Regression modelling with priors using Fisher information covariance kernels (I-priors)

Md. Haziq Md. Jamil

A thesis submitted to the Department of Statistics of the London School of Economics and Political Science for the degree of Doctor of Philosophy

7 June 2018

Abstract

Regression analysis is undoubtedly an important tool to understand the relationship between one or more explanatory and independent variables of interest. In this thesis, we explore a novel methodology for fitting a wide-range of parametric and non-parametric regression models, called the I-prior methodology.

We assume that the regression function belongs to a reproducing kernel Hilbert or Krein space of functions, and by doing so, it allows us to utilise the convenient topologies of these vector spaces. This is important for the derivation of the Fisher information of the regression function, which might be infinite-dimensional. Based on the principle of maximum entropy, an I-prior is an objective Gaussian process prior for the regression function with covariance function proportional to its Fisher information.

Our work focusses on the statistical methodology and computational aspects of fitting I-priors models. In the first part of the thesis, we examine a likelihood-based approach (direct optimisation and EM algorithm) for fitting I-prior models with normally distributed errors. The culmination of this work is the R package **iprior** which has been made publicly available on the Comprehensive R Archive Network (CRAN). In the second part, the normal I-prior methodology is extended to fit categorical response models, achieved by "squashing" the regression functions through a probit sigmoid function. Estimation of I-probit models, as we call it, proves challenging due to the intractable integral involved in computing the likelihood. We apply a fully Bayesian treatment with a variational approximation in order to obtain the required posterior distributions. Finally, in the third part, we again turn to a fully Bayesian approach of variable selection for linear models using I-priors.

We illustrate the use of I-priors in various simulated and real-data examples. Our study advocates the I-prior methodology as being a simple, intuitive and comparable, though often times better, alternative to similar leading state-of-the-art models.

Keywords: Fisher information, Gaussian process, regression, binary, multinomial, variational inference, empirical Bayes, empirical Bayes, expectation maximisation, EM algorithm

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 65,689 words.

I confirm that Chapters 2 and 3 were jointly co-authored with Dr. Wicher Bergsma, and I contributed 60% of these works.

To-do list

Contents

Figures

Tables

Theorems

Definitions

Nomenclature

As much as possible, and unless otherwise stated, the following conventions are used throughout this thesis.

Conventions

$\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$	Boldface lower case letters denote real vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$	Boldface upper case letters denote real matrices
$\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$	Calligraphic upper case letters denote sets
x'	Primes are used to distinguish elements (not indicate derivatives)
$\hat{\theta}$	Hats are used to denote estimators of parameters

Indexing

$\mathbf{A}_{ij}$, A_{ij}, a_{ij}	The (i, j)'th element of the matrix $\mathbf{A}$
$\mathbf{A}_{i\cdot}$	The i'th row of the matrix $\mathbf{A}$ as a tall vector (transposed row vector)
$\mathbf{A}_{\cdot j}$	The j'th column vector of the matrix $\mathbf{A}$

Symbols

$\mathbb{N}$	The set of natural numbers (excluding zero)
$\mathbb{Z}$	The set of integers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_{>0}$	The set of positive real numbers, $\{x \in \mathbb{R} \vert x > 0\}$
$\mathbb{R}_{\geq 0}$	The set of non-negative real numbers, $\{x \in \mathbb{R} \vert x \geq 0\}$
$\mathbb{R}^d$	The d-dimensional Euclidean space
$\mathcal{A}^c$	The complement of a set $\mathcal{A}$
$\mathcal{P}(\mathcal{A})$	The power set of the set $\mathcal{A}$
$\{\}, \emptyset$	The empty set
$\mathbf{0}$	A vector of zeroes
$\mathbf{1}_n$	A length n vector of ones
$\mathbf{I}_n$	The $n \times n$ identity matrix
$\exists$	(short hand) There exists
$\forall$	(short hand) For all
$\lim_{n \to \infty}$	The limit as n tends to infinity
$\xrightarrow{\text{dist.}}$	Convergence in distribution
$O(n)$	Computational complexity (time or storage)
Δx	A quantity representing a change in x

Relations

$a \approx b$	a is approximately or almost equal to b
$a \propto b$	a is equivalent to b up to a constant of proportionality
$a \equiv b$	a is identical to b
$A \Rightarrow B$	The statement B being true is predicated on A being true
$A \Leftrightarrow B$	The statement A is true if and only if B is true
$a \in \mathcal{A}$	a is an element of the set $\mathcal{A}$
$\mathcal{A} \subseteq \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$ which may include itself
$\mathcal{A} \subset \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$ which does not include itself
$a := b$, $a \leftarrow b$	a is assigned the value b
$X \sim p(X)$	The random variable X is distributed according to the pdf $p(X)$
$X \sim D$	The random variable X is distributed according to the pdf specified by the distribution D, e.g. $D \equiv \mathrm{N}(0, 1)$
$X_1,...,X_n \overset{\text{iid}}{\sim} D$	Each random variable X_i, $i = 1, \ldots, n$ is independently and identically distributed according to the pdf specified by the distribution D
$X \vert Y$	The (random) variable X given/conditional on Y

Functions

$\inf \mathcal{A}$	The infimum of a set $\mathcal{A}$
$\sup \mathcal{A}$	The supremum of a set $\mathcal{A}$
$\min \mathcal{A}$	The minimum value of a set $\mathcal{A}$
$\max \mathcal{A}$	The maximum value of a set $\mathcal{A}$
$\arg\min_x f(x)$	The value of x which minimises the function $f(x)$
$\arg\max_x f(x)$	The value of x which maximises the function $f(x)$
$\vert a \vert$ with $a \in \mathbb{R}$	The absolute value of a; $\vert a \vert = a$ if a is positive, and $-a$ if a is negative, and $\vert 0 \vert = 0$
$\delta_{xx'}$	The Kronecker delta; $\delta_{xx'} = 1$ if $x = x'$, and 0 otherwise
$[A]$	The Iverson bracket; $[A] = 1$ if the logical proposition A is true, and 0 otherwise
$\mathbb{1}_{\mathcal{A}}(x)$	The indicator function; $\mathbb{1}_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$, and 0 otherwise
e^x, $\exp(x)$	The natural exponential function
$\log(x)$	The natural logarithmic function
$\frac{\mathrm{d}}{\mathrm{d}x} f(x)$, $\dot{f}(x)$	The derivative of f with respect to x
$f \circ g$	Composition of functions, i.e. g following f

Abstract vector space operations and notations

$\mathcal{V}^{\perp}$	The orthogonal complement of the space $\mathcal{V}$
$\mathcal{V}^{\vee}$	The algebraic dual space of $\mathcal{V}$
$\mathcal{V}^{*}$	The continuous dual space of $\mathcal{V}$
$\overline{\mathcal{V}}$	The closure of the space $\mathcal{V}$
$\mathcal{B}(\mathcal{V})$	The Borel σ-algebra of $\mathcal{V}$
$\mathrm{L}^p(\mathcal{X}, \nu)$	The set of p-integrable functions over the space $\mathcal{X}$ with measure ν
$\mathrm{L}(\mathcal{V}; \mathcal{W})$	The set of bounded, linear operators from $\mathcal{V}$ to $\mathcal{W}$
$\dim(\mathcal{V})$	The dimensions of the vector space $\mathcal{V}$
$\langle x, y \rangle_{\mathcal{V}}$	The inner product between x and y in the vector space $\mathcal{V}$

$\|x\|_{\mathcal{V}}$	The norm of x in the vector space $\mathcal{V}$
$D(x, y)$	The distance between x and y
$x \otimes y$	The tensor product of x and y which are elements of a vector space
$\mathcal{F} \otimes \mathcal{G}$	The tensor product space of two vector spaces
$\mathcal{F} \oplus \mathcal{G}$	The direct sum (or tensor sum) of two vector spaces
$\mathrm{d}f(x)$, $\mathrm{d}^2 f(x)$	The first and second Fréchet differentials of f at x
$\partial_v f(x)$, $\partial_v^2 f(x)$	The first and second Gâteaux differentials of f at x in the direction v
$\nabla f(x)$, $\nabla^2 f(x)$	The gradient and Hessian of f at x in the direction v (f is a mapping of a Hilbert space)

Matrix and vector operations

$\mathbf{a}^\top$, $\mathbf{A}^\top$	The transpose of a vector $\mathbf{a}$ or matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	The inverse of a square matrix $\mathbf{A}$
$\|\mathbf{a}\|^2$	The squared 2-norm the vector $\mathbf{a}$, equivalent to $\mathbf{a}^\top\mathbf{a}$
$\|\mathbf{A}\|$	The determinant of a matrix $\mathbf{A}$
$\mathrm{tr}(\mathbf{A})$	The trace of a square matrix $\mathbf{A}$
$\mathrm{diag}(\mathbf{A})$	The diagonal elements of a square matrix $\mathbf{A}$
$\mathrm{rank}(\mathbf{A})$	The rank of a matrix $\mathbf{A}$
$\mathrm{vec}(\mathbf{A})$	The column-wise vectorisation of a matrix $\mathbf{A}$
$\mathbf{a} \otimes \mathbf{b}$	The outer product of two vectors $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{A} \otimes \mathbf{B}$	The Kronecker product of matrix $\mathbf{A}$ with matrix $\mathbf{B}$
$\mathbf{A} \circ \mathbf{B}$	The Hadamard product two matrices $\mathbf{A}$ and $\mathbf{B}$

Statistical functions

$\mathrm{P}(A)$	The probability of event A occurring
$p(X\|\theta)$	The probability density function of X given parameters θ
$L(\theta\|X)$	The log-likelihood of θ given data X, sometimes simply $L(\theta)$
$\mathrm{BF}(M, M')$	Bayes factor for comparing two models M and M'
$\mathcal{I}(\theta)$	The Fisher information for θ
$\mathrm{E}[X]$, $\mathrm{E}\,X$	The expectation[1] of the random element X
$\mathrm{Var}[X]$, $\mathrm{Var}\,X$	The variance[1] of the random element X
$\mathrm{Cov}[X, Y]$	The covariance[1] between two random elements X and Y
$H(p)$	The entropy of the distribution $p(X)$
$\mathrm{D}_{\mathrm{KL}}\big(q(x)\|p(x)\big)$	The Kullback-Leibler divergence from $p(x)$ to $q(x)$, denoted also by $\mathrm{D}_{\mathrm{KL}}(q\|p)$ for short

Statistical distributions

$\mathrm{N}(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2
$\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\phi(z)$	The standard normal pdf
$\Phi(z)$	The standard normal cdf

[1] When there is ambiguity as to which random element the expectation or variance is taken under or what its distribution is, this is explicated by means of subscripting, e.g. $\mathrm{E}_{X\sim\mathrm{N}(0,1)}\,X$ to denote the expectation of a standard normal random variable.

$\phi(x\|\mu,\sigma^2)$	The pdf of $\mathrm{N}(\mu,\sigma^2)$
$\phi(\mathbf{x}\|\boldsymbol{\mu},\boldsymbol{\Sigma})$	The pdf of $\mathrm{N}_d(\boldsymbol{\mu},\boldsymbol{\Sigma})$
$\mathrm{MN}_{n,m}(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Psi})$	Matrix normal distribution with mean $\boldsymbol{\mu}$ and row variances $\boldsymbol{\Sigma}\in\mathbb{R}^{n\times n}$ and column variances $\boldsymbol{\Psi}\in\mathbb{R}^{m\times m}$
${}^{\mathrm{t}}\mathrm{N}(\mu,\sigma^2,a,b)$	Truncated univariate normal distribution with mean μ and variance σ^2 restricted to the interval (a,b)
$\mathrm{N}_+(0,1)$	The half-normal distribution with variance σ^2
$\mathrm{N}_+(0,\sigma^2)$	The folded-normal distribution with variance σ^2
${}^{\mathrm{t}}\mathrm{N}_d(\boldsymbol{\mu},\boldsymbol{\Sigma},\mathcal{A})$	Truncated d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ restricted to the set $\mathcal{A}$
$\Gamma(s,r)$	Gamma distribution with shape s and rate r parameters
$\Gamma^{-1}(s,\sigma)$	Inverse gamma distribution with shape s and scale σ parameters
χ^2_d	Chi-squared distribution with d degrees of freedom
$\mathrm{Bern}(p)$	Bernoulli distribution with probability of success p
$\mathrm{Cat}(p_1,\ldots,p_m)$	Categorical distribution with m categories, and each category has probability of success p_j

Abbreviations

ANOVA	Analysis of variance
cdf	cumulative distribution function
CRAN	Comprehensive R Archive Network
DAG	directed acyclic graph
EM	expectation-maximisation
fBm	Fractional Brownian motion
GPR	Gaussian process regression
HMC	Hamiltonian Monte Carlo
HPM	highest probability model
IIA	independent of irrelevant alternatives
iid	Identical and independently distributed
Lasso	Least absolute shrinkage and selection operator
MCMC	Markov chain Monte Carlo
ML	maximum likelihood
MMSE	minimum mean squared error
OLS	ordinary least squares
pd/p.d.	positive definite
pdf	probability density function
PIP	posterior inclusion probability
pmf	probability mass function
PMP	posterior model probability
RKHS	Reproducing kernel Hilbert space
RKKS	Reproducing kernel Kreĭn space
RSS	residual sum of squares
SE	Squared exponential (kernel)

Chapter 1

Introduction

Regression analysis is undoubtedly one of the most important tools available at a practitioner's disposal to understand the relationship between one or more explanatory variables x, and the independent variable of interest, y. This relationship is usually expressed as $y \approx f(x|\theta)$, where f is called the *regression function*, and this is dependent on one or more parameters denoted by θ. Regression analysis concerns the estimation of said regression function, and once a suitable estimate $\hat{f}$ has been found, post-estimation procedures such as prediction and inference surrounding f or θ, may be performed.

Estimation of the regression function may be done in many ways. This thesis concerns the use of *I-priors* (Bergsma, 2018), in a semi-Bayesian manner, for regression modelling. The I-prior is an objective, entropy-maximising prior for the regression function which makes use of its Fisher information. The essence of regression modelling using I-priors is introduced briefly below, but as the development of I-priors is fairly recent, we dedicate two full chapters (Chapters 2 and 3) to describe the concept fully.

This thesis has three main chapters which we hope to present as methodological innovations surrounding the use of I-priors for regression modelling. Chapter 4 describes the I-prior modelling framework and computational methods relating to the estimation of I-prior models. Chapter 5 extends the I-prior methodology to fit categorical outcome models. Chapter 6 discusses the use of I-priors in variable selection for linear models. In addition to introducing the statistical model of interest and motivating the use of I-priors, this current introductory chapter ultimately provides a summary outline of the thesis.

1.1 Regression models

For subject $i \in \{1, \ldots, n\}$, assume a real-valued response y_i has been observed, as well as a row vector of p covariates $x_i = (x_{i1}, \ldots, x_{ip})$, where each x_{ik} belongs to some set $\mathcal{X}_k$, for $k = 1, \ldots, p$. Let $\mathcal{S} = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ denote this observed sample of size n. Consider then the following regression model, which stipulates the dependence of the y_i on the x_i:

$$y_i = \alpha + f(x_i) + \epsilon_i, \tag{1.1}$$

where f is some regression function to be estimated, and α is an intercept. Additionally, it is assumed that the errors ϵ_i are normally distributed according to

$$(\epsilon_1, \ldots, \epsilon_n)^\top \sim \mathrm{N}_n(0, \boldsymbol{\Psi}^{-1}). \tag{1.2}$$

where $\boldsymbol{\Psi} = (\psi_{ij})_{i,j=1}^n$ is the precision matrix. We shall often refer to model (1.1) subject to (1.2) as the *normal regression model*. The choice of multivariate normal errors is not only a convenient one (as far as distributional assumptions go), but one that is motivated by the principle of maximum entropy (Jaynes, 1957a, 1957b, 2003).

Interestingly, a wide variety of statistical models can be captured by the seemingly humble normal regression model, simply by varying the form of the regression function f. For instance, when f can be parameterised linearly as $f(x_i) = x_i^\top \beta$, $\beta \in \mathbb{R}^p$, we then have the ordinary linear regression—a staple problem in statistics and other quantitative fields.

We might also have data that is separated naturally into groups or levels by design, for example, data from stratified sampling, students within schools, or longitudinal measurements over time. In such a case, we might want to consider a regression function with additive components

$$f(x_i^{(j)}, j) = f_1(x_i^{(j)}) + f_2(j) + f_{12}(x_i^{(j)}, j)$$

where $x_i^{(j)}$ denotes the p-dimensional i'th observation for group $j \in \{1, \ldots, m\}$. Again, assuming a linear parameterisation, this is recognisable as the standard multilevel or random-effects linear model (Rabe-Hesketh and Skrondal, 2012), with f_2 representing the varying intercept via $f_2(j) = \alpha_j$, f_{12} representing the varying slopes via $f_{12}(x_{ij}, j) = x_i^\top \beta_j$, with $\beta_j \in \mathbb{R}^p$, and f_1 representing the fixed-effects linear component $x_i^\top \beta$ as above.

Moving on from linear models, smoothing models may be of interest as well. A myriad of models exist for this type of problem, with most classed as nonparametric regression (Wassermann, 2006), and the more popular ones include LOcal regrESSion (LOESS), kernel regression, and smoothing splines (Wahba, 1990). Semiparametric regression

models, on the other hand, combines the linear component of a regression model with a non-parameteric component.

Further, the regression problem is made more intriguing when the set of covariates $\mathcal{X}$ is functional—in which case the linear regression model aims to estimate coefficient functions $\beta : \mathcal{T} \to \mathbb{R}$ from the model

$$y_i = \int_{\mathcal{T}} x_i(t)\beta(t)\,\mathrm{d}t + \epsilon_i.$$

Nonparametric and semiparametric regression with functional covariates (Ramsay and Silverman, 2005) have also been widely explored. Models of this nature still fall under the remit of the normal regression model by selecting a regression functional with domain over the functional covariates.

1.2 Vector space of functions

It would be beneficial to prescribe some sort of structure for which estimation of the regression function can be carried out easily and reliably. This needed structure is given to us by assuming that our regression function f for the normal model lies in some reproducing kernel Hilbert or Kreĭn space (RKHS/RKKS) $\mathcal{F}$ equipped with the reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Often, the reproducing kernel (or simply kernel, for short) is shaped by one or more parameters which we shall denote by η. Correspondingly, the kernel is rightfully denoted h_η to indicate the dependence of the parameters on the kernels, though where this is seemingly obvious, might be omitted. For I-prior modelling, which is the focus of this thesis, we make the assumption that our regression function lies in a RKKS $\mathcal{F}$.

RKKSs, and more popularly RKHSs, provide a geometrical advantage to learning algorithms: projections of the inputs to a richer and more informative (and usually higher dimensional) *feature space*, where learning is more likely to be successful, need not be figured out explicitly. Instead, *feature maps* are implicitly calculated by the use of kernel functions. This is known as the "kernel trick" in the machine learning literature (Hofmann et al., 2008), and it has facilitated the success of kernel methods for learning, particularly in algorithms with inner products involving the transformed inputs.

Due to the one-to-one mapping between the set of kernel functions and the set of RKHSs, choosing the space in which the regression function lies is equivalent to choosing a particular kernel function, and this is chosen according to the desired effects of the covariates on the regression function. RKKSs on the other hand also possess unique kernels, but every (generalised) kernel[1] is associated to *at least* one RKKS. An in-depth

[1] By generalised kernels, we mean kernels that are not necessarily positive definite in nature.

discussion (including the motivation for their use) on kernels, RKHSs and RKKSs will be provided later in Chapter 2, but for now, it suffices to say that kernels which invoke either a linear, smooth or categorical dependence, or any combinations thereof, are of interest. This would allow us to fit the various models described earlier within this RKHS/RKKS framework.

1.3 Estimating the regression function

Having decided on a vector space $\mathcal{F}$, we now turn to the task of choosing the best $f \in \mathcal{F}$ that fits the data sample $\mathcal{S}$. 'Best' here could mean a great deal of things, such as choosing f which minimises an empirical risk measure[2] defined by

$$\hat{\mathrm{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \Lambda\big(y_i, f(x_i)\big)$$

for some loss function $\Lambda : \mathbb{R}^2 \to [0, \infty)$. A common choice for the loss function is the *squared loss function*

$$\Lambda\big(y_i, f(x_i)\big) = \sum_{j=1}^{n} \psi_{ij}\big(y_i - f(x_i)\big)\big(y_j - f(x_j)\big),$$

and when used, defines the *(generalised) least squares regression*. For the normal model, the minimiser of the empirical risk measure under the squared loss function is also the maximum likelihood (ML) estimate of f, since $\hat{\mathrm{R}}(f)$ would be twice the negative log-likelihood of f, up to a constant.

The ML estimator of f typically interpolates the data if the dimension of $\mathcal{F}$ is at least n, so is of little use. The most common method to overcome this issue is *Tikhonov regularisation*, whereby a regularisation term is added to the risk function, with the aim of imposing a penalty on the complexity of f. In particular, smoothness assumptions on f can be represented by using its reproducing kernel Hilbert space (RKHS) norm $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}$ as the regularisation term[3]. Therefore, the solution to the regularised least squares problem—call this f_{reg}—is the minimiser of the mapping from $\mathcal{F}$ to $\mathbb{R}$ defined by

$$f \mapsto \overbrace{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij}\big(y_i - f(x_i)\big)\big(y_j - f(x_j)\big)}^{\text{data fit term}} + \overbrace{\lambda^{-1}\|f - f_0\|_{\mathcal{F}}^2}^{\text{penalty term}}, \tag{1.3}$$

[2] More appropriately, the risk functional $\mathrm{R}(f) = \int \Lambda(y, f(x))\,\mathrm{dP}(y, x)$, i.e. the expectation of the loss function under some probability measure of the observed sample, should be used. Often the true probability measure is not known, so the empirical risk measure is used instead.

which also happens to be the *penalised maximum likelihood* solution. Here, $f_0 \in \mathcal{F}$ can be thought of a prior 'best guess' for the function f. The $\lambda^{-1} > 0$ parameter—known as the regularisation parameter—controls the trade-off between the data-fit term and the penalty term in (1.3), and is not usually known a priori and must be estimated.

An attractive consequence of the representer theorem (Kimeldorf and Wahba, 1970) for Tikhonov regularisation implies that f_{reg} admits the form

$$f_{\text{reg}} = f_0 + \sum_{i=1}^{n} h(\cdot, x_i) w_i, \qquad w_i \in \mathbb{R}, \ \forall i = 1, \dots, n, \tag{1.4}$$

even if $\mathcal{F}$ is infinite dimensional. This simplifies the original minimisation problem from a search for f over a possibly infinite-dimensional domain, to a search for the optimal coefficients w_i in n dimensions.

Tikhonov regularisation also has a well known Bayesian interpretation, whereby the regularisation term encodes prior information about the function f. For the normal regression model with $f \in \mathcal{F}$, an RKHS, it can be shown that f_{reg} is the posterior mean of f given a *Gaussian process prior* (Rasmussen and Williams, 2006) with mean f_0 and covariance kernel $\text{Cov}\big(f(x_i), f(x_j)\big) = \lambda h(x_i, x_j)$. The exact solution for the coefficients $\mathbf{w} := (w_1, \dots, w_n)^\top$ are in fact $\mathbf{w} = \big(\mathbf{H} + \mathbf{\Psi}^{-1}\big)^{-1}(\mathbf{y} - \mathbf{f}_0)$, where $\mathbf{H} = \big(h(x_i, x_j)\big)_{i,j=1}^{n}$ (often referred to as the Gram matrix or kernel matrix) and $(\mathbf{y} - \mathbf{f}_0) = (y_1 - f_0(x_1), \dots, y_n - f_0(x_n))^\top$.

1.4 Regression using I-priors

Building upon the Bayesian interpretation of regularisation, Bergsma (2018) proposes an original prior distribution for the regression function such that its realisations admit the form for the solution given in the representer theorem. The *I-prior* for the regression function f in (1.1) subject to (1.2) and $f \in \mathcal{F}$, a RKKS with kernel h_η, is defined as the distribution of a random function of the form (1.4) when the w_i are distributed according to

$$(w_1, \dots, w_n)^\top \sim \text{N}_n(\mathbf{0}, \mathbf{\Psi}),$$

where $\mathbf{0}$ is a length n vector of zeroes. As a result, we may view the I-prior for f as having the Gaussian process distribution

$$\mathbf{f} := \big(f(x_1), \dots, f(x_n)\big)^\top \sim \text{N}_n(\mathbf{f}_0, \mathbf{H}_\eta \mathbf{\Psi} \mathbf{H}_\eta) \tag{1.5}$$

[3] Concrete notions of complexity penalties can be introduced if $\mathcal{F}$ is a normed space, though RKHSs are typically used as it gives great conveniences (see Chapter 2).

with $\mathbf{H}_\eta$ an $n \times n$ matrix with (i, j) entries equal to $h_\eta(x_i, x_j)$, and $\mathbf{f}_0$ a vector containing the $f_0(x_i)$'s. The covariance matrix of this multivariate normal prior is related to the Fisher information for f, and hence the name I-prior—the 'I' stands for information. Furthermore, the I-prior happens to be an entropy-maximising prior, subject to certain constraints. Chapter 3 contains details of the derivation of I-priors for the normal regression model.

As with Gaussian process regression (GPR), the function f is estimated by its posterior mean. For the normal model, the posterior distribution for the regression function conditional on the responses $\mathbf{y} = (y_1, \ldots, y_n)$,

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}}, \tag{1.6}$$

can easily be found, and it is in fact normally distributed. The posterior mean for f evaluated at a point $x \in \mathcal{X}$ is given by

$$\mathrm{E}\left[f(x)\middle|\mathbf{y}\right] = f_0(x) + \mathbf{h}_\eta^\top(x)\overbrace{\boldsymbol{\Psi}\mathbf{H}_\eta\big(\mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}\big)^{-1}(\mathbf{y} - \mathbf{f}_0)}^{\tilde{\mathbf{w}}} \tag{1.7}$$

where we have defined $\mathbf{h}_\eta(x)$ to be the vector of length n with entries $h_\eta(x, x_i)$ for $i = 1, \ldots, n$. Incidentally, the elements of the n-vector $\tilde{\mathbf{w}}$ defined in (1.7) are the posterior means of the random variables w_i in the formulation (1.4). The point-evaluation posterior variance for f is given by

$$\mathrm{Var}\left[f(x)\middle|\mathbf{y}\right] = \mathbf{h}_\eta^\top(x)\big(\mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}\big)^{-1}\mathbf{h}_\eta^\top(x). \tag{1.8}$$

Prediction for a new data point $x_{\text{new}} \in \mathcal{X}$ then concerns obtaining the *posterior predictive distribution*

$$p(y_{\text{new}}|\mathbf{y}) = \int p(y_{\text{new}}|f_{\text{new}}, \mathbf{y})p(f_{\text{new}}|\mathbf{y})\,\mathrm{d}f_{\text{new}},$$

where we had defined $f_{\text{new}} := f(x_{\text{new}})$. This is again a normal distribution in the case of the normal model, with similar mean and variance as in (1.7). For a derivation, see Section 4.2 in Chapter 4 for details.

There is also the matter of optimising model parameters θ, which in our case, collectively refers to the kernel parameters η and the precision matrix of the errors $\boldsymbol{\Psi}$. Model parameters θ may be estimated in several ways, either by likelihood-based methods or fully Bayesian methods. The former includes methods such as direct maximisation of the (marginal) likelihood, $L(\theta) = \int p(\mathbf{y}|\theta, \mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}$, and the expectation-maximisation (EM) algorithm. Both are seen as a form of *empirical Bayes* estimation, or a type-II ML estimation (Bishop, 2006), as it is known in machine learning. In a fully Bayesian setting on the other hand, Markov chain Monte Carlo (MCMC) may be employed, assuming prior distributions on the model parameters.

1.5 Advantages and limitations of I-priors

The I-prior methodology has the following advantages:

1. **A unifying methodology for various regression models.**

   The I-prior methodology has the ability to fit a multitude of regression models simply by choosing the RKKS to which the regression function belongs. As such, it can be seen as a unifying methodology for various parametric and non-parametric regression models including additive models, multilevel models and models with one or more functional covariates.

2. **Simple estimation procedure.**

   Estimation of model parameters using the aforementioned methods are very simple to implement, barring any computational and numerical hurdles, which will be discussed.

3. **Parsimonious specification.**

   I-prior models are most typically specified using only RKHS scale parameters and the error precision. This encourages parsimony in model building; for example, smoothing models can be fitted using only two parameters, while linear multilevel models can be fitted with notably fewer parameters than the standard versions.

4. **Prevents over-fitting and under-smoothing.**

   As alluded to earlier, any function f that passes through the data points is a least squares solution. Regularising the problem with the use of I-priors prevents overfitting, with the added advantage that the posterior solution under an I-prior does not tend to under smooth as much as Tikhonov regularisation does (Bergsma, 2018). Under smoothing can adversely impact the estimate of f, and in real terms might even show features and artefacts that are not really there.

5. **Better prediction.**

   Empirical studies and real-data examples show that predictive performance of I-priors are comparative to, and often better than, other leading state-of-the-art models, including the closely related GPR.

6. **Straightforward inference.**

   Marginal likelihoods after integrating out the I-prior are easily obtained, making model selection via likelihood comparison a viable option. This method of comparing marginal likelihood with maximum likelihood estimate plug-ins of the model parameters, is viewed as empirical Bayes factors comparison in the Bayesian literature.

The main drawback of using I-prior models is computational in nature, namely, the requirement of working with an $n \times n$ matrix and its inverse, as seen in equations (1.7) and (1.8), regardless of estimation method (ML or Bayes). Analysis of data sets that are not more than a few thousand in size can be considered feasible; anything more than this is debilitatingly slow to compute. In addition, care must be taken to avoid numerical instabilities when calculating the marginal log-likelihood during parameter estimation, which can affect gradient based optimisation or the EM algorithm.

Another issue when performing likelihood-based methods is that the optimisation objective may be non-convex such that multiple local optima may exist. In such cases, multiple restarts from different initialisations may ultimately lead to a global maximum, although difficulties may be faced if numerical instabilities occur.

Lastly, a remark on model assumptions, which are twofold: 1) the assumption of $f \in \mathcal{F}$ an RKKS; and 2) normality of errors. Of the two, the latter is more likely to be violated, especially when dealing with discrete responses, e.g. in classification. Deviating from the normality assumption would require approximation techniques to be implemented in order to obtain the posterior distributions of interest.

1.6 Outline of thesis

This thesis is structured as follows:

- Following this introductory chapter, **Chapter 2** provides a brief overview of functional analysis, and in particular, descriptions of interesting function spaces for regression. In **Chapter 3**, the concept of the Fisher information is extended to potentially infinite-dimensional parameters. This allows us to define the Fisher information for the regression function which parameterises the normal regression model, and we explain how this relates to the I-prior.
- The aforementioned computational methods relating to the estimation of I-prior models are explored in **Chapter 4**, namely the direct optimisation of the log-likelihood, the EM algorithm, and MCMC methods. The goal is to describe stable and efficient algorithms for estimating I-prior models. The R package **iprior** (Jamil, 2017) is the culmination of the effort put in towards completing this chapter, which has been made publicly available on the CRAN.
- Many models of interest involve response variables of a categorical nature. A naïve implementation of the I-prior model is certainly possible, but proper ways do exist to handle non-normality of errors. **Chapter 5** extends the I-prior methodology to discrete outcomes. There, the non-Gaussian likelihood that arises in the posteriors

are approximated by way of variational inference. The advantages of the I-prior in normal regression models carry over into categorical response models.

- **Chapter 6** is a contribution to the area of variable selection. Specifically for linear models with p variables to select from, model comparison requires elucidation of 2^p marginal likelihoods, and this becomes infeasible when p is large. To circumvent this issues, we use a stochastic search method to choose models that have high posterior probabilities of occurring, equivalent to choosing models that have large Bayes factors. We experiment with the use of I-priors to improve false selections, especially in the presence of multicollinearity.

Chapters 4 to 6 contain R computer implementations of the statistical methodologies described therein, and the code for replication are made available at `http://myphdcode.haziqj.ml`.

Familiarity with basic estimation concepts (maximum likelihood, Bayes, empirical Bayes) and their corresponding estimation methods (gradient-based methods, Newton, quasi-Newton methods, MCMC, EM algorithm) are assumed throughout. Brief supplementary chapters are attached for readers who wish to familiarise themselves with topics such as variational inference and Hamiltonian Monte Carlo, which are used in Chapters 4 and 5. These brief readings are designed to be supplementary in nature, and are not strictly essential for the chapters. Additionally, Appendices A–I contain references to several statistical distributions and their properties, derivations, and proofs of the algorithms described in this thesis.

On a closing note, a dedicated website for this PhD project has been created, and it can be viewed at `http://phd.haziqj.ml`.

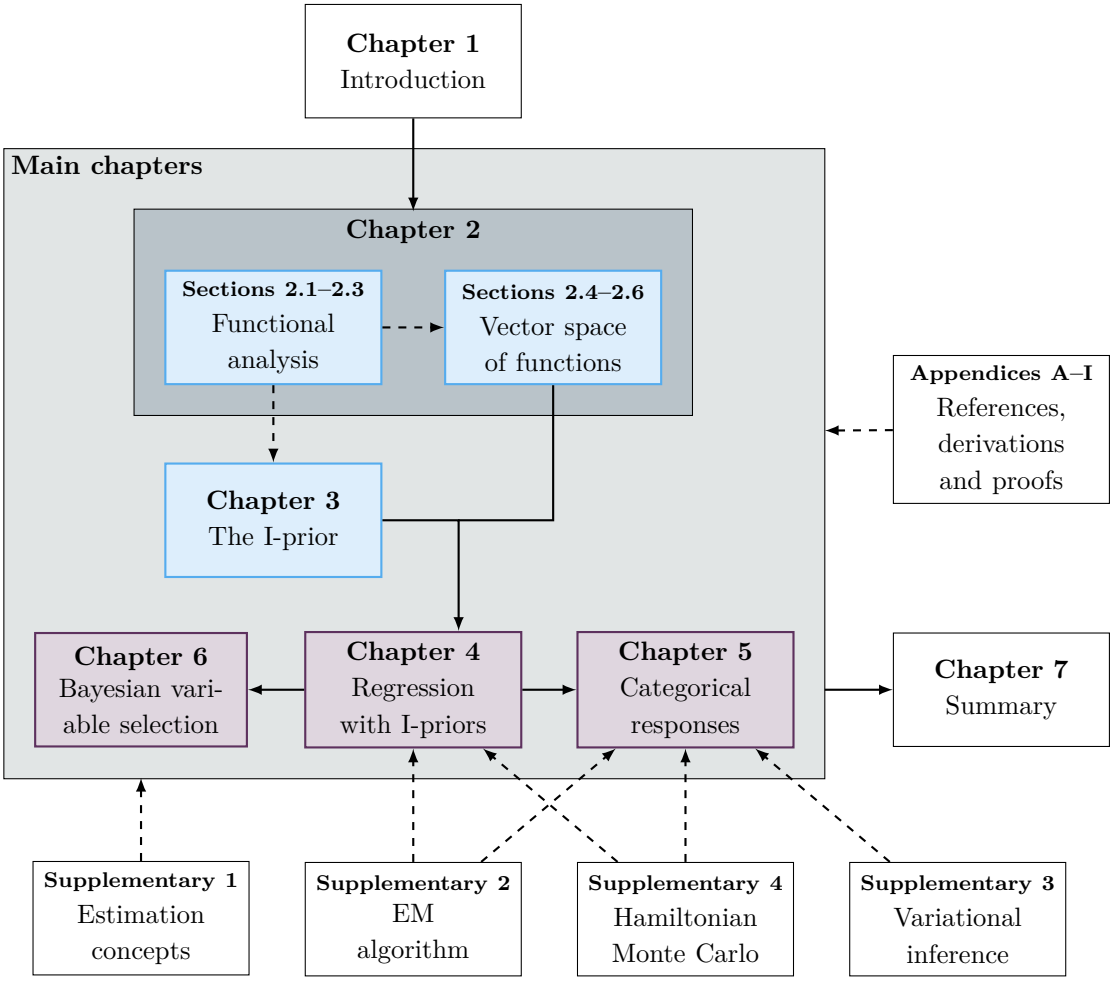


Figure 1.1: Schematic representation of the organisation of the chapters of this thesis. Solid lines indicate requisite relevances, while dashed lines indicate supporting and supplementary relevances. Chapters indicated by **blue** boxes are theoretical in nature, while those in **purple** are methodological.

Chapter 2

Vector space of functions

For regression modelling with I-priors, it is assumed that the regression functions lie in some vector space of functions. The purpose of this chapter is to provide a concise review of functional analysis leading up to the theory of reproducing kernel Hilbert and Kreĭn spaces (RKHS/RKKS). The interest with these RKHSs and RKKSs is that these spaces have well established mathematical structure and offer desirable topologies. In particular, it allows the possibility of deriving the Fisher information for regression functions—this will be covered in Chapter 3. As we shall see, RKHSs are also extremely convenient in that they may be specified completely via their reproducing kernels. Several of these function spaces are of interest to us, for example, spaces of linear functions, smoothing functions, and functions whose inputs are nominal values and even functions themselves. RKHSs are widely studied in the applied statistical and machine learning literature, but perhaps RKKSs are less so. To provide an early insight, RKKSs are simply a generalisation of RKHSs, and are defined as the difference between two RKHSs. The flexibility provided by RKKSs will prove both useful and necessary, especially when considering sums and products of scaled function spaces, as is done in I-prior modelling.

It is emphasised that a deep knowledge of functional analysis, including RKHS and RKKS theory, is not at all necessary for I-prior modelling, so perhaps the advanced reader may wish to skip Sections 2.1 to 2.3. Section 2.4 describes the fundamental RKHS of interest for I-prior regression, which we refer to as the "building block" RKHSs. The reason for this is that it is possible to construct new function spaces from existing ones, and this is described in Section 2.5.

Two remarks before starting. Firstly, on notation: sets and vector spaces are denoted by calligraphic letters, and as much as possible, we shall stick to the convention that $\mathcal{F}$ denotes a function space, and $\mathcal{X}$ denotes the set of covariates or function inputs. Occasionally, we will describe a generic Hilbert space denoted by $\mathcal{H}$. Elements of the vector space of real functions over a set $\mathcal{X}$ are denoted $f(\cdot)$, but more commonly and

simply f. This distinguishes them from the actual evaluation of the function at an input point $x \in \mathcal{X}$, denoted $f(x) \in \mathbb{R}$. For a much cleaner read, we dispense with boldface notation for vectors and matrices when talking about them, without ambiguity, in the abstract sense. Secondly, on bibliography: references will be minimised throughout the presentation of this chapter, but a complete annotated bibliography at the end in Section 2.6.

2.1 Some functional analysis

The core study of functional analysis revolves around the treatment of functions as objects in vector spaces over a field[1]. Vector spaces, or linear spaces as they are sometimes known, may be endowed with some kind of structure so as to allow ideas such as closeness and limits to be conceived. Of particular interest to us is the structure brought about by *inner products*, which allow the rigorous mathematical study of various geometrical concepts such as lengths, directions, and orthogonality, among other things. We begin with the definition of an inner product.

Definition 2.1 (Inner products)**.** Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is said to be an inner product on $\mathcal{F}$ if all of the following are satisfied:

- **Symmetry**. $\langle f, g \rangle_{\mathcal{F}} = \langle g, f \rangle_{\mathcal{F}}$, $\forall f, g \in \mathcal{F}$.
- **Linearity**. $\langle \lambda_1 f_1 + \lambda_2 f_2, g \rangle_{\mathcal{F}} = \lambda_1 \langle f_1, g \rangle_{\mathcal{F}} + \lambda_2 \langle f_2, g \rangle_{\mathcal{F}}$, $\forall f_1, f_2, g \in \mathcal{F}$, $\forall \lambda_1, \lambda_2 \in \mathbb{R}$.
- **Non-degeneracy**. $\langle f, f \rangle_{\mathcal{F}} = 0 \Leftrightarrow f = 0$.

Additionally, an inner product is said to be *positive definite* if $\langle f, f \rangle_{\mathcal{F}} \geq 0$, $\forall f \in \mathcal{F}$. Inner products need not necessarily be positive definite, and we shall revisit this fact later when we cover Kreĭn spaces. However, for the purposes of the forthcoming discussion, the inner products that are referenced are the positive definite kind, unless otherwise stated.

We can always define a *norm* on $\mathcal{F}$ using the inner product as

$$\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}}. \tag{2.1}$$

Norms are another form of structure that specifically captures the notion of length. This is defined below.

Definition 2.2 (Norms)**.** Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A non-negative function $|| \cdot ||_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to [0, \infty)$ is said to be a norm on $\mathcal{F}$ if all of the following are satisfied:

- **Absolute homogeneity**. $||\lambda f||_{\mathcal{F}} = |\lambda| \, ||f||_{\mathcal{F}}$, $\forall \lambda \in \mathbb{R}$, $\forall f \in \mathcal{F}$

[1] In this thesis, this will be $\mathbb{R}$ exclusively.

- **Subadditivity**. $||f+g||_{\mathcal{F}} \leq ||f||_{\mathcal{F}} + ||g||_{\mathcal{F}}$, $\forall f, g \in \mathcal{F}$
- **Point separating**. $||f||_{\mathcal{F}} = 0 \Leftrightarrow f = 0$

The subadditivity property is also known as the *triangle inequality*. Also note that since $\|-f\|_{\mathcal{F}} = \|f\|_{\mathcal{F}}$, and by the triangle inequality and point separating property, we have that $\|f\|_{\mathcal{F}} = \frac{1}{2}\|f\|_{\mathcal{F}} + \frac{1}{2}\|-f\|_{\mathcal{F}} \geq \frac{1}{2}\|f-f\|_{\mathcal{F}} = 0$, thus implying non-negativity of norms. Several important relationships between norms and inner products hold in linear spaces, namely, the *Cauchy-Schwarz inequality*

$$|\langle f, g\rangle_{\mathcal{F}}| \leq \|f\|_{\mathcal{F}} \|g\|_{\mathcal{F}};$$

the *parallelogram law*

$$\|f+g\|_{\mathcal{F}}^2 + \|f-g\|_{\mathcal{F}}^2 = 2\|f\|_{\mathcal{F}}^2 + 2\|g\|_{\mathcal{F}}^2;$$

and the *polarisation identity* (in various forms)

$$\begin{aligned} \|f+g\|_{\mathcal{F}}^2 - \|f-g\|_{\mathcal{F}}^2 &= 4\langle f, g\rangle_{\mathcal{F}}, \\ \|f+g\|_{\mathcal{F}}^2 - \|f\|_{\mathcal{F}}^2 - \|g\|_{\mathcal{F}}^2 &= 2\langle f, g\rangle_{\mathcal{F}}, \\ -\|f-g\|_{\mathcal{F}}^2 + \|f\|_{\mathcal{F}}^2 + \|g\|_{\mathcal{F}}^2 &= 2\langle f, g\rangle_{\mathcal{F}}, \end{aligned}$$

for any $f, g \in \mathcal{F}$.

A vector space endowed with an inner product (c.f. norm) is called an inner product space (c.f. normed vector space). As a remark, inner product spaces can always be equipped with a norm using (2.1), but not always the other way around. A norm needs to satisfy the parallelogram law for an inner product to be properly defined.

The norm $||\cdot||_{\mathcal{F}}$, in turn, induces a metric (a notion of distance) on $\mathcal{F}$, i.e. $D(f, g) = ||f-g||_{\mathcal{F}}$, for $f, g \in \mathcal{F}$. With these notions of distances, one may talk about sequences of functions in $\mathcal{F}$ which are *convergent*, and sequences whose elements become arbitrarily close to one another as the sequence progresses (*Cauchy*).

Definition 2.3 (Convergent sequence)**.** A sequence $\{f_n\}_{n=1}^{\infty}$ of elements of a normed vector space $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ is said to *converge* to some $f \in \mathcal{F}$, if for every $\epsilon > 0$, $\exists N = N(\epsilon) \in \mathbb{N}$, such that $\forall n > N$, $||f_n - f||_{\mathcal{F}} < \epsilon$.

Definition 2.4 (Cauchy sequence)**.** A sequence $\{f_n\}_{n=1}^{\infty}$ of elements of a normed vector space $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ is said to be a Cauchy sequence if for every $\epsilon > 0$, $\exists N = N(\epsilon) \in \mathbb{N}$, such that $\forall n, m > N$, $||f_n - f_m||_{\mathcal{F}} < \epsilon$.

Every convergent sequence is Cauchy (from the triangle inequality), but the converse is not true. If the limit of the Cauchy sequence exists within the vector space, then the

sequence converges to it. A vector space is said to be *complete* if it contains the limits of all Cauchy sequences, or in other words, if every Cauchy sequence converges. There are special names given to complete vector spaces. A complete inner product space is known as a *Hilbert space*, while a complete normed space is called a *Banach space*. Out of interest, an inner product space that is not complete is sometimes known as a *pre-Hilbert space*, since its completion with respect to the norm induced by the inner product is a Hilbert space.

A subset $\mathcal{G} \subseteq \mathcal{F}$ is a *closed subspace* of $\mathcal{F}$ if it is closed under addition and multiplication by a scalar. That is, for any $g, g' \in \mathcal{G}$, $\lambda_1 g + \lambda_2 g'$ is also in $\mathcal{G}$, for $\lambda_1, \lambda_2 \in \mathbb{R}$. For Hilbert spaces, each closed subspace is also complete, and thus a Hilbert space in its own right. Although, as a remark, not every Hilbert subspace need be closed, and therefore complete.

Being vectors in a vector space, we can discuss mapping of vectors onto a another space, or in essence, having a function acted upon them. To establish terminology, we define linear and bilinear maps (operators).

Definition 2.5 (Linear map/operator)**.** Let $\mathcal{F}$ and $\mathcal{G}$ be two Hilbert spaces over $\mathbb{R}$. An operator A is a map from $\mathcal{F}$ to $\mathcal{G}$, and we denote its action on a function $f \in \mathcal{F}$ as $A(f) \in \mathcal{G}$, or simply $Af \in \mathcal{G}$. A *linear operator* satisfies $A(f + f') = A(f) + A(f')$ and $A(\lambda f) = \lambda A(f)$, for all $f, f' \in \mathcal{F}$ and $\lambda \in \mathbb{R}$. If $\mathcal{G}$ is the base field ($\mathbb{R}$ in our case), then the linear operator A is called a *linear functional*.

Definition 2.6 (Bilinear map/operator)**.** Let $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$ be Hilbert spaces over $\mathbb{R}$. A *bilinear operator* $B : \mathcal{F} \times \mathcal{G} \to \mathcal{H}$ is linear in each argument separately, i.e.

- $B(\lambda_1 f + \lambda_2 f', h) = \lambda_1 B(f, h) + \lambda_2 B(f', h)$; and
- $B(f, \lambda_1 g + \lambda_2 g') = \lambda_1 B(f, g) + \lambda_2 B(f, g')$,

for all $f, f' \in \mathcal{F}$, $g, g' \in \mathcal{G}$ and $\lambda_1, \lambda_2 \in \mathbb{R}$. In other words, the mappings $B_g : f \mapsto B(f, g)$ for any $g \in \mathcal{G}$, and $B_f : g \mapsto B(f, g)$ for any $f \in \mathcal{F}$, are both linear maps. If $\mathcal{F} \equiv \mathcal{G}$, then the bilinear map is *symmetric*. If $\mathcal{H}$ is the base field ($\mathbb{R}$ in our case), then B is called a *bilinear form*.

An interesting property of these operators to look at, besides linearity, is whether or not they are *continuous*.

Definition 2.7 (Continuity)**.** Let $\mathcal{F}$ and $\mathcal{G}$ be two Hilbert spaces. A function $A : \mathcal{F} \to \mathcal{G}$ is said to be *continuous at* $g \in \mathcal{F}$, if for every $\epsilon > 0$, $\exists \delta = \delta(\epsilon, g) > 0$ such that

$$\|f - g\|_{\mathcal{F}} < \delta \quad \Rightarrow \quad \|Af - Ag\|_{\mathcal{G}} < \epsilon.$$

A is *continuous* on $\mathcal{F}$, if it is continuous at every point $g \in \mathcal{F}$. If, in addition, δ depends on ϵ only, A is said to be *uniformly continuous*.

Continuity in the sense of linear operators here means that a convergent sequence in $\mathcal{F}$ can be mapped to a convergent sequence in $\mathcal{G}$. For a particular linear operator, the evaluation functional, this means that closeness in norm implies pointwise closeness—this relates to RKHSs, which is discussed in Section 2.2. There is an even stronger notion of continuity called *Lipschitz continuity*.

Definition 2.8 (Lipschitz continuity)**.** Let $\mathcal{F}$ and $\mathcal{G}$ be two Hilbert spaces. A function $A : \mathcal{F} \to \mathcal{G}$ is *Lipschitz continuous* if $\exists M > 0$ such that $\forall f, f' \in \mathcal{F}$,

$$\|Af - Af'\|_{\mathcal{G}} \leq M\|f - f'\|_{\mathcal{F}}.$$

Clearly, Lipschitz continuity implies uniform continuity: choose $\delta = \delta(\epsilon) := \epsilon/M$ and replace this in Definition 2.7. A continuous, linear operator is also one that is bounded.

Definition 2.9 (Bounded operator)**.** The linear operator $A : \mathcal{F} \to \mathcal{G}$ between two Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ is said to be *bounded* if there exists some $M > 0$ such that

$$\|Af\|_{\mathcal{G}} \leq M\|f\|_{\mathcal{F}}.$$

The smallest such M is defined to be the *operator norm*, denoted $\|A\| := \sup_{f \in \mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$.

Lemma 2.1 (Equivalence of boundedness and continuity)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be two Hilbert spaces, and $A : \mathcal{F} \to \mathcal{G}$ a linear operator. A is bounded if and only if it is continuous.*

Proof. Suppose that A is bounded. Then, $\forall f, f' \in \mathcal{F}$, $\exists M > 0$ such that $\|A(f - f')\|_{\mathcal{G}} \leq M\|f - f'\|_{\mathcal{G}}$, so A is Lipschitz continuous. Conversely, let A be a continuous linear operator, especially at the zero vector. In other words, $\exists \delta > 0$ such that $\|A(f)\|_{\mathcal{G}} = \|A(f+0-0)\|_{\mathcal{G}} = \|A(f) - A(0)\| \leq 1$, $\forall f \in \mathcal{F}$ whenever $\|f\|_{\mathcal{F}} \leq \delta$. Thus, for all non-zero $f \in \mathcal{F}$,

$$\begin{aligned}
\|A(f)\|_{\mathcal{G}} &= \left\| \frac{\|f\|_{\mathcal{F}}}{\delta} A\left(\frac{\delta}{\|f\|_{\mathcal{F}}} f\right) \right\|_{\mathcal{G}} \\
&= \left| \frac{\|f\|_{\mathcal{F}}}{\delta} \right| \left\| A\left(\frac{\delta}{\|f\|_{\mathcal{F}}} f\right) \right\|_{\mathcal{G}} \\
&\leq \frac{\|f\|_{\mathcal{F}}}{\delta} \cdot 1,
\end{aligned}$$

and therefore A is bounded. ∎

So important is the concept of linearity and continuity, that there are specially named spaces which contain linear and continuous functionals.

Definition 2.10 (Dual spaces)**.** Let $\mathcal{F}$ be a Hilbert space. The space $\mathcal{F}^\vee$ of *linear functionals* is called the *algebraic dual space* of $\mathcal{F}$. The space $\mathcal{F}^*$ of *continuous linear functionals* is called the *continuous dual space* or alternatively, the *topological dual space*, of $\mathcal{F}$.

As it turns out, the algebraic dual space and continuous dual space coincide in finite-dimensional Hilbert spaces: take any $A \in \mathcal{F}^\vee$; since A is finite-dimensional, it is bounded, and therefore continuous (see Lemma 2.1), so $A \in \mathcal{F}^*$ and $\mathcal{F}^\vee \subseteq \mathcal{F}^*$; but $\mathcal{F}^* \subseteq \mathcal{F}^\vee$ trivially, so $\mathcal{F}^\vee \equiv \mathcal{F}^*$. For infinite-dimensional Hilbert spaces, this is not so, but in any case, we will only be considering the continuous dual space in this thesis. The following result is an important one, which states that continuous linear functionals of an inner product space are nothing more than inner products.

Theorem 2.2 (Riesz-Fréchet)**.** *Let $\mathcal{F}$ be a Hilbert space. Every element A of the continuous dual space $\mathcal{F}^*$, i.e. all continuous linear functionals $A : \mathcal{F} \to \mathbb{R}$, can be uniquely written in the form $\langle \cdot, g\rangle_\mathcal{F} =: A_g \in \mathcal{F}^*$, for some $g \in \mathcal{F}$. Moreover, $\|g\|_\mathcal{F} = \|A_g\|_{\mathcal{F}^*}$.*

Proof. Omitted—see Yamamoto (2012, Theorem 4.2.1) for a proof. ■

Remark 2.1. The Riesz-Fréchet theorem is also commonly referred to as the *Riesz representation theorem* for Hilbert spaces.

The notion of isometry (transformation that preserves distance) is usually associated with metric spaces; two metric spaces being isometric means that they identical as far as their metric properties are concerned. For Hilbert spaces (and more generally, for normed spaces), there is an analogous concept as well in *isometric isomorphism* (a bijective isometry), such that two Hilbert spaces being isometrically isomorphic imply that they have exactly the same geometric structure, but may very well contain fundamentally different objects.

Definition 2.11 (Isometric isomorphism)**.** Two Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ are said to be *isometrically isomorphic*, symbolised $\mathcal{F} \cong \mathcal{G}$, if there is a linear bijective map $U : \mathcal{F} \to \mathcal{G}$ which preserves the inner product, i.e. for any $f, f' \in \mathcal{F}$,

$$\langle f, f'\rangle_\mathcal{F} = \langle Uf, Uf'\rangle_\mathcal{G}.$$

A consequence of the Riesz-Fréchet theorem is that it gives us a canonical isometric isomorphism $U : g \mapsto \langle \cdot, g\rangle_\mathcal{F} =: A_g$ between $\mathcal{F}$ and its continuous dual $\mathcal{F}^*$: A_g is obviously linear (bilinearity of inner products), and using the polarisation identity,

$$\begin{aligned} 2\langle Ug, Ug'\rangle_{\mathcal{F}^*} &= \|U(g)\|^2_{\mathcal{F}^*} + \|U(g')\|^2_{\mathcal{F}^*} - \|U(g - g')\|^2_{\mathcal{F}^*} \\ &= \|g\|^2_\mathcal{F} + \|g'\|^2_\mathcal{F} - \|g - g'\|^2_\mathcal{F} \\ &= 2\langle g, g'\rangle_\mathcal{F}. \end{aligned}$$

Implicitly, this means that $\mathcal{F}^*$ is a Hilbert space as well.

Another important type of mapping is the mapping P of an element in $\mathcal{F}$ onto a closed subspace $\mathcal{G} \subset \mathcal{F}$, such that $Pf \in \mathcal{G}$ is closest to f. This mapping is called the *orthogonal projection*, due to the fact that such projections yield perpendicularity in the sense that $\langle f - Pf, g\rangle_{\mathcal{F}} = 0$ for any $g \in \mathcal{G}$. Consequently, we see that $\|f\|_{\mathcal{F}}^2 = \|Pf\|_{\mathcal{F}}^2 + \|f - Pf\|_{\mathcal{F}}^2$ from the polarisation identity. The remainder $f - Pf$ belongs to the *orthogonal complement* of $\mathcal{G}$.

Definition 2.12 (Orthogonal complement)**.** Let $\mathcal{F}$ be a Hilbert space and $\mathcal{G} \subset \mathcal{F}$ be a closed subspace. The linear subspace $\mathcal{G}^\perp = \{f \,|\, \langle f, g\rangle_{\mathcal{F}} = 0, \forall g \in \mathcal{G}\}$ is called the orthogonal complement of $\mathcal{G}$ in $\mathcal{F}$.

Theorem 2.3 (Orthogonal decomposition)**.** *Let $\mathcal{F}$ be a Hilbert space and $\mathcal{G} \subset \mathcal{F}$ be a closed subspace. For every $f \in \mathcal{F}$, we can write $f = g + g^c$, where $g \in \mathcal{G}$ and $g^c \in \mathcal{G}^\perp$, and this decomposition is unique.*

Proof. Omitted—see Rudin (1987, Theorem 4.11) for a proof. ■

We can write $\mathcal{F} = \mathcal{G} \oplus \mathcal{G}^\perp$, where the $\oplus$ symbol denotes the *direct sum*, and such a decomposition is called a *tensor sum decomposition*. In infinite-dimensional Hilbert spaces, some subspaces are not closed, but all orthogonal complements are closed. In such spaces, the orthogonal complement of the orthogonal complement of $\mathcal{G}$ is the closure of $\mathcal{G}$, i.e. $(\mathcal{G}^\perp)^\perp =: \overline{\mathcal{G}}$, and we say that $\mathcal{G}$ is *dense* in $\overline{\mathcal{G}}$. Another interesting fact regarding the orthogonal complement is that $\mathcal{G} \cap \mathcal{G}^\perp = \{0\}$, since any $g \in \mathcal{G} \cap \mathcal{G}^\perp$ must be orthogonal to itself, i.e. $\langle g, g\rangle_{\mathcal{G}} = 0$ implying that $g = 0$.

The following theorem states that orthogonal decompositions are unique.

Corollary 2.3.1. *Let $\mathcal{G}$ be a subspace of a Hilbert space $\mathcal{F}$. Then, $\mathcal{G}^\perp = \{0\}$ if and only if $\mathcal{G}$ is dense in $\mathcal{F}$.*

Proof. If $\mathcal{G}^\perp = \{0\}$ then $(\mathcal{G}^\perp)^\perp = \overline{\mathcal{G}} = \mathcal{F}$. Conversely, since $\mathcal{G}$ is dense in $\mathcal{F}$, we have $\mathcal{G}^\perp = \overline{\mathcal{G}}^\perp = \mathcal{F}^\perp = \{0\}$. ■

Besides tensor sums, of importance is the concept of *tensor products*, which can be thought of as a generalisation of the outer product in Euclidean space.

Definition 2.13 (Tensor products)**.** Let $x_1 \in \mathcal{H}_1$ and $x_2 \in \mathcal{H}_2$ be two elements of two real Hilbert spaces. Then, the tensor product $x_1 \otimes x_2 : \mathcal{H}_1 \times \mathcal{H}_2 \to \mathbb{R}$, is a bilinear form defined as

$$(x_1 \otimes x_2)(y_1, y_2) = \langle x_1, y_1\rangle_{\mathcal{H}_1} \langle x_2, y_2\rangle_{\mathcal{H}_2}$$

for any $(y_1, y_2) \in \mathcal{H}_1 \times \mathcal{H}_2$.

Correspondingly, we may also define the *tensor product space*.

Definition 2.14 (Tensor product space)**.** The tensor product space $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the completion of the space

$$\mathcal{A} = \left\{ \sum_{j=1}^{J} x_{1j} \otimes x_{2j} \,\middle|\, x_{1j} \in \mathcal{H}_1, x_{2j} \in \mathcal{H}_2, J \in \mathbb{N} \right\}.$$

with respect to the norm induced by the inner product

$$\left\langle \sum_{j=1}^{J} x_{1j} \otimes x_{2j}, \sum_{k=1}^{K} y_{1k} \otimes y_{2k} \right\rangle_{\mathcal{A}} = \sum_{j=1}^{J} \sum_{k=1}^{K} \langle x_{1j}, y_{1k} \rangle_{\mathcal{H}_1} \langle x_{2j}, y_{2k} \rangle_{\mathcal{H}_2}.$$

Interestingly, the tensor product can be viewed as an operator between two Hilbert spaces. That is, for each pair of elements $(x_1, x_2) \in \mathcal{H}_1 \times \mathcal{H}_2$, we define the operator $A_{x_1,x_2} : \mathcal{H}_1 \to \mathcal{H}_2$ in the following way:

$$\begin{aligned} A_{x_1,x_2} : \mathcal{H}_1 &\to \mathcal{H}_2 \\ y_1 &\mapsto \langle x_1, y_1 \rangle_{\mathcal{H}_1} x_2. \end{aligned}$$

Incidentally, an operator defined in such a way is called a *rank one* operator. Indeed, for some $y_1 \in \mathcal{H}_1$ and $y_2 \in \mathcal{H}_2$, we have that

$$\begin{aligned} \langle A_{x_1,x_2}(y_1), y_2 \rangle_{\mathcal{H}_2} &= \left\langle \langle x_1, y_1 \rangle_{\mathcal{H}_1} x_2, y_2 \right\rangle_{\mathcal{H}_2} \\ &= \langle x_1, y_1 \rangle_{\mathcal{H}_1} \langle x_2, y_2 \rangle_{\mathcal{H}_2} \\ &= (x_1 \otimes x_2)(y_1, y_2). \end{aligned}$$

We now have three distinct interpretations of the tensor product. For $x_1, y_1 \in \mathcal{H}_1$ and $x_2, y_2 \in \mathcal{H}_2$, these are:

- **General form** (as an element in the tensor product space).

$$x_1 \otimes x_2 \in \mathcal{H}_1 \otimes \mathcal{H}_2.$$

- **Operator**.

$$\begin{aligned} x_1 \otimes x_2 : \mathcal{H}_1 &\to \mathcal{H}_2 \\ y_1 &\mapsto \langle x_1, y_1 \rangle_{\mathcal{H}_1} x_2 \end{aligned}$$

- **Bilinear form** (as per Definition 2.13).

$$\begin{aligned} x_1 \otimes x_2 : \mathcal{H}_1 \times \mathcal{H}_2 &\to \mathbb{R} \\ (y_1, y_2) &\mapsto \langle x_1, y_1 \rangle_{\mathcal{H}_1} \langle x_2, y_2 \rangle_{\mathcal{H}_2} \end{aligned}$$

Remark 2.2. As explained by Kokoszka and Reimherr (2017, §10.5, p. 227), tensors are often thought of as generalisations of matrices. For example, in Euclidean space, a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, formed by two vectors $x_1 \in \mathbb{R}^n$ and $x_2 \in \mathbb{R}^m$ via $\mathbf{A} = x_1 x_2^\top =: x_1 \otimes x_2$, can be viewed in at least three ways: 1) as a traditional matrix in the space $\mathbb{R}^n \otimes \mathbb{R}^m = \mathbb{R}^{n\times m}$; 2) as a linear transformation in Euclidean space $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ (or the reverse) by multiplying $\mathbf{A}$ from the left or right by a vector; or 3) as a bilinear mapping $\mathbf{A} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ in the form of $\mathbf{A}(y_1, y_2) = y_1^\top \mathbf{A} y_2 = y_1^\top x_1 x_2^\top y_2 = (y_1^\top x_1)(y_2^\top x_2)$, for some $y_1 \in \mathbb{R}^n$ and $y_2 \in \mathbb{R}^m$, arising often in the study of quadratic forms.

For the last part of this introductory section on functional analysis, we discuss measures on Hilbert spaces, and in particular, a probability measure. Let $\mathcal{H}$ be a real Hilbert space. As discussed earlier, we can define a metric on $\mathcal{H}$ using $D(x, x') = \|x - x'\|_{\mathcal{H}}$, where the norm on $\mathcal{H}$ is the norm induced by the inner product. A collection Σ of subsets of $\mathcal{H}$ is called a *σ-algebra* if $\emptyset \in \Sigma$, $S \in \Sigma$ implies its complement $S^c \in \Sigma$, and $S_j \in \Sigma$, $j \geq 1$ implies $\bigcup_{j=1}^{\infty} S_j \in \Sigma$. The smallest σ-algebra containing all open subsets of $\mathcal{H}$ is called the *Borel σ-algebra*, and its members the Borel sets. Denote by $\mathcal{B}(\mathcal{H})$ the Borel σ-algebra of $\mathcal{H}$.

Recall that a function $\nu : \Sigma \to [0, \infty]$ is called a *measure* if it satisfies

- **Non-negativity:** $\nu(S) \geq 0$ for all S in Σ;
- **Null empty set:** $\nu(\emptyset) = 0$; and
- **σ-additivity:** for all countable, mutually disjoint sets $\{S_i\}_{i=1}^{\infty}$,

$$\nu\left(\bigcup_{i=1}^{\infty} S_i\right) = \sum_{i=1}^{\infty} \nu(S_i).$$

A measure ν on $\big(\mathcal{H}, \mathcal{B}(\mathcal{H})\big)$ is called a *Borel measure* on $\mathcal{H}$. We shall only concern ourselves with finite Borel measures. In addition, if $\nu(\mathcal{H}) = 1$ then ν is a *(Borel) probability measure* and the measure space $\big(\mathcal{H}, \mathcal{B}(\mathcal{H}), \nu\big)$ is a *(Borel) probability space*.

Let $(\Omega, \mathcal{E}, \mathrm{P})$ be a probability space. We say that a mapping $X : \Omega \to \mathcal{H}$ is a *random element* in $\mathcal{H}$ if $X^{-1}(B) \in \mathcal{E}$ for every Borel set, i.e., X is a function such that for every $B \in \mathcal{B}(\mathcal{H})$, its preimage $X^{-1}(B) = \{\omega \in \Omega \,|\, X(\omega) \in B\}$ lies in $\mathcal{E}$. This is simply a generalisation of the definition of random variables in regular Euclidean space. From this definition, we can also properly define random functions f in a Hilbert space of functions $\mathcal{F}$. In any case, every random element X induces a probability measure on $\mathcal{H}$ defined by

$$\nu(B) = \mathrm{P}\left(X^{-1}(B)\right) = \mathrm{P}\left(\omega \in \Omega | X(\omega) \in B\right) = \mathrm{P}(X \in B).$$

The measure ν is called the *distribution* of X. The *density* p of X is a measurable function with the property that

$$\mathrm{P}(X \in B) = \int_{X^{-1}(B)} \omega \,\mathrm{dP}(\omega) = \int_B p(x) \,\mathrm{d}\nu(x).$$

Definition 2.15 (Mean vector)**.** Let ν be a Borel probability measure on a real Hilbert space $\mathcal{H}$. Supposing that a random element X of $\mathcal{H}$ is *integrable*, that is to say

$$\mathrm{E}\|X\|_{\mathcal{H}} = \int_{\mathcal{H}} \|z\|_{\mathcal{H}} \,\mathrm{d}\nu(z) < \infty,$$

then the unique element $\mu \in \mathcal{H}$ satisfying

$$\langle \mu, x \rangle = \int_{\mathcal{X}} \langle z, x \rangle_{\mathcal{X}} \,\mathrm{d}\nu(z) = \mathrm{E}\langle X, x \rangle_{\mathcal{H}}$$

for all $x \in \mathcal{H}$ is called the *mean vector*.

Definition 2.16 (Covariance operator)**.** Let ν be a Borel probability measure on a real Hilbert space $\mathcal{H}$. Suppose that a random element X of $\mathcal{H}$ is *square integrable*, i.e., $\mathrm{E}\|X\|_{\mathcal{H}}^2 < \infty$, and let μ be the mean vector of X. Then the *covariance operator* C is defined by the mapping

$$\begin{aligned} C : \mathcal{H} &\to \mathcal{H} \\ x &\mapsto \mathrm{E}\left[\langle X - \mu, x \rangle_{\mathcal{H}} (X - \mu)\right]. \end{aligned}$$

The covariance operator C is also an element of $\mathcal{H} \otimes \mathcal{H}$ that satisfies

$$\begin{aligned} \langle C, x \otimes x' \rangle_{\mathcal{H} \otimes \mathcal{H}} &= \int_{\mathcal{H}} \langle z - \mu, x \rangle_{\mathcal{H}} \langle z - \mu, x' \rangle_{\mathcal{H}} \,\mathrm{d}\nu(z) \\ &= \mathrm{E}\left[\langle X - \mu, x \rangle_{\mathcal{H}} \langle X - \mu, x' \rangle_{\mathcal{H}}\right] \end{aligned}$$

for all $x, x' \in \mathcal{H}$.

From the definition of the covariance operator, we see that it induces a symmetric, bilinear form, which we shall denote by $\mathrm{Cov} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, through

$$\begin{aligned} \langle Cx, x' \rangle_{\mathcal{H}} &= \left\langle \mathrm{E}\left[\langle X - \mu, x \rangle_{\mathcal{H}} (X - \mu)\right], x' \right\rangle_{\mathcal{H}} \\ &= \mathrm{E}\left[\langle X - \mu, x \rangle_{\mathcal{H}} \langle X - \mu, x' \rangle_{\mathcal{H}}\right] \\ &=: \mathrm{Cov}[x, x']. \end{aligned}$$

Definition 2.17 (Gaussian vectors)**.** A random element X is called *Gaussian* if $\langle X, x \rangle_{\mathcal{H}}$ has a normal distribution for all fixed $x \in \mathcal{H}$. A Gaussian vector X is characterised by its mean element $\mu \in \mathcal{H}$ and its covariance $C \in \mathcal{H} \otimes \mathcal{H}$.

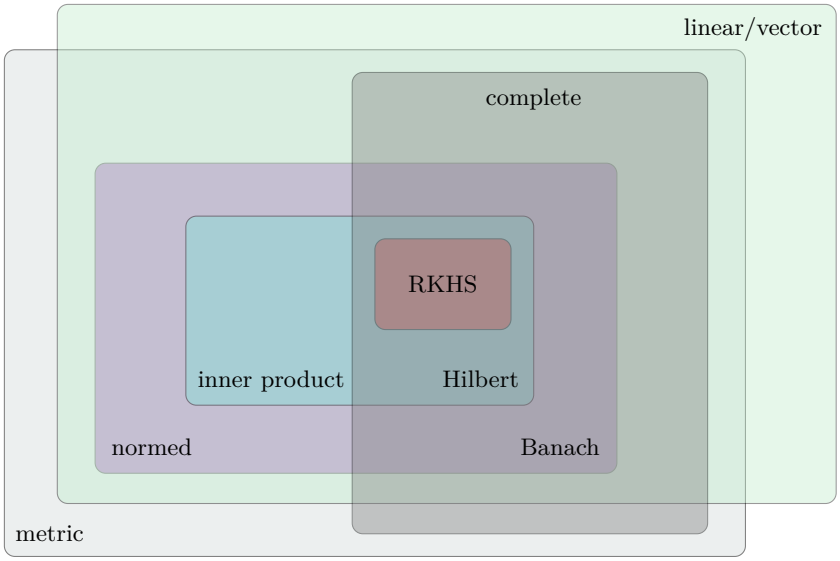


Figure 2.1: A hierarchy of vector spaces[2].

2.2 Reproducing kernel Hilbert space theory

The introductory section sets us up nicely to discuss the coveted reproducing kernel Hilbert space. This is a subset of Hilbert spaces for which its evaluation functionals are continuous (by definition, in fact). The majority of this section, apart from defining RKHSs, is an exercise in convincing ourselves that each and every RKHS of functions can be specified solely through its reproducing kernel. To begin, we consider a fundamental linear functional on a Hilbert space of functions $\mathcal{F}$, that assigns a value to $f \in \mathcal{F}$ for each $x \in \mathcal{X}$, called the *evaluation functional*.

Definition 2.18 (Evaluation functional)**.** Let $\mathcal{F}$ be a vector space of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$. For a fixed $x \in \mathcal{X}$, the functional $\delta_x : \mathcal{F} \to \mathbb{R}$ as defined by $\delta_x(f) = f(x)$ is called the (Dirac) evaluation functional at x.

It is easy to see that evaluation functionals are always linear: $\delta_x(\lambda f + g) = (\lambda f + g)(x) = \lambda f(x) + g(x) = \lambda\delta_x(f) + \delta_x(g)$ for $\lambda \in \mathbb{R}$, $f, g \in \mathcal{F}$ real functions over $\mathcal{X}$. Humble as they may seem, the entirety of the evaluation functionals over the domain $\mathcal{X}$ determines f uniquely, and thus are of great importance in understanding the space $\mathcal{F}$. Core topological properties like convergence are hinged on continuity, and it is therefore important that evaluation functionals are continuous. As it turns out, RKHSs by definition provide exactly this.

[2]Reproduced from the lecture slides of Dino Sejdinovic and Arthur Gretton entitled 'Foundations of Reproducing Kernel Hilbert Spaces: Advanced Topics in Machine Learning', 2014. URL: http://www.stats.ox.ac.uk/~sejdinov/teaching/atml14/Theory_slides2_2014.pdf.

Definition 2.19 (Reproducing kernel Hilbert space)**.** A Hilbert space $\mathcal{F}$ of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ on a non-empty set $\mathcal{X}$ is called a *reproducing kernel Hilbert space* if the evaluation functional $\delta_x : f \mapsto f(x)$ is continuous (equivalently, bounded) on $\mathcal{F}$, $\forall x \in \mathcal{X}$.

Continuity (boundedness) of evaluation functionals in an RKHS means that functions that are close in RKHS norm imply that they are also close pointwise, since $|\delta_x(f) - \delta_x(g)| = |\delta_x(f - g)| \leq M\|f - g\|_{\mathcal{F}}$ for some real $M > 0$. Note that the converse is not necessarily true. RKHSs are particularly well behaved in this respect, compared to other Hilbert spaces, and this property in particular has desirable consequences for a wide variety of applications, including nonparametric curve estimation, learning and decision theory, and many more.

While the continuity condition by definition is what makes an RKHS, it is neither easy to check this condition in practice, nor is it intuitive as to the meaning of its name. In fact, there isn't even any mention of what a reproducing kernel actually is. In order to benefit from the desirable continuity property of RKHS, we should look at this from another, more intuitive, perspective. By invoking the Riesz representation theorem, we see that for all $x \in \mathcal{X}$, there exists a unique element $h_x \in \mathcal{F}$ such that

$$f(x) = \delta_x(f) = \langle f, h_x \rangle_{\mathcal{F}}, \forall f \in \mathcal{F}$$

holds. Since h_x itself is a function in $\mathcal{F}$, it holds that for every $x' \in \mathcal{X}$ there exists a $h_{x'} \in \mathcal{F}$ such that

$$h_x(x') = \delta_{x'}(h_x) = \langle h_x, h_{x'} \rangle_{\mathcal{F}}.$$

This leads us to the definition of a *reproducing kernel* of an RKHS—the very notion that inspires its name.

Definition 2.20 (Reproducing kernels)**.** Let $\mathcal{F}$ be a Hilbert space of functions over a non-empty set $\mathcal{X}$. A symmetric, bivariate function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *kernel*, and it is a *reproducing kernel* of $\mathcal{F}$ if h satisfies

- $\forall x \in \mathcal{X}$, $h(\cdot, x) \in \mathcal{F}$; and
- $\forall x \in \mathcal{X}$, $f \in \mathcal{F}$, $\langle f, h(\cdot, x) \rangle_{\mathcal{F}} = f(x)$ (the reproducing property).

In particular, for any $x, x' \in \mathcal{X}$,

$$h(x, x') = \langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}}.$$

An important property for reproducing kernels of a RKHS is that they are positive definite functions. That is, $\forall a_1, \ldots, a_n \in \mathbb{R}$ and $\forall x_1, \ldots, x_n \in \mathcal{X}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0.$$

Proposition 2.4 (Reproducing kernels of RKHS are positive-definite)**.** *Let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel for a Hilbert space $\mathcal{F}$. Then h is a symmetric and positive definite function.*

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \langle h(\cdot, x_i), h(\cdot, x_j) \rangle_{\mathcal{F}} \\
&= \left\langle \sum_{i=1}^{n} a_i h(\cdot, x_i), \sum_{j=1}^{n} a_j h(\cdot, x_j) \right\rangle_{\mathcal{F}} \\
&= \left\| \sum_{i=1}^{n} a_i h(\cdot, x_i) \right\|_{\mathcal{F}}^2 \\
&\geq 0
\end{aligned}$$

■

Remark 2.3. In the kernel method literature, a *kernel* $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is usually defined as the inner product between inputs in feature space. That is, take $\phi : \mathcal{X} \to \mathcal{V}$, $x \mapsto \phi(x)$, where $\mathcal{V}$ is a Hilbert space. Then the kernel is defined as $h(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}$, for any $x, x' \in \mathcal{X}$. The space $\mathcal{V}$ is known as the *feature space* and the mapping ϕ the *feature map*. In many mathematical models involving feature space mappings, elucidation of the feature map and feature space is not necessary, and thus computation is made simpler by the use of kernels (known as the *kernel trick*—Hofmann et al., 2008). Note that kernels defined in this manner are positive definite, while in this thesis, we opt for a more general definition allowing kernels to not necessarily be positive. The relevance of this generality will be appreciated when we discuss reproducing kernel Kreĭn spaces in Section 2.3.

Introducing the following definition of the *kernel matrix* (also known as the *Gram matrix*) is useful at this point.

Definition 2.21 (Kernel matrix)**.** Let $\{x_1, \ldots, x_n\}$ be a sample of points, where each $x_i \in \mathcal{X}$, and h a kernel over $\mathcal{X}$. Define the *kernel matrix* $\mathbf{H}$ for h as the $n \times n$ matrix with (i, j) entries equal to $h(x_i, x_j)$.

Obviously, $\mathbf{H}$ is a positive definite matrix if the kernel that defines it is positive definite: $\mathbf{a}^\top \mathbf{H} \mathbf{a} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0$ for any choice of $a_1, \ldots, a_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$.

So far, we have seen that reproducing kernels of a RKHS are positive-definite functions, and that RKHSs are Hilbert spaces with continuous evaluation functionals, but one might wonder what exactly the relationship between a reproducing kernel and a RKHS is. We assert the following:

- For every RKHS $\mathcal{F}$ of functions over a set $\mathcal{X}$, there corresponds a unique, positive-definite reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and vice-versa. That is, a Hilbert space is a RKHS if it possesses a unique, reproducing kernel.
- For every positive-definite function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there corresponds a unique RKHS $\mathcal{F}$ that has h as its reproducing kernel.

Pictorially, the following relationships are established:

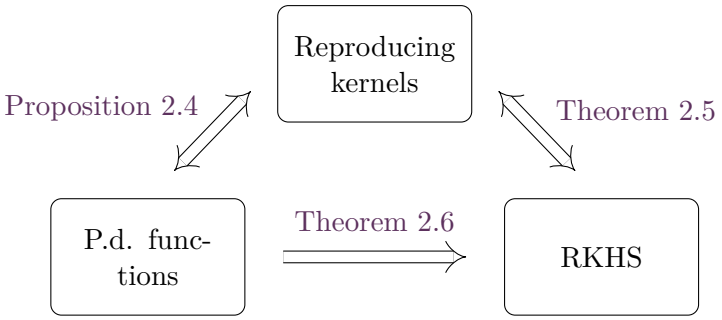


Figure 2.2: Relationships between positive definite functions, reproducing kernels, and RKHSs.

In essence, the notion of positive-definite functions and reproducing kernels of RKHSs are equivalent, and that there is a bijection between the set of positive-definite kernels and the set of RKHSs. The rest of this section is a consideration of these assertions, addressed by the two theorems that follow.

