10,10]$
(2) $s_0 = 5$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-2,2]$
(3) $s_0 = 10$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-3,3]$
(4) $s_0 = 20$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-5,5]$
(5) $s_0 = 2$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-5,5]$
(6) $s_0 = 2$, $\theta_{0,1:2} = 5$
(7) $s_0 = 3$, $\theta_{0,1:3} = 5$
(8) $s_0 = 4$, $\theta_{0,1:4} = 5$

Table 6.5.: Comparison of scalable Bayesian methods for sparse logistic regression

	TPR	FDR	ℓ_2	MSPE	Time
Laplace	**0.77 ± 0.08**	0.01 ± 0.02	2.35 ± 0.84	0.10 ± 0.02	43.79 ± 4.76
	0.67 ± 0.06	0.01 ± 0.01	10.04 ± 1.37	0.15 ± 0.02	41.39 ± 4.10
	0.90 ± 0.13	0.01 ± 0.04	0.65 ± 0.40	0.04 ± 0.02	43.72 ± 3.12
Gauss	**0.77 ± 0.08**	0.01 ± 0.02	2.45 ± 0.61	**0.09 ± 0.02**	**40.04 ± 0.87**
	0.68 ± 0.06	0.01 ± 0.01	9.70 ± 1.21	**0.14 ± 0.02**	**39.87 ± 0.96**
	0.90 ± 0.13	0.01 ± 0.05	1.08 ± 0.55	**0.04 ± 0.01**	**39.26 ± 0.46**
VarBVS	0.74 ± 0.08	**0.00 ± 0.00**	**2.08 ± 0.61**	0.11 ± 0.02	200.30 ± 42.89
	0.64 ± 0.06	**0.00 ± 0.01**	**9.41 ± 1.24**	0.16 ± 0.02	240.40 ± 50.88
	0.89 ± 0.14	**0.00 ± 0.00**	**0.62 ± 0.34**	0.04 ± 0.02	121.49 ± 44.23
BhGLM			3.36 ± 0.64	0.11 ± 0.02	127.72 ± 49.34
			11.76 ± 1.14	0.19 ± 0.02	183.00 ± 50.87
			0.90 ± 0.46	0.04 ± 0.01	133.38 ± 33.04

All tests: $X \in \mathbb{R}^{1000 \times 2000}$ with $X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.
(1) $s_0 = 25$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-3,3]$
(2) $s_0 = 50$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-4,4]$
(3) $s_0 = 5$, $\theta_{0,S_0} \overset{i.i.d.}{\sim}$ Unif$[-5,5]$

Bibliography

[ALP20] A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. “Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics”. English. In: *Probab. Theory Relat. Fields* 177.1-2 (2020), pp. 323–368. ISSN: 0178-8051; 1432-2064/e.

[Amb+18] Luca Ambrogioni et al. “Wasserstein Variational Inference”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: `https://proceedings.neurips.cc/paper/2018/file/2c89109d42178de8a367c0228f169bf8-Paper.pdf`.

[AR20] Pierre Alquier and James Ridgway. “Concentration of tempered posteriors and of their variational approximations”. English. In: *Ann. Stat.* 48.3 (2020), pp. 1475–1497. ISSN: 0090-5364; 2168-8966/e.

[Atc17] Yves A. Atchadé. “On the contraction properties of some high-dimensional quasi-posterior distributions”. English. In: *Ann. Stat.* 45.5 (2017), pp. 2248–2273. ISSN: 0090-5364; 2168-8966/e.

[Ay+17] Nihat Ay et al. *Information geometry*. English. Vol. 64. Cham: Springer, 2017, pp. xi + 407. ISBN: 978-3-319-56477-7/hbk; 978-3-319-56478-4/ebook.

[BCG21] Sayantan Banerjee, Ismaël Castillo, and Subhashis Ghosal. *Bayesian inference in high-dimensional models*. 2021. arXiv: `2101.04491 [math.ST]`.

[BE13] Douglas Bates and Dirk Eddelbuettel. “Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package”. In: *Journal of Statistical Software* 52.5 (2013), pp. 1–24. URL: `https://www.jstatsoft.org/v52/i05/`.