Theorem 2.5 (RKHS uniqueness)**.** *Let $\mathcal{F}$ be a Hilbert space of functions over $\mathcal{X}$. $\mathcal{F}$ is a RKHS if and only if $\mathcal{F}$ has a reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and that h is unique to $\mathcal{F}$.*

Proof. First we tackle existence, i.e. we prove that $\mathcal{F}$ is a RKHS if and only if $\mathcal{F}$ has a reproducing kernel. Suppose $\mathcal{F}$ is a Hilbert space of functions, and $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel for $\mathcal{F}$. Then, choosing $\delta = \epsilon / \|h(\cdot, x)\|_{\mathcal{F}}$, for any $f \in \mathcal{F}$ such that $\|f - g\|_{\mathcal{F}} < \delta$, we have

$$\begin{aligned} |\delta_x(f) - \delta_x(g)| &= |(f - g)(x)| \\ &= |\langle f - g, h(\cdot, x)\rangle_{\mathcal{F}}| \quad \text{(reproducing property)} \\ &\leq \|h(\cdot, x)\|_{\mathcal{F}} \, \|f - g\|_{\mathcal{F}} \quad \text{(Cauchy-Schwarz)} \\ &= \epsilon. \end{aligned}$$

Thus, the evaluation functional is (uniformly) continuous on $\mathcal{F}$, and by definition, $\mathcal{F}$ is a RKHS. Now suppose that $\mathcal{F}$ is a RKHS, and h is a kernel function over $\mathcal{X} \times \mathcal{X}$. The reproducing property of h is had by following the argument preceding Definition 2.20.

As for uniqueness, assume that the RKHS $\mathcal{F}$ has two reproducing kernels h_1 and h_2. Then, $\forall f \in \mathcal{F}$ and $\forall x \in \mathcal{X}$,

$$\langle f, h_1(\cdot, x) - h_2(\cdot, x) \rangle_{\mathcal{F}} = f(x) - f(x) = 0.$$

In particular, if we take $f = h_1(\cdot, x) - h_2(\cdot, x)$, we obtain $\|h_1(\cdot, x) - h_2(\cdot, x)\|_{\mathcal{F}}^2 = 0$. Thus, $h_1 = h_2$. ■

Theorem 2.6 (Moore-Aronszajn)**.** *If $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive-definite function then there exists a unique RKHS whose reproducing kernel is h.*

Sketch proof. Most of the details here have been omitted, except for the parts which we feel are revealing as to the properties of a RKHS. For a complete proof, see Gu (2013, Theorem 2.3). Start with the linear space

$$\mathcal{F}_0 = \left\{ f_n : \mathcal{X} \to \mathbb{R} \,\middle|\, f_n = \sum_{i=1}^{n} w_i h(\cdot, x_i), x_i \in \mathcal{X}, w_i \in \mathbb{R}, n \in \mathbb{N} \right\}$$

and endow this linear space with the following inner product:

$$\left\langle \sum_{i=1}^{n} w_i h(\cdot, x_i), \sum_{j=1}^{m} w'_j h(\cdot, x'_j) \right\rangle_{\mathcal{F}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} w_i w'_j h(x_i, x'_j).$$

It may be shown that this is indeed a valid inner product satisfying the conditions laid in Definition 2.1. At this point, the reproducing property is already had:

$$\begin{aligned} \big\langle f_n, h(\cdot, x) \big\rangle_{\mathcal{F}_0} &= \left\langle \sum_{i=1}^{n} w_i h(\cdot, x_i), h(\cdot, x) \right\rangle_{\mathcal{F}_0} \\ &= \sum_{i=1}^{n} w_i h(x, x_i) \\ &= f_n(x), \end{aligned}$$

for any $f_n \in \mathcal{F}_0$.

Let $\mathcal{F}$ be the completion of $\mathcal{F}_0$ with respect to this inner product. In other words, define $\mathcal{F}$ to be the set of functions $f : \mathcal{X} \to \mathbb{R}$ for which there exists a Cauchy sequence $\{f_n\}_{n=1}^{\infty}$ in $\mathcal{F}_0$ converging pointwise to $f \in \mathcal{F}$. The inner product for $\mathcal{F}$ is defined to be

$$\langle f, f' \rangle_{\mathcal{F}} = \lim_{n \to \infty} \langle f_n, f'_n \rangle_{\mathcal{F}_0}.$$

The sequence $\{\langle f_n, f'_n\rangle_{\mathcal{F}_0}\}_{n=1}^{\infty}$ is convergent and does not depend on the sequence chosen, but only on the limits f and f' (Berlinet and Thomas-Agnan, 2004, Lemma 5). We may check that this indeeds defines a valid inner product. The reproducing property carries over to the completion:

$$\begin{aligned}\langle f, h(\cdot, x)\rangle_{\mathcal{F}} &= \lim_{n\to\infty} \langle f_n, h(\cdot, x)\rangle_{\mathcal{F}_0} \\ &= \lim_{n\to\infty} f_n(x) \\ &= f(x).\end{aligned}$$

To prove uniqueness, let $\mathcal{G}$ be another RKHS with reproducing kernel h. $\mathcal{F}$ has to be a closed subspace of $\mathcal{G}$, since $h(\cdot, x) \in \mathcal{G}$ for all $x \in \mathcal{X}$, and because $\mathcal{G}$ is complete and contains $\mathcal{F}_0$ and hence its completion. Using the orthogonal decomposition theorem, we have $\mathcal{G} = \mathcal{F} \oplus \mathcal{F}^{\perp}$, i.e. any $g \in \mathcal{G}$ can be decomposed as $g = f + f^c$, $f \in \mathcal{F}$ and $f^c \in \mathcal{F}^{\perp}$. For each element $g \in \mathcal{G}$ we have that, for all $x \in \mathcal{X}$,

$$\begin{aligned}g(x) &= \langle g, h(\cdot, x)\rangle_{\mathcal{G}} \\ &= \left\langle f + f^c, h(\cdot, x)\right\rangle_{\mathcal{G}} \\ &= \left\langle f, h(\cdot, x)\right\rangle_{\mathcal{G}} + \cancelto{0}{\left\langle f^c, h(\cdot, x)\right\rangle_{\mathcal{G}}} \\ &= f(x)\end{aligned}$$

so therefore $g \in \mathcal{F}$ too. It must be that $\mathcal{F} \equiv \mathcal{G}$. ■

A consequence of the above proof is that we can show that any function f in a RKHS $\mathcal{F}$ with kernel h can be written in the form $f(x) = \sum_{i=1}^{n} h(x, x_i)w_i$, with some $(w_1, \ldots, w_n) \in \mathbb{R}^n$, $n \in \mathbb{N}$. More precisely, $\mathcal{F}$ is the completion of the space $\mathcal{G} = \operatorname{span}\{h(\cdot, x) \,|\, x \in \mathcal{X}\}$ endowed with the inner product as stated in Section 2.2.

2.3 Reproducing kernel Kreĭn space theory

In this section, we review elementary Kreĭn and reproducing kernel Kreĭn space theory, and comment on the similarity and differences between it and RKHSs. Kreĭn spaces are linear spaces endowed with a Hilbertian topology, characterised by an inner product which is non-positive.

Definition 2.22 (Negative and indefinite inner products)**.** Let $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ be an inner product of a vector space $\mathcal{F}$, as per Definition 2.1. An inner product is said to be *negative-definite* if for all $f \in \mathcal{F}$, $\langle f, f\rangle_{\mathcal{F}} \leq 0$. It is *indefinite* if it is neither positive- nor negative-definite.

Definition 2.23 (Kreĭn space)**.** An inner product space $\left(\mathcal{F}, \langle\cdot,\cdot\rangle_{\mathcal{F}}\right)$ is a *Krein space* if there exists two Hilbert spaces $\left(\mathcal{F}_+, \langle\cdot,\cdot\rangle_{\mathcal{F}_+}\right)$ and $\left(\mathcal{F}_-, \langle\cdot,\cdot\rangle_{\mathcal{F}_-}\right)$ spanning $\mathcal{F}$ such that

- All $f \in \mathcal{F}$ can be decomposed into $f = f_+ + f_-$, where $f_+ \in \mathcal{F}_+$ and $f_- \in \mathcal{F}_-$.
- This decomposition is orthogonal, i.e. $\mathcal{F}_+ \cup \mathcal{F}_- = \{0\}$, and $\langle f_+, f_-\rangle_{\mathcal{F}} = 0$ for all $f_+ \in \mathcal{F}_+$ and $f_- \in \mathcal{F}_-$, with the inner product on $\mathcal{F}$ defined below.
- $\forall f, f' \in \mathcal{F}$, $\langle f, f'\rangle_{\mathcal{F}} = \langle f_+, f'_+\rangle_{\mathcal{F}_+} - \langle f_-, f'_-\rangle_{\mathcal{F}_-}$.

Remark 2.4. Any Hilbert space is also a Kreĭn space, which is seen by taking $\mathcal{F}_- = \{0\}$ in the above Definition 2.23.

Let P be the projection of the Kreĭn space $\mathcal{F}$ onto $\mathcal{F}_+$, and $Q = I - P$ the projection onto $\mathcal{F}_-$, where I is the identity map. These are caleld the *fundamental projections* of $\mathcal{F}$. We shall refer to $\mathcal{F}_+$ as the *positive subspace*, and $\mathcal{F}_-$ as the *negative subspace*. These monikers stem from the fact that for all $f, f' \in \mathcal{F}$, $\langle Pf, Pf'\rangle_{\mathcal{F}_+} \geq 0$ while $\langle Qf, Qf'\rangle_{\mathcal{F}_-} \leq 0$. We introduce the notation $\ominus$ to refer to the Kreĭn space decomposition: $\mathcal{F} = \mathcal{F}_+ \ominus \mathcal{F}_-$. There is then a notion of an *associated Hilbert space*.

Definition 2.24 (Associated Hilbert space)**.** Let $\mathcal{F}$ be a Kreĭn space with decomposition into Hilbert spaces $\mathcal{F}_+$ and $\mathcal{F}_-$. Denote by $\mathcal{F}_{\mathcal{H}}$ the associated Hilbert space defined by $\mathcal{F}_{\mathcal{H}} = \mathcal{F}_+ \oplus \mathcal{F}_-$, with inner product

$$\langle f, f'\rangle_{\mathcal{F}_{\mathcal{H}}} = \langle f_+, f'_+\rangle_{\mathcal{F}_+} + \langle f_-, f'_-\rangle_{\mathcal{F}_-},$$

for all $f, f' \in \mathcal{F}$.

The associated Hilbert space can be found via the linear operator $J = P - Q$ called the *fundamental symmetry*. That is, a Kreĭn space $\mathcal{F}$ can be turned into its associated Hilbert space by using the positive-definite inner product of the associated Hilbert space as $\langle f, f'\rangle_{\mathcal{F}_{\mathcal{H}}} = \langle f, Jf'\rangle_{\mathcal{F}}$, for all $f, f' \in \mathcal{F}$. The converse is true too: starting from a Hilbert space $\mathcal{F}_{\mathcal{H}}$ and an operator J, the vector space endowed with the inner product $\langle f, f'\rangle_{\mathcal{F}} = \langle f, Jf'\rangle_{\mathcal{F}_{\mathcal{H}}}$, for all $f, f' \in \mathcal{F}$, is a Kreĭn space.

We realise that for a Kreĭn space $\mathcal{F}$, $|\langle f, f\rangle_{\mathcal{F}}| \leq \|f\|^2_{\mathcal{F}_{\mathcal{H}}}|$ for all $f \in \mathcal{F}$ (we say that $\mathcal{F}_{\mathcal{H}}$ majorises the $\mathcal{F}$), and in fact it is the smallest Hilbert space to do so. The strong topology on $\mathcal{F}$ is defined to be the topology arising from the norm of $\mathcal{F}_{\mathcal{H}}$, and this does not depend on the decomposition chosen (Ong et al., 2004). Now, we define a RKKS.

Definition 2.25 (Reproducing kernel Kreĭn space)**.** A Krein space $\mathcal{F}$ of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ on a non-empty set $\mathcal{X}$ is called a *reproducing kernel Krein space* if the evaluation functional $\delta_x : f \mapsto f(x)$ is continuous on $\mathcal{F}$, $\forall x \in \mathcal{X}$, endowed with its strong topology (i.e. the topology of its associated Hilbert space $\mathcal{F}_{\mathcal{H}}$).

One might wonder whether the uniqueness theorem (Theorem 2.5) holds for RKKS. Indeed, for every RKKS $\mathcal{F}$ of functions over a set $\mathcal{X}$, there corresponds a unique reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Lemma 2.7 (Uniqueness of kernel for RKKS)**.** *Let $\mathcal{F}$ be a RKKS of functions over a set $\mathcal{X}$, with $\mathcal{F} = \mathcal{F}_+ \ominus \mathcal{F}_-$. Then, $\mathcal{F}_+$ and $\mathcal{F}_-$ are both RKHS with kernel h_+ and h_-, and the kernel $h = h_+ - h_-$ is a unique, symmetric, reproducing kernel for $\mathcal{F}$.*

Proof. Since $\mathcal{F}$ is a RKKS, evaluation functionals are continuous on $\mathcal{F}$ with respect to topology of the associated Hilbert space $\mathcal{F}_\mathcal{H} = \mathcal{F}_+ \oplus \mathcal{F}_-$. Therefore, $\mathcal{F}_\mathcal{H}$ is a RKHS, and so too are $\mathcal{F}_+$ and $\mathcal{F}_-$ with respective kernels h_+ and h_-.

Furthermore, $h(\cdot, x) \in \mathcal{F}$ since $h_+(\cdot, x) \in \mathcal{F}_+$ and $h_-(\cdot, x) \in \mathcal{F}_-$ for some $x \in \mathcal{X}$. Then, for any $f \in \mathcal{F}$,

$$
\begin{aligned}
\langle f, h(\cdot, x)\rangle_\mathcal{F} &= \langle f, h_+(\cdot, x)\rangle_\mathcal{F} - \langle f, h_-(\cdot, x)\rangle_\mathcal{F} \\
&= \langle f_+, h_+(\cdot, x)\rangle_{\mathcal{F}_+} - \cancelto{0}{\langle f_-, h_+(\cdot, x)\rangle_{\mathcal{F}_-}} \\
&\quad - \cancelto{0}{\langle f_+, h_-(\cdot, x)\rangle_{\mathcal{F}_+}} + \langle f_-, h_-(\cdot, x)\rangle_{\mathcal{F}_-} \\
&= f_+(x) + f_-(x) \\
&= f(x)
\end{aligned}
$$

The last two lines are achieved by linearity of evaluation functionals ($\delta_x(f_+) + \delta_x(f_-) = \delta_x(f_+ + f_-)$), and the fact that $f = f_+ + f_-$ (by the Kreĭn space decomposition). We have that $h = h_+ - h_-$ is a reproducing kernel for $\mathcal{F}$. Uniqueness follows as a consequence of the non-degeneracy condition of the respective inner products for $\mathcal{F}_+$ and $\mathcal{F}_-$. ■

Remark 2.5. Unlike reproducing kernels of RKHSs, reproducing kernels of RKKSs may not be positive definite.

The analogue of the Moore-Aronszajn theorem holds partially for RKKS, up to uniqueness. That is, there is *at least* one associated RKKS with kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if and only if h can be decomposed as the difference between two positive kernels h_+ and h_- over $\mathcal{X}$, i.e. $h = h_+ - h_-$. The proof of this statement is rather involved, so is omitted in the interest of maintaining coherence to the discussion at hand. This subject has been studied by various authors; one may refer to works by Alpay (1991, Theorem 2 & Example in Section 4), and Mary (2003, Theorem 2.28).

The take-away message as we close this section is that there is no bijection, but a surjection, between the set of RKKS and the set of bivariate, symmetric functions over $\mathcal{X} \times \mathcal{X}$. In any case, Hilbertian topology applies to Kreĭn spaces via the associated Hilbert space, and in particular, RKKS provide a functional space for which evaluation

functionals are continuous. The motivation for the use of Kreĭn spaces will become clear when constructing function spaces out of (scaled) building block RKHS later in Section 2.5.

2.4 RKHS building blocks

This section describes what we refer to as the "building block" RKHSs of functions. In the context of regression modelling using I-priors, we may assume that the regression function lies in any one of these single RKHSs, although it may be more appropriate to consider function spaces built upon these RKHSs for more complex models. Construction of new function spaces from these building block RKHSs will be discussed in the next section.

2.4.1 The RKHS of constant functions

The vector space of constant functions $\mathcal{F}$ over a set $\mathcal{X}$ contains the functions $f : \mathcal{X} \to \mathbb{R}$ such that $f(x) = c_f \in \mathbb{R}$, $\forall x \in \mathcal{X}$. These functions would be useful to model an overall average, i.e. an "intercept effect". The space $\mathcal{F}$ can be equipped with a norm to form an RKHS, as shown in the following proposition.

Proposition 2.8 (RKHS of constant functions)**.** *The space $\mathcal{F}$ as described above endowed with the norm $\|f\|_{\mathcal{F}} = |c_f|$ forms an RKHS with the reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as defined, rather simply, by*

$$h(x, x') = 1,$$

known as the constant kernel.

Proof. If $\mathcal{F}$ is an RKHS with kernel h as described, then $\mathcal{F}$ is spanned by the functions $h(\cdot, x) = 1$, so it is clear that $\mathcal{F}$ consists of constant functions over $\mathcal{X}$. On the other hand, if the space $\mathcal{F}$ is equipped with the inner product $\langle f, f' \rangle_{\mathcal{F}} = c_f c_{f'}$, then the reproducing property follows, since $\langle f, h(\cdot, x) \rangle_{\mathcal{F}} = c_f = f(x)$. Hence, $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}} = |c_f|$. ■

Remark 2.6. In I-prior modelling, it is simpler to consider the intercept of a regression model as a parameter to be estimated, rather than a separate function within an RKHS of constant functions for which its posterior is to be estimated. See Section 4.2.1 in Chapter 4 for further details.

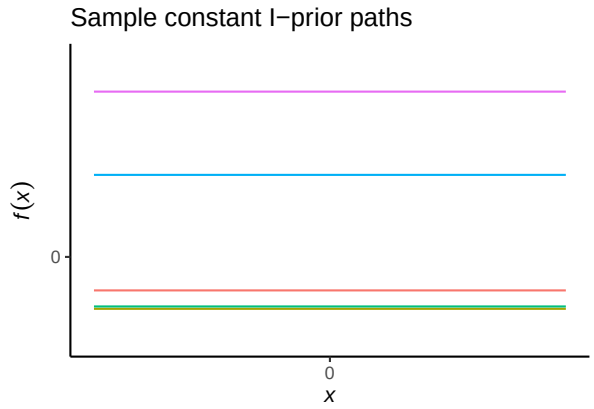


Figure 2.3: Sample I-prior paths from the RKHS of constant functions.

2.4.2 The canonical (linear) RKHS

Consider a function space $\mathcal{F}$ over $\mathcal{X}$ which consists of functions of the form $f_\beta : \mathcal{X} \to \mathbb{R}$, $f_\beta : x \mapsto \langle x, \beta \rangle_\mathcal{X}$ for some $\beta \in \mathbb{R}$. Suppose that $\mathcal{X} \equiv \mathbb{R}^p$, then $\mathcal{F}$ consists of the linear functions $f_\beta(x) = x^\top \beta$. More generally, if $\mathcal{X}$ is a Hilbert space, then its continuous dual consists of elements of the form $f_\beta = \langle \cdot, \beta \rangle_\mathcal{X}$ by the Riesz representation theorem. We can show that the continuous dual space of $\mathcal{X}$ is a RKHS which consists of these linear functions.

Proposition 2.9 (The canonical RKHS)**.** *The continuous dual space a Hilbert space $\mathcal{X}$, denoted by $\mathcal{X}^*$, is a RKHS of linear functions over $\mathcal{X}$ of the form $\langle \cdot, \beta \rangle_\mathcal{X}$, $\beta \in \mathcal{X}$. Its reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by*

$$h(x, x') = \langle x, x' \rangle_\mathcal{X}.$$

Proof. Define $f_\beta := \langle \cdot, \beta \rangle_\mathcal{X}$ for some $\beta \in \mathcal{X}$. Clearly this is linear and continuous, so $f_\beta \in \mathcal{X}^*$, and so $\mathcal{X}^*$ is a Hilbert space containing functions $f : \mathcal{X} \to \mathbb{R}$ of the form $f_\beta(x) = \langle x, \beta \rangle_\mathcal{X}$. By the Riesz representation theorem, every element of $\mathcal{X}^*$ has the form f_β. It also gives us a natural isometric isomorphism such that the following is true:

$$\langle \beta, \beta' \rangle_\mathcal{X} = \langle f_\beta, f_{\beta'} \rangle_{\mathcal{X}^*}.$$

Hence, for any $f_\beta \in \mathcal{X}^*$,

$$\begin{aligned} f_\beta(x) &= \langle x, \beta \rangle_\mathcal{X} \\ &= \langle f_x, f_\beta \rangle_{\mathcal{X}^*} \\ &= \big\langle \langle \cdot, x \rangle_\mathcal{X}, f_\beta \big\rangle_{\mathcal{X}^*}. \end{aligned}$$

Thus, $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as defined by $h(x, x') = \langle x, x' \rangle_\mathcal{X}$ is the reproducing kernel of $\mathcal{X}^*$. ■

In many other literature, the kernel $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$ is also known as the *linear kernel*. The use of the term 'canonical' is fitting not just due to the relation between a Hilbert space and its continuous dual space. Let $\phi : \mathcal{X} \to \mathcal{V}$ be the feature map from the space of covariates (inputs) to some feature space $\mathcal{V}$. Suppose both $\mathcal{X}$ and $\mathcal{V}$ are Hilbert spaces, then a kernel is defined as

$$h(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}}.$$

Taking the feature map to be $\phi(x) = \langle \cdot, x \rangle_{\mathcal{X}}$, we can prove the reproducing property to obtain $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$, which implies $\phi(x) = h(\cdot, x)$, and thus ϕ is the *canonical feature map* (Steinwart and Christmann, 2008, Lemma 4.19).

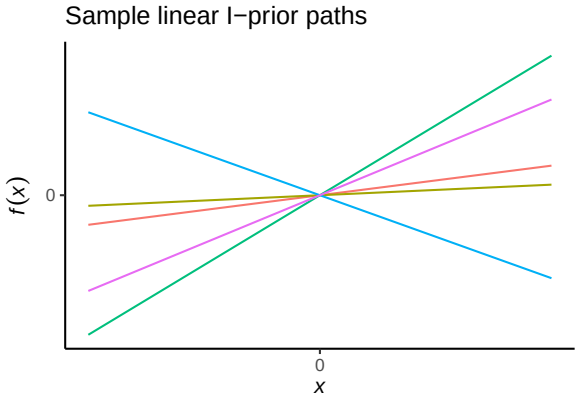


Figure 2.4: Sample paths from the canonical RKHS.

The origin of a Hilbert space may be arbitrary, in which case a centring may be appropriate. We define the centred canonical RKHS as follows.

Definition 2.26 (Centred canonical RKHS)**.** Let $\mathcal{X}$ be a Hilbert space, P be a probability measure over $\mathcal{X}$, and $\mu \in \mathcal{X}$ be the mean of a random element $X \in \mathcal{X}$. Define $(\mathcal{X} - \mu)'$, the continuous dual space of $\mathcal{X} - \mu$, to be the *centred canonical RKHS*. $(\mathcal{X} - \mu)'$ consists of the centred linear functions $f_\beta(x) = \langle x - \mu, \beta \rangle_{\mathcal{X}}$, for $\beta \in \mathcal{X}$, such that $\mathrm{E}[f_\beta(X)] = 0$. The reproducing kernel of $(\mathcal{X} - \mu)'$ is

$$h(x, x') = \langle x - \mu, x' - \mu \rangle_{\mathcal{X}}.$$

That the centred canonical RKHS consists of zero mean function, $\mathrm{E}[f_\beta(X)] = 0$, consider the following argument:

$$\begin{aligned} \mathrm{E}[f_\beta(X)] &= \mathrm{E}\langle X - \mu, \beta \rangle_{\mathcal{X}} \\ &= \mathrm{E}\langle X, \beta \rangle_{\mathcal{X}} - \langle \mu, \beta \rangle_{\mathcal{X}}, \end{aligned}$$

and since $\mathrm{E}\langle X, \beta \rangle_{\mathcal{X}} = \langle \mu, \beta \rangle_{\mathcal{X}}$ for any $\beta \in \mathcal{X}$, the results follows.

Remark 2.7. In practice, the probability measure P over $\mathcal{X}$ is unknown, so we find it useful to use the empirical distribution over $\mathcal{X}$ instead, so that $\mathcal{X}$ is centred by the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2.4.3 The fractional Brownian motion RKHS

Brownian motion, which also goes by the name Wiener process, has been an inquisitive subject in the mathematical sciences, and here, we describe a function space motivated by a generalised version of Brownian motion paths.

Suppose $B_\gamma(t)$ is a continuous-time Gaussian process on $[0, T]$, i.e. for any finite set of indices $t_1, \ldots, t_k$, where each $t_j \in [0, T]$, $\big(B_\gamma(t_1), \ldots, B_\gamma(t_k)\big)$ is a multivariate normal random variable. $B_\gamma(t)$ is said to be a *fractional Brownian motion* (fBm) if $\mathrm{E}[B_\gamma(t)] = 0$ for all $t \in [0, T]$ and

$$\mathrm{Cov}\left[B_\gamma(t), B_\gamma(s)\right] = \frac{1}{2}\big(|t|^{2\gamma} + |s|^{2\gamma} - |t-s|^{2\gamma}\big) \qquad \forall t, s \in [0, T],$$

where $\gamma \in (0, 1)$ is called the *Hurst index*, *Hurst parameter* or even *Hurst coefficient*. Introduced by Mandelbrot and Ness (1968), fBms are a generalisation of Brownian motion. The Hurst parameter plays two roles: 1) it describes the raggedness of the resultant motion, with higher values leading to smoother motion; and 2) it determines the type of process the fBm is, as past increments of $B_\gamma(t)$ are weighted by $(t-s)^{\gamma-1/2}$. When $\gamma = 1/2$ exactly, the fBm is a standard Brownian motion and its increments are independent; when $\gamma > 1/2$ (resp. $\gamma < 1/2$) its increments are positively (resp. negatively) correlated.

Now, let $\mathcal{X}$ be a Hilbert space. Schoenberg (1937, Theorem 3) has shown that, for $0 < \gamma \leq 1$, there exists a Hilbert space $\mathcal{V}$ and a function $\phi_\gamma : \mathcal{X} \to \mathcal{V}$ such that $\forall x, x' \in \mathcal{X}$,

$$\big\|\phi_\gamma(x) - \phi_\gamma(x')\big\|_{\mathcal{V}} = \|x - x'\|_{\mathcal{X}}^{\gamma}.$$

Using the polarisation identity, we find that the kernel of the RKHS with feature space $\mathcal{V}$ and feature map ϕ_γ defines a kernel function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ identical to the fBm covariance kernel.

Definition 2.27 (Fractional Brownian motion RKHS)**.** The fractional Brownian motion (fBm) RKHS $\mathcal{F}$ is the space of functions on the Hilbert space $\mathcal{X}$ possessing the reproducing kernel $h_\gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$h_\gamma(x, x') = \big\langle \phi_\gamma(x), \phi_\gamma(x') \big\rangle_{\mathcal{V}} = \frac{1}{2}\big(\|x\|_{\mathcal{X}}^{2\gamma} + \|x'\|_{\mathcal{X}}^{2\gamma} - \|x - x'\|_{\mathcal{X}}^{2\gamma}\big),$$

which depends on the Hurst coefficient $\gamma \in (0, 1)$. We shall reference this space as the fBm-γ RKHS.

Remark 2.8. When $\gamma = 1$, by the polarisation identity we get $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$, which is the (reproducing) kernel of the canonical RKHS.

From its construction, it is clear that the fBm kernel is positive definite, and thus defines an RKHS. That the fBm RKHS describes a space of functions is proved in Cohen (2002), who studied this space in depth. It is also noted in the collection of examples of Berlinet and Thomas-Agnan (2004, pp.71 & 319).

The Hurst coefficient γ controls the "smoothness" of the functions in the RKHS. We can talk about smoothness in the context of Hölder continuity of functions.

Definition 2.28 (Hölder condition)**.** A function f over a set $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is said to be *Hölder continuous* of order $0 < \gamma \leq 1$ if there exists a $C > 0$ such that $\forall x, x' \in \mathcal{X}$,

$$|f(x) - f(x')| \leq C\|x - x'\|^{\gamma}.$$

Functions in the Hölder space $\mathrm{C}^{k,\gamma}(\mathcal{X})$, where $k \geq 0$ is an integer, consists of those functions over $\mathcal{X}$ having continuous derivatives up to order k and such that the kth partial derivatives are Hölder continuous of order γ. Unlike realisations of actual fBm paths with Hurst index γ, which are well-known to be almost surely Hölder continuous of order less than γ (Embrechts and Maejima, 2002, Theorem 4.1.1), functions in its namesake RKHS are strictly smoother.

Proposition 2.10. *The fBm-γ RKHS $\mathcal{F}$ of functions over $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ are Hölder continuous of order γ.*

Proof. For some $f \in \mathcal{F}$ we have $f(x) = \langle f, h(\cdot, x) \rangle_{\mathcal{F}}$ by the reproducing property of the kernel h of $\mathcal{F}$. It follows from the Cauchy-Schwarz inequality that for any $x, x' \in \mathcal{X}$,

$$\begin{aligned} |f(x) - f(x')| &= |\langle f, h(\cdot, x) - h(\cdot, x') \rangle_{\mathcal{F}}| \\ &\leq \|f\|_{\mathcal{F}} \left\| h(\cdot, x) - h(\cdot, x') \right\|_{\mathcal{F}} \\ &= \|f\|_{\mathcal{F}} \|x - x'\|_{\mathcal{X}}^{\gamma}, \end{aligned}$$

since

$$\begin{aligned} \left\| h(\cdot, x) - h(\cdot, x') \right\|_{\mathcal{F}}^2 &= \left\| h(\cdot, x) \right\|_{\mathcal{F}}^2 + \left\| h(\cdot, x') \right\|_{\mathcal{F}}^2 - 2\langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}} \\ &= h(x, x) + h(x', x') - 2h(x, x') \\ &= \|x - x'\|_{\mathcal{X}}^{2\gamma}, \end{aligned}$$

and thus proving the proposition. ■

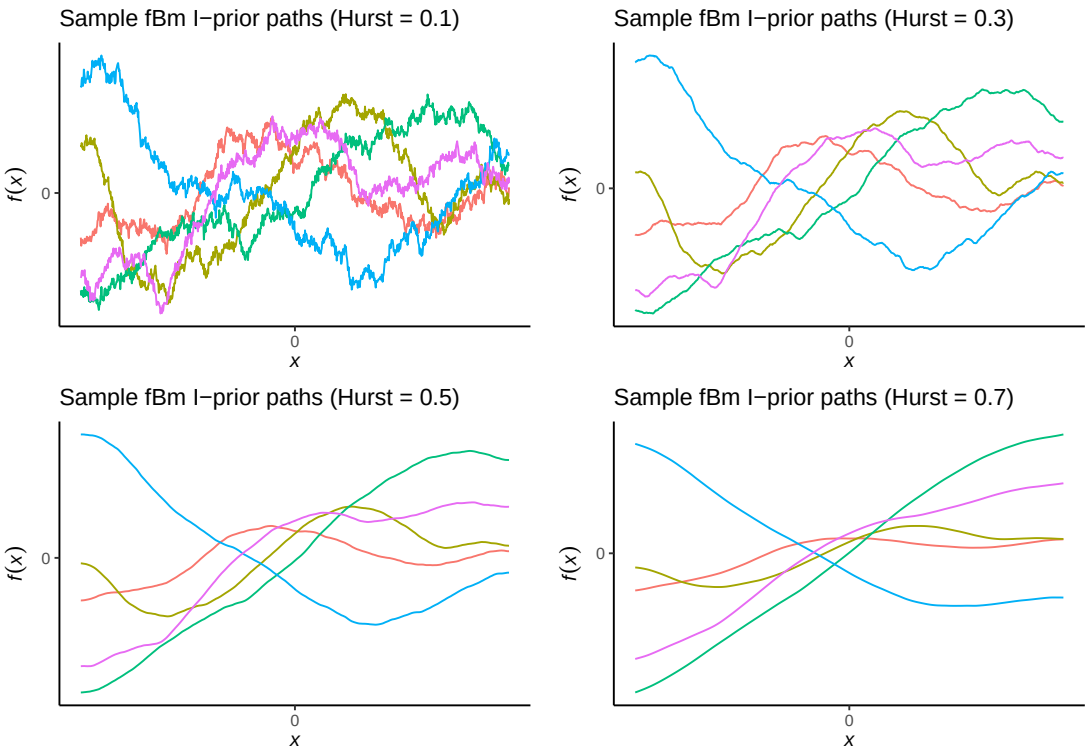


Figure 2.5: Sample I-prior paths from the fBm RKHS with varying Hurst coefficients. Note that the fBm-γ RKHS contains functions that are rougher than these I-prior paths.

The fBm-γ RKHS is spanned by the functions $h(\cdot, x)$, which means that $f(0) = 0$ for all $f \in \mathcal{F}$, which may be undesirable. We define the centred fBm RKHS as follows.

Definition 2.29 (Centred fBm RKHS)**.** Let $\mathcal{X}$ be a Hilbert space, P be a probability measure over $\mathcal{X}$, and $\mu \in \mathcal{X}$ be the mean with respect to this probability measure. The kernel $\bar{h}_\gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$\bar{h}_\gamma(x, x') = \frac{1}{2} \mathrm{E}\left[\|x - X\|_{\mathcal{X}}^{2\gamma} + \|x' - X'\|_{\mathcal{X}}^{2\gamma} - \|x - x'\|_{\mathcal{X}}^{2\gamma} - \|X - X'\|_{\mathcal{X}}^{2\gamma}\right]$$

is the reproducing kernel of the *centred* fBm-γ RKHS, which consists of functions f in the fBm-γ RKHS such that $\mathrm{E}[f(X)] = 0$. In the above definition, $X, X' \sim \mathrm{P}$ are two independent copies of a random vector $X \in \mathcal{X}$.

Remark 2.9. Again, when $\gamma = 1$, we get the reduction

$$\begin{aligned}
\bar{h}_{\gamma=1}(x, x') &= \frac{1}{2} \mathrm{E}\left[\|x - X\|_{\mathcal{X}}^2 + \|x' - X'\|_{\mathcal{X}}^2 - \|x - x'\|_{\mathcal{X}}^2 - \|X - X'\|_{\mathcal{X}}^2\right] \\
&= \frac{1}{2} \mathrm{E}\left[\langle X, X\rangle_{\mathcal{X}} + \langle X', X'\rangle_{\mathcal{X}} + 2\langle x, x'\rangle_{\mathcal{X}} - 2\langle x, X\rangle_{\mathcal{X}} - 2\langle x', X'\rangle_{\mathcal{X}}\right] \\
&= \langle \mu, \mu\rangle_{\mathcal{X}} + \langle x, x'\rangle_{\mathcal{X}} - \langle x, \mu\rangle_{\mathcal{X}} - \langle \mu, x'\rangle_{\mathcal{X}} \\
&= \langle x - \mu, x' - \mu\rangle_{\mathcal{X}},
\end{aligned}$$

which is the (reproducing) kernel of the centred canonical RKHS.

Remark 2.10. For posterity, a general centring of any (positive-definite) kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be achieved via

$$\bar{h}(x, x') = h(x, x') - \mathrm{E}[h(x, X')] - \mathrm{E}[h(X, x')] + \mathrm{E}[h(X, X')],$$

where expectations are taken for the random elements $X, X' \overset{\text{iid}}{\sim} \mathrm{P}$, a probability measure over $\mathcal{X}$. This centred kernel gives rise to the centred RKHS $\bar{\mathcal{F}}$ of centred functions $\mathrm{E}[f(X)]$, $f \in \bar{\mathcal{F}}$. As per Remark 2.7, the empirical distribution of P can be used to approximate the unknown, true P.

2.4.4 The squared exponential RKHS

The squared exponential (SE) kernel function is indeed known to be the default kernel used for Gaussian process regression in machine learning. It is a positive definite function, and hence defines an RKHS. The definition of the SE RKHS is as follows.

Definition 2.30 (Squared exponential RKHS)**.** The squared exponential (SE) RKHS $\mathcal{F}$ of functions over some set $\mathcal{X} \subseteq \mathbb{R}^p$ equipped with the 2-norm $\|\cdot\|_2$ is defined by the positive definite kernel $h_l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$h_l(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{2l^2}\right).$$

The real-valued parameter $l > 0$ is called the *lengthscale* parameter, and is a smoothing parameter for the functions in the RKHS.

It is known by many other names, including the Gaussian kernel, due to its semblance to the kernel of the Gaussian pdf. Especially in the machine learning literature, the term Gaussian radial basis functions (RBF) is used, and commonly the simpler parameterisation $\gamma = (2l^2)^{-1}$ is utilised. Duvenaud (2014) remarks that "exponentiated quadratic" is a more aptly descriptive name for this kernel.

Despite being used extensively for learning algorithms using kernels, an explicit study of the RKHS defined by the SE kernel was not done until recently by Steinwart et al. (2006). In that work, the authors describe the nature of real-valued functions in the SE RKHS by considering a a real restriction on the SE RKHS of functions over complex values. Their derivation of an orthonormal basis of such an RKHS proved the SE kernel to be the reproducing kernel for the SE RKHS.

SE kernels are known to be "universal". That is, it satisfies the following definition of universal kernels due to Micchelli et al. (2006).

Definition 2.31 (Universal kernel)**.** Let $\mathrm{C}(\mathcal{X})$ be the space of all continuous, complex-valued functions $f : \mathcal{X} \to \mathbb{C}$ equipped with the maximum norm $\|\cdot\|_\infty$, and denote $\mathcal{K}(\mathcal{X})$

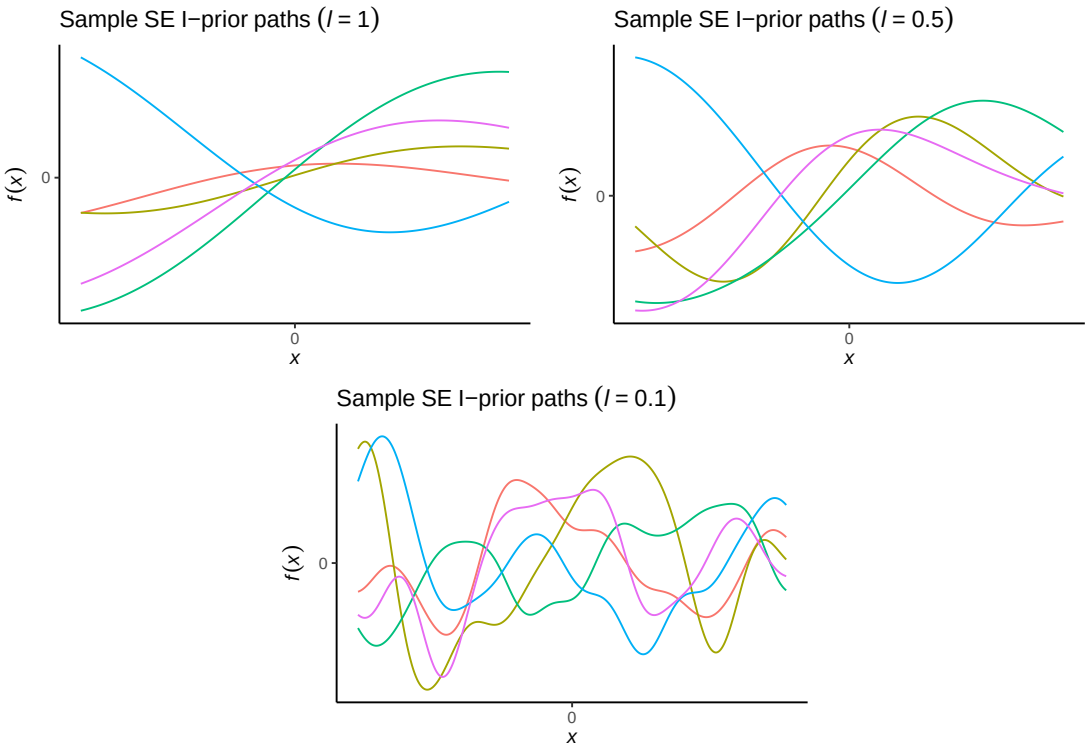


Figure 2.6: Sample paths from the SE RKHS with varying values for the lengthscale.

as the space of *kernel sections* $\overline{\text{span}}\{h(\cdot, x)|x \in \mathcal{X}\}$, where here, h is a complex-valued kernel function. A kernel h is said to be *universal* if given any compact subset $\mathcal{Z} \subset \mathcal{X}$, any positive number ϵ and any function $f \in \mathrm{C}(\mathcal{Z})$, there is a function $g \in \mathcal{K}(\mathcal{Z})$ such that $\|f - g\|_{\mathcal{Z}} \leq \epsilon$.

The consequence of this universal property vis-à-vis regression modelling is that any (continuous) regression function f may be approximated very well by a function $\hat{f}$ belonging to the SE RKHS, and these two functions can get arbitrarily close to each other in the max norm sense. This, together with the convenient computational advantages that the SE kernel brings (Raykar and Duraiswami, 2007), is a testament to the popularity of SE kernels.

In a similar manner to the two previous subsections, we may also derive the *centred* SE RKHS.

Definition 2.32 (Centred SE RKHS)**.** Let $\mathcal{X} \subseteq \mathbb{R}^p$ be equipped with the 2-norm $\|\cdot\|_2$, and let P denote the distribution over $\mathcal{X}$. Assuming integrability of $h(x, X)$, for any $x \in \mathcal{X}$ and a random element $X \in \mathcal{X}$, the *centred* squared exponential (SE) RKHS (with lengthscale l) of functions over $\mathcal{X}$ is defined by the positive definite kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$h(x, x') = e^{-\frac{\|x-x'\|_2^2}{2l^2}} - \mathrm{E}\left[e^{-\frac{\|x-X'\|_2^2}{2l^2}}\right] - \mathrm{E}\left[e^{-\frac{\|X-x'\|_2^2}{2l^2}}\right] + \mathrm{E}\left[e^{-\frac{\|X-X'\|_2^2}{2l^2}}\right],$$

where $X, X' \sim \mathrm{P}$ are two independent random elements of $\mathcal{X}$. This ensures that $\mathrm{E}[f(X)] = 0$ for any f in this RKHS.

2.4.5 The Pearson RKHS

In all of the previous RKHSs of functions, the domain $\mathcal{X}$ was taken to be some Euclidean space. The Pearson RKHS is a vector space of functions whose domain $\mathcal{X}$ is a finite set. Let P be a probability measure over the finite set $\mathcal{X}$. The Pearson RKHS is defined as follows.

Definition 2.33 (Pearson RKHS)**.** The *Pearson RKHS* is the RKHS of functions over a finite set $\mathcal{X}$ defined by the reproducing kernel

$$h(x, x') = \frac{\delta_{xx'}}{\mathrm{P}(X = x)} - 1,$$

where $X \sim \mathrm{P}$ and δ is the Kronecker delta.

The Pearson RKHS contains functions which are centred, and has the desirable property that the contribution of $[f(x)]^2$ to the squared norm of f is proportional to $\mathrm{P}(X = x)$.

Proposition 2.11. *Let $\mathcal{F}$ be the Pearson RKHS of functions over a finite set $\mathcal{X}$. Then,*

$$\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} \mid \mathrm{E}[f(X)] = 0\}$$

with

$$\|f\|_{\mathcal{F}}^2 = \mathrm{Var}[f(X)] = \sum_{x \in \mathcal{X}} \mathrm{P}(X = x)[f(x)]^2, \ \forall f \in \mathcal{F}.$$

Proof. Write $p_x = \mathrm{P}(X = x)$. The set of functions $\{h(\cdot, x) \mid x \in \mathcal{X}\}$ form a basis for $\mathcal{F}$, and thus each $f \in \mathcal{F}$ can be written as $f(x) = \sum_{x' \in \mathcal{X}} w_{x'} h(x, x')$ for some scalars $w_i \in \mathbb{R}$, $i \in \mathcal{X}$. But $\mathrm{E}[h(X, x')] = \mathrm{E}[\delta_{Xx'}]/p_{x'} - 1 = p_{x'}/p_{x'} - 1 = 0$, and thus $\mathrm{E}[f(X)] = 0$. Conversely, suppose $f : \mathcal{X} \to \mathbb{R}$ is such that $\mathrm{E}[f(X)] = 0$. Taking $w_x = p_x f(x)$, we see that

$$\begin{aligned}\sum_{x' \in \mathcal{X}} w_{x'} h(x, x') &= \frac{w_x}{p_x} - \sum_{x' \in \mathcal{X}} w_{x'} \\ &= \frac{f(x)\cancel{p_x}}{\cancel{p_x}} - \cancelto{\mathrm{E}[f(X)] = 0}{\sum_{x' \in \mathcal{X}} p_{x'} f(x')} \qquad = f(x)\end{aligned}$$

and thus $h(\cdot, x)$ spans $\mathcal{F}$ so $f \in \mathcal{F}$.

The second part is proved as follows. Noting that with the choice $w_x = p_x f(x)$ and due to the reproducing property of h for the RKHS $\mathcal{F}$, the squared norm is

$$\begin{aligned}
\langle f, f\rangle_{\mathcal{F}} &= \left\langle \sum_{x\in\mathcal{X}} w_x h(\cdot, x), \sum_{x'\in\mathcal{X}} w_{x'} h(\cdot, x') \right\rangle_{\mathcal{F}} \\
&= \sum_{x\in\mathcal{X}} \sum_{x'\in\mathcal{X}} w_x w_{x'} \left\langle h(\cdot, x), h(\cdot, x')\right\rangle_{\mathcal{F}} \\
&= \sum_{x\in\mathcal{X}} \sum_{x'\in\mathcal{X}} w_x w_{x'} h(x, x') \\
&= \sum_{x\in\mathcal{X}} w_x f(x) \\
&= \sum_{x\in\mathcal{X}} \mathrm{P}(X = x)[f(x)]^2,
\end{aligned}$$

which is also the variance of $f(X)$. ■

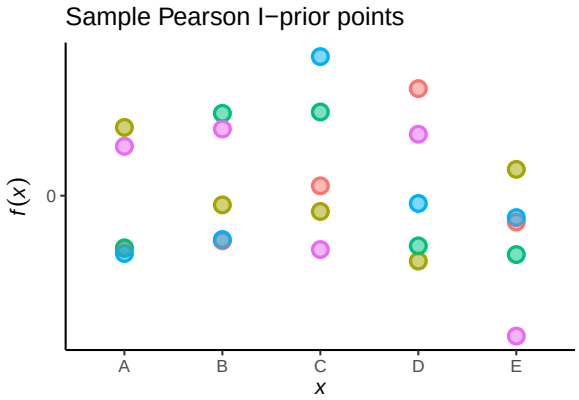


Figure 2.7: Sample I-prior "paths" from the Pearson RKHS. These are represented as points over a finite set. Similarly coloured points are from the same "path", and since they are zero-mean functions, they sum to zero.

2.5 Constructing RKKSs from existing RKHSs

The previous section outlined all of the basic RKHSs of functions that will form the building blocks when constructing more complex function spaces. We will see, at the outset, that sums of kernels are kernels and products of kernels are also kernels. This provides us a platform for constructing new function spaces from existing ones. To be more flexible in the specification of these new function spaces, we do not restrict ourselves to positive-definite kernels only, thereby necessitating us to use the theory of RKKSs.

2.5.1 Sums, products and scaling of RKHS

Sums of positive definite kernels are also positive definite kernels, and the product of positive definite kernel is a positive definite kernel. They each, in turn, are associated with a RKHS that is defined by the sum of kernels and product of kernels, respectively. The two lemmas below formalise these two facts.

Lemma 2.12 (Sum of kernels)**.** *If h_1 and h_2 are positive-definite kernels on $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively, then $h = h_1 + h_2$ is a positive-definite kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. Moreover, denote $\mathcal{F}_1$ and $\mathcal{F}_2$ the RKHS defined by h_1 and h_2 respectively. Then $\mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2$ is a RKHS defined by $h = h_1 + h_2$, where*

$$\mathcal{F}_1 \oplus \mathcal{F}_2 = \{f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R} \,|\, f = f_1 + f_2, f_1 \in \mathcal{F}_1 \text{ and } f_2 \in \mathcal{F}_2\}.$$

For all $f \in \mathcal{F}$,

$$\|f\|_{\mathcal{F}}^2 = \min_{f_1+f_2=f} \left\{ \|f_1\|_{\mathcal{F}_1}^2 + \|f_2\|_{\mathcal{F}_2}^2 \right\}.$$

Proof. That h_1+h_2 is a positive-definite kernel should be obvious, as the sum of two positive definite functions is also positive definite. For a proof of the remaining statements, see Berlinet and Thomas-Agnan (2004, Theorem 5). ■

Lemma 2.13 (Products of kernels)**.** *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two RKHS of functions over $\mathcal{X}_1$ and $\mathcal{X}_2$, with respective reproducing kernels h_1 and h_2. Then, $h = h_1 h_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. Moreover, the tensor product space $\mathcal{F}_1 \otimes \mathcal{F}_2$ is a RKHS with reproducing kernel h.*

Proof. Fix $n \in \mathbb{N}$, and let $\mathbf{H}_1$ and $\mathbf{H}_2$ be the kernel matrices for h_1 and h_2 respectively. Since these kernel matrices are symmetric and positive definite by virtue of h_1 and h_2 being symmetric and positive-definite functions, we can write $\mathbf{H}_1 = \mathbf{A}^\top \mathbf{A}$ and $\mathbf{H}_1 = \mathbf{B}^\top \mathbf{B}$ for some matrices $\mathbf{A}$ and $\mathbf{B}$: perform an (orthogonal) eigendecomposition of each of the kernel matrices, and take square roots of the eigenvalues. Let $\mathbf{H}$ be the kernel matrix for $h = h_1 h_2$. With $x_i = (x_{i1}, x_{i2})$, its (i, j) entries are

$$\begin{aligned} h(x_i, x_j) &= h_1(x_{i1}, x_{i2})\, h_2(x_{j1}, x_{j2}) \\ &= (\mathbf{A}^\top \mathbf{A})_{ij}\, (\mathbf{B}^\top \mathbf{B})_{ij} \\ &= \sum_{k=1}^{n} a_{ik} a_{jk} \sum_{l=1}^{n} b_{il} b_{jl}, \end{aligned}$$

where we have denoted b_{ij} and c_{ij} to be the (i,j)'th entries of $\mathbf{B}$ and $\mathbf{C}$ respectively Then,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} h(x_i,x_j) &= \sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j a_{ik}a_{jk}b_{il}b_{jl}\\ &= \sum_{k=1}^{n}\sum_{l=1}^{n}\left(\sum_{i=1}^{n}\lambda_i a_{ik}b_{il}\right)\left(\sum_{j=1}^{n}\lambda_j a_{jk}b_{jl}\right)\\ &= \sum_{k=1}^{n}\sum_{l=1}^{n}\left(\sum_{i=1}^{n}\lambda_i a_{ik}b_{il}\right)^2\\ &\geq 0\end{aligned}$$

Again, for the remainder of the statement in the lemma, we refer to Berlinet and Thomas-Agnan (2004, Theorem 13). ■

A familiar fact from linear algebra is realised here from Lemmas 2.12 and 2.13: 1) the addition of positive (semi-)definite matrices is a positive-definite matrix; and 2) the *Hadamard product*[3] of two positive (semi-)definite matrices is a positive (semi-)definite matrix.

The scale of a RKHS of functions $\mathcal{F}$ over a set $\mathcal{X}$ with kernel h may be arbitrary. To resolve this issue, a scale parameter $\lambda \in \mathbb{R}$ for the kernel h may be introduced, which will typically need to be estimated from the data. If h is a positive definite-kernel on $\mathcal{X}\times\mathcal{X}$, and $\lambda \geq 0$ a scalar, then this yields a scaled RKHS $\mathcal{F}_\lambda = \{\lambda f \,|\, f \in \mathcal{F}\}$ with reproducing kernel λh, where $\mathcal{F}$ is the RKHS defined by h.

Restricting λ to the positive reals is arbitrary and unnecessarily restrictive. Especially when considering sums and products of scaled RKHSs, having negative scale parameters also give additional flexibility. The resulting kernels from summation and/or multiplication with negative kernels may no longer be positive definite, and in such cases, they give rise to RKKSs instead.

Remark 2.11. Recall that a RKKS $\mathcal{F}$ of functions over $\mathcal{X}$ can be uniquely decomposed as the difference between two RKHSs $\mathcal{F}_+$ and $\mathcal{F}_-$, and its associated Hilbert space $\mathcal{F}_\mathcal{H}$ is the RKHS $\mathcal{F}_+ \oplus \mathcal{F}_-$. It is important to note that both $\mathcal{F}$ and $\mathcal{F}_\mathcal{H}$ contain identical functions over $\mathcal{X}$, but different topologies. That is to say, functions that are close with respect to the norm of $\mathcal{F}$ may not be close to each other in the norm of $\mathcal{F}_\mathcal{H}$.

[3]The Hadamard product is an element-wise multiplication of two matrices $\mathbf{A}$ and $\mathbf{B}$ of identical dimensions, denoted $\mathbf{A}\circ\mathbf{B}$. That is, $(\mathbf{A}\circ\mathbf{B})_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$.

2.5.2 The polynomial RKKS

A polynomial construction based on a particular RKHS building block is considered here. For example, using the canonical RKHS in the polynomial construction would allow us to easily add higher order effects of the covariates $x \in \mathcal{X}$. In particular, we only require a single scale parameter in polynomial kernel construction.

Definition 2.34 (The polynomial RKKS)**.** Let $\mathcal{X}$ be a Hilbert space. The kernel function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ obtained through the d-degree polynomial construction of linear kernels is

$$h_\lambda(x, x') = \big(\lambda\langle x, x'\rangle_{\mathcal{X}} + c\big)^d,$$

where $\lambda \in \mathbb{R}$ is a scale parameter for the linear kernel, and $c \in \mathbb{R}$ is a real constant called the *offset*. This kernel defines the *polynomial RKKS* of degree d.

Write

$$h_\lambda(x, x')_{\mathcal{F}} = \sum_{k=0}^{d} \frac{d!}{k!(d-k)!} c^{k-d} \lambda^k \langle x, x'\rangle_{\mathcal{X}}^k.$$

Evidently, as the name suggests, this is a polynomial involving the canonical kernel. In particular, each of the k-powered kernels (i.e., $\langle x, x'\rangle_{\mathcal{X}}^k$) defines a RKHS of their own (since these are merely products of kernels), and therefore the sum of these k-powered kernels define the polynomial RKHS.

The offset parameter influences trade-off between the higher-order versus lower-order terms in the polynomial. It is sometimes known as the bias term.

Proposition 2.14. *The polynomial RKKS $\mathcal{F}$ of real functions over $\mathcal{X}$ contains polynomial functions of the form $f(x) = \sum_{k=0}^{d} \beta_k x^k$.*

Proof. By construction, $\mathcal{F} = \mathcal{F}_\emptyset \oplus \bigoplus_{i=1}^{d} \bigotimes_{j=1}^{i} \mathcal{F}_j$, where each $\mathcal{F}_j, j \neq 0$ is the canonical RKHS, and $\mathcal{F}_\emptyset$ is the RKHS of constant functions. Each $f \in \mathcal{F}$ can therefore be written as $f = \beta_0 + \sum_{i=1}^{d} \prod_{j=1}^{i} f_j$, and $f_j(x) = b_j x$ as they are functions from the canonical RKHS, where b_j is a constant. Therefore, $f(x) = \sum_{k=0}^{d} \beta_k x^k$. ■

Remark 2.12. We may opt to use other RKHSs as the building blocks of the polynomial RKHS. In particular, using the centred canonical kernel seems natural, so that each of the functions in the constituents of the direct sum of spaces is centred. However, the polynomial RKKS itself will not be centred.

2.5.3 The ANOVA RKKS

We find it useful to begin this subsection by spending some time to elaborate on the classical analysis of variance (ANOVA) decomposition, and the associated notions of main effects and interactions. This will go a long way in understanding the thinking behind constructing an ANOVA-like RKKS of functions.

The classical ANOVA decomposition

The standard one-way ANOVA is essentially a linear regression model which allows comparison of means from two or more samples. Given sets of observations $y_j = \{y_{1j}, \ldots, y_{n_j j}\}$, $j = 1, \ldots, m$, we consider the linear model $y_{ij} = \mu_j + \epsilon_{ij}$, where ϵ_{ij} are independent, univariate, normal random variables with a common variance. This covariate-less model is used to make inferences about the *treatment means* μ_j. Often, the model is written in the *overparameterised* form by substituting $\mu_j = \mu + \tau_j$. This gives a different, arguably better, interpretability to the model: the τ_j's, referred to as the *treatment effects*, now represent the amount of deviation from the grand, *overall mean* μ. Estimating all τ_j's and μ separately is not possible because there is one degree of freedom that needs to be addressed in the model: there are $p+1$ mean parameters to estimate but only information from p means. A common fix to this identification issue is to set one of the μ_j's, say the first one μ_1, to zero, or impose the restriction $\sum_{j=1}^m \mu_j = 0$. The former treats one of the m levels as the control, while the latter treats all treatment effects symmetrically.

Now write the ANOVA model slightly differently, as $y_i = f(x_i) + \epsilon_i$, where f is defined on the discrete domain $\mathcal{X} = \{1, \ldots, m\}$, and i indexes all of the $n := \sum_{j=1}^m n_j$ observations. Here, f represents the group-level mean, returning μ_j for some $j \in \mathcal{X}$. In a similar manner, we can perform the ANOVA decomposition on f as

$$f = Af + (I - A)f = f_o + f_t,$$

where A is an averaging operator that "averages out" its argument x and returns a constant, and I is the identity operator. $f_o = Af$ is a constant function representing the *overall mean*, whereas $f_t = (I - A)f$ is a function representing the *treatment effects* τ_j. Here are two choices of A:

- $\boldsymbol{Af(x) = f(1) = \mu_1}$. This implies $f(x) = f(1) + \big(f(x) - f(1)\big)$. The overall mean μ is the group mean μ_1, which corresponds to setting the restriction $\mu_1 = 0$.
- $\boldsymbol{Af(x) = \sum_{x=1}^m f(x)/m =: \bar{\alpha}}$. This implies $f(x) = \bar{\alpha} + \big(f(x) - \bar{\alpha}\big)$. The overall mean is $\mu = \sum_{j=1}^m \alpha_j / m$, which corresponds to the restriction $\sum_{j=1}^m \mu_j = 0$.

By definition, $AAf = A^2 f = Af$, because averaging a constant returns that constant. We must have that $Af_t = A(I - A)f = Af - A^2 f = 0$. The choice of A is arbitrary, as is the choice of restriction, so long as it satisfies the condition that $Af_t = 0$.

The multiway ANOVA can be motivated in a similar fashion. Let $x = (x_1, \ldots, x_p) \in \prod_{k=1}^{p} \mathcal{X}_k$, and consider functions that map $\prod_{k=1}^{p} \mathcal{X}_k$ to $\mathbb{R}$. Let A_j be an averaging operator on $\mathcal{X}_k$ that averages the k'th component of x from the active argument list, i.e. $A_k f$ is constant on the $\mathcal{X}_k$ axis but not necessarily an overall constant function. An ANOVA decomposition of f is

$$f = \left(\prod_{k=1}^{p} (A_k + I - A_k) \right) f = \sum_{\mathcal{K} \in \mathcal{P}_p} \left(\prod_{k \in \mathcal{K}} (I - A_k) \prod_{k \notin \mathcal{K}} A_k \right) f = \sum_{\mathcal{K} \in \mathcal{P}_p} f_{\mathcal{K}} \tag{2.2}$$

where we had denoted $\mathcal{P}_p = \mathcal{P}(\{1, \ldots, p\})$ to be the power set of $\{1, \ldots, p\}$ whose cardinality is 2^p. The summands $f_{\mathcal{K}}$ will compose of the overall effect, main effects, two-way interaction terms, and so on. Each of the terms will satisfy the condition $A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p \backslash \{\}$.

Example 2.1 (Two-way ANOVA decomposition)**.** Let $p = 2$, $\mathcal{X}_1 = \{1, \ldots, m_1\}$, and $\mathcal{X}_2 = \{1, \ldots, m_2\}$. The power set $\mathcal{P}_2$ is $\big\{\{\}, \{1\}, \{2\}, \{1, 2\}\big\}$. The ANOVA decomposition of f (with indices derived trivially from the power set) is

$$f = f_\emptyset + f_1 + f_2 + f_{12}.$$

Here are two choices for the averaging operator A_k analogous to the previous illustration in the one-way ANOVA.

- Let $A_1 f(x) = f(1, x_2)$ and $A_2 f(x) = f(x_1, 1)$. Then,

$$\begin{aligned} f_\emptyset(x) &= A_1 A_2 f && = f(1,1) \\ f_1(x) &= (I - A_1) A_2 f && = f(x_1, 1) - f(1,1) \\ f_2(x) &= A_1 (I - A2) f && = f(1, x_2) - f(1,1) \\ f_{12}(x) &= (I - A_1)(I - A2) f && = f(x_1, x_2) - f(x_1, 1) - f(1, x_2) + f(1,1). \end{aligned}$$

- Let $A_k f(x) = \sum_{x_k=1}^{m_k} f(x_1, x_2)/m_k, k = 1, 2$. Then,

$$\begin{aligned} f_\emptyset(x) &= A_1 A_2 f && = f_{\cdot\cdot} \\ f_1(x) &= (I - A_1) A_2 f && = f_{x_1 \cdot} - f_{\cdot\cdot} \\ f_2(x) &= A_1 (I - A_2) f && = f_{\cdot x_2} - f_{\cdot\cdot} \\ f_{12}(x) &= (I - A_1)(I - A_2) f && = f - f_{x_1 \cdot} - f_{\cdot x_2} + f_{\cdot\cdot}, \end{aligned}$$

  where $f_{\cdot\cdot} = \sum_{x_1, x_2} f(x_1, x_2)/m_1 m_2$, $f_{x_1 \cdot} = \sum_{x_2} f(x_1, x_2)/m_2$, and $f_{\cdot x_1} = \sum_{x_1} f(x_1, x_2)/m_1$.

It is also easy to convince ourselves that $A_1 f_1 = A_2 f_2 = A_1 f_{12} = A_2 f_{12} = 0$ in either choice of the averaging operator A_k.

Functional ANOVA decomposition

Let us now extend the ANOVA decomposition idea to a general function $f : \mathcal{X} \to \mathbb{R}$ in some vector space $\mathcal{F}$. We shall jump straight into the multiway ANOVA analogue for functional decomposition, and to that end, consider $x = (x_1, \ldots, x_p) \in \prod_{k=1}^{p} \mathcal{X}_k =: \mathcal{X}$ a measurable space, where each of the spaces $\mathcal{X}_k$ has measure ν_k, and $\nu = \nu_1 \times \cdots \times \nu_p$ is the product measure on $\mathcal{X}$. In the following, denote by $\mathcal{F}_k$ the vector space of functions over the set $\mathcal{X}_k$, $k = 1, \ldots, p$, and $\mathcal{F}_\emptyset$ the vector space of constant functions.

As $\mathcal{X}$ need not necessarily be a collection of finite sets, we need to figure out a suitable linear operator that performs an "averaging" of some sort. Consider the linear operator $A_k : \mathcal{F} \to \mathcal{F}_{-k}$, where $\mathcal{F}_{-k}$ is a vector space of functions for which the kth component is constant over $\mathcal{X}$, defined by

$$A_k f(x) = \int_{\mathcal{X}_k} f(x_1, \ldots, x_p) \, \mathrm{d}\nu_k(x_k). \tag{2.3}$$

Thus, for the one-way ANOVA ($p = 1$), we get

$$f(x) = \overbrace{\int_{\mathcal{X}} f(x) \, \mathrm{d}\nu(x)}^{f_\emptyset(x)} + \overbrace{\left(f - \int_{\mathcal{X}} f(x) \, \mathrm{d}\nu(x) \right)}^{f_1(x)} \tag{2.4}$$

and for the two-way ANOVA ($p = 2$), we have $f = f_\emptyset + f_1 + f_2 + f_{12}$, with

$$\begin{aligned}
f_\emptyset(x) &= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu_1(x_1) \, \mathrm{d}\nu_2(x_2) \\
f_1(x) &= \int_{\mathcal{X}_2} \left(f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2) \, \mathrm{d}\nu_1(x_1) \right) \mathrm{d}\nu_2(x_2) \\
f_2(x) &= \int_{\mathcal{X}_1} \left(f(x_1, x_2) - \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu_2(x_2) \right) \mathrm{d}\nu_1(x_1) \\
f_{12}(x) &= f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2) \, \mathrm{d}\nu_1(x_1) - \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu_2(x_2) \\
&\quad + \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu_1(x_1) \, \mathrm{d}\nu_2(x_2).
\end{aligned}$$

The averaging operator A_k defined in (2.3) generalises the concept of the previous subsection's averaging operator. We must then also have, as before, that $A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p \backslash \{\}$. For the one-way functional ANOVA decomposition in (2.4), it must be that f_1 is a zero-mean function. As for the two-way ANOVA, it is the case that $\int_{\mathcal{X}_k} f_{\mathcal{K}}(x_1, x_2) \, \mathrm{d}\nu_k(x_k) = 0, k = 1, 2$, and $\mathcal{K} \in \big\{\{1\}, \{2\}, \{1, 2\}\big\}$ (Durrande et al., 2013).

This is highly suggestive as to what the ANOVA decomposition of the space $\mathcal{F}$ should look like in general. Starting with $p = 1$, any $f \in \mathcal{F}$ can be decomposed as a sum of a constant plus a zero mean function, so we have the decomposition of the vector space $\mathcal{F} = \mathcal{F}_\emptyset \oplus \bar{\mathcal{F}}_1$, where a bar over $\mathcal{F}_k$, $k = 1, \ldots, p$ will be used to denote the vector space of zero mean functions over $\mathcal{X}_k$. For $p \geq 2$ we can argue something similar. Take the vector space space $\mathcal{F}$ of functions over $\prod_{k=1}^p \mathcal{X}_k$ to be the tensor product space $\mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_p$ whose elements are identified as being tensor product functions $f_1 \otimes \cdots \otimes f_p$, where each $f_k : \mathcal{X}_k \to \mathbb{R}$ belongs to $\mathcal{F}_k$. This is constructed by repeatedly taking the completion of linear combinations of the tensor product $f_k \otimes f_j$, $k, j \in \{1, \ldots, p\}$ as per Definition 2.14. Considered individually, each $\mathcal{F}_k$ can then be decomposed as $\mathcal{F}_k = \mathcal{F}_{\emptyset k} \oplus \bar{\mathcal{F}}_k$, where $\mathcal{F}_{\emptyset k}$ is the space of functions constant along the k'th axis. Expanding out under the distributivity rule of tensor products and rearranging slightly, we obtain

$$\begin{aligned} \mathcal{F} &= \big(\mathcal{F}_{\emptyset 1} \oplus \bar{\mathcal{F}}_1\big) \otimes \cdots \otimes \big(\mathcal{F}_{\emptyset p} \oplus \bar{\mathcal{F}}_p\big) \\ &= \mathcal{F}_\emptyset \;\oplus\; \bigoplus_{j=1}^{p} \Big(\bigotimes_{i \neq j} \mathcal{F}_{\emptyset i} \otimes \bar{\mathcal{F}}_j \Big) \;\oplus\; \bigoplus_{\substack{j,k=1 \\ j<k}}^{p} \Big(\bigotimes_{i \neq j,k} \mathcal{F}_{\emptyset i} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k \Big) \\ &\quad \oplus \;\cdots\; \oplus \; \Big(\bar{\mathcal{F}}_1 \otimes \cdots \otimes \bar{\mathcal{F}}_p \Big), \end{aligned} \tag{2.5}$$

where '$\bigoplus$' and '$\bigotimes$' represent the summation and product operator for direct/tensor sums and products, respectively. To clarify,

- $\mathcal{F}_\emptyset$ is the space of constant functions $f_\emptyset : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$;
- $\Big(\bigotimes_{i \neq j} \mathcal{F}_{\emptyset i} \otimes \bar{\mathcal{F}}_j \Big)$ is the space of functions that are constant on all coordinates except the j'th coordinate of x, and the functions are centred on the j'th coordinate;
- $\Big(\bigotimes_{i \neq j,k} \mathcal{F}_{\emptyset i} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k \Big)$ is the space of functions that are constant on all coordinates except the jth and kth coordinate of x, and the functions are centred on these two coordinates;
- $\bar{\mathcal{F}}_1 \otimes \cdots \otimes \bar{\mathcal{F}}_p$ is the space of zero-mean functions $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$;

and so on for for the rest of the spaces in the summand, of which there are 2^p members all together. Therefore, given an arbitrary function $f \in \mathcal{F}$, the projection of f onto the above respective spaces in (2.5) leads to the *functional ANOVA representation*

$$f(x) = \alpha + \sum_{j=1}^{p} f_j(x_j) + \sum_{\substack{j,k=1 \\ j<k}}^{p} f_{jk}(x_j, x_k) + \cdots + f_{1 \cdots p}(x), \tag{2.6}$$

where α is the grand intercept (a constant).

Definition 2.35 (Functional ANOVA representation)**.** Let $\mathcal{P}_p = \mathcal{P}(\{1,\ldots,p\})$, the power set of $\{1,\ldots,p\}$. For any function $f \in \mathcal{F} \equiv \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_p$, with each $\mathcal{F}_k$ a space of functions over $\mathcal{X}_k$, $k = 1,\ldots,p$, the formula for f in (2.6) is known as the *functional ANOVA representation* of f if $\forall k \in \mathcal{K} \in \mathcal{P}_p \backslash \{\}$,

$$A_k f_{\mathcal{K}}(x) = \int_{\mathcal{X}_k} f_{\mathcal{K}}(x)\, \mathrm{d}\nu_k(x_k) = 0. \tag{2.7}$$

In other words, the integral of $f_{\mathcal{K}}$ with respect to any of the variables indexed by the elements in $\mathcal{K}$, is zero.

For the constant term, main effects, and two-way interaction terms, the familiar classical expressions are obtained:

$$\begin{aligned} f_\emptyset &= \int f \,\mathrm{d}\nu; \\ f_j &= \int f \prod_{i\neq j} \mathrm{d}\nu_i - f_\emptyset; \\ f_{jk} &= \int f \prod_{i\neq j,k} \mathrm{d}\nu_i - f_j - f_k - f_\emptyset. \end{aligned}$$

The ANOVA kernel

At last, we come to the section of deriving the ANOVA RKKS, and, rest assured, the preceding long build-up will prove to not be in vain. The main idea is to construct a RKKS such that the functions that lie in them will have the ANOVA representation in (2.6). The bulk of the work has been done, and in fact we know exactly how this ANOVA RKKS should be structured—it is the space as specified in (2.5). The ANOVA RKKS will be constructed by a similar manipulation of the individual kernels representing the RKHS building blocks.

Definition 2.36 (The ANOVA RKKS)**.** For $k = 1,\ldots,p$, let $\mathcal{F}_k$ be centred RKHSs of functions over the set $\mathcal{X}_k$ with kernel $h_k : \mathcal{X}_k \times \mathcal{X}_k \to \mathbb{R}$. Let $\lambda_k, k = 1,\ldots,p$ be real-valued scale parameters. The ANOVA RKKS of functions $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$ is specified by the ANOVA kernel, defined by

$$h_\lambda(x,x') = \prod_{k=1}^{p} \big(1 + \lambda_k h_k(x_k, x'_k)\big). \tag{2.8}$$

It is interesting to note that an ANOVA RKKS is constructed very simply through multiplication of univariate kernels. Expanding out equations (2.8), we see that it is in

fact a sum of products of kernels with increasing orders of interaction:

$$h_\lambda(x,x') = 1 + \sum_{j=1}^{p} \lambda_j h_j(x_j, x'_j) + \sum_{\substack{j,k=1 \\ j<k}}^{p} \lambda_j \lambda_k h_j(x_j, x'_j) h_k(x_k, x'_k) + \cdots + \prod_{j=1}^{p} \lambda_j h_j(x_j, x'_j).$$

It is now clear from this expansion that the ANOVA RKKS yields functions that resemble those with the ANOVA representation in (2.6): the mean value of the function stems from the '1', i.e. it lies in a RKHS of constant functions; the main effects are represented by the sum of the individual kernels; the two-way interaction terms are represented by the second-order kernel interactions; and so on.

Example 2.2. Consider two RKKSs $\mathcal{F}_k$ with kernel $\lambda_k h_k$, $k = 1, 2$. The ANOVA kernel defining the ANOVA RKKS $\mathcal{F}$ is

$$h_\lambda\big((x_1,x_2),(x'_1,x'_2)\big) = 1 + \lambda_1 h_1(x_1,x'_1) + \lambda_2 h_2(x_2,x'_2) + \lambda_1\lambda_2 h_1(x_1,x'_1) h_2(x_2,x'_2).$$

Suppose that $\mathcal{F}_1$ and $\mathcal{F}_2$ are the centred canonical RKKS of functions over $\mathbb{R}$. Then, functions in $\mathcal{F} = \mathcal{F}_0 \oplus \mathcal{F}_1 \oplus \mathcal{F}_2 \oplus (\mathcal{F}_1 \otimes \mathcal{F}_2)$ are of the form

$$f(x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

As a remark, not all of the components of the ANOVA RKKS need to be included in the construction. The selective exclusion of certain interactions characterises many interesting statistical models. Excluding certain terms of the ANOVA RKKS is equivalent to setting the scale parameter for those relevant components to be zero, i.e. they play no role in the decomposition of the function. With this in mind, the ANOVA RKKS then gives us an objective way of model-building, from linear regression, to multilevel models, longitudinal models, and so on.

Finally, we note that the functional ANOVA decomposition of a RKKS is orthogonal. Without loss of generality, assume that all scale parameters are positive. For $p = 1$, we have that $\mathcal{F} = \mathcal{F}_\emptyset \oplus \mathcal{F}_1$, where each of $\mathcal{F}_\emptyset$ and $\mathcal{F}_1$ is a RKHS. From Lemma 2.12, the squared norm of $\mathcal{F}$ is given by

$$\|f\|^2_{\mathcal{F}} = \|f_\emptyset\|^2_{\mathcal{F}_\emptyset} + \|f_1\|^2_{\mathcal{F}_1},$$

if the decomposition $f = f_\emptyset + f_1$ is minimal. Hence, the decomposition is orthogonal. An inductive argument can be used to extend and generalise to any $p \geq 2$.

2.6 Summary

The review of functional analysis allows us to describe the theory of RKHSs and RKKSs, which are of interest to us because the topology endowed on such spaces gives appreciable assurances—in particular, all evaluation functionals are continuous in these spaces. Moreover, RKHSs and RKKSs can be specified completely through kernel functions, with new and complex function spaces built simply by manipulation of these kernel functions. Of particular importance is the ANOVA functional decomposition, for which we realise provides an objective way of constructing various function spaces for regression and modelling. Such models will be described later on in detail in Chapter 4.

An annotated collection of bibliographical references used for this chapter is as follows.

- **Functional analysis**. On the introductory material relating to functional analysis in Section 2.1, the lecture notes by Sejdinovic and Gretton (2012) is recommended, and forms the basis for most of our material. Additionally, Kokoszka and Reimherr (2017), Rudin (1987), and Yamamoto (2012) provides a complementary reading.
- **RKHS theory**. There are certainly no shortages of introductory texts relating to the theory of RKHS: Steinwart and Christmann (2008), Berlinet and Thomas-Agnan (2004), and Gu (2013) to name a few. The concise sketch proof for the Moore-Aronszajn theorem was mostly inspired by Hein and Bousquet (2004, Theorem 4).
- **Kreĭn space and RKKS theory**. The innovation of indefinite inner product spaces perhaps started in mathematical physics literature, for which the theory of special relativity depends. Four-dimensional space-time is an often cited example. In any case, we referred to mainly Ong et al. (2004), which gives an overview in the context of learning using indefinite kernels. Alpay (1991) and Zafeiriou (2012) were also useful for understanding the fundamental concepts of RKKSs.
- **RKHS building blocks**. The main building block RKHSs, i.e. the canonical RKHS, the fBm RKHS and the Pearson RKHS, are described in the manuscript of Bergsma (2018).
- **ANOVA and functional ANOVA**. Classical ANOVA is pretty much existent in every fundamental statistical textbook. These texts have extremely well written introductions to this very important concept: Casella and R. L. Berger (2002, Ch. 11), Dean and Voss (1999, Ch. 3). On the relation between classical ANOVA and functional ANOVA decomposition, Gu (2013) offers novel insights. There is diverse literature concerning functional ANOVA, namely from the fields of statistical learning (e.g. Wahba, 1990), applied mathematics (e.g. F. Y. Kuo et al., 2010), and sensitivity analysis (e.g. Durrande et al., 2013; Sobol, 2001).

Chapter 3

Fisher information and the I-prior

We are interested in calculating the Fisher information for our unknown regression function f (the parameter to be estimated) in (1.1), subject to (1.2) and $f \in \mathcal{F}$, a RKKS. Unlike in the traditional case, $\mathcal{F}$ may be infinite dimensional, and hence care must be taken when computing derivatives with respect to f when this is the case. If $\mathcal{F}$ possesses an orthonormal basis, then one could define the derivative of the functional $\rho : \mathcal{F} \to \mathbb{R}$ component-wise with respect to the orthonormal basis, as in the finite dimensional case. This is analogous to the usual concept of partial derivatives.

However, the notion of partial derivatives does not generalise to arbitrary topological vector spaces for two reasons. Firstly, general spaces may not have an orthonormal basis (Tapia, 1971, §5, pp. 76). Secondly, component-wise derivatives, which are in essence limits taken component-wise using the usual definition of derivatives, may not coincide with the overall limit taken with respect to the topology of the vector space. For these reasons, there is a need to consider the rigorous concepts of differentiation suitable for infinite-dimensional vector spaces provided by Fréchet and Gâteaux derivatives. These concepts are introduced in Section 3.2, prior to the actual derivation of the Fisher information of the regression function in Section 3.3.

In the remaining sections, we discuss the notion of prior distributions for regression functions, and how one might assign a suitable prior. In our case, we choose an objective prior following (Jaynes, 1957a, 1957b, 2003): in the absence of any prior knowledge, a prior distribution which maximises entropy should be used. As it turns out, the entropy maximising prior for f is Gaussian with mean chosen a priori and covariance kernel proportional to the Fisher information. We call such a distribution on f an *I-prior distribution* for f. The I-prior has a simple, intuitive appeal: much information about f corresponds to a larger prior covariance, and thus less influence of the prior mean, and more of the data, in informing the posterior, and vice versa.

3.1 The traditional Fisher information

It was Fisher (1922) who introduced the method of maximum likelihood (ML) as an objective way of conducting statistical inference. This method of inference is distinguished from the Bayesian school of thought in that only the data may inform deductive reasoning, but not any sort of prior probabilities. Towards the later stages of his career[1], his work reflected the view that the likelihood is to be more than simply a device to obtain parameter estimates; it is also a vessel that carries uncertainty about estimation. In this light and in the absence of the possibility of making probabilistic statements, one should look to the likelihood in order to make rational conclusions about an inference problem. Specifically, we may ask two things of the likelihood function: where is the maxima and what does the graph around the maxima look like? The first of these two problems is maximum likelihood estimation, while the second concerns the Fisher information.

In simple terms, the Fisher information measures the amount of information that an observable random variable Y carries about an unknown parameter θ of the statistical model that models Y. To make this concrete, Y has the density function $p(\cdot|\theta)$ which depends on θ. Write the log-likelihood function of θ as $L(\theta) = \log p(Y|\theta)$, and the gradient function of the log-likelihood (the *score function*) with respect to θ as $S(\theta) = \partial L(\theta)/\partial\theta$. The *Fisher information* about the parameter θ is defined to be expectation of the second moment of the score function,

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log p(Y|\theta)\right)^2\right].$$

Here, expectation is taken with respect to the random variable Y under its true distribution. Under certain regularity conditions, it can be shown that $\mathrm{E}[S(\theta)] = 0$, and thus the Fisher information is in fact the variance of the score function, since $\mathrm{Var}[S(\theta)] = \mathrm{E}[S(\theta)^2] - \mathrm{E}^2[S(\theta)]$. Further, if $\log p(Y|\theta)$ is twice differentiable with respect to θ, then it can be shown that under certain regularity conditions,

$$\mathcal{I}(\theta) = \mathrm{E}\left[-\frac{\partial^2}{\partial\theta^2}\log p(Y|\theta)\right].$$

Many textbooks provides a proof of this fact—see, for example, Wasserman (2004, §9.7).

From the last equation above, we see that the Fisher information is related to the curvature or concavity of the graph of the log-likelihood function, averaged over the random variable Y. The curvature, defined as the second derivative on the graph[2] of a function, measures how quickly the function changes with changes in its input values.

[1] The introductory chapter of Pawitan (2001) and the citations therein give a delightful account of the evolution of the Fisherian view regarding statistical inference.

[2] Formally, the graph of a function g is the set of all ordered pairs $(x, g(x))$.

This then gives an intuition regarding the uncertainty surrounding θ at its maximal value; high Fisher information is indicative of a sharp peak at the maxima and therefore small variance, while low Fisher information is indicative of a shallow maxima for which many θ share similar log-likelihood values.

3.2 Fisher information in Hilbert space

We extend the idea beyond thinking about parameters as merely numbers in the usual sense, to abstract objects in Hilbert spaces. This generalisation allows us to extend the concept of Fisher information to regression functions in RKKSs later. The score and Fisher information is derived in a familiar manner, but extra care is required when taking derivatives with respect to elements in Hilbert spaces. We discuss a generalisation of the concept of differentiability from real-valued functions of a single, real variable, as is common in calculus, to functions between Hilbert spaces.

Definition 3.1 (Fréchet derivative)**.** Let $\mathcal{V}$ and $\mathcal{W}$ be two Hilbert spaces, and $\mathcal{U} \subseteq \mathcal{V}$ be an open subset. A function $\rho : \mathcal{U} \to \mathcal{W}$ is called *Fréchet differentiable* at $x \in \mathcal{U}$ if there exists a bounded, linear operator $T : \mathcal{V} \to \mathcal{W}$ such that

$$\lim_{v \to 0} \frac{\left\| \rho(x+v) - \rho(x) - Tv \right\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} = 0$$

If this relation holds, then the operator T is unique, and we write $\mathrm{d}\rho(x) := T$ and call it the *Fréchet derivative* or *Fréchet differential* of ρ at x. If ρ is differentiable at every point $\mathcal{U}$, then ρ is said to be *(Fréchet) differentiable* on $\mathcal{U}$.

Remark 3.1. Since $\mathrm{d}\rho(x)$ is a bounded, linear operator, by Lemma 2.1 (p. 41), it is also continuous.

Remark 3.2. While the Fréchet derivative is most commonly defined as the derivative of functions between Banach spaces, the definition itself also applies to Hilbert spaces, since complete inner product spaces are also complete normed spaces. Since our main focus are RKHSs and RKKSs, i.e. spaces with Hilbertian topology (recall that RKKSs are endowed with the topology of its associated Hilbert space), it is beneficial to present the material using Hilbert spaces. We appeal to the works of Balakrishnan (1981, Definition 3.6.5) and Bouboulis and Theodoridis (2011, §6) in this regard.

Remark 3.3. The use of the open subset $\mathcal{U}$ in the definition above for the domain of the function ρ is so that the notion of ρ being differentiable is possible even without having it defined on the entire space $\mathcal{V}$.

The intuition here is similar to that of regular differentiability, in that the linear operator T well approximates the change in ρ at x (the numerator), relative to the change in x (the denominator)—the fact that the limit exists and is zero, it must mean that the numerator converges faster to zero than the denominator does. In Landau notation, we have the familiar expression $\rho(x+v) = \rho(v) + \mathrm{d}\rho(x)(v) + o(v)$, that is, the derivative of ρ at x gives the best linear approximation to ρ near x. Note that the limit in the definition is meant in the usual sense of convergence of functions with respect to the norms of $\mathcal{V}$ and $\mathcal{W}$.

For the avoidance of doubt, $\mathrm{d}\rho(x)$ is not a vector in $\mathcal{W}$, but is an element of the set of bounded, linear operators from $\mathcal{V}$ to $\mathcal{W}$, denoted $\mathrm{L}(\mathcal{V};\mathcal{W})$. That is, if $\rho : \mathcal{U} \to \mathcal{W}$ is a differentiable function at all points in $\mathcal{U} \subseteq \mathcal{V}$, then its derivative is a linear map

$$\begin{aligned} \mathrm{d}\rho : \mathcal{U} &\to \mathrm{L}(\mathcal{V};\mathcal{W}) \\ x &\mapsto \mathrm{d}\rho(x). \end{aligned}$$

It follows that this function may also have a derivative, which by definition will be a linear map as well. This is the *second Fréchet derivative* of ρ, defined by

$$\begin{aligned} \mathrm{d}^2\rho : \mathcal{U} &\to \mathrm{L}\big(\mathcal{V};\mathrm{L}(\mathcal{V};\mathcal{W})\big) \\ x &\mapsto \mathrm{d}^2\rho(x). \end{aligned}$$

To make sense of the space on the right-hand side, consider the following argument.

- Take any $\phi(\cdot) \in \mathrm{L}\big(\mathcal{V};\mathrm{L}(\mathcal{V};\mathcal{W})\big)$. For all $v \in \mathcal{V}$, $\phi(v) \in \mathrm{L}(\mathcal{V};\mathcal{W})$, and $\phi(v)$ is linear in v.
- Since $\phi(v) \in \mathrm{L}(\mathcal{V};\mathcal{W})$, it is itself a linear operator taking elements from $\mathcal{V}$ to $\mathcal{W}$. We can write it as $\phi(v)(\cdot)$ for clarity.
- So, for any $v' \in \mathcal{V}$, $\phi(v)(v') \in \mathcal{W}$, and it depends linearly on v' too. Thus, given any two $v, v' \in \mathcal{V}$, we obtain an element $\phi(v)(v') \in \mathcal{W}$ which depends linearly on both v and v'.
- It is therefore possible to identify $\phi \in \mathrm{L}\big(\mathcal{V};\mathrm{L}(\mathcal{V};\mathcal{W})\big)$ with an element $\xi \in \mathrm{L}(\mathcal{V} \times \mathcal{V}, \mathcal{W})$ such that for all $v, v' \in \mathcal{V}$, $\phi(v)(v') = \xi(v, v')$.

To summarise, there is an isomorphism between the space on the right-hand side and the space $\mathrm{L}(\mathcal{V} \times \mathcal{V}, \mathcal{W})$ of all continuous, bilinear maps from $\mathcal{V}$ to $\mathcal{W}$. The second derivative $\mathrm{d}^2\rho(x)$ is therefore a bounded, symmetric, bilinear operator from $\mathcal{V} \times \mathcal{V}$ to $\mathcal{W}$.

Another closely related type of differentiability is the concept of *Gâteaux differentials*, which is the formalism of functional derivatives in calculus of variations. Let $\mathcal{V}$, $\mathcal{W}$ and $\mathcal{U}$ be as before, and consider the function $\rho : \mathcal{U} \to \mathcal{W}$.

Definition 3.2 (Gâteaux derivative)**.** The *Gâteaux differential* or the *Gâteaux derivative* $\partial_v \rho(x)$ of ρ at $x \in \mathcal{U}$ in the direction $v \in \mathcal{V}$ is defined as

$$\partial_v \rho(x) = \lim_{t \to 0} \frac{\rho(x + tv) - \rho(x)}{t},$$

for which this limit is taken relative to the topology of $\mathcal{W}$. The function ρ is said to be *Gâteaux differentiable* at $x \in \mathcal{U}$ if ρ has a directional derivative along all directions at x. We name the operator $\partial\rho(x) : \mathcal{V} \to \mathcal{W}$ which assigns $v \mapsto \partial_v \rho(x) \in \mathcal{W}$ the *Gâteaux derivative* of ρ at x, and the operator $\partial\rho : \mathcal{U} \to (\mathcal{V}; \mathcal{W}) = \{A \,|\, A : \mathcal{V} \to \mathcal{W}\}$ which assigns $x \mapsto \partial\rho(x)$ simply the *Gâteaux derivative* of ρ.

Remark 3.4. For Gâteaux derivatives, $\mathcal{V}$ need only be a vector space, while $\mathcal{W}$ a topological space. Tapia (1971, p. 55) wrote that for quite some time analysis was simply done using the topology of the real line when dealing with functionals. As a result, important concepts such as convergence could not be adequately discussed.

Remark 3.5. Tapia (1971, p. 52) goes on to remark that the space $(\mathcal{V}; \mathcal{W})$ of operators from $\mathcal{V}$ to $\mathcal{W}$ is not a topological space, and there is no obvious way to define a topology on it. Consequently, we cannot consider the Gâteaux derivative of the Gâteaux derivative.

Unlike the Fréchet derivative, which is by definition a linear operator, the Gâteaux derivative may fail to satisfy the additive condition of linearity[3]. Even if it is linear, it may fail to depend continuously on some $v' \in \mathcal{V}$ if $\mathcal{V}$ and $\mathcal{W}$ are infinite dimensional. In this sense, Fréchet derivatives are more demanding than Gâteaux derivatives. Nevertheless, the reasons we bring up Gâteaux derivatives is because it is usually simpler to calculate Gâteaux derivatives than Fréchet derivatives, and the two concepts are connected by the lemma below.

Lemma 3.1 (Fréchet differentiability implies Gâteaux differentiability)**.** *If ρ is Fréchet differentiable at $x \in \mathcal{U}$, then $\rho : \mathcal{U} \to \mathcal{W}$ is Gâteaux differentiable at that point too, and* $\mathrm{d}\rho(x) = \partial\rho(x)$.

Proof. Since ρ is Fréchet differentiable at $x \in \mathcal{U}$, we can write $\rho(x+v) \approx \rho(x) + \mathrm{d}\rho(x)(v)$ for some $v \in \mathcal{V}$. Then,

$$\begin{aligned} \lim_{t \to 0} \left\| \frac{\rho(x + tv) - \rho(x)}{t} - \mathrm{d}\rho(x)(v) \right\|_{\mathcal{W}} &= \lim_{t \to 0} \frac{1}{t} \left\| \rho(x + tv) - \rho(x) - \mathrm{d}\rho(x)(tv) \right\|_{\mathcal{W}} \\ &= \lim_{t \to 0} \frac{\left\| \rho(x + tv) - \rho(x) - \mathrm{d}\rho(x)(tv) \right\|_{\mathcal{W}}}{\|tv\|_{\mathcal{V}}} \|v\|_{\mathcal{V}} \end{aligned} \tag{3.1}$$

[3] Although, for all scalars $\lambda \in \mathbb{R}$, the Gâteaux derivative is homogenous: $\partial_{\lambda v} \rho(x) = \lambda \partial_v \rho(x)$.

converges to 0 since ρ is Fréchet differentiable at x, and $t \to 0$ if and only if $\|tv\|_{\mathcal{V}} \to 0$. Thus, ρ is Gâteaux differentiable at x, and the Gâteaux derivative $\partial_v \rho(x)$ of ρ at x in the direction v coincides with the Fréchet derivatiave of ρ at x evaluated at v. ■

On the other hand, Gâteaux differentiability does not necessarily imply Fréchet differentiability. A sufficient condition for Fréchet differentiability is that the Gâteaux derivative is continuous at the point of differentiation, i.e. the map $\partial\rho : \mathcal{U} \to (\mathcal{V}; \mathcal{W})$ is continuous at $x \in \mathcal{U}$. In other words, if $\partial\rho(x)$ is a bounded linear operator and the convergence in (3.1) is uniform with respect to all v such that $\|v\|_{\mathcal{V}} = 1$, then $\mathrm{d}\rho(x)$ exists and $\mathrm{d}\rho(x) = \partial\rho(x)$ (Tapia, 1971, p. 57 & 66).

Consider now the function $\mathrm{d}\rho(x) : \mathcal{V} \to \mathcal{W}$ and suppose that ρ is twice Fréchet differentiable at $x \in \mathcal{U}$, i.e. $\mathrm{d}\rho(x)$ is Fréchet differentiable at $x \in \mathcal{U}$ with derivative $\mathrm{d}^2\rho(x) : \mathcal{V} \times \mathcal{V} \to \mathcal{W}$. Then, $\mathrm{d}\rho(x)$ is also Gâteaux differentiable at the point x and the two differentials coincide. In particular, we have

$$\left\| \frac{\mathrm{d}\rho(x + tv)(v') - \mathrm{d}\rho(x)(v')}{t} - \mathrm{d}^2\rho(x)(v, v') \right\|_{\mathcal{W}} \to 0 \text{ as } t \to 0, \tag{3.2}$$

by a similar argument in the proof of Lemma 3.1 above. We will use this fact when we describe the Hessian in a little while.

There is also the concept of *gradients* in Hilbert space. Recall that, as a consequence of the Riesz-Fréchet theorem, the mapping $U : \mathcal{V} \to \mathcal{V}^*$ from the Hilbert space $\mathcal{V}$ to its continuous dual space $\mathcal{V}^*$ defined by $U : v \mapsto \langle \cdot, v \rangle_{\mathcal{V}}$ is an isometric isomorphism. Again, let $\mathcal{U} \subseteq \mathcal{V}$ be an open subset, and let $\rho : \mathcal{U} \to \mathbb{R}$ be a Fréchet differentiable function with derivative $\mathrm{d}\rho : \mathcal{U} \to \mathrm{L}(\mathcal{V}; \mathbb{R}) \equiv \mathcal{V}^*$. We define the gradient as follows.

Definition 3.3 (Gradients in Hilbert space)**.** The *gradient* of ρ is the operator $\nabla\rho : \mathcal{U} \to \mathcal{V}$ defined by $\nabla\rho = U^{-1} \circ \mathrm{d}\rho$. Thus, for $x \in \mathcal{U}$, the gradient of ρ at x, denoted $\nabla\rho(x)$, is the unique element of $\mathcal{V}$ satisfying

$$\langle \nabla\rho(x), v \rangle_{\mathcal{V}} = \mathrm{d}\rho(x)(v)$$

for any $v \in \mathcal{V}$. Note that $\nabla\rho$ being a composition of two continuous functions, is itself continuous.

Remark 3.6. Alternatively, the gradient can be motivated using the Riesz representation theorem in Definition 3.1 of the Fréchet derivative. Since $\mathcal{V}^* \ni T : \mathcal{V} \to \mathbb{R}$, there is a unique element $v^* \in \mathcal{V}$ such that $T(v) = \langle v^*, v \rangle_{\mathcal{V}}$ for any $v \in \mathcal{V}$. The element $v^* \in \mathcal{V}$ is called the gradient of ρ at x.

Since the gradient of ρ is an operator on $\mathcal{U}$ to $\mathcal{V}$, it may itself have a Fréchet derivative. Assuming existence, i.e. ρ is twice Fréchet differentiable at $x \in \mathcal{U}$, we call this derivative

the *Hessian* of ρ. From (3.2), it must be that

$$\begin{aligned}
\mathrm{d}^2\rho(x)(v,v') &= \lim_{t\to 0} \frac{\mathrm{d}\rho(x+tv)(v') - \mathrm{d}\rho(x)(v')}{t} \\
&= \lim_{t\to 0} \frac{\langle \nabla\rho(x+tv), v'\rangle_{\mathcal{V}} - \langle \nabla\rho(x), v'\rangle_{\mathcal{V}}}{t} \\
&= \lim_{t\to 0} \left\langle \frac{\nabla\rho(x+tv) - \nabla\rho(x)}{t}, v' \right\rangle_{\mathcal{V}} \\
&= \left\langle \partial_v \nabla\rho(x), v' \right\rangle_{\mathcal{V}}.
\end{aligned}$$

The second line follows from the definition of gradients, the third line by linearity of inner products, and the final line by definition of Gâteaux derivatives and continuity of inner products[4]. Since $\nabla\rho$ is continuous, its Fréchet and Gâteaux differentials coincide, and we have that $\partial_v \nabla\rho(x) = \mathrm{d}\nabla\rho(x)(v)$. Letting $\mathcal{V}$, $\mathcal{W}$ and $\mathcal{U}$ be as before, we now define the Hessian for the function $\rho : \mathcal{U} \to \mathcal{W}$.

Definition 3.4 (Hessian)**.** The Fréchet derivative of the gradient of ρ is known as the *Hessian* of ρ. Denoted $\nabla^2\rho$, it is the mapping $\nabla^2\rho : \mathcal{U} \to \mathrm{L}(\mathcal{V};\mathcal{V})$ defined by $\nabla^2\rho = \mathrm{d}\nabla\rho$, and it satisfies

$$\left\langle \nabla^2\rho(x)(v), v' \right\rangle_{\mathcal{V}} = \mathrm{d}^2\rho(x)(v,v').$$

for $x \in \mathcal{U}$ and $v, v' \in \mathcal{V}$.

Remark 3.7. Since $\mathrm{d}^2\rho(x)$ is a bilinear form in $\mathcal{V}$, we can equivalently write

$$\mathrm{d}^2\rho(x)(v,v') = \langle \mathrm{d}^2\rho(x), v \otimes v' \rangle_{\mathcal{V}\otimes\mathcal{V}}$$

following the correspondence between bilinear forms and tensor product spaces.

With the differentiation tools above, we can now derive the Fisher information that we set out to obtain at the beginning of this section. Let Y be a random variable with density in the parametric family $\{p(\cdot|\theta) \,|\, \theta \in \Theta\}$, where Θ is now assumed to be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_\Theta$. If $p(Y|\theta) > 0$, the log-likelihood function of θ is the real-valued function $L(\cdot|Y) : \Theta \to \mathbb{R}$ defined by $\theta \mapsto \log p(Y|\theta)$. The score S, assuming existence, is defined to be the (Fréchet) derivative of $L(\cdot|Y)$ at θ, i.e. $S : \Theta \to \mathrm{L}(\Theta;\mathbb{R}) \equiv \Theta^*$ defined by $S = \mathrm{d}L(\cdot|Y)$. The second (Fréchet) derivative of $L(\cdot|Y)$ at θ is then $\mathrm{d}^2L(\cdot|Y) : \Theta \to \mathrm{L}(\Theta\times\Theta;\mathbb{R})$. We now prove the following proposition.

Proposition 3.2 (Fisher information in Hilbert spaces)**.** *Assume that both $p(Y|\cdot)$ and $\log p(Y|\cdot)$ are Fréchet differentiable at θ. Then, the Fisher information for $\theta \in \Theta$ is the element in the tensor product space $\Theta \otimes \Theta$ defined by*

$$\mathcal{I}(\theta) = \mathrm{E}[\nabla L(\theta|Y) \otimes \nabla L(\theta|Y)].$$

[4]For any continuous function $g : \mathbb{R} \to \mathbb{R}$, $\lim_{x\to a} g(x) = g(\lim_{x\to a} x) = g(a)$.

Equivalently, assuming further that $\log p(Y|\cdot)$ *is twice Fréchet differentiable at* θ*, the Fisher information can be written as*

$$\mathcal{I}(\theta) = \mathrm{E}[-\nabla^2 L(\theta|Y)].$$

Note that both expectations are taken under the true distribution of random variable Y*.*

Proof. The Gâteaux derivative of $L(\cdot|Y) = \log p(Y|\cdot)$ at $\theta \in \Theta$ in the direction $b \in \Theta$, which is also its Fréchet derivative, is

$$\begin{aligned}
\partial_b L(\theta|Y) &= \frac{\mathrm{d}}{\mathrm{d}t} \log p(Y|\theta + tb)\bigg|_{t=0} \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}t} p(Y|\theta + tb)\big|_{t=0}}{p(Y|\theta)} \\
&= \frac{\partial_b p(Y|\theta)}{p(Y|\theta)}.
\end{aligned}$$

Since it assumed that $p(Y|\cdot)$ is Fréchet differentiable at θ, $\mathrm{d}p(Y|\theta)(b) = \partial_b p(Y|\theta)$. The expectation of the score for any $b \in \Theta$ is shown to be

$$\begin{aligned}
\mathrm{E}[\mathrm{d}L(\theta|Y)(b)] &= \mathrm{E}\left[\frac{\mathrm{d}p(Y|\theta)(b)}{p(Y|\theta)}\right] \\
&= \int \frac{\mathrm{d}p(Y|\theta)(b)}{\cancel{p(Y|\theta)}} \cancel{p(Y|\theta)}\, \mathrm{d}Y \\
&= \mathrm{d}\left(\int p(Y|\theta)\, \mathrm{d}Y\right)(b) \\
&= 0.
\end{aligned}$$

The interchange of Lebesgue integrals and Fréchet differentials is allowed under certain conditions[5], which are assumed to be satisfied here. The derivative of $\int p(Y|\cdot)\, \mathrm{d}Y$ at any value of $\theta \in \Theta$ is the zero vector, as it is the derivative of a constant (i.e. 1).

Using the classical notion that the Fisher information is the variance of the score function, then, for fixed $b, b' \in \Theta$, combined with the fact that $\mathrm{E}[\mathrm{d}L(\theta|Y)]$ is a zero-

[5]Following Kammar (2016), the conditions are:

1. $L(\cdot|Y)$ is Frechét differentiable on $\mathcal{U} \subseteq \Theta$ for almost every $Y \in \mathbb{R}$.
2. $L(\theta|Y)$ and $\mathrm{d}L(\theta|Y)(b)$ are both integrable with respect to Y, for any $\theta \in \mathcal{U} \subseteq \Theta$ and $b \in \Theta$.
3. There is an integrable function $g(Y)$ such that $L(\theta|Y) \leq g(Y)$ for all $\theta \in \Theta$ and almost every $Y \in \mathbb{R}$.

These conditions as stated are analogous to the measure theoretic requirements for Leibniz's integral rule to hold (differentiation under the integral sign). For nice and well-behaved probability densities, such as the normal density that we will be working with, there aren't issues with interchanging integrals and derivatives.

mean function, we have that

$$\begin{aligned}\mathcal{I}(\theta)(b,b') &= \mathrm{E}[\mathrm{d}L(\theta|Y)(b)\cdot \mathrm{d}L(\theta|Y)(b')]\\ &= \mathrm{E}\left[\,\langle \nabla L(\theta|Y), b\rangle_\Theta \left\langle \nabla L(\theta|Y), b'\right\rangle_\Theta\,\right]\\ &= \left\langle \mathrm{E}[\nabla L(\theta|Y)\otimes \nabla L(\theta|Y)], b\otimes b'\right\rangle_{\Theta\otimes\Theta}.\end{aligned}$$

Hence, $\mathcal{I}(\theta)$ as a bilinear form corresponds to the element $\mathrm{E}[\nabla L(\theta|Y)\otimes\nabla L(\theta|Y)] \in \Theta\otimes\Theta$.

The Gâteaux derivative of the Fréchet differential is the second Fréchet derivative, since $L(\cdot|Y)$ is assumed to be twice differentiable at $\theta\in\Theta$:

$$\begin{aligned}\mathrm{d}^2L(\theta|Y)(b,b') &= \partial_{b'}\mathrm{d}L(\theta|Y)(b)\\ &= \partial_{b'}\left(\frac{\mathrm{d}p(Y|\theta)(b)}{p(Y|\theta)}\right)\\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}p(Y|\theta+tb')(b)}{p(Y|\theta+tb')}\right)\Bigg|_{t=0}\\ &= \frac{p(Y|\theta)\mathrm{d}^2p(Y|\theta)(b,b') - \mathrm{d}p(Y|\theta)(b)\mathrm{d}p(Y|\theta)(b')}{p(Y|\theta)^2}\\ &= \frac{\mathrm{d}^2p(Y|\theta)(b,b')}{p(Y|\theta)} - \mathrm{d}L(\theta|Y)(b)\,\mathrm{d}L(\theta|Y)(b').\end{aligned}$$

Taking expectations of the first term in the right-hand side, we get that

$$\begin{aligned}\mathrm{E}\left[\frac{\mathrm{d}^2p(Y|\theta)(b,b')}{p(Y|\theta)}\right] &= \int \frac{\mathrm{d}\big(\mathrm{d}p(Y|\theta)\big)(b,b')}{\cancel{p(Y|\theta)}}\cancel{p(Y|\theta)}\,\mathrm{d}Y\\ &= \mathrm{d}^2\left(\int p(Y|\theta)\,\mathrm{d}Y\right)(b,b')\\ &= 0.\end{aligned}$$

Thus, we see that from the first result obtained,

$$\begin{aligned}\mathrm{E}[-\mathrm{d}^2L(\theta|Y)(b,b')] &= \mathrm{E}[\mathrm{d}L(\theta|Y)(b)\,\mathrm{d}L(\theta|Y)(b')]\\ &= \mathcal{I}(\theta)(b,b'),\end{aligned}$$

while

$$\begin{aligned}\mathrm{E}[-\mathrm{d}^2L(\theta|Y)(b,b')] &= -\,\mathrm{E}\langle\nabla^2L(\theta|Y)(b),b'\rangle_\Theta\\ &= \langle -\,\mathrm{E}\,\nabla^2L(\theta|Y)(b),b'\rangle_\Theta.\end{aligned}$$

It would seem that $\mathrm{E}[-\nabla^2L(\theta|Y)(b)]$ is an operator from Θ onto itself which also induces a bilinear form equivalent to $\mathrm{E}[-\mathrm{d}^2L(\theta|Y)]$. Therefore, $\mathcal{I}(\theta) = \mathrm{E}[-\nabla^2L(\theta|Y)]$. ■

The Fisher information $\mathcal{I}(\theta)$ for θ, much like the covariance operator, can be viewed in one of three ways:

1. As its general form, i.e. an element in $\Theta \otimes \Theta$;
2. As an operator $\mathcal{I}(\theta) : \Theta \to \Theta$ defined by $\mathcal{I}(\theta)(b) = \mathrm{E}[-\nabla^2 L(\theta|Y)](b)$; and finally
3. As a bilinear form $\mathcal{I}(\theta) : \Theta \times \Theta \to \mathbb{R}$ defined by $\mathcal{I}(\theta)(b, b') = \langle -\mathrm{E}\,\nabla^2 L(\theta|Y)(b), b' \rangle_\Theta$ $= \mathrm{E}[-\mathrm{d}^2 L(\theta|Y)(b, b')]$.

In particular, viewed as a bilinear form, the evaluation of the Fisher information for θ at two points b and b' in Θ is seen as the Fisher information between two continuous, linear functionals of θ. For brevity, we denote this $\mathcal{I}(\theta_b, \theta_{b'})$, where $\theta_b = \langle \theta, b \rangle_\theta$ for some $b \in \Theta$. The natural isometry between Θ and Θ^* then allows us to write

$$\mathcal{I}(\theta_b, \theta_{b'}) = \langle \mathcal{I}(\theta), b \otimes b' \rangle_{\Theta \otimes \Theta} = \big\langle \mathcal{I}(\theta), \langle \cdot, b \rangle_\Theta \otimes \langle \cdot, b' \rangle_\Theta \big\rangle_{\Theta^* \otimes \Theta^*}. \tag{3.3}$$

3.3 Fisher information for regression functions

We are now equipped to derive the Fisher information for our regression function. For convenience, we restate the regression model and its assumptions. The regression model relating response variables $y_i \in \mathbb{R}$ and the covariates $x_i \in \mathcal{X}$, for $i = 1, \ldots, n$ is

$$y_i = \alpha + f(x_i) + \epsilon_i \qquad \text{(from 1.1)}$$
$$(\epsilon_1, \ldots, \epsilon_n)^\top \sim \mathrm{N}_n(0, \boldsymbol{\Psi}^{-1}) \qquad \text{(from 1.2)}$$

where $\alpha \in \mathbb{R}$ is an intercept and f is in a RKKS $\mathcal{F}$ with kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Note that the dependence of the kernel on parameters η is implicitly assumed.

Lemma 3.3 (Fisher information for regression function)**.** *For the regression model (1.1) subject to (1.2) and $f \in \mathcal{F}$ where $\mathcal{F}$ is a RKKS with kernel h, the Fisher information for f is given by*

$$\mathcal{I}(f) = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} h(\cdot, x_i) \otimes h(\cdot, x_j)$$

where ψ_{ij} are the (i, j)'th entries of the precision matrix $\boldsymbol{\Psi}$ of the normally distributed model errors. More generally, suppose that $\mathcal{F}$ has a feature space $\mathcal{V}$ such that the mapping $\phi : \mathcal{X} \to \mathcal{V}$ is its feature map, and if $f(x) = \langle \phi(x), v \rangle_\mathcal{V}$, then the Fisher information $I(v) \in \mathcal{V} \otimes \mathcal{V}$ for v is

$$\mathcal{I}(v) = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} \phi(x_i) \otimes \phi(x_j).$$

Proof. For $x \in \mathcal{X}$, let $k_x : \mathcal{V} \to \mathbb{R}$ be defined by $k_x(v) = \langle \phi(x), v \rangle_{\mathcal{V}}$. Clearly, k_x is linear and continuous. Hence, the Gâteaux derivative of $k_x(v)$ in the direction u is

$$\begin{aligned}
\partial_u k_x(v) &= \lim_{t \to 0} \frac{k(v + tu) - k(v)}{t} \\
&= \lim_{t \to 0} \frac{\langle \phi(x), v + tu \rangle_{\mathcal{V}} - \langle \phi(x), v \rangle_{\mathcal{V}}}{t} \\
&= \lim_{t \to 0} \frac{t \langle \phi(x), u \rangle_{\mathcal{V}}}{t} \\
&= \langle \phi(x), u \rangle_{\mathcal{V}}.
\end{aligned}$$

Since clearly $\partial_u k_x(v)$ is a continuous linear operator for any $u \in \mathcal{V}$, it is bounded, so the Fréchet derivative exists and $\mathrm{d}k_x(v) = \partial k_x(v)$. Let $\mathbf{y} = \{y_1, \ldots, y_n\}$, and denote the hyperparameters of the regression model by $\theta = \{\alpha, \mathbf{\Psi}, \eta\}$. Without loss of generality, assume $\alpha = 0$, and even if this is not so, we can always add back α to the y_i's later. Regardless, both α and $\mathbf{y}$ are constant in the differential of $L(v|\mathbf{y}, \theta)$. The log-likelihood of v is given by

$$L(v|\mathbf{y}, \theta) = \text{const.} - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} \big(y_i - k_{x_i}(v)\big) \big(y_j - k_{x_j}(v)\big)$$

and the score by

$$\begin{aligned}
\mathrm{d}L(\cdot|\mathbf{y}, \theta) &= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} \, \mathrm{d}(k_{x_i} k_{x_j} - y_j k_{x_i} - y_i k_{x_j} + y_i y_j) \\
&= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} (k_{x_j} \mathrm{d}k_{x_i} + k_{x_i} \mathrm{d}k_{x_j} - y_j \mathrm{d}k_{x_i} - y_i \mathrm{d}k_{x_j}).
\end{aligned}$$

Differentiating again gives

$$\begin{aligned}
\mathrm{d}^2 L(\cdot|\mathbf{y}, \theta) &= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} (\mathrm{d}k_{x_j} \mathrm{d}k_{x_i} + \mathrm{d}k_{x_i} \mathrm{d}k_{x_j}) \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} \, \mathrm{d}k_{x_i} \mathrm{d}k_{x_j},
\end{aligned}$$

since the derivative of $\mathrm{d}k_x$ is zero (it is the derivative of a constant). We can then calculate the Fisher information to be

$$\begin{aligned}\mathcal{I}(v) = -\operatorname{E}\left[\mathrm{d}^2 L(v|\mathbf{y},\theta)\right] &= \operatorname{E}\left[\sum_{i=1}^n\sum_{j=1}^n \psi_{ij}\langle\phi(x_i),\cdot\rangle_{\mathcal{V}}\,\langle\phi(x_j),\cdot\rangle_{\mathcal{V}}\right] \\ &= \sum_{i=1}^n\sum_{j=1}^n \psi_{ij}\,\langle\phi(x_i)\otimes\phi(x_j),\cdot\rangle_{\mathcal{V}\otimes\mathcal{V}} \\ &= \sum_{i=1}^n\sum_{j=1}^n \psi_{ij}\cdot\phi(x_i)\otimes\phi(x_j).\end{aligned}$$

Here, we had treated $\phi(x_i)\otimes\phi(x_j)$ as a bilinear operator, since $\mathcal{I}(v)\in\mathcal{V}\otimes\mathcal{V}$ as well. Also, the expectation is free of the random variable under expectation (i.e. $\mathbf{y}$), which makes the second line possible.

By taking the canonical feature $\phi(x) = h(\cdot,x)$, we have that $\phi\equiv h(\cdot,x):\mathcal{X}\to\mathcal{F}\equiv\mathcal{V}$ and therefore for $f\in\mathcal{F}$, the reproducing property gives us $f(x) = \langle h(\cdot,x), f\rangle_{\mathcal{F}}$, so the formula for $\mathcal{I}(f)\in\mathcal{F}\otimes\mathcal{F}$ follows. ■

The above lemma gives the form of the Fisher information for f in a rather abstract fashion. Consider the following example of applying Lemma 3.3 to obtain the Fisher information for a standard linear regression model.

Example 3.1 (Fisher information for linear regression)**.** As before, suppose model (1.1) subject to (1.2) and $f\in\mathcal{F}$, a RKHS. For simplicity, we assume iid errors, i.e. $\boldsymbol{\Psi} = \psi\mathbf{I}_n$. Let $\mathcal{X} = \mathbb{R}^p$, and the feature space $\mathcal{V} = \mathbb{R}^p$ be equipped with the usual dot product $\langle\cdot,\cdot\rangle_{\mathcal{V}}:\mathcal{V}\otimes\mathcal{V}\to\mathbb{R}$ defined by $v^\top v$. Consider also the identity feature map $\phi:\mathcal{X}\to\mathcal{V}$ defined by $\phi(\mathbf{x}) = \mathbf{x}$. For some $\boldsymbol{\beta}\in\mathcal{V}$, the linear regression model is such that $f(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\beta} = \langle\phi(\mathbf{x}),\boldsymbol{\beta}\rangle_{\mathcal{V}}$. Therefore, according to Lemma 3.3, the Fisher information for $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta}) &= \sum_{i=1}^n\sum_{j=1}^n \psi\cdot\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_j) \\ &= \psi\sum_{i=1}^n\sum_{j=1}^n \mathbf{x}_i\otimes\mathbf{x}_j \\ &= \psi\mathbf{X}^\top\mathbf{X}.\end{aligned}$$

Note that the operation '$\otimes$' on two vectors in Euclidean space is simply their outer product. The resulting $\mathbf{X}$ is a $n\times p$ matrix containing the entries $\mathbf{x}_1^\top,\ldots,\mathbf{x}_n^\top$ row-wise. This is of course recognised as the Fisher information for the regression coefficients in the standard linear regression model.

We can also compute the Fisher information for linear functionals of f, and in particular, for point evaluation functionals of f, thereby allowing us to compute the Fisher information at two points $f(x)$ and $f(x')$.

Corollary 3.3.1 (Fisher information between two linear functionals of f)**.** *For our regression model as defined in (1.1) subject to (1.2) and f belonging to a RKKS $\mathcal{F}$ with kernel h, the Fisher information at two points $f(x)$ and $f(x')$ is given by*

$$\mathcal{I}\big(f(x), f(x')\big) = \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij} h(x, x_i) h(x', x_j).$$

Proof. In a RKKS $\mathcal{F}$, the reproducing property gives $f(x) = \langle f, h(\cdot, x)\rangle_{\mathcal{F}}$ and in particular, $\langle h(\cdot, x), h(\cdot, x')\rangle_{\mathcal{F}} = h(x, x')$. By (3.3), we have that

$$\begin{aligned}
\mathcal{I}(f)\big(h(\cdot, x), h(\cdot, x')\big) &= \big\langle \mathcal{I}(f), h(\cdot, x) \otimes h(\cdot, x')\big\rangle_{\mathcal{F}\otimes\mathcal{F}} \\
&= \left\langle \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij} h(\cdot, x_i) \otimes h(\cdot, x_j) \ , \ h(\cdot, x) \otimes h(\cdot, x') \right\rangle_{\mathcal{F}\otimes\mathcal{F}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij} \big\langle h(\cdot, x_i), h(\cdot, x)\big\rangle_{\mathcal{F}} \big\langle h(\cdot, x_j), h(\cdot, x')\big\rangle_{\mathcal{F}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij} h(x, x_i) h(x', x_j).
\end{aligned}$$

The second to last line follows from the definition of the usual inner product for tensor spaces, and the last line follows by the reproducing property. ■

An inspection of the formula in Corollary 3.3.1 reveals the fact that the Fisher information for $f(x)$, $\mathcal{I}\big(f(x), f(x)\big)$, is positive if and only if $h(x, x_i) \neq 0$ for at least one $i \in \{1, \ldots, n\}$. In practice, this condition is often satisfied for all x, so this result might be considered both remarkable and reassuring, because it suggests we can estimate f over its entire domain, no matter how big, even though we only have a finite amount of data points.

3.4 The induced Fisher information RKHS

From Lemma 3.3, the formula for the Fisher information uses n points of the observed data $x_i \in \mathcal{X}$. This seems to suggest that the Fisher information only exists for a finite subspace of the RKKS $\mathcal{F}$. Indeed, this is the case, and we will be specific about the subspace for which there is Fisher information. Consider the following set, a similar one

considered in the proof of the Moore-Aronszajn theorem (Theorem 2.6, p. 51):

$$\mathcal{F}_n = \left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, f(x) = \sum_{i=1}^{n} h(x, x_i) w_i,\ w_i \in \mathbb{R},\ i = 1, \ldots, n \right\}. \tag{3.4}$$

Since $h(\cdot, x_i) \in \mathcal{F}$, any $f \in \mathcal{F}_n$ is also in $\mathcal{F}$ by linearity, and thus $\mathcal{F}_n$ is a subset of $\mathcal{F}$. Further, $\mathcal{F}_n$ is closed under addition and multiplication by a scalar, and is therefore a subspace of $\mathcal{F}$. Unlike Theorem 2.6, $\mathcal{F}_n$ defined here is a finite subspace of dimension n.

Let $\mathcal{F}_n^\perp$ be the orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$. By the orthogonal decomposition theorem (Theorem 2.3, p. 43), any regression function $f \in \mathcal{F}$ can be uniquely decomposed as $f = f_n + r$, with $f_n \in \mathcal{F}_n$ and $r \in \mathcal{F}_n^\perp$, where $\mathcal{F} = \mathcal{F}_n \oplus \mathcal{F}_n^\perp$. We saw in the proof of Theorem 2.6 that $\mathcal{F}$ is the closure of $\mathcal{F}_n$, so therefore $\mathcal{F}$ is dense in $\mathcal{F}_n$, and hence by Corollary 2.3.1 (p. 43) we have that $\mathcal{F}_n^\perp = \{0\}$. Alternatively, we could have argued that any $r \in \mathcal{F}_n^\perp$ is orthogonal to each of the $h(\cdot, x_i) \in \mathcal{F}$, so by the reproducing property of h, $r(x_i) = \langle r, h(\cdot, x_i) \rangle_{\mathcal{F}} = 0$. This suggests the following corollary.

Corollary 3.3.2. *With $g \in \mathcal{F}$, the Fisher information for g is zero if and only if $g \in \mathcal{F}_n^\perp$, i.e. if and only if $g(x_1) = \cdots = g(x_n) = 0$.*

Proof. Let $\mathcal{I}(f)$ be the Fisher information for f. The Fisher information for $\langle f, r \rangle_{\mathcal{F}}$ is

$$\begin{aligned}
\mathcal{I}(f)(r, r) &= \langle \mathcal{I}(f), r \otimes r \rangle_{\mathcal{F} \otimes \mathcal{F}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} \langle h(\cdot, x_i), r \rangle_{\mathcal{F}} \langle h(\cdot, x_j), r \rangle_{\mathcal{F}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} r(x_i) r(x_j).
\end{aligned}$$

So if $r \in \mathcal{F}_n^\perp$, then $r(x_1) = \cdots = r(x_n) = 0$, and thus the Fisher information at $r \in \mathcal{F}_n^\perp$ is zero. Conversely, if the Fisher information is zero, it must necessarily mean that $r(x_1) = \cdots = r(x_n) = 0$ since $\psi_{ij} > 0$, and thus $r \in \mathcal{F}_n^\perp$. ∎

The above corollary implies that the Fisher information for our regression function $f \in \mathcal{F}$ exists only on the n-dimensional subspace $\mathcal{F}_n$. More subtly, as there is no Fisher information for $r \in \mathcal{F}_n^\perp$, r cannot be estimated from the data. Thus, in estimating f, we will only ever consider the finite subspace $\mathcal{F}_n \subset \mathcal{F}$ where there is information about f.

As it turns out, $\mathcal{F}_n$ can be identified as a RKHS with reproducing kernel equal to the Fisher information for f. That is, the real, symmetric, and positive-definite function h_n over $\mathcal{X} \times \mathcal{X}$ defined by $h_n(x, x') = \mathcal{I}\big(f(x), f(x')\big)$ is associated to the RKHS which is $\mathcal{F}_n$, equipped with the squared norm $\|f\|_{\mathcal{F}_n}^2 = \sum_{i,j=1}^{n} w_i (\boldsymbol{\Psi}^{-1})_{ij} w_j$. This is stated in the next lemma.

Lemma 3.4. *Let $\mathcal{F}_n$ as in (3.4) be equipped with the inner product*

$$\langle f, f' \rangle_{\mathcal{F}_n} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i (\mathbf{\Psi}^{-1})_{ij} w'_j = \mathbf{w}^\top \mathbf{\Psi} \mathbf{w}' \tag{3.5}$$

for any two $f = \sum_{i=1}^{n} h(\cdot, x_i) w_i$ and $f' = \sum_{j=1}^{n} h(\cdot, x_j) w'_j$ in $\mathcal{F}_n$. Then, $h_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as defined by

$$h_n(x, x') = \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} h(x, x_i) h(x', x_j)$$

is the reproducing kernel of $\mathcal{F}_n$.

Proof. What needs to be proven is the reproducing property of h_n for $\mathcal{F}_n$. First note that by defining $w_j(x) = \sum_{k=1}^{n} \psi_{jk} h(x, x_k)$, we see that

$$h_n(x, \cdot) = \sum_{j=1}^{n} \sum_{k=1}^{n} \psi_{jk} h(x, x_j) h(\cdot, x_k) = \sum_{j=1}^{n} w_j(x) h(\cdot, x_j)$$

Furthermore, writing $h(\cdot, x_j) = \sum_{k=1}^{n} \delta_{jk} h(\cdot, x_k)$, with δ being the Kronecker delta, we see that $h(\cdot, x_j)$ is also an element of $\mathcal{F}_n$, and in particular,

$$\big\langle h(\cdot, x_i), h(\cdot, x_k) \big\rangle_{\mathcal{F}_n} = \sum_{j=1}^{n} \sum_{l=1}^{n} \delta_{ij} (\mathbf{\Psi}^{-1})_{jl} \delta_{lk} = (\mathbf{\Psi}^{-1})_{ik}.$$

Denote by ψ^-_{ij} the (i,j)'th element of $\mathbf{\Psi}^{-1}$. A fact we will use later is $\sum_{k=1}^{n} \psi_{jk} \psi^-_{ik} = (\mathbf{\Psi}\mathbf{\Psi}^{-1})_{ji} = (\mathbf{I}_n)_{ji} = \delta_{ji}$. In the mean time,

$$\begin{aligned}
\langle f, h_n(x, \cdot) \rangle_{\mathcal{F}_n} &= \left\langle \sum_{i=1}^{n} h(\cdot, x_i) w_i, \sum_{j=1}^{n} \sum_{k=1}^{n} \psi_{jk} h(x, x_j) h(\cdot, x_k) \right\rangle_{\mathcal{F}_n} \\
&= \sum_{i=1}^{n} w_i \sum_{j=1}^{n} \sum_{k=1}^{n} \psi_{jk} h(x, x_j) \big\langle h(\cdot, x_i), h(\cdot, x_k) \big\rangle_{\mathcal{F}_n} \\
&= \sum_{i=1}^{n} w_i \sum_{j=1}^{n} \sum_{k=1}^{n} \psi_{jk} h(x, x_j) \psi^-_{ik} \\
&= \sum_{i=1}^{n} w_i \sum_{j=1}^{n} \delta_{ji} h(x, x_j) \\
&= \sum_{i=1}^{n} w_i h(x, x_i) \\
&= f(x).
\end{aligned}$$

Therefore, h_n is a reproducing kernel for $\mathcal{F}_n$. Obviously, h_n is positive definite (it is a squared kernel), and hence it defines the RKHS $\mathcal{F}_n$. ■

3.5 The I-prior

In the introductory chapter (Chapter 1), we discussed that unless the regression function f is regularised (for instance, using some prior information), the ML estimator of f is likely to be inadequate. In choosing a prior distribution for f, we appeal to the principle of maximum entropy (Jaynes, 1957a, 1957b, 2003), which states that the probability distribution which best represents the current state of knowledge is the one with largest entropy. In this section, we aim to show the relationship between the Fisher information for f and its maximum entropy prior distribution. Before doing this, we recall the definition of entropy and derive the maximum entropy prior distribution for a parameter which has unrestricted support. Let (Θ, D) be a metric space and let $\nu = \nu_D$ be a volume measure induced by D (e.g. Hausdorff measure). In addition, assume ν is a probability measure over Θ so that $(\Theta, \mathcal{B}(\Theta), \nu)$ is a Borel probability space.

Definition 3.5 (Entropy)**.** Denote by p a probability density over Θ relative to ν. Suppose that $\int p \log p \, \mathrm{d}\nu < \infty$, i.e. $p \log p$ is Lebesgue integrable and belongs to the space $\mathrm{L}^1(\Theta, \nu)$. The entropy of a distribution p over Θ relative to a measure ν is defined as

$$H(p) = -\int_\Theta p(\theta) \log p(\theta) \, \mathrm{d}\nu(\theta). \tag{3.6}$$

In deriving the maximum entropy distribution, we will need to maximise the functional H with respect to p. Typically, this is done using calculus of variations techniques, and standard calculations (Appendix A, p. 259) reveal that the functional derivative of $H(p)$ with respect to p, denoted $\partial H / \partial p$, is equal to $-1 - \log p$. We now present another well known result from information theory, regarding the form of the maximum entropy distribution.

Lemma 3.5 (Maximum entropy distribution)**.** *Let (Θ, D) be a metric space, $\nu = \nu_D$ be a volume measure induced by D, and p be a probability density function on Θ. The entropy maximising density $\tilde{p}$, which satisfies*

$$\underset{p \in L^2(\Theta, \nu)}{\arg\max} \, H(p) = -\int_\Theta \tilde{p}(\theta) \log \tilde{p}(\theta) \, \mathrm{d}\nu(\theta),$$

subject to the constraints

$$\mathrm{E}\left[D(\theta, \theta_0)^2\right] = \int_\Theta D(\theta, \theta_0)^2 p(\theta) \, \mathrm{d}\nu(\theta) = \textit{const.}, \qquad \int_\Theta p(\theta) \, \mathrm{d}\nu(\theta) = 1,$$
$$\textit{and} \quad p(\theta) \geq 0, \forall \theta \in \Theta,$$

is the density given by

$$\tilde{p}(\theta) \propto \exp\left(-\frac{1}{2} D(\theta, \theta_0)^2\right),$$

for some fixed $\theta_0 \in \Theta$. If (Θ, D) is a Euclidean space and ν a flat (Lebesgue) measure then $\tilde{p}$ represents a (multivariate) normal density.

Sketch proof. This follows from standard calculus of variations, though we provide a sketch proof here. Set up the Langrangian

$$\begin{aligned}\mathcal{L}(p, \gamma_1, \gamma_2) = &- \int_\Theta p(\theta) \log p(\theta) \, \mathrm{d}\nu(\theta) + \gamma_1 \left(\int_\Theta D(\theta, \theta_0)^2 p(\theta) \, \mathrm{d}\nu(\theta) - \text{const.} \right) \\ &+ \gamma_2 \left(\int_\Theta p(\theta) \, \mathrm{d}\nu(\theta) - 1 \right).\end{aligned}$$

Taking derivatives with respect to p (see Appendix A, p. 259 for definition of functional derivatives) yields

$$\frac{\partial}{\partial p} \mathcal{L}(p, \gamma_1, \gamma_2)(\theta) = -1 - \log p(\theta) + \gamma_1 D(\theta, \theta_0)^2 + \gamma_2.$$

Set this to zero, and solve for $p(\theta)$:

$$\begin{aligned}p(\theta) &= \exp\left(\gamma_1 D(\theta, \theta_0)^2 + \gamma_2 - 1\right) \\ &\propto \exp\left(\gamma_1 D(\theta, \theta_0)^2\right).\end{aligned}$$

This density is positive for any values of γ_1 (and γ_2), and it normalises to one if $\gamma_1 < 0$. As γ_1 can take any value less than zero, we choose $\gamma_1 = -1/2$.

Now, if $\Theta \equiv \mathbb{R}^m$ and ν is the Lebesgue measure, then $D(\theta, \theta_0)^2 = \|\theta - \theta_0\|^2_{\mathbb{R}^m}$, so $\tilde{p}$ is recognised as a multivariate normal density centred at θ_0 with identity covariance matrix. ■

Returning to the normal regression model of (1.1) subject to (1.2), we shall now derive the maximum entropy prior for f in some RKKS $\mathcal{F}$. One issue that we have is that the set $\mathcal{F}$ is potentially "too big" for the purpose of estimating f, that is, for certain pairs of functions $\mathcal{F}$, the data do not allow an assessment of whether one is closer to the truth than the other. In particular, the data do not contain information to distinguish between two functions f and g in $\mathcal{F}$ for which $f(x_i) = g(x_i), i = 1, \ldots, n$. Since the Fisher information for a linear functional of a non-zero $f \in \mathcal{F}_n$ is non-zero, there is information to allow a comparison between any pair of functions in $f_0 + \mathcal{F}_n := \{f_0 + f_n \,|\, f_0 \in \mathcal{F}, f_n \in \mathcal{F}_n\}$. A prior for f therefore need not have support $\mathcal{F}$, instead it is sufficient to consider priors with support $f_0 + \mathcal{F}_n$, where $f_0 \in \mathcal{F}$ is fixed and chosen a priori as a "best guess" of f. We now state and prove the main I-prior theorem.

Theorem 3.6 (The I-prior)**.** *Let $\mathcal{F}$ be a RKKS with kernel h, and consider the finite dimensional subspace $\mathcal{F}_n$ of $\mathcal{F}$ equipped with an inner product as per (3.5). Let ν be a volume measure induced by the norm $\|\cdot\|_{\mathcal{F}_n} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{F}_n}}$. With $f_0 \in \mathcal{F}$, let $\mathcal{D}_0$ be the class*

of distributions p such that

$$\mathrm{E}\left[\|f - f_0\|^2_{\mathcal{F}_n}\right] = \int_{\mathcal{F}_n} \|f - f_0\|^2_{\mathcal{F}_n}\, p(f)\,\mathrm{d}\nu(f) = \textit{const.}$$

Denote by $\tilde{p}$ the density of the entropy maximising distribution among the class of distributions within $\mathcal{D}_0$. Then, $\tilde{p}$ is Gaussian over $\mathcal{F}$ with mean f_0 and covariance function equal to the reproducing kernel of $\mathcal{F}_n$, i.e.

$$\mathrm{Cov}\left[f(x), f(x')\right] = h_n(x, x').$$

We call $\tilde{p}$ the I-prior *for f.*

Proof. Recall the fact that any $f \in \mathcal{F}$ can be decomposed into $f = f_n + r$, with $f_n \in \mathcal{F}_n$ and $r \in \mathcal{F}_n^\perp$. Also recall that there is no Fisher information about any $r \in \mathcal{R}_n$, and therefore it is not possible to estimate r from the data. Therefore, $p(r) = 0$, and one needs only consider distributions over $\mathcal{F}_n$ when building distributions over $\mathcal{F}$.

The norm on $\mathcal{F}_n$ induces the metric $D(f, f') = \|f - f'\|_{\mathcal{F}_n}$. Consider functions in the set $f_0 + \mathcal{F}_n$, i.e. functions of the form

$$f = f_0 + \sum_{i=1}^n h(\cdot, x_i) w_i,$$

such that $(f - f_0) \in \mathcal{F}_n$. Compute the squared distance between f and f_0:

$$\begin{aligned} D(f, f_0)^2 &= \|f - f_0\|^2_{\mathcal{F}_n} \\ &= \left\| \sum_{i=1}^n h(\cdot, x_i) w_i \right\|^2_{\mathcal{F}_n} \\ &= \mathbf{w}^\top \boldsymbol{\Psi}^{-1} \mathbf{w}. \end{aligned}$$

Thus, by Lemma 3.5, the maximum entropy distribution for $f - f_0 = \sum_{i=1}^n h(\cdot, x_i) w_i$ is

$$(w_1, \ldots, w_n)^\top \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}).$$

This implies that f is Gaussian, since

$$\begin{aligned} \langle f, f' \rangle_{\mathcal{F}} &= \left\langle f_0 + \sum_{i=1}^n h(\cdot, x_i) w_i \,,\, f' \right\rangle_{\mathcal{F}} \\ &= \langle f_0, f' \rangle_{\mathcal{F}} + \sum_{i=1}^n w_i \left\langle h(\cdot, x_i), f' \right\rangle_{\mathcal{F}} \end{aligned}$$

is a sum of normal random variables, and therefore $\langle f, f' \rangle_{\mathcal{F}}$ is normally distributed for any $f' \in \mathcal{F}$. The mean $\mu \in \mathcal{F}$ of this random vector f satisfies $\mathrm{E}\langle f, f' \rangle_{\mathcal{F}} = \langle \mu, f' \rangle_{\mathcal{F}}$ for all $f' \in \mathcal{F}_n$, but

$$\begin{aligned}
\mathrm{E}\langle f, f' \rangle_{\mathcal{F}} &= \langle f_0, f' \rangle_{\mathcal{F}} + \mathrm{E}\left[\sum_{i=1}^{n} w_i \left\langle h(\cdot, x_i), f' \right\rangle_{\mathcal{F}} \right] \\
&= \langle f_0, f' \rangle_{\mathcal{F}} + \sum_{i=1}^{n} \cancelto{0}{\mathrm{E}[w_i]} \left\langle h(\cdot, x_i), f' \right\rangle_{\mathcal{F}} \\
&= \langle f_0, f' \rangle_{\mathcal{F}},
\end{aligned}$$

so $\mu \equiv f_0$.

Following Definition 2.16 (p. 46), the covariance between two evaluation functionals of f is shown to satisfy

$$\begin{aligned}
\mathrm{Cov}\left[f(x), f(x') \right] &= \mathrm{Cov}\left[\langle f, h(\cdot, x) \rangle_{\mathcal{F}}, \langle f, h(\cdot, x') \rangle_{\mathcal{F}} \right] \\
&= \mathrm{E}\left[\langle f - f_0, h(\cdot, x) \rangle_{\mathcal{F}} \langle f - f_0, h(\cdot, x') \rangle_{\mathcal{F}} \right].
\end{aligned}$$

Then, making use of the reproducing property of h for $\mathcal{F}$, we have that

$$\begin{aligned}
\mathrm{Cov}\left[f(x), f(x') \right] &= \mathrm{E}\left[\left\langle \sum_{i=1}^{n} h(\cdot, x_i) w_i, h(\cdot, x) \right\rangle_{\mathcal{F}} \left\langle \sum_{j=1}^{n} h(\cdot, x_j) w_j, h(\cdot, x') \right\rangle_{\mathcal{F}} \right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \left\langle h(\cdot, x), h(\cdot, x_i) \right\rangle_{\mathcal{F}} \left\langle h(\cdot, x'), h(\cdot, x_j) \right\rangle_{\mathcal{F}} \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} h(x, x_i) h(x', x_j),
\end{aligned}$$

which is the reproducing kernel for $\mathcal{F}_n$. ■

In closing, we reiterate the fact that the I-prior for f in the normal regression model subject to f belonging to some RKKS $\mathcal{F}$ with kernel h_η has the simple representation

$$\begin{aligned}
f(x_i) &= f_0(x_i) + \sum_{k=1}^{n} h_\eta(x_i, x_k) w_k \\
(w_1, \ldots, w_n)^\top &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}).
\end{aligned} \tag{3.7}$$

Equivalently, this may be written as a Gaussian process-like prior

$$\big(f(x_1), \ldots, f(x_n) \big)^\top \sim \mathrm{N}(\mathbf{f}_0, \mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta), \tag{3.8}$$

where $\mathbf{f}_0 = \big(f_0(x_1), \ldots, f_0(x_n) \big)^\top$ is the vector of prior mean functional evaluations, and $\mathbf{H}_\eta = \big(h_\eta(x_i, x_j) \big)_{i,j=1}^{n}$ is the kernel matrix.

3.6 Conclusion

In estimating the regression function f of the normal model in (1.1) subject to (1.2) and f belonging to a RKKS $\mathcal{F}$, we established that the entropy maximising prior distribution for f is Gaussian with some chosen prior mean f_0, and covariance function proportional[6] to the Fisher information for f. We call this the I-prior for f.

The dimension of the function space $\mathcal{F}$ could be huge, infinite-dimensional even, while the task of estimating $f \in \mathcal{F}$ only relies on a finite amount of data point. However, we are certain that the Fisher information for f exists only for the finite subspace $\mathcal{F}_n$ as defined in (3.4), and it is zero everywhere else. This suggests that the data only allows us to provide an estimation to the function $f \in \mathcal{F}$ by considering functions in an (at most) n-dimensional subspace instead. In other words, it would be futile to consider functions in a space larger than this, and hence there is an element of dimension reduction here, especially when $\dim(\mathcal{F}) \gg n$.

By equipping the subspace $\mathcal{F}_n$ with the inner product (3.5), $\mathcal{F}_n$ is revealed to be a RKHS with reproducing kernel equal to the Fisher information for f. Importantly, functions in the subspace $\mathcal{F}_n$ are structurally similar to the functions in the parent space $\mathcal{F}$ (though their topologies may differ). The problem at hand then boils down to a Gaussian process regression using the kernel of the RKHS $\mathcal{F}_n$, which is the Fisher information for f.

[6] Proportionality, rather than equality, is a consequence of any RKHS scale parameters that $\mathcal{F}$ may have.

Chapter 4

Modelling with I-priors

In the previous chapter, we defined an I-prior for the normal regression model (1.1) subject to (1.2) and f belonging to a reproducing kernel Hilbert or Krein space of functions $\mathcal{F}$, as a Gaussian distribution on f with covariance function proportional to the Fisher information for f. We also saw how new function spaces can be constructed via the polynomial and ANOVA reproducing kernel Kreĭn spaces (RKKSs). In this chapter, we shall describe various regression models, and identify them with appropriate RKKSs, so that an I-prior may be defined on it.

Methods for estimating I-prior models are described in Section 4.2. Estimation here refers to obtaining the posterior distribution of the regression function under an I-prior, while optimising the kernel parameters of $\mathcal{F}$ and the error precision $\boldsymbol{\Psi}$. Likelihood based methods, namely direct optimisation of the likelihood and the expectation-maximisation (EM) algorithm, are the preferred estimation methods of choice. Having said this, it is also possible to estimate I-prior models under a full Bayesian paradigm by employing Markov chain Monte Carlo methods to sample from the relevant posterior densities. Once estimation is completed, post-estimation procedures such as inference and prediction for a new data point can be done. This is described in Section 4.4.

Careful considerations of the computational aspects are required to ensure efficient estimation of I-prior models, and these are discussed in Section 4.3. The culmination of the computational work on I-prior estimation is the **iprior** package (Jamil, 2017), which is a publicly available R package that has been published to the Comprehensive R Archive Network (CRAN).

Finally, several examples of I-prior modelling are presented in Section 4.5: in particular, a multilevel data set, a longitudinal data set, and a data set involving a functional covariate, are analysed using the I-prior methodology. Code for replication is available at `http://myphdcode.haziqj.ml`.

4.1 Various regression models

In the introductory chapter (Section 1.1), we described several interesting regression models. The goal of this section is to formulate the I-prior model that describes each of these models. This is done by carefully choosing the RKHS/RKKS $\mathcal{F}$ of real functions over a set $\mathcal{X}$ to which the regression function f belongs. Without loss of generality and for simplicity, assume a prior mean of zero for the I-prior distribution.

4.1.1 Multiple linear regression

Let $\mathcal{X} \equiv \mathbb{R}^p$ be equipped with the regular Euclidean dot product, and $\mathcal{F}_\lambda$ be the scaled canonical RKHS of functions over $\mathcal{X}$ with kernel $h_\lambda(\mathbf{x}, \mathbf{x}') = \lambda \mathbf{x}^\top \mathbf{x}'$, for any two $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$. Then, an I-prior on f implies that

$$\begin{aligned} f(\mathbf{x}_i) &= \sum_{j=1}^{n} \lambda \mathbf{x}_i^\top \mathbf{x}_j w_j \\ &= \sum_{j=1}^{n} \lambda \left(\sum_{k=1}^{p} x_{ik} x_{jk} \right) w_j \\ &= \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \end{aligned}$$

where each $\beta_k := \lambda \sum_{j=1}^{n} x_{jk} w_j$. This implies a multivariate normal prior distribution for the regression coefficients

$$\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p) \sim \mathrm{N}_p(\mathbf{0}, \lambda^2 \mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X}), \tag{4.1}$$

where $\mathbf{X}$ is the $n \times p$ design matrix for the covariates, excluding the column of ones at the beginning typically reserved for the intercept. As expected, the covariance matrix for $\boldsymbol{\beta}$ is recognised as the scaled Fisher information matrix for the regression coefficients.

If the covariates are not measured similarly, e.g. weights in kilograms, heights in metres, etc., then it makes sense to introduce scale parameters λ_k to account for the difference in scale. One could decompose the regression function into

$$f(\mathbf{x}_i) = f_1(x_{i1}) + \cdots + f_p(x_{ip})$$

for which $f \in \mathcal{F}_\lambda \equiv \mathcal{F}_{\lambda_1} \oplus \cdots \oplus \mathcal{F}_{\lambda_p}$, and $\mathcal{F}_{\lambda_k}$, $k = 1, \ldots, p$ are unidimensional canonical RKHSs with kernels $h_{\lambda_k}(x_{ik}, x_{jk}) = \lambda_k x_{ik} x_{jk}$. In effect, we now have p scale parameters, one for each of the RKHSs associated with the p covariates. The RKKS $\mathcal{F}_\lambda$ therefore has kernel

$$h(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} \lambda_k x_{ik} x_{jk},$$

and hence each regression coefficient can now be written as $\beta_k = \sum_{j=1}^{n} \lambda_k x_{jk} w_j$, for which we see the λ_k's scaling role on the x_{jk}'s. Thus, the corresponding I-prior for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, \mathbf{X}^\top \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}\mathbf{X}),$$

with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Note that $\mathcal{F}_\lambda$ can be seen as a special case of the ANOVA RKKS, in which only the main effects are considered, in which case the *centred canonical RKHSs* should be considered instead. This approach is disadvantageous when p is large, in which case there would be numerous scale parameters to estimate.

Remark 4.1. Of course, one could simply turn to standardisation of the $\mathbf{X}$ variables, so as to make the variables measure on the same scale. We feel this is a rather ad-hoc approach which creates meaningless units (they are standard deviations) for the covariates which are then fiddly to interpret. On the other hand, there is a balance to be made between elegance and feasibility. With large p, standardising is much simpler and computationally less burdensome than estimating p individual scale parameters. In Chapter 6, where we tackle the problem of Bayesian variable selection using I-priors in linear models, standardisation of the variables is done for the sake of streamlining the Gibbs sampler.

Remark 4.2. The I-prior for $\boldsymbol{\beta}$ in (4.1) bears resemblance to the g-prior (Zellner, 1986), and in fact, the g-prior can be interpreted as an I-prior if the inner product of $\mathcal{X}$ is the Mahalonobis inner product. See Appendix E for a discussion.

4.1.2 Multilevel linear modelling

Let $\mathcal{X} \equiv \mathbb{R}^p$, and suppose that alongside the covariates, there is information on group levels $\mathcal{M} = \{1, \ldots, m\}$ for each unit i. That is, every observation for unit i is known to belong to a specific group j, and we write $\mathbf{x}_i^{(j)}$ to indicate this. Let n_j denote the sample size for cluster j, and the overall sample size be $n = \sum_{j=1}^{m} n_j$. When modelled linearly with the responses $y_i^{(j)}$, the model is known as a multilevel (linear) model, although it is known by many other names: random-effects models, random coefficient models, hierarchical models, and so on. As this model is seen as an extension of linear models, applications are plenty, especially in research designs for which the data varies at more than one level.

Consider a functional ANOVA decomposition of the regression function as follows:

$$f(\mathbf{x}_i^{(j)}, j) = \alpha + f_1(\mathbf{x}_i^{(j)}) + f_2(j) + f_{12}(\mathbf{x}_i^{(j)}, j). \tag{4.2}$$

To mimic the standard linear multilevel model, assume $f_1 \in \mathcal{F}_1$ the Pearson RKHS, $f_2 \in \mathcal{F}_2$ the centred canonical RKHS, and $f_{12} \in \mathcal{F}_{12} = \mathcal{F}_1 \otimes \mathcal{F}_2$, the tensor product

space of $\mathcal{F}_1$ and $\mathcal{F}_2$. As we know, α is the overall intercept, and the varying intercepts are given by the function f_2. While f_1 is the (main) linear effect of the covariates, f_{12} provides the varying linear effect of the covariates by each group. The I-prior for $f - \alpha$ is assumed to lie in the function space $\mathcal{F} - \alpha$, which is an ANOVA RKKS with kernel

$$h_\lambda\big((\mathbf{x}_i^{(j)}, j), (\mathbf{x}_i^{(j')}, j')\big) = \lambda_1 h_1(\mathbf{x}_i^{(j)}, \mathbf{x}_{i'}^{(j')}) + \lambda_2 h_2(j, j') + \lambda_1\lambda_2 h_1(\mathbf{x}_i^{(j)}, \mathbf{x}_{i'}^{(j')}) h_2(j, j'),$$

with h_1 the centred canonical kernel and h_2 the Pearson kernel. The reason for not including an RKHS of constant functions in $\mathcal{F}$ is because the overall intercept is usually simpler to estimate as an external parameter (see Section 4.2.1).

We can show that the regression function (4.2) corresponds to the standard way of writing the multilevel model,

$$f(\mathbf{x}_i^{(j)}, j) = \beta_0 + \mathbf{x}_i^{(j)\top}\boldsymbol{\beta}_1 + \beta_{0j} + \mathbf{x}_i^{(j)\top}\boldsymbol{\beta}_{1j}. \tag{4.3}$$

and determine the prior distributions on $(\beta_{0j}, \boldsymbol{\beta}_{1j}^\top)^\top \in \mathbb{R}^{p+1}$. For the interested reader, the details are in Appendix F.1. The standard multilevel random effects assumption is that $(\beta_{0j}, \boldsymbol{\beta}_{1j}^\top)^\top$ is normally distributed with mean zero and covariance matrix $\boldsymbol{\Phi}$. In total, there are $p + 1$ regression coefficients and $(p + 1)(p + 2)/2$ covariance parameters in Φ to be estimated. In contrast, the I-prior model is parameterised by only two RKKS scale parameters—one for $\mathcal{F}_1$ and one for $\mathcal{F}_2$—and the error precision ψ. While the estimation procedure for $\boldsymbol{\Phi}$ in the standard multilevel model can result in non-positive covariance matrices, the I-prior model has the advantage that positive definiteness is taken care of automatically[1].

As a remark, the following regression functions are nested

- $f(\mathbf{x}_i^{(j)}, j) = \alpha + f_1(\mathbf{x}_i^{(j)}) + f_2(j)$ (random intercept model);
- $f(\mathbf{x}_i^{(j)}, j) = \alpha + f_1(\mathbf{x}_i^{(j)})$ (linear regression model);
- $f(\mathbf{x}_i^{(j)}, j) = \alpha + f_2(j)$ (ANOVA model);
- $f(\mathbf{x}_i^{(j)}, j) = \alpha$ (intercept only model),

and thus one may compare likelihoods to ascertain the best fitting model. In addition, one may add flexibility to the model in two possible ways:

1. **More than two levels**. The model can be easily adjusted to reflect the fact that that the data is structured in a hierarchy containing three or more levels. For the three level case, let the indices $j \in \{1, \ldots, m_1\}$ and $k \in \{1, \ldots, m_2\}$ denote the

[1] By virtue of the estimate of the regression function belonging to $\mathcal{F}_n$, an RKHS with a positive definite kernel equal to the Fisher information for f.

two levels, and simply decompose the regression function accordingly:

$$f(\mathbf{x}_i^{(j,k)}, j, k) = \alpha + f_1(\mathbf{x}_i^{(j,k)}) + f_2(j) + f_3(k) + f_{12}(\mathbf{x}_i^{(j,k)}, j) + f_{13}(\mathbf{x}_i^{(j,k)}, k) \\ + f_{23}(j,k) + f_{123}(\mathbf{x}_i^{(j,k)}, j, k).$$

2. **Covariates not varying with levels**. Suppose now we would like to add covariates with a fixed effect to the model, i.e. covariates $\mathbf{z}_i^{(j)}$ which are not assumed to affect the responses differently in each group. The regression function would be:

$$f(\mathbf{x}_i^{(j)}, j, \mathbf{z}_j) = \alpha + f_1(\mathbf{x}_i^{(j)}) + f_2(j) + f_3(\mathbf{z}_i^{(j)}) + f_{12}(\mathbf{x}_i^{(j)}, j).$$

This can be seen as a limited functional ANOVA decomposition of f.

4.1.3 Longitudinal modelling

Longitudinal or panel data observes covariate measurements $x_i \in \mathcal{X}$ and responses $y_i(t) \in \mathbb{R}$ for individuals $i = 1, \dots, n$ across a time period $t \in \{1, \dots, T\} =: \mathcal{T}$. Often, the time indexing set $\mathcal{T}$ may be unique to each individual i, so measurements for unit i happens across a time period $\{t_{i1}, \dots, t_{iT_i}\} =: \mathcal{T}_i$—this is known as an unbalanced panel. It is also possible that covariate measurements vary across time too, so appropriately they are denoted $x_i(t)$. For example, $x_i(t)$ could be repeated measurements of the variable x_i at time point $t \in \mathcal{T}_i$. The relationship between the response variables $y_i(t)$ at time $t \in \mathcal{T}_i$ is captured through the equation

$$y_i(t) = f\big(i, x_i, t\big) + \epsilon_i(t)$$

where the distribution of $\boldsymbol{\epsilon}_i = \big(\epsilon_i(t_{i1}), \dots, \epsilon_i(t_{iT_i})\big)^\top$ is Gaussian with mean zero and covariance matrix $\boldsymbol{\Psi}_i$. Assuming $\boldsymbol{\Psi}_i = \psi_i \mathbf{I}_{T_i}$ or even $\boldsymbol{\Psi}_i = \psi \mathbf{I}_{T_i}$ are perfectly valid choices, even though this seemingly ignores any time dependence between the observations. In reality, the I-prior induces time dependence of the observations via the kernels in the prior covariance matrix for f. Additionally, the random vectors $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_{i'}$ are assumed to be independent for any two distinct $i, i' \in \{1, \dots, n\}$.

Motivated by a functional ANOVA decomposition, we obtain

$$f(i, x_i, t) = \alpha + f_1(i) + f_2(x_i) + f_3(t) + f_{13}(i,t) + f_{23}(x_i, t) + f_{12}(i, x_i) \\ + f_{123}(i, x_i, t) \tag{4.4}$$

where α is an overall constant, and each of the ANOVA component functions belongs to the appropriate (tensor product) space as described in Section 2.5.3 (p. **??**). $\mathcal{F}_1$ is the Pearson RKHS, but choices for $\mathcal{F}_2$ and $\mathcal{F}_3$ are plentiful. In fact, any of the

RKHS/RKKS described in Chapter 3 can be used to either model a linear dependence (canonical RKHS), nominal dependence (Pearson RKHS), polynomial dependence (polynomial RKKS) or smooth dependence (fBm or SE RKHS) on the x_i's and t's on f.

4.1.4 Classification

We describe a naïve classification model using I-priors. Here, the responses are categorical $y_i \in \{1, \dots, m\} =: \mathcal{M}$, and additionally, write $\mathbf{y}_{i\cdot} = (y_{i1}, \dots, y_{im})^\top$ where the class responses y_{ij} equal one if individual i's response category is $y_i = j$, and zero otherwise. In other words, there is exactly a single '1' at the j'th position in the vector $\mathbf{y}_{i\cdot}$, and zeroes everywhere else. For $j = 1, \dots, m$, we model

$$\begin{aligned} y_{ij} &= \alpha + \overbrace{\alpha_j + f_j(x_i)}^{f(x_i, j)} + \epsilon_{ij} \\ (\epsilon_{i1}, \dots, \epsilon_{im})^\top &\overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}). \end{aligned} \tag{4.5}$$

The idea here being that we attempt to model the class responses y_{ij} using class-specific regression functions f_j, and the class responses are assumed to be independent among individuals, but may or may not be correlated among classes for each individual. The class correlations are manifest themselves in the variance of the errors $\boldsymbol{\Psi}^{-1}$, which is an $m \times m$ matrix.

Denote the regression function f in (4.5) on the set $\mathcal{X} \times \mathcal{M}$ as $f(x_i, j) = \alpha_j + f_j(x_i)$. This regression function corresponds to an ANOVA decomposition of the spaces $\mathcal{F}_\mathcal{M}$ and $\mathcal{F}_\mathcal{X}$ of functions over $\mathcal{M}$ and $\mathcal{X}$ respectively. That is, $\mathcal{F} = \mathcal{F}_\mathcal{M} \oplus (\mathcal{F}_\mathcal{M} \otimes \mathcal{F}_\mathcal{X})$ is a decomposition into the main effects of 'class', and an interaction effect of the covariates for each class. Let $\mathcal{F}_\mathcal{M}$ and $\mathcal{F}_\mathcal{X}$ be RKHSs respectively with kernels $a : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ and $b_\eta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then, the ANOVA RKKS $\mathcal{F}$ possesses the reproducing kernel $h_\eta : (\mathcal{X} \times \mathcal{M})^2 \to \mathbb{R}$ as defined by

$$h_\eta\big((x, j), (x', j')\big) = a(j, j') + a(j, j') b_\eta(x, x'). \tag{4.6}$$

The kernel b_η may be any of the kernels described in this thesis, ranging from the linear kernel, to the fBm kernel, or even an ANOVA kernel. Choices for $a : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ include

1. **The Pearson kernel** (as defined in Definition 2.33, p. 63). With $J \sim \mathrm{P}$, a probability measure over $\mathcal{M}$,

$$a(j, j') = \frac{\delta_{jj'}}{\mathrm{P}(J = j)} - 1.$$

2. **The centred identity kernel**. With δ denoting the Kronecker delta function,

$$a(j, j') = \delta_{jj'} - 1/m.$$

The purpose of either of these kernels is to contribute to the class intercepts α_j, and to associate a regression function in each class. The only difference between the two is the inverse probability weighting per class that is applied in the Pearson kernel, but not in the identity kernel.

With $f \in \mathcal{F}$ (the RKKS with kernel h_η), it is straightforward to assign an I-prior on f. It is in fact

$$\begin{gathered} f(x_i, j) = \sum_{j'=1}^{m} \sum_{i'=1}^{n} a(j, j')\big(1 + b_\eta(x_i, x_{i'})\big) w_{i'j'} \\ (w_{i'1}, \ldots, w_{i'm})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}) \end{gathered} \tag{4.7}$$

assuming a zero prior mean $f_0(x, j) = 0$. The model then classifies the i'th data point to class j if $\hat{y}_{ij} = \max(\hat{y}_{i1}, \ldots, \hat{y}_{im})$, where $\hat{y}_{ik} = \hat{\alpha} + \hat{f}(x_i, k)$, the prediction for the k'th component of y_i.

There are several drawbacks to using the model described above. Unlike in the case of continuous response variables, the normal I-prior model is highly inappropriate for categorical responses. For one, it violates the normality and homoscedasticity assumptions of the errors. For another, predicted values may be out of the range $[0, m]$ and thus poorly calibrated. Furthermore, it would be more suitable if the class probabilities—the probability of an observation belonging to a particular class—were also part of the model. In Chapter 5, we propose an improvement to this naïve I-prior classification model by considering a probit-like transformation of the regression functions.

4.1.5 Smoothing models

Single- and multi-variable smoothing models can be fitted under the I-prior methodology using the fBm RKHS. In standard kernel based smoothing methods, the squared exponential kernel is often used, and the corresponding RKHS contains analytic functions. There are several attractive properties of using the fBm RKHS, and for one-dimensional smoothing, these are discussed below.

Assume that, up to a constant, the regression function lies in the scaled, centred fBm RKHS $\mathcal{F}$ of functions over $\mathcal{X} \equiv \mathbb{R}$ with Hurst index 1/2. Thus, with a centring with respect to the empirical distribution P_n of $\{x_1, \ldots, x_n\}$ and using the absolute norm on

$\mathbb{R}$, $\mathcal{F}$ has kernel

$$h_\lambda(x, x') = \frac{\lambda}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \big(|x - x_i| + |x' - x_j| - |x - x'| - |x_i - x_j|\big).$$

As proven by van der Vaart and van Zanten (2008, Section 10), $\mathcal{F}$ contains absolutely continuous functions possessing a square integrable weak derivative satisfying $f(0) = 0$. The norm is given by $\|f\|_{\mathcal{F}}^2 = \int \dot{f}^2 \, \mathrm{d}x$. The posterior mean of f based on an I-prior is then a (one-dimensional) smoother for the data. For f of the form $f = \sum_{i=1}^{n} h(\cdot, x_i) w_i$, i.e. $f \in \mathcal{F}_n$, the finite subspace of $\mathcal{F}$ as in Section 3.4 (p. 87), then Bergsma (2018) shows that f can be represented as

$$f(x) = \int_{-\infty}^{x} \beta(t) \, \mathrm{d}t \tag{4.8}$$

where

$$\beta(t) = \sum_{i : x_i \le t} w_i = \frac{f(x_{i_t+1}) - f(x_{i_t})}{x_{i_t+1} - x_{i_t}} \tag{4.9}$$

with $i_t = \max_{x_i \le t} i$. Under the I-prior with an iid assumption on the errors, the w_i's are zero mean normal random variables with variance ψ, so that β as defined above is an ordinary Brownian bridge with respect to the empirical distribution P_n. The I-prior for f is piecewise linear with knots at $x_1, \ldots, x_n$, and the same holds true for the posterior mean. The implication is that the I-prior automatically adapts to irregularly spaced x_i: in any region where there are no observations, the resulting smoother is linear. This is explained by the reduced Fisher information about the derivative of the regression curve in regions with no observation.

In Bergsma (2018), it is shown that the covariance function for β is

$$\mathrm{Cov}\big[\beta(x), \beta(x')\big] = n\big(\min\{\mathrm{P}_n(X < x), \mathrm{P}_n(X_n < x')\} - \mathrm{P}_n(X < x)\,\mathrm{P}_n(X_n < x')\big)$$

From this, notice that $\mathrm{Var}\big[\beta(x)\big] = \mathrm{P}_n(X_n < x)\big(1 - \mathrm{P}_n(X_n < x)\big)$, which shows an automatic boundary correction: close to the boundary there is little Fisher information on the derivative of the regression function $\beta(x)$, so the prior variance is small. This will lead to more shrinkage of the posterior derivative of f towards the derivative of the prior mean f_0.