[BE18] James Joseph Balamuta and Dirk Eddelbuettel. *RcppEnsmallen: Header-Only C++ Mathematical Optimization Library for ’Armadillo’*. 2018.

[Bha+18] Shikhar Bhardwaj et al. “ensmallen: a flexible C++ library for efficient function optimization”. In: *CoRR* abs/1810.09361 (2018). DOI: `10.5281/zenodo.2008650`. arXiv: `1810.09361`. URL: `http://arxiv.org/abs/1810.09361`.

[Bis06] Christopher M. Bishop. *Pattern recognition and machine learning*. English. New York, NY: Springer, 2006, pp. xx + 738. ISBN: 0-387-31073-8/hbk.

[BKM17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. “Variational Inference: A Review for Statisticians”. In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: `10.1080/01621459.2017.1285773`.

[BLM16] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities. A nonasymptotic theory of independence. Corrected paperback edition*. English. Corrected paperback edition. Oxford: Oxford University Press, 2016, pp. x + 481. ISBN: 978-0-19-876765-7/pbk.

[Bro70a] C. G. Broyden. "The convergence of a class of double-rank minimization algorithms. I: General considerations". English. In: *J. Inst. Math. Appl.* 6 (1970), pp. 76–90. ISSN: 0020-2932. DOI: 10.1093/imamat/6.1.76.

[Bro70b] C. G. Broyden. "The convergence of a class of double-rank minimization algorithms. II: The new algorithm". English. In: *J. Inst. Math. Appl.* 6 (1970), pp. 222–231. ISSN: 0020-2932. DOI: 10.1093/imamat/6.3.222.

[BTW07] Florentina Bunea, Alexandre Tsybakov, and Marten H. Wegkamp. "Sparsity oracle inequalities for the Lasso". English. In: *Electron. J. Stat.* 1 (2007), pp. 169–194. ISSN: 1935-7524/e.

[Bv11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data. Methods, theory and applications*. English. Berlin: Springer, 2011, pp. xvii + 556. ISBN: 978-3-642-20191-2/hbk.

[CJ11] T. Tony Cai and Tiefeng Jiang. "Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices". English. In: *Ann. Stat.* 39.3 (2011), pp. 1496–1525. ISSN: 0090-5364. DOI: 10.1214/11-AOS879.

[CS12] Peter Carbonetto and Matthew Stephens. "Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies". English. In: *Bayesian Anal.* 7.1 (2012), pp. 73–108. ISSN: 1931-6690/e.

[CSR21] Gabriel Clara, Botond Szabo, and Kolyan Ray. *sparsevb: Spike-and-Slab Variational Bayes for Linear and Logistic Regression*. R package version 0.1.0. 2021. URL: https://CRAN.R-project.org/package=sparsevb.

[CSv15] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. "Bayesian linear regression with sparse priors". English. In: *Ann. Stat.* 43.5 (2015), pp. 1986–2018. ISSN: 0090-5364; 2168-8966/e.

[Cv12] Ismaël Castillo and Aad van der Vaart. "Needles and straw in a haystack: posterior concentration for possibly sparse sequences". English. In: *Ann. Stat.* 40.4 (2012), pp. 2069–2101. ISSN: 0090-5364; 2168-8966/e.

[DET06] David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. "Stable recovery of sparse overcomplete representations in the presence of noise". English. In: *IEEE Trans. Inf. Theory* 52.1 (2006), pp. 6–18. ISSN: 0018-9448.

[Dud02] R. M. Dudley. *Real analysis and probability. Repr*. English. Repr. Vol. 74. Cambridge: Cambridge University Press, 2002, pp. x + 555. ISBN: 0-521-00754-2/pbk; 0-521-80972-X/hbk.

[EB18] Dirk Eddelbuettel and James Joseph Balamuta. “Extending R with C++: A Brief Introduction to Rcpp”. In: *The American Statistician* 72.1 (2018), pp. 28–36. DOI: 10.1080/00031305.2017.1375990. URL: https://doi.org/10.1080/00031305.2017.1375990.