Another advantage of the I-prior methodology is the ability to fit single or multi-dimensional smoothing models with just two parameters to be estimated: the RKHS scale parameter λ and the error precision $\boldsymbol{\Psi}$. The Hurst parameter $\gamma \in (0, 1)$ of the fBm RKHS can also be treated as a free parameter for added flexibility, but for most practical applications, we find that the default setting of $\gamma = 1/2$ performs sufficiently well.

Remark 4.3. From (4.8), the prior process for f is thus an integrated Brownian bridge. This shows a close relation with cubic spline smoothers, which can be interpreted as the posterior mean when the prior is an integrated Wiener process (Wahba, 1990). Unlike I-priors however, cubic spline smoothers do not have automatic boundary corrections, and typically the additional assumption is made that the smoothing curve is linear at the boundary knots.

4.1.6 Regression with functional covariates

Suppose that we have functional covariates x in the real domain, and that $\mathcal{X}$ is a set of differentiable functions. If so, it is reasonable to assume that $\mathcal{X}$ is a Hilbert-Sobolev space with inner product

$$\langle x, x' \rangle_{\mathcal{X}} = \int \dot{x}(t)\dot{x}'(t)\,\mathrm{d}t,$$

so that we may apply the linear, fBm or any other kernels which make use of inner products by making use of the polarisation identity. Furthermore, let $z \in \mathbb{R}^T$ be the discretised realisation of the function $x \in \mathcal{X}$ at regular intervals $t = 1, \ldots, T$. Then

$$\langle x, x' \rangle_{\mathcal{X}} \approx \sum_{t=1}^{T-1} (z_{t+1} - z_t)(z'_{t+1} - z'_t).$$

For discretised observations at non-regular intervals $\{t_1, \ldots, t_T\}$ then a more general formula to the above one might be used, for instance,

$$\langle x, x' \rangle_{\mathcal{X}} \approx \sum_{i=1}^{T-1} \frac{(z_{t_{i+1}} - z_{t_i})(z'_{t_{i+1}} - z'_{t_i})}{t_{i+1} - t_i}.$$

4.2 Estimation

After selecting a RKHS/RKKS $\mathcal{F}$ of functions over $\mathcal{X}$ suitable for the regression problem at hand, one then proceeds to estimate the posterior distribution of the regression function. The I-prior model (1.1) subject to (1.2) and $f \in \mathcal{F}$ has the simple and convenient representation

$$\begin{aligned} y_i &= \alpha + \overbrace{f_0(x_i) + \sum_{k=1}^{n} h_\eta(x_i, x_k) w_k}^{f(x_i)} + \epsilon_i \\ (\epsilon_1, \ldots, \epsilon_n)^\top &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}^{-1}) \\ (w_1, \ldots, w_n)^\top &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}), \end{aligned} \tag{4.10}$$

where $f_0 : \mathcal{X} \to \mathbb{R}$ is a function chosen a priori representing the 'best guess' of f, and the dependence of the kernel of $\mathcal{F}$ on parameters η is emphasised through the subscript in $h_\eta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

The parameters of the I-prior model are collectively denoted by $\theta = \{\alpha, \eta, \boldsymbol{\Psi}\}$. Given θ and a prior choice for f_0, the posterior regression function is determined solely by the posterior distribution of the w_i's. Using standard multivariate normal results, one finds that the posterior distribution for $\mathbf{w} := (w_1, \ldots, w_n)^\top$ is $\mathbf{w}|\mathbf{y} \sim \mathrm{N}_n(\tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$, where

$$\tilde{\mathbf{w}} = \boldsymbol{\Psi}\mathbf{H}_\eta \mathbf{V}_y^{-1}(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{f}_0) \quad \text{and} \quad \tilde{\mathbf{V}}_w = \left(\mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}\right)^{-1} = \mathbf{V}_y^{-1}, \tag{4.11}$$

using the familiar notation that we introduced in Section 1.4. For a derivation, see Appendix G.1 (p. 285). By linearity, the posterior distribution for f is also normal.

In each modelling scenario, there are a number of kernel parameters η that need to be estimated from the data. Assuming that the covariate space is $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$, and there is an ANOVA like decomposition of the function space $\mathcal{F}$ into its constituents spaces $\mathcal{F}_1, \ldots, \mathcal{F}_p$, then at the very least, there are p scale parameters $\lambda_1, \ldots, \lambda_p$ for each of the RKHSs. Depending on the RKHS used, there could be more kernel parameters that need to be optimised, for instance, the Hurst index for the fBm RKHS, the lengthscale for the SE RKHS, and/or the offset for the polynomial RKKS. However, these may be treated as fixed parameters as well.

The following subsections describe possible estimation procedures for the hyperparameters of the model. Henceforth, for simplicity, the following additional standing assumptions are imposed on the I-prior model (4.10):

A1 **Centred responses**. Set $\alpha = 0$ and replace the responses by their centred versions $y_i \mapsto \tilde{y}_i = y_i - \frac{1}{n}\sum_{i=1}^n$.

A2 **Zero prior mean**. Assume a zero prior mean $f_0(x) = 0$ for all $x \in \mathcal{X}$.

A3 **Iid errors**. Assume identical and independent error random variables, i.e. $\boldsymbol{\Psi} = \psi\mathbf{I}_n$.

Assumptions A1 and A2 are motivated by the discussion in Section 4.2.1. Although assumption A3 is not strictly necessary, it is often a reasonable one and one that simplifies the estimation procedure greatly.

4.2.1 The intercept and the prior mean

In most statistical models, an intercept is a necessary inclusion which aids interpretation. In the context of the I-prior model (4.10), a lack of an intercept would fail to account for the correct locational shift of the regression function along the y-axis. Further, when

zero-mean functions are considered, the intercept serves as being the 'grand mean' value of the responses.

The handling of an intercept to the regression model may be viewed in one of two ways. The first is to view it as a function belonging to the RKHS of constant functions $\mathcal{F}_\emptyset$, and thereby tensor summing this space to $\mathcal{F}$. The second is to simply treat the intercept as a parameter of the model to be estimated. In the polynomial and ANOVA RKKSs, we saw that an intercept is naturally induced by the inclusion of a RKHS of constant functions in their construction. In any of the other RKHSs described in Chapter 2, an intercept would need to be added separately. These two methods convey slightly different interpretations of the intercept: in the first method, the intercept is shrunk by an I-prior, while in the second, it is not. Estimation is also entirely different for the two methods.

In the first method, the intercept-less RKHS/RKKS $\mathcal{F}$ with kernel h is made to include an intercept by modifying the kernel to be $1 + h$. The intercept will then be implicitly taken care of without having dealt with it explicitly. However, it can be obtained by realising that for $\alpha \in \mathcal{F}_\emptyset$ the RKHS of constant functions, then $\alpha = \sum_{i=1}^n w_i$.

On the other hand, consider the intercept as a parameter α to be estimated. Obtaining an estimate α using a likelihood-based argument is rather simple. From (4.10), $\mathrm{E}[y_i] = \alpha + f_0(x_i)$ for all $i = 1, \dots, n$, so the maximum likelihood (ML) estimate for $\mathrm{E}[y]$ is its sample mean $\bar{y} = \frac{1}{n}\sum_{i=1} y_i$, and hence the ML estimate for α is $\hat{\alpha} = \bar{y} - \frac{1}{n}\sum_{i=1}^n f_0(x_i)$. Thus, assumption A1 therefore implies that the ML estimate for the intercept is the sample mean of the responses.

4.2.2 Direct optimisation

Under assumptions A1–A3, a direct optimisation of the parameters $\theta = \{\eta, \boldsymbol{\Psi} = \psi\mathbf{I}_n\}$ using the log-likelihood of θ is straightforward to implement. Denote $\boldsymbol{\Sigma}_\theta := \psi\mathbf{H}_\eta^2 + \psi^{-1}\mathbf{I}_n = \mathbf{V}_y$. From (4.10), the marginal log-likelihood of θ is given by

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y}|\mathbf{w}, \theta) p(\mathbf{w}|\theta) \,\mathrm{d}\mathbf{w} \\ &= -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}\tilde{\mathbf{y}}^\top \boldsymbol{\Sigma}_\theta^{-1}\tilde{\mathbf{y}}, \end{aligned} \tag{4.12}$$

which is the log-likelihood of a zero mean multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}_\theta$. This closed-form expression of the integral (4.12) stems from the fact that the (conditional) likelihood and the I-prior are both Gaussian. Note that the term 'marginal' refers to the fact that we are averaging out the random function represented by $\mathbf{w}$.

For Gaussian processes, direct optimisation is typically done using the conjugate gradients method with a Cholesky decomposition on the covariance kernel to maintain stability (Rasmussen and Williams, 2006), but we opt for an eigendecomposition of the kernel matrix $\mathbf{H}_\eta = \mathbf{V}\,\text{diag}(u_1, \ldots, u_n)\,\mathbf{V}^\top$ instead. Further, since $\mathbf{H}_\eta$ is a symmetric matrix, we have that $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_n$, and thus

$$\begin{aligned}\mathbf{V}_y &= \psi\mathbf{V}\,\text{diag}(u_1^2, \ldots, u_n^2)\,\mathbf{V}^\top + \psi^{-1}\mathbf{V}\mathbf{V}^\top \\ &= \mathbf{V}\,\text{diag}(\psi u_1^2 + \psi^{-1}, \ldots, \psi u_n^2 + \psi^{-1})\,\mathbf{V}^\top\end{aligned}$$

for which the inverse and log-determinant is easily obtainable. To be explicit, the log-likelihood is given by

$$\begin{aligned}L(\theta) = &-\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}\log(\psi u_i^2 + \psi^{-1}) \\ &-\frac{1}{2}\tilde{\mathbf{y}}^\top\mathbf{V}\,\text{diag}\left(\frac{1}{\psi u_1^2 + \psi^{-1}}, \ldots, \frac{1}{\psi u_n^2 + \psi^{-1}}\right)\mathbf{V}^\top\tilde{\mathbf{y}}\end{aligned} \tag{4.13}$$

This method is relatively robust to numerical instabilities and is better at ensuring positive definiteness of the covariance kernel. The direct optimisation method can be prone to local optima, in which case repeating the optimisation at different starting points and choosing the one which yields the highest likelihood is one way around this. On a practical note, parameters are best transformed so that optimisation of these parameters are done on an unrestricted scale (e.g. $\log\psi$).

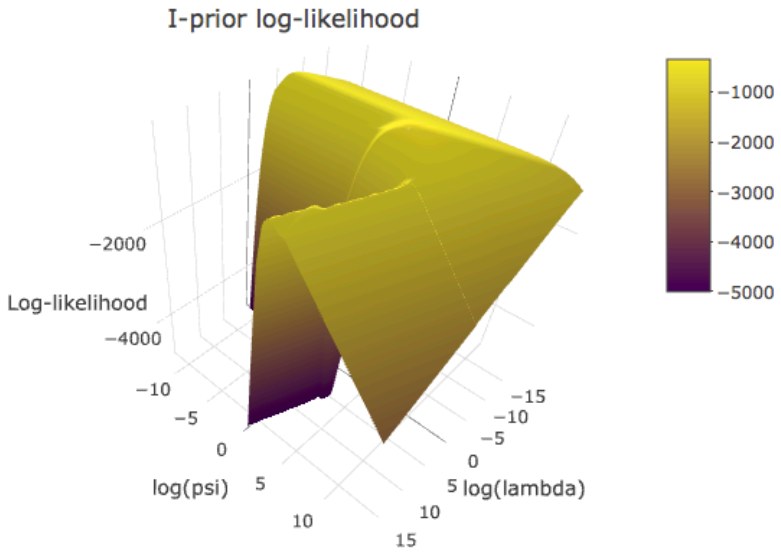


Figure 4.1: A typical log-likelihood surface plot of I-prior models, in which there are two ridges. The maximum occurs along one of the two ridges, or sometimes at the intersection. Clearly, different initialisations can lead optimisation algorithms to either ridge and possibly converge to a local optima.

Let $\mathbf{U}$ be the Fisher information matrix for $\theta \in \mathbb{R}^q$. Standard calculations (Appendix C.1) show that under the marginal distribution $\tilde{\mathbf{y}} \sim \mathrm{N}_n\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right)$, the (i,j)'th coordinate of $\mathbf{U}$ is

$$u_{ij} = \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_\theta^{-1}\frac{\partial \boldsymbol{\Sigma}_\theta}{\partial \theta_i}\boldsymbol{\Sigma}_\theta^{-1}\frac{\partial \boldsymbol{\Sigma}_\theta}{\partial \theta_j}\right), \qquad i,j = 1,\ldots,q, \tag{4.14}$$

where the derivative of a matrix with respect to a scalar is the element-wise derivative of the matrix. With $\hat{\theta}$ denoting the ML estimate for θ, under suitable conditions, $\sqrt{n}(\hat{\theta} - \theta)$ has an asymptotic multivariate normal distribution with mean zero and covariance matrix $\mathbf{U}^{-1}$ (Casella and R. L. Berger, 2002). In particular, the standard error for θ_k is the k'th diagonal element of $\mathbf{U}^{-1/2}$.

4.2.3 Expectation-maximisation algorithm

Evidently, the model in (4.10) resembles a random-effects model, for which the EM algorithm is easily employed to estimate its hyperparameters. Assume A1–A3 holds. By treating the complete data as $\{\mathbf{y}, \mathbf{w}\}$ and the w_i's as "missing", the t'th iteration of the E-step entails computing

$$\begin{aligned} Q(\theta) &= \mathrm{E}_{\mathbf{w}}\left(\log p(\mathbf{y}, \mathbf{w}|\theta)\middle|\, \mathbf{y}, \theta^{(t)}\right) \\ &= \mathrm{E}_{\mathbf{w}}\left(\text{const.} + \cancel{\frac{n}{2}\log\psi} - \frac{\psi}{2}\|\tilde{\mathbf{y}} - \mathbf{H}_\eta \mathbf{w}\|^2 - \cancel{\frac{n}{2}\log\psi} - \frac{\psi^{-1}}{2}\|\mathbf{w}\|^2 \,\middle|\, \mathbf{y}, \theta^{(t)}\right) \\ &= \text{const.} - \frac{\psi}{2}\tilde{\mathbf{y}}^\top\tilde{\mathbf{y}} - \frac{1}{2}\operatorname{tr}\left(\overbrace{(\psi\mathbf{H}_\eta^2 + \psi^{-1}\mathbf{I}_n)}^{\boldsymbol{\Sigma}_\theta}\tilde{\mathbf{W}}^{(t)}\right) + \psi\tilde{\mathbf{y}}^\top\mathbf{H}_\eta\tilde{\mathbf{w}}^{(t)}, \end{aligned} \tag{4.15}$$

where $\tilde{\mathbf{w}}^{(t)} = \mathrm{E}[\mathbf{w}|\mathbf{y}, \theta^{(t)}]$ and $\tilde{\mathbf{W}}^{(t)} = \mathrm{E}[\mathbf{w}\mathbf{w}^\top|\mathbf{y}, \theta^{(t)}]$ are the first and second posterior moments of $\mathbf{w}$ calculated at the tth EM iteration. These can be computed directly from (4.11), substituting for $\theta^{(t)} = \{\eta^{(t)}, \psi^{(t)}\}$ as appropriate.

The M-step assigns $\theta^{(t+1)}$ the value of θ which maximises the Q function above. This boils down to solving the first order conditions

$$\frac{\partial Q}{\partial \eta} = -\frac{1}{2}\operatorname{tr}\left(\frac{\partial \boldsymbol{\Sigma}_\theta}{\partial \eta}\tilde{\mathbf{W}}^{(t)}\right) + \psi\tilde{\mathbf{y}}^\top\frac{\partial \mathbf{H}_\eta}{\partial \eta}\tilde{\mathbf{w}}^{(t)} \tag{4.16}$$

$$\frac{\partial Q}{\partial \psi} = -\frac{1}{2}\tilde{\mathbf{y}}^\top\tilde{\mathbf{y}} - \operatorname{tr}\left(\frac{\partial \boldsymbol{\Sigma}_\theta}{\partial \psi}\tilde{\mathbf{W}}^{(t)}\right) + \tilde{\mathbf{y}}^\top\mathbf{H}_\eta\tilde{\mathbf{w}}^{(t)} \tag{4.17}$$

equated to zero. As $\partial\boldsymbol{\Sigma}_\theta/\partial\psi = \mathbf{H}_\eta^2 - \psi^{-2}\mathbf{I}_n$, the solution to (4.17) for ψ admits a closed form given values for η:

$$\psi^{(t+1)} = \left\{\frac{\operatorname{tr}\tilde{\mathbf{W}}^{(t)}}{\tilde{\mathbf{y}}^\top\tilde{\mathbf{y}} + \operatorname{tr}(\mathbf{H}_\eta^2\tilde{\mathbf{W}}^{(t)}) - 2\tilde{\mathbf{y}}^\top\mathbf{H}_\eta\tilde{\mathbf{w}}^{(t)}}\right\}^{1/2}. \tag{4.18}$$

We use this fact to form a sequential updating scheme $\eta^{(t)} \to \psi^{(t+1)} \to \eta^{(t+1)} \to \cdots$, and this form of the EM algorithm is known as the *expectation conditional maximisation* algorithm (Meng and Rubin, 1993). Now, the solution to (4.16) can also be found in closed-form given values ψ, for many models, but in general, this is not the case. In cases where closed-form solutions do exist for η, then it is just a matter of iterating the update equations until a suitable convergence criterion is met (e.g. no more sizeable increase in successive log-likelihood values). In cases where closed-form solutions do not exist for η, the Q function is again optimised with respect to η using the gradient-based algorithms.

In our experience, the EM algorithm is more stable than direct maximisation, in the sense that the EM steps increase the likelihood in a gentle manner that prevents sudden explosions of the likelihood. In contrast, the search direction using gradient-based methods can grow the likelihood too quickly and potentially causes numerical errors to creep in. As such, the EM is especially suitable if there are many scale parameters to estimate, but on the flip side, it is typically slow to converge. The **iprior** package provides a method to automatically switch to the direct optimisation method after running several EM iterations. This then combines the stability of the EM with the speed of direct optimisation.

4.2.4 Markov chain Monte Carlo methods

For completeness, it should be mentioned that a full Bayesian treatment of the model is possible, with additional priors on the set of hyperparameters. Markov chain Monte Carlo (MCMC) methods can then be employed to sample from the posteriors of the hyperparameters, with point estimates obtained using the posterior mean or mode, for instance. Additionally, the posterior distribution encapsulates the uncertainty about the parameter, for which inference can be made. Posterior sampling can be done using Gibbs-based methods in **WinBUGS** (Lunn et al., 2000) or **JAGS** (Plummer, 2003), and both have interfaces to R via **R2WinBUGS** (Sturtz et al., 2005) and **runjags** (Denwood, 2016) respectively. Hamiltonian Monte Carlo (HMC) sampling is also a possibility, and the Stan project (Carpenter et al., 2017) together with the package **rstan** (Stan Development Team, 2016) makes this possible in R.

On the software side, all of these MCMC packages require the user to code the model individually, and we are not aware of the existence of MCMC-based packages which are able to estimate GPR models. This makes it inconvenient for GPR and I-prior models, because in addition to the model itself, the kernel functions need to be coded as well and ensuring computational efficiency would be a difficult task.

Speaking of efficiency, it is more advantageous to marginalise the I-prior and work with the marginal model (4.12), rather than the hierarchical specification (4.10). The reason for this is that the latter model has a parameter space whose dimension is $O(n)$, while the former only samples the hyperparameters. Note that the marginal model (4.12) cannot then be sampled efficiently using a Gibbs procedure as the Gibbs conditionals are not of closed-form. Instead, Hamiltonian MC should be used, which does not depend on model conjugacy.

4.2.5 Comparison of estimation methods

Consider a one-dimensional smoothing example, for which $n = 150$ data pairs (y_i, x_i) have been generated according to the relationship

$$y_i = \text{const.} + \overbrace{0.35\,\phi(x_i|1, 0.8^2) + 0.65\,\phi(x_i|4, 1.5^2) + \mathbb{1}(x_i > 4.5)\,e^{1.25(x_i-4.5)}}^{f_{\text{true}}(x_i)} + \epsilon_i, \tag{4.19}$$

where $\phi(\cdot|\mu, \sigma^2)$ is the probability density function of a normal distribution with mean μ and variance σ^2. The observed y_i's are thought to be noisy versions of the true points, in which ϵ_i follows an indescript, not necessarily normal, distribution. The predictors $x_1, \ldots, x_n$ have been sampled roughly from the interval $(-1, 6)$, and the sampling was intentionally not uniform so that there is slight sparsity in the middle. Figure 4.2 plots the sampled points and the true regression function.

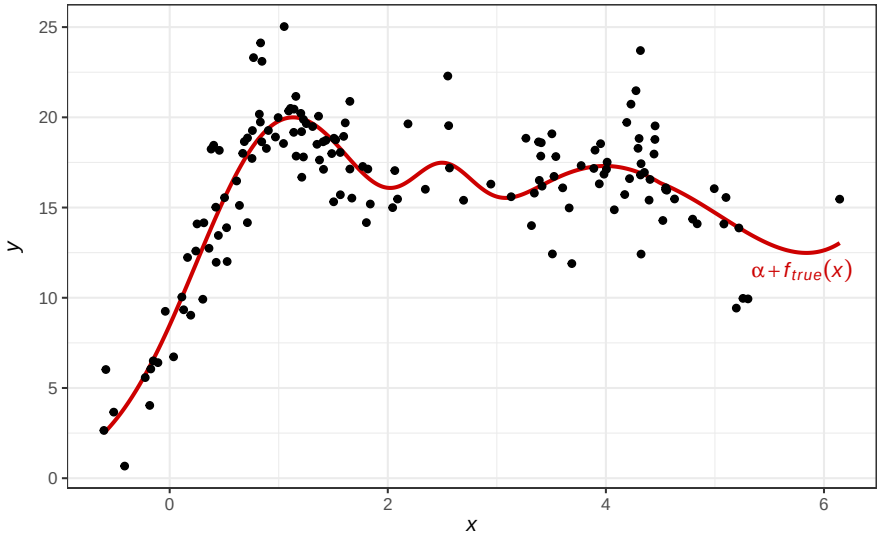


Figure 4.2: A plot of the sampled data points according to equation (4.19), with the true regression function superimposed.

We attempt to estimate f_{true} by a function f belonging to the fBm-0.5 RKHS $\mathcal{F}_\lambda$, with an I-prior on f. There are two parameters that need to be estimated: the scale parameter λ for the fBm-0.5 RKHS, and the error precision ψ. These can be estimated using the maximum likelihood methods described above, namely by direct optimisation using a quasi-Newton algorithm, and the EM algorithm. These two methods are implemented in the **iprior** package. A full Bayesian treatment is possible, and we use the **rstan** implementation of Stan to perform Hamiltonian Monte Carlo sampling of the posterior densities. A vague prior choice for λ and ψ are prescribed, namely

$$\lambda, \psi \overset{\text{iid}}{\sim} \text{N}_{\geq 0}(0, 100),$$

where $\text{N}_+(0, \sigma^2)$ represents the *folded-normal* distribution[2,3]. We have also set an improper prior density $p(\alpha) \propto$ const. for the intercept. The advantage of HMC is that efficiency is not dictated by conjugacy, so there is freedom to choose any appropriate prior choice on the parameters.

Table 4.1: Table comparing the estimated parameter values, (marginal) log-likelihood values, and also time taken for the three estimation methods.

	Direct optimisation	EM algorithm	Hamiltonian MC
Intercept (α)	16.1 (0.35)	16.1 (0.35)	16.1 (0.17)
Scale (λ)	5.01 (1.23)	5.01 (1.26)	5.61 (1.42)
Precision (ψ)	0.236 (0.03)	0.236 (0.03)	0.237 (0.03)
Log-density	-339.7	-339.7	-341.1
Predictive RMSE	0.574	0.575	0.582
Iterations	12	266	2000
Time taken (s)	0.96	3.65	232

Table 4.1 tabulates the estimated parameter values, (marginal) log-likelihood values, and also time taken for the three estimation methods. The three methods concur on the estimated parameter values, although the scale parameter has been estimated slightly differently, which is possibly attributed to the effect of the prior for λ. The resulting log-likelihood value for the Bayesian method is lower than the ML methods, which also took the longest to compute. Although the EM algorithm took longer than the direct optimisation method to compute, the time taken per iteration is significantly shorter than one Newton iteration.

[2]The random variable $X \sim \text{N}_+(\mu, \sigma^2)$ has the density $p(x) = \phi(x|\mu, \sigma^2)\,\mathbb{1}(x \geq 0)$.

[3]Note that a single scale parameter λ is not identified in sign, and is thus constrained to the positive reals. This is applicable in both likelihood-based and Bayesian methods.

4.3 Computational considerations and implementation

Computational complexity for estimating I-prior models (and in fact, for GPR in general) is dominated by the inversion (by way of eigendecomposition in our case) of the $n \times n$ matrix $\mathbf{\Sigma}_\theta = \psi\mathbf{H}_\eta^2 + \psi^{-1}\mathbf{I}_n$, which scales as $O(n^3)$ in time. For the direct optimisation method, this matrix inversion is called when computing the log-likelihood, and thus must be computed at each Newton step. For the EM algorithm, this matrix inversion appears when calculating $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{W}}$, the first and second posterior moments of the I-prior random effects. Furthermore, storage requirements for I-priors models are similar to that of GPR models, which is $O(n^2)$.

4.3.1 The Nyström approximation

The shared computational issues of I-prior and GPR models allow us to delve into machine learning literature, which is rich in ways to resolve these issue, as summarised by Quiñonero-Candela and Rasmussen (2005). One such method is to exploit low rank structures of kernel matrices. The idea is as follows. Let $\mathbf{Q}$ be a matrix with rank $q < n$, and suppose that $\mathbf{QQ}^\top$ can be used sufficiently well to represent the kernel matrix $\mathbf{H}_\eta$. Then

$$(\psi\mathbf{H}_\eta^2 + \psi^{-1}\mathbf{I}_n)^{-1} \approx \psi\left[\mathbf{I}_n - \mathbf{Q}\left(\left(\psi^2\mathbf{Q}^\top\mathbf{Q}\right)^{-1} + \mathbf{Q}^\top\mathbf{Q}\right)^{-1}\mathbf{Q}^\top\right],$$

obtained via the Woodbury matrix identity, is potentially a much cheaper operation which scales $O(nq^2)$: $O(q^3)$ to do the inversion, and $O(nq)$ to do the multiplication (because typically the inverse is premultiplied to a vector). When using the linear kernel for a low-dimensional covariate then the above method is exact ($\mathbf{Q} = \mathbf{X}$, where $\mathbf{X}$ is the design matrix). This fact is clearly demonstrated by the equivalence of the p-dimensional linear model implied by (4.1) with the n-dimensional I-prior model using the canonical RKHS. If $p \ll n$ then certainly using the linear representation is much more efficient.

However, other interesting kernels such as the fractional Brownian motion (fBm) kernel or the squared exponential kernel results in kernel matrices which are full rank. An approximation to the kernel matrix using a low-rank matrix is the Nyström method (Williams and Seeger, 2001). The theory has its roots in approximating eigenfunctions, but this has since been adopted to speed up kernel machines. The main idea is to obtain an (approximation to the true) eigendecomposition of $\mathbf{H}_\eta$ based on a small subset $q \ll n$ of the data points.

Let $\mathbf{H}_\eta = \mathbf{VUV}^\top = \sum_{i=1}^n u_i\mathbf{v}_i\mathbf{v}_i^\top$ be the (orthogonal) decomposition of the symmetric matrix $\mathbf{H}_\eta$. As mentioned, avoiding this expensive $O(n^3)$ eigendecomposition is desired, and this is achieved by selecting a subset $\mathcal{Q}$ of size q of the n data points $\{1, \ldots, n\}$, so that $\mathbf{H}_\eta$ may be approximated using the rank q matrix $\mathbf{H}_\eta \approx \sum_{i\in\mathcal{Q}} \tilde{u}_i\tilde{\mathbf{v}}_i\tilde{\mathbf{v}}_i^\top$.

Without loss of generality, reorder the rows and columns of $\mathbf{H}_\eta$ so that the data points indexed by $\mathcal{Q}$ are used first:

$$\mathbf{H}_\eta = \begin{pmatrix} \mathbf{A}_{q\times q} & \mathbf{B}_{q\times(n-q)} \\ \mathbf{B}^\top_{q\times(n-q)} & \mathbf{C}_{(n-q)\times(n-q)} \end{pmatrix}.$$

In other words, the data points indexed by $\mathcal{Q}$ forms the smaller $q \times q$ kernel matrix $\mathbf{A}$. Let $\mathbf{A} = \mathbf{V}_q\mathbf{U}_q\mathbf{V}_q^\top = \sum_{i=1}^q u_i^{(q)}\mathbf{v}_i^{(q)}\mathbf{v}_i^{(q)\top}$ be the eigendeceomposition of $\mathbf{A}$. The Nyström method provides the formulae for $\tilde{u}_i$ and $\tilde{\mathbf{v}}_i$ (Rasmussen and Williams, 2006, §8.1, equations 8.2 and 8.3) as

$$\begin{aligned} \tilde{u}_i &:= \frac{n}{q}u_i^{(q)} \in \mathbb{R} \\ \tilde{\mathbf{v}}_i &:= \sqrt{\frac{q}{n}}\frac{1}{u_i^{(q)}}\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix}^\top \mathbf{v}_i^{(q)} \in \mathbb{R}^n. \end{aligned}$$

Denoting $\mathbf{U}_q$ as the diagonal matrix of eigenvalues $u_1^{(q)}, \ldots, u_m^{(q)}$, and $\mathbf{V}_q$ the corresponding matrix of eigenvectors $\mathbf{v}_i^{(q)}$, we have

$$\mathbf{H}_\eta \approx \overbrace{\begin{pmatrix}\mathbf{V}_q \\ \mathbf{B}^\top\mathbf{V}_q\mathbf{U}_q^{-1}\end{pmatrix}}^{\bar{\mathbf{V}}}\mathbf{U}_q\overbrace{\begin{pmatrix}\mathbf{V}_q^\top & \mathbf{U}_q^{-1}\mathbf{V}_q^\top\mathbf{B}\end{pmatrix}}^{\bar{\mathbf{V}}^\top}.$$

Unfortunately, it may be the case that $\bar{\mathbf{V}}\bar{\mathbf{V}}^\top \neq \mathbf{I}_n$, while orthogonality is crucial in order to easily calculate the inverse of $\mathbf{\Sigma}_\theta$. An additional step is required to obtain an orthogonal version of the Nyström decomposition, as studied by Fowlkes et al. (2004, 2001). Let $\mathbf{K} = \mathbf{A} + \mathbf{A}^{-\frac{1}{2}}\mathbf{B}^\top\mathbf{B}\mathbf{A}^{-\frac{1}{2}}$, where $\mathbf{A}^{-\frac{1}{2}} = \mathbf{V}_m\mathbf{U}_m^{-\frac{1}{2}}\mathbf{V}_m$, and obtain the eigendecomposition of this $m \times m$ matrix $\mathbf{K} = \mathbf{R}\hat{\mathbf{U}}\mathbf{R}^\top$. Defining

$$\hat{\mathbf{V}} = \begin{pmatrix}\mathbf{A} \\ \mathbf{B}^\top\end{pmatrix}\mathbf{A}^{-\frac{1}{2}}\mathbf{R}\hat{\mathbf{U}}^{-\frac{1}{2}} \in \mathbb{R}^n \times \mathbb{R}^m,$$

then we have that $\mathbf{H}_\eta \approx \hat{\mathbf{V}}\hat{\mathbf{U}}\hat{\mathbf{V}}^\top$ such that $\hat{\mathbf{V}}\hat{\mathbf{V}}^\top = \mathbf{I}_n$ (Fowlkes et al., 2004, Appx. A). Estimating I-prior models with the Nyström method including the orthogonalisation step takes roughly $O(nm^2)$ time and $O(nm)$ storage.

The issue of selecting the subset $\mathcal{Q}$ remains. The simplest method, and that which is implemented in the **iprior** package, would be to uniformly sample a subset of size q from the n points. Although this works well in practice, the quality of approximation might suffer if the points do not sufficiently represent the training set. In this light, greedy approximations have been suggested to select the q points, so as to reduce some error criterion relating to the quality of approximation. For a brief review of more sophisticated methods of selecting $\mathcal{Q}$, see Rasmussen and Williams (2006, §8.1).

4.3.2 Front-loading kernel matrices for the EM algorithm

The evaluation of the Q function in (4.15) is $O(n^3)$, because a change in the values of θ requires evaluating $\mathbf{\Sigma}_\theta = \psi\mathbf{H}_\eta^2 + \psi^{-1}\mathbf{I}_n$, for which squaring $\mathbf{H}_\eta$ takes the bulk of the computational time. This is disadvantageous because, a Newton or quasi-Newton algorithm used for the M-step would require multiple evaluations of Q in order to complete an EM update.

In this section, we describe an efficient method of evaluating Q if the I-prior model only involves estimating the RKHS scale parameters and the error precision under assumptions A1–A3. The premise is this: squaring an ANOVA kernel matrix can be made more efficient because it is a linear combination of several other kernel matrices, which can be pre-calculated and stored for multiple use throughout the EM algorithm. We now describe the procedure in detail.

Corresponding to p building block RKHSs $\mathcal{F}_1, \ldots, \mathcal{F}_p$ of functions over $\mathcal{X}_1, \ldots, \mathcal{X}_p$, there are p scale parameters $\lambda_1, \ldots, \lambda_p$ and reproducing kernels $h_1, \ldots, h_p$. Assume that only the scale parameters are to be estimated, and the rest of the kernel parameters (Hurst coefficient, lengthscale, or offset) are fixed. Write $\theta = \{\lambda_1, \ldots, \lambda_p, \psi\}$. The most common modelling scenarios that will be encountered are listed below:

1. **Single scale parameter**. With $p = 1$, $f \in \mathcal{F} \equiv \lambda_1\mathcal{F}_1$ of functions over a set $\mathcal{X}$. $\mathcal{F}$ may be any of the building block RKHSs. Note that $\mathcal{X}_1$ itself may be more than one-dimensional. The kernel over $\mathcal{X}_1 \times \mathcal{X}_1$ is therefore

$$h_\lambda = \lambda_1 h_1.$$

2. **Multiple scale parameters**. Here, $\mathcal{F}$ is a RKKS of functions $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$, and thus $\mathcal{F} \equiv \lambda_1\mathcal{F}_1 \oplus \cdots \oplus \lambda_p\mathcal{F}_p$, where each $\mathcal{F}_k$ is one of the building block RKHSs. The kernel is

$$h_\lambda = \lambda_1 h_1 + \cdots + \lambda_p h_p.$$

3. **Multiple scale parameters with level-2 interactions**. This occurs commonly with multilevel and longitudinal models. Suppose that $\mathcal{X}_1$ is the set of 'levels' and there are $p - 1$ covariate sets $\mathcal{X}_k$, $k = 2, \cdots, p$. The function space $\mathcal{F}$ is a special case of the ANOVA RKKS containing only main and two-way interaction effects, and its kernel is

$$h_\lambda = \sum_{j=1}^{p} \lambda_j h_j + \sum_{j<k} \lambda_j\lambda_k h_j h_k,$$

where $\mathcal{F}_1$ is the Pearson RKHS, and the remaining are any of the building block RKHSs.

4. **Polynomial RKKS**. When using the polynomial RKKS of degree d to incite a polynomial relationship of the covariate set $\mathcal{X}_1$ on the function $f \in \mathcal{F}$ (excluding an intercept), then the kernel of $\mathcal{F}$ is

$$h_\lambda = \sum_{k=1}^{d} b_k \lambda_1^k h_1^k.$$

where $b_k = \frac{d!}{k!(d-k)!}$, $k = 1, \ldots, d$ are constants.

Of course, many other models are possible, such as the ANOVA RKKS with all p levels of interactions. What we realise is that any of these scenarios are simply a sum-product of a manipulation of the set of scale parameters $\lambda = \{\lambda_1, \ldots, \lambda_p\}$ and the set of kernel functions $h = \{h_1, \ldots, h_p\}$.

Let us be more concrete about what we mean by 'manipulation' of the sets λ and h. Define an 'instruction operator' which expands out both sets identically as required by the modelling scenario. Computationally speaking, this instruction could be carried out through an instructive list $\mathcal{Q}$ containing the indices to multiply out. For the four scenarios above, the list $\mathcal{Q}$ are as follows:

1. $\mathcal{Q} = \big\{\{1\}\big\}$.
2. $\mathcal{Q} = \big\{\{1\}, \ldots, \{p\}\big\}$.
3. $\mathcal{Q} = \big\{\{1\}, \ldots, \{p\}, \{1,2\}, \ldots, \{p-1, p\}\big\}$.
4. $\mathcal{Q} = \big\{\{1\}, \{1,1\}, \ldots, \{\overbrace{1, \ldots, 1}^{d}\}\big\}$.

For the polynomial RKKS in the fourth example, one must also multiply the constants b_k to the λ's as appropriate. Let q be the cardinality of the set $\mathcal{Q}$, which is the number of summands required to construct the kernel for $\mathcal{F}$. Denote the instructed sets as $\xi = \{\xi_1, \ldots, \xi_q\}$ for λ and $a = \{a_1, \ldots, a_q\}$ for h. We can write the kernel h_λ as a linear combination of ξ and a,

$$h_\lambda = \xi_1 a_1 + \cdots + \xi_q a_q.$$

The reason this is important is because changes in λ for h_λ only changes the ξ_k's, but not the a_k's. This allows us to compute and store all of the required $n \times n$ kernel matrices $\mathbf{A}_1, \ldots, \mathbf{A}_q$ by application of the instruction set $\mathcal{Q}$ on h, evaluated at all pairs of data points $(x_i, x_j) \in \mathcal{X} \times \mathcal{X}$. This process of initialisation need only be done once prior to commencing the EM algorithm—a step we refer to as 'kernel loading'.

Notice that

$$\begin{aligned}
\operatorname{tr}\left(\boldsymbol{\Sigma}_\theta \tilde{\mathbf{W}}^{(t)}\right) &= \operatorname{tr}\left((\psi \mathbf{H}_\eta^2 + \psi^{-1}\mathbf{I}_n)\tilde{\mathbf{W}}^{(t)}\right) \\
&= \psi \operatorname{tr}(\mathbf{H}_\eta^2 \tilde{\mathbf{W}}^{(t)}) + \psi^{-1} \operatorname{tr} \tilde{\mathbf{W}}^{(t)} \\
&= \psi \operatorname{tr}\left(\sum_{j,k=1}^{q} \xi_j \xi_k \big(\mathbf{A}_j \mathbf{A}_k + (\mathbf{A}_j \mathbf{A}_k)^\top\big)\tilde{\mathbf{W}}^{(t)}\right) + \psi^{-1} \operatorname{tr} \tilde{\mathbf{W}}^{(t)} \\
&= 2\psi \sum_{j,k=1}^{q} \xi_j \xi_k \operatorname{tr}\left(\mathbf{A}_j \mathbf{A}_k \tilde{\mathbf{W}}^{(t)}\right) + \psi^{-1} \operatorname{tr} \tilde{\mathbf{W}}^{(t)}.
\end{aligned}$$

Provided that we have the matrices $\mathbf{A}_{jk} = \mathbf{A}_j \mathbf{A}_k$, $j, k = 1, \ldots, q$ in addition to $\mathbf{A}_1, \ldots, \mathbf{A}_q$ pre-calculated and stored, then evaluating $\operatorname{tr}\left(\mathbf{A}_{jk}\tilde{\mathbf{W}}^{(t)}\right) = \operatorname{vec}(\mathbf{A}_{jk})^\top \operatorname{vec}(\tilde{\mathbf{W}}^{(t)})$ is $O(n^2)$, although this only need to be done once per EM iteration. Thus, with the kernels loaded, the overall time complexity to evaluate Q is $O(n^2)$ at the beginning of each iteration, but roughly linear in ξ thereafter.

In conclusion, we have achieved efficiency at the expense of storage and a potentially long initialisation phase of kernel loading. In the **iprior** package, kernel loading is performed using the `kernL()` command. The storing of the kernel matrices can be very expensive, especially if the sample size is very large; Figure 4.3 shows the storage cost of front-loading the kernel matrices for varying number of ANOVA components $p = 1, \ldots, 5$ and sample sizes. On the bright side, once the kernel matrices are stored in hard memory, the **iprior** package allows them to be reused again and again. A practical situation where this might be useful is when we would like to repeat the EM at various initial values. Although front-loading of kernel matrices increase storage requirements, this is manageable in practice in modern computer systems for sample sizes of $n \leq 5,000$, and there is a clear advantage of doing so.

Remark 4.4. The sign of the scale parameters itself are not identified in the model (this is easily seen when having a single scale parameter in the model since the scale is squared when it appears in the likelihood) but *relative signs of the scale parameters with respect to each other* is.

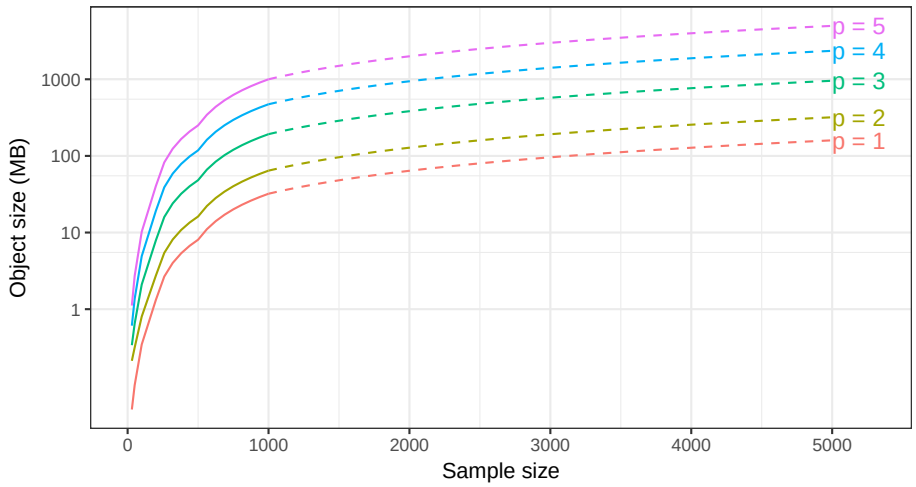


Figure 4.3: Storage cost of front-loading the kernel matrices for varying number of ANOVA components $p = 1, \ldots, 5$ and sample sizes. Solid lines indicate actual values, while dotted lines are (linear) extrapolations. Storage requirements increases exponentially, since for p ANOVA components, there are 2^{p+1} kernel matrices to store in memory.

4.3.3 The exponential family EM algorithm

In the original EM paper by Dempster et al. (1977), the EM algorithm was demonstrated to be easily administered to complete data likelihoods belonging to the exponential family for which the maximum likelihood estimates are easily computed. If this is the case, then the M-step simply involves replacing the unknown sufficient statistics in the ML estimates with their *conditional expectations*. Certain I-prior models admit this property, namely regression functions belonging to the full or limited ANOVA RKKS. For such models, we can reduce the EM algorithm to a sequential updating scheme of the latent variables (missing data) and parameters, bypassing the need for a gradient-based optimisation in the M-step. We describe the implementation of this exponential family EM below.

Assume A1–A3 applies, and that only the error precision ψ and the RKHS scale parameters $\lambda_1, \ldots, \lambda_p$ need to be estimated, i.e. all other kernel parameters are fixed—a similar situation was described in the previous subsection. For the full ANOVA RKKS, the kernel can be written in the form

$$
\begin{aligned}
h_\lambda &= \sum_{i=1}^{p} \lambda_i h_i + \sum_{i<j} \lambda_i \lambda_j h_i h_j + \cdots + \prod_{i=1}^{p} \lambda_i h_i \\
&= \lambda_k \overbrace{\left(h_k + \sum_{i} \lambda_i h_i h_k + \cdots + h_k \prod_{i \neq k} \lambda_i h_i \right)}^{\text{terms of } \lambda_k} + \overbrace{\sum_{i \neq k} \lambda_i h_i + \sum_{i,j \neq k} \lambda_i \lambda_j h_i h_j + \cdots + 0}^{\text{no } \lambda_k \text{ here}} \\
&= \lambda_k r_k + s_k
\end{aligned}
$$

where r_k and s_k are both functions over $\mathcal{X} \times \mathcal{X}$, defined respectively as the terms of the ANOVA kernel involving λ_k, and the terms not involving λ_k. The reason for splitting h_λ like this will become apparently momentarily.

Programmatically, this looks complicated to implement in software, but in fact it is not. Consider again the instruction list $\mathcal{Q}$ for the ANOVA RKKS (Example 3, Section 4.3.2). We can split this list into two: $\mathcal{R}_k$ as those elements of $\mathcal{Q}$ which involve the index k, and $\mathcal{S}_k$ as those elements of $\mathcal{Q}$ which do not involve the index k. Let ζ_k, e_k be the sets of λ and h after applying the instructions of $\mathcal{R}_k$, and let ξ_k and a_k be the sets of λ and h after application of the instruction list $\mathcal{S}_k$. Now, we have

$$r_k = \frac{1}{\lambda_k} \sum_{l=1}^{|\mathcal{R}_k|} \zeta_{lk} e_{lk} \quad \text{and} \quad s_k = \sum_{l=1}^{|\mathcal{S}_k|} \xi_{lk} a_{lk},$$

as real-valued functions defined over $\mathcal{X} \times \mathcal{X}$. Defining $\mathbf{R}_k$ and $\mathbf{S}_k$ as the kernel matrices with (i, j) entries $r_k(x_i, x_j)$ and $s_k(x_i, x_j)$ respectively, for $i, j = 1, \ldots, n$, we have that

$$\mathbf{H}_\eta^2 = \lambda_k^2 \mathbf{R}_k^2 + \lambda_k \overbrace{\left(\mathbf{R}_k \mathbf{S}_k + (\mathbf{R}_k \mathbf{S}_k)^\top\right)}^{\mathbf{U}_k} + \mathbf{S}_k^2.$$

Consider now the full data log-likelihood for λ_k, $k = 1, \ldots, p$, conditionally dependent on the rest of the unknown parameters $\lambda_{-k} = \{\lambda_1, \ldots, \lambda_p\} \backslash \{\lambda_k\}$ and ψ:

$$\begin{aligned} L(\lambda_k | \mathbf{y}, \mathbf{w}, \lambda_{-k}, \psi) &= \text{const.} - \frac{1}{2} \operatorname{tr}\left((\psi \mathbf{H}_\eta^2 + \psi^{-1} \mathbf{I}_n) \mathbf{w}\mathbf{w}^\top \right) + \psi \tilde{\mathbf{y}}^\top \mathbf{H}_\eta \mathbf{w} \\ &= \text{const.} - \lambda_k^2 \frac{\psi}{2} \operatorname{tr}(\mathbf{R}_k^2 \mathbf{w}\mathbf{w}^\top) + \lambda_k \left(\psi \tilde{\mathbf{y}}^\top \mathbf{R}_k \mathbf{w} - \frac{\psi}{2} \operatorname{tr}(\mathbf{U}_k \mathbf{w}\mathbf{w}^\top) \right). \end{aligned} \tag{4.20}$$

Notice that the above likelihood is an exponential family distribution with the natural parameterisation $\beta = (-\lambda_k^2, \lambda_k)$ and sufficient statistics T_1 and T_2 defined by

$$T_1 = \frac{\psi}{2} \operatorname{tr}(\mathbf{R}_k^2 \mathbf{w}\mathbf{w}^\top) \quad \text{and} \quad T_2 = \psi \tilde{\mathbf{y}}^\top \mathbf{R}_k \mathbf{w} - \frac{\psi}{2} \operatorname{tr}(\mathbf{U}_k^2 \mathbf{w}\mathbf{w}^\top).$$

This likelihood is maximised at $\hat{\lambda}_k = T_2 / 2T_1$, but of course, the variables $w_1, \ldots, w_n$ are never observed. As per the exponential family EM routine, replace occurrences of $\mathbf{w}$ and $\mathbf{w}\mathbf{w}^\top$ with their respective conditional expectations, i.e. $\mathbf{w} \mapsto \mathrm{E}[\mathbf{w}|\mathbf{y}] = \tilde{\mathbf{w}}$ and $\mathbf{w}\mathbf{w}^\top \mapsto \mathrm{E}[\mathbf{w}\mathbf{w}^\top|\mathbf{y}] = \tilde{\mathbf{V}}_w + \tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top$ as defined in (4.11). That the λ_k's have closed-form expressions, together with the closed-form expression for ψ in (4.18), greatly simplifies the EM algorithm. At the M-step, one simply updates the parameters in turn, and as such, there is no maximisation per se.

The exponential family EM algorithm for ANOVA-type I-prior models is summarised in Algorithm 1. It requires $O(n^3)$ computational time at each step, which is spent on

computing the matrix inverse in the E-step. The M-step takes at most $O(n^2)$ time to compute. Algorithm 1 also requires front-loading of the kernel matrices, which increases storage requirements. As a remark, it is not necessary that h_λ is the full ANOVA RKKS; any of the examples 1–3 in Section 4.3.2 can be estimated using this method, since they are seen as special cases of the ANOVA decomposition.

Algorithm 1 Exponential family EM for ANOVA-type I-prior models

procedure INITIALISATION
  Initialise $\lambda_1^{(0)}, \ldots, \lambda_p^{(0)}, \psi^{(0)}$
  Compute and store matrices as per $\mathcal{R}_k$ and $\mathcal{S}_k$.
  $t \leftarrow 0$
end procedure

while not converged **do**
  procedure E-STEP
    $\tilde{\mathbf{w}} \leftarrow \psi^{(t)}\mathbf{H}_{\eta^{(t)}}\left(\psi^{(t)}\mathbf{H}^2_{\eta^{(t)}} + \psi^{-(t)}\mathbf{I}_n\right)^{-1}\tilde{\mathbf{y}}$
    $\tilde{\mathbf{W}} \leftarrow \left(\psi^{(t)}\mathbf{H}^2_{\eta^{(t)}} + \psi^{-(t)}\mathbf{I}_n\right)^{-1} + \tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top$
  end procedure

  procedure M-STEP
    for $k = 1, \ldots, p$ **do**
      $T_{1k} \leftarrow \frac{1}{2}\operatorname{tr}(\mathbf{R}_k^2\tilde{\mathbf{W}})$
      $T_{2k} \leftarrow \tilde{\mathbf{y}}^\top\mathbf{R}_k\tilde{\mathbf{w}} - \frac{1}{2}\operatorname{tr}(\mathbf{U}_k^2\tilde{\mathbf{W}}^\top)$
      $\lambda_k^{(t+1)} \leftarrow T_{2k}/2T_{1k}$
    end for
    $T_3 \leftarrow \tilde{\mathbf{y}}^\top\tilde{\mathbf{y}} + \operatorname{tr}(\mathbf{H}^2_{\eta^{(t)}}\tilde{\mathbf{W}}^{(t)}) - 2\tilde{\mathbf{y}}^\top\mathbf{H}_{\eta^{(t)}}\tilde{\mathbf{w}}^{(t)}$
    $\psi^{(t+1)} \leftarrow \operatorname{tr}\tilde{\mathbf{W}}^{(t)}/T_3$
  end procedure
  $t \leftarrow t + 1$
end while

$\{\hat{\lambda}_1, \ldots, \hat{\lambda}_p, \hat{\psi}\} \leftarrow \{\lambda_1^{(t)}, \ldots, \lambda_p^{(t)}, \psi^{(t)}\}$
return Estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_p, \hat{\psi}$

Remark 4.5. Another compelling reason to use Algorithm 1 is conjugacy of the exponential family of distributions. Realise that $\lambda_k|\mathbf{y}, \mathbf{w}, \lambda_{-k}, \psi$ is in fact normally distributed, with mean and variance given by $T_2/2T_1$ and $1/2T_1$ respectively. If we were so compelled to assign a normal prior on each of the λ_k's, then the conditionally dependent log-likelihood of λ_k, $L(\lambda_k|\mathbf{y}, \mathbf{w}, \lambda_{-k}, \psi)$, would have a normal prior log-density involving λ_k added on. Importantly, viewed as a posterior log-density for λ_k, the λ_k is normally distributed. The exponential family EM is thus easily modified to compute maximum a posteriori (MAP) estimates (or penalised ML estimates) of the scale parameters.

Remark 4.6. The restriction to ANOVA RKKSs is due to the fact that as soon as higher degrees of the λ_k's come into play, e.g. using the polynomial kernel, then the ML estimates for the λ_k's involve solving a polynomial of degree $2d-1$ for FOC equations. Although this is not in itself hard to do, the elegance of the algorithm, especially viewed as having the normal conjugacy property for the λ_k's, is lost.

4.4 Post-estimation

One of the perks of a (semi-)Bayesian approach to regression modelling is that we are able to use Bayesian post-estimation machinery involving the relevant posterior distributions. With the normal I-prior model, there is the added benefit that posterior distributions are easily obtained in closed form. We describe post-estimation procedures such as prediction of a new data point, inference surrounding the prediciton, and model comparison. The plots that are shown in this subsection is a continuation of the example from Section 4.2.5.

Recall that for the I-prior model (4.10), a regression function $f(x) = \sum_{i=1}^n h_{\hat\eta}(x, x_i)\tilde w_i$ has the posterior Gaussian distribution specified by the mean and variance of the multivariate normal $\tilde w_i$'s given in (4.11). Denote by $\mathbf{h}_{\hat\eta}(x)$ the n-vector with entries equal to $h_{\hat\eta}(x, x_i)$. Precisely, the posterior density for the regression function is

$$f(x)|\mathbf{y} \sim \mathrm{N}\left(\mathbf{h}_{\hat\eta}(x)\hat{\mathbf{w}}, \mathbf{h}_{\hat\eta}(x)^\top \left(\mathbf{H}_{\hat\eta}\hat{\boldsymbol\Psi}\mathbf{H}_{\hat\eta} + \hat{\boldsymbol\Psi}^{-1}\right)^{-1}\mathbf{h}_{\hat\eta}(x)\right) \tag{4.21}$$

for any x in the domain of the regression function. Here, the hats on the parameters indicate the use of the optimised model parameters, i.e. the ML or MAP estimates.

Prediction of a new data point is now described. A priori, assume that $y_{\text{new}} = \hat\alpha + f(x_{\text{new}}) + \epsilon_{\text{new}}$, where $\epsilon_{\text{new}} \sim \mathrm{N}(0, \psi_{\text{new}}^{-1})$, and $f \sim$ I-prior. Denote the covariance between ϵ_{new} and $\boldsymbol\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$ by $\boldsymbol\sigma_{\text{new}}^\top \in \mathbb{R}^n$. Under an iid model (assumption A3), then $\psi_{\text{new}} = \psi = \operatorname{Var}\epsilon_i$ for any $i \in \{1, \dots, n\}$, and $\boldsymbol\sigma_{\text{new}}^\top = \mathbf{0}$, but otherwise, these extra parameters need to be dealt with somehow, either by specifying them a priori or estimating them again, which seems excessive. In any case, using a linearity argument, the posterior distribution for y_{new} is normal, with mean and variance given by

$$\mathrm{E}[y_{\text{new}}|\mathbf{y}] = \hat\alpha + \mathrm{E}\left[f(x_{\text{new}})|\mathbf{y}\right] + \text{correction term} \tag{4.22}$$

and

$$\operatorname{Var}[y_{\text{new}}|\mathbf{y}] = \operatorname{Var}\left[f(x_{\text{new}})|\mathbf{y}\right] + \psi_{\text{new}}^{-1} + \text{correction term}. \tag{4.23}$$

A derivation is presented in Appendix G.2. Note, that the mean and variance correction term vanishes under an iid assumption A3. The posterior distribution for y_{new} can be used in several ways. Among them, is to construct a $100(1-\alpha)\%$ credibility interval for

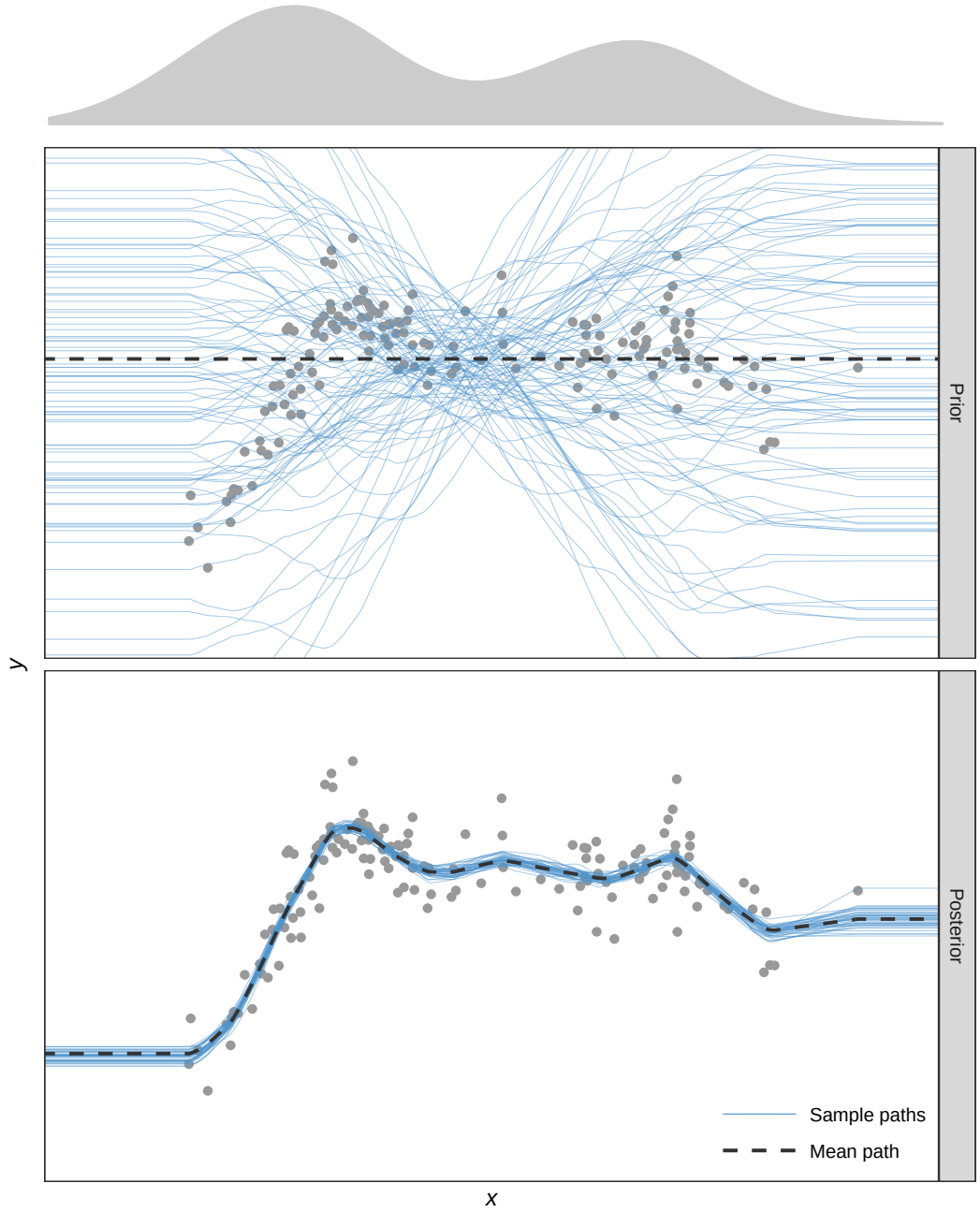


Figure 4.4: Prior (top) and posterior (bottom) sample path realisations of regression functions drawn from their respective distributions when $\mathcal{F}$ is a fBm-0.5 RKHS. At the very top of the figure, a smoothed density estimate of the x's is overlaid. In regions with few data points (near the centre), there is little Fisher information, and hence a conservative prior closer to zero, the prior mean, for this region.

the (mean) predicted value y_{new} using

$$\mathrm{E}[y_{\text{new}}|\mathbf{y}] \pm \Phi^{-1}(1-\alpha/2)\ \mathrm{Var}[y_{\text{new}}|\mathbf{y}]^{\frac{1}{2}},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. One could also perform a posterior predictive density check of the data $\mathbf{y}$, by repeatedly sampling n points from its posterior distribution. This provides a visual check of whether there are any systematic deviances between what the model predicts, and what is observed from the data.

Lastly, we discuss model comparison. Recall that the marginal distribution for $\mathbf{y}$ after integrating out the I-prior for f in model (4.10) is normal. Suppose that we are interested in comparing two candidate models M_0 and M_1, each with parameter sets θ_0 and θ_1. Commonly, we would like to test whether or not particular terms in the ANOVA RKKS are significant contributors in explaining the relationship between the responses and predictors. A log-likelihood comparison is possible using an asymptotic chi-squared distribution, with degrees of freedom equal to the difference between the number of parameters in M_1 and M_0. This is assuming model M_0 is nested within M_1, which is the case for ANOVA-type constructions. Note that if two models have the same number of parameters, then the model with the higher likelihood is preferred.

Remark 4.7. This method of comparing marginal likelihoods can be seen as Bayesian model selection using *empirical Bayes factors*, where the Bayes factor of comparing model M_1 against model M_0 is defined as

$$\mathrm{BF}(M_1, M_0) = \frac{\int p(\mathbf{y}|\hat{\theta}_1, \mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}}{\int p(\mathbf{y}|\hat{\theta}_0, \mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}}.$$

Bayes factor values of greater than one indicate more support for model M_1 over M_0. The term 'empirical' stems from the fact that the parameters are estimated via an empirical Bayes approach (maximum marginal likelihood), as opposed to assuming prior distributions on them and integrating them out.

4.5 Examples

We demonstrate I-prior modelling on a toy data set to illustrate the Nyström method, as well as three other real-data examples. All of the analyses were conducted in R, and I-prior model estimation was done using the **iprior** package (Jamil, 2017). The **iprior** package comes documented with usage examples in the vignette. The complete source code for replication is found at `http://myphdcode.haziqj.ml`. Note that in all of these examples, A1–A3 were assumed.

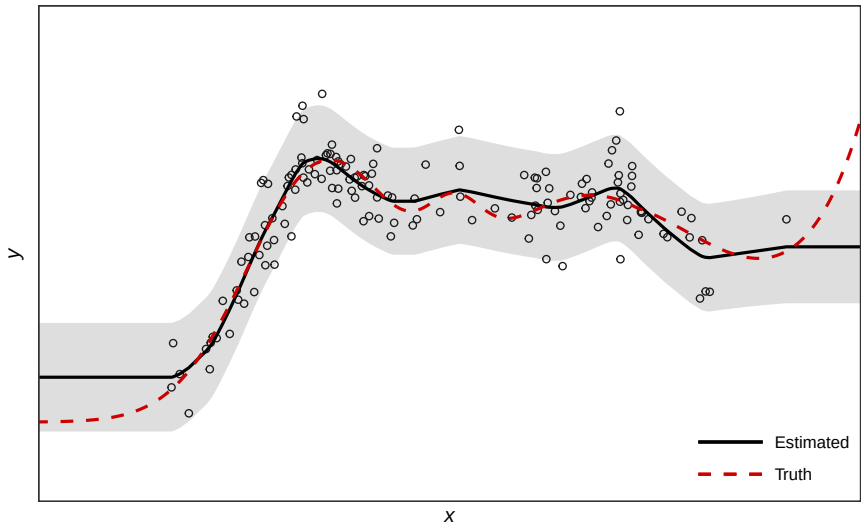


Figure 4.5: The estimated regression line (solid black) is the posterior mean estimate of the regression function (shifted by the intercept), which also gives the posterior mean estimate for the responses y. The shaded region is the 95% credibility interval for predictions. The true regression line (dashed red) is shown for comparison.

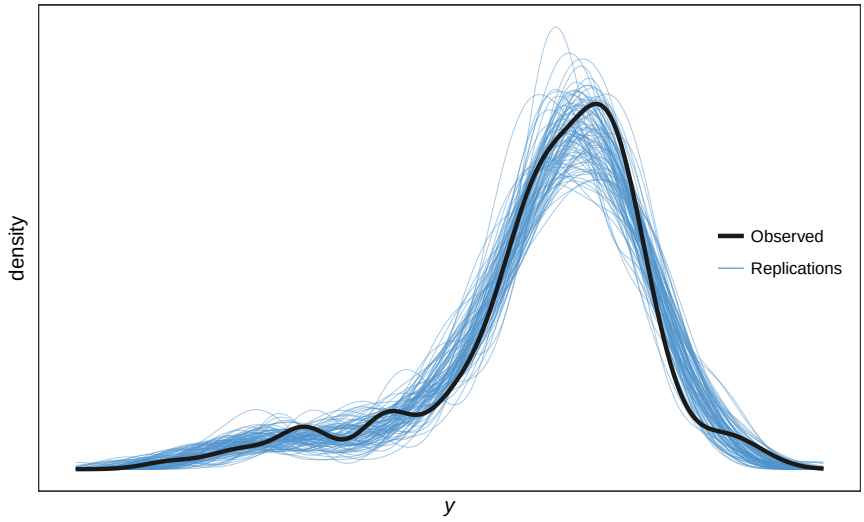


Figure 4.6: Posterior predictive density checks of the responses: repeated sampling from the posterior density of the y_i's and plotting their densities allows us to compare model predictions against observed samples.

4.5.1 Random effects models

In this section, a comparison between a standard random effects model and the I-prior approach for estimating varying intercept and slopes model is illustrated. The example concerns control data[4] from several runs of radioimmunoassays (RIA) for the protein insulin-like growth factor (IGF-I) (explained in further detail in Davidian and Giltinan, 1995, §3.2.1). RIA is an in vitro assay technique which is used to measure concentration of antigens—in our case, the IGF-I proteins. When an RIA is run, control samples at known concentrations obtained from a particular lot are included for the purpose of assay quality control. It is expected that the concentration of the control material remains stable as the machine is used, up to a maximum of about 50 days, at which point control samples from a new batch is used to avoid degradation in assay performance.

```
R> data(IGF, package = "nlme")
R> head(IGF)
```

```
## Grouped Data: conc ~ age | Lot
##   Lot age conc
## 1   1   7 4.90
## 2   1   7 5.68
## 3   1   8 5.32
## 4   1   8 5.50
## 5   1  13 4.94
## 6   1  13 5.19
```

The data consists of IGF-I concentrations (`conc`) from control samples from 10 different lots measured at differing `age`s of the lot. The data were collected with the aim of identifying possible trends in control values `conc` with `age`, ultimately investigating whether or not the usage protocol of maximum sample age of 50 days is justified. Pinheiro and Bates (2000) remarks that this is not considered a longitudinal problem because different samples were used at each measurement.

We shall model the IGF data set using the I-prior methodology using the ANOVA-decomposed regression function

$$f(\texttt{age}, \texttt{Lot}) = f_1(\texttt{age}) + f_2(\texttt{Lot}) + f_{12}(\texttt{age}, \texttt{Lot})$$

where f_1 lies in the linear RKHS $\mathcal{F}_1$, f_2 in the Pearson RKHS $\mathcal{F}_2$ and f_{12} in the tensor product space $\mathcal{F}_{12} = \mathcal{F}_1 \otimes \mathcal{F}_2$. The regression function f then lies in the RKHS $\mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2 \oplus \mathcal{F}_{12}$ with kernel equal to the sum of the kernels from each of the RKHSs. The explanation here is that the `conc` levels are assumed to be related to both `age` and `Lot`, and in particular, the contribution of `age` on `conc` varies with each individual `Lot`. This

[4] This data is available in the R package **nlme** (Pinheiro et al., 2017).

gives the intended effect of a linear mixed-effects model, which is thought to be suitable in this case, in order to account for within-lot and between-lot variability. We first fit the model using the **iprior** package, and then compare the results with the standard random effects model using the R command `lme4::lmer()`. The command to fit the I-prior model using the EM algorithm is

```
R> mod.iprior <- iprior(conc ~ age * Lot, IGF, method = "em")

## ========================================
## Converged after 57 iterations.

R> summary(mod.iprior)

## Call:
## iprior(formula = conc ~ age * Lot, data = IGF, method = "em")
##
## RKHS used:
## Linear (age)
## Pearson (Lot)
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -4.4889 -0.3798 -0.0090  0.2563  4.3973
##
## Hyperparameters:
##           Estimate   S.E.      z P[|Z>z|]
## lambda[1]   0.0000 0.0002 -0.004    0.997
## lambda[2]   0.0007 0.0030  0.238    0.812
## psi         1.4576 0.1366 10.672   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Closed-form EM algorithm. Iterations: 57/100
## Converged to within 1e-08 tolerance. Time taken: 2.882966 secs
## Log-likelihood value: -291.9033
## RMSE of prediction: 0.8273639 (Training)
```

To make inference on the covariates, we look at the scale parameters `lambda`. We see that both scale parameters for `age` and `Lot` are close to zero, and a test of significance is not able to reject the hypothesis that these parameters are indeed null. We conclude that neither `age` nor `Lot` has a linear effect on the `conc` levels. The plot of the fitted regression line in Figure 4.7 does show an almost horizontal line for each `Lot`.

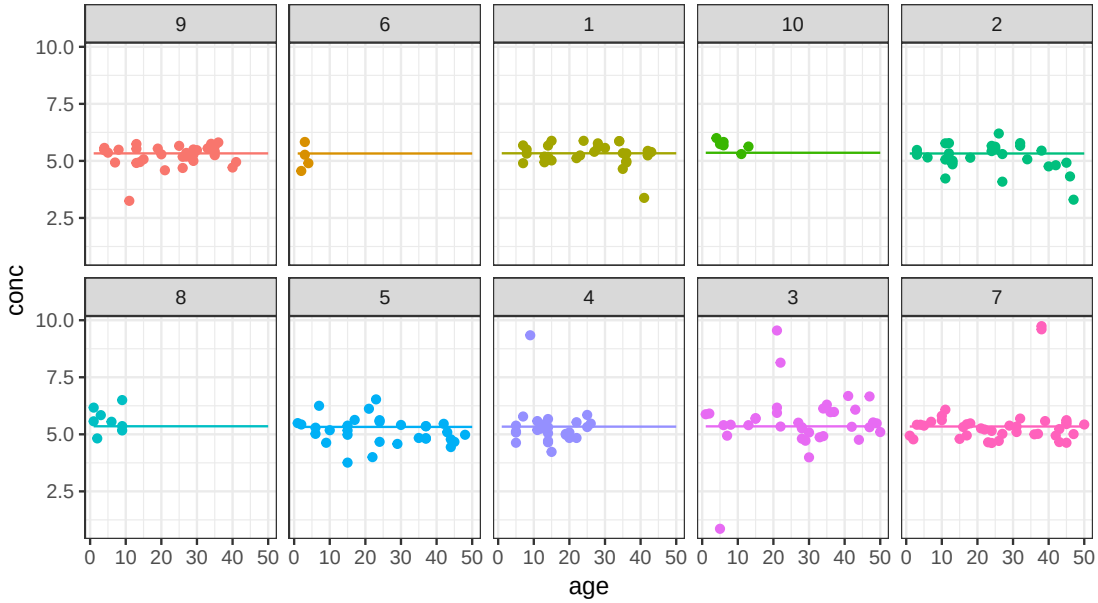


Figure 4.7: Plot of fitted regression line for the I-prior model on the IGF data set, separated into each of the 10 lots.

The standard random effects model, as explored by Davidian and Giltinan (1995) and Pinheiro and Bates (2000), is

$$\begin{aligned} \texttt{conc}_{ij} &= \beta_{0j} + \beta_{1j}\texttt{age}_{ij} + \epsilon_{ij} \\ \begin{pmatrix}\beta_{0j}\\ \beta_{1j}\end{pmatrix} &\sim \mathrm{N}\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{pmatrix}\right) \\ \epsilon_{ij} &\sim \mathrm{N}(0, \sigma^2) \end{aligned}$$

for $i = 1, \ldots, n_j$ and the index j representing the 10 `Lots`. Fitting this model using `lmer`, we can test for the significance of the fixed effect β_0, for which we find that it is not (p-value $= 0.627$), and arrive at the same conclusion as in the I-prior model.

```
R> (mod.lmer <- lmer(conc ~ age + (age | Lot), IGF))

## Linear mixed model fit by REML ['lmerModLmerTest']
## Formula: conc ~ age + (age | Lot)
##    Data: IGF
## REML criterion at convergence: 594.3662
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Lot      (Intercept) 0.082507
##           age         0.008092 -1.00
##  Residual             0.820628
```

```
## Number of obs: 237, groups:  Lot, 10
## Fixed Effects:
## (Intercept)          age
##    5.374974    -0.002535
```

```
R> round(coef(summary(mod.lmer)), 4)
```

```
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   5.3750     0.1075 41.5757 50.0053   0.0000
## age          -0.0025     0.0050  9.5136 -0.5025   0.6267
```

However, we notice that the package reports a perfect negative correlation between the random effects, σ_{01}. This indicates a potential numerical issue when fitting the model—a value of exactly -1, 0 or 1 is typically imposed by the package to force through estimation in the event of non-positive definite covariance matrices arising. We can inspect the eigenvalues of the covariance matrix for the random effects to check that they are indeed non-positive definite. One of the eigenvalues was found to be negative, so the covariance matrix is non-positive definite.

```
R> eigen(VarCorr(mod.lmer)$Lot)
```

```
## eigen() decomposition
## $values
## [1]  6.872939e-03 -1.355253e-20
##
## $vectors
##             [,1]        [,2]
## [1,] -0.99522490 -0.09760839
## [2,]  0.09760839 -0.99522490
```

Degenerate covariance matrices often occur in models with a large number of random coefficients, and in cases where values of the variance components are estimated at the boundary. These are typically solved by setting restrictions which then avoids overparameterising the model. One advantage of the I-prior method for varying intercept/slopes model is that the positive-definiteness is automatically taken care of. Furthermore, I-prior models typically require fewer parameters to fit a similar varying intercept/slopes model—in the above example, the I-prior model estimated only three parameters, while the standard random effects model estimated a total of six parameters.

It is also possible to "recover" the estimates of the standard random effects model from the I-prior model, albeit in a slighly manual fashion (refer to Section 4.1.2). Denote by f^j the individual linear regression lines for each of the $j = 1, \ldots, 10$ `Lots`. Then, each of these f^j has a slope and intercept for which we can estimate from the fitted values

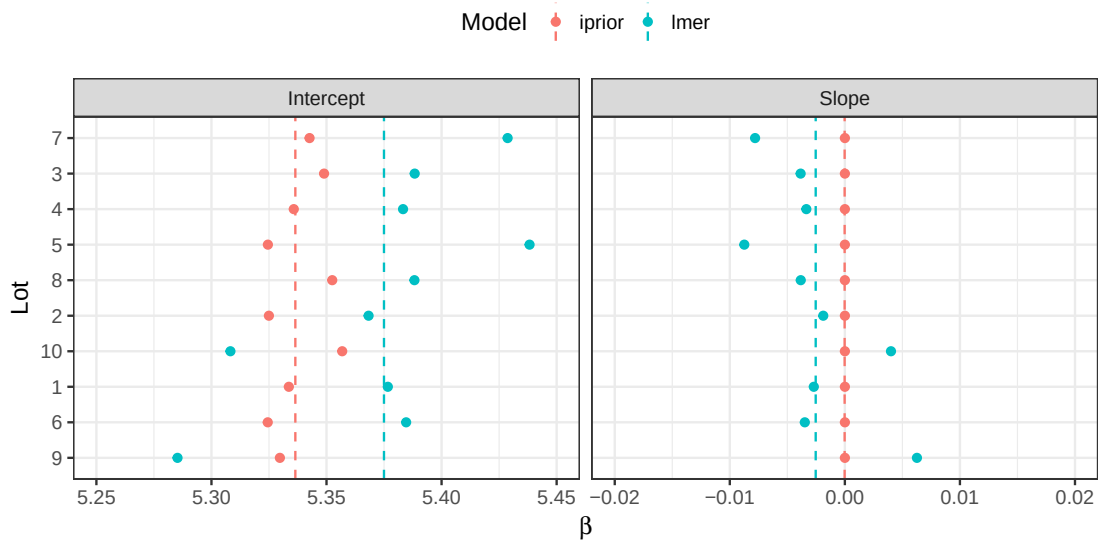


Figure 4.8: A comparison of the estimates for random intercepts and slopes (denoted as points) using the I-prior model and the standard random effects model. The dashed vertical lines indicate the fixed effect values.

$\hat{f}^j(x_{ij})$, $i = 1, \ldots, n_j$. This would give us the estimate of the posterior mean of the random intercepts and slopes; these would typically be obtained using empirical-Bayes methods in the case of the standard random effects model.

Furthermore, σ_0^2 and σ_1^2 gives a measure of variability of the intercepts and slopes of the different groups, and this can be calculated from the estimates of the random intercepts and slopes. In the same spirit, $\rho_{01} = \sigma_{01}/(\sigma_0\sigma_1)$, which is the correlation between the random intercept and slope, can be similarly calculated. Finally, the fixed effects can be estimated from the intercept and slope of the best fit line running through the I-prior estimated `conc` values. The intuition for this is that the fixed effects are essentially the ordinary least squares (OLS) of a linear model if the groupings are disregarded. Figure 4.8 illustrates the differences in the estimates for the random coefficients, while Table 4.2 illustrates the differences in the estimates for the covariance matrix. Minor differences do exist, with the most noticeable one being that the slopes in the I-prior model are categorically estimated as zero, and the sign of the correlation ρ_{01} being opposite in both models. Even so, the conclusions from both models are similar.

Table 4.2: A comparison of the estimates for the covariance matrix of the random effects using the I-prior model and the standard random effects model.

Parameter	`iprior`	`lmer`
σ_0	0.012	0.083
σ_1	0.000	0.008
ρ_{01}	0.690	-1.000

4.5.2 Longitudinal data analysis

We consider a balanced longitudinal data set consisting of weights in kilograms of 60 cows, 30 of which were randomly assigned to treatment group A, and the remaining 30 to treatment group B. The animals were weighed 11 times over a 133-day period; the first 10 measurements for each animal were made at two-week intervals and the last measurement was made one week later. This experiment was reported by Kenward (1987), and the data set is included as part of the package **jmcm** (J. Pan and Y. Pan, 2017) in R. The variable names have been renamed for convenience.

```
R> data(cattle, package = "jmcm")
R> names(cattle) <- c("id", "time", "group", "weight")
R> cattle$id <- as.factor(cattle$id)  # convert to factors
R> levels(cattle$group) <- c("Treatment A", "Treatment B")
R> str(cattle)
```

```
## 'data.frame': 660 obs. of  4 variables:
##  $ id    : Factor w/ 60 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1..
##  $ time  : num  0 14 28 42 56 70 84 98 112 126 ...
##  $ group : Factor w/ 2 levels "Treatment A",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ weight: int  233 224 245 258 271 287 287 287 290 293 ...
```

The response variable of interest are the `weight` growth curves, and the aim is to investigate whether a treatment effect is present. The usual approach to analyse a longitudinal data set such as this one is to assume that the observed growth curves are realizations of a Gaussian process. For example, Kenward (1987) assumed a so-called ante-dependence structure of order k, which assumes an observation depends on the previous k observations, but given these, is independent of any preceeding observations.

Using the I-prior, it is not necessary to assume the growth curves were drawn randomly. Instead, it suffices to assume that they lie in an appropriate function class. For this example, we assume that the function class is the fBm RKHS, i.e. we assume a smooth effect of time on weight. The growth curves form a multidimensional (or functional) response equivalent to a "wide" format of representing repeated measures data. In our analysis using the **iprior** package, we used the "long" format and thus our (unidimensional) sample size n is equal to 60 cows $\times$ 11 repeated measurements. We also have two covariates potentially influencing growth, namely the cow subject `id` and also treatment `group`. The regression model can then be thought of as

$$\begin{aligned}\texttt{weight} &= \alpha + f(\texttt{id}, \texttt{group}, \texttt{time}) + \epsilon \\ \epsilon &\sim \mathrm{N}(0, \psi^{-1}).\end{aligned}$$

Table 4.3: A brief description of the five models fitted using I-priors.

Model	Explanation	Formula (`weight ~ ...`)
1	Growth does not vary with treatment nor among cows	`time`
2	Growth varies among cows only	`id * time`
3	Growth varies with treatment only	`group * time`
4	Growth varies with treatment and among cows	`id * time + group * time`
5	Growth varies with treatment and among cows, with an interaction effect between treatment and cows	`id * group * time`

We assume iid errors, and in addition to a smooth effect of `time`, we further assume a nominal effect of both cow `id` and treatment `group` using the Pearson RKHS. In the **iprior** package, factor type objects are treated with the Pearson kernel automatically, and the only `model` option we need to specify is the `kernel = "fbm"` option for the `time` variable. We have opted not to estimate the Hurst coefficient in the interest of computational time, and instead left it at the default value of 1/2. Table 4.3 explains the five models we have fitted.

The simplest model fitted was one in which the growth curves do not depend on the treatment effect or individual cows. We then added treatment effect and the cow `id` as covariates, separately first and then together at once. We also assumed that both of these covariates are time-varying, and hence added also the interaction between these covariates and the `time` variable. The final model was one in which an interaction between treatment effect and individual cows was assumed, which varied over time.

All models were fitted using the `mixed` estimation method. Compared to the EM algorithm alone, we found that the combination of direct optimisation with the EM algorithm fits the model about six times faster for this data set due to slow convergence of EM algorithm. Here is the code and output for fitting the first model:

```
R> # Model 1: weight ~ f(time)
R> set.seed(456)
R> (mod1 <- iprior(weight ~ time, cattle, kern = "fbm", method = "mixed"))

## Running 5 initial EM iterations
## ================================================================
## Now switching to direct optimisation
## final  value 1394.615062
## converged
## Log-likelihood value: -2789.231
```

```
##
##  lambda     psi
## 0.83592 0.00375
```

Table 4.4: Summary of the five I-prior models fitted to the cow data set. Error S.D. refers to the inverse square root of the error precision, $\psi^{-1/2}$.

Model	Formula (`weight ~ ...`)	Log-likelihood	Error S.D.	Number of parameters
1	`time`	-2789.23	16.33	1
2	`id * time`	-2791.42	16.39	2
3	`group * time`	-2295.16	3.68	2
4	`id * time + group * time`	-2270.85	3.39	3
5	`id * group * time`	-2249.26	3.90	3

The results of the model fit are summarised in Table 4.4. We can test for a treatment effect by testing Model 4 against the alternative that Model 2 is true. The log-likelihood ratio test statistic is $D = -2(-2791.42 - (-2270.85)) = 1041.14$, which has an asymptotic chi-squared distribution with $3 - 2 = 1$ degree of freedom. The p-value for this likelihood ratio test is less than 10^{-6}, so we conclude that Model 4 is significantly better.

We can next investigate whether the treatment effect differs among cows by comparing Models 5 and 4. As these models have the same number of parameters, we can simply choose the one with the higher likelihood, which is Model 5. We conclude that treatment does indeed have an effect on growth, and that the treatment effect differs among cows. A plot of the fitted regression curves onto the cow data set is shown in Figure 4.9.

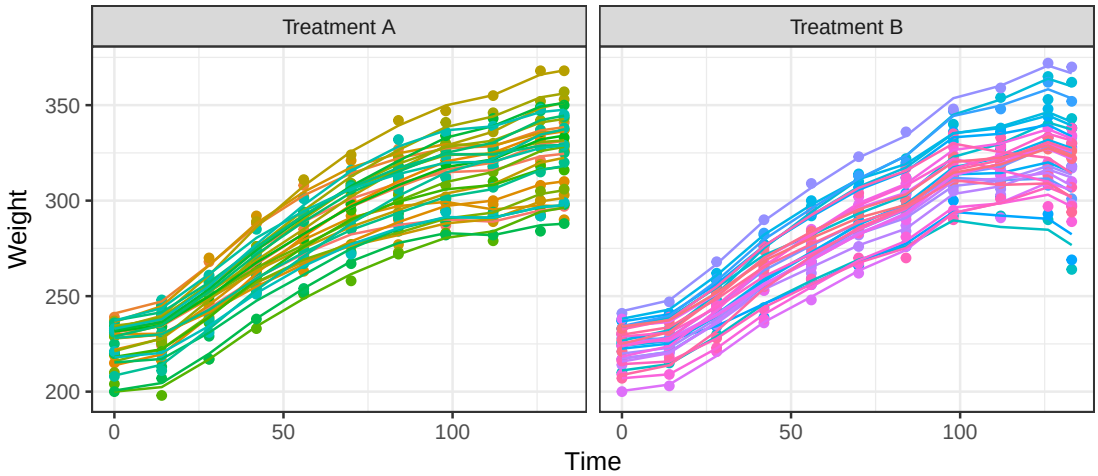


Figure 4.9: A plot of the I-prior fitted regression curves from Model 5. In this model, growth curves differ among cows and by treatment effect (with an interaction between cows and treatment effect), thus producing these 60 individual lines, one for each cow, split between their respective treatment groups (A or B).

4.5.3 Regression with a functional covariate

We illustrate the prediction of a real valued response with a functional covariate using a widely analysed data set for quality control in the food industry. The data[5] contain samples of spectrometric curve of absorbances of 215 pieces of finely chopped meat, along with their water, fat and protein content. These data are recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850–1050 nm by the Near Infrared Transmission (NIT) principle. Absorption data has not been measured continuously, but instead 100 distinct wavelengths were obtained. Figure 4.10 shows a sample of 10 such spectrometric curves.

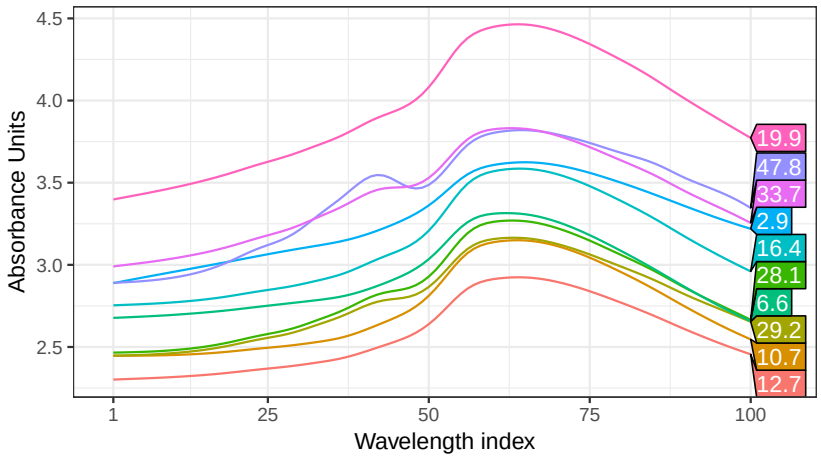


Figure 4.10: Sample of spectrometric curves used to predict fat content of meat. For each meat sample the data consists of a 100 channel spectrum of absorbances and the contents of moisture, fat (numbers shown in boxes) and protein measured in percent. The absorbance is $-\log 10$ of the transmittance measured by the spectrometer. The three contents, measured in percent, are determined by analytic chemistry.

For our analyses and many others' in the literature, the first 172 observations in the data set are used as a training sample for model fitting, and the remaining 43 observations as a test sample to evaluate the predictive performance of the fitted model. The focus here is to use the **iprior** package to fit several I-prior models to the Tecator data set, and calculate out-of-sample predictive error rates. We compare the predictive performance of I-prior models against Gaussian process regression and the many other different methods applied on this data set. These methods include neural networks (Thodberg, 1996), kernel smoothing (Ferraty and Vieu, 2006), single and multiple index functional regression models (D. Chen et al., 2011), sliced inverse regression (SIR) and sliced average variance estimation (SAVE), multivariate adaptive regression splines (MARS), partial least squares (PLS), and functional additive model with and without

[5]Obtained from Tecator (see `http://lib.stat.cmu.edu/datasets/tecator` for details). We used the version made available in the dataframe `tecator` from the R package **caret** (Kuhn et al., 2017).

component selection (FAM & CSEFAM). An analysis of this data set using the SIR and SAVE methods were conducted by Lian and Li (2014), while the MARS, PLS and (CSE)FAM methods were studied by Zhu et al. (2014). Table 4.5 tabulates the all of the results from these various references.

Assuming a regression model as in (4.10), we would like to model the `fat` content y_i using the spectral curves x_i. Let $x_i(t)$ denote the absorbance for wavelength $t = 1, \ldots, 100$. From Figure 4.10, it appears that the curves are smooth enough to be differentiable, and therefore it is reasonable to assume that they lie in the Sobolev-Hilbert space as discussed in Section 4.1.6. We take first differences of the 100-dimensional matrix, which leaves us with the 99-dimensional covariate saved in the object named `absorp`. The `fat` and `absorp` data have been split into `*.train` and `*.test` samples, as mentioned earlier. Our first modelling attempt is to fit a linear effect by regressing the responses `fat.train` against a single high-dimensional covariate `absorp.train` using the linear RKHS and the direct optimisation method.

```
R> # Model 1: Canonical RKHS (linear)
R> (mod1 <- iprior(y = fat.train, absorp.train))
```

```
## iter   10 value 222.653144
## final  value 222.642108
## converged
## Log-likelihood value: -445.2844
##
##     lambda        psi
## 4576.86595    0.11576
```

Our second and third model uses polynomial RKHSs of degrees two and three, which allows us to model quadratic and cubic terms of the spectral curves respectively. We also opted to estimate a suitable offset parameter, and this is called to `iprior()` with the option `est.offset = TRUE`. Each of the two models has a single scale parameter, an offset parameter, and an error precision to be estimated. The direct optimisation method has been used, and while both models converged regularly, it was noticed that there were multiple local optima that hindered the estimation (output omitted).

```
R> # Model 2: Polynomial RKHS (quadratic)
R> mod2 <- iprior(y = fat.train, absorp.train, kernel = "poly2",
+                 est.offset = TRUE)
R> # Model 3: Polynomial RKHS (cubic)
R> mod3 <- iprior(y = fat.train, absorp.train, kernel = "poly3",
+                 est.offset = TRUE)
```

Next, we attempt to fit a smooth dependence of fat content on the spectrometric curves using the fBm RKHS. By default, the Hurst coefficient for the fBm RKHS is set to be 0.5. However, with the option `est.hurst = TRUE`, the Hurst coefficient is included in the estimation procedure. We fit models with both a fixed value for Hurst (at 0.5) and an estimated value for Hurst. For both of these models, we encountered numerical issues when using the direct optimisation method. The L-BFGS algorithm kept on pulling the hyperparameter towards extremely high values, which in turn made the log-likelihood value greater than the machine's largest normalised floating-point number (`.Machine$double.xmax = 1.797693e+308`). To circumvent this issue, we used the EM algorithm to estimate the fixed Hurst model, and the `mixed` method for the estimated Hurst model. For both models, the `stop.crit` was relaxed and set to `1e-3` for quicker convergence, though this did not affect the predictive abilities compared to a more stringent `stop.crit`.

```
R> # Model 4: fBm RKHS (default Hurst = 0.5)
R> (mod4 <- iprior(y = fat.train, absorp.train, kernel = "fbm",
+                  method = "em", control = list(stop.crit = 1e-3)))

## ============================================
## Converged after 65 iterations.
## Log-likelihood value: -204.4592
##
##     lambda        psi
##    3.24112 1869.32897

R> # Model 5: fBm RKHS (estimate Hurst)
R> (mod5 <- iprior(fat.train, absorp.train, kernel = "fbm", method = "mixed",
+                  est.hurst = TRUE, control = list(stop.crit = 1e-3)))

## Running 5 initial EM iterations
## ============================================================================
## Now switching to direct optimisation
## iter   10 value 115.648462
## final  value 115.645800
## converged
## Log-likelihood value: -231.2923
##
##    lambda     hurst       psi
## 204.97184   0.70382   9.96498
```

Finally, we fit an I-prior model using the SE RKHS with lengthscale estimated. Here we illustrate the use of the `restarts` option, in which the model is fitted repeatedly

from different starting points. In this case, eight random initial parameter values were used and these jobs were parallelised across the eight available cores of the machine. The additional `par.maxit` option in the `control` list is an option for the maximum number of iterations that each parallel job should do. We have set it to 100, which is the same number for `maxit`, but if `par.maxit` is less than `maxit`, the estimation procedure continues from the model with the best likelihood value. We see that starting from eight different initial values, direct optimisation leads to (at least) two log-likelihood optima sites, -231.5 and -680.5.

```
R> # Model 6: SE kernel
R> (mod6 <- iprior(fat.train, absorp.train, est.lengthscale = TRUE,
+                  kernel = "se", control = list(restarts = TRUE,
+                                                par.maxit = 100)))

## Performing 8 random restarts on 8 cores
## ======================================================================
## Log-likelihood from random starts:
##     Run 1     Run 2     Run 3     Run 4     Run 5     Run 6     Run 7
## -231.5440 -680.4636 -680.4636 -680.4637 -680.4637 -231.5440 -231.5440
## Continuing on Run 6
## final  value 115.771932
## converged
## Log-likelihood value: -231.544
##
##      lambda lengthscale         psi
##    96.11515     0.09269     6.15426
```

Predicted values of the test data is obtained using `predict()`. An example for obtaining the first model's predicted values is shown below. The `predict()` method for `ipriorMod` objects also return the test MSE if the vector of test data is supplied.

```
R> predict(mod1, newdata = list(absorp.test), y.test = fat.test)

## Test RMSE: 2.890353
##
## Predicted values:
##  [1] 43.607 20.444  7.821  4.491  9.044  8.564  7.935 11.615 13.807
## [10] 17.359
## # ... with 33 more values
```

These results are summarised in Table 4.5. For the I-prior models, a linear effect of the functional covariate gives a training RMSE of 2.89, which is improved by both the qudratic and cubic model. The training RMSE is improved further by assuming a

smooth RKHS of functions for f, i.e. the fBm and SE RKHSs. When it comes to out-of-sample test error rates, the cubic model gives the best RMSE out of the I-prior models for this particular data set, with an RMSE of 0.58. This is followed closely by the fBm RKHS with estimated Hurst coefficient (fBm-0.70) and also the fBm RKHS with default Hurst coefficient (fBm-0.50). The best performing I-prior model is only outclassed by the neural networks of Thodberg (1996), who also performed model selection using automatic relevance determination (ARD). The I-prior models also give much better test RMSE than Gaussian process regression.

Table 4.5: A summary of the root mean squared error (RMSE) of prediction for the I-prior models and various other methods in literature conducted on the Tecator data set. Values for the methods under *Others* were obtained from the corresponding references cited earlier.

	RMSE	
Model	Train	Test
I-prior		
Linear	2.89	2.89
Quadratic	0.72	0.97
Cubic	0.37	0.58
Smooth (fBm-0.50)	0.00	0.68
Smooth (fBm-0.70)	0.19	0.63
Smooth (SE-0.09)	0.35	1.85
Gaussian process regression[a]		
Linear	0.18	2.36
Smooth (SE-7.04)	0.17	2.10
Others		
Neural network[b]		0.36
Kernel smoothing[c]		1.49
Single/multiple indices model[d]		1.55
Sliced inverse regression		0.90
Sliced average variance estimation		1.70
MARS[e]		0.88
Partial least squares[e]		1.01
FAM[e]		0.92
CSEFAM[e]		0.85

[a] GPR models were fit using `gausspr()` in **kernlab**.
[b] Neural network best results with automatic relevance determination (ARD) quoted.
[c] Data set used was a 160/55 training/test split.
[d] These are results of a leave-one-out cross-validation scheme.
[e] Data set used was an extended version with $n = 240$, and a random 185/55 training/test split.

4.5.4 Using the Nyström method

We investigate the use of the Nyström method of approximating the kernel matrix in estimating I-prior models. Let us revisit the data set generated by (4.19) described in Section 4.2.5. The features of this regression function are two large bumps at the centres of the mixed Gaussian PDFs, and also a small bump right after $x > 4.5$ caused by the additional exponential function. The true regression function tends to positive infinity as x increases, and to zero as x decreases. Samples of (x_i, y_i), $i = 1, \ldots, 2000$ have been generated by the built-in `gen_smooth()` function, of which the first few lines of the data are shown below.

```
R> dat <- gen_smooth(n = 2000, xlim = c(-1, 5.5), seed = 1)
R> head(dat)
```

```
##            y         X
## 1  0.6803514 -2.608953
## 2  3.6747031 -2.554039
## 3 -1.1563508 -2.381275
## 4  2.2657657 -2.280259
## 5  2.5398243 -2.214122
## 6  1.2929592 -2.170532
```

One could fit the regression model using all available data points, with an I-prior from the fBm-0.5 RKHS of functions as follows (note that the `silent` option is used to suppress the output from the `iprior()` function):

```
R> (mod.full <- iprior(y ~ X, dat, kernel = "fbm",
+                       control = list(silent = TRUE)))
```

```
## Log-likelihood value: -4355.075
##
##  lambda     psi
## 2.30244 0.23306
```

To implement the Nyström method, the option `nystrom = 50` was added to the function call, which uses 50 randomly selected data points for the Nyström approximation.

```
R> (mod.nys <- iprior(y ~ X, dat, kernel = "fbm", nystrom = 50,
+                      control = list(silent = TRUE)))
```

```
## Log-likelihood value: -1945.33
##
##  lambda     psi
## 1.64833 0.13538
```

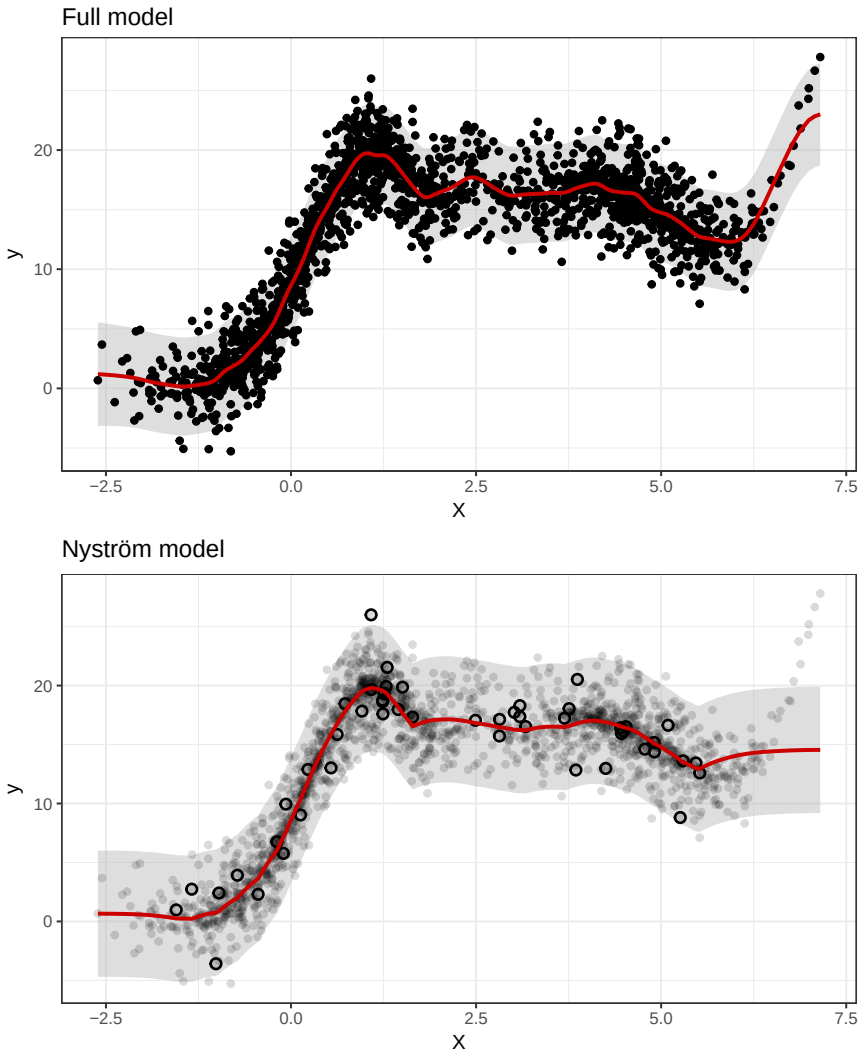


Figure 4.11: Plot of predicted regression function for the full model (top) and the Nyström approximated method (bottom). For the Nyström plot, the data points that were active are shown by circles with bold outlines.

```
R> get_time(mod.full); get_size(mod.full, "MB"); get_prederror(mod.full)

## 12.10819 mins
## 128.2 MB
## Training RMSE
##      2.054232

R> get_time(mod.nys); get_size(mod.nys); get_prederror(mod.nys)

## 1.287808 secs
## 982.2 kB
## Training RMSE
##      2.171928
```

The hyperparameters estimated for both models are slightly different. The log-likelihood is also different, but this is attributed to information loss due to the approximation procedure. Nevertheless, we see from Figure 4.11 that the estimated regression functions are quite similar in both the full model and the approximated model. The main difference is that the the Nyström method was not able to extrapolate the right hand side of the plot well, because it turns out that there were no data points used from this region. This can certainly be improved by using a more intelligent sampling scheme. The full model took a little over 12 minutes to converge, while the Nyström method took seconds without compromising too much on root mean squared error of predictions. Storage savings is significantly higher with the Nyström method as well.

4.6 Conclusion

The steps for I-prior modelling are essentially three-fold:

1. Select an appropriate function space (equivalently, kernels) for which specific effects are desired on the covariates.
2. Estimate the posterior regression function and optimise the hyperparameters, which include the RKHS scale parameter(s), error precision, and any other kernel parameters such as the Hurst index.
3. Perform post-estimation procedures such as
    - Posterior predictive checks;
    - Model comparison via log-likelihood ratio tests/empirical Bayes factors; and
    - Prediction of new data point.

The main sticking point with the estimation procedure is the involvement of the $n \times n$ kernel matrix, for which its inverse is needed. This requires $O(n^2)$ storage and $O(n^3)$ computational time. The computational issue faced by I-priors are mirrored in Gaussian process regression, so the methods to overcome these computational challenges in GPR can be explored further. However, most efficient computational solutions exploit the nature of the SE kernel structure, which is the most common kernel used in GPR. Nonetheless, we suggest the following as considerations for future work:

1. **Sparse variational approximations**. Variational methods have seen an active development in recent times. By using inducing points (Titsias, 2009) or stochastic variational inference (Hensman et al., 2013), such methods can greatly reduce computational storage and speed requirements. A recent paper by Cheng and Boots (2017) also suggests a variational algorithm with linear complexity for GPR-type models.

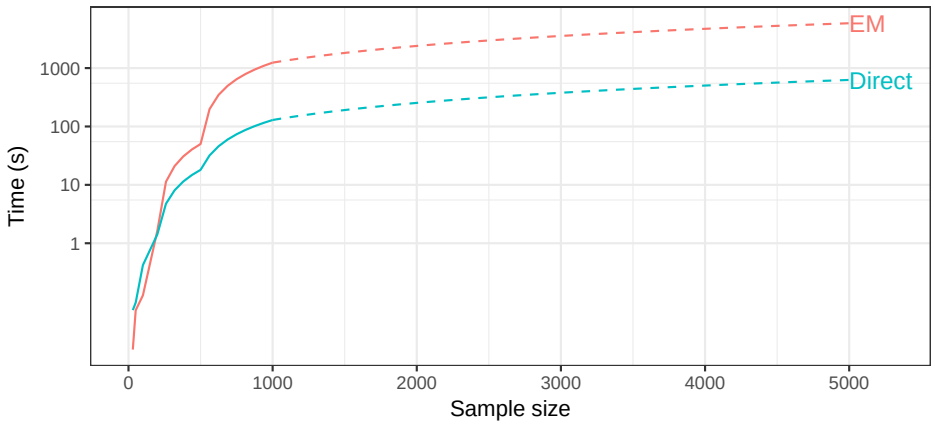


Figure 4.12: Average time taken to complete the estimation of an I-prior model (EM algorithm and direct optimisation) of varying sample sizes. The solid line represents actual timings, while the dotted lines are linear extrapolations.

2. **Accelerating the EM algorithm**. Two methods can be explored. The first is called parameter-expansion EM algorithm (PXEM) by Liu et al. (1998), which has been shown to be promising for random-effects type models. It involves correcting the M-step by a 'covariance adjustment', so that extra information can be capitalised on to improve convergence rates. The second is a quasi-Newton acceleration of the EM algorithm as proposed by Lange (1995). A slight change to the EM gradient algorithm in the M-step steers the EM algorithm to the Newton-Raphson algorithm, thus exploiting the benefits of the EM algorithm in the early stages (monotonic increase in likelihood) and avoiding the pitfalls of Newton-Raphson (getting stuck in local optima). Both algorithms require an in-depth reassessment of the EM algorithm to be tailored to I-prior models.

Chapter 5

I-priors for categorical responses

Consider polytomous response variables $\mathbf{y} = \{y_1, \ldots, y_n\}$, where each y_i takes on exactly one of the values from the set of m possible choices $\{1, \ldots, m\}$. Modelling categorical response variables is of profound interest in statistics, econometrics and machine learning, with applications aplenty. In the social sciences, categorical variables often arise from survey responses, and one may be interested in studying correlations between explanatory variables and the categorical response of interest. Economists are frequently interested in discrete choice models to explain and predict choices between several alternatives, such as consumers' choices of goods or modes of transport. In this age of big data, machine learning algorithms are used for classification of observations based on what is usually a large set of variables or features.

The model (1.1) subject to normality assumptions (1.2) is not entirely appropriate for polytomous variables $\mathbf{y}$. As an extension to the I-prior methodology, we propose a flexible modelling framework suitable for regression of categorical response variables. In the spirit of generalised linear models (McCullagh and Nelder, 1989), we relate class probabilities of the observations to a normal I-prior regression model via a *link function*. Perhaps though, it is more intuitive to view it as machine learners do: since the regression function is ranged on the entire real line, it is necessary to "squash" it through some sigmoid function to conform it to the interval $[0, 1]$ suitable for probability ranges.

Expanding on this idea further, assume that the y_i's follow a categorical distribution, $i = 1, \ldots, n$, denoted by

$$y_i \sim \text{Cat}(p_{i1}, \ldots, p_{im}),$$

with the class probabilities satisfying $p_{ij} \geq 0, \forall j = 1, \ldots, m$ and $\sum_{j=1}^{m} p_{ij} = 1$. The probability mass function (pmf) of y_i is given by

$$p(y_i) = p_{i1}^{[y_i=1]} \cdots p_{im}^{[y_i=m]},$$

where the notation $[\cdot]$ refers to the Iverson bracket[1]. The dependence of the class probabilities on the covariates is specified through the relationship

$$g(p_{i1}, \ldots, p_{im}) = \big(\alpha_1 + f_1(x_i), \ldots, \alpha_m + f_m(x_i)\big),$$

where $g : [0,1]^m \to \mathbb{R}^m$ is some specified link function. As we will see later, an underlying normal regression model as in (1.1) subject to (1.2) naturally implies a *probit* link function. With an I-prior assumed on the f_j's, we call this method of probit regression using I-priors the *I-probit* regression model.

Due to the nature of the model assumptions, unfortunately the posterior distribution of the regression functions cannot be found in closed form. In particular, marginalising the I-prior from the joint likelihood involves a high-dimensional intractable integral (c.f. equation 5.10). Similar problems are encountered in mixed logistic or probit multinomial models (Breslow and Clayton, 1993; McCulloch et al., 2000) and also Gaussian process classification (Neal, 1999; Rasmussen and Williams, 2006). In these models, Laplace approximation for maximum likelihood estimation or Markov chain Monte Carlo (MCMC) methods for Bayesian estimation are used. We instead explore a *variational approximation* to the marginal log-likelihood, and thus, to the posterior density of the regression function. The main idea is to replace the difficult posterior distribution with an approximation that is tractable to be used within an EM framework. As such, the computational work derived in the previous section is applicable for estimation of I-probit models as well.

As in the normal I-prior model, the I-probit model estimated using a *variational EM* algorithm is seen as an empirical Bayes method of estimation, since the model parameters are replaced with their ML estimates. It is emphasised again, that working in such a semi-Bayesian framework allows fast estimation of the model in comparison to traditional MCMC, yet provides us with the conveniences that come with Bayesian machinery. For example, inferences around log-odds is usually cumbersome for probit models, but a credibility interval can easily be obtained by resampling methods from the posterior distribution of the regression function, which as we shall see, is approximated to be normally distributed.

By choosing appropriate RKHSs/RKKSs for the regression functions, we are able to fit a multitude of binary and multinomial models, including multilevel or random-effects models, linear and non-linear classification models, and even spatio-temporal models. Examples of these models applied to real-world data is shown in Section 5.7. We find that the many advantages of the normal I-prior methodology transfer over quite well to the I-probit model for binary and multinomial regression.

[1] $[A]$ returns 1 if the proposition A is true, and 0 otherwise. The Iverson bracket is a generalisation of the Kronecker delta.

5.1 A latent variable motivation: the I-probit model

We derive the I-probit model through a latent variable motivation. It is convenient, as we did in Section 4.1.4, to again think of the responses $y_i \in \{1, \dots, m\}$ as comprising of a binary vector $\mathbf{y}_{i\cdot} = (y_{i1}, \dots, y_{im})^\top$, with a single '1' at the position corresponding to the value that y_i takes. That is,

$$y_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{if } y_i \neq j. \end{cases}$$

With $y_i \overset{\text{iid}}{\sim} \text{Cat}(p_{i1}, \dots, p_{im})$ for $i = 1, \dots, n$, each y_{ij} is distributed as Bernoulli with probability p_{ij}, $j = 1, \dots, m$ according to the above formulation. Now, assume that, for each $y_{i1}, \dots, y_{im}$, there exists corresponding *continuous, underlying, latent variables* $y^*_{i1}, \dots, y^*_{im}$ such that

$$y_i = \begin{cases} 1 & \text{if } y^*_{i1} \geq y^*_{i2}, y^*_{i3}, \dots, y^*_{im} \\ 2 & \text{if } y^*_{i2} \geq y^*_{i1}, y^*_{i3}, \dots, y^*_{im} \\ \vdots & \\ m & \text{if } y^*_{im} \geq y^*_{i2}, y^*_{i3}, \dots, y^*_{i\,m-1}. \end{cases} \tag{5.1}$$

In other words, $y_{ij} = \arg\max_{k=1}^m y^*_{ik}$. Such a formulation is common in economic choice models, and is rationalised by a utility-maximisation argument: an agent faced with a choice from a set of alternatives will choose the one which benefits them most. In this sense, the y^*_{ij}'s represent individual i's *latent propensities* for choosing alternative j.

Instead of modelling the observed y_{ij}'s directly, we model instead the n latent variables in each class $j = 1, \dots, m$ according to the regression problem

$$\begin{aligned} y^*_{ij} &= \alpha + \alpha_j + f_j(x_i) + \epsilon_{ij} \\ (\epsilon_{i1}, \dots, \epsilon_{im})^\top &\overset{\text{iid}}{\sim} \text{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}), \end{aligned} \tag{5.2}$$

with α being the grand intercept, α_j group or class intercepts, and $f_j : \mathcal{X} \to \mathbb{R}$ a regression function belonging to some RKKS $\mathcal{F}$ of functions over the covariate set $\mathcal{X}$ with reproducing kernel h_η. We can see some semblance of this model with the one in (4.7), and ultimately the aim is to assign I-priors to the regression function of these latent variables, which we shall describe shortly. For now, write $\boldsymbol{\mu}(x_i) \in \mathbb{R}^m$ whose j'th component is $\alpha + \alpha_j + f_j(x_i)$, and realise that each $\mathbf{y}^*_{i\cdot} = (y^*_{i1}, \dots, y^*_{im})^\top$ has the distribution $\text{N}_m(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1})$, conditional on the data x_i, the intercepts $\alpha, \alpha_1, \dots, \alpha_m$, the evaluations of the functions at x_i for each class $f_1(x_i), \dots, f_m(x_i)$, and the error covariance matrix $\boldsymbol{\Psi}^{-1}$.

The probability p_{ij} of observation i belonging to class j is then calculated as

$$\begin{aligned} p_{ij} &= \mathrm{P}(y_i = j) \\ &= \mathrm{P}\left(\{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}\right) \\ &= \int_{\{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}} \cdots \int \phi(y_{i1}^*, \ldots, y_{im}^* | \boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}) \, \mathrm{d}y_{i1}^* \cdots \mathrm{d}y_{im}^*, \end{aligned} \tag{5.3}$$

where $\phi(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density of the multivariate normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. This is the probability that the normal random variable $\mathbf{y}_{i\cdot}^*$ belongs to the set $\mathcal{C}_j := \{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}$, which are cones in $\mathbb{R}^m$. Since the union of these cones is the entire m-dimensional space of reals, the probabilities add up to one and hence they represent a proper probability mass function (pmf) for the classes. For reference, we define our *probit link function* $g_j^{-1}(\cdot|\boldsymbol{\Psi}) : \mathbb{R}^m \to [0, 1]$ by the mapping

$$\boldsymbol{\mu}(x_i) \mapsto \int_{\mathcal{C}_j} \phi(\mathbf{y}^* | \boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}) \, \mathrm{d}\mathbf{y}^*. \tag{5.4}$$

While this does not have a closed-form expression and highlights one of the difficulties of working with probit models, the integral is by no means impossible to compute—see Section 5.6.1 for a note regarding this matter.

Now, we'll see how to specify an I-prior on the regression problem (5.2). In the naïve I-prior classification model (Section 4.1.4, p. 100), we wrote $f(x_i, j) = \alpha_j + f_j(x_i)$, and called for f to belong to an ANOVA RKKS with kernel defined in (4.6). Instead of doing the same, we take a different approach. Treat the α_j's in (5.2) as intercept parameters to estimate with the additional requirement that $\sum_{j=1}^m \alpha_j = 0$. Further, let $\mathcal{F}$ be a (centred) RKHS/RKKS of functions over $\mathcal{X}$ with reproducing kernel h_η. Now, consider putting an I-prior on the regression functions $f_j \in \mathcal{F}$, $j = 1 \ldots, m$, defined by

$$f_j(x_i) = f_0(x_i, j) + \sum_{k=1}^n h_\eta(x_i, x_k) w_{ik}$$

with $\mathbf{w}_{i\cdot} := (w_{i1}, \ldots, w_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}(0, \boldsymbol{\Psi})$. This is similar to the naïve I-prior specification (4.7), except that the intercept have been treated as parameters rather than accounting for them using an RKHS of functions (Pearson RKHS or identity kernel RKHS). Importantly, the overall regression relationship still satisfies the ANOVA functional decomposition, because the α_j's sum to zero. We find that this approach, rather than the I-prior specification described in the naïve classification, bodes well down the line computationally.

We call the multinomial probit regression model of (5.1) subject to (5.2) and I-priors on $f_j \in \mathcal{F}$, the *I-probit model*. For completeness, this is stated again: for $i = 1, \ldots, n$,

$y_i = \arg\max_{k=1}^m y_{ik}^* \in \{1, \dots, m\}$, where, for $j = 1, \dots, m$,

$$
\begin{aligned}
y_{ij}^* &= \alpha + \alpha_j + \overbrace{f_0(x_i, j) + \sum_{k=1}^{n} h_\eta(x_i, x_k) w_{ik}}^{f_j(x_i)} + \epsilon_{ij} \\
\boldsymbol{\epsilon}_{i\cdot} &:= (\epsilon_{i1}, \dots, \epsilon_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}) \\
\mathbf{w}_{i\cdot} &:= (w_{i1}, \dots, w_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}).
\end{aligned}
\tag{5.5}
$$

The parameters of the I-probit model are denoted by $\theta = \{\alpha_1, \dots, \alpha_m, \eta, \boldsymbol{\Psi}\}$. To establish notation, let

- $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries ϵ_{ij}, whose rows are $\boldsymbol{\epsilon}_{i\cdot}$, columns are $\boldsymbol{\epsilon}_{\cdot j}$, and is distributed $\boldsymbol{\epsilon} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1})$;
- $\mathbf{w} \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries w_{ij}, whose rows are $\mathbf{w}_{i\cdot}$, columns are $\mathbf{w}_{\cdot j}$, and is distributed $\mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$;
- $\mathbf{f}, \mathbf{f}_0 \in \mathbb{R}^{n \times m}$ denote the matrices containing (i, j) entries $f_j(x_i)$ and $f_0(x_i, j)$ respectively, so that $\mathbf{f} = \mathbf{f}_0 + \mathbf{H}_\eta \mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{1}_n \mathbf{f}_0^\top, \mathbf{H}_\eta^2, \boldsymbol{\Psi})$;
- $\boldsymbol{\alpha} = (\alpha + \alpha_1, \dots, \alpha + \alpha_m)^\top \in \mathbb{R}^m$ be the vector of intercepts;
- $\boldsymbol{\mu} = \mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{f}$, whose (i, j) entries are $\mu_j(x_i) = \alpha + \alpha_j + f_j(x_i)$; and
- $\mathbf{y}^* \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries y_{ij}^*, that is, $\mathbf{y}^* = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, so $\mathbf{y}^* | \mathbf{w} \sim \mathrm{MN}_{n,m}(\boldsymbol{\mu} = \mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{H}_\eta \mathbf{w}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1})$ and $\operatorname{vec} \mathbf{y}^* \sim \mathrm{N}_{nm}\left(\operatorname{vec}(\mathbf{1}_n \boldsymbol{\alpha}^\top), \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n\right)$—note that the marginal distribution of $\mathbf{y}^*$ cannot be expressed as a matrix normal, except when $\boldsymbol{\Psi} = \mathbf{I}_m$.

In the above, we have made use of matrix normal distributions,denoted by $\mathrm{MN}(\cdot, \cdot)$. The definition and properties of matrix normal distributions can be found in (Appendix C.2, p. 266).

Before proceeding with estimating the I-probit model (5.5), we lay out several standing assumptions:

A4 **Centred responses**. Set $\alpha = 0$.

A5 **Zero prior mean**. Assume a zero prior mean $f_0(x) = 0$ for all $x \in \mathcal{X}$.

A6 **Fixed error precision**. Assume $\boldsymbol{\Psi}$ is fixed.

Assumption A4 is a requirement for identifiability, while A5 is motivated by a similar argument to assumption A2 in the normal I-prior model. While estimation of $\boldsymbol{\Psi}$ would add flexibility to the model, several computational issues were not able to be resolved within the time limitations of completing this project (see Section 5.6.3).

5.2 Identifiability and IIA

The parameters in the standard linear multinomial probit model is well known to be unidentified (Keane, 1992; Train, 2009), and we find this to be the case in the I-probit model as well. Unrestricted probit models are not identified for two reasons. Firstly, an addition of a non-zero constant $a \in \mathbb{R}$ to the latent variables y_{ij}^*'s in (5.1) will not change which latent variable is maximal, and therefore leaves the model unchanged. It is for this reason that assumptions A4 and A5 are imposed. Secondly, all latent variables can be scaled by some positive constant $c \in \mathbb{R}_{>0}$ without changing which latent variable is largest. This means that m-variate normal distribution $\mathrm{N}_m\left(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}\right)$ of the underlying latent variables $\mathbf{y}_{i\cdot}^*$ would yield the same class probabilities as the multivariate normal distribution $\mathrm{N}_m\left(a\mathbf{1}_m + c\boldsymbol{\mu}(x_i), c^2\boldsymbol{\Psi}^{-1}\right)$, according to (5.3). Therefore, the multinomial probit model is not identified as there exists more than one set of parameters for which the categorical likelihood $\prod_{i,j} p_{ij}$ is the same.

Identification issues in the probit model is resolved by setting one restriction on the intercepts $\alpha_1, \ldots, \alpha_m$ (location) and $m+1$ restrictions on the precision matrix $\boldsymbol{\Psi}$ (scale). Restrictions on the intercepts include $\sum_{j=1}^m \alpha_j = 0$ or setting one of the intercepts to zero. In this work, we apply the former restriction to the I-probit model, as this is analogous to the requirement of zero-mean functions in the functional ANOVA decomposition. If A6 holds, then location identification is all that is needed to achieve identification. However, if $\boldsymbol{\Psi}$ is a free parameter to be estimated, only $m(m-1)/2 - 1$ parameters are identified. Many possible specifications of the restriction on $\boldsymbol{\Psi}$ is possible, depending on the number of alternatives m and the intended effect of $\boldsymbol{\Psi}$ (to be explained shortly):

- **Case $m = 2$** (minimum number of restrictions = 3).

$$\boldsymbol{\Psi} = \begin{pmatrix} 1 & \\ 0 & 0 \end{pmatrix}, \text{ or } \boldsymbol{\Psi} = \begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}$$

- **Case $m = 3$** (minimum number of restrictions = 4).

$$\boldsymbol{\Psi} = \begin{pmatrix} 1 & & \\ \psi_{12} & \psi_{22} & \\ 0 & 0 & 0 \end{pmatrix}, \text{ or } \boldsymbol{\Psi} = \begin{pmatrix} 1 & & \\ 0 & \psi_{22} & \\ 0 & 0 & \psi_{33} \end{pmatrix}$$

- **Case $m \geq 4$** (minimum number of restrictions = $m + 1$).

$$\boldsymbol{\Psi} = \begin{pmatrix} 1 & & & & \\ \psi_{12} & \psi_{22} & & & \\ \vdots & \vdots & \ddots & & \\ \psi_{1,m-1} & \psi_{2,m-1} & \cdots & \psi_{m-1,m-1} & \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \text{ or } \boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & & & \\ & \psi_{22} & & \\ & & \ddots & \\ & & & \psi_{mm} \end{pmatrix}$$

Remark 5.1. Identification is most commonly achieved by fixing the latent propensities of one of the classes to zero and fixing one element the covariance matrix (Bunch, 1991; Dansie, 1985). Fixing the last class, say, to zero, i.e. $y^*_{im} = 0, \forall i = 1, \dots, n$ has the effect of shrinking $\boldsymbol{\Psi}$ to an $(m-1)$ matrix, and thus one more restriction needs to be made (typically, $\boldsymbol{\Psi}_{11}$ is set to one). This speaks to the fact that the absolute values of the latent propensities themselves do not matter, and only their relative differences do. We also remark that for the binary case ($m = 2$), setting the latent propensities for the second class to zero and fixing the remaining variance parameter to unity yields

$$\begin{aligned} p_{i1} &= \mathrm{P}(y^*_{i1} > y^*_{i2} = 0) \\ &= \mathrm{P}\left(\alpha_1 + f_1(x_i) + \epsilon_{i1} > 0 \,|\, \epsilon_{i1} \overset{\text{iid}}{\sim} \mathrm{N}(0,1)\right) \\ &= \Phi\big(\alpha_1 + f_1(x_i)\big) \end{aligned} \tag{5.6}$$

and $p_{i2} = 1 - \Phi\big(\alpha_1 + f_1(x_i)\big)$, $i = 1, \dots, n$—the familiar binary probit model. Note that in the binary case only one set of latent propensities need to be estimated, so we can drop the subscript '1' in the above equations. In fact, for m classes, only $m-1$ sets of regression functions need to be estimated (since one of them needs to be fixed), but in the multinomial presentation of this thesis we define regression functions for each class.

Now, we turn to a discussion of the role of $\boldsymbol{\Psi}$ in the model. In decision theory, the independence axiom states that an agent's choice between a set of alternatives should not be affected by the introduction or elimination of a choice option. The probit model is suitable for modelling multinomial data where the independence axiom, which is also known as the *independence of irrelevant alternatives* (IIA) assumption, is not desired. Such cases arise frequently in economics and social science, and the famous Red-Bus-Blue-Bus example is often used to illustrate IIA: suppose commuters face the decision between taking cars and red busses. The addition of blue busses to commuters' choices should, in theory, be more likely chosen by those who prefer taking the bus over cars. That is, assuming commuters are indifferent about the colour of the bus, commuters who are predisposed to taking the red bus would see the blue bus as an identical alternative. Yet, if IIA is imposed, then the three choices are distinct, and the fact that red and blue busses are substitutable is ignored.

To put it simply, the model is IIA if choice probabilities depend only on the choice in consideration, and not on any other alternatives. In the I-probit model, or rather, in probit models in general, choice dependency is controlled by the error precision matrix $\boldsymbol{\Psi}$. Specifically, the off-diagonal elements $\boldsymbol{\Psi}_{jk}$ capture the correlations between alternatives j and k. Allowing all $m(m+1)/2$ covariance elements of $\boldsymbol{\Psi}$ to be non-zero leads to the *full I-probit model*, and would not assume an IIA position. Figure 5.1 illustrates the covariance structure for the marginal distribution of the latent propensities, $\mathbf{V}_{y^*} = \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n$, and of the I-prior $\mathbf{V}_f = \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2$.

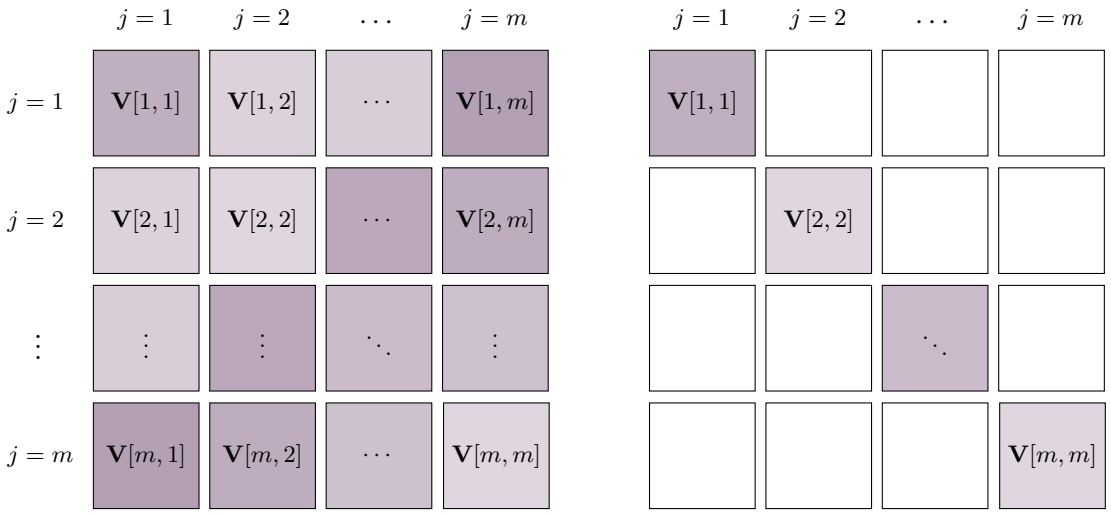


Figure 5.1: Illustration of the covariance structure of the full I-probit model (left) and the independent I-probit model (right). The full model has m^2 blocks of $n \times n$ symmetric matrices, and the blocks themselves are arranged symmetrically about the diagonal. The independent model, on the other hand, has a block diagonal structure, and its sparsity induces simpler computational methods for estimation.

While it is an advantage to be able to model the correlations across choices (unlike in logistic models), there are applications where the IIA assumption would not adversely affect the analysis, such as classification tasks. Some analyses might also be indifferent as to whether or not choice dependency exists. In these situations, it would be beneficial, algorithmically speaking, to reduce the I-probit model to a simpler version by assuming $\boldsymbol{\Psi} = \text{diag}(\psi_1, \ldots, \psi_m)$, which would trigger an IIA assumption in the I-probit model. We refer to this model as the *independent I-probit model*. The independence structure causes the distribution of the latent variables to be $y^*_{ij} \sim \text{N}(\mu_k(x_i), \sigma_j^2)$ independently for $j = 1, \ldots, m$, where $\sigma_j^2 = \psi_j^{-1}$. As a continuation of line (5.3), we can show the class probabilities p_{ij} to be

$$\begin{aligned} p_{ij} &= \int\limits_{\{y^*_{ij} > y^*_{ik} | \forall k \neq j\}} \cdots \int \prod_{k=1}^{m} \left\{ \phi(y^*_{ik} | \mu_k(x_i), \sigma_k^2) \, \mathrm{d}y^*_{ik} \right\} \\ &= \int \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{y^*_{ij} - \mu_k(x_i)}{\sigma_k} \right) \phi(y^*_{ij} | \mu_j(x_i), \sigma_j^2) \, \mathrm{d}y^*_{ij} \\ &= \text{E}_Z \left[\prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{\sigma_j}{\sigma_k} Z + \frac{\mu_j(x_i) - \mu_k(x_i)}{\sigma_k} \right) \right] \end{aligned} \tag{5.7}$$

where $Z \sim \text{N}(0, 1)$, $\Phi(\cdot)$ its cdf, and $\phi(\cdot | \mu, \sigma^2)$ is the pdf of $X \sim \text{N}(\mu, \sigma^2)$. The equation (5.3) is thus simplified to a unidimensional integral involving the Gaussian pdf and cdf, which can be computed fairly efficiently using quadrature methods.

5.3 Estimation

The premise of the I-probit model is having regression functions capture the dependence of the covariates on a latent, continuous scale using I-priors, and then transforming these regression functions onto a probability scale. Therefore, as with the normal I-prior model, an estimate of the posterior regression function with optimised hyperparameters is sought. A schematic diagram depicting the I-probit model is shown in Figure 5.2.

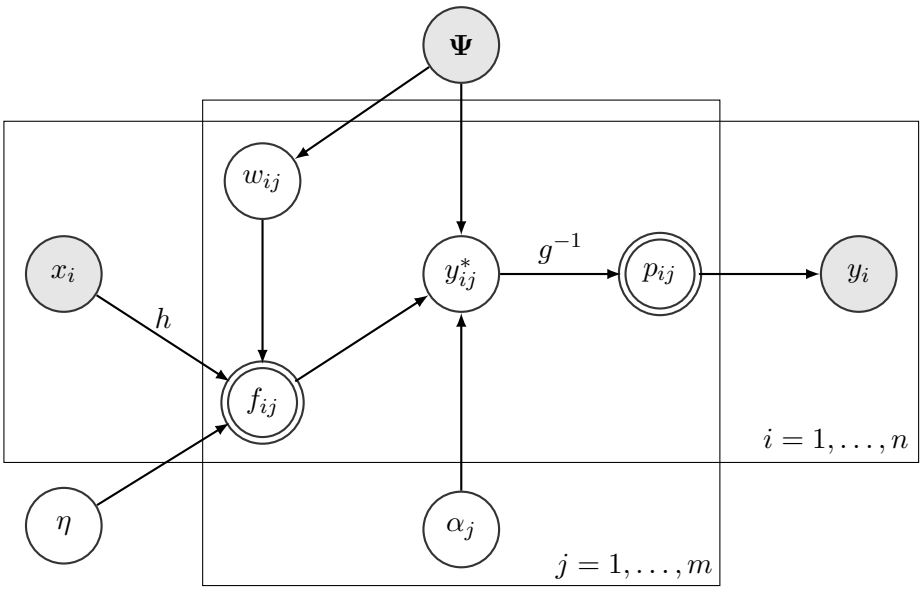


Figure 5.2: A directed acyclic graph (DAG) of the I-probit model. Observed or fixed nodes are shaded, while double-lined nodes represents calculable quantities.

The log likelihood function for θ using all n observations $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ is obtained by performing the following integration:

$$L(\theta|\mathbf{y}) = \log \iint p(\mathbf{y}|\mathbf{y}^*, \theta) p(\mathbf{y}^*|\mathbf{w}, \theta) p(\mathbf{w}|\theta) \, \mathrm{d}\mathbf{y}^* \, \mathrm{d}\mathbf{w}. \tag{5.8}$$

Here, $p(\mathbf{w}|\theta)$ is the pdf of $\mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$, $p(\mathbf{y}^*|\mathbf{w}, \theta)$ is the pdf of $\mathrm{MN}_{n,m}(\mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{H}_\eta \mathbf{w}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1})$, and $p(\mathbf{y}|\mathbf{y}^*, \theta) = \prod_{i=1}^n \prod_{j=1}^m \left[y^*_{ij} = \max \mathbf{y}^*_{i\cdot} \right]^{[y_i = j]}$, with $0^0 := 1$. Note that, given the corresponding latent propensities $\mathbf{y}^*_{i\cdot} = (y^*_{i1}, \ldots, y^*_{im})^\top$, the distribution $y_i|\mathbf{y}^*_{i\cdot}$ is tantamount to a degenerate categorical distribution: with knowledge of which latent propensities is largest, the outcome of the categorical response becomes a certainty.

The integral appearing in (5.8) is of order $2nm$, and so presents a massive computational challenge for classical numerical integration methods. This can be reduced by either integrating out the random effects $\mathbf{w}$ or the latent propensities $\mathbf{y}^*$ separately.

Continuing on (5.8) gets us to either

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y}|\mathbf{y}^*, \theta) p(\mathbf{y}^*|\theta)\, \mathrm{d}\mathbf{y}^* \\ &= \log \int \left\{ \prod_{i=1}^{n} \prod_{j=1}^{m} \left[y_{ij}^* = \max \mathbf{y}_{i\cdot}^* \right]^{[y_i=j]} \right\} \phi(\mathbf{y}^*|\mathbf{1}_n \boldsymbol{\alpha}^\top, \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\, \mathrm{d}\mathbf{y}^* \\ &= \log \int_{\bigcap_{i=1}^{n} \{y_{iy_i}^* > y_{ik}^* | \forall k \neq y_i\}} \phi(\mathbf{y}^*|\mathbf{1}_n \boldsymbol{\alpha}^\top, \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\, \mathrm{d}\mathbf{y}^*, \end{aligned} \tag{5.9}$$

by recognising that $\int p(\mathbf{y}^*|\mathbf{w}, \theta) p(\mathbf{w}|\theta)\, \mathrm{d}\mathbf{w}$ has a closed-form expression since it is an integral involving two Gaussian densities, or

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y}|\mathbf{w}, \theta)\, p(\mathbf{w}|\theta)\, \mathrm{d}\mathbf{w} \\ &= \log \int \prod_{i=1}^{n} \left\{ \prod_{j=1}^{m} \left(g_j^{-1}(\overbrace{\boldsymbol{\alpha} + \mathbf{w}^\top \mathbf{h}_\eta(x_i)}^{\boldsymbol{\mu}(x_i)} \,|\, \boldsymbol{\Psi}) \right)^{[y_i=j]} \phi(\mathbf{w}_{i\cdot}|\mathbf{0}, \boldsymbol{\Psi})\, \mathrm{d}\mathbf{w}_{i\cdot} \right\}, \end{aligned} \tag{5.10}$$

where we have denoted the class probabilities p_{ij} from (5.3) using the function $g_j^{-1}(\cdot|\boldsymbol{\Psi}) : \mathbb{R}^m \to [0, 1]$. Unfortunately, neither of these two simplifications are particularly helpful. In (5.9), the integral represents the probability of a mn-dimensional normal variate which is not straightforward to calculate, because its covariance matrix is dense. In (5.10), the integral has no apparent closed-form. The unavailability of an efficient, reliable way of calculating the log-likelihood hampers hope of obtaining parameter estimates via direct likelihood maximisation methods.

Furthermore, the posterior density of the regression function $\mathbf{f} = \mathbf{H}_\eta \mathbf{w}$, which requires the posterior density of $\mathbf{w}$ obtained via $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w}) p(\mathbf{w})$, has normalising constant equal to $L(\theta)$, which is intractable. The challenge of estimation is then to first overcome this intractability by means of a suitable approximation of the marginalising integral. We present three possible avenues to achieve this aim, namely the Laplace approximation, a variational EM algorithm, and Markov chain Monte Carlo (MCMC) methods.

5.3.1 Laplace approximation

The focus here is to obtain the posterior density $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w}) p(\mathbf{w}) =: e^{R(\mathbf{w})}$ which has normalising constant equal to the marginal density of $\mathbf{y}$, $p(\mathbf{y}) = \int e^{R(\mathbf{w})}\, \mathrm{d}\mathbf{w}$, as per (5.10). Note that the dependence of the pdfs on θ is implicit, but is dropped for clarity. Laplace's method (Kass and Raftery, 1995, §4.1.1, pp. 777–778) entails expanding a Taylor series for R about its posterior mode $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w}) p(\mathbf{w})$, which gives the

relationship

$$\begin{aligned} R(\mathbf{w}) &= R(\hat{\mathbf{w}}) + \cancelto{0}{(\mathbf{w}-\hat{\mathbf{w}})^\top \nabla R(\hat{\mathbf{w}})} - \frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w}-\hat{\mathbf{w}}) + \cdots \\ &\approx R(\hat{\mathbf{w}}) + -\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w}-\hat{\mathbf{w}}), \end{aligned}$$

because, assuming that R has a unique maximum, ∇R evaluated at its mode is zero. This is recognised as the logarithm of an unnormalised Gaussian density, implying $\mathbf{w}|\mathbf{y} \sim \mathrm{N}_n(\hat{\mathbf{w}}, \boldsymbol{\Omega}^{-1})$. Here, $\boldsymbol{\Omega} = -\nabla^2 R(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}}$ is the negative Hessian of Q evaluated at the posterior mode, and is typically obtained as a byproduct of the maximisation routine of R using gradient or quasi-gradient based methods.

The marginal distribution is then approximated by

$$\begin{aligned} p(\mathbf{y}) &= \int \exp \overbrace{R(\mathbf{w})}^{\approx\, R(\hat{\mathbf{w}}) - \frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w}-\hat{\mathbf{w}})} \mathrm{d}\mathbf{w} \\ &\approx (2\pi)^{n/2} |\boldsymbol{\Omega}|^{-1/2} e^{R(\hat{\mathbf{w}})} \int (2\pi)^{-n/2} |\boldsymbol{\Omega}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w}-\hat{\mathbf{w}})\right) \mathrm{d}\mathbf{w} \\ &= (2\pi)^{n/2} |\boldsymbol{\Omega}|^{-1/2} p(\mathbf{y}|\hat{\mathbf{w}}) p(\hat{\mathbf{w}}). \end{aligned}$$

The log marginal density of course depends on the parameters θ, which becomes the objective function to maximise in a likelihood maximising approach. Note that, should a fully Bayesian approach be undertaken, i.e. priors prescribed on the model parameters using $\theta \sim p(\theta)$, then this approach is viewed as a maximum a posteriori approach.

In any case, each evaluation of the objective function $L(\theta) = \log p(\mathbf{y}|\boldsymbol{\theta})$ involves finding the posterior modes $\hat{\mathbf{w}}$. This is a slow and difficult undertaking, especially for large sample sizes n—even assuming computation of the class probabilities is efficient—because the dimension of this integral is exactly the sample size.

Standard errors for the parameters can be obtained from diagonal entries of the information matrix involving the second derivatives of $\log p(\mathbf{y})$. However, it is not known whether the asymptotic variance of the parameters are affected by a Laplace approximation to the likelihood.

Lastly, as a comment, Laplace's method only approximates the true marginal likelihood well if the true posterior density function is small far away from the mode. In other words, a second order approximation of $R(\mathbf{w})$ must be be reliable for Laplace's method to be successful. This is typically the case if the posterior distribution is symmetric about the mode and falls quickly in the tails.

5.3.2 Variational EM algorithm

We turn to variational methods as a means of approximating the posterior densities of interest and obtain parameter estimates. Variational methods are widely discussed in the machine learning literature, but there have been efforts to popularise it in statistics (Blei et al., 2017). Although variational inference is typically seen as a fully Bayesian method, whereby approximate posterior densities are sought for the latent variables and parameters, our goal is to apply variational inference to facilitate a pseudo maximum likelihood approach.

Consider employing an EM algorithm, similar to the one seen in the previous chapter, to estimate I-probit models. This time, treat both the latent propensities $\mathbf{y}^*$ and the I-prior random effects $\mathbf{w}$ as 'missing', so the complete data is $\{\mathbf{y}, \mathbf{y}^*, \mathbf{w}\}$. Now, due to the independence of the observations $i = 1, \dots, n$, the complete data log-likelihood is

$$\begin{aligned} L(\theta|\mathbf{y}, \mathbf{y}^*, \mathbf{w}) &= \log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}|\theta) \\ &= \sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{i\cdot}^*) + \log p(\mathbf{y}^*|\mathbf{w}) + \log p(\mathbf{w}) \\ &= \text{const.} + \cancel{\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})^\top(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})\right) \\ &\quad \cancel{-\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}^{-1}\mathbf{w}^\top\mathbf{w}\right) \end{aligned} \tag{5.11}$$

which looks like the complete data log-likelihood seen previously in (4.15) (**??**, p. 107), except that here, together with $\mathbf{w}$, the $\mathbf{y}_{i\cdot}^*$'s are not observed.

For the E-step, it is of interest to determine the posterior density $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}) = p(\mathbf{y}^*|\mathbf{w}, \mathbf{y})p(\mathbf{w}|\mathbf{y})$. We have discerned from the discussion at the beginning of this section that this is hard to obtain, since it involves an intractable marginalising integral. We thus seek a suitable approximation

$$p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}, \theta) \approx \tilde{q}(\mathbf{y}^*, \mathbf{w}),$$

where $\tilde{q}$ satisfies $\tilde{q} = \arg\min_q \mathrm{D}_{\mathrm{KL}}(q\|p) = \arg\min_q \int \log \frac{q(\mathbf{y}^*, \mathbf{w})}{p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}, \theta)} q(\mathbf{y}^*, \mathbf{w})\, \mathrm{d}\mathbf{z}$, subject to certain constraints. The constraint considered by us in this thesis is that q satisfies a *mean-field* factorisation

$$q(\mathbf{y}^*, \mathbf{w}) = q(\mathbf{y}^*)q(\mathbf{w}).$$

Under this scheme, the variational distribution for $\mathbf{y}^*$ is found to be a *conically truncated multivariate normal* distribution, and for $\mathbf{w}$, a multivariate normal distribution.

It can be shown that, for any variational density q, the marginal log-likelihood is an upper-bound for the quantity $\mathcal{L}_q(\theta) := \mathcal{L}(q, \theta)$ defined by

$$\log p(\mathbf{y}|\theta) \geq \mathrm{E}_{\mathbf{y}^*,\mathbf{w}\sim q}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}|\theta)] - \mathrm{E}_{\mathbf{y}^*,\mathbf{w}\sim q}[\log q(\mathbf{y}^*, \mathbf{w})] =: \mathcal{L}(q, \theta),$$

a quantity often referred to as the *evidence lower bound* (ELBO). It turns out that minimising $\mathrm{D}_{\mathrm{KL}}(q\|p)$ is equivalent to maximising the ELBO, a quantity that is more practical to work with than the KL divergence, and certainly more tractable than the log marginal density. Hence, if q approximates the true posterior well, then the ELBO is a suitable proxy for the marginal log-likelihood.

In practice, obtaining ML parameter estimates and the posterior density $q(\mathbf{y}^*, \mathbf{w})$ which maximises the ELBO is achieved using a *variational EM algorithm*, an EM algorithm in which the conditional distribution are replaced with a variational approximation. The t'th E-step entails obtaining the density $q^{(t+1)}$ as a solution to $\arg\max_q \mathcal{L}(q, \theta)$, keeping θ fixed at the current estimate $\theta^{(t)}$. Let $\bar{\mathbf{y}}^* = \mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top$. The objective function to be maximised is computed as

$$\begin{aligned} Q(\theta) = {} & \mathrm{E}_{\mathbf{y}^*,\mathbf{w}\sim q^{(t+1)}}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}|\theta)] \\ = {} & \text{const.} - \frac{1}{2}\operatorname{tr}\Big(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top\mathbf{w}]\Big) \\ & - \frac{1}{2}\operatorname{tr}\Big(\boldsymbol{\Psi}\big\{\,\mathrm{E}[\mathbf{y}^{*\top}\mathbf{y}^*] + n\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - 2\boldsymbol{\alpha}\mathbf{1}_n^\top\,\mathrm{E}[\mathbf{y}^*] - 2\,\mathrm{E}[\mathbf{w}^\top]\mathbf{H}_\eta[\mathrm{E}[\mathbf{y}^*] - \mathbf{1}_n\boldsymbol{\alpha}^\top]\big\}\Big), \end{aligned} \tag{5.12}$$

and this is maximised with respect to θ in the M-step to obtain $\theta^{(t+1)}$ The algorithm alternates between the E- and M-step until convergence of the ELBO. A full derivation of the variational EM algorithm used by us will be described in Section 5.4.

5.3.3 Markov chain Monte Carlo methods

Markov chain Monte Carlo (MCMC) methods is the tool of choice for a complete Bayesian analysis of multinomial probit models (McCulloch et al., 2000; Nobile, 1998). Albert and Chib (1993) showed that a data augmentation approach, i.e. the latent variable approach, to probit models can be analysed using exact Bayesian methods, due to the underlying normality structure. Paired with corresponding conjugate prior choices, sampling from the posterior is very simple using a Gibbs sampling approach. That is, assuming a prior distribution on the parameters $\theta \sim p(\theta)$, the model with likelihood given by (5.8) obtains posterior samples $\{\mathbf{y}^{*(t)}, \mathbf{w}^{(t)}, \theta^{(t)}\}_{t=1}^T$ from their respective Gibbs conditional distributions. In particular, $\mathbf{y}^*|\mathbf{y}, \mathbf{w}, \theta$ is distributed according to a truncated multivariate normal, while $\mathbf{w}|\mathbf{y}, \mathbf{y}^*, \theta$ a multivariate normal. These conditional distributions are exactly of the same form as the ones obtained under a variational scheme.

The difference is that in MCMC, sampling from posterior distributions is performed, whereas in a variational inference framework, a deterministic update of the variational distributions is performed.

A downside to the data augmentation scheme for probit models in a MCMC framework is that it enlarges the variable space by an additional nm dimensions, which is memory inefficient for large n. The models with likelihood (5.9) or (5.10) after integrating out $\mathbf{w}$ and $\mathbf{y}^*$ respectively, is less demanding for MCMC sampling than the model with likelihood (5.8). However, as mentioned already, (5.9) contains an integral involving a mn-variate normal distribution whose covariance matrix is dense, and as far as we are aware, the Kronecker product structure cannot be exploited for efficiency in sampling. This leaves (5.10), a non-conjugate model whose full conditional densities are not of recognisable form. Hamiltonian Monte Carlo is another possibility, since it does not require conjugacy. For binary models, this is a feasible approach because the class probabilities normal cdfs (c.f. equation 5.6), which means that it is doable using off-the-shelf software such as Stan. However, with multinomial responses, the arduous task of computing class probabilities, which involve integration of an at most m-dimensional normal density, must be addressed separately.

5.3.4 Comparison of estimation methods

In this subsection, we utilise a toy binary classification data set which has been simulated according to a spiral pattern, as in Figure 5.3. The predictor variables are X_1 and X_2, each of which are scaled similarly. Following (5.6), the binary I-probit model that is fitted is

$$
\begin{aligned}
y_i &\sim \text{Bern}(p_i) \\
\Phi^{-1}(p_i) &= \alpha + \overbrace{\sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k}^{f(x_i)} \\
w_1, \ldots, w_n &\overset{\text{iid}}{\sim} \text{N}(0, 1),
\end{aligned}
$$

where h_λ is the (scaled) kernel of the fBm-0.5 RKHS $\mathcal{F}$ to which f belongs.

We carry out the three estimation precodures described above (Laplace's method, variational EM, and Hamiltonian MC) to compare parameter estimates, (training) error rates, and runtime. The Laplace and variational EM methods were performed in the **iprobit** package, while Stan was used to code the Hamiltonian MC sampler. Prior choices for the fully Bayesian methods were: 1) a vague folded normal prior $\lambda \sim \text{N}_+(0, 100)$ for the RKHS scale parameter, and 2) a diffuse prior for the intercept $p(\alpha) \propto \text{const}$. Note that the restriction of λ to the positive orthant is required for identifiability. The results are presented in Table 5.1.

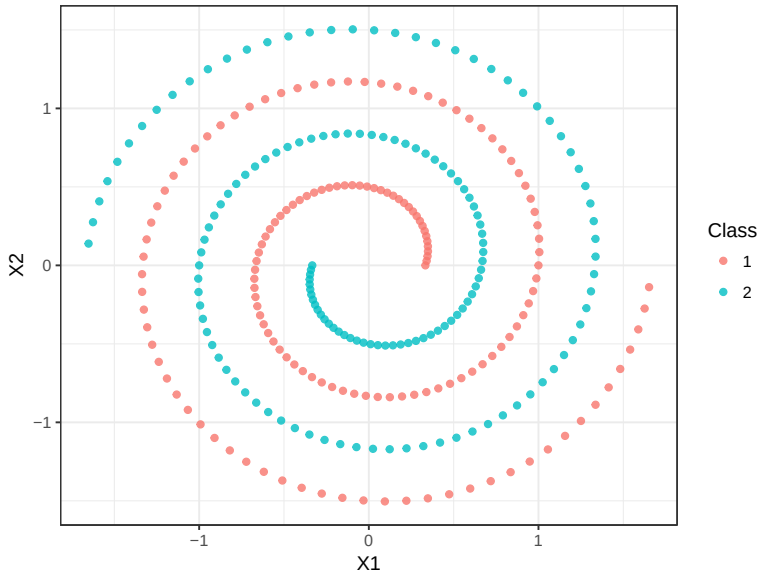


Figure 5.3: A plot of simulated spiral data set.

The three methods pretty much concur on the estimation of the intercept, but not on the RKHS scale parameter. As a result, the log-density value calculated at the parameter estimates is also different in all three methods. Notice the high posterior standard deviation for the scale parameter in the HMC method. The posterior density for λ was very positively skewed, and this contributed to the large posterior mean.

Table 5.1: Table comparing the estimated parameter values, (marginal) log-likelihood values, and also time taken for the three estimation methods.

	Laplace approximation	Variational EM	Hamiltonian MC
Intercept (α)	-0.02 (0.03)	0.00 (0.06)	0.00 (0.58)
Scale (λ)	0.85 (0.01)	5.67 (0.23)	29.3 (5.21)
Log-density	-171.8	-43.2	-8.5
Error rate (%)	44.7	0.00	0.00
Brier score	0.20	0.02	0.01
Iterations	20	56	2000
Time taken (s)	>3600	5.32	>1800

A plot of the log-likelihood (or ELBO) surface for three methods in Figure 5.4 reveals some insight. The variational likelihood has two ridges, with the maxima occurring around the intersection of these two ridges. The Laplace likelihood seems to indicate a similar shape—in both the Laplace and variational method, the posterior distribution of $\mathbf{w}$ is approximated by a Gaussian distribution, with different means and variances. However, parts of the Laplace likelihood are poorly approximated resulting in a loss of fidelity around the supposed maxima, which might have contributed to the set of values that were estimated. Laplace's method is known to yield poor approximations

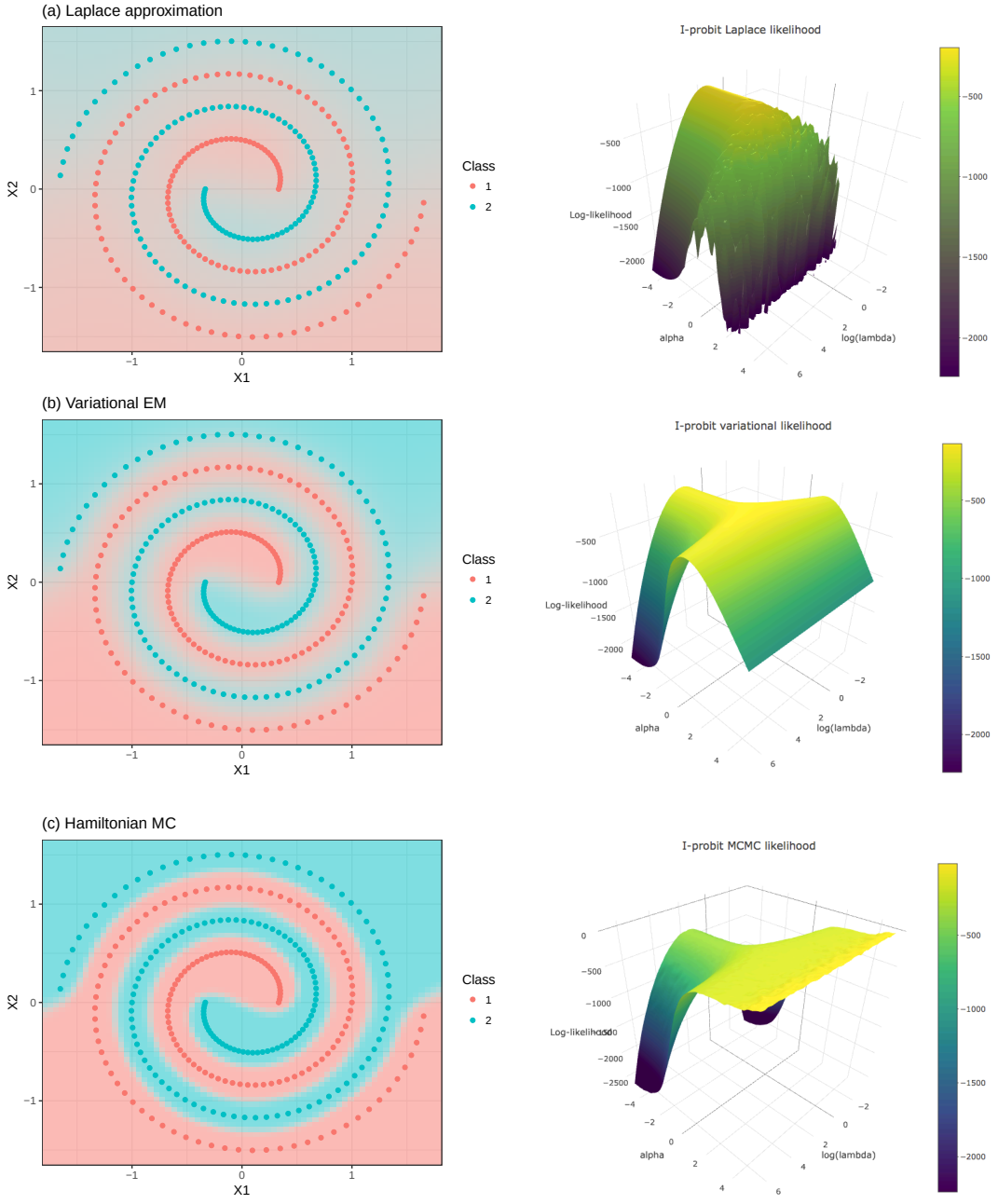


Figure 5.4: Plots showing predicted probabilities (shaded region) for belonging to class '1' or '2' indicated by colour and intensity, and log-likelihood/ELBO surface plots for (a) Laplace's method, (b) variational EM, and (c) Hamiltonian MC. For the Hamiltonian MC plot, parameters are treated as fixed, and the mean log-density of the I-probit model recorded.

to probit model likelihoods (Kuss and Rasmussen, 2005). On the other hand, the log-likelihood calculated using a Hamiltonian MC sampler (treating parameters as fixed values) yields a slightly different graph: the log-likelihood increases as values of α become larger, resulting in the upwards inflection of the log-likelihood surface (as opposed to a downward inflection seen in the variational and Laplace likelihood).

In terms of predictive abilities, both the variational and Hamiltonian MC methods, even though the posteriors are differently estimated, have good predictive performance as indicated by their error rates and Brier scores[2]. Figure 5.4 shows that HMC is more confident of new data predictions compared to variational inference, as indicated by the intensity of the shaded regions (HMC is shaded stronger than variational EM). Laplace's method gave poor predictive performance.

Finally, on the computational side, variational inference was by far the fastest method to fit the model. Sampling using Hamiltonian MC was very slow, because the parameter space is in effect $O(n+2)$ (parameters are $\{w_1, \ldots, w_n, \alpha, \lambda\}$ under the model with likelihood (5.10), i.e. without the data augmentation scheme). As for Laplace, each Newton step involves obtaining posterior modes of the w_i's, and this contributed to the slowness of this method. The reality is that variational inference takes seconds to complete what either the Laplace or full MCMC methods would take minutes or even hours to. The predictive performance, while not as good as HMC, is certainly an acceptable compromise in favour of speed.

5.4 The variational EM algorithm for I-probit models

We present an EM algorithm to estimate the I-probit latent variables $\mathbf{y}^*$ and $\mathbf{w}$, in which the E-step consists of a mean-field variational approximation of the conditional density $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}, \theta) = q(\mathbf{y}^*)q(\mathbf{w})$. As per assumptions *A4*, *A5* and *A6*, the parameters of the I-probit model consists of $\theta = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top, \eta\}$.

The algorithm cycles through a variational inference E-step, in which the variational density $q(\mathbf{y}^*, \mathbf{w}) = \prod_{i=1}^n q(\mathbf{y}_i^*)q(\mathbf{w})$ is optimised with respect to the Kullback-Leibler divergence $\mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{y}^*, \mathbf{w}) \| p(\mathbf{y}^*, \mathbf{w}|\mathbf{y})\big)$, and an M-step, in which the approximate expected joint density (5.12) is maximised with respect to the parameters θ. Convergence is assessed by monitoring the ELBO. Apart from the fact that the variational EM algorithm uses approximate conditional distributions and involves matrices $\mathbf{y}^*$ and $\mathbf{w}$, it is very similar to the EM described in Chapter 4, and as such, the efficient computational work derived there is applicable.

[2]The Brier score is defined as $\frac{1}{n}\sum_{i=1}^n \sum_{j=1}^m (y_{ij} - \hat{p}_{ij})$ with $y_{ij} = 1$ if $y_i = j$ and zero otherwise, and $\hat{p}_{ij}$ is the fitted probability $\hat{\mathrm{P}}(y_i = j)$. It gives a better sense of "training/test error", compared to simple misclassification rates, by accounting for the forecasted probabilities of the events happening. The Brier score is a proper scoring rule, i.e. it is uniquely minimised by the true probabilities.

5.4.1 The variational E-step

Let $\tilde{q}(\mathbf{y}^*, \mathbf{w})$ be the pdf that minimises the Kullback-Leibler divergence $\mathrm{D_{KL}}\left(q \| p\right)$ subject to the mean-field constraint $q(\mathbf{y}^*, \mathbf{w}) = q(\mathbf{y}^*)q(\mathbf{w})$. By appealing to Bishop (2006, equation 10.9, p. 466), the optimal mean-field variational density $\tilde{q}$ for the latent variables $\mathbf{y}^*$ and $\mathbf{w}$ satisfy

$$\log \tilde{q}(\mathbf{y}^*) = \mathrm{E}_{\mathbf{w}\sim\tilde{q}}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w})] + \text{const.} \tag{5.13}$$

$$\log \tilde{q}(\mathbf{w}) = \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w})] + \text{const.} \tag{5.14}$$

where $p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}) = p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\mathbf{w})p(\mathbf{w})$ is as per (5.8). We now present the variational densities $\tilde{q}(\mathbf{y}^*)$ and $\tilde{q}(\mathbf{w})$. For further details on the derivation of these densities, please refer to Appendix H (p. 289).

Variational distribution for the latent propensities $\mathbf{y}^*$

The fact that the rows $\mathbf{y}^*_{i\cdot} \in \mathbb{R}^m$, $i = 1, \ldots, n$ of $\mathbf{y}^* \in \mathbb{R}^{n\times m}$ are independent can be exploited, and this results in a further induced factorisation $q(\mathbf{y}^*) = \prod_{i=1}^n q(\mathbf{y}^*_i)$. Define the set $\mathcal{C}_j = \{y^*_{ij} > y^*_{ik} \,|\, \forall k \neq j\}$. Then $q(\mathbf{y}^*_{i\cdot})$ is the density of a multivariate normal distribution with mean $\tilde{\boldsymbol{\mu}}_{i\cdot} = \boldsymbol{\alpha} + \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i)$, where $\tilde{\mathbf{w}} = \mathrm{E}_{\mathbf{w}\sim\tilde{q}}[\mathbf{w}]$, and variance $\boldsymbol{\Psi}^{-1}$, subject to a truncation of its components to the set $\mathcal{C}_{y_i}$. That is, for each $i = 1, \ldots, n$ and noting the observed categorical response $y_i \in \{1, \ldots, m\}$ for the i'th observation, the $\mathbf{y}^*_i$'s are distributed according to

$$\mathbf{y}^*_{i\cdot} \overset{\text{iid}}{\sim} \begin{cases} \mathrm{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}) & \text{if } y^*_{iy_i} > y^*_{ik}, \forall k \neq y_i \\ 0 & \text{otherwise.} \end{cases} \tag{5.15}$$

We denote this by $\mathbf{y}^*_{i\cdot} \overset{\text{iid}}{\sim} {}^{\mathrm{t}}\mathrm{N}(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, and the important properties of this distribution are explored in the appendix.