[Edd13] Dirk Eddelbuettel. *Seamless R and C++ integration with Rcpp*. English. Vol. 64. New York, NY: Springer, 2013, pp. xxviii + 220. ISBN: 978-1-4614-6867-7/pbk; 978-1-4614-6868-4/ebook.

[EF11] Dirk Eddelbuettel and Romain François. “Rcpp: Seamless R and C++ Integration”. In: *Journal of Statistical Software* 40.8 (2011), pp. 1–18. DOI: 10.18637/jss.v040.i08. URL: https://www.jstatsoft.org/v40/i08/.

[ES14] Dirk Eddelbuettel and Conrad Sanderson. “RcppArmadillo: accelerating R with high-performance C++ linear algebra”. English. In: *Comput. Stat. Data Anal.* 71 (2014), pp. 1054–1063. ISSN: 0167-9473. DOI: 10.1016/j.csda.2013.02.005.

[ES21] Tim van Erven and Botond Szabó. “Fast Exact Bayesian Inference for Sparse Signals in the Normal Sequence Model”. In: *Bayesian Analysis* (2021), pp. 1–28. DOI: 10.1214/20-BA1227. URL: https://doi.org/10.1214/20-BA1227.

[FHT10] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software, Articles* 33.1 (2010), pp. 1–22. ISSN: 1548-7660. DOI: 10.18637/jss.v033.i01. URL: https://www.jstatsoft.org/v033/i01.

[Fle70] R. Fletcher. “A new approach to variable metric algorithm”. English. In: *Comput. J.* 13 (1970), pp. 317–322. ISSN: 0010-4620. DOI: 10.1093/comjnl/13.3.317.

[GGv00] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. “Convergence rates of posterior distributions”. English. In: *Ann. Stat.* 28.2 (2000), pp. 500–531. ISSN: 0090-5364; 2168-8966/e.

[GJ+20] Gaël Guennebaud, Benoît Jacob, et al. *Eigen 3.3.9*. http://eigen.tuxfamily.org. 2020.

[Gol70] D. Goldfarb. “A family of variable-metric methods derived by variational means”. English. In: *Math. Comput.* 24 (1970), pp. 23–26. ISSN: 0025-5718. DOI: 10.2307/2004873.

[Goo+20] Ben Goodrich et al. *rstanarm: Bayesian applied regression modeling via Stan.* R package version 2.21.1. 2020. URL: https://mc-stan.org/rstanarm.

[Gv17] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of nonparametric Bayesian inference*. English. Vol. 44. Cambridge: Cambridge University Press, 2017, pp. xxiv + 646. ISBN: 978-0-521-87826-5; 978-1-139-02983-4. DOI: 10.1017/9781139029834.

[HB15] Matthew Hoffman and David Blei. “Stochastic Structured Variational Inference”. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Guy Lebanon and S. V. N. Vishwanathan. Vol. 38. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, Sept. 2015, pp. 361–369. URL: `http://proceedings.mlr.press/v38/hoffman15.pdf`.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning. Data mining, inference, and prediction. 2nd ed*. English. 2nd ed. New York, NY: Springer, 2009, pp. xxii + 745. ISBN: 978-0-387-84857-0/hbk; 978-0-387-84858-7/ebook.

[JJ00] Tommi Jaakkola and Michael Jordan. “Bayesian Parameter Estimation Via Variational Methods”. In: *Statistics and Computing* 10 (Aug. 2000). DOI: `10.1023/A:1008932416310`.

[JRH20] Prateek Jaiswal, Vinayak Rao, and Harsha Honnappa. “Asymptotic consistency of α-Rényi-approximate posteriors”. English. In: *J. Mach. Learn. Res.* 21 (2020). Id/No 156, p. 42. ISSN: 1532-4435; 1533-7928/e.

[KL51] S. Kullback and R. A. Leibler. “On information and sufficiency”. English. In: *Ann. Math. Stat.* 22 (1951), pp. 79–86. ISSN: 0003-4851.