The required expectation $\tilde{\mathbf{y}}^* := \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}[\mathbf{y}^*_{i\cdot}] = \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}[y^*_{i1}, \ldots, y^*_{im}]^\top$ in the M-step can be tricky to obtain. One strategy that can be considered is Monte Carlo integration: using samples from $\mathrm{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1})$, disregard those that do not satisfy the condition $y^*_{iy_i} > y^*_{ik}, \forall k \neq j$, and then take the sample average. This works reasonably well so long as the truncation region does not fall into the extreme tails of the multivariate normal. Alternatively, a Gibbs-based approach (Robert, 1995) for sampling from a truncated multivariate normal can be implemented, and this is detailed in Appendix C.4.

If the independent I-probit model is under consideration, whereby the covariance matrix has the independent structure $\boldsymbol{\Psi} = \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_m^{-2})$, then the first moment

can be considered component-wise. Each component of this expectation is given by

$$\tilde{y}_{ik}^* = \begin{cases} \tilde{\mu}_{ik} - \sigma_k C_i^{-1} \int \phi_{ik}(z) \prod_{l \neq k, y_i} \Phi_{il}(z)\phi(z)\,\mathrm{d}z & \text{if } k \neq y_i \\ \tilde{\mu}_{iy_i} - \sigma_{y_i} \sum_{k \neq y_i} \left(\tilde{y}_{ik}^* - \tilde{\mu}_{ik}\right) & \text{if } k = y_i \end{cases} \tag{5.16}$$

with

$$\phi_{ik}(Z) = \phi\left(\frac{\sigma_{y_i}}{\sigma_k} Z + \frac{\tilde{\mu}_{iy_i} - \tilde{\mu}_{ik}}{\sigma_k}\right)$$
$$\Phi_{ik}(Z) = \Phi\left(\frac{\sigma_{y_i}}{\sigma_k} Z + \frac{\tilde{\mu}_{iy_i} - \tilde{\mu}_{ik}}{\sigma_k}\right)$$
$$C_i = \int \prod_{l \neq j} \Phi_{il}(z)\phi(z)\,\mathrm{d}z$$

and $Z \sim \mathrm{N}(0,1)$ with pdf and cdf $\phi(\cdot)$ and $\Phi(\cdot)$ respectively. The integrals that appear above are functions of a unidimensional Gaussian pdf, and these can be computed rather efficiently using quadrature methods.

Variational distribution for the I-prior random effects w

Given that both $\operatorname{vec}\mathbf{y}^*|\operatorname{vec}\mathbf{w}$ and $\operatorname{vec}\mathbf{w}$ are normally distributed as per the model (5.5), we find that the full conditional distribution $p(\mathbf{w}|\mathbf{y}^*, \mathbf{y}) \propto p(\mathbf{y}^*, \mathbf{y}, \mathbf{w}) \propto p(\mathbf{y}^*|\mathbf{w})p(\mathbf{w})$ is also normal. The variational density q for $\operatorname{vec}\mathbf{w} \in \mathbb{R}^{nm}$ is found to be Gaussian with mean and precision given by

$$\operatorname{vec}\tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta)\operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top) \quad \text{and} \quad \tilde{\mathbf{V}}_w^{-1} = \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n = \mathbf{V}_{y^*}. \tag{5.17}$$

As a computational remark, computing the inverse $\tilde{\mathbf{V}}_w^{-1}$ presents a challenge, as this takes $O(n^3m^3)$ time if computed naïvely. By exploiting the Kronecker product structure in $\tilde{\mathbf{V}}_w$, we are able to efficiently compute the required inverse in roughly $O(n^3m)$ time—see Section 5.6.2 for details. Storage requirement is $O(n^2m^2)$, as a result of the covariance matrix in (5.17).

If the independent I-probit model is assumed, i.e. $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \ldots, \psi_m)$, then the posterior covariance matrix $\tilde{\mathbf{V}}_w$ has a simpler structure which implies column independence in the matrix $\mathbf{w}$. By writing $\mathbf{w}_{\cdot j} = (w_{1j}, \ldots, w_{nj})^\top \in \mathbb{R}^n$, $j = 1, \ldots, m$, to denote the column vectors of $\mathbf{w}$, and with a slight abuse of notation, we have that

$$\mathrm{N}_{nm}(\operatorname{vec}\mathbf{w}|\operatorname{vec}\tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w) = \prod_{j=1}^{m} \mathrm{N}_n(\mathbf{w}_{\cdot j}|\tilde{\mathbf{w}}_{\cdot j}, \tilde{\mathbf{V}}_{w_j}),$$

where $\mathrm{N}_d(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the pdf of $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and

$$\tilde{\mathbf{w}}_{\cdot j} = \psi_j \tilde{\mathbf{V}}_{w_j} \mathbf{H}_\eta(\tilde{\mathbf{y}}_j^* - \alpha_j \mathbf{1}_n) \quad \text{and} \quad \tilde{\mathbf{V}}_{w_j} = \left(\psi_j \mathbf{H}_\eta^2 + \psi_j^{-1}\mathbf{I}_n\right)^{-1}.$$

We note the similarity between (5.17) above and the posterior distribution for the I-prior random effects in a normal model (4.11) seen in the previous chapter, with the difference being (5.17) uses the continuous latent propensities $\mathbf{y}^*$ instead of the the observations $\mathbf{y}$. The consequence of this is that the posterior regression functions are class independent, the exact intended effect by specifying a diagonal precision matrix $\boldsymbol{\Psi}$. Storage requirement is $O(n^2m)$, since we need $\mathbf{V}_{w_1}, \dots, \mathbf{V}_{w_m}$.

Remark 5.2. The variational distribution $q(\mathbf{w})$ which approximates $p(\mathbf{w}|\mathbf{y})$ is in fact exactly $p(\mathbf{w}|\mathbf{y}^*)$, the conditional density of the I-prior random effects given the latent propensities. By the law of total expectations,

$$\mathrm{E}[r(\mathbf{w})|\mathbf{y}] = \mathrm{E}_{\mathbf{y}^*}\left[\,\mathrm{E}[r(\mathbf{w})|\mathbf{y}^*]\,\middle|\,\mathbf{y}\right],$$

where $r(\cdot)$ is some function of $\mathbf{w}$, and expectations are taken under the posterior distribution of $\mathbf{y}^*$. Hypothetically, if the true pdf $p(\mathbf{y}^*|\mathbf{y})$ were tractable, then the E-step can be computed using the true conditional distribution. Since it is not tractable, we resort to an approximation, and in the case of a variational approximation, (5.17) is obtained.

5.4.2 The M-step

From (5.12), the function to be maximised in the M-step is

$$\begin{aligned} Q(\theta) =\ & \mathrm{E}_{\mathbf{y}^*,\mathbf{w}\sim q^{(t+1)}}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}|\theta)] \\ =\ & \text{const.} - \frac{1}{2}\operatorname{tr}\Big(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top\mathbf{w}]\Big) \\ & - \frac{1}{2}\operatorname{tr}\Big(\boldsymbol{\Psi}\big\{\,\mathrm{E}[\mathbf{y}^{*\top}\mathbf{y}^*] + n\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - 2\boldsymbol{\alpha}\mathbf{1}_n^\top\,\mathrm{E}[\mathbf{y}^*] - 2\,\mathrm{E}[\mathbf{w}^\top]\mathbf{H}_\eta[\mathrm{E}[\mathbf{y}^*] - \mathbf{1}_n\boldsymbol{\alpha}^\top]\big\}\Big), \end{aligned}$$

where expectations are taken with respect to the variational distributions of $\mathbf{y}^*$ and $\mathbf{w}$. Note that since $\boldsymbol{\Psi}$ is treated as fixed, the term $\mathrm{E}[\mathbf{y}^{*\top}\mathbf{y}^*]$ is absorbed into the constant. On closer inspection, the trace involving the second moments of $\mathbf{w}$ is found to be

$$\operatorname{tr}\Big(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top\mathbf{w}]\Big) = \sum_{i,j=1}^{m}\Big\{\psi_{ij}\operatorname{tr}(\mathbf{H}_\eta^2\tilde{\mathbf{W}}_{ij}) + \psi_{ij}^-\operatorname{tr}(\tilde{\mathbf{W}}_{ij})\Big\}$$

by the results of the derivations in Appendix H.1.2 (p. 293). In the above, we had defined ψ_{ij}^- to be the (i,j)'th element of $\boldsymbol{\Psi}^{-1}$, and

$$\tilde{\mathbf{W}}_{ij} = \mathrm{E}[\mathbf{w}_{\cdot i}\mathbf{w}_{\cdot j}^\top] = \mathbf{V}_w[i,j] + \tilde{\mathbf{w}}_{\cdot i}\tilde{\mathbf{w}}_{\cdot j}^\top,$$

where $\mathbf{V}_w[i,j] \in \mathbb{R}^{n\times n}$ refers to the (i,j)'th submatrix block of $\mathbf{V}_w$, and the n-vector $\tilde{\mathbf{w}}_{\cdot j} = \big(\mathrm{E}\,w_{ij}\big)_{i=1}^n$ is the expected value of the random effects for class j. Specifically, when the error precision is of the form $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \dots, \psi_m)$, this trace reduces to

$$\operatorname{tr}\Big(\boldsymbol{\Psi}\,\mathrm{E}(\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}) + \boldsymbol{\Psi}^{-1}\,\mathrm{E}(\mathbf{w}^\top\mathbf{w})\Big) = \sum_{j=1}^{m}\Big\{\psi_j\operatorname{tr}(\mathbf{H}_\eta^2\tilde{\mathbf{W}}_{jj}) + \psi_j^{-1}\operatorname{tr}(\tilde{\mathbf{W}}_{jj})\Big\}$$
$$= \sum_{j=1}^{m}\operatorname{tr}\big(\overbrace{(\psi_j\mathbf{H}_\eta^2 + \psi_j^{-1}\mathbf{I}_n)}^{\boldsymbol{\Sigma}_{\theta,j}}\tilde{\mathbf{W}}_{jj}\big)$$

The bulk of the computational effort required to evaluate $Q(\theta)$ stems from the trace involving the second moments of $\mathbf{w}$, and the fact that $\mathbf{H}_\eta^2$ needs to be reevaluated each time $\theta = \{\boldsymbol{\alpha}, \eta\}$ changes. As discussed previously, each E-step takes $O(n^3m)$ time to compute the required first and second (approximate) posterior moments of $\mathbf{w}$. Once this is done, we can use the 'front-loading of the kernel matrices' trick described in Section 4.3.2, which effectively renders the evaluation of Q to be linear in θ (after an initial $O(n^2)$ procedure at the beginning).

As in the normal linear model, we employ a sequential update of the parameters (à la expectation conditional maximisation algorithm) by solving the first order conditions

$$\frac{\partial}{\partial\eta}Q(\eta|\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i,j=1}^{m}\psi_{ij}\operatorname{tr}\left(\frac{\partial\mathbf{H}_\eta^2}{\partial\eta}\tilde{\mathbf{W}}_{ij}\right) + \operatorname{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top\frac{\partial\mathbf{H}_\eta}{\partial\eta}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) \tag{5.18}$$

$$\frac{\partial}{\partial\boldsymbol{\alpha}}Q(\boldsymbol{\alpha}|\eta) = 2n\boldsymbol{\Psi}\boldsymbol{\alpha} - 2\sum_{i=1}^{n}\boldsymbol{\Psi}\big(\mathbf{y}_{i\cdot}^* - \tilde{\mathbf{w}}^\top\mathbf{h}_\eta(x_i)\big) \tag{5.19}$$

equated to zero, where $\mathbf{h}_\eta(x_i) \in \mathbb{R}^n$ is the i'th row of the kernel matrix $\mathbf{H}_\eta$. We now present the update equations for the parameters.

Update for kernel parameters η

When only ANOVA RKHS scale parameters are involved, then the conditional solution of η to (5.18) can be found in closed-form, much like in the exponential family EM algorithm described in Section 4.3.3 (p. 116). Under the same setting as in that subsection, assume that only $\eta = \{\lambda_1, \ldots, \lambda_p\}$ need be estimated, and for each $k = 1, \ldots, p$, we can decompose the kernel matrix as $\mathbf{H}_\eta = \lambda_k\mathbf{R}_k + \mathbf{S}_k$ and its square as $\mathbf{H}_\eta^2 = \lambda_k^2\mathbf{R}_k^2 + \lambda_k\mathbf{U}_k + \mathbf{S}_k^2$. As a follow-on from (5.18), the conditional solution for λ_k given the rest of the parameters is obtained by solving

$$\begin{aligned}\frac{\partial}{\partial\lambda_k}Q(\lambda_k|\boldsymbol{\alpha},\boldsymbol{\lambda}_{-k}) =& -\frac{1}{2}\sum_{i,j=1}^{m}\psi_{ij}\operatorname{tr}\Big((2\lambda_k\mathbf{R}_k^2 + \mathbf{U}_k)\tilde{\mathbf{W}}_{ij}\Big) + \operatorname{tr}\Big(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top\mathbf{R}_k(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\Big)\\ =& -\lambda_k\sum_{i,j=1}^{m}\psi_{ij}\operatorname{tr}(\mathbf{R}_k^2\tilde{\mathbf{W}}_{ij}) - \frac{1}{2}\sum_{i,j=1}^{m}\psi_{ij}\operatorname{tr}(\mathbf{U}_k\tilde{\mathbf{W}}_{ij})\\ &+ \operatorname{tr}\Big(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top\mathbf{R}_k(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\Big)\end{aligned}$$

equals zero. This yields the solution

$$\hat{\lambda}_k = \frac{\operatorname{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top \mathbf{R}_k(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) - \frac{1}{2}\sum_{i,j=1}^m \psi_{ij}\operatorname{tr}(\mathbf{U}_k\tilde{\mathbf{W}}_{ij})}{\sum_{i,j=1}^m \psi_{ij}\operatorname{tr}(\mathbf{R}_k^2\tilde{\mathbf{W}}_{ij})}$$

In the case of the independent I-probit model, where $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \ldots, \psi_m)$, $\hat{\lambda}_k$ has the form

$$\hat{\lambda}_k = \frac{\sum_{j=1}^m \psi_j \left(\tilde{\mathbf{w}}_{\cdot j}^\top \mathbf{R}_k(\tilde{\mathbf{y}}_{\cdot j}^* - \alpha_j\mathbf{1}_n) - \frac{1}{2}\operatorname{tr}(\mathbf{U}_k\tilde{\mathbf{W}}_{jj})\right)}{\sum_{j=1}^m \psi_j \operatorname{tr}(\mathbf{R}_k^2\tilde{\mathbf{W}}_{jj})}.$$

Remark 5.3. There is no closed-form solution for η when the polynomial kernel is used, or when there are kernel parameters to optimise (e.g. Hurst coefficient or SE kernel lengthscale). In these situations, solutions for η are obtained using numerical methods (i.e. employ quasi-Newton methods such as L-BFGS algorithm for optimising $Q(\eta)$).

Update for intercepts $\boldsymbol{\alpha}$

It is easy to see that the unique solution to (5.19) is

$$\hat{\boldsymbol{\alpha}} = \frac{1}{n}\boldsymbol{\Psi}^{-1}\left(\sum_{i=1}^n \boldsymbol{\Psi}\big(\mathbf{y}_{i\cdot}^* - \tilde{\mathbf{w}}^\top\mathbf{h}_\eta(x_i)\big)\right) = \frac{1}{n}\sum_{i=1}^n \big(\mathbf{y}_{i\cdot}^* - \tilde{\mathbf{w}}^\top\mathbf{h}_\eta(x_i)\big) \in \mathbb{R}^m.$$

Being free of $\boldsymbol{\Psi}$, the solution is the same whether the full or independent I-probit model is assumed. Furthermore, we must have that $\sum_{j=1}^m \alpha_j = 0$ for identifiability, so as an additional step to satisfy this condition, the solution $\hat{\boldsymbol{\alpha}}$ is centred.

5.4.3 Summary

Notice that the evaluation of each component of the posterior depends on knowing the posterior distribution of the other, i.e. $q(\mathbf{y}^*)$ depends on $q(\mathbf{w})$ and vice-versa. Similarly, each parameter update is obtained conditional upon the value of the rest of the parameters. These circular dependencies are dealt with by way of an iterative updating scheme: with arbitrary starting values for the distributions $q^{(0)}(\mathbf{y}^*)$ and $q^{(0)}(\mathbf{w})$, and for the parameters $\theta^{(0)}$, each are updated in turn according to the above derivations.

The updating sequence is repeated until no significant increase in the convergence criterion, the ELBO, is observed. The ELBO for the I-probit model is given by the quantity

$$\mathcal{L}_q(\theta) = \frac{nm}{2} + \sum_{i=1}^n \log C_i(\theta) + \frac{1}{2}\log|\tilde{\mathbf{V}}_w| - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i,j=1}^m \psi_{ij}^- \operatorname{tr}(\tilde{\mathbf{W}}_{ij}), \qquad (5.20)$$

where $C_i(\theta)$ is the normalising constant of the distribution ${}^{\mathrm{t}}\mathrm{N}_m(\boldsymbol{\alpha}+\tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i), \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, with $\mathcal{C}_{y_i} = \{y^*_{iy_i} > y^*_{ik} | \forall k \neq y_i\}$, and ψ^-_{ij}. That is,

$$C_i(\theta) = \int_{\{y^*_{iy_i} > y^*_{ik} \,|\, \forall k \neq y_i\}} \cdots \int \phi(y^*_{i1}, \ldots, y^*_{im} | \boldsymbol{\alpha} + \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i), \boldsymbol{\Psi}^{-1}) \, \mathrm{d}y^*_{i1} \cdots \mathrm{d}y^*_{im}.$$

Similar to the EM algorithm, each iteration of the algorithm increases the ELBO to a stationary point (Blei et al., 2017). Unlike the EM algorithm though, the variational EM algorithm does *not* guarantee an increase in the marginal log-likelihood at each step, nor does it guarantee convergence to the global maxima of the log-likelihood.

Further, the ELBO expression to be maximised is often not convex, which means the CAVI algorithm may terminate at local modes, for which they may be many. Note that the variational distribution with the higher ELBO value is the distribution that is closer, in terms of the KL divergence, to the true posterior distribution. In our experience, multiple random starts alleviates this issue for the I-probit model.

5.5 Post-estimation

Post-estimation procedures such as obtaining predictions for a new data point, the credibility interval for such predictions, and model comparison, are of interest. These are performed in an empirical Bayes manner using the variational posterior density of the regression function obtained from the output of the variational EM algorithm.

We first describe prediction of a new data point x_{new}. Step one is to determine the distribution of the posterior regression functions in each class, $\mathbf{f}(x_{\text{new}}) = \mathbf{w}^\top \mathbf{h}_\eta(x_{\text{new}})$, where $\mathbf{h}_\eta(x_{\text{new}})$ is the vector of length n containing entries $h_\eta(x_i, x_{\text{new}})$, given values for the parameters θ of the I-probit model. To this end, we use the ELBO estimates for θ, i.e. $\hat{\theta} = \arg\max_\theta \mathcal{L}_q(\theta)$, as obtained from the variational EM algorithm. As we know, the variational distribution of $\operatorname{vec} \mathbf{w}$ is normally distributed with mean and variance according to (5.17). By writing $\operatorname{vec} \tilde{\mathbf{w}} = (\tilde{\mathbf{w}}_{\cdot 1}, \ldots, \tilde{\mathbf{w}}_{\cdot m})^\top$ to separate out the I-prior random effects per class, we have that $\mathbf{w}_{\cdot j} | \hat{\theta} \sim \mathrm{N}_n(\tilde{\mathbf{w}}_{\cdot j}, \tilde{\mathbf{V}}_w[j,j])$, and $\operatorname{Cov}[\mathbf{w}_{\cdot j}, \mathbf{w}_{\cdot k}] = \tilde{\mathbf{V}}_w[j,k]$, where the '$[\cdot,\cdot]$' indexes the $n \times n$ sub-block of the block matrix structured matrix $\mathbf{V}_w$. Thus, for each class $j = 1, \ldots, m$ and any $x \in \mathcal{X}$,

$$f_j(x) | \mathbf{y}, \hat{\theta} \sim \mathrm{N}\left(\mathbf{h}_{\hat{\eta}}(x)^\top \tilde{\mathbf{w}}_{\cdot j} \,,\, \mathbf{h}_{\hat{\eta}}(x)^\top \tilde{\mathbf{V}}_w[j,j] \mathbf{h}_{\hat{\eta}}(x) \right),$$

and the covariance between the regression functions in two different classes is

$$\operatorname{Cov}\left[f_j(x), f_k(x) | \mathbf{y}, \hat{\theta} \right] = \mathbf{h}_{\hat{\eta}}(x)^\top \tilde{\mathbf{V}}_w[j,k] \tilde{\mathbf{h}}_{\hat{\eta}}(x).$$

Algorithm 2 Variational EM for the I-probit model (fixed $\boldsymbol{\Psi}$)

procedure INITIALISATION
  Initialise $\theta^{(0)} \leftarrow \{\boldsymbol{\alpha}^{(0)}, \eta^{(0)}\}$
  $\tilde{q}^{(0)}(\mathbf{w}) \leftarrow \text{MN}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$
  $\tilde{q}^{(0)}(\mathbf{y}^*_{i\cdot}) \leftarrow {}^{\text{t}}\text{N}_m(\tilde{\boldsymbol{\alpha}}^{(0)}, \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$
  $t \leftarrow 0$
end procedure

while not converged **do**
  procedure VARIATIONAL E-STEP
    for $i = 1, \ldots, n$ **do** ▷ Update $\mathbf{y}^*$
      $\tilde{q}^{(t+1)}(\mathbf{y}^*_{i\cdot}) \leftarrow {}^{\text{t}}\text{N}_m\left(\tilde{\boldsymbol{\alpha}}^{(t)} + \tilde{\mathbf{w}}^{(t)\top}\mathbf{h}_{\eta^{(t)}}(x_i), \boldsymbol{\Psi}, \mathcal{C}_{y_i}\right)$
      $\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} \leftarrow \text{E}_{q^{(t+1)}}(\mathbf{y}^*_{i\cdot})$
    end for

    $\tilde{\mathbf{V}}_w^{(t+1)} \leftarrow \left((\boldsymbol{\Psi} \otimes \mathbf{H}^2_{\eta^{(t)}}) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right)^{-1}$ ▷ Update $\mathbf{w}$
    $\text{vec}\,\tilde{\mathbf{w}}^{(t+1)} \leftarrow \tilde{\mathbf{V}}_w^{(t+1)}(\boldsymbol{\Psi} \otimes \mathbf{H}_{\eta^{(t)}})\,\text{vec}(\tilde{\mathbf{y}}^{*(t+1)} - \mathbf{1}_n\boldsymbol{\alpha}^{(t)\top})$
    $\tilde{q}^{(t+1)}(\mathbf{w}) \leftarrow \text{N}_{nm}\left(\text{vec}\,\tilde{\mathbf{w}}^{(t+1)}, \tilde{\mathbf{V}}_w^{(t+1)}\right)$
  end procedure

  procedure M-STEP
    if ANOVA kernel (closed-form updates) **then** ▷ Update η
      for $k = 1, \ldots, p$ **do**
        $T_{1k} \leftarrow \sum_{i,j=1}^m \psi_{ij}\,\text{tr}(\mathbf{R}_k^2\tilde{\mathbf{W}}_{ij})$
        $T_{2k} \leftarrow \text{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top\mathbf{R}_k(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) - \frac{1}{2}\sum_{i,j=1}^m \psi_{ij}\,\text{tr}(\mathbf{U}_k\tilde{\mathbf{W}}_{ij})$
        $\lambda_k^{(t+1)} \leftarrow T_{2k}/T_{1k}$
      end for
    else
      $\eta^{(t+1)} \leftarrow \arg\max_\eta Q(\eta|\boldsymbol{\alpha}^{(t)})$ by L-BFGS algorithm
    end if

    $\mathbf{a} \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} - \tilde{\mathbf{w}}^{(t+1)\top}\tilde{\mathbf{h}}_{\eta^{(t+1)}}(x_i)\right)$ ▷ Update $\boldsymbol{\alpha}$
    $\boldsymbol{\alpha}^{(t+1)} \leftarrow \mathbf{a} - \frac{1}{m}\sum_{j=1}^m a_j$
  end procedure

  Calculate ELBO $\mathcal{L}^{(t+1)}$
  $t \leftarrow t + 1$

  $\{\tilde{q}(\mathbf{y}^*), \tilde{q}(\mathbf{w}), \hat{\theta}\} \leftarrow \{\tilde{q}^{(t)}(\mathbf{y}^*), \tilde{q}^{(t)}(\mathbf{w}), \theta^{(t)}\}$
  return Variational densities $\{\tilde{q}(\mathbf{y}^*), \tilde{q}(\mathbf{w})\}$
  return Estimates $\{\hat{\boldsymbol{\alpha}}, \hat{\eta}\}$
  return ELBO $\mathcal{L}_q(\theta) = \mathcal{L}^{(t)}$
end while

Then, in step two, using the results obtained in the previous chapter in Section 4.4 (p. 119), we have that the latent propensities $y^*_{\text{new},j}$ for each class are normally distributed with mean, variance, and covariances

$$\begin{aligned}
\mathrm{E}[y^*_{\text{new},j}|\mathbf{y},\hat{\theta}] &= \hat{\alpha}_j + \mathrm{E}\left[f_j(x_{\text{new}})|\mathbf{y},\hat{\theta}\right] &&=: \hat{\mu}_j(x_{\text{new}}) \\
\mathrm{Var}[y^*_{\text{new},j}|\mathbf{y},\hat{\theta}] &= \mathrm{Var}\left[f(x_{\text{new}})|\mathbf{y},\hat{\theta}\right] + \boldsymbol{\Psi}^{-1}_{jj} &&=: \hat{\sigma}^2_j(x_{\text{new}}) \\
\mathrm{Cov}[y^*_{\text{new},j}, y^*_{\text{new},k}|\mathbf{y},\hat{\theta}] &= \mathrm{Cov}\left[f_j(x), f_k(x)|\mathbf{y},\hat{\theta}\right] + \boldsymbol{\Psi}^{-1}_{jk} &&=: \hat{\sigma}_{jk}(x_{\text{new}}).
\end{aligned}$$

From here, step three would be to extract class information of data point x_{new}, which are contained in the normal distribution $\mathrm{N}_m\left(\hat{\boldsymbol{\mu}}_{\text{new}}, \hat{\mathbf{V}}_{\text{new}}\right)$, where

$$\hat{\boldsymbol{\mu}}_{\text{new}} = \big(\mu_1(x_{\text{new}}), \dots, \mu_m(x_{\text{new}})\big)^\top \quad \text{and} \quad \hat{\mathbf{V}}_{\text{new},jk} = \begin{cases} \hat{\sigma}^2_j(x_{\text{new}}) & \text{if } j = k \\ \hat{\sigma}_{jk}(x_{\text{new}}) & \text{if } j \neq k. \end{cases}$$

The predicted class is inferred from the latent variables via

$$\hat{y}_{\text{new}} = \arg\max_k \hat{\mu}_k(x_{\text{new}}),$$

while the probabilities for each class are obtained via integration of a multivariate normal density, as per (5.3):

$$\hat{p}_{\text{new},j} = \int\limits_{\{y^*_j > y^*_k | \forall k \neq j\}} \cdots \int \phi(y^*_1, \dots, y^*_m | \hat{\boldsymbol{\mu}}_{\text{new}}, \hat{\mathbf{V}}_{\text{new}}) \, \mathrm{d}y^*_1 \cdots \mathrm{d}y^*_m. \tag{5.21}$$

For the independent I-probit model, class probabilities are obtained in a more compact manner via

$$\hat{p}_{\text{new},j} = \mathrm{E}_Z \left[\prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{\hat{\sigma}_j(x_{\text{new}})}{\hat{\sigma}_k(x_{\text{new}})} Z + \frac{\hat{\mu}_j(x_{\text{new}}) - \hat{\mu}_k(x_{\text{new}})}{\hat{\sigma}^2_k(x_{\text{new}})} \right) \right],$$

as per (5.7), since the m components of $\mathbf{f}(x_{\text{new}})$, and hence the $\mathbf{y}^*_{\text{new},j}$'s, are independent of each other ($\boldsymbol{\Psi}$ and $\hat{\mathbf{V}}_{\text{new}}$ are diagonal). Prediction of a single new data point takes $O(n^2 m)$ time, because there are essentially m I-prior posterior regression functions, and each take $O(n^2)$ to evaluate. This is assuming negligible time to compute the class probabilities.

We are able to take advantage of the Bayesian machinery to obtain credibility intervals for probability estimates or any transformation of these probabilities (e.g. log odds or odds ratios). The procedure is as follows. First, obtain samples $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(T)}$ by drawing from its variational posterior distribution $\operatorname{vec} \mathbf{w}^{(t)}|\hat{\theta} \sim \mathrm{N}_{nm}(\operatorname{vec} \tilde{\mathbf{w}}, \mathbf{V}_w)$. Then, obtain samples of class probabilities $\{p^{(1)}_{xj}, \dots, p^{(T)}_{xj}\}_{j=1}^m$, for a given data point $x \in \mathcal{X}$ by

evaluating

$$p_{xj}^{(t)} = \int_{\{y_j^* > y_k^* | \forall k \neq j\}} \cdots \int \phi\big(y_1^*, \ldots, y_m^* | \hat{\boldsymbol{\mu}}^{(t)}(x), \hat{\mathbf{V}}(x)\big)\, \mathrm{d}y_1^* \cdots \mathrm{d}y_m^*,$$

where $\hat{\boldsymbol{\mu}}^{(t)}(x) = \hat{\boldsymbol{\alpha}} + \mathbf{w}^{(t)\top}\mathbf{h}_{\hat{\eta}}(x)$, and $\hat{\mathbf{V}}(x)_{jk}$ equals $\hat{\sigma}_j^2(x)$ if $j = k$, and $\hat{\sigma}_{jk}(x)$ otherwise. To obtain a statistic of interest, say, a 95% credibility interval of a function $r(p_{xj})$ of the probabilities, simply take the empirical lower 2.5th and upper 97.5th percentile of the transformed sample $\big\{r(p_{xj}^{(1)}), \ldots, r(p_{xj}^{(T)})\big\}$.

Remark 5.4. Unfortunately, with the variational EM algorithm, standard errors for the parameters θ are not so easy to obtain. We could not ascertain as to the availability of an unbiased estimate of the asymptotic covariance matrix for θ under a variational framework. One strategy for obtaining standard errors is bootstrap (Y.-C. Chen et al., 2018):

1. Obtain $\hat{\theta} = \arg\max_\theta \mathcal{L}_q(\theta)$ using $\mathcal{S} = \{(y_1, x_1), \ldots, (y_n, x_n)\}$.
2. For $t = 1, \ldots, T$, do
   (a) Obtain $\mathcal{S}^{(t)} = \{(y_1^{(t)}, x_1^{(t)}), \ldots, (y_n^{(t)}, x_n^{(t)})\}$ by sampling n points with replacement from $\mathcal{S}$.
   (b) Compute $\hat{\theta}^{(t)} = \arg\max_\theta \mathcal{L}_q(\theta)$ using the data $\mathcal{S}^{(t)}$.
3. For the l-th component of θ, compute its variance estimator using

$$\widehat{\mathrm{Var}}(\hat{\theta}_l) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_l^{(t)} - \bar{\theta}_l)^2 \quad \text{where} \quad \bar{\theta}_l = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_l^{(t)}.$$

The obvious downside to this bootstrap scheme is computational time.

Finally, a discussion on model comparison, which, in the variational inference literature, is achieved by comparing ELBO values of competing models (Beal and Ghahramani, 2003). The rationale is that the ELBO serves as a conservative estimate for the log marginal likelihood, which would allow model selection via (empirical) Bayes factors. This stems from the fact that

$$\log p(\mathbf{y}|\theta) = \mathcal{L}_q(\theta) + \mathrm{D}_{\mathrm{KL}}(q\|p) > \mathcal{L}_q(\theta),$$

since the Kullback-Leibler divergence from the true posterior density $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y})$ to the variational density $q(\mathbf{y}^*, \mathbf{w})$ is strictly positive (it is zero if and only if the two densities are equivalent), and is minimised under a variational inference scheme. Kass and Raftery (1995) suggest Section 5.5 as a way of interpreting observed Bayes factor values $\mathrm{BF}(M_1, M_0)$ for comparing model M_1 against model M_0, where $\mathrm{BF}(M_1, M_0)$ is approx-

imated by

$$\text{BF}(M_1, M_0) \approx \frac{\mathcal{L}_q(\theta|M_1)}{\mathcal{L}_q(\theta|M_0)},$$

and $\mathcal{L}_q(\theta|M_k)$, $k = 0, 1$, is the ELBO for model M_k. It should be noted that while this works in practice, there is no theoretical basis for model comparison using the ELBO (Blei et al., 2017).

Table 5.2: Guidelines for interpreting Bayes factors (Kass and Raftery, 1995).

$2 \log \text{BF}(M_1, M_0)$	$\text{BF}(M_1, M_0)$	Evidence against M_0
0–2	1–3	Not worth more than a bare mention
2–6	3–20	Positive
6–10	20–150	Strong
>10	>150	Very strong

Remark 5.5. In the previous chapter on normal I-prior models, the I-prior could be integrated out of the model completely, resulting in a normal log-likelihood for the parameters. Model comparison can be validly done using likelihood ratio tests and asymptotic chi-square distributions. Here however, we only have a lower bound to the log-likelihood, and most likely the asymptotic results of likelihood ratio tests do not hold. Then, the concept of approximate (empirical) Bayes factors seem most intuitive, even if not rooted in theory.

5.6 Computational considerations

Computational challenges for the I-probit model stems from two sources: 1) calculation of the class probabilities (5.3); and 2) storage and time requirements for the variational EM algorithm. Ways in which to overcome these challenges are discussed. In addition, we also discuss considerations to take into account if estimation of the error precision $\boldsymbol{\Psi}$ is desired, and thus pave the way for future work.

5.6.1 Efficient computation of class probabilities

The issue at hand here is that for $m > 4$, the evaluation of the class probabilities in (5.3) is computationally burdensome using classical methods such as quadrature methods Geweke et al. (1994). As such, simulation techniques (Monte Carlo integration) are employed instead. The simplest strategy to overcome this is a frequency simulator (otherwise known as Monte Carlo integration): obtain random samples from $\text{N}_m\left(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}\right)$, and calculate how many of these samples fall within the required region. This method is fast and yields unbiased estimates of the class probabilities. However, in an extensive comparative study of various probability simulators, Hajivassiliou et al. (1996) concluded

that the Geweke-Hajivassiliou-Keane (GHK) probability simulator (Geweke, 1989; Hajivassiliou and McFadden, 1998; Keane and Wolpin, 1994) is the most reliable under a multitude of scenarios. This is now described, and for clarity, we drop the subscript i denoting individuals.

Suppose that an observation $y = j$ has been made. Reformulate $\mathbf{y}^*$ in (5.1) by anchoring on the j'th latent variable y_j^* to obtain

$$\mathbf{z} := (\overbrace{y_1^* - y_j^*}^{z_1}, \dots, \overbrace{y_{j-1}^* - y_j^*}^{z_{j-1}}, \overbrace{y_{j+1}^* - y_j^*}^{z_j}, \dots, \overbrace{y_m^* - y_j^*}^{z_{m-1}},)^\top \in \mathbb{R}^{m-1}.$$

Note that we have indexed the vector $\mathbf{z}$ using $j' = k$ if $k < j$, and $j' = k - 1$ if $k > j$ for $k = 1, \dots, m$, so that the index j' runs from 1 to $m - 1$. Let $\mathbf{Q} \in \mathbb{R}^{(m-1)\times m}$ be a matrix formed by inserting a column of minus ones at the j'th position in an $(m - 1)$ identity matrix. We can then write $\mathbf{z} = \mathbf{Q}\mathbf{y}^*$, and thus we have that $\mathbf{z} \sim \mathrm{N}_{m-1}(\boldsymbol{\nu}_{(j)}, \boldsymbol{\Omega}_{(j)})$, where $\boldsymbol{\nu}_{(j)} = \mathbf{Q}\boldsymbol{\mu}(x_i)$ and $\boldsymbol{\Omega}_{(j)} = \mathbf{Q}\boldsymbol{\Psi}^{-1}\mathbf{Q}^\top$. These are indexed by '$(j)$' because the transformation is dependent on which latent variable the $\mathbf{z}$'s are anchored on.

Remark 5.6. Incidentally, the probit model in (5.1) is equivalently represented by

$$y_i = \begin{cases} 1 & \text{if } \max(y_{i2}^* - y_{i1}^*, \dots, y_{im}^* - y_{i1}^*) < 0 \\ j & \text{if } \max(y_{i2}^* - y_{i1}^*, \dots, y_{im}^* - y_{i1}^*) = y_{ij}^* - y_{i1}^* \geq 0, \end{cases} \tag{5.22}$$

which is obtained by anchoring on the first latent variable (referred to as the reference category), although the choice of reference category is arbitrary. This is similar to fixing the latent variables of the reference category to zero, and thus, as discussed previously in Section 5.2, full identification is achieved by fixing one more element of the covariance matrix.

For the symmetric and positive definite covariance matrix $\boldsymbol{\Omega}_{(j)}$, obtain its Cholesky decomposition as $\boldsymbol{\Omega}_{(j)} = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a lower triangular matrix. Then, $\mathbf{z} = \boldsymbol{\nu}_{(j)} + \mathbf{L}\boldsymbol{\zeta}$, where $\boldsymbol{\zeta} \sim \mathrm{N}_{m-1}(\mathbf{0}, \mathbf{I}_{m-1})$. That is,

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_{m-1} \end{pmatrix} = \begin{pmatrix} \nu_{(j)1} \\ \nu_{(j)2} \\ \vdots \\ \nu_{(j)m-1} \end{pmatrix} + \begin{pmatrix} L_{11} & & & \\ L_{21} & L_{22} & & \\ \vdots & \vdots & \ddots & \\ L_{m-1,1} & L_{m-1,2} & \cdots & L_{m-1,m-1} \end{pmatrix} \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_{m-1} \end{pmatrix}$$
$$= \begin{pmatrix} \nu_{(j)1} + L_{11}\zeta_1 \\ \nu_{(j)2} + \sum_{k=1}^{2} L_{k2}\zeta_k \\ \vdots \\ \nu_{(j)m-1} + \sum_{k=1}^{m-1} L_{k,m-1}\zeta_k \end{pmatrix}.$$

With this setup, the probability p_j of an observation belonging to class j, which is equivalent to the probability that each $z_{j'} < 0$, $j' = 1, \dots, m-1$, can be expressed as

$$\begin{aligned} p_j &= \mathrm{P}(z_1 < 0, \dots, z_{m-1} < 0) \\ &= \mathrm{P}(\zeta_1 < u_1, \dots, \zeta_{m-1} < u_{m-1}) \\ &= \mathrm{P}(\zeta_1 < u_1)\,\mathrm{P}(\zeta_2 < u_2 | \zeta_1 < u_1)\,\mathrm{P}(\zeta_3 < u_3 | \zeta_1 < u_1, \zeta_2 < u_2) \cdots \\ &\quad \cdots \mathrm{P}(\zeta_{m-1} < u_{m-1} | \zeta_1 < u_1, \dots, \zeta_{m-2} < u_{m-2}), \end{aligned}$$

where

$$u_{j'} = u_{j'}(\zeta_1, \dots, \zeta_{j'-1}) = \begin{cases} -\nu_{(j)1}/L_{11} & \text{for } j' = 1 \\ -\big(\nu_{(j)j'} + \sum_{k=1}^{j'-1} L_{kj'}\zeta_k\big)/L_{j'j'} & \text{for } j' = 2, \dots, m-1 \end{cases}$$

The GHK algorithm entails making draws from one-sided right truncated standard normal distributions (for instance, using an inverse transform method detailed in Appendix C.3, p. 267):

- Draw $\tilde{\zeta}_1 \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, u_1)$.
- Draw $\tilde{\zeta}_2 \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_2)$, where $\tilde{u}_2 = u_2(\tilde{\zeta}_1)$.
- Draw $\tilde{\zeta}_3 \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_3)$, where $\tilde{u}_3 = u_3(\tilde{\zeta}_1, \tilde{\zeta}_2)$.
- $\cdots$
- Draw $\tilde{\zeta}_{m-1} \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_{m-2})$, where $\tilde{u}_{m-1} = u_m(\tilde{\zeta}_1, \dots, \tilde{\zeta}_{m-2})$.

These values are then used in the following manner:

- Use $\tilde{\zeta}_1$ to obtain a "draw" of $\mathrm{P}(\zeta_2 < u_2 | \zeta_1 < \zeta_1)$,

$$\begin{aligned} \widetilde{\mathrm{P}}(\zeta_2 < u_2 | \zeta_1 < \zeta_1) &= \mathrm{P}(\zeta_2 < u_2 | \zeta_1 = \tilde{\zeta}_1) \\ &= \Phi\Big(- \big(\nu_{(j)2} + L_{12}\tilde{\zeta}_1\big)/L_{22}\Big) \end{aligned}$$

- Use $\tilde{\zeta}_1$ and $\tilde{\zeta}_2$ to obtain a "draw" of $\mathrm{P}(\zeta_3 < u_3 | \zeta_1 < u_1, \zeta_2 < u_2)$,

$$\begin{aligned} \widetilde{\mathrm{P}}(\zeta_3 < u_3 | \zeta_1 < u_1, \zeta_2 < u_2) &= \mathrm{P}(\zeta_3 < u_3 | \zeta_1 = \tilde{\zeta}_1, \zeta_2 = \tilde{\zeta}_2) \\ &= \Phi\Big(- \big(\nu_{(j)3} + L_{13}\tilde{\zeta}_1 + L_{23}\tilde{\zeta}_2\big)/L_{33}\Big) \end{aligned}$$

- And so on.

Therefore, a simulated probability for p_j (denoted with a tilde) is obtained as

$$\tilde{p}_j = \Phi\left(-\nu_{(j)1}/L_{11}\right) \prod_{j'=2}^{m-1} \Phi\left(-\left(\nu_{(j)j'} + \textstyle\sum_{k=1}^{j'-1} L_{kj'}\tilde{\zeta}_k\right)/L_{j'j'}\right). \tag{5.23}$$

By performing the above scheme T number of times to obtain T such simulated probabilities $\{p_j^{(1)}, \ldots, p_j^{(T)}\}$, the actual probability of interest p_j is then approximated by the sample mean of the draws,

$$\hat{p}_j = \frac{1}{T}\sum_{t=1}^{T} p_j^{(t)}.$$

If it so happens that one of the standard normal cdfs in (5.23) is extremely small, this can cause loss of significance due to floating-point errors (catastrophic cancellation). It is better to work on a log-probability scale, so the products in (5.23) turn into sums, and revert back by exponentiating.

Remark 5.7. The GHK algorithm provides reasonably fast and accurate calculations of class probabilities when $\boldsymbol{\Psi}$ is dense. As we alluded to earlier in the chapter, the class probabilities condense to a unidimensional integral involving products of normal cdfs (c.f. equation 5.7) if $\boldsymbol{\Psi}$ is diagonal. Note that if $\boldsymbol{\Psi}$ is diagonal, then the transformed $\boldsymbol{\Omega}_{(j)} = \mathbf{Q}\boldsymbol{\Psi}^{-1}\mathbf{Q}^\top$ is certainly not: the components of $\mathbf{z}$ are correlated because they are all anchored on the same random variable. Thus, direct evaluation of (5.7) using quadrature methods avoids the Cholesky step and random sampling employed by the GHK method.

5.6.2 Efficient Kronecker product inverse

As with the normal I-prior model, the time complexity of the variational inference algorithm for I-probit models is dominated by the step involving the posterior evaluation of the I-prior random effects $\mathbf{w}$, which essentially is the inversion of an $nm \times nm$ matrix. The matrix in question is

$$\mathbf{V}_w = \left[(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right]^{-1}. \qquad \text{(from 5.17)}$$

We can actually exploit the Kronekcer product structure to compute the inverse efficiently. Perform an orthogonal eigendecomposition of $\mathbf{H}_\eta$ to obtain $\mathbf{H}_\eta = \mathbf{V}\mathbf{U}\mathbf{V}^\top$ and of $\boldsymbol{\Psi}$ to obtain $\boldsymbol{\Psi} = \mathbf{Q}\mathbf{P}\mathbf{Q}^\top$. This process takes $O(n^3 + m^3) \approx O(n^3)$ time if $m \ll n$ or if done in parallel, and needs to be performed once per CAVI iteration. Then, manipulate

$\mathbf{V}_w^{-1}$ as follows:

$$\begin{aligned}\mathbf{V}_w^{-1} &= (\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \\ &= (\mathbf{Q}\mathbf{P}\mathbf{Q}^\top \otimes \mathbf{V}\mathbf{U}^2\mathbf{V}^\top) + (\mathbf{Q}\mathbf{P}^{-1}\mathbf{Q}^\top \otimes \mathbf{V}\mathbf{V}^\top) \\ &= (\mathbf{Q} \otimes \mathbf{V})(\mathbf{P} \otimes \mathbf{U}^2)(\mathbf{Q}^\top \otimes \mathbf{V}^\top) + (\mathbf{Q} \otimes \mathbf{V})(\mathbf{P}^{-1} \otimes \mathbf{I}_n)(\mathbf{Q}^\top \otimes \mathbf{V}^\top) \\ &= (\mathbf{Q} \otimes \mathbf{V})(\mathbf{P} \otimes \mathbf{U}^2 + \mathbf{P}^{-1} \otimes \mathbf{I}_n)(\mathbf{Q}^\top \otimes \mathbf{V}^\top)\end{aligned}$$

Its inverse is

$$\begin{aligned}\mathbf{V}_w &= (\mathbf{Q}^\top \otimes \mathbf{V}^\top)^{-1}(\mathbf{P} \otimes \mathbf{U}^2 + \mathbf{P}^{-1} \otimes \mathbf{I}_n)^{-1}(\mathbf{Q} \otimes \mathbf{V})^{-1} \\ &= (\mathbf{Q} \otimes \mathbf{V})(\mathbf{P} \otimes \mathbf{U}^2 + \mathbf{P}^{-1} \otimes \mathbf{I}_n)^{-1}(\mathbf{Q}^\top \otimes \mathbf{V}^\top)\end{aligned}$$

which is easy to compute since the middle term is an inverse of diagonal matrices. This brings time complexity of the variational EM algorithm down to a similar requirement as if $\boldsymbol{\Psi}$ were diagonal. Unfortunately, storage requirements remain at $O(n^2m^2)$ when $\boldsymbol{\Psi}$ is dense, because the entire $nm \times nm$ matrix $\mathbf{V}_w$ is needed to evaluate the posterior mean of $\operatorname{vec} \mathbf{w}$.

5.6.3 Estimation of $\boldsymbol{\Psi}$ in future work

Suppose that $\boldsymbol{\Psi} \in \mathbb{R}^{m \times m}$ is a free parameter to be estimated, bearing in mind that only $m(m-1)/2-1$ variance components are identified in the I-probit model (see Section 5.2). If so, the Q function from (5.12) conditional on the rest of the parameters can be written as

$$Q(\boldsymbol{\Psi}|\boldsymbol{\alpha}, \eta) = \text{const.} - \frac{1}{2} \operatorname{tr}\Big(\boldsymbol{\Psi}\overbrace{\operatorname{E}\left((\mathbf{y}^* - \boldsymbol{\mu})^\top(\mathbf{y}^* - \boldsymbol{\mu})\right)}^{\mathbf{G}_1} + \boldsymbol{\Psi}^{-1}\overbrace{\operatorname{E}(\mathbf{w}^\top\mathbf{w})}^{\mathbf{G}_2}\Big)$$

with $\boldsymbol{\mu} = \mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}$. This can be solved using numerical methods, though it must be ensured that the identifiability constraints and positive-definiteness are satisfied. Specifically in the case where $\boldsymbol{\Psi}$ is a diagonal matrix $\operatorname{diag}(\psi_1, \ldots, \psi_m)$, then

$$\begin{aligned}Q(\boldsymbol{\Psi}|\boldsymbol{\alpha}, \eta) = \text{const.} &- \frac{1}{2}\sum_{j=1}^{m} \psi_j \operatorname{tr} \operatorname{E}\left((\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j})(\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j})^\top\right) \\ &- \frac{1}{2}\sum_{j=1}^{m} \psi_j^{-1} \operatorname{tr} \operatorname{E}(\mathbf{w}_{\cdot j}\mathbf{w}_{\cdot j}^\top)\end{aligned}$$

is maximised, for $j = 1, \ldots, m$, at

$$\hat{\psi}_j = \left(\frac{\operatorname{E}(\mathbf{w}_{\cdot j}^\top\mathbf{w}_{\cdot j})}{\operatorname{E}\left((\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j})^\top(\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j})\right)}\right)^{\frac{1}{2}},$$

independently of the rest of the other ψ_k's, $k \neq j$. As per the derivations in Appendix H.1.2 (p. 293), the numerator of this expression is equal to $\text{tr}(\tilde{\mathbf{V}}_{w_j} + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot j}^\top) = \text{tr}(\tilde{\mathbf{W}}_{jj})$. The denominator on the other hand is

$$\text{E}(\mathbf{y}_{\cdot j}^{*\top}\mathbf{y}_{\cdot j}^*) - n\alpha_j^2 - \text{tr}(\mathbf{H}_\eta^2\tilde{\mathbf{W}}_{jj}) - 2\mathbf{y}_{\cdot j}^{*\top}\mathbf{H}_\eta\tilde{\mathbf{w}}_{\cdot j} - 2\alpha_j\sum_{i=1}^n\sum_{i'=1}^n(y_{ij}^* - h_\eta(x_i, x_{i'})\tilde{w}_{ij}).$$

In either the full or I-probit model, solving $\boldsymbol{\Psi}$ involves the second moments of a truncated normal distribution. In the case where $\boldsymbol{\Psi}$ is dense, this is obtained by Monte Carlo methods, where samples from a truncated multivariate normal distribution are obtained using Gibbs sampling. Although this strategy can be used when $\boldsymbol{\Psi}$ is diagonal, we show that the form for the second moments involve integration of standard normal cdfs and pdfs (Lemma C.5, p. 269), much like the formula for the first moments.

5.7 Examples

We present analyses of real-data examples using the I-probit model for a variety of applicaitons, namely binary and multiclass classification, meta-analysis, and spatio-temporal modelling of point processes. Examples in this section have been analysed using in R using the in-development **iprobit** package written by us. Code for replication is provided at `http://myphdcode.haziqj.ml`. All of these examples had assumed a fixed error precision $\boldsymbol{\Psi} = \mathbf{I}_m$.

5.7.1 Predicting cardiac arrhythmia

Statistical learning tools are being used in the field of medicine as a means to aid medical diagnosis of diseases. In this example, factors determining the presence or absence of heart diseses are studied. Traditionally, cardiologists inspect patients' cardiac activity (ECG data) in order to reach a diagnosis, which remains the "gold standard" method of obtaining diagnoses. The study by Guvenir et al. (1997) aimed to predict cardiac abnormalities by way of machine learning, and minimise the difference between the gold standard and computer-based classifications.

The data set[3] at hand contains a myriad of ECG readings and other patient attributes such as age, height, and weight. Altogether, there are $n = 451$ observations and $p = 279$ predictors. In order for a valid comparison to be made to other studies, we excluded nominal covariates, leaving us with $p = 194$ continuous predictors, which we then standardised. In the original data set, there are 13 distinct classes of cardiac

[3]Data is made publicly available at `https://archive.ics.uci.edu/ml/datasets/arrhythmia`.

arrhythmia—again, following the lead of other studies, we had combined all forms of cardiac diseases to form a single class, thus reducing the problem to a binary classification task (normal vs. arrhythmia).

Following (5.6), the relationship between patient i's probability of having a form of cardiac arrhthmia p_i and the predictors $x_i \in \mathcal{X} \equiv \mathbb{R}^{194}$ is modelled as

$$\Phi(p_i) = \alpha + f(x_i).$$

Further, assuming $f \in \mathcal{F}$ a suitable RKHS with kernel h_λ, we may assign an I-prior on the (latent) regression function f. We consider three RKHSs: the canonical (linear) RKHS, the fBm-0.5 RKHS and the SE RKHS. The first of these three assumes an underlying linear relationship of the covariates and the probabilities, while the other two assumes a smooth relationship. As all covariates had been standardised, it is sufficient to assign a single scale parameter λ for the I-probit model.

For reference, fitting an I-probit model on the full data set takes about 4 seconds only, with convergence reached in at most 15 iterations. Figure 5.5 plots the variational lower bound value over time and iterations for the cardiac arrhythmia data set. As expected, the lower bound value increases over time until a convergence criterion is reached.

To measure predictive ability, we fit the I-probit models on a random subset of the data and obtain the out-of-sample test error rates from the remaining held-out observations. We then compare the results against popular machine learning classifiers, namely: 1) linear and quadratic discriminant analysis (LDA/QDA); 2) k-nearest neighbours; 3) support vector machines (SVM) (Steinwart and Christmann, 2008); 4) Gaussian process classification (Rasmussen and Williams, 2006); 5) random forests (Breiman, 2001); 6) nearest shrunken centroids (NSC) (Tibshirani et al., 2002); and 7) L-1 penalised logistic regression (Friedman et al., 2001). The experiment is set up as follows:

1. Form a training set by sub-sampling $s \in \{50, 100, 200\}$ observations.
2. The remaining unsampled data is used as the test set.
3. Fit model on training set, and obtain test error rates defined as

$$\text{test error rate} = \frac{1}{s}\sum_{i=1}^{n}[y_i^{\text{pred}} \neq y_i^{\text{test}}] \times 100\%.$$

4. Repeat steps 1-3 100 times to obtain the *average* test error rates and standard errors.

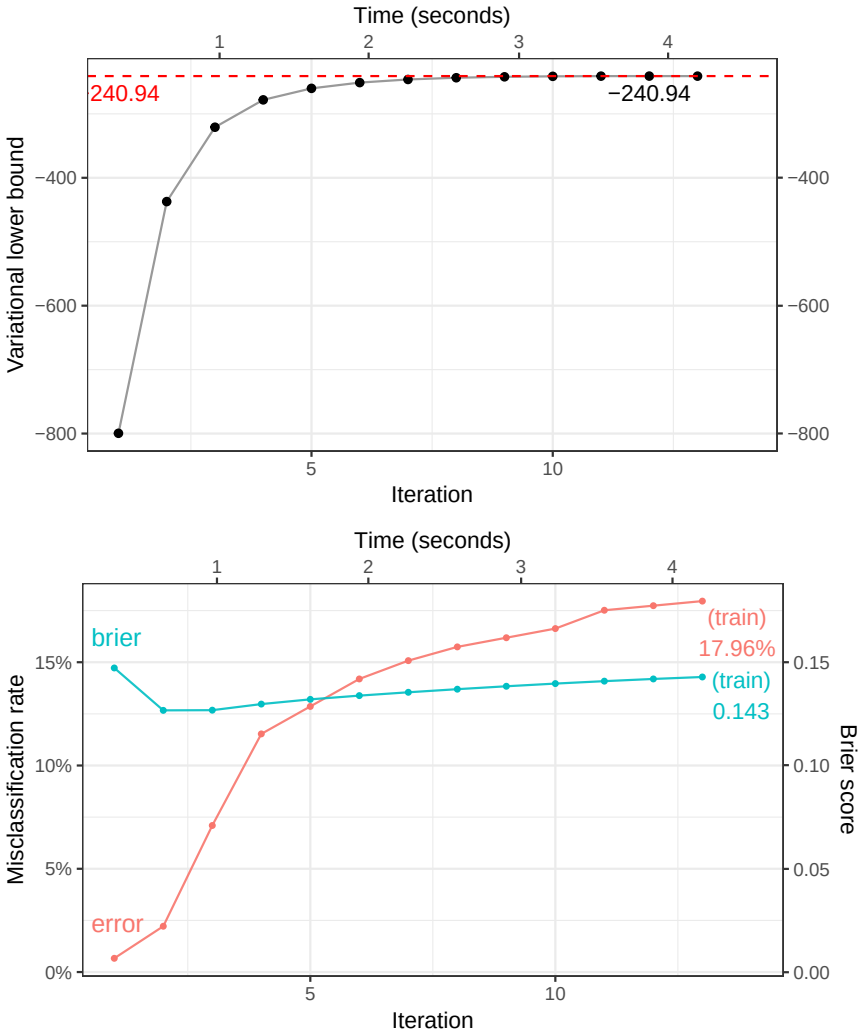


Figure 5.5: Plot of variational lower bound over time (left), and plot of training error rate and Brier scores over time (right).

Results for the methods listed above were extracted from the in-depth study by Cannings and Samworth (2017), who also conducted identical experiments using their random projection (RP) ensemble classification method. These are all tabulated in Table 5.3.

Of the three I-probit models, the fBm model performed the best. That it performed better than the canonical linear I-probit model is unsurprising, since an underlying smooth function to model the latent variables is expected to generalise better than a rigid straight line function. The poor performance of the SE I-probit model may be due to the fact that the lengthscale parameter was not estimated (fixed at $l = 1$), but then again, we notice reliable performance of the fBm even with fixed Hurst index ($\gamma = 0.5$). It can be seen that the fBm I-probit model also outperform the more popular machine learning algorithms out there including k-nearest neighbours, support vector machines and Gaussian process classification. It came second only to random forests, an

Table 5.3: Mean out-of-sample misclassification rates and standard errors in parantheses for 100 runs of various training (s) and test ($451 - s$) sizes for the cardiac arrhythmia binary classification task.

	Misclassification rate (%)		
Method	$s = 50$	$s = 100$	$s = 200$
I-probit			
Linear	35.52 (0.44)	31.35 (0.33)	29.45 (0.38)
Smooth (fBm-0.5)	33.64 (0.66)	28.12 (0.34)	24.33 (0.24)
Smooth (SE-1.0)	48.26 (0.40)	48.32 (0.43)	47.11 (0.37)
Others			
RP-LDA	33.24 (0.42)	30.19 (0.35)	27.49 (0.30)
RP-QDA	30.47 (0.33)	28.28 (0.26)	26.31 (0.28)
RP-k-NN	33.49 (0.40)	30.18 (0.33)	27.09 (0.31)
Random forests	31.65 (0.39)	26.72 (0.29)	22.40 (0.31)
SVM (linear)	36.16 (0.47)	35.61 (0.39)	35.20 (0.35)
SVM (Gaussian)	48.39 (0.49)	47.24 (0.46)	46.85 (0.43)
GP (Gaussian)	37.28 (0.42)	33.80 (0.40)	29.31 (0.35)
NSC	34.98 (0.46)	33.00 (0.40)	31.08 (0.41)
L-1 logistic	34.92 (0.42)	30.48 (0.34)	26.12 (0.27)

ensemble learning method, which is also generally faster to train than Gaussian process-like regressions such as I-prior models. The time complexity of a random forest algorithm is $O(pqn\log(n))$ (Louppe, 2014), where p is the number of variables used for training, q is the number of random decision trees, and n is the number of observations.

5.7.2 Meta-analysis of smoking cessation

Conider the smoking cessation data set, as described in Skrondal and Rabe-Hesketh (2004). It contains observations from 27 separate smoking cessation studies in which participants are subjected to either a nicotine gum treatment or a placebo. The interest is to estimate the treatment effect size, and whether it is statistically significant, i.e. whether or not nicotine gum is an effective treatment for quitting smoking. The studies are conducted at different times and due to various reasons such as funding and cultural effects, the results from all of the studies may not be in agreement. The number of effective participants plays a major role in determining the power of the statistical tests performed in individual studies. The question then becomes how do we meaningfully aggregate all the data to come up with one summary measure?

Several methods exist to analyse such data sets. One may consider a fixed-effects model, similar to a classical one-way ANOVA model to establish whether or not the

effect size is significant. Because of the study-specific characteristics, it is natural to consider multilevel or random-effects models as a means to estimate the effect size. Regardless of method, the approach of analysing study-level treatment effects instead of patient-level data only is the paradigm for meta-analysis, and our I-prior model takes this approach as well.

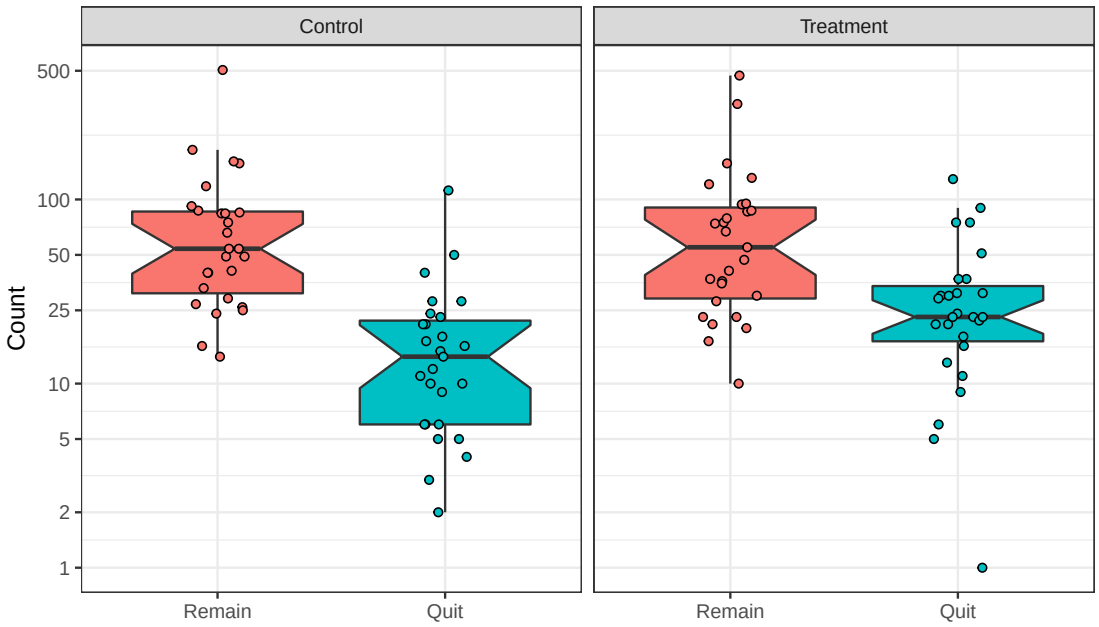


Figure 5.6: Comparative box-plots of the distribution of patients who successfully quit smoking and those who remained smokers, in the two treatment groups. It is evident that there are more successful patients quitting smoking in the treatment group than in the control group. The raw odds ratio of quitting smoking (treatment vs. control) is 1.66.

A summary of the data is displayed by the box-plot in Figure 5.6. On the whole, there are a total of 5,908 patients, and they are distributed roughly equally among the control and treatment groups (46.3% and 53.7% respectively, on average). From the box-plots, it is evident that there are more patients who quit smoking in the treatment group as compared to the placebo control group. There are various measures of treatment effect size, such as risk ratio or risk differences, but we shall concentrate on *odds ratios* as defined by

$$\text{odds ratio} = \frac{\text{odds of quitting smoking in } \textit{treatment} \text{ group}}{\text{odds of quitting smoking in } \textit{control} \text{ group}}.$$

The odds of quitting smoking in either group is defined as

$$\text{odds} = \frac{\text{P(quit smoking)}}{1 - \text{P(quit smoking)}},$$

and these probabilities, odds and ultimately odds ratio can be estimated from sample proportions. This raw odds ratio for all study groups is calculated as $1.66 = e^{0.50}$. It is also common for the odds ratio to be reported on the log scale (usually as a remnant of logistic models). A value greater than one for the odds ratio (or equivalently, greater than zero for the log odds ratio) indicates a significant treatment effect.

A random-effects analysis using a multilevel logistic model has been considered by Agresti and Hartzel (2000). Let $i = 1, \ldots, n_k$ index the patients in study group $k \in \{1, \ldots, 27\}$. For patient i in study j, p_{ik} denotes the probability that the patient has successfully quit smoking. Additionally, x_{ik} is the centred dummy variable indicating patient i's treatment group in study k. These take on two values: 0.5 for treated patients and -0.5 for control patients. The logistic random-effects model is

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}x_{ij}$$

with

$$\begin{pmatrix}\beta_{0j}\\ \beta_{1j}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{pmatrix}\right)$$

Agresti and Hartzel (2000) also made the additional assumption $\sigma_{01} = 0$, so that, coupled with the contrast coding used for x_{ik}, the total variance $\mathrm{Var}(\beta_{0k} + \beta_{1j}x_{ik})$ would be constant in both treatment groups. The overall log odds ratio is represented by β_1, and this is estimated as $0.57 \approx \log 1.76$.

In an I-prior model, the Bernoulli probabilities p_{ik} are regressed against the treatment group indicators x_{ik} and also the patients' study group k via the regression function f and a probit link:

$$\begin{aligned}\Phi^{-1}(p_{ik}) &= f(x_{ik}, k)\\ &= f_1(x_{ik}) + f_2(k) + f_{12}(x_{ik}, j).\end{aligned}$$

We have decomposed our function f into three parts: f_1 represents the treatment effect, f_2 represents the effect of the study groups, and f_{12} represents the interaction effect between the treatment and study group on the modelled probabilities. As both x_{ik} and k are nominal variables, the functions f_1 and f_2 both lie in the Pearson RKHS of functions $\mathcal{F}_1$ and $\mathcal{F}_2$, each with RKHS scale parameters λ_1 and λ_2. As such, it does not matter how the x_{ik} variables are coded (dummy coding 0, 1 vs. centred coding -0.5, 0.5) as the scaling of the function is determined by the RKHS scale parameters. The interaction effect f_{12} lies in the RKHS tensor product $\mathcal{F}_1 \otimes \mathcal{F}_2$. In the I-probit model, there are only two parameters to estimate, while in the standard logistic random-effects model, there are six. The results of the I-prior fit are summarised in the table below.

Table 5.4: Results of the I-probit model fit for three models.

Model	ELBO	Error rate (%)	Brier score	No. of parameters
f_1	-3210.76	23.65	0.179	1
$f_1 + f_2$	-3142.24	29.30	0.206	2
$f_1 + f_2 + f_{12}$	-3091.20	23.48	0.168	2

The approximated marginal log-likelihood value for the I-prior model (i.e. variational lower bound), the Brier score for each model and the number of RKHS scale parameters estimated in the model are reported in Table 5.4. Three models were fitted: 1) a model with only the treatment effect; 2) a model with a treatment effect and a study group effect; and 3) Model 2 with the additional assumption that treatment effect varies across study groups. Model 1 disregards the study group effects, while Model 2 assumes that the effectiveness of the nicotine gum treatment does not vary across study groups (akin to a varying-intercept model). A model comparison using the evidence lower bound indicates that Model 3 has the highest value, and the difference is significant from a Bayes factor standpoint—$\text{BF}(M_3, M_1)$ and $\text{BF}(M_3, M_2)$ are both greater than 150. The misclassification rate and Brier score indicates good predictive performance of the models, and there is not much to distinguish between the three although Model 3 is the best out of the three models.

Unlike in the logistic random-effects model, where the log odds ratio can be read off directly from the coefficients, with an I-prior probit model the log odds ratio needs to be calculated manually from the fitted probabilities. The probabilities of interest are the probabilities of quitting smoking under each treatment group for each study group k—call these $p_k(\text{treatment})$ and $p_k(\text{control})$. That is,

$$p_k(\text{treatment}) = \Phi\big(\hat{\nu}(\text{treatment}, k)\big)$$
$$p_k(\text{control}) = \Phi\big(\hat{\nu}(\text{control}, k)\big),$$

where $\hat{\nu}$ represents the standardised posterior mean estimate for the regression functions which are distributed according to

$$f(x_{ik}, k)|\mathbf{y}, \hat{\theta} \sim \text{N}\left(\hat{\mu}(x_{ik}, k), \hat{\sigma}^2(x_{ij}, k)\right),$$

with $x_{ik} \in \{\text{treatment}, \text{control}\}$ and $k \in \{1, \ldots, 27\}$ (see details in Section 5.5). The log odds ratio for each study group can then be calculated as usual. For the overall log odds ratio, the probabilities that are used are the averaged probabilities weighted according to the sample sizes in each group. This has been calculated as $0.51 \approx \log 1.66$, slightly lower than both the raw log odds ratio and the log odds ratio estimated by the

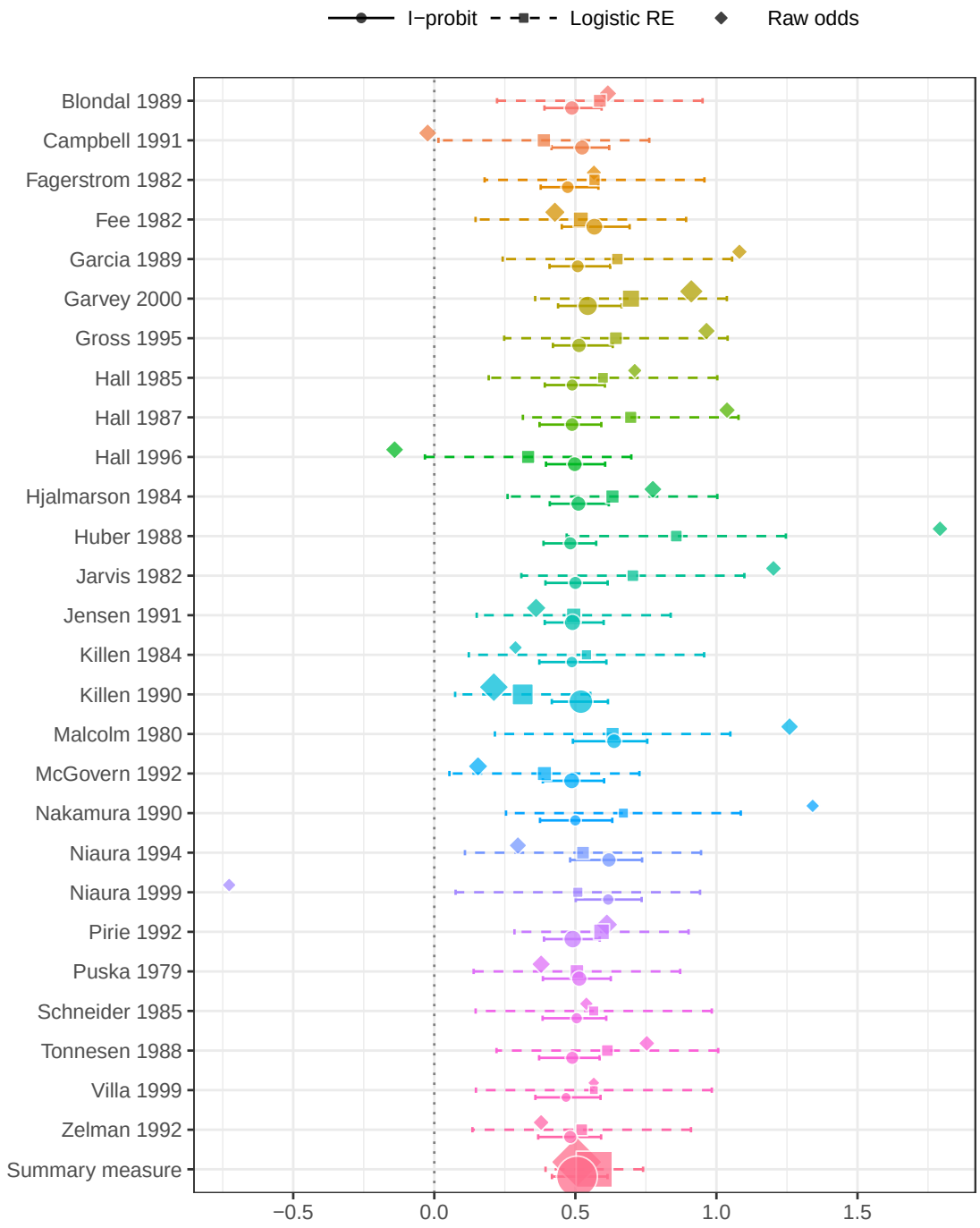


Figure 5.7: Forest plot of effect sizes (log odds ratios) in each group as well as the overall effect size together with their 95% confidence bands. The plot compares the raw log odds ratios, the logistic random-effect estimates, and the I-prior estimates. Sizes of the points indicate the relative sample sizes per study group.

logistic random-effects model. This can perhaps be attributed to some shrinkage of the estimated probabilities due to placing a prior with zero mean on the regression functions.

The credibility intervals for the log odds ratios in the forest plot of Figure 5.7 are also noticeably narrower under an I-prior compared to the fitted multilevel model. One explanation is that empirical Bayes estimates, such as the I-probit estimates under a variational EM framework, tend to underestimate the variability in the estimates because the variability in the parameters are ignored when point estimates are used, compared to distributions in a true Bayesian estimation framework.

5.7.3 Multiclass classification: Vowel recognition data set

We illustrate multiclass classification using I-priors on a speech recognition data set[4] with $m = 11$ classes to be predicted from digitized low pass filtered signals generated from voice recordings. Each class corresponds to a vowel sound made when pronouncing a specific word. The words that make up the vowel sounds are shown in Table 5.5. Each word was uttered once by multiple speakers, and the data are split into a training and a test set. Four males and four female speakers contributed to the training set, while four male and three female speakers contributed to the test set. The recordings were manipulated using speech processing techniques, such that each speaker yielded six frames of speech from the eleven vowels, each with a corresponding 10-dimensional numerical input vector (the predictors). This means that the size of the training set is $8 \times 6 \times 11 = 528$, while $7 \times 6 \times 11 = 462$ data points are available for testing the predictive performance of the models. This data set is also known as Deterding's vowel recognition data (after the original collector, Deterding, 1990). Machine learning methods such as neural networks and nearest neighbour methods were analysed by Robinson (1989).

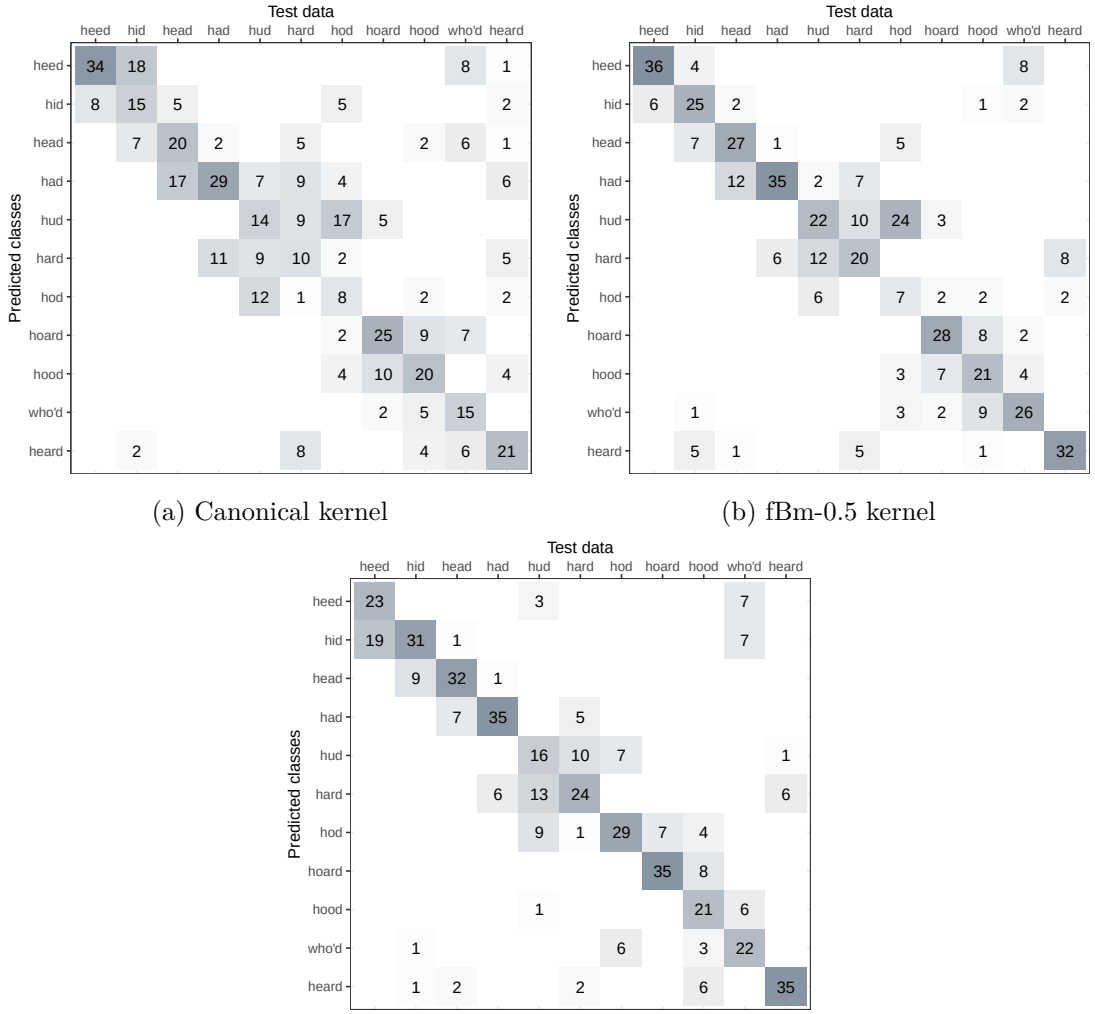
Table 5.5: The eleven words that make up the classes of vowels.

Class	Label	Vowel	Word	Class	Label	Vowel	Word
1	hid	iː	heed	7	hOd	ɒ	hod
2	hId	ɪ	hid	8	hod	ɔː	hoard
3	hEd	ɛ	head	9	hUd	ʊ	hood
4	hAd	a	had	10	hud	uː	who'd
5	hYd	ʌ	hud	11	hed	ɜː	heard
6	had	ɑː	hard				

We will fit the data using an I-probit model with the canonical linear kernel, fBm-0.5 kernel, and the SE kernel with lengthscale $l = 1$. Each model took roughly 13 seconds per iteration in fitting the training data set ($n = 528$). In particular, the canonical kernel

[4]Data is publicaly available from the UCI Machine Learning Repository, URL: https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Vowel+Recognition+-+Deterding+Data).

model took a long time to converge, with each variational inference iteration improving the lower bound only slighly each time. In contrast, both the fBm-0.5 and SE model were quicker to converge. Multiple restarts from different random seeds were conducted, and we found that they all converged to a similar lower bound value. This alleviates any concerns that the model might have converged to different multiple local optima.



(a) Canonical kernel

(b) fBm-0.5 kernel

(c) SE kernel

Figure 5.8: Confusion matrices for the vowel classification problem in which predicted values were obtained from the I-probit models. The maximum value for any cell is 42 (seven speakers delivered six frames of speech per vowel). Blank cells indicate nil values.

A good way to visualise the performance of model predictions is through a confusion matrix, as shown in Figure 5.8. The numbers in each row indicate the instances of a predicted class, while the numbers in the column indicate instances of the actual classes, while nil values are indicated by blank cells.

Table 5.6: Results of various classification methods for the vowel data set.