[Kul97] Solomon Kullback. *Information theory and statistics. Reprint of the 2nd ed. ’68*. English. Reprint of the 2nd ed. ’68. Mineola, NY: Dover Publications, Inc., 1997, pp. xvi + 399. ISBN: 0-486-69684-7.

[Lan02] Serge Lang. *Algebra.* English. Vol. 211. New York, NY: Springer, 2002, pp. xv + 914. ISBN: 0-387-95385-X.

[LT16] Yingzhen Li and Richard E Turner. “Rényi Divergence Variational Inference”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: `https://proceedings.neurips.cc/paper/2016/file/7750ca3559e5b8e1f44210283368fc16-Paper.pdf`.

[MSW16] Patrick McDermott, John Snyder, and Rebecca Willison. *Methods for Bayesian Variable Selection with Binary Response Data using the EM Algorithm*. 2016. arXiv: `1605.05429 [stat.CO]`.

[NR20] Richard Nickl and Kolyan Ray. “Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions”. English. In: *Ann. Stat.* 48.3 (2020), pp. 1383–1408. ISSN: 0090-5364; 2168-8966/e.

[NSH19] Naveen N. Narisetty, Juan Shen, and Xuming He. “Skinny Gibbs: a consistent and scalable Gibbs sampler for model selection”. English. In: *J. Am. Stat. Assoc.* 114.527 (2019), pp. 1205–1217. ISSN: 0162-1459. DOI: `10.1080/01621459.2018.1482754`.

[NW06] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. English. New York, NY: Springer, 2006, pp. xxii + 664. ISBN: 0-387-30303-0.

[PV17] Juho Piironen and Aki Vehtari. “Sparsity information and regularization in the horseshoe and other shrinkage priors”. English. In: *Electron. J. Stat.* 11.2 (2017), pp. 5018–5051. ISSN: 1935-7524. DOI: 10.1214/17-EJS1337SI.

[R C21a] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org.

[R C21b] R Core Team. *Writing R Extensions*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org.

[Ran+16] Rajesh Ranganath et al. “Operator Variational Inference”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/d947bf06a885db0d477d707121934ff8-Paper.pdf.

[RC04] Christian P. Robert and George Casella. *Monte Carlo statistical methods*. English. New York, NY: Springer, 2004, pp. xxx + 645. ISBN: 0-387-21239-6.

[RG14] Veronika Ročková and Edward I. George. “EMVS: the EM approach to Bayesian variable selection”. English. In: *J. Am. Stat. Assoc.* 109.506 (2014), pp. 828–846. ISSN: 0162-1459. DOI: 10.1080/01621459.2013.869223.

[Rob07] Christian P. Robert. *The Bayesian choice. From decision-theoretic foundations to computational implementation.* English. New York, NY: Springer, 2007, pp. xxv + 602. ISBN: 978-0-387-71598-8. DOI: 10.1007/0-387-71599-1.

[RS21] Kolyan Ray and Botond Szabó. “Variational Bayes for High-Dimensional Linear Regression With Sparse Priors”. In: *Journal of the American Statistical Association* 0.0 (2021), pp. 1–12. DOI: 10.1080/01621459.2020.1847121.

[RSC20] Kolyan Ray, Botond Szabo, and Gabriel Clara. “Spike and slab variational Bayes for high dimensional logistic regression”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 14423–14434. URL: https://proceedings.neurips.cc/paper/2020/file/a5bad363fc47f424ddf5091c8471480a-Paper.pdf.

[SBK20] Abhijoy Saha, Karthik Bharath, and Sebastian Kurtek. “A Geometric Variational Approach to Bayesian Inference”. In: *Journal of the American Statistical Association* 115.530 (2020). PMID: 33041402, pp. 822–835. DOI: 10.1080/01621459.2019.1585253.

[SC16] Conrad Sanderson and Ryan Curtin. “Armadillo: a template-based C++ library for linear algebra”. In: *Journal of Open Source Software* 1.2 (2016), p. 26. DOI: 10.21105/joss.00026. URL: https://doi.org/10.21105/joss.00026.

[SC18] Conrad Sanderson and Ryan Curtin. “A user-friendly hybrid sparse matrix class in C++”. English. In: *Mathematical software – ICMS 2018. 6th international conference, South Bend, IN, USA, July 24–27, 2018. Proceedings*. Cham: Springer, 2018, pp. 422–430. ISBN: 978-3-319-96417-1; 978-3-319-96418-8. DOI: 10.1007/978-3-319-96418-8_50.

[Sch65] Lorraine Schwartz. “On Bayes procedures”. English. In: *Z. Wahrscheinlichkeitstheor. Verw. Geb.* 4 (1965), pp. 10–26. ISSN: 0044-3719.

[Sha71] D. F. Shanno. “Conditioning of quasi-Newton methods for function minimization”. English. In: *Math. Comput.* 24 (1971), pp. 647–656. ISSN: 0025-5718. DOI: 10.2307/2004840.

[Sta20] Stan Development Team. *RStan: the R interface to Stan*. R package version 2.21.2. 2020. URL: http://mc-stan.org/.

[Sta21] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*. Version 2.27. 2021. URL: https://mc-stan.org.

[Stu03] Karl-Theodor Sturm. “Probability measures on metric spaces of nonpositive curvature”. English. In: *Heat kernels and analysis on manifolds, graphs, and metric spaces. Lecture notes from a quarter program on heat kernels, random walks, and analysis on manifolds and graphs, April 16–July 13, 2002, Paris, France*. Providence, RI: American Mathematical Society (AMS), 2003, pp. 357–390. ISBN: 0-8218-3383-9/pbk.

[SZ13] Shai Shalev-Shwartz and Tong Zhang. “Stochastic dual coordinate ascent methods for regularized loss minimization”. English. In: *J. Mach. Learn. Res.* 14 (2013), pp. 567–599. ISSN: 1532-4435.

[Tib96] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. English. In: *J. R. Stat. Soc., Ser. B* 58.1 (1996), pp. 267–288. ISSN: 0035-9246.

[TS11] Richard Eric Turner and Maneesh Sahani. “Two problems with variational expectation maximisation for time series models”. In: *Bayesian Time Series Models*. Ed. by David Barber, A. Taylan Cemgil, and SilviaEditors Chiappa. Cambridge University Press, 2011, pp. 104–124. DOI: 10.1017/CBO9780511984679.006.

[Wan+13] Qian Wang et al. “AUGEM: Automatically generate high performance Dense Linear Algebra kernels on x86 CPUs”. In: *SC ’13: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. 2013, pp. 1–12. DOI: 10.1145/2503210.2503219.

[WB19] Yixin Wang and David M. Blei. “Frequentist consistency of variational Bayes”. English. In: *J. Am. Stat. Assoc.* 114.527 (2019), pp. 1147–1161. ISSN: 0162-1459; 1537-274X/e.

[XQY12] Zhang Xianyi, Wang Qian, and Zhang Yunquan. “Model-driven Level 3 BLAS Performance Optimization on Loongson 3A Processor”. In: *2012 IEEE 18th International Conference on Parallel and Distributed Systems*. 2012, pp. 684–691. DOI: 10.1109/ICPADS.2012.97.

[Yi+18] Nengjun Yi et al. "BhGLM: Bayesian hierarchical GLMs and survival models, with applications to genomics and epidemiology". In: *Bioinformatics* 35.8 (Sept. 2018), pp. 1419–1421. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty803. eprint: https://academic.oup.com/bioinformatics/article-pdf/35/8/1419/36613783/bioinformatics_35_8_1419.pdf. URL: https://doi.org/10.1093/bioinformatics/bty803.

[YPB20] Yun Yang, Debdeep Pati, and Anirban Bhattacharya. "α-variational inference with statistical guarantees". English. In: *Ann. Stat.* 48.2 (2020), pp. 886–905. ISSN: 0090-5364; 2168-8966/e.

[ZG20] Fengshuo Zhang and Chao Gao. "Convergence rates of variational posterior distributions". English. In: *Ann. Stat.* 48.4 (2020), pp. 2180–2207. ISSN: 0090-5364; 2168-8966/e.