	Error rate (%)	
Model	Train	Test
I-probit		
Linear	29	54
Smooth (fBm-0.5)	22	40
Smooth (SE-1.0)	7	34
Others		
Linear regression	48	67
Logistic regression	22	51
Linear discriminant analysis	32	56
Quadratic discriminant analysis	1	53
Decision trees	5	54
Neural networks		45
k-nearest neighbours		44
FDA/BRUTO	6	44
FDA/MARS	13	39

Comparisons to other methods that had been used to analyse this data set is given in Table 5.6. In particular, the I-probit model is compared against 1) linear regression; 2) logistic linear regression; 3) linear and quadratic discriminant analysis; 4) decision trees; 5) neural networks; 6) k-nearest neighbours; and 7) flexible discriminant analysis. All of these methods are described in further detail in Friedman et al. (2001, Ch.4 & 12, Table 12.3). The I-probit model using both the fBm-0.5 and SE kernel offers one of the best out-of-sample classification error rates (34.4%) of all the methods compared. The linear I-probit model is seen to be comparable to logistic regression, linear and quadratic discrimant analysis, and also decision trees. It also provides significant improvement over multiple linear regression.

5.7.4 Spatio-temporal modelling of bovine tuberculosis in Cornwall

Data containing the number of breakdows of bovine tubercolosis (BTB) in Cornwall, the locations of the infected animals, and the year of occurence is analysed. The interest, as motivated by veterinary epidimiology, is to understand whether or not there is spatial segregation of the infection of the herds, and whether there is a time-element to the presence or absence of this spatial segregation. There has been previous work done to analyse this data set. Diggle et al. (2005) developed a non-parametric method to estimate spatial segregation using a multivariate point process. The occurrences are modelled as Poisson point processes, and spatial segregation is said to have occured if

the model-estimated type-specific breakdown probabilities at any given location are not significantly different from the sample proportions. The authors estimated the probabilities via kernel regression, and the test statistic of interest had to be estimated via Monte Carlo methods. Other works include Diggle et al. (2013), who used a fully Bayesian approach for spatio-temporal multivariate log-Gaussian Cox processes, which is implemented in the R package **lgcp** (Taylor et al., 2013).

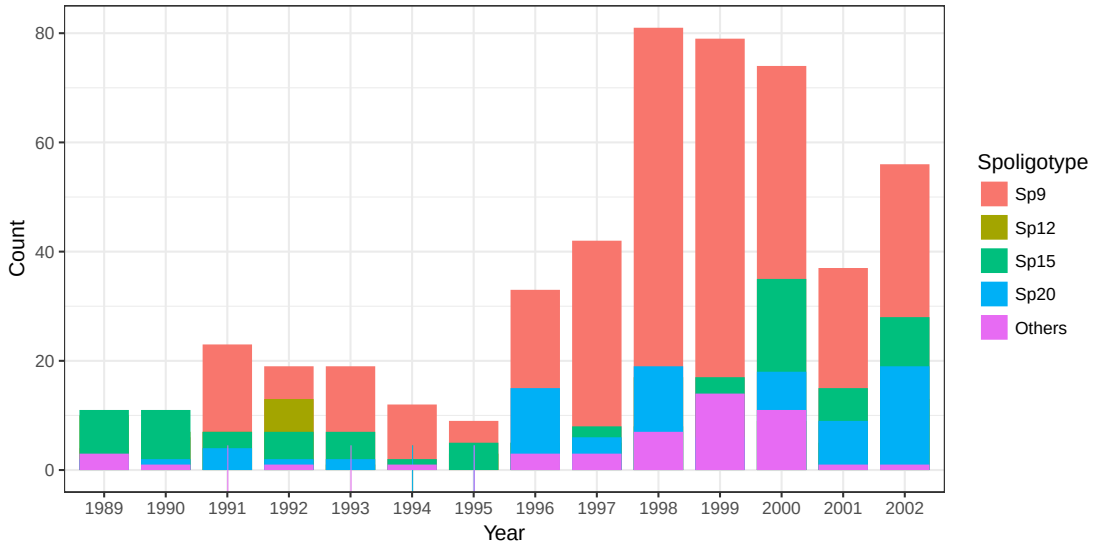


Figure 5.9: Distribution of the different types (Spoligotypes) of bovine tubercolosis affecting herds in Cornwall over the period 1989 to 2002.

The data set contains $n = 919$ recorded cases over a span of 14 years. For each of the cases, spatial data pertaining to the location of the farm (Northings and Eastings, measured in kilometres) are available. Originally, 11 unique spoligotypes were recorded in the data, with the four most common spoligotypes being Sp9 ($m = 1$), Sp12 ($m = 2$), Sp15 ($m = 3$) and Sp20 ($m = 4$), as shown by the histogram in Figure 5.9. We had grouped the remaining seven spoligotypes into an 'Others' category ($m = 5$), so that the problem becomes a multinomial regression with five distinct outcomes.

We are able to investigate any spatio-temporal patterns of infection using I-priors rather simply. Let p_{ij} denote the probability that a particular farm i is infected with a BTB disease with spoligotype $j \in \{1, \dots, 5\}$. We model the transformed probabilities $g_j(p_{ij})$ as following a function which takes two covariates, i.e. the spatial data $x_1 \in \mathbb{R}^2$, and the temporal data x_2 (year of infection):

$$\begin{aligned} p_{ij} &= g_j^{-1}\big(f_k(x_1, x_2)\big)_{k=1}^m \\ &= g_j^{-1}\big(f_{1k}(x_1) + f_{2k}(x_2) + f_{12k}(x_1, x_2)\big)_{k=1}^m, \end{aligned}$$

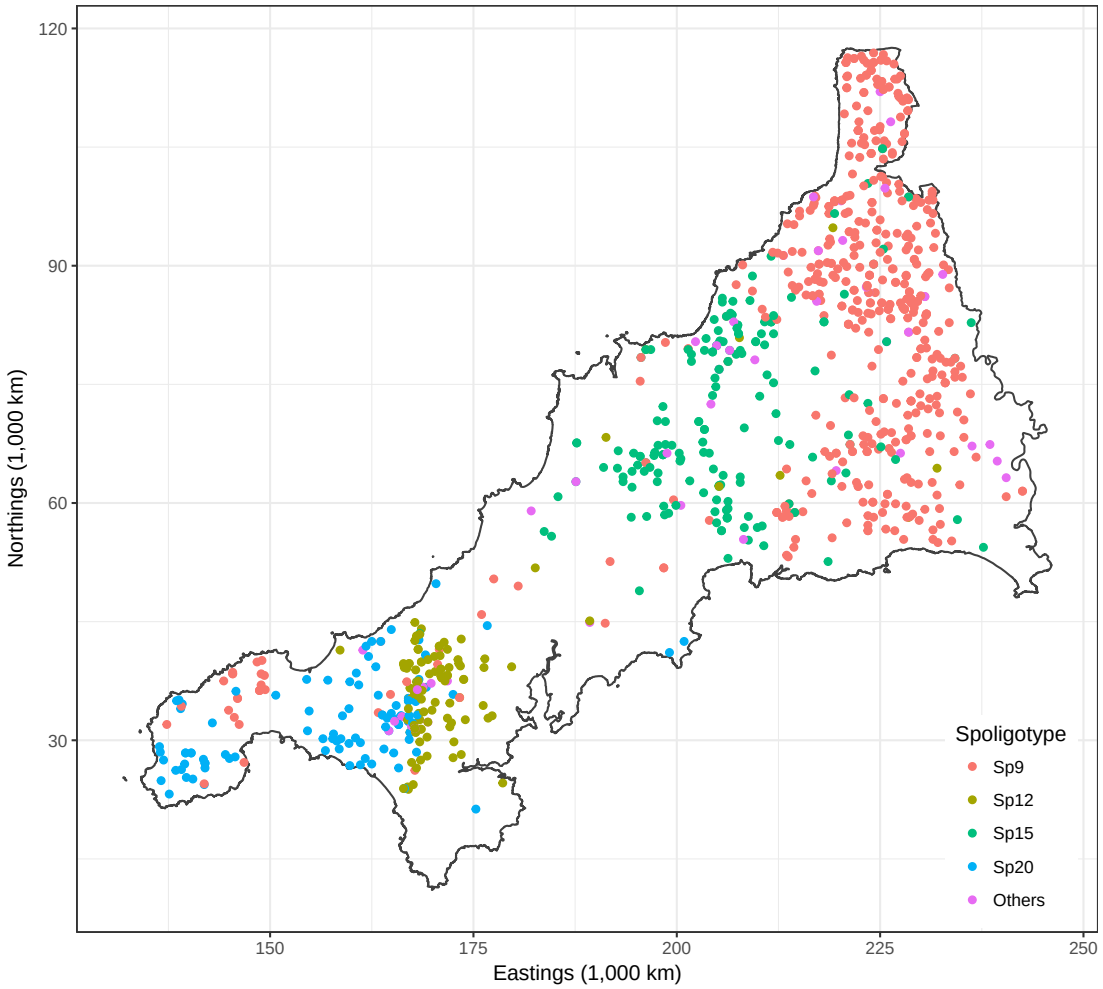


Figure 5.10: Spatial distribution of all cases over the 14 years.

where the function $g_j^{-1} : \mathbb{R}^m \to [0, 1]$ is the same squashing function used in equation (5.10). We assume a smooth effect of space and time on the probabilities, and appropriate RKHSs for the functions $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$ are the fBm-0.5 RKHS. Alternatively, as per Diggle et al. (2005), divide the data into four distinct time periods: 1) 1996 and earlier; 2) 1997 to 1998; 3) 1999 to 2000; and finally 4) 2001 to 2002. In this case, x_2 would indicate which period the infection took place in, and thus would have a nominal effect on the probabilities. An appropriate RKHS for f_2 in such a case would be the Pearson RKHS. In either case, the function $f_{12} \in \mathcal{F}_1 \otimes \mathcal{F}_2$ would be the "interaction effect", meaning that with such an effect present, the spatial distribution of the diseases are assumed to vary across the years.

We fitted four different models:

- $\boldsymbol{M_0}$**: Intercept only**.

$$p_{ij} = g_j^{-1}(\alpha_k)_{k=1}^m$$

Table 5.7: Results of the fitted I-probit models. Estimates of the class intercepts and scale parameters, together with their respective bootstrap standard errors, are presented. For model comparison, we can look at ELBOs, error misclassification rates, and Brier scores.

	M_0: Intercepts only		M_1: Spatial only		M_2: Spatio-temporal		M_3: Spatio-period	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept (Sp9)	0.948	0.000	1.364	0.015	1.401	0.079	1.395	0.103
Intercept (Sp12)	-0.173	0.000	-0.435	0.013	-0.506	0.017	-0.463	0.045
Intercept (Sp15)	0.103	0.000	-0.020	0.011	-0.008	0.059	-0.010	0.094
Intercept (Sp20)	-0.202	0.000	-0.775	0.051	-0.795	0.223	-0.783	0.343
Intercept (Others)	-0.676	0.000	-0.134	0.016	-0.091	0.077	-0.139	0.104
Scale (spatial)			0.194	0.008	-0.176	0.178	0.172	0.169
Scale (temporal)					-0.006	0.003	-0.004	0.006
ELBO	-1187.47		-564.33		-537.23		-543.94	
Error rate (%)	46.25		19.26		18.06		18.50	
Brier score	0.249		0.143		0.136		0.138	

- $\boldsymbol{M_1}$**: Spatial segregation**.

$$p_{ij} = g_j^{-1}\big(\alpha_k + f_{1k}(x_i)\big)_{k=1}^m$$

$f_{1k} \in \mathcal{F}_1$ Pearson RKHS.

- $\boldsymbol{M_2}$**: Spatio-temporal**.

$$p_{ij} = g_j^{-1}\big(\alpha_k + f_{1k}(x_i) + f_{2k}(t_i) + f_{12k}(x_i, t_i)\big)_{k=1}^m$$

$f_{1k} \in \mathcal{F}_1$ Pearson RKHS, $f_{2k} \in \mathcal{F}_2$ fBm-0.5 RKHS, and $f_{12k} \in \mathcal{F}_1 \otimes \mathcal{F}_2$

- $\boldsymbol{M_3}$**: Spatio-period**.

$$p_{ij} = g_j^{-1}\big(\alpha_k + f_{1k}(x_i) + f_{2k}(t_i) + f_{12k}(x_i, t_i)\big)_{k=1}^m$$

$f_{1k} \in \mathcal{F}_1$ Pearson RKHS, $f_{2k} \in \mathcal{F}_2$ Pearson RKHS, and $f_{12k} \in \mathcal{F}_1 \otimes \mathcal{F}_2$

Model M_0 corresponds to a model which ignores any spatial or temporal effects (the baseline intercept only model). Model M_1 takes into account only spatial effects. Both models M_2 and M_3 account for spatio-temporal effects, but M_2 assumes a smooth effect of time, while M_3 segregates the points into four distinct time periods for analysis. Model comparison is easily done, and Table 5.7 indicates that model M_2 has the highest log-likelihood of the four models, making it the preferable model.

For a more visual approach, we can look at the plots of the surface probabilities. To obtain these probabilities, we first determined the spatial points (Northings and Eastings) which fall inside the polygon which makes up Cornwall. We then obtained predicted probabilities for each class of disease at each location. Figure 5.11 was obtained using the model with spatial covariates only, thus ignoring any temporal effects. In the case of the spatio-temporal model, we used the model which had the period formulation for time (model M_3). This way, we can display the surface probabilities of the time periods in four plots only, which is more economical to exhibit within the margins of this thesis. Note that there is no issue with using the continuous time model—we have produced an animated gif image at `http://phd.haziqj.ml/examples/`, showing the evolution of the surface probabilities over time.

As the plots suggests, there is indeed spatial segregation for the four most common spoligotypes, and this is also very prominently seen from Figure 5.11. In comparing the distribution of the spoligotypes over the years, we may refer to Figure 5.12, a series of predicted probability surface plots over the four time periods obtained from model M_3. For each time period, we also superimposed the actual observations onto the predicted surface probabilities. In addition, coloured dotted lines are displayed to indicate the

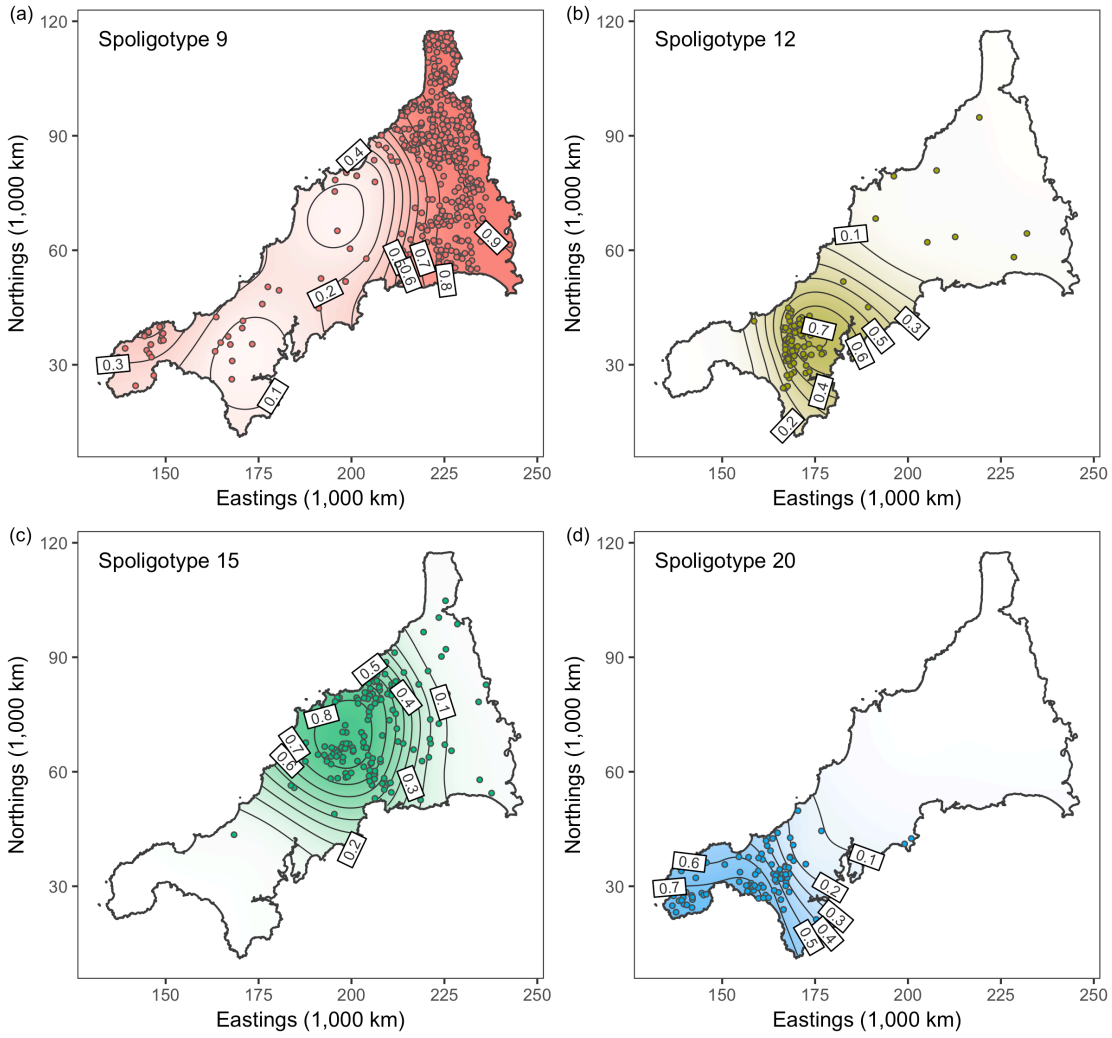


Figure 5.11: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium *Mycobacterium bovis* over the entire time period using model M_1.

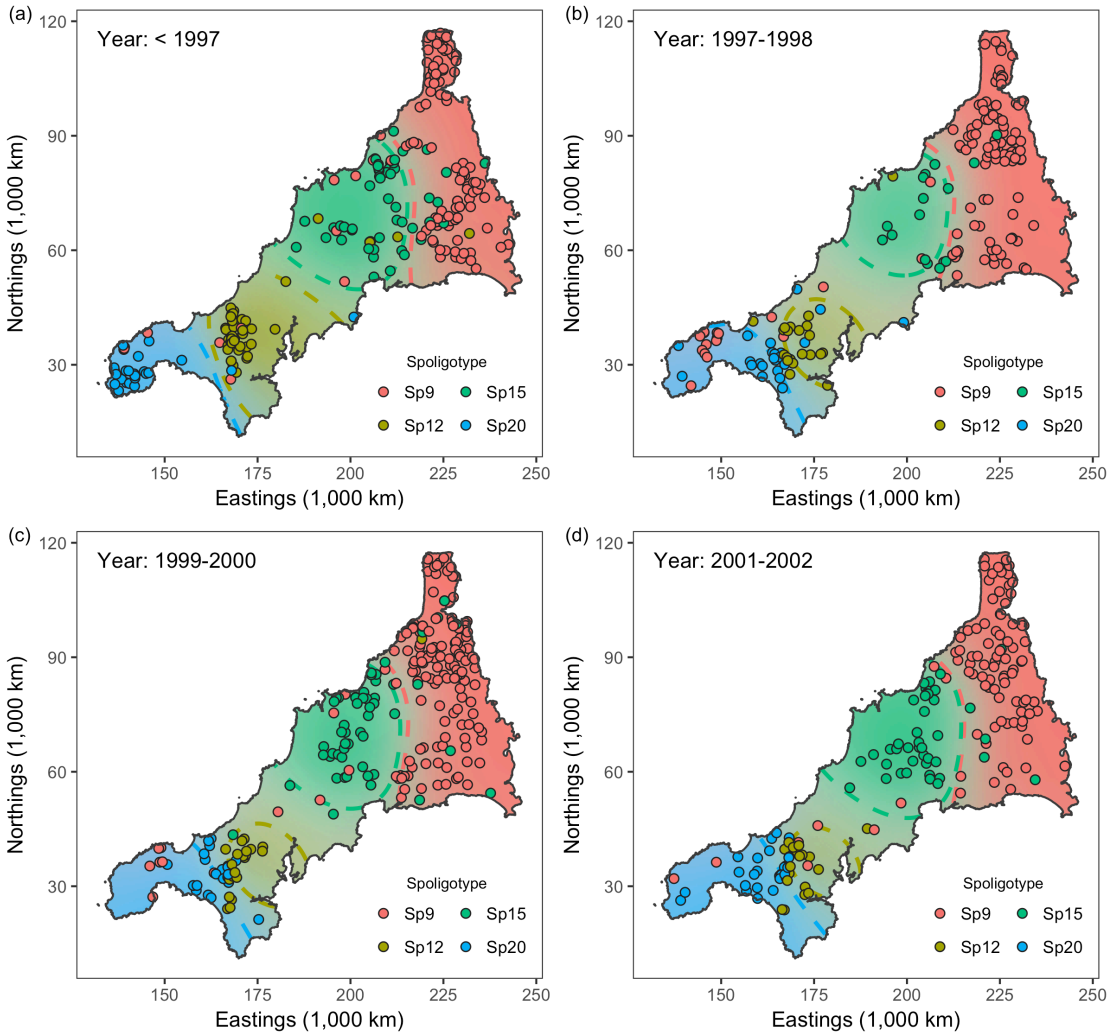


Figure 5.12: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium *Mycobacterium bovis* over four different time periods using model M_3.

"decision boundaries" for each of the four spoligotypes. The most evident change is seen to the spatial distribution of spoligotype 12, with the decision boundary giving it a large area in years 1996 and earlier, but this steadily shrunk over the years. Spoligotype 9, which is most commonly seen in the east of Cornwall, seems to have made its way down to the south-west over the years. The other two spoligotypes seem to be rather constant over the years.

5.8 Conclusion

This work presents an extension of the normal I-prior methodology to fit categorical response models using probit link functions—a methodology we call the I-probit. The main motivation behind this work is to overcome the drawbacks of modelling probabilities using the normal I-prior model. We assumed latent variables that represent 'class propensities' exist, modelled these using a normal I-prior, and simply transformed them via a probit link function. In this way, all of the advantages of the I-prior methodology seen for the normal model are preserved for binary and multinomial regression as well.

The core of this work explores ways in which to overcome the intractable integral presented by the I-probit model in (5.8). Techniques such as quadrature methods, Laplace approximation and MCMC tend to fail, or are unsatisfactorily slow to accomplish. The main reason for this is the dimension of this integral, which is nm, and thus for large sample sizes and/or number of classes, is unfeasible with such methods. We turned to variational inference in the face of an intractable posterior density that hampers an EM algorithm, and the result is a sequential updating scheme, similar in time and storage requirements to the EM algorithm.

In terms of similarity to other works, the generalised additive models (GAMs) of Hastie and Tibshirani (1986) comes close. The setup of GAMs is near identical to the I-probit model, although estimation is done differently. GAMs do not assume smooth functions from any RKHS, but instead estimates the f's using a local scoring method or a local likelihood method. Kernel methods for classification are extremely popular in computer science and machine learning; examples include support vector machines (Schölkopf and Smola, 2002) and Gaussian process classification (Rasmussen and Williams, 2006), with the latter being more closely related to the I-probit method. However, Gaussian process classification typically uses the logistic sigmoid function, and estimation most commonly performed using Laplace approximation, but other methods such as expectation propagation (Minka, 2001) and MCMC (Neal, 1999) have been explored as well. Variational inference for Gaussian process probit models have been studied by Girolami and Rogers (2006), with their work providing a close reference to the variational algorithm employed by us.

Suggestions for future work include:

1. **Estimation of $\boldsymbol{\Psi}$**. A limitation we had to face in this work was to treat $\boldsymbol{\Psi}$ as fixed. The discussion in Section 5.6.3 shows that estimation of $\boldsymbol{\Psi}$ is possible, however, the specific nature of implementing this in computer code could not be explored in time. In particular, for the full I-probit model, the best method of imposing positive-definite constraints for $\boldsymbol{\Psi}$ in the M-step has not been fully researched.

2. **Inclusion of class-specific covariates**. Throughout the chapter, we assumed that covariates were unit-specific, rather than class-specific. To illustrate, consider modelling the choice of travel mode between two destinations (car, coach, train or aeroplane) as a function of disposable income and travel time. Individuals' income as a predictor of transportation choice is unit-specific, but clearly, travel time depends on the mode of transport. To incorporate class-specific covariates z_{ij}, the regression on the latent propensities in (5.2) could be extended as such:

$$y_{ij}^* = \overbrace{\alpha_j + f_j(x_i) + e(z_{ij})}^{f(x_i, z_{ij}, j)} + \epsilon_{ij}$$

   An I-prior would then be applied as usual, with careful consideration of the RKKS used to model f.

3. **Improving computational efficiency**. The $O(n^3m)$ time requirement for estimating I-probit models hinder its use towards large-data applications. In a limited study, we did not obtain reliable improvements using low-rank approximations of the kernel matrix such as the Nyström method. The key to improving computational efficiency could lie in sparse variational methods, a suggestion that was made to improve normal I-prior models as well.

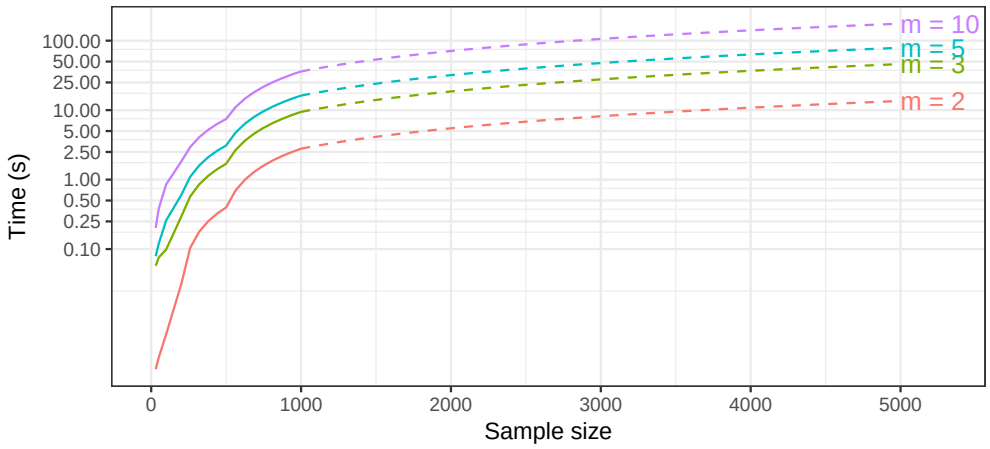


Figure 5.13: Time taken to complete a single variational inference iteration for varying sample sizes and number of classes m. The solid line represents actual timings, while the dotted lines are linear extrapolations.

As a final remark, we note that variational Bayes, which entails a fully Bayesian treatment of the model (setting priors on model parameters θ), is a viable alternative to variational EM. The output of such a variational inference algorithm would be approximate posterior densities for θ, in addition to $q(\mathbf{y}^*)$ and $q(\mathbf{w})$, instead of point estimates for θ. Posterior inferences surrounding the parameters would then be possible, such as obtaining posterior standard deviations, credibility intervals, and so on. However, a variational Bayes route has its cons:

1. **Tedious derivations**. As the parameters now have a distribution $\theta = \{\boldsymbol{\alpha}, \eta, \boldsymbol{\Psi}\} \sim q(\boldsymbol{\alpha}, \eta, \boldsymbol{\Psi})$, quantities such as

   - $\mathrm{E}(\log|\boldsymbol{\Psi}|)$;
   - $\mathrm{E}(\mathbf{H}_\eta^2)$; and
   - $\operatorname{tr}\mathrm{E}\left((\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})^\top\right)$,

   among others, will need to be derived for the variational inference algorithm, and these can be tricky to compute.

2. **Suited only to conjugate exponential family models**. When conjugate exponential family models are considered, the approximate variational densities (under a mean-field assumption) are easily recognised, as they themselves belong to the same exponential family as the model or prior. However, I-prior does not always admit conjugacy for the kernel parameters η (only for ANOVA RKKSs scale parameters), and most certainly not for $\boldsymbol{\Psi}$ (at least not in the current parameterisation). When this happens, techniques such as importance sampling or Metropolis algorithms need to be employed to obtain the posterior means required for the variational algorithm to proceed.

3. **Prior specification and sensitivity**. It is not clear how best to specify prior information (from a subjectivist's standpoint) for the RKHS scale parameters, intercepts, and perhaps the error precision, because these are parameters relating to the latent propensities which are not very meaningful or interpretable. Of course, one could easily specify vague or even diffuse priors. The concern is that the model could be sensitive to prior choices.

In consideration of the above, we opted to employ a variational EM algorithm for estimation of I-probit models, instead of a full variational Bayes estimation. In any case, computational complexity is expected to be the same between the two methods. An interesting point to note is that the RKHS scale parameters and intercept would admit a normal posterior under a variational Bayes scheme. This means that the posterior mode and the posterior mean coincide, so point estimates under a variational EM algorithm are exactly the same as the posterior mean estimates under a variational Bayes framework when a diffuse prior is used.

Chapter 6

Bayesian variable selection using I-priors

Earlier in Section 4.1 (p. 96), we saw that model (1.1) subject to normal assumptions (1.2), model assumptions A1–A3, and f belonging to the canonical RKHS of functions over $\mathcal{X} \equiv \mathbb{R}^p$ yields the standard multiple regression model

$$\begin{aligned} y_i &= \alpha + \sum_{k=1}^{p} x_{ik}\beta_k + \epsilon_i \\ \epsilon_i &\overset{\text{iid}}{\sim} \mathrm{N}_n(0, \sigma^2). \end{aligned} \tag{6.1}$$

In this chapter, we use the notation $\sigma^2 = \psi^{-1}$ to denote the error variance. Furthermore, an I-prior on the regression coefficient entails prescribing the following normal prior the β_k's:

$$\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)^\top \sim \mathrm{N}(\mathbf{0}, \kappa\sigma^2\mathbf{X}^\top\mathbf{X}).$$

This follows from (4.1) after a slight reparameterisation of the RKHS scale parameter $\kappa \mapsto \lambda^2/\sigma^4$. Throughout this chapter, we assume that the columns of the design matrix $\mathbf{X} = (X_1, \ldots, X_p)$ have been standardised, so that a single RKHS scale parameter is sufficient for the p covariates.

The topic of interest for this chapter is model selection for linear regression models. That is, from a set of p covariates or predictors $\{X_1, \ldots, X_p\}$, the task is to determine the best choice of subset(s) of variables that should be included in a regression model used to explain the variation in the response variable. As such, the term *variable selection* is synonymous to model selection for linear regression models. Fundamental to this notion of variable selection is an inherent belief in sparseness of the true data generative process surrounding the response variable, i.e. not all of the variables need be used to predict the response. Model selection is indeed a huge topic to cover fully. We broadly

classify variable selection into three categories: 1) (pairwise) model comparison using some criterion; 2) shrinkage to induce sparsity; and 3) Bayesian model selection. We understand that different categorisations and indeed categories of model selection exist in the literature, but our focus is on the discussion of the three types as mentioned.

Model selection criteria, both from a frequentist and Bayesian standpoint, can either be of a predictive nature (R^2, mean squared error of prediction (MSEP), C_p (Mallows, 1973), k-fold cross-validation MSEP, etc.), or a likelihood-based information criterion (likelihood ratios, Bayes factors, Akaike information criterion (AIC, Akaike, 1973), Bayesian information criterion (BIC, Schwarz, 1978), etc.). Selecting a model based on either of these criteria requires comparison of all 2^p criteria, which is not feasible for large p. Typically, these criteria are used in conjunction with step-wise procedures such as forward-selection or backward-deletion to restrict attention to a smaller number of potential subsets (George and McCulloch, 1993; Miller, 2002).

On the other hand, regularised least squares regression (ridge regression (Hoerl and Kennard, 1970), Lasso (Tibshirani, 1996), or a convex combination of the two via elastic nets (Zou and Hastie, 2005), etc.) provides additional information to the regression model in order to provide a sparse solution to linear system of equations in $\boldsymbol{\beta}$. These methods are proven to be popular as they are fast and perform exceptionally well in many situations, even in cases where $p > n$. Additionally, the Lasso produces solutions for $\boldsymbol{\beta}$ which are exactly zero. However, the Lasso in general produces estimates which are biased towards zero, are inconsistent, and have no valid standard errors (Friedman et al., 2001; Kyung et al., 2010). Further criticisms of the Lasso include its inability to select more than n predictors in a $p > n$ situation, and poor performance when multicollinearity exists among the covariates.

From a Bayesian perspective, regularisation is akin to placing priors on the β_k's to shrink the effects of the β_k's: the ridge regression has a Bayesian interpretation of placing normal priors on the regression coefficients, while the Lasso a Laplace or double exponential prior (Park and Casella, 2008). The term adaptive shrinkage has been used for the method in which hyper-priors are placed on the scale parameter of the prior for the β_k's. The idea is to adaptively shape the prior depending on the importance of the variable in the regression model. Bayesian shrinkage includes the task of specifying tuning parameters, which could potentially affect chain mixing in a Markov chain Monte Carlo method (MCMC) procedure (which is often used).

Bayesian model selection is probabilistic in nature: a priori, one assigns probabilities over the set of models, and then after observing the data, posterior model probabilities (PMPs) are used to discern which of the models was likeliest to have been behind the data generative process of the observed responses. Of course, with large p then calculation of all 2^p posterior model probabilities to ascertain which is highest will be a challenge,

if not impossible. But, as with most Bayesian applications, MCMC can be applied as a practical means of overcoming this intractability. This stochastic approach to variable selection was pioneered by George and McCulloch (1993), and studied by others such as Dellaportas et al. (2002), L. Kuo and Mallick (1998), and Ntzoufras (2011). Unlike shrinkage methods, Bayesian model selection is able to quantify the amount of times a variable 'enters the model' (inclusion probabilities), and thereby measuring its worth as a predictor.

Note that, in addition to model probabilities and inclusion probabilities, estimates of regression coefficients are obtained simultaneously in Bayesian variable selection. When several competing models have high posterior probabilities, regression coefficients from each model, or indeed any quantity of interest, may be combined and weighted using their posterior model probabilities, a technique known as *Bayesian model averaging* (Hoeting et al., 1999; Madigan and Raftery, 1994). Averaging over a set of models takes into account the uncertainty surrounding model selection, which other standard statistical procedures ignore upon selection of a single model from which to do inference. It is known to be the case that predictive accuracy of the model-averaged quantity is improved, as measured by a logarithmic scoring rule (Raftery et al., 1997).

Bayesian model selection is not without criticism, however. For complex models with many predictors or samples, MCMC is slow and may mix poorly (O'Hara and Sillanpää, 2009). Often, there are a lot of tuning parameters that need to be set correctly for the problem at hand.

The plan for this chapter is to describe a fully Bayesian model for variable selection using I-priors. The approach that we take is a stochastic search of the model space due to L. Kuo and Mallick (1998), realised through a simple Gibbs sampling procedure. The main motivation behind using I-priors in Bayesian variable selection is its suitability in accommodating to datasets with strong multicollinearity and being able to run with little to no prior information about the parameters. A simulation study is conducted and several real-world examples presented to demonstrate this fact.

6.1 Preliminary: model probabilities, model evidence and Bayes factors

The paradigm of model selection is as follows. From a finite set of models $\mathcal{M} = \{M_1, \ldots, M_K\}$, pairs of data $\{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$, $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathcal{X} \equiv \mathbb{R}^p$, had been generated according to the generative process dictated by one of the models $M_k \in \mathcal{M}$ and its respective parameters Θ_k. Having observed only this data set, the goal is to infer which of the models had generated the data, and consequently obtain estimates for the parameters. It is perhaps most natural to ponder which of the models is most likely to

be the "true" one given the data presented, and thus, this natural way of thinking leads one to the concept of *model probabilities*. From a Bayesian perspective in particular, *posterior model probabilities* allow us to quantify the certainty to which any model is behind the data generative process, after taking into account relevant evidence (observation of the data) and prior beliefs about model and parameter uncertainty.

Let $p(M_1), \ldots, p(M_K)$ be prior probabilities assigned to the model space $\mathcal{M}$, and $p(\Theta_k|M_k)$ be the prior on the parameters of model M_k. For any model $M_k \in \mathcal{M}$, the posterior model probability for model m is

$$p(M_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k)p(M_k)}{\sum_{k=1}^{K} p(\mathbf{y}|M_k)p(M_k)} \tag{6.2}$$

where

$$p(\mathbf{y}|M_k) = \int p(\mathbf{y}|M_k, \Theta_k)p(\Theta_k|M_k)\,\mathrm{d}\Theta_k \tag{6.3}$$

is known as the marginal likelihood, or *evidence*, for model M_k. As a remark, the prior distributions for the parameters do not necessarily need to depend on the model, so we might have that $p(\Theta_k|M_k) = p(\Theta_k)$. A natural strategy for model selection is to select the model such that $p(M_k|\mathbf{y})$ is largest (the *highest probability model*, HPM), but several models rather than just a single one may be reported to convey model uncertainty (Chipman et al., 2001).

Note, that models may be pairwise compared based on these posterior model probabilities, for which the posterior odds

$$\frac{p(M_k|\mathbf{y})}{p(M_0|\mathbf{y})} = \overbrace{\frac{p(\mathbf{y}|M_k)}{p(\mathbf{y}|M_0)}}^{\text{Bayes factor}} \times \overbrace{\frac{p(M_k)}{p(M_0)}}^{\text{prior odds}} \tag{6.4}$$

provide a point summary for comparing model M_k against model M_0. The first term on the right hand side is the Bayes factor for comparing any model $M_k \in \mathcal{M}$ to another model $M_0 \in \mathcal{M}$, and is denoted by $\mathrm{BF}(M_k, M_0)$. Thus, model selection based on posterior model probabilities can be formalised as the Bayesian alternative to classical hypothesis testing using Bayes factors (Kass and Raftery, 1995).

The issue that is faced with Bayesian model selection is that all posterior model probabilities must be calculated in order for a full comparison to be made. When the model space is very large, this can prove to be an insurmountable task. In the case of linear regression, where each of the p variables may be selected or not, the size of the model space is 2^p. Even for moderate sized p this can already be a challenge computationally. In the coming sections, we shall see that this problem is alleviated by the use of MCMC methods to evaluate posterior model probabilities.

6.2 The Bayesian variable selection model

We shall loosely refer to a model as a subset of variables selected from the full set of variables $\{X_1, \ldots, X_p\}$. It would be useful to be able to index each of these 2^p possible models somehow, and we achieve this by the use of indicator variables $\gamma \in \{0,1\}^p$. Let $\gamma_j = 1$ if the variable X_j is selected, and $\gamma_j = 0$ otherwise, for $j = 1, \ldots, p$. As an example, the full model, where all the variables are included in the model, is denoted by $\gamma = (1, \ldots, 1)$, while the intercept only model is denoted by $\gamma = (0, \ldots, 0)$. Note that we do not consider the intercept to be selectable.

Following L. Kuo and Mallick (1998), the linear model in (6.1) is expanded to include the indicator variables to form

$$\begin{aligned} y_i &= \alpha + \sum_{k=1}^{p} x_{ik}\gamma_k\beta_k + \epsilon_i \\ \epsilon_i &\overset{\text{iid}}{\sim} \mathrm{N}_n(0, \sigma^2). \end{aligned} \tag{6.5}$$

Hence, in addition to the usual model parameters $(\boldsymbol{\beta}, \sigma, \alpha)$, we are also interested in conducting model inferences through the posterior distribution of the γ's. The priors for the parameters are described below:

- **Model indicators** γ_j. An independent Bernoulli prior is specified for the model indicators
$$p(\gamma) = \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}. \tag{6.6}$$
We may choose to set all $\pi_j = 0.5$ a priori to reflect equally likely probabilities that any model may be chosen. Alternatively, we might have some subjective beliefs about which predictor is more likely or unlikely to be included in the model. We may also choose to include π_j in the estimation procedure by assigning a hyperprior on π_j such as the Beta(1, 1) (uniform distribution), Beta(1/2, 1/2) (Jeffreys prior), or any other suitable hyperprior. In any case, in this thesis we consider the simplest case of setting all $\pi_j = 0.5$.
- **Regression coefficients** $\boldsymbol{\beta}$. The L. Kuo and Mallick (1998) model is often known as the independent sampler due to the independence of model parameters and the indicator variables, i.e., $p(\boldsymbol{\beta}, \gamma) = p(\boldsymbol{\beta})p(\gamma)$. As such, prior choices for the regression coefficients can be any of the usual priors on $\boldsymbol{\beta}$, including
  - the independent prior $\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, c^2\mathbf{I}_p)$ for some choice of c (e.g. $c = 10$);
  - the g-prior $\boldsymbol{\beta}|\sigma, g \sim \mathrm{N}_p(\mathbf{0}, g\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$ for some g either chosen a priori or estimated (Bayes or empirical Bayes); or
  - the I-prior $\boldsymbol{\beta}|\sigma, \kappa \sim \mathrm{N}_p(\mathbf{0}, \kappa\sigma^2\mathbf{X}^\top\mathbf{X})$, which is the focus of this chapter.

- **Intercept** α. A normal prior $\alpha \sim \mathrm{N}(0, \sigma^2 A)$.
- **Scale** σ. An inverse gamma prior $\sigma \sim \Gamma^{-1}(c, d)$.

Priors for the intercept and scale parameters are chosen so as to maintain conjugacy to the normal regression model. Choices for the prior hyperparameters depend on the user's prior beliefs, but it is reasonable to set vague and uninformative hyperparameters to let the data speak as much as it can, especially in the absence of prior information. With this in mind, we may choose large values of A (e.g. 100) and small values of the shape and scale parameters for the inverse gamma (e.g. 0.001). Note that as $c, d \to 0$ in the inverse gamma distribution we get the Jeffreys prior[1] for scale parameters.

Remark 6.1. The BVS model (6.5) together with the choice of Bernoulli priors on γ and a normal prior $\mathrm{N}_p(\mathbf{0}, \mathbf{V}_\beta)$ for $\boldsymbol{\beta}$ can be seen a *spike-and-slab* prior for linear regression models, a mixture of a point mass at zero and a normal density (Geweke, 1996; Mitchell and Beauchamp, 1988). Write $\boldsymbol{\theta} = (\gamma_1\beta_1, \ldots, \gamma_p\beta_p)^\top$, which are interpreted as the 'model-specific regression coefficients'. Then, the prior on $\boldsymbol{\theta}$ is equivalently written

$$\boldsymbol{\theta}|\gamma \sim \begin{cases} \mathrm{N}_p(\mathbf{0}, \mathbf{V}_\beta) & \text{w.p. } p(\gamma) \\ 0 & \text{w.p. } 1 - p(\gamma). \end{cases}$$

A subtle fact of these spike-and-slab priors is that the posterior distribution for $\boldsymbol{\theta}$ will also be a combination of a point mass and a normal density (with appropriate posterior parameters). Looking at it from this perspective, regression coefficients are assigned zero values with positive probability, and it is this fact that allows covariates to be dropped from the model. As pointed out by L. Kuo and Mallick (1998), the form of the variable selection model allows the selection of important variables, while simultaneously shrinking the coefficients via prior information.

6.3 Gibbs sampling for the I-prior BVS model

The Bayesian variable selection model can be estimated using Gibbs sampling, as demonstrated originally by L. Kuo and Mallick (1998). In this section, we describe the Gibbs sampling procedure to obtain posterior samples of the parameters. For the I-prior specifically, the joint density of the responses and the priors is

$$p(\mathbf{y}, \gamma, \boldsymbol{\beta}, \alpha, \sigma^2, \kappa) = p(\mathbf{y}|\gamma, \boldsymbol{\beta}, \alpha, \sigma^2)p(\boldsymbol{\beta}|\sigma^2, \kappa)p(\alpha|\sigma^2)p(\gamma)p(\sigma^2)p(\kappa),$$

where the distribution of the model $p(\mathbf{y}|\gamma, \boldsymbol{\beta}, \alpha, \sigma^2)$ and of the priors have been described in the previous section (except for κ, which we now assign an inverse gamma distribu-

[1] The Jeffreys prior for a parameter θ is defined as $p(\theta) \propto |\mathcal{I}(\theta)|^{1/2}$ (Jeffreys, 1946).

tion). Let us denote $\Theta = \{\alpha, \boldsymbol{\beta}, \gamma, \sigma^2, \kappa\}$ to be the full set of parameters that we wish to obtain posterior samples for. Starting with suitable initial values $\Theta^{(0)}$, we then proceed to obtain samples $\Theta^{(1)}, \ldots, \Theta^{(T)}$ by sampling each parameter from the conditional posterior density of that parameter given the rest of the parameters. A suggested set of initial values are the maximum likelihood (ML) estimates of Θ or the posterior mean estimate of Θ under the full model $\gamma = (1, \ldots, 1)$ after an initial MCMC run.

The Gibbs conditional densities are straightforward to obtain on account of model conjugacy (details of derivation are in Appendix I, p. 295). We start with $\boldsymbol{\beta}$: the conditional density of $\boldsymbol{\beta}$ given $\alpha, \gamma, \sigma^2, \kappa$ is multivariate normal with mean $\tilde{\mathbf{B}}(\mathbf{y} - \alpha\mathbf{1}_n)$ and covariance matrix $\sigma^2\tilde{\mathbf{B}}$, where $\tilde{\mathbf{B}} = \mathbf{X}_\gamma^\top\mathbf{X}_\gamma + (\kappa\mathbf{X}^\top\mathbf{X})^{-1}$, and $\mathbf{X}_\gamma = (\gamma_1 X_1 \cdots \gamma_p X_p)$. Interestingly, when X_j is dropped from the model ($\gamma_j = 0$), the posterior mean and variance for j'th component of $\boldsymbol{\beta}$ is entirely informed by the prior (L. Kuo and Mallick, 1998). The data-driven I-prior incorporates information from the covariates into the prior, which then informs the posterior. In a similar manner, the conditional density for the intercept α is found to be $\mathrm{N}\left(\sum_{i=1}^n (y_i - \mathbf{x}_i^\top\boldsymbol{\theta})/\tilde{A}, \sigma^2\tilde{A}\right)$, where $\tilde{A} = n + A^{-1}$ and A is the prior variance for α.

The (conditional) posterior samples of $\gamma = (\gamma_1, \ldots, \gamma_p)$ are obtained component-wise, and each conditional probability mass function for γ_j is Bernoulli with success probability $\tilde{\pi}_j = u_j/(u_j + v_j)$, where

$$u_j = \pi_j \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}_j^{[1]}\|^2\right)$$

and

$$v_j = (1 - \pi_j) \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}_j^{[0]}\|^2\right).$$

Here, we have used the notation $\boldsymbol{\theta}_j^{[1]}$ to indicate the vector $\boldsymbol{\theta}$ with the j'th component replaced by $\boldsymbol{\beta}_j$, and $\boldsymbol{\theta}_j^{[0]}$ to indicate the vector $\boldsymbol{\theta}$ with the jth component replaced by 0. Values of 1 for γ are more likely to be sampled when the ratio u_j/v_j is greater than the prior odds $\pi_j/(1 - \pi_j)$. Specifically when the prior probabilities π_j are all set to be 0.5, then γ_j will be more likely to be sampled as '1' if $u_j > v_j$, i.e. if the residual sum of squares (RSS) $\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2$ is *smaller* when the jth component is non-zero, compared to the RSS when the j'th component of $\boldsymbol{\theta}$ is zero.

We can in fact draw parallels to a Bayesian hypothesis test, with the null hypothesis being $\mathrm{H}_0 : \beta_j = 0$ and the alternative being $\mathrm{H}_1 : \beta_j \neq 0$, conditional on knowing all other values of the parameters. Under H_k, $\mathbf{y}|\Theta \sim \mathrm{N}_n(\alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\theta}_j^{[k]}, \sigma^2\mathbf{I}_n)$, $k = 0, 1$. The conditional Bayes factor comparing the model in the alternative hypothesis M_1 to the

model in the null hypothesis M_0 is therefore

$$\text{BF}(M_1, M_0) = \frac{u_j/\pi_j}{v_j/(1-\pi_j)} = \frac{\tilde{\pi}_j}{1-\tilde{\pi}_j} \Big/ \frac{\pi_j}{1-\pi_j}.$$

Thus, it can be seen that the decision to include or exclude the j'th variable from the model relates a hypothesis test using the Bayes factor rule, and this decision is embedded in the conditional posterior probabilities $\tilde{\pi}_j$. The Gibbs sampling procedure does something that can be described as "an automated stochastic F-test for subset selection" (L. Kuo and Mallick, 1998).

Both scale parameters σ^2 and κ follow the conditional inverse gamma distributions

$$\sigma^2|\alpha, \beta, \gamma, \kappa \sim \Gamma^{-1}\big(n/2 + c_\sigma + 1, \|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2/2 + d_\sigma\big)$$

and

$$\kappa|\alpha, \beta, \gamma, \sigma^2 \sim \Gamma^{-1}\big(p/2 + c_\kappa + 1, \boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta}/\sigma^2 + d_\kappa\big).$$

Note that the inverse gamma distribution that we specify here is defined by its shape and scale parameter, and has the density function described in Appendix C.6. Here, $\{c_\sigma, d_\sigma\}$ and $\{c_\kappa, d_\kappa\}$ are the shape and scale hyperparameters of the inverse gamma priors on σ^2 and κ respectively.

6.4 Posterior inferences

Having obtained posterior samples $\Theta^{(t)} = \{\alpha^{(t)}, \boldsymbol{\beta}^{(t)}, \gamma^{(t)}, \sigma^{2(t)}, \kappa^{(t)}\}$, there are two quantities of interest in relation to model inferences. The first is an estimate of posterior model probabilities, given by

$$\widehat{\text{P}}(\gamma = \gamma'|\mathbf{y}) = \frac{1}{T}\sum_{i=1}^{T}[\gamma^{(t)} = \gamma'], \tag{6.7}$$

where $[\cdot]$ is the Iverson bracket. This gives an estimate of the probability of a model coded by γ' appearing in the posterior state space of models. The second is a quantification of the posterior inclusion for each of the p variables $X_1, \ldots, X_p$, known as *posterior inclusion probabilities* (PIPs) for a variable being selected in any model. This is given by

$$\widehat{\text{P}}(\gamma_j = 1|\mathbf{y}) = \frac{1}{T}\sum_{i=1}^{T}[\gamma_j^{(t)} = 1], \qquad j = 1, \ldots, p. \tag{6.8}$$

Posterior inclusion probabilities are the marginals of the posterior model probabilities across each variable.

Table 6.1: Illustration of samples of γ from the Gibbs sampler for $p = 3$. As an example, to estimate the posterior model probability of $\{X_1,X_3\}$, we count the occurrences of the combination $\gamma^{(t)} = (1, 0, 1)$ in the sample and divide by T. To estimate posterior inclusion probabilities for any of the three variables, we take the sample mean of the binary variates column-wise.

t	$\gamma_1^{(t)}$	$\gamma_2^{(t)}$	$\gamma_3^{(t)}$
1	1	0	1
2	1	0	0
3	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	1	0	1

Note, that the regression coefficient of interest is not $\boldsymbol{\beta}$, but rather the "model averaged" regression coefficients $\boldsymbol{\theta} = (\gamma_1\beta_1, \ldots, \gamma_p\beta_p)^\top$ (Madigan and Raftery, 1994). Posterior variances for $\boldsymbol{\theta}$ will typically be larger than variances for $\boldsymbol{\beta}$, because posterior estimates surrounding $\boldsymbol{\theta}$ will have incorporated model uncertainty, but $\boldsymbol{\beta}$ on the other hand, will not. Thus, any inferential procedure surrounding the regression coefficients avoids the risk of over-confidence. Note that, since $\boldsymbol{\theta}$ will contain values of exactly zero when predictors are dropped out of the model, the posterior density for $\boldsymbol{\theta}$ is a mixture of a point mass at zero and a normal density. In any case, the likelihood only provides sufficient information to identify the product of $\boldsymbol{\beta}$ and γ, but not each of them separately (L. Kuo and Mallick, 1998).

Finally, any quantity of interest Δ can be incorporated as part of the Gibbs sampling procedure. That is, at each Gibbs iteration $t = 1, \ldots, T$, calculate $\Delta^{(t)}$ as a function of the parameter values at iteration t. This can be done during the Gibbs sampling process, or even after the fact as part of a post-processing procedure. Any inference on the posterior of Δ will then have incorporated the model uncertainty from a model averaging standpoint, as discussed earlier. As an example, suppose we are interested in the predicted value at a new covariate value $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$. For each Gibbs sample, calculate

$$y_{\text{new}}^{(t)} = \alpha^{(t)} + \mathbf{x}_{\text{new}}^\top(\gamma_1\beta_1, \ldots, \gamma_p\beta_p),$$

and obtain a point estimate $\hat{y}_{\text{new}}^{(t)}$ using the posterior mean or mode. The uncertainty for this estimate is contained in the standard deviation calculated from the sample $y_{\text{new}}^{(1)}, \ldots, y_{\text{new}}^{(T)}$, from which a 95% credibility interval for this estimate can be obtained from the empirical upper and lower 0.025 cut off points.

6.5 Two stage procedure

The variable selection procedure can be improved by a "pre-selection" of variables to trim off unimportant variables which reduces the size of the model space being explored. Without appealing to other external pre-selection methods, there is actually information that we could use from Bayesian variable selection models in the form of posterior inclusion probabilities. The procedure would work as follows:

1. Run the Bayesian variable selection model and obtain posterior inclusion probabilities for each variable.
2. Discard variables with inclusion probablities less than a certain treshold, τ.
3. Re-run the Bayesian variable selection model on the set of reduced variables.

A natural choice for τ would be 0.5, and therefore a two-stage approach to Bayesian variable selection can then be motivated as selecting the subset of variables which constitutes what is known as the *median probability model*. The median probability model is obtained by selecting all variables with a posterior inclusion probability of greater than or equal to a half. Barbieri and J. O. Berger (2004) show that the median probability model has the property of being optimally predictive (minimises squared error loss for predictions) under certain strict conditions.

The notion of a two-stage approaches are not new, as many variable selection methods in the literature generally employ a pre-selection method of some kind before running their selection process proper. This can be based on subjective preconceptions about which variables to retain, substantive theory, or even an objective pre-selection criterion. Two-stage procedures for Bayesian variable selection models have been used in works by Fouskakis and Draper (2008) and Ntzoufras (2011).

In the simulation studies conducted and observations from real-data examples, this two-stage approach does seem to provide a benefit. The complexity of estimating all model probabilities grows exponentially with p, therefore reducing this benefits the model selection procedure because the search of the model space is less cluttered. Of course, this idea works if the 'correct' variables are deleted when proceeding to the second stage. We posit that the p posterior inclusion probabilities are easier to estimate than the 2^p posterior model probabilities from the same amount of information coming from the MCMC samples. As a result, information summarised through the posterior inclusion probabilities are more precise than the posterior model probabilities.

6.6 Simulation study

In this section, we conduct a simulation study to compare the performance of different priors in the Bayesian variable selection framework described above. The priors on $\boldsymbol{\beta}$ that are compared are those mentioned in Section 6.2, i.e. the I-prior, the independent prior with large prior variance (flat/uninformative prior), and the g-prior with $g = n$ (unit information prior, Ntzoufras, 2011). We also make a comparison the variable selection performance of the Lasso, which, from a Bayesian perspective, is similar to setting a double-exponential or Laplace priors on the regression coefficients (Park and Casella, 2008). For clarity, the Lasso model employed in the simulations is of a frequentist regularisation framework as per Tibshirani (1996), and is neither a Bayesian variable selection model as described earlier, nor a fully Bayes implementation as per Park and Casella (2008). We felt it interesting to compare the Lasso as it is widely used for variable selection of linear models.

The experiment is to select from $p = 100$ variables of a $n = 150$ sample size artificial dataset which has pairwise correlations between the variables. This was inspired by the studies done by George and McCulloch (1993) and L. Kuo and Mallick (1998) in their respective papers, albeit on a larger scale (in theirs, $p = 30$). Five different scenarios were looked at. For each scenario, only s out of 100 variables were selected to form the "true" model and generate the responses according to the linear model $\mathbf{y} \sim \mathrm{N}_{100}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_{150})$. The signal-to-noise ratio (SNR) as a percentage is defined as s%, and the five scenarios are made up of varying SNR from high to low: 90%, 75%, 50%, 25%, and 10%. Variables that were included in the model had true $\boldsymbol{\beta}$ coefficients equal to one. That is, $\boldsymbol{\beta}_{\text{true}} = (\mathbf{1}_s, \mathbf{0}_{100-s})^\top$, where $\mathbf{1}_s$ is a row-vector of s ones, and $\mathbf{0}_{100-s}$ is a row-vector of $100 - s$ zeroes. The data generation process is summarised as follows:

- Draw $\mathbf{Z}_1, \ldots, \mathbf{Z}_{100} \overset{\text{iid}}{\sim} \mathrm{N}_{150}(\mathbf{0}, \mathbf{I}_{150})$.
- Draw $\mathbf{U} \sim \mathrm{N}_{150}(\mathbf{0}, \mathbf{I}_{150})$.
- Set $\mathbf{X} = (\mathbf{Z}_1 + \mathbf{U}, \ldots, \mathbf{Z}_{100} + \mathbf{U})$. This induces pairwise correlations of about $1/2$ between the columns of $\mathbf{X}$.[2]
- Draw $\mathbf{y} \sim \mathrm{N}_{150}(\mathbf{X}\boldsymbol{\beta}_{\text{true}}, \sigma^2\mathbf{I}_{150})$, with $\sigma = 2$.

In each scenario, we are interested in obtaining the highest probability model and counting the number of false choices made in this model after a two-stage procedure of variable selection. False choices can either be selecting variables wrongly (false inclusion) or failing to select a variable (false exclusion). Each scenario was repeated a total of 100 times to account for variability in the data generation process, and the results averaged.

A summary of the results is presented in Table 6.2. The overall results are also plotted in the form a frequency polygon (see Figure 6.1).

Table 6.2: Simulation results (proportion of false choices) for the Bayesian variable selection experiment using the I-prior, an independent prior, the g-prior and the Lasso across varying SNR.

	Signal-to-noise ratio				
False choices	90%	75%	50%	25%	10%
I-prior					
0-2	**0.93** (0.03)	**0.92** (0.03)	**0.90** (0.03)	**0.79** (0.04)	**0.55** (0.05)
3-5	0.07 (0.03)	0.07 (0.03)	0.10 (0.03)	0.20 (0.04)	0.27 (0.04)
>5	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	0.01 (0.01)	0.18 (0.04)
Ind. prior					
0-2	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**0.44** (0.05)	**1.00** (0.00)
3-5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.30 (0.05)	0.00 (0.00)
>5	**1.00** (0.00)	**1.00** (0.00)	**1.00** (0.00)	0.26 (0.04)	0.00 (0.00)
g-prior					
0-2	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	**0.78** (0.04)	**0.86** (0.03)
3-5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.14 (0.03)	0.13 (0.03)
>5	**1.00** (0.00)	**1.00** (0.00)	**1.00** (0.00)	0.08 (0.03)	0.01 (0.01)
Lasso					
0-2	0.03 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
3-5	0.19 (0.04)	0.02 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
>5	**0.78** (0.04)	**0.98** (0.01)	**1.00** (0.00)	**1.00** (0.00)	**1.00** (0.00)

The simulation results seem to indicate that the I-prior performs consistently well across all five scenarios, making no more than five false choices out of 100 (i.e. a 95% correct selection rate) in at least 82% of the time in the worst scenario. We do not observe much difference between the g-prior and the independent prior, and while they behave poorly in high SNR scenarios, these two priors seem to perform extremely well in low SNR scenarios. A high propensity to drop variables in these scenarios is a likely explanation, which does not necessarily indicate good performance—they perform well by contentiously omitting of a large number of unnecessary variables, especially in a two-stage procedure. Finally, the Lasso is well known to yield poor selection performance under multicollinearity, so the results are expected. The Lasso procedure was not subject to a two-stage approach because the Lasso does not provide information regarding posterior inclusion probabilities for individual variables.

[2]For any row of $\mathbf{X}$, $\text{Cov}[X_j, X_k] = \text{Cov}[Z_j+U, Z_k+U] = \text{Var}\ [U] = 1$, and $\text{Var}[X_j] = \text{Var}[Z_j+U] = 2$. Thus, $\text{Corr}[X_j, X_k] = \text{Cov}[X_j, X_k]/(\text{Var}[X_j]\text{Var}[X_k])^{1/2} = 1/2$.

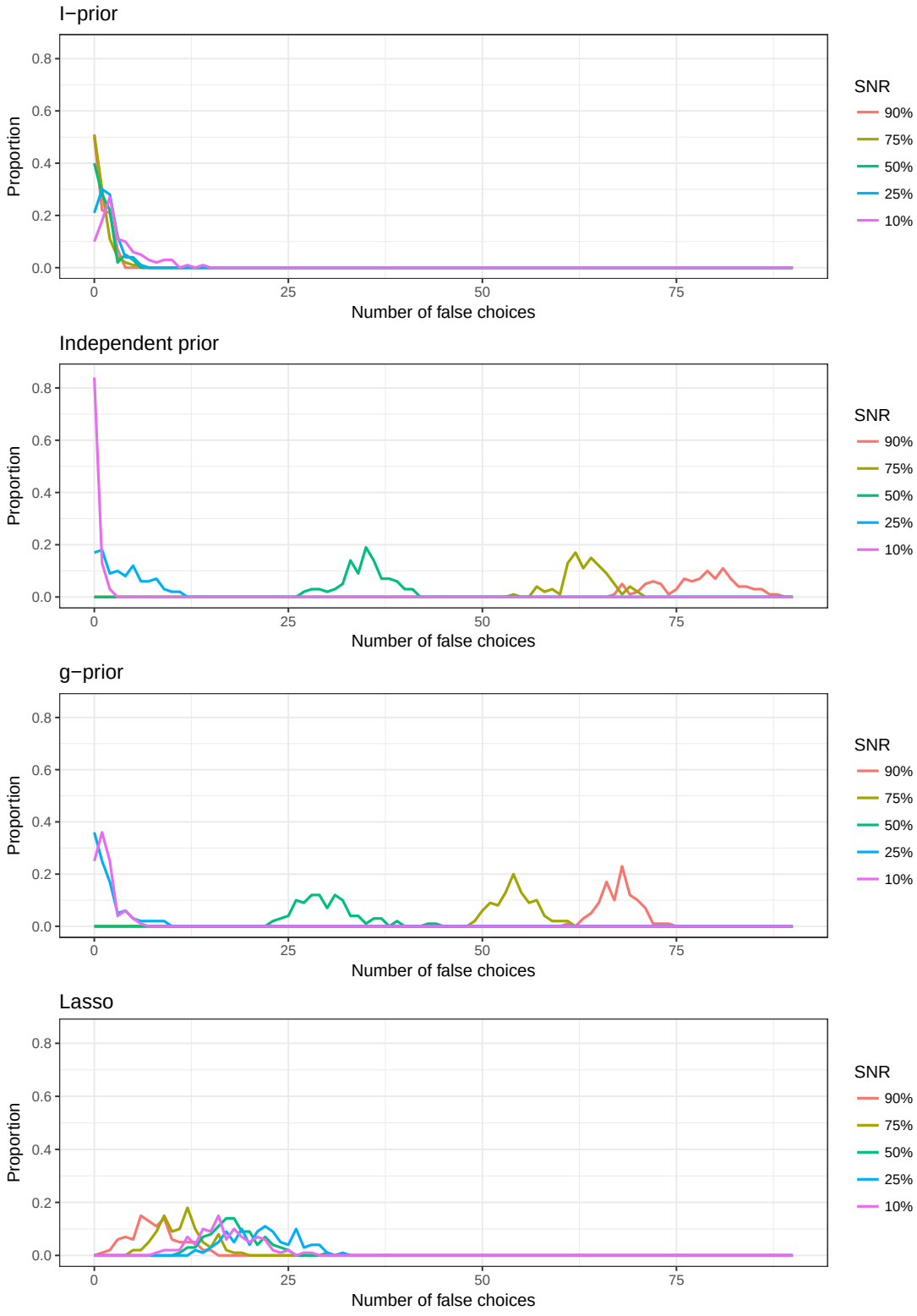


Figure 6.1: Frequency polygons for the number of false choices for each of the four priors. The I-prior performs robustly well across the five scenarios tested, mostly yielding five or fewer false inclusions or exclusions. Spurious exclusions led to the independent and g-prior simultaneously performing well in low SNR and badly in high SNR scenarios. The Lasso is known to be unreliable in the presence of collinearity.

We also inspect the sensitivity of the hyperprior choice of π_j for the indicator variables on the number of false choices made. Figure 6.2 plots the mean number of false choices made in each of the five SNR scenarios with varying hyperprior setting for π_j. From the plot, it is seen that for high SNR scenarios, setting π_j too low increases the number of false exclusions. Conversely, for low SNR scenarios, setting π_j too high increases the number of false inclusions. This makes sense: when the true model size is small, then setting π_j too high encourages variables to be retained in the model. While the optimal π_j corresponds directly to the true SNR (e.g. SNR = 10% performs best under $\pi_j = 0.10$), Figure 6.2 makes a case for $\pi_j = 0.5$ to be a 'safe choice' in the face of prior ignorance on model size.

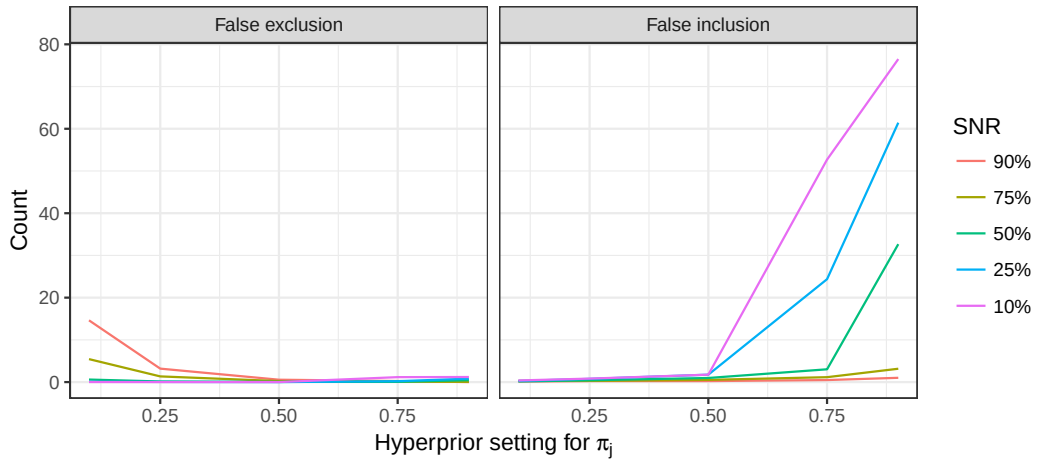


Figure 6.2: Average number of false choices (false inclusions or false exlusions) for the five different scenarios (SNR varied between 90%, 75%, 50%, 25% and 10%) with different hyperprior setting for $\gamma_j \sim \text{Bern}(\pi_j)$.

6.7 Examples

Now, we apply our I-prior Bayesian variable selection model to three real-world data sets that have all been previously analysed in the variable selection literature. All examples were analysed in R using our **ipriorBVS** package (Jamil, 2018) which contains a wrapper to JAGS (Plummer, 2003). Reproducible code is available at `http://myphdcode.haziqj.ml`. In all analyses, a two-stage procedure was conducted for the I-prior model, where each stage consists of obtaining 15,000 MCMC samples (including 5,000 for burn-in).

6.7.1 Aerobic data set

This dataset appeared in the *SAS/STAT® User's Guide* (SAS Institute Inc., 2008) and was also analysed by L. Kuo and Mallick (1998). It involves understanding the factors

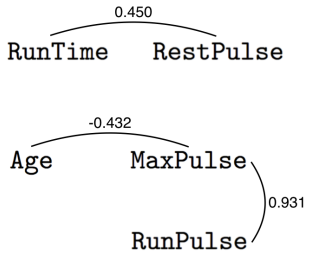


Figure 6.3: The sample correlations of interest in the aerobic fitness dataset. These show variables with correlations greater than 0.4 in magnitude.

which affect aerobic fitness, which is measured by the ability to consume oxygen. A sample of $n = 30$ male participants' had their physical fitness measured by means of simple exercise tests. The response variable contains measurement of oxygen uptake rate in mL/kg body weight per minute. The six covariates were the participants' age (X_1), weight (X_2), time taken to run one mile (X_3), resting heart rate (X_4), heart rate while running (X_5), and maximum heart rate during the exercise (X_6). This dataset, although small in size, is interesting to analyse because of the correlations between the variables, mainly due to the measurements being taken during the same exercise test. The sample correlations of interest are shown in Figure Figure 6.3.

Table 6.3: Results for variable selection of the Aerobic data set. Note that the Bayes factors reported are the Bayes factors comparing any of the models to Model 1 (base model).

	PIP	$\boldsymbol{\theta}$ est. (SD)	Model 1	Model 2	Model 3	Model 4
X_1	0.685	−0.169 (0.14)	✓		✓	
X_2	0.205	−0.017 (0.05)				
X_3	1.000	−0.745 (0.12)	✓	✓	✓	✓
X_4	0.168	−0.013 (0.05)				
X_5	0.663	−0.163 (0.15)	✓			✓
X_6	0.275	0.003 (0.10)				
		PMP	0.564	0.235	0.105	0.096
		BF	1.000	0.418	0.187	0.170

Notice that Table 6.3 reports only on four of a possible $2^6 = 64$ models, and realise that the sum of the posterior model probabilities add to one. Naturally, models which are deemed important by virtue of data evidence are sampled more often, and in fact, models which are unpromising may not even get sampled. So, MCMC methods does not need to list out all possible models because models which are never visited in the posterior state space are assigned a probability of zero. The highest posterior model was found to be the model with the variables X_1, X_3 and X_5 (PMP = 0.564). In Figure 6.4, we can see that the point mass at zero overwhelms the rest of the values in the density plots for X_2, X_4 and X_6, and hence these variables were dropped.

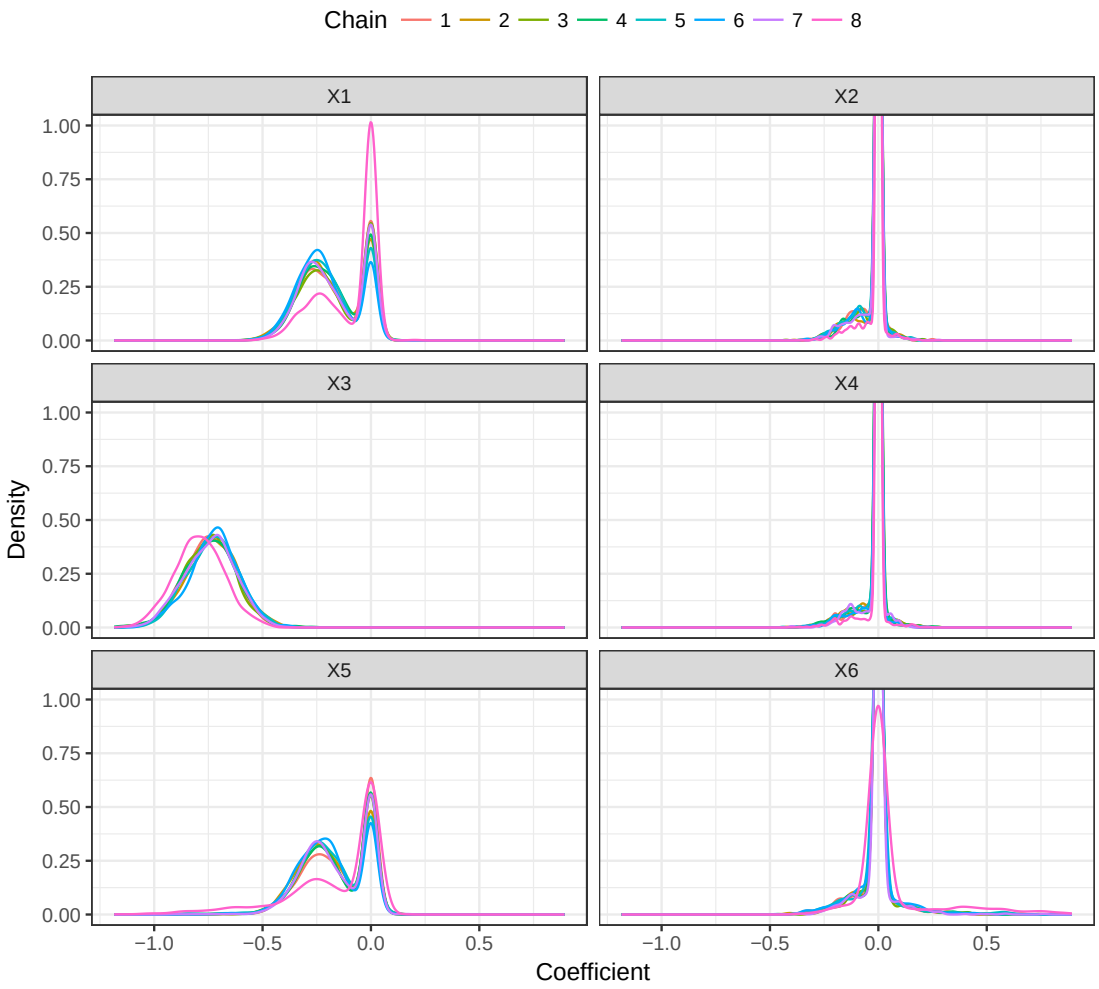


Figure 6.4: Posterior density plots of the regression coefficients $\boldsymbol{\theta}$ for the aerobic data set. The 'spike' at zero observed in the density plots for X_2, X_4 and X_6 is indicative of these variable being dropped often in the posterior samples.

6.7.2 Mortality and air pollution data

The next real world application comes from a paper by McDonald and Schwing (1973). In it, the effects of air pollution on mortality in a US metropolitan area ($n = 60$ and $p = 15$) were studied. The response variable is the total age adjusted mortality rate, and the main pollution effects of interest were that of hydrocarbons (HC), oxides of nitrogen (NO_x) and sulphur dioxide (SO_2). Several other environmental and socioeconomic considerations were taken into account, otherwise the model may include unexplained variation caused by factors other than pollution. For example, a metropolitan area with a high proportion of the elderly should expect to have a higher mortality rate than one with a low proportion. All of the variables can be considered as continuous and real; Table 6.4 provides a description of the variables.

Table 6.4: Description of the air pollution data set.

Variable	Description
Mortality	Total age adjusted mortality rate
Precipitation	Mean annual precipitation (in)
Relative humidity	Percent relative humidity, annual average at 1 p.m.
January temperature	Mean January temperature (°F)
July temperature	Mean July temperature (°F)
Population density	Population per square mile in urbanised area
Household size	Population per household
Education	Median school years completed for those over 25
Sound housing units	Percentage of sound housing units (no defects)
Age >65 years	Percent of population that is 65 years of age or over
Non-white	Percent of urbanised area population that is non-white
White collar	Percent employment in white-collar urbanised occupations
Income <$3,000	Percent of families with income under $3,000
HC	Relative population potential of hydrocarbons
NO_x	Relative population potential of oxides of nitrogen
SO_2	Relative population potential of sulphur dioxide

This dataset also contains several highly correlated variables which impedes a meaningful regression analysis. When the full model is fitted using ordinary least squares, none of the pollutant effects were found to be significant. Clearly, a variable selection method was required. McDonald and Schwing (1973) used a ridge regression technique to determine which variables to select and eliminate "unstable" coefficients found from a trace analysis. In addition, the authors also looked at a variable elimination method based on total squared error via Mallow's C_p. The results are summarised in Table 6.5.

In this case, the I-prior BVS model concurred with the overall finding of McDonald and Schwing (1973), in that SO_2 was found to be a significant contributing factor towards mortality rates, while the rest of the pollutants were not. the I-prior BVS model also obtained a model with the largest R^2 and the smallest size. We note that the effect size for SO_2 is slightly larger under an I-prior, but generally, the rest of the I-prior coefficients are similar in magnitude and sign to the coefficients of the other two models.

Table 6.5: A comparison of the coefficient values obtained using ordinary least squares (full model), McDonald and Schwing's minimum C_p and ridge analysis, and the I-prior. Standard errors/posterior standard deviations are given in parentheses. Values shaded grey indicate OLS regression coefficients not significant at the 10% level.

	Full model	Min. C_p	Ridge	I-prior
Environmental factors				
Precipitation	0.306 (0.14)	0.247 (0.07)	0.230 (0.07)	0.254 (0.12)
Relative humidity	0.009 (0.10)			
January temperature	-0.318 (0.18)	-0.164 (0.06)	-0.172 (0.06)	-0.195 (0.11)
July temperature	-0.237 (0.15)	-0.073 (0.07)		
Demographic factors				
Population density	0.084 (0.09)		0.091 (0.06)	
Household size	-0.232 (0.15)			
Education	-0.233 (0.16)	-0.190 (0.06)	-0.171 (0.07)	-0.151 (0.12)
Sound housing units	-0.052 (0.15)			
Age >65 years	-0.213 (0.20)			
Non-white	0.640 (0.19)	0.481 (0.07)	0.462 (0.07)	0.517 (0.10)
White collar	-0.014 (0.12)			
Income <$3,000	-0.009 (0.22)			
Pollution potential				
HC	-0.979 (0.72)			
NO_x	0.983 (0.75)			
SO_2	0.090 (0.15)	0.255 (0.06)	0.232 (0.06)	0.302 (0.09)
Size	15	6	6	5
R^2	0.764	0.541	0.553	0.676

6.7.3 Ozone data set

In this section, we replicate the Bayesian variable selection analysis of the Ozone dataset done by Casella and Moreno (2006) which appeared initially in Breiman and Friedman (1985), and also show how Bayesian variable selection can help select important interaction terms. The data consists of daily ozone readings and various meteorological quantities, and the aim was to see which of these quantities contributed to the ozone concentration. The variables considered are explained in Table 6.6.

The data contains 366 points, one for each day of the leap year 1976. There are 163 data points containing missing data on some of the predictors, so we did a complete case analysis on the remaining 203 samples. Out of these 203, we randomly set aside 25 to use for validation, thus the n used to train the model was $n = 178$. The training and test set were repeated multiple times and results averaged in order to make a comparison to the unknown training and test set used in the other studies. Out-of-sample prediction RMSE were obtained, as well as the coefficient of determination R^2.

Table 6.6: Description of the ozone data set for the analysis done in Section 6.7.3

Variable	Description
y	Daily maximum one-hour-average ozone reading (ppm) at Upland, CA
X_1	Month: $1 =$ January, $\ldots, 12 =$ December
X_2	Day of month: $1, 2, \ldots$
X_3	Day of week: $1 =$ Monday, $\ldots, 7 =$ Sunday
X_4	500-millibar pressure height (m) measured at Vandenberg AFB
X_5	Wind speed (mph) at Los Angeles International Airport (LAX)
X_6	Humidity (%) at LAX
X_7	Temperature (°F) measured at Sandberg, CA
X_8	Inversion base height (feet) at LAX
X_9	Pressure gradient (mmHg) from LAX to Daggett, CA
X_{10}	Visibility (mi) measured at LAX
X_{11}	Temperature (°F) measured at El Monte, CA
X_{12}	Inversion base temperature (degrees Fahrenheit) at LAX

Casella and Moreno removed the variables X_{11} and X_{12} before running their selection model, citing multicollinearity causing ill-conditioned design matrices. Upon inspection, there are indeed correlations among the variables as high as 0.93 for some of them, but not enough to cause rank deficiency in the design matrix and a degenerate $\mathbf{X}^\top\mathbf{X}$ matrix. The correlations $\text{Corr}(X_7, X_{11}) = 0.91$ and $\text{Corr}(X_{11}, X_{12}) = 0.93$ seemed to drive the decision to drop the two variables, and while it is a valid concern, we will conduct variable selection on the full set of 12 variables. We can then see the performance of I-priors in the presence of multicollinearity in this real-world data set. On another note, the variables X_1, X_2 and X_3 were presumably intended to be categorical as in modelling seasonality in a time series data, but these were treated as continuous, as did Casella and Moreno. The results are compared in Table 6.7.

Table 6.7: Results for variable selection of the Ozone data set using only linear predictors.

Method	Variables	Size	R^2	RMSE
I-prior	X_1, X_6, X_{11}	3	0.708	0.554
Casella and Moreno (C&M)	X_6, X_7, X_8	3	0.686	0.992
Breiman and Friedman (B&F)	X_7, X_8, X_9, X_{10}	4	0.669	1.056

What we found was that the model selected using the I-prior does better in terms of R^2 as well as RMSE compared to the methods used by C&M and B&F. The average posterior model probability for X_1, X_6, X_{11} as found by the I-prior was 0.722[3]. One thing to note is that the I-prior model selected the variable X_{11} instead of its highly correlated proxy X_7, which is what C&M selected. These two variables are temperature measurements at different locations in California. As C&M excluded X_{11} from the model search it was of course never considered in their model selection process, and because

we included it in ours, the variable selection method was able to consider both variables together and decide on the more appropriate one.

Interestingly, it turns out that Sandberg, CA (location of temperature measurement for X_7) is about 121km away from Upland, CA (location of ozone reading), but El Monte, CA (location of temperature measurement for X_{11}) is only 35km away from Upland, CA. It stands to reason that X_{11} provides more geographical reliability than X_7. Unless there is a strong insistence on deleting variables beforehand, we might not know for sure whether the variable was rightfully removed from consideration, as this example seems to prove. Out of curiosity, running the variable selection model on the reduced variable space as C&M did, we arrive at the same results as theirs.

We then used the I-prior method to select between the squared terms and all level two interactions, in addition to all the variables, in an effort to improve model prediction. For 12 such variables, the number of variables to select becomes $12+12+12(12-1)/2 = 90$. By doing so, we were able to improve the model to give a slightly better predictive ability. The results are shown below in Table 6.8. The I-prior again selected a model which was superior in terms of R^2 and RMSE compared to that obtained by C&M.

Table 6.8: Results for variable selection of the Ozone data set using linear, squared and two-way interaction terms.

Method	Variables	Size	R^2	RMSE
I-prior	$X_1, X_5, X_6, X_{11}, X_{12}, X_1^2, X_9^2, X_6X_{11}, X_6X_{12}, X_7X_9$	10	0.812	0.503
C&M	$X_2, X_1^2, X_7^2, X_9^2, X_1X_5, X_2X_6, X_3X_7, X_4X_6, X_6X_8, X_6X_{10}$	10	0.758	0.873

6.8 Conclusion

The model selection problem is an important one in statistics, but highly contentious. Miller (2002) writes that many statisticians view model selection as 'unclean' or 'distasteful', and that "terms such as 'fishing expeditions', 'torturing the data until they confess', 'data mining', and others are used as descriptions of these practices". The disagreement with the principle of model selection stems from the belief in the mantra that models should be built by thoughtfully choosing variables which are expected to influence the response by appealing to substantive theory, and not by virtue of optimising some model selection criterion. However, variable selection as an exploratory study is certainly justified by many practical applications, especially when there is a genuine desire to know the most reasonable, parsimonious and interpretable model. Through

[3]Since the total model space used was different between our method, C&M and B&F, it does not make sense to compare posterior model probabilities which we obtained. C&M reported a model probability of 0.491 for their model, but this model was not selected at all using the I-prior.

variable selection exercises, we can learn which covariates are important, and which are negligible, in explaining the variation in the response.

The Bayesian variable selection method that we have seen has the appeal of reducing the problem of model search into one of estimation. At the outset, we aimed to seek a model which: 1) requires little tuning on the part of the user; 2) would work well in the presence of multicollinearity; and 3) is able to work well with little to no prior information. The I-prior on the regression coefficients in L. Kuo and Mallick's spike-and-slab stochastic search framework achieves this aim.

The attractive feature of a Bayesian approach to variable selection is the ability to simultaneously shrink and select predictors, thereby incorporating model uncertainty in the regressors. Sparsification is not "hard coded", in the sense that regression coefficients are assigned a value of zero with some positive probability in the posterior. This is unlike the regularisation or penalised log-likelihood approach to variable selection using the Lasso, elastic net, and so on, whereby sparsity is induced at the mode, but not in the posterior distribution (Scott and Varian, 2014). This translates to being provided with a single variable selection decision, rather than information that is coded through a probability distribution.

We discuss three areas to concentrate on for future research and improvement:

1. $\boldsymbol{p > n}$ **cases**. Typically, when there is insufficient information in the data to inform the estimation, then additional information is sought from the priors. In our case, the I-prior covariance involves the inverse of a low rank matrix which is not invertible. A p-variate normal distribution with a singular covariance matrix will only have a probability distribution defined on a low dimensional subspace. The issue may however be computational—it might be worth exploring the generalised inverse, or study ways in which to avoid the inverse computation in the Gibbs sampler. As a matter of fact, we note that the posterior precision for $\boldsymbol{\beta}$ can be written as

$$\begin{aligned}\tilde{\mathbf{B}}^{-1} &= \big(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma + (\kappa\mathbf{X}^\top\mathbf{X})^{-1}\big)^{-1} \\ &= \mathbf{X}_\gamma^\top\mathbf{X}_\gamma\big((\mathbf{X}_\gamma^\top\mathbf{X}_\gamma)^2 + \kappa\mathbf{I}_p\big)^{-1}\end{aligned}$$

   which avoids the need for inverting the low-rank matrix $\mathbf{X}_\gamma^\top\mathbf{X}_\gamma$.

2. **Improvement in computational time**. Although the model itself is not computationally intensive to run (roughly $O(np^2)$ in time per Gibbs iteration), the main bottleneck is the reliance on a stochastic sampling algorithm. As in the previous chapter, variational inference is a promising area to look into, especially given that the Gibbs conditional distributions were straightforward to obtain, and these might be similar to a mean-field variational distribution. If this is successful,

then it is expected to reduce computational time and avoid convergence issues that comes with traditional MCMCs. Variational inference with spike-and-slab priors on regression coefficients was studied by Ormerod et al. (2017).

3. **Extension to generalised linear models**. L. Kuo and Mallick (1998) in their paper already provided a sketch of how the variable selection model would work. With the ideas in Chapter 5, we can extend the I-prior variable selection to categorical responses when the continuous latent propensities are modelled using linear functions. Such an approach has been implemented in gene selection studies, for which the variables are gene expressions and the responses are presence of a particular disease (Lee et al., 2003).

Finally, it should be mentioned that more complex variable selection models can be coded with the γ indicators. For instance, in selecting squared or interaction terms, we can insist on having the model select the main term if the squared or interaction term is selected:

$$y_i = \alpha + \gamma_1\beta_1 x_{1i} + \gamma_2\beta_2 x_{2i} + \gamma_1\gamma_2\gamma_3\beta_3 x_{1i}x_{2i}.$$

Or perhaps, we could use a single γ indicator for the dummy variables which make up a single categorical covariate, which we would then infer on the selection of the single covariate rather than each individual category of the covariate.

Chapter 7

Summary

The work done in this thesis explores the concept of regression modelling using priors with Fisher information covariance kernels (I-priors, Bergsma, 2018). It is best seen as a flexible regression technique which is able to fit both parametric and nonparametric models, and bears similarity to Gaussian process regression. For the regression model (1.1) subject to (1.2), stated again here for convenience,

$$
\begin{aligned}
y_i &= \alpha + f(x_i) + \epsilon_i && \text{(from 1.1)} \\
(\epsilon_1, \dots, \epsilon_n) &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}^{-1}) && \text{(from 1.2)} \\
i &= 1, \dots, n,
\end{aligned}
$$

and it is assumed that the regression function f lies in some reproducing kernel Hilbert or Krein space $\mathcal{F}$ with kernel h_η defined over the set of covariates $\mathcal{X}$. In Chapter 2, we built a primer on basic functional analysis, and described various interesting RKHS/RKKS for regression modelling.

We then ascertained the form of the Fisher information for f, treated as a parameter of the model to be estimated, and from Corollary 3.3.1, it is

$$
\begin{aligned}
\mathcal{I}\big(f(x), f(x')\big) &= \sum_{i,j=1}^{n} \psi_{ij} h_\eta(x, x_i) h_\eta(x', x_j) \\
&= \mathbf{h}_\eta(x)^\top \boldsymbol{\Psi} \mathbf{h}_\eta(x'),
\end{aligned}
$$

for any two points x, x' in the domain of f, obtained using appropriate calculus for topological spaces detailed in Chapter 3. An I-prior for f is defined as Gaussian with mean function f_0 chosen a priori, and covariance function equal to the Fisher information.

The I-prior for f has the simple representation

$$f(x_i) = f_0(x_i) + \sum_{k=1}^{n} h_\eta(x_i, x_k) w_k$$
$$(w_1, \ldots, w_n)^\top \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi})$$
$$i = 1, \ldots, n,$$

and is written equivalently as the Gaussian process prior

$$\big(f(x_1), \ldots, f(x_n)\big)^\top \sim \mathrm{N}_n(\mathbf{f}_0, \mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta),$$

where $\mathbf{H}_\eta = \big(h_\eta(x_i, x_j)\big)_{i,j=1}^n$.

In Chapter 4, we looked how the I-prior model has wide-ranging applications, from multilevel modelling, to longitudinal modelling, and modelling with functional covariates. Estimation was conducted mainly using a simple EM algorithm, although direct optimisation and fully Bayesian estimation using MCMC is also possible. In the case of polytomous responses, we used a latent variable framework in Chapter 5 to assign I-priors to latent propensities which drive the outcomes under a probit-transform scheme. An extension of the EM algorithm was considered, in which the E-step was replaced with variational inference, so as to overcome the intractability brought about by the conditional distributions. For both continuous and categorical response I-prior models, we find advantages of using I-priors, namely that model building and estimation is simple, inference straightforward, and predictions comparable, if not better, to similar state-of-the-art techniques.

Finally, in Chapter 6, we dealt with the problem of model selection, specifically for linear regression models. There, we used a fully Bayesian approach for estimating model probabilities in which regression coefficients are assigned an I-prior. We devised a model that requires minimal tuning on the part of the user, yet performs well in simulated and real-data examples, especially if multicollinearity exists among the covariates.

7.1 Summary of contributions

We give a summary of the novel contributions of this thesis.

- **Fisher information for infinite-dimensional parameters**. When the RKHS/ RKKS $\mathcal{F}$ is infinite-dimensional (e.g. covariates are themselves functions), then the Fisher information involves derivatives with respect to an infinite-dimensional vector. Finite-dimensional results using component-wise/partial derivatives may fail in infinite dimensions. The technology of Fréchet and Gâteaux differentials

accommodate for the fact that f may be infinite-dimensional, which, at minimum, requires $\mathcal{F}$ to be a normed vector space. We foresee the work of Section 3.2 being applicable elsewhere, such as learning in (reproducing kernel) Banach spaces (Haizhang Zhang et al., 2009; Haizhang Zhang and J. Zhang, 2012), or in the theory of parameter estimation for general exponential family type distributions of the form

$$p(X|\theta) = B(X)\exp\big(\langle\theta, T(X)\rangle_{\mathcal{H}} - A(\theta)\big),$$

for which θ lies in some inner-product space $\mathcal{H}$ which might be infinite-dimensional (Sriperumbudur et al., 2017).

- **Efficient estimation methods for normal I-prior models**. The preferred estimation method for normal I-prior models for stability is the EM algorithm. Implementing the EM algorithm can be computationally costly, due to the squaring and inversion of the kernel matrices in the Q function in (4.15). Unfortunately, not much can be done about the inversion part, but we explored some ways to perform the squaring methodically. Combining a 'front-loading method' of the kernel matrices (Section 4.3.2) and an exponential family ECM (expectation conditional maximisation) algorithm (Meng and Rubin, 1993), the estimation procedure is streamlined. Our computational work culminated in the publicly available and well-documented R package **iprior** (Jamil, 2017) published on CRAN.

- **Methodological extension of I-priors to categorical responses**. Extension of the I-prior methodology to fit categorical responses is of great interest. We proposed a latent variable framework, for which there corresponds latent propensities corresponding to each category of the observed response variable. Instead of modelling the responses directly, the latent propensities are modelled using an I-prior, and class probabilities obtained using a normal integral. We named this model the I-probit model. The challenge of estimation is overcoming said integral, and we used a variational EM algorithm in which the E-step uses a variational approximation to intractable conditional density. The variational EM algorithm was preferred over a fully Bayesian variational inference algorithm for two reasons: 1) the work done in the continuous case EM algorithm applies directly; and 2) prior specification for hyperparameter can be dispensed with. Classification, meta-analysis and spatio-temporal modelling are specific examples of the applications of the I-probit models.

- **Some distributional results for truncated normals**. In deriving the variational algorithm, some properties related to the conically truncated multivariate independent normal distribution (as defined in Appendix C.4) were required. A small contribution of ours was to derive the closed-form expressions for its first and second moments, and its entropy (Lemma C.5). We have only seen closed-

form expressions of the mean of such a distribution being used before (Girolami and Rogers, 2006) but not for the variance, nor an explicit derivation of these quantities.

- **Bayesian variable selection under collinearity**. Model comparison using likelihood ratio tests or Bayes factors is fine when the number of models under consideration is fairly small. Under a fully Bayesian scheme, we use MCMC to approximate posterior model probabilities of competing linear models. At the outset, we sought a model which required minimal intervention on the part of the user. The I-prior achieved this, with the added advantage of performing well under multicollinearity.

7.2 Open questions

In closing, we briefly discuss several questions which remain open during the course of completing this project.

- **Initialisation of EM or gradient-based methods**. Figure 4.1 indicates the impact that starting values can have on gradient-based optimisation. One can end up at a local optima on one of the two ridges. Usually, one of the ridges will have a higher maximum than the other, but it is not clear how to direct the algorithm in the direction of the 'correct' ridge.

  Importantly, the interpretation of a flat ridge in the likelihood is that there is insufficient information coming from the data to inform parameter estimation. In the EM algorithm, estimation is usually characterised by a fast increase in likelihood in the first few steps (as it climbs up the ridge), and then later iterations only improve the likelihood ever so slightly (as it moves along the ridge in search of the maximum). In some real-data cases (e.g. Tecator data set), we noticed that the EM sequence veers to the boundary of the parameter space, where the likelihood is infinite (e.g. $L(\psi) \to \infty$ as $\psi \to 0, \infty$).

  Ill-posed problems similar to this are resolved by adding penalty terms to the log-likelihood. As to what penalty terms are appropriate remains an open question.

- **Standard errors for variational approximation**. Under a variational scheme, the log-likelihood function $L(\theta)$ is replaced with the ELBO $\mathcal{L}_q(\theta)$ which serves as a conservative approximation to it. The question we have is whether the approximation degrades the asymptotic properties of the estimators obtained via variational inference? In particular, are the standard errors obtained from the information matrix involving $\mathcal{L}_q(\theta)$ reliable? This question has also been posed by Bickel et al. (2013), Y.-C. Chen et al. (2018), and Hall et al. (2011).

Variational methods for maximum likelihood learning can be seen as a deliberate misspecification of the model to achieve tractibility. As such, the variational EM has been referred to as obtaining pseudo- or quasi-ML estimates. The quasi-likelihood literature has results relating to efficiency of parameter estimates (adjustments to the information matrix is needed), and we wonder if these are applicable for variational inference.

Also, obtaining standard errors directly from an EM algorithm is of interest, especially under a variational EM setting. Though this is described in McLachlan and Krishnan (2007, Ch. 4), we have not seen this implemented widely.

- **Comparison of logistic and probit links**. For general binary and multinomial models, the logistic link function sees more prevalent use than its probit counterpart. Of course, we chose the probit as it has distributional advantages which we can exploit for estimation using variational inference. However, is there a difference between the behaviour of the probit and logistic model? We know that there is a difference between the logistic and normal distribution, especially in scaling and behaviour in the tails, but do these affect the outcome of I-prior models?

- **Consistency of I-prior Bayesian variable selection**. We wondered about model selection consistency for I-priors in Bayesian variable selection. That is, assuming that model M_{true} is behind the true data generative process, do

$$\lim_{n\to\infty} \mathrm{P}(M_{\text{true}}|\mathbf{y}) = 1 \quad \text{and} \quad \lim_{n\to\infty} \mathrm{P}(M_k|\mathbf{y}) = 0, \forall M_k \neq M_{\text{true}}$$

  hold for the I-prior Bayesian variable selection methodology? In machine learning, this property is referred to as the *oracle property*. For the g-prior specifically, model consistency results were obtained by Fernández et al. (2001) and Liang et al. (2008). Casella et al. (2009) also looks at consistency of Bayesian procedures for a wide class of prior distributions, but we have yet to examine whether the I-prior falls under the remit of their work.

Supplementary S1

Basic estimation concepts

Statistics concerns what can be learned from data (Davison, 2003). A statistical model comprises of a probabilistic component which drives the data generative process, in addition to a systematic or deterministic component, which sets it apart from pure mathematical models. Real-valued observations $\mathbf{y} := \{y_1, \ldots, y_n\}$ are treated as realisations from an assumed probability distribution with parameters $\theta = (\theta_1, \ldots, \theta_p)^\top \in \Theta$. The crux of statistical inference is to estimate θ given the observed values, so that this optimised value may be used in the model to make deductions. We describe the *frequentist* and *Bayesian* paradigms for parameter estimation.

S1.1 Maximum likelihood estimation

In the frequentist setting, the *likelihood* function, or simply likelihood, is a function of the parameters θ which measures the plausibility of the parameter value given the observed data to fit a statistical model. It is defined as the mapping $\theta \mapsto p(\mathbf{y}|\theta)$, where $p(\mathbf{y}|\theta)$ is the probability density function (or in the case of discrete observations, the probability mass function) of the modelled distribution of the observations.

It is logical to consider the parameter set which provides the largest likelihood value,

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta} p(\mathbf{y}|\theta). \tag{S1.1}$$

The value $\hat{\theta}$ is referred to as the *maximum likelihood estimate* for θ. For convenience, the *log-likelihood* function $L(\theta) = \log p(\mathbf{y}|\theta)$ is maximised instead; as the logarithm is a monotonically increasing function, the maximiser of the log-likelihood function is exactly the maximiser of the likelihood function itself.

When ML estimates are unable to be found in closed-form, the maximisation problem of (S1.1) requires iterative, numerical methods to find the maximum. These methods are often *gradient-based* methods, algorithms that make use of the gradient of the objective function to be optimised. Examples include Newton's method, Fisher's scoring, quasi-Newton methods, gradient descent, and conjugate gradient methods. As the name suggests, these methods require

evaluation of gradients or approximate gradients, and in some cases, the Hessian. Depending on the situation, gradients or Hessians can be expensive or inconvenient to compute or approximate. In cases of multi-modality of the objective function, the algorithms can potentially converge to a local optima, as it is known that the algorithms are quite sensitive to starting locations.

Besides invariance, the ML estimate comes with the attractive limiting property $\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta_{\text{true}}) \xrightarrow{\text{dist.}} \text{N}_p\left(0, \mathcal{I}(\theta)^{-1}\right)$ (Casella and R. L. Berger, 2002) as sample size $n \to \infty$, where $\mathcal{I}(\theta)$ is the Fisher information for θ. Other asymptotic properties of the ML estimate include consistency, i.e. $\text{P}(\|\hat{\theta}_{\text{ML}} - \theta_{\text{true}}\| > \epsilon) \xrightarrow{\text{prob.}} 0$ for any $\epsilon > 0$, and efficiency, i.e. it achieves the Cramér-Rao lower bound $\text{Var}[\theta_{\text{ML}}] \geq \mathcal{I}(\theta)^{-1}$.

As the likelihood measures the plausibility of a parameter value given the data, it can be used to compare two competing models. Let $\Theta_0 = \{\theta \,|\, \theta_{d+1} = \theta_{d+1,0}, \ldots, \theta_p = \theta_{p,0}\}$ be the set of parameters with restrictions on the last d components of θ. The *likelihood ratio test* statistic for testing the null hypothesis $\text{H}_0 : \theta \in \Theta_0$ against the alternative $\text{H}_1 : \theta \notin \Theta_0$ is

$$\lambda = -2 \log \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = -2\big(\log L(\hat{\theta}_0) - \log L(\hat{\theta})\big), \tag{S1.2}$$

where $\hat{\theta}_0 = \arg\max_{\theta \in \Theta_0} \log p(\mathbf{y}|\theta)$. Wilks' theorem states that λ has an asymptotic chi-squared distribution with degrees of freedom equal to the number of restrictions imposed (or rather, the difference in dimensionality of Θ and Θ_0). This gives a convenient way of comparing nested models.

As a remark, models with more parameters will always have higher, or similar, log-likelihood, than models with fewer parameters, because the model has a better ability to fit the data with more free parameters. In a linear regression setting, this relates to overfitting: a linear model with as many explanatory variables as there are data points ($n = p$) will extrapolate every point in the data set. Overfitting is an oft cited problem of maximum likelihood.

S1.2 Bayesian estimation

The *Bayesian* approach to estimating θ takes a different outlook, in that it supplements what is already known from the data with additional information in the form of prior beliefs about the parameters. This usually means treating the parameters as random, following some distribution dictated by a *prior density* $p(\theta)$. There are many ways of categorising different types of priors. Broadly speaking, priors, and hence Bayesian analysis (Kadane, 2011; Robert, 2007), can be either *subjective* or *objective*, with the demonyms 'subjectivists' and 'objectivists' used to refer to those subscribing to each respective principle. Subjectivists assert that probabilities are merely opinions, while objectivists, in contrast, view probabilities as an extension of logic. In this regard, objectives Bayes seek to minimise the statistician's contribution to inference and 'let data speak for itself', while subjective Bayes does the opposite.

In either case, inference about the parameters are then performed using the *posterior density*

$$p(\theta|\mathbf{y}) \propto \overbrace{p(\mathbf{y}|\theta)}^{\text{likelihood}} \times \overbrace{p(\theta)}^{\text{prior}}, \tag{S1.3}$$

rather than through a single point estimate such as the ML estimate in the frequentist case. The posterior density encapsulates the uncertainty surrounding the parameters θ after observing the data $\mathbf{y}$. The *posterior mean*

$$\tilde{\theta} = \int \theta p(\theta|\mathbf{y}) \, \mathrm{d}\theta \tag{S1.4}$$

is normally taken to be the point estimate for θ, with its uncertainty usually reported in the form of a *credible interval*: if θ_k is the k'th component of θ, then a $(1-\alpha) \times 100\%$ credible interval for θ_k is (θ_k^l, θ_k^u), where $\mathrm{P}(\theta_k^l \leq \theta_k \leq \theta_k^u) = (1-\alpha) \times 100\%$. Under a quadratic loss function, $\tilde{\theta}$ minimises the expected loss $\mathrm{E}[(\theta - \theta_{\text{true}})^2]$ (J. O. Berger, 1985, §4.4.2, Result 3), and is hence also viewed as the *minimum mean squared error* (MMSE) estimator.

On a practical note, integration over the parameter space may be intractable, for instance, the model consists of a large number of parameters for which we would like the posterior mean of, or the marginalising integral cannot be found in closed form. Markov chain Monte Carlo (MCMC) methods are the standard way of approximating such integrals, by way of random sampling from the posterior. The sample $\{\theta^{(1)}, \ldots, \theta^{(T)}\}$ is then manipulated in a way to derive its approximation. In the case of the posterior mean,

$$\hat{\mathrm{E}}[\theta|\mathbf{y}] = \frac{1}{T} \sum_{i=1}^{T} \theta^{(t)} \tag{S1.5}$$

gives an approximation, and its $(1-\alpha) \times 100\%$ credible interval can be approximated using the lower $\alpha/2 \times 100\%$ and upper $(1-\alpha/2) \times 100\%$ quantile of the sample.

The normalising constant is the marginal likelihood over the distribution of the parameters, $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta) \, \mathrm{d}\theta$. The quantity $p(\mathbf{y})$ is also known as the *model evidence*, or simply, *evidence*. As its name suggests, model evidence is used as a measure of how much support there is for a particular model. As such, it is used as a basis for model comparison. Let $p(\mathbf{y}|M_0)$ and $p(\mathbf{y}|M_1)$ be the model evidence for two competing models M_0 and M_1 respectively. Define the *Bayes factor* for comparing model M_0 against an alternative model M_1 as

$$\mathrm{BF}(M_0, M_1) = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)}. \tag{S1.6}$$

Values of $\mathrm{BF}(M_0, M_1) < 1$ would suggest that the data provides more evidence for model M_1 over M_0.

Note that the model evidence is free of θ because the parameters have been marginalised out, or put another way, considered in entirety and averaged over all possible values of θ drawn from its prior density. Thus, model comparison using Bayes factors differs from the frequentist likelihood ratio comparison in that it does not depend on any one particular set of values for the parameters.

S1.3 Maximum a posteriori estimation

One may also find the value of θ which maximises the posterior,

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\mathbf{y}|\theta)p(\theta), \tag{S1.7}$$

which is the mode of the posterior distribution. This quantity is known as the *maximum a posteriori* (MAP) estimate. It is different from the ML estimate in that the maximisation objective is augmented with with the prior density for θ. In this sense, MAP estimation can be seen as regularisation of the ML estimation procedure, whereby a 'penalty' term is added to avoid overfitting.

MAP estimation is often criticised for not being representative of Bayesian methods. That is, MAP estimation returns a point estimate with no apparent way of quantifying its uncertainty. Furthermore, unlike ML estimators, MAP estimators are not invariant under reparameterisation. If θ is a random variable with density $p(\theta)$, then the pdf of $\xi := g(\theta)$, where $g : \theta \mapsto g(\theta)$ is a one-to-one transformation, is

$$p_{\xi}(\xi) = p_{\theta}\big(g^{-1}(\xi)\big) \left| \frac{\mathrm{d}}{\mathrm{d}\xi} g^{-1}(\xi) \right|. \tag{S1.8}$$

The second term in (S1.8) is called the *Jacobian (determinant)*. Therefore, a different parameterisation of θ will impact the location of the maximum because of the introduction of the Jacobian into the optimisation objective (S1.7).

S1.4 Empirical Bayes

The term *empirical Bayes* (Casella, 1985; Robbins, 1956) refers to procedure in which features of the prior is informed by the data. This is realised by parameterising the prior by a hyper-parameter η, i.e. $\theta \sim p(\theta|\eta)$. Values for the hyper-parameter are clearly important, because they appear in the posterior for θ:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta|\eta)}{p(\mathbf{y}|\eta)} \tag{S1.9}$$

To avoid the subjectivist's approach of specifying values for η a priori, one instead turns to the data for guidance. Information concerning η is contained in the marginal likelihood $p(\mathbf{y}|\eta) = \int p(\mathbf{y}|\theta)p(\theta|\eta)\,\mathrm{d}\theta$. This paves the way for using the *maximum marginal likelihood* estimate

$$\hat{\eta} = \arg\max_{\eta} p(\mathbf{y}|\eta) \tag{S1.10}$$

in place of η in the equation of (S1.9). This procedure is also coined *maximum likelihood type-II* (Bishop, 2006), and is commonly referred to as such in the machine learning literature. It is also commonplace in statistics, especially in random-effects or latent variable models which employ a maximum likelihood procedure such as EM algorithm.

As a remark, estimation of η itself can be made to conform to Bayesian philosophy, i.e., by placing priors on it and inferring η through its posterior. Such a procedure is referred to

as *Bayesian hierarchical modelling*. A motivation for doing this is because the ML estimate of η ignores any uncertainty in it. Of course, the hyper-prior for η could be parameterised by a hyper-hyper-parameter, and itself have a prior, and so on and so forth. Evidently the model is specified until such a point where there are parameters of the model which are left 'unoptimised' and must be specified in subjective manner (Beal, 2003).

Supplementary S2

The EM algorithm

Often times, there are unobserved, random variables $\mathbf{w} = \{w_1, \ldots, w_n\}$ that are assumed to make up the data generative process, prescribed in the statistical model through the *joint pdf* $p(\mathbf{y}, \mathbf{w}|\theta)$. Examples of models that include latent variables are plenty: Gaussian mixture models, latent class analysis, factor models, random coefficient models, and so on. In order to obtain ML estimates through a direct maximisation of the likelihood, it is necessary to first marginalise out the latent variables via

$$p(\mathbf{y}|\theta) = \int \overbrace{p(\mathbf{y}|\mathbf{w}, \theta)p(\mathbf{w}|\theta)}^{p(\mathbf{y},\mathbf{w}|\theta)} \, \mathrm{d}\mathbf{w} \tag{S2.1}$$

and obtain the *marginal likelihood*. Note that the integral is replaced by a summation over all possible values in the case of discrete latent variables $\mathbf{w}$.

Direct maximisation of the marginal (log-)likelihood might not be favourable due to intractability in obtaining ML solutions. The form of the marginal likelihood might might not be conducive for closed-form estimates to be found, necessitating the use of numerical, gradient-based methods which is subject to its own undesirable quirks. Moreover, when the evaluation of the (log-)likelihood, gradient and/or Hessian are expensive to compute, then numerical methods are burdensome to execute.

It is usually the case that if the latent variables $\mathbf{w}$ were somehow known, estimation would be made simpler. That is, the solution to $\arg\max_\theta \log p(\mathbf{y}, \mathbf{w}|\theta)$ can be obtained in a simple manner. The expectation-maximisation algorithm (Dempster et al., 1977), commonly known as the EM algorithm, is an iterative procedure which exploits the fact that the so-called *complete data likelihood* is easier to work with. Correspondingly, in EM terminology, the marginal likelihood is referred to as the *incomplete data likelihood*.

We describe a derivation of both a general EM algorithm and an EM algorithm for models whose data generative pdf belongs to an exponential family of pdfs. Interestingly, the EM algorithm can be modified to obtain maximum a posteriori estimates or penalised log-likelihood solutions. As a note, the EM algorithm is not an algorithm per se, in that it does not provide exact instructions as to what the E- and M-steps should comprise of. Rather, it is a generic device to obtain parameter estimates (McLachlan and Krishnan, 2007).

S2.1 Derivation of the EM algorithm

For want of an iterative procedure to obtain maximum likelihood estimates, we seek a solution to

$$\arg\max_{\theta}\{L(\theta|\mathbf{y}) - L(\theta^{(t)}|\mathbf{y}) \geq 0\}, \tag{S2.2}$$

where the solution to (S2.2) yields an improvement to the current t'th iteration of the log-likelihood value $L(\theta^{(t)}|\mathbf{y})$. Note that the objective function in (S2.2) forms an upper bound for the quantity $Q(\theta|\theta^{(t)})$, as shown below:

$$\begin{aligned}
L(\theta|\mathbf{y}) - L(\theta^{(t)}|\mathbf{y}) &= \log \int p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)\,\frac{p(\mathbf{w}|\mathbf{y},\theta^{(t)})}{p(\mathbf{w}|\mathbf{y},\theta^{(t)})}\,\mathrm{d}\mathbf{w} - \log p(\mathbf{y}|\theta^{(t)}) \\
&\geq \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y},\theta^{(t)})}\,\mathrm{d}\mathbf{w} \quad \text{(Jensen's inequality)} \\
&\quad - \log p(\mathbf{y}|\theta^{(t)}) \int p(\mathbf{w}|\mathbf{y},\theta^{(t)})\,\mathrm{d}\mathbf{w} \\
&= \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y},\theta^{(t)})p(\mathbf{y}|\theta^{(t)})}\,\mathrm{d}\mathbf{w} \\
&=: Q(\theta|\theta^{(t)}).
\end{aligned}$$

Evidently, to maximise $L(\theta|\mathbf{y})$, we can't do any worse than maximising $Q(\theta|\theta^{(t)})$ in θ. Denote by $\theta^{(t+1)}$ as the maximiser of $Q(\theta|\theta^{(t)})$. Then,

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y},\theta^{(t)})p(\mathbf{y}|\theta^{(t)})}\,\mathrm{d}\mathbf{w} \\
&= \arg\max_{\theta} \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log p(\mathbf{y}|\mathbf{w},\theta)p(\mathbf{w}|\theta)\,\mathrm{d}\mathbf{w} \\
&= \arg\max_{\theta} \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log p(\mathbf{y},\mathbf{w}|\theta)\,\mathrm{d}\mathbf{w} \\
&= \arg\max_{\theta} \mathrm{E}_{\mathbf{w}}\left[\log p(\mathbf{w},\mathbf{y}|\theta)|\mathbf{y},\theta^{(t)}\right]
\end{aligned}$$

We arrive at an iterative procedure summarised succintly as the following:

Algorithm 3 EM algorithm

1: **initialise** $\boldsymbol{\theta}^{(0)}$ and $t \leftarrow 0$
2: **while** not converged **do**
3: E-step: compute $Q(\theta|\theta^{(t)}) = \mathrm{E}_{\mathbf{w}}\left[\log p(\mathbf{w},\mathbf{y}|\theta)|\mathbf{y},\theta^{(t)}\right]$
4: M-step: $\theta^{(t+1)} \leftarrow \arg\max_{\theta} Q(\theta|\theta^{(t)})$
5: $t \leftarrow t+1$
6: **end while**

Notice that the log-likelihood function satisfies

$$L(\theta|\mathbf{y}) \geq L(\theta^{(t)}|\mathbf{y}) + Q(\theta|\theta^{(t)}), \tag{S2.3}$$

for which equality is achieved when $\theta = \theta^{(t)}$, since

$$\begin{aligned} Q(\theta^{(t)}|\theta^{(t)}) &= \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w},\theta^{(t)})p(\mathbf{w}|\theta^{(t)})}{p(\mathbf{w}|\mathbf{y},\theta^{(t)})p(\mathbf{y}|\theta^{(t)})}\,\mathrm{d}\mathbf{w} \\ &= \int p(\mathbf{w}|\mathbf{y},\theta^{(t)}) \log \frac{p(\mathbf{y},\mathbf{w}|\theta^{(t)})}{p(\mathbf{y},\mathbf{w}|\theta^{(t)})}\,\mathrm{d}\mathbf{w} \\ &= 0. \end{aligned}$$

This implies that the EM algorithm improves the log-likelihood values at each iteration, since

$$L(\theta^{(t+1)}|\mathbf{y}) - L(\theta^{(t)}|\mathbf{y}) \geq Q(\theta^{(t+1)}|\theta^{(t)}) \geq 0$$

and $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) = 0$ since $\theta^{(t+1)}$ maximises $Q(\cdot|\theta^{(t)})$.

The expectation in the E-step involves the conditional pdf $p(\mathbf{w}|\mathbf{y},\theta^{(t)})$. Viewed through a Bayesian lens, this is the posterior density of the latent variables using the t'th iteration parameter values. The success of the E-step is predicated on the availability of the conditional pdf for the expectation. If not, approximations to the E-step can be explored, for example using Monte Carlo methods (Wei and Tanner, 1990) or a variational approximation (Beal, 2003).

The solution to the M-step usually, but not always, exists in closed form. Maximising the Q function over all possible values of θ may not be feasible (McLachlan and Krishnan, 2007). In such situations, the generalised EM algorithm (as defined by Dempster et al., 1977) requires only that $\theta^{(t+1)}$ be chosen in a way that

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}).$$

That is, $\theta^{(t+1)}$ is chosen so as to increase the value of the Q function at its current parameter value. As seen in the argument above, this requirement is sufficient for a guaranteed increase in the log-likelihood function at each iteration.

S2.2 Exponential family EM algorithm

Consider the density function $p(\cdot|\boldsymbol{\theta})$ of the complete data $\mathbf{z} = \{\mathbf{y}, \mathbf{w}\}$, which depends on parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)^\top \in \Theta \subseteq \mathbb{R}^s$, belonging to an exponential family of distributions. This density takes the form $p(\mathbf{z}|\boldsymbol{\theta}) = B(\mathbf{z}) \exp\big(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{T}(\mathbf{z})\rangle - A(\boldsymbol{\theta})\big)$, where $\boldsymbol{\eta} : \mathbb{R}^s \mapsto \mathbb{R}$ is a link function, $\mathbf{T}(\mathbf{z}) = \big(T_1(\mathbf{z}), \ldots, T_s(\mathbf{z})\big)^\top \in \mathbb{R}^s$ are the sufficient statistics of the distribution, and $\langle \cdot, \cdot \rangle$ is the usual Euclidean dot product. It is often easier to work in the *natural parameterisation* of the exponential family distribution

$$p(\mathbf{z}|\boldsymbol{\eta}) = B(\mathbf{z}) \exp\big(\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z})\rangle - A^*(\boldsymbol{\eta})\big) \tag{S2.4}$$

by defining $\boldsymbol{\eta} := \big(\eta_1(\boldsymbol{\theta}), \ldots, \eta_r(\boldsymbol{\theta})\big) \in \mathcal{E}$, and $\exp A^*(\boldsymbol{\eta}) = \int B(\mathbf{z}) \exp\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z})\rangle \,\mathrm{d}\mathbf{z}$ to ensure the density function normalises to one. As an aside, the set $\mathcal{E} := \big\{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_s) \,|\, \int \exp A^*(\boldsymbol{\eta}) < \infty\big\}$ is called the *natural parameter space*. If $\dim \mathcal{E} = r < s = \dim \Theta$, then the the pdf belongs

to the *curved exponential family* of distributions. If $\dim \mathcal{E} = r = s = \dim \Theta$, then the family is a *full exponential family.*

Assuming the latent $\mathbf{w}$ variables are observed and working with the natural parameterisation, then the complete maximum likelihood (ML) estimate for $\boldsymbol{\eta}$ is obtained by solving

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{z}|\boldsymbol{\eta}) = \mathbf{T}(\mathbf{z}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = 0. \tag{S2.5}$$

Of course, the variable $\mathbf{w}$ are never observed, so the ML estimate for $\boldsymbol{\eta}$ can only be informed from what is observed. Let $p(\mathbf{y}|\boldsymbol{\eta}) = \int p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \, \mathrm{d}\mathbf{w}$ represent the marginal density of the observations $\mathbf{y}$. Now, the ML estimate for $\boldsymbol{\eta}$ is obtained by solving

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}|\boldsymbol{\eta}) &= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}|\boldsymbol{\eta}) \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} \left(\int p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \, \mathrm{d}\mathbf{w} \right) \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \int \left(\frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) \mathrm{d}\mathbf{w} \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \int \left(p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) \mathrm{d}\mathbf{w} \\
&= \int \left(\mathbf{T}(\mathbf{y}, \mathbf{w}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) \right) p(\mathbf{w}|\mathbf{y}, \boldsymbol{\eta}) \, \mathrm{d}\mathbf{w} \\
&= \mathrm{E}_{\mathbf{w}} \left[\mathbf{T}(\mathbf{y}, \mathbf{w}) | \mathbf{y} \right] - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta})
\end{aligned} \tag{S2.6}$$

equated to zero. Note that we are allowed to change the order of integration and differentation provided the integrand is continuously differentiable. So the only difference between the first order condition of (S2.5) and that of (S2.6) is that the sufficient statistics involving the unknown $\mathbf{w}$ are replaced by their conditional or posterior expectations.

A useful identity to know is that $\frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = \mathrm{E}_{\mathbf{z}}[\mathbf{T}(\mathbf{z})]$ (Casella and R. L. Berger, 2002, Theorem 3.4.2 & Exercise 3.32(a)), which can be expressed in terms of the original parameters $\boldsymbol{\theta}$. As a consequence, solving for the ML estimate for $\boldsymbol{\theta}$ from the FOC equations (S2.6) is possible without having to deal with the derivative of A^* with respect to the natural parameters. Having said this, an analytical solution in $\boldsymbol{\theta}$ may not exist, because the relationship of $\boldsymbol{\theta}$ could be implicit in the set of equations $\mathrm{E}_{\mathbf{w}} \left[\mathbf{T}(\mathbf{w}, \mathbf{y}) | \mathbf{y}, \boldsymbol{\theta} \right] = \mathrm{E}_{\mathbf{y},\mathbf{w}} \left[\mathbf{T}(\mathbf{y}, \mathbf{w}) | \boldsymbol{\theta} \right]$. One way around this is to employ an iterative procedure, as detailed in Algorithm 4.

Algorithm 4 Exponential family EM

1: **initialise** $\boldsymbol{\theta}^{(0)}$ and $t \leftarrow 0$
2: **while** not converged **do**
3: E-step: $\tilde{\mathbf{T}}^{(t+1)}(\mathbf{y}, \mathbf{w}) \leftarrow \mathrm{E}_{\mathbf{w}} \left[\mathbf{T}(\mathbf{w}, \mathbf{y}) | \mathbf{y}, \boldsymbol{\theta}^{(t)} \right]$
4: M-step: $\boldsymbol{\theta}^{(t+1)} \leftarrow$ solution to $\tilde{\mathbf{T}}^{(t+1)}(\mathbf{y}, \mathbf{w}) = \mathrm{E}_{\mathbf{y},\mathbf{w}} \left[\mathbf{T}(\mathbf{y}, \mathbf{w}) | \boldsymbol{\theta} \right]$
5: $t \leftarrow t + 1$
6: **end while**

To see how Algorithm 4 motivates the EM algorithm, consider the following argument. Recall that for the EM algorithm, the function $Q_t(\boldsymbol{\eta}) = \mathrm{E}_{\mathbf{w}}[\log p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) | \mathbf{y}, \boldsymbol{\eta}^{(t)}]$ is maximised at each

iteration t. For exponential families of the form (S2.4), the Q_t function turns out to be

$$Q_t(\boldsymbol{\eta}) = \mathrm{E}_{\mathbf{w}}\left[\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z})\rangle | \mathbf{y}, \boldsymbol{\eta}^{(t)}\right] - A^*(\boldsymbol{\eta}) + \log B(\mathbf{z}),$$

and this is maximised at the value of $\boldsymbol{\eta}$ satisfying

$$\frac{\partial}{\partial \boldsymbol{\eta}} Q_t(\boldsymbol{\eta}) = \mathrm{E}_{\mathbf{w}}\left[\mathbf{T}(\mathbf{y}, \mathbf{w}) | \mathbf{y}, \boldsymbol{\eta}^{(t)}\right] - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = 0,$$

a similar condition to (S2.6) when obtaining ML estimate of $\boldsymbol{\eta}$. Thus, Q_t is maximised by the solution to line 4 in Algorithm 4.

S2.3 Bayesian EM algorithm

A simple modification of the EM algorithm can be done to obtain maximum a posteriori estimates, or maximum penalised likelihood estimates. Under a Bayesian framework, a prior is assigned on the model parameters, $\theta \sim p(\theta)$. Recall that the MAP estimate is obtained as the maximiser of the log-density $\log p(\mathbf{y}|\theta) + \log p(\theta)$.

The EM algorithm works as before, but replaces the E-step with

$$\mathrm{E}_{\mathbf{w}}\left[\log p(\mathbf{w}, \mathbf{y}|\theta) + \log p(\theta) | \mathbf{y}, \theta^{(t)}\right] = Q(\theta|\theta^{(t)}) + \log p(\theta) \tag{S2.7}$$

since $\log p(\theta)$ has no terms involving the latent variables $\mathbf{w}$. The M-step now maximises (S2.7) with respect to θ, which includes the log prior density (or a penalty term). It would seem that the regular EM algorithm maximises (S2.7) such that $p(\theta) \propto$ const. is a diffuse prior for θ. Beal and Ghahramani (2003) discuss a more Bayesian extension of EM, in which the output of the so-called *variational Bayes EM* algorithm are (approximate) posterior distributions of the parameters, rather than MAP estimates discussed here.

Supplementary S3

Variational inference

Consider a statistical model parameterised by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ for which we have observations $\mathbf{y} := \{y_1, \ldots, y_n\}$, but also some latent variables $\mathbf{w}$. Typically, in such models, there is a want to to evaluate the integral

$$I = \int p(\mathbf{y}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}, \tag{S3.1}$$

Marginalising out the latent variables in (S3.1) is usually a precursor to obtaining a log-likelihood function to be maximised in a frequentist setting, whereby there is an implicit dependence on the model parameters in the evaluation of I. In Bayesian analysis, priors are specified on the model parameters $\theta \sim p(\theta)$. By concatenating the latent variables and model parameters to form $\mathbf{w}$, the I corresponds to the marginal density for $\mathbf{y}$, on which the posterior depends.

In many instances, for one reason or another, evaluation of (S3.1) or is difficult, in which case inference is halted unless a way of overcoming the intractability is found. In this chapter, we discus *variational inference* (VI) as a means of approximating the integral. The literature on variational inference is typically presented in a Bayesian light (Bishop, 2006; Blei et al., 2017; Jordan et al., 1999), and as such, it is commonly known as *variational Bayes* method. The main attraction from a Bayesian point of view is that it provides a deterministic way of obtaining (approximate) posteriors, i.e. it does not involve sampling from posteriors.

Variational inference can be used in conjunction with an EM algorithm, in which the E-step is replaced with a variational E-step. This *variational EM algorithm* is used for maximum likelihood learning, but can modified to obtain maximum a posteriori estimates. In the works of (Beal, 2003; Beal and Ghahramani, 2003), the authors realised that the EM algorithm can be extended easily to obtain posterior densities of the latent variables and parameters if the statistical model is conjugate exponential family. They refer to this as the *variational Bayes EM algorithm*, but in fact this is really just variational inference in which the algorithm resembles an EM algorithm with clear E- and M-steps.

We first briefly introduce variational methods for approximating the intractable integral, and this is usually considered a fully Bayesian treatment of the model. We then describe variational EM, and provide a comparison of the two methods.

S3.1 A brief introduction to variational inference

The crux of variational inference is this: find a suitably close distribution function $q(\mathbf{w})$ that approximates the true posterior $p(\mathbf{w}|\mathbf{y})$, where closeness here is defined in the Kullback-Leibler divergence sense,

$$\mathrm{D}_{\mathrm{KL}}(q\|p) = \int \log \frac{q(\mathbf{w})}{p(\mathbf{w}|\mathbf{y})} q(\mathbf{w})\,\mathrm{d}\mathbf{w}.$$

Posterior inference is then conducted using $q(\mathbf{w})$ in lieu of $p(\mathbf{w}|\mathbf{y})$. Advantages of this method are that 1) it is fast to implement computationally (compared to MCMC); 2) convergence is assessed simply by monitoring a single convergence criterion; and 3) it works well in practice, as attested to by the many studies implementing VI.

Briefly, we present the motivation behind variational inference and the minimisation of the KL divergence. Denote by $q(\cdot)$ some density function of $\mathbf{w}$. One may show that log marginal density (the log of the intractable integral (S2.1)) holds the following bound:

$$\begin{aligned}
\log p(y) &= \log p(\mathbf{y}, \mathbf{w}) - \log p(\mathbf{w}|\mathbf{y}) \quad \text{(Bayes' theorem)} \\
&= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} - \log \frac{p(\mathbf{w}|\mathbf{y})}{q(\mathbf{w})} \right\} q(\mathbf{w})\,\mathrm{d}\mathbf{w} \quad \text{(expectations both sides)} \\
&= \mathcal{L}(q) + \mathrm{D}_{\mathrm{KL}}(q\|p) \\
&\geq \mathcal{L}(q)
\end{aligned} \tag{S3.2}$$

since the KL divergence is a non-negative quantity. The functional $\mathcal{L}(q)$ given by

$$\begin{aligned}
\mathcal{L}(q) &= \int \log \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} q(\mathbf{w})\,\mathrm{d}\mathbf{w} \\
&= \mathrm{E}_{\mathbf{w}\sim q}[\log p(\mathbf{y}, \mathbf{w})] + H(q),
\end{aligned} \tag{S3.3}$$

where H is the entropy functional, is known as the *evidence lower bound* (ELBO). Evidently, the closer q is to the true p, the better, and this is achieved by maximising $\mathcal{L}$, or equivalently, minimising the KL divergence from p to q. Note that the bound (S3.2) achieves equality if and only if $q(\mathbf{w}) \equiv p(\mathbf{w}|\mathbf{y})$, but of course the true form of the posterior is unknown to us—see Section S3.2 for a discussion. Maximising $\mathcal{L}(q)$ or minimising $\mathrm{D}_{\mathrm{KL}}(q\|p)$ with respect to the density q is a problem of calculus of variations, which incidentally, is where variational inference takes its name. The astute reader will realise that $\mathrm{D}_{\mathrm{KL}}(q||p)$ is impossible to compute, since one does not know the true distribution $p(\mathbf{w}|\mathbf{y})$. Efforts are concentrated on maximising the ELBO instead.

Maximising $\mathcal{L}$ over all possible density functions q is not possible without considering certain constraints. Two such constraints are described. The first, is to make a distributional assumption regarding q, for which it is parameterised by ν. For instance, we might choose the closest normal distribution to the posterior $p(\mathbf{w}|\mathbf{y})$ in terms of KL divergence. In this case, the task is to find optimal mean and variance parameters of a normal distribution.

[1] Reproduced from the talk by David Blei entitled 'Variational Inference: Foundations and Innovations', 2017. URL: https://simons.berkeley.edu/talks/david-blei-2017-5-1.

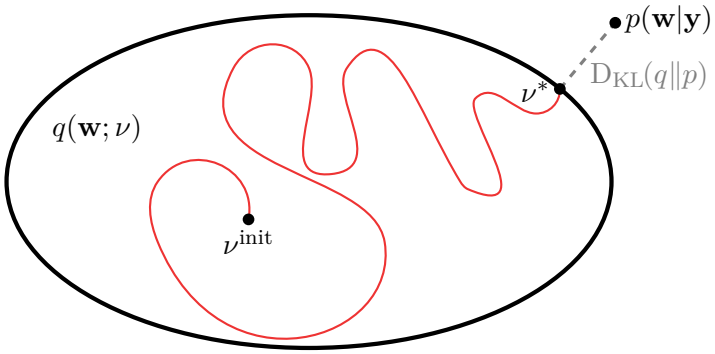


Figure S1: Schematic view of variational inference[1]. The aim is to find the closest distribution q (parameterised by a variational parameter ν) to p in terms of KL divergence within the set of variational distributions, represented by the ellipse.

The second type of constraint, and the one considered in this thesis, is simply an assumption that the approximate posterior q factorises into M disjoint factors. Partition $\mathbf{w}$ into M disjoint groups $\mathbf{w} = (w_{[1]}, \ldots, w_{[M]})$. Note that each factor $w_{[k]}$ may be multidimensional. Then, the structure

$$q(\mathbf{w}) = \prod_{k=1}^{M} q_k(w_{[k]})$$

for q is considered. This factorised form of variational inference is known in the statistical physics literature as the *mean-field theory* (Itzykson and Drouffe, 1991).

Remark S3.1. The choice of factorisation is completely arbitrary, although forcing a factorisation also induces independence between the factors in the posterior, and this may or may not be suitable for the problem at hand. Landing the correct choice of factorisation is rather experimental, as the aim is to balance tractability and model misspeficication. In a model with both latent variables and random parameters (in a Bayesian setting), then a good starting point would be to factorise the latent variables and parameters.

Let us denote the distributions which minimise the Kullback-Leibler divergence (maximise the variational lower bound) by the use of tildes. The impact of the mean-field factorisation on the ELBO is inspected:

$$\begin{aligned}\mathcal{L}(q) &= \int \cdots \int \log \frac{p(\mathbf{y}, \mathbf{w})}{\prod_{k=1}^{M} q_k(\mathbf{w})} \prod_{k=1}^{m} \left\{ q_k(w_{[k]}) \, \mathrm{d}w_{[k]} \right\} \\ &= \int \cdots \int \left(\log p(\mathbf{y}, \mathbf{w}) - \sum_{k=1}^{M} \log q_k(\mathbf{w}) \right) \prod_{k=1}^{M} \left\{ q_k(w_{[k]}) \, \mathrm{d}w_{[k]} \right\}\end{aligned}$$

and rearranging slightly for terms involving the j'th component only, we get

$$\begin{aligned}
\mathcal{L}(q) &= \int \cdots \int \big(\log p(\mathbf{y}, \mathbf{w}) - \log q_j(w_{[j]}) + \text{const.}\big)\, q_j(w_{[j]})\, \mathrm{d}w_{[j]} \prod_{k \neq j} \big\{q_k(w_{[k]})\, \mathrm{d}w_{[k]}\big\} \\
&= \int \overbrace{\left(\int \cdots \int \log p(\mathbf{y}, \mathbf{w}) \prod_{k \neq j} \big\{q_k(w_{[k]})\, \mathrm{d}w_{[k]}\big\} \right)}^{\log \tilde{p}(\mathbf{y}, w_{[j]}) + \text{const.}} q_j(w_{[j]})\, \mathrm{d}w_{[j]} \\
&\quad - \int \log q_j(w_{[j]}) q_j(w_{[j]})\, \mathrm{d}w_{[j]} + \text{const.} \\
&= -\mathrm{D}_{\mathrm{KL}}(q_{[j]} \| \tilde{p}) + \text{const.}
\end{aligned}$$

The task of maximising $\mathcal{L}$ is then equivalent to maximising $-\mathrm{D}_{\mathrm{KL}}(q_{[j]} \| \tilde{p})$, where $\tilde{p}$ is defined in the overbrace of the second line in the equation above. Thus, for each $w_{[k]}$, $k = 1, \ldots, M$, $\tilde{q}_k$ satisfies

$$\log \tilde{q}_k(w_{[k]}) = \mathrm{E}_{-k}[\log p(\mathbf{y}, \mathbf{w})] + \text{const.} \tag{S3.4}$$

where expectation of the joint log density of $\mathbf{y}$ and $\mathbf{w}$ is taken with respect to all of the unknowns $\mathbf{w}$, except the one currently in consideration $w_{[k]}$, under their respective $\tilde{q}_k$ densities. For further details, refer to Bishop (2006, eq. 10.9, p. 466).

In practice, rather than an explicit calculation of the normalising constant, one simply needs to inspect (S3.4) to recognise it as a known log-density function, which is the case when exponential family distributions are considered. That is, suppose that each complete conditional $p(w_{[k]} | \mathbf{w}_{-k}, \mathbf{y})$, where $\mathbf{w}_{-k} = \{w_{[i]} | i \neq k\}$, follows an exponential family distribution

$$p(w_{[k]} | \mathbf{w}_{-k}, \mathbf{y}) = B(w_{[k]}) \exp\big(\langle \zeta_k(\mathbf{w}_{-k}, \mathbf{y}), w_{[k]} \rangle - A(\zeta_k)\big).$$

Then, from (S3.4),

$$\begin{aligned}
\tilde{q}(w_{[k]}) &\propto \exp\big(\mathrm{E}_{-k} \log p(w_{[k]} | \mathbf{w}_{-k}, \mathbf{y})\big) \\
&= \exp\Big(\log B(w_{[k]}) + \mathrm{E}\langle \zeta_k(\mathbf{w}_{-k}, \mathbf{y}), w_{[k]} \rangle - \mathrm{E}[A(\zeta_k)]\Big) \\
&\propto B(w_{[k]}) \exp \mathrm{E}\langle \zeta_\xi(\mathbf{w}_{-k}, \mathbf{y}), w_{[k]} \rangle
\end{aligned}$$

is also in the same exponential family. In situations where there is no closed form expression for $\tilde{q}$, then one resorts to sampling methods such as a Metropolis random walk to estimate quantities of interest. This stochastic step within a deterministic algorithm has been explored before in the context of a Monte Carlo EM algorithm—see Meng and Van Dyk (1997, §4, pp. 537–538) and references therein.

One notices that the optimal mean-field variational densities for each component are coupled with one another, in the sense that the distribution $\tilde{q}_k$ depends on the moments of the rest of the components $\mathbf{w}_{-k}$. For very simple problems, an exact solution for each $\tilde{q}_k$ can be found, but usually, the way around this is to employ an iterative procedure. The *coordinate ascent mean-field variational inference* (CAVI) algorithm cycles through each of the distributions in turn, updating them in sequence starting with arbitrary distributions as initial values.

Algorithm 5 The CAVI algorithm

initialise Variational factors $q_k(w_{[k]})$
while ELBO $\mathcal{L}(q)$ not converged **do**
  for $k = 1, \ldots, M$ **do**
    $\tilde{q}_k(w_{[k]}) \leftarrow \text{const.} \times \exp\left(\mathrm{E}_{-k} \log p(\mathbf{y}, \mathbf{w})\right)$ ▷ from (S3.4)
  end for
  $\mathcal{L}(q) \leftarrow \mathrm{E}_{\mathbf{w} \sim \prod_k \tilde{q}_k} \log p(\mathbf{y}, \mathbf{w}) + \sum_{k=1}^{m} H\left[q_k(w_{[k]})\right]$ ▷ Update ELBO
end while
return $\tilde{q}(\mathbf{w}) = \prod_{k=1}^{M} \tilde{q}_j(w_{[k]})$

Each iteration of the CAVI brings about an improvement in the ELBO (hence the name coordinate ascent). The algorithm terminates when there is no more significant improvement in the ELBO, indicating a convergence of the CAVI. Blei et al. (2017) notes that the ELBO is typically a non-convex function, in which case convergence may be to (one of possibly many) local optima. A simple solution would be to restart the CAVI at multiple initial values, and the solution giving the highest ELBO is the distribution that is closest to the true posterior.

S3.2 Variational EM algorithm

Consider again the latent variable setup described in Supplementary Chapter S2, in which the goal is to maximise the (marginal) log-likelihood of the parameters θ of the model, after integrating out the latent variables, as given by (S2.1). We will see how the EM algorithm relates to minimising the KL divergence between a density $q(\mathbf{w})$ and the posterior of $\mathbf{w}$, and connect this idea to variational methods.

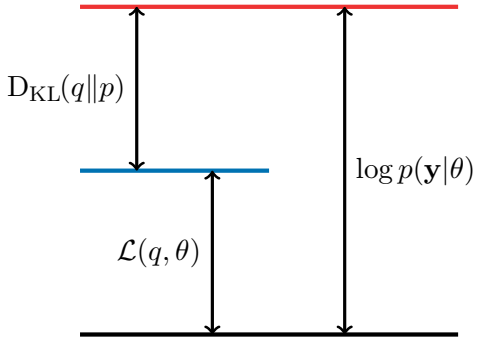


Figure S2: Illustration[2] of the decomposition of the log-likelihood into $\mathcal{L}(q, \theta)$ and $\mathrm{D}_{\mathrm{KL}}(q\|p)$. The quantity $\mathcal{L}(q, \theta)$ is a lower bound for the log-likelihood.

[2]Reproduced from Bishop (2006, Figure 9.11).

As we did in deriving (S3.2), we decompose the (marginal) log-likelihood as

$$\begin{aligned}\log p(\mathbf{y}|\theta) &= \log p(\mathbf{y}, \mathbf{w}|\theta) - \log p(\mathbf{w}|\mathbf{y}, \theta) \\ &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{w}|\theta)}{q(\mathbf{w})} - \log \frac{p(\mathbf{w}|\mathbf{y}, \theta)}{q(\mathbf{w})} \right\} q(\mathbf{w}) \, \mathrm{d}\mathbf{w} \\ &= \underbrace{\mathrm{E}_{\mathbf{w}\sim q}\left[\log \frac{p(\mathbf{y}, \mathbf{w}|\theta)}{q(\mathbf{w})} \right]}_{\mathcal{L}(q,\theta)} - \underbrace{\mathrm{E}_{\mathbf{w}\sim q}\left[\log \frac{p(\mathbf{w}|\mathbf{y}, \theta)}{q(\mathbf{w})} \right]}_{-\mathrm{D}_{\mathrm{KL}}(q\|p)},\end{aligned}$$

where $q(\mathbf{w})$ is any density function over the latent variables. This decomposition is shown in Figure S2. The interest is then to have a density function $q(\mathbf{w})$ which is as close as possible to the true posterior density $p(\mathbf{y}|\mathbf{w}, \theta)$ in the KL divergence sense. Since the KL divergence is non-negative, minimising $\mathrm{D}_{\mathrm{KL}}(q\|p)$ is equivalent to maximising $\mathcal{L}(q, \theta)$.

As a remark, the above line of thought should be familiar as it is the exact same one made for variational inference. The twist here is that we will peruse a distribution which tightens the lower bound $\mathcal{L}(q, \theta)$ to the marginal log-likelihood, and this happens when $\mathrm{D}_{\mathrm{KL}}(q\|p)$ is exactly zero, and this in turn happens when q is exactly the true posterior density. That is, for some parameter value, $\theta = \theta^{(t)}$ say, the solution to

$$\arg\max_{q} \mathcal{L}(q, \theta^{(t)}) \tag{S3.5}$$

is $q^{(t+1)}(\mathbf{w}) = p(\mathbf{w}|\mathbf{y}, \theta^{(t)})$, because

$$\mathrm{D}_{\mathrm{KL}}(q\|p) = \mathrm{E}\left[\log \frac{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})} \right] = 0.$$

At this stage, we have the equality

$$\log p(\mathbf{y}|\theta) = \mathcal{L}(q^{(t+1)}, \theta) \tag{S3.6}$$

$$= \mathrm{E}_{\mathbf{w}\sim q^{(t+1)}}\left[\log \frac{p(\mathbf{y}, \mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})} \right] \tag{S3.7}$$

$$= \underbrace{\mathrm{E}_{\mathbf{w}\sim q^{(t+1)}}\left[\log p(\mathbf{y}, \mathbf{w}|\theta) \right]}_{Q(\theta|\theta^{(t)})} - \underbrace{\mathrm{E}_{\mathbf{w}\sim q^{(t+1)}}\left[\log p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \right]}_{-H(q^{(t+1)})}, \tag{S3.8}$$

The term on the left is recognised as the Q function of the E-step

$$Q(\theta) = Q(\theta|\theta^{(t)}) = \mathrm{E}_{\mathbf{w}}\left[\log p(\mathbf{y}, \mathbf{w}|\theta) \,\middle|\, \mathbf{y}, \theta^{(t)} \right],$$

while the term on the left is an entropy term which does not depend on θ. Thus, minimising the KL divergence, or maximising the lower bound $\mathcal{L}$ with respect to q, corresponds to the E-step in the EM algorithm.

Furthermore, since equality between the log-likelihood and the lower bound is achieved after the E-step, increasing $\mathcal{L}(q^{(t+1)}, \theta)$ with respect to θ is sure to bring about an increase in the

log-likelihood. That is, for any θ, we find that

$$\begin{aligned}\log p(\mathbf{y}|\theta) - \log p(\mathbf{y}|\theta^{(t)}) &= Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + \Delta\,\text{entropy} \\ &\geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}).\end{aligned}$$

because entropy differences are positive by Gibbs' inequality. We see that maximising Q with respect to θ (the M-step) brings about an improvement to the log-likelihood value.

To summarise, given initial values $q^{(0)}$ for the distribution and $\theta^{(0)}$ for the parameters, the EM algorithm is seen as iterating between

- **E-step**: $q^{(t+1)} \leftarrow \arg\max_q \mathcal{L}(q, \theta^{(t)})$, i.e., maximise $\mathcal{L}(q, \theta)$ with respect to q, keeping θ fixed. This is equivalent to minimising the KL divergence $\mathrm{D}_{\mathrm{KL}}(q\|p)$.
- **M-step**. $\theta^{(t+1)} \leftarrow \arg\max_\theta \mathcal{L}(q^{(t+1)}, \theta)$, i.e., maximise $\mathcal{L}(q, \theta)$ with respect to θ, keeping $q(\mathbf{w})$ fixed.

When the true posterior distribution $p(\mathbf{w}|\mathbf{y})$ is not tractable, then the E-step becomes intractable as well. By constraining the maximisation in the E-step to consider q belonging to a family of tractable densities, the E-step yields a variational approximation $\tilde{q}$ to the true posterior. In Section S3.1, we saw that constraining q to be of a factorised form, then $\tilde{q}$ is a mean-field density. After a variational E-step, the M-step proceeds as normal. This form of the EM is known as *variational EM algorithm* (VEM) (Beal, 2003). The variational EM algorithm can also be modified to obtain MAP estimates by including the log prior density to the maximisation objective in the M-step.

Due to an approximation to the true posterior being used in the E-step, there is no guarantee that the log-likelihood value will increase at each iteration. This is seen pictorially in Section S3.2: since the bound on the log-likelihood is not tight, increasing this bound will not necessarily cause an increase in log-likelihood value (Scenario C), and even if it did, it may not give as much an increase as it would under the true posterior density (Scenario B). Scenario A depicts an ideal case whereby the increase in log-likelihood is as much as it would be if the true posterior density was used.

On a practical note, if the posterior density is intractable, then so is the marginal likelihood, which means that we're unable to determine convergence of the EM using the log-likelihood. Instead, the lower bound $\mathcal{L}(q, \theta)$ should be used, which monotonically increases to a local optima (as in the CAVI algorithm).

S3.3 Comparing variational inference and variational EM

Variational inference is a fully Bayesian treatment of the model, for which the goal is to obtain approximate posterior densities for all latent variables and parameters. Variational EM algorithm on the other hand has the objective of obtaining ML or MAP estimates of the parameters using an EM algorithm in which the E-step is replaced with a variational E-step. In some cases, the CAVI algorithm can resemble an EM algorithm, especially when there is a distinction between

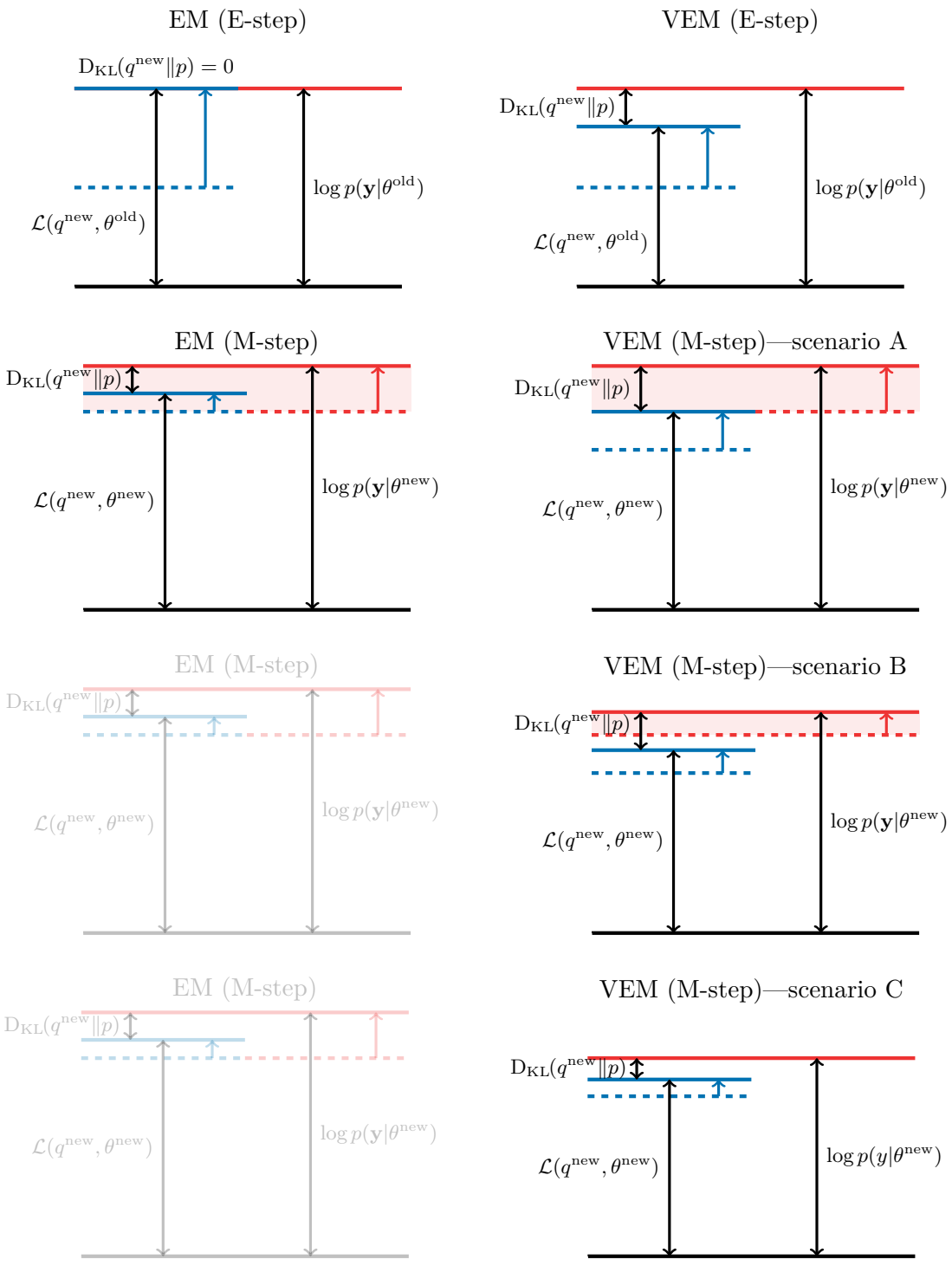


Figure S3: Illustration of EM vs VEM. Whereas the EM guarantees an increase in log-likelihood value (red shaded region), the VEM does not.

latent variables and parameters, and a conjugate exponential family model is involved (Blei et al., 2017).

Variational inference can yield exactly similar point estimates as variational EM if the approximate posterior is symmetric, e.g. a normal distribution. Under a normal posterior, its mean is used as a point estimate, which coincides with the mode, which is a MAP estimate, or in the case of diffuse priors, a ML estimate. However, since the output of variational inference are posterior densities instead of a single point estimate, one is able to obtain posterior standard deviations or credibility intervals about the parameters, something which is not so straightforward under a variational EM or even EM framework.

Derivation of the CAVI algorithm and ELBO for specific models is certainly more tedious than the derivation of the variational EM algorithm. Often, quantities that are required in the derivation include $\mathrm{E}[\theta]$, $\mathrm{E}[\theta^2]$, $\mathrm{E}[\theta^{-1}]$, $\mathrm{E}[\log\theta]$ or any other moment of some function of θ, where expectations are taken under the approximating q posterior density. For certain distributions $q(\theta)$ these quantities can be awkward to compute, and may need approximating themselves.

The computational time and storage requirements of variational methods is virtually the same as EM algorithm (Beal, 2003; Blei et al., 2017). Consider the mean-field variational approximation. In variational inference or variational EM, the updating step for the factors involve

$$\tilde{q}_k^{(t+1)}(w_{[k]}) \leftarrow \text{const.} \times \exp\left(\mathrm{E}_{\mathbf{w}_{-k}\sim\tilde{q}^{(t)}}[\log p(\mathbf{y},\mathbf{w})]\right), \tag{S3.9}$$

for each of the factors of the approximate posterior $q(\mathbf{w}) = \prod_{k=1}^{M} q_k(w_{[k]})$. In the EM algorithm E-step, one obtains the Q function

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{\mathbf{w}}[\log p(\mathbf{y},\mathbf{w})|\mathbf{y},\theta^{(t)}]. \tag{S3.10}$$

We can see that in both equations (S3.9) and (S3.10), there is a need to compute the expectation of the joint log density, but the difference between the variational inference and EM or variational EM lies in the M-step. In variational inference one seeks a distribution, while in EM or variational EM one seeks a point estimate (posterior mode) of this distribution.

Table S1: Comparison between variational inference and variational EM.

Variational inference	Variational EM
GOAL: Posterior densities for $(\mathbf{w}, \theta)$	**GOAL**: ML/MAP estimates for θ
Variational approximation for latent variables and parameters $q(\mathbf{w}, \theta) \approx p(\mathbf{w}, \theta \vert \mathbf{y})$	Variational approximation for latent variables only $q(\mathbf{w}) \approx p(\mathbf{w} \vert \mathbf{y})$
Priors required on θ	Priors not necessary for θ
Derivation can be tedious	Derivation less tedious
Inference on θ through posterior density $q(\theta)$	Asymptotic distribution of θ not well studied; standard errors for θ not easily obtained
Suited to conjugate exponential family models: posteriors will be easily recognisable	Suited to conjugate exponential family models, but not necessary

Supplementary S4

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo had its beginnings in statistical physics, with the 1987 paper by Duane et al., using what they called 'Hybrid Monte Carlo' in lattice models of quantum theory. Their work merged the approaches of molecular dynamics and Markov chain Monte Carlo methods. As an interesting side note, their method abbreviates also to 'HMC', but throughout the statistical literature, it is more commonly referred to by its more descriptive name Hamiltonian Monte Carlo. Incidentally, the use of HMC started with applications to neural networks as early as 1996 (see Neal, 2011 for an excellent review of the subject matter). It was not until 2011 when active development of the method, and in particular, software for for statistical applications began. The Stan initiative (Carpenter et al., 2017) began in response to difficulties faced when performing full Bayesian inference on multilevel generalised linear models. These difficulties mainly involved poor efficiency in usual MCMC samplers, particularly due to high autocorrelations in the posterior chains, which meant that many chains and many iterations were required to get an adequate sample. It was a case of exhausting all possible algorithmic remedies for existing samplers (Gibbs samplers, Metropolis samplers, etc.), and realising that fundamentally not much improvement can be had unless a novel sampling technique was discovered.

The basic idea behind HMC is to use Hamiltonian dynamics to propose new states in the posterior sampling, rather than relying on 'random walks'. If one were to understand and use the geometry of the posterior density to one's benefit, then it should be possible to generate new proposal states with high probabilities of acceptance and move far away from the current state. Hamiltonian dynamics, like classical Newtonian mechanics, provides a framework for modelling the motion of a body in space across time t. Additionally, Hamiltonian dynamics concatenates the position vector x with its momentum z, and the motion of x in d-dimensional space is then described through Hamilton's equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial H}{\partial z} \quad \text{and} \quad \frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{\partial H}{\partial x}, \tag{S4.1}$$

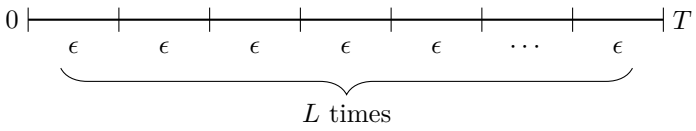
where $H = H(x, z)$ is called the Hamiltonian of the system. The Hamiltonian is an operator which encapsulates the total energy of the system. In a closed system, one can express the sum of operators corresponding to the kinetic energy $K(p)$ and the potential energy $U(z)$ of the system

$$H(x, z) = K(z) + U(x). \tag{S4.2}$$

Substituing (S4.2) into (S4.1), we get the system of partial differential equations (PDEs)

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial}{\partial z} K(z) \quad \text{and} \quad \frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{\partial}{\partial x} U(x). \tag{S4.3}$$

To describe the evolution of $\big(x(t), z(t)\big)$ from time t to $t+T$, it is necessary to discretise time, and split $T = L\epsilon$. The quantity L is known as the number of *leapfrogs*, and ϵ the *step size*.



The system of PDEs is solved using Euler's method, or the more commonly used leapfrog integration, which is a three-step process:

1. **Half-step momentum.** $z(t + \epsilon/2) = z(t) - \frac{\epsilon}{2} \frac{\partial}{\partial x} U\big(x(t)\big)$
2. **Full-step position.** $x(t + \epsilon) = x(t) + \epsilon \frac{\partial}{\partial z} K\big(z(t + \epsilon/2)\big)$
3. **Half-step momentum.** $z(t + \epsilon) = z(t + \epsilon/2) = z(t) - \frac{\epsilon}{2} \frac{\partial}{\partial x} U\big(x(t)\big)$

in which steps 1–3 are repeated L times.

Having knowing the formula for how particles move in space, we can use this information to treat random points drawn from some probability density as 'particles'. Randomness of position and momentum are prescribed through probability densities on each. Given some energy function $E(\theta)$ over states θ, the *canonical distribution* of the states θ (otherwise known as the *canonical ensemble*) is given by the probability density function

$$p(\theta) \propto \exp\left(-\frac{E(\theta)}{k\tau}\right),$$

where k is Boltzmann's constant, τ is the absolute temperature of the system. The Hamiltonian is one such energy function over states (x, z). By replacing $E(\theta)$ by (S4.2) in the pdf above, we realise that the distribution for x and z are independent. The system can be manipulated such that $k\tau = 1$—in any case, these are constants which can be absorbed into one of the terms in the pdf anyway.

Using a *quadratic kinetic energy* function $K(z) = z^\top M^{-1} z/2$[1], we find that the probability density function for z is

$$p(z) \propto \exp\left(-\frac{1}{2} z^\top M^{-1} z\right),$$

implying $z \sim \mathrm{N}_d(0, M)$. Here, $M = \mathrm{diag}(m_1, \ldots, m_d)$ is called the *mass matrix*, which obviously serves as the variance for the randomly distributed z. As for the potential energy, choose a function such that $U(x) = -\log p(x)$, implying $p(x) \propto \exp\big(-U(x)\big)$. Here, $p(x)$ represents the target density from which we wish to sample, for instance, a posterior density of interest. Thus, to sample variables x from $p(x)$, one artificially introduces momentum variables z and

sample jointly instead from $p(x, z) = p(z)p(z)$, and discarding z thereafter. The HMC algorithm is summarised in Algorithm 6.

Algorithm 6 Hamiltonian Monte Carlo

1: **initialise** $x^{(0)}$, $z^{(0)}$ and choose values for L, ϵ and M
2: **for** $t = 1, \ldots, T$ **do**
3: Draw $z \sim \mathrm{N}_d(0, M)$ ▷ Perturb momentum
4: Move $(x^{(t)}, z^{(t)}) \mapsto (x^*, z^*)$ using Hamiltonian dynamics ▷ Proposal state
5: Accept/reject proposal state, i.e. ▷ Metropolis update

$$(x^{(t+1)}, z^{(t+1)}) \leftarrow \begin{cases} (x^*, z^*) & \text{w.p. } \min(1, A) \\ (x^{(t)}, z^{(t)}) & \text{otherwise} \end{cases}$$

where

$$A = \frac{p(x^*, z^*)}{(x^{(t)}, z^{(t)})} = \exp\left(H(x, z) - H(x^{(t)}, z^{(t)})\right)$$

6: **end for**
7: **return** Samples $\left\{x^{(1)}, \ldots, x^{(T)}\right\}$

HMC is often times superior to standard Gibbs sampling, for a variety of reasons. For one, conjugacy does not play any role in the efficiency of the HMC sampler, thus freeing the modeller to choose more appropriate and more intuitive prior densities for the parameters of the model. For another, the HMC sampler is designed to incite little autocorrelations between samples, and thus increasing efficiency.

Several drawbacks do exist with the HMC sampler. Firstly, it is impossible to directly sample from discrete distributions $p(x)$. More concretely, HMC requires that the domain of $p(x)$ is continuous and that $\partial \log p(x)/\partial x$ is inexpensive to compute. To work around this, one must reformulate the model by marginalising out the discrete variables, and obtain them back later by separately sampling from their posteriors. Alternatively, a Gibbs sampler specifically for the discrete variables could be augmented with the HMC sampler. The other drawback of HMC is that there are many tuning parameters (leapfrog L, step-size ϵ, mass matrix M, etc.) that is not immediately easy to perfect, at least not to the novice user.

The implementation of HMC by the programming language Stan, which interfaces many other programming languages including R, Python, MATLAB, Julia, Stata and Mathematica, is a huge step forward in computational Bayesian analysis. Stan takes the liberty of performing all the tuning necessary, and the practitioner is left with simply specifying the model. A vast library of differentiable probability functions are available, with the ability to bring your own code as well. Development is very active and many improvements and optimisations have been made since its inception.

[1]Thinking back to elementary mechanics, this is the familiar $\frac{1}{2}mv^2$ formula for kinetic energy and substituting in the identity $z = mv$, where m is the mass of the object, and v is its velocity.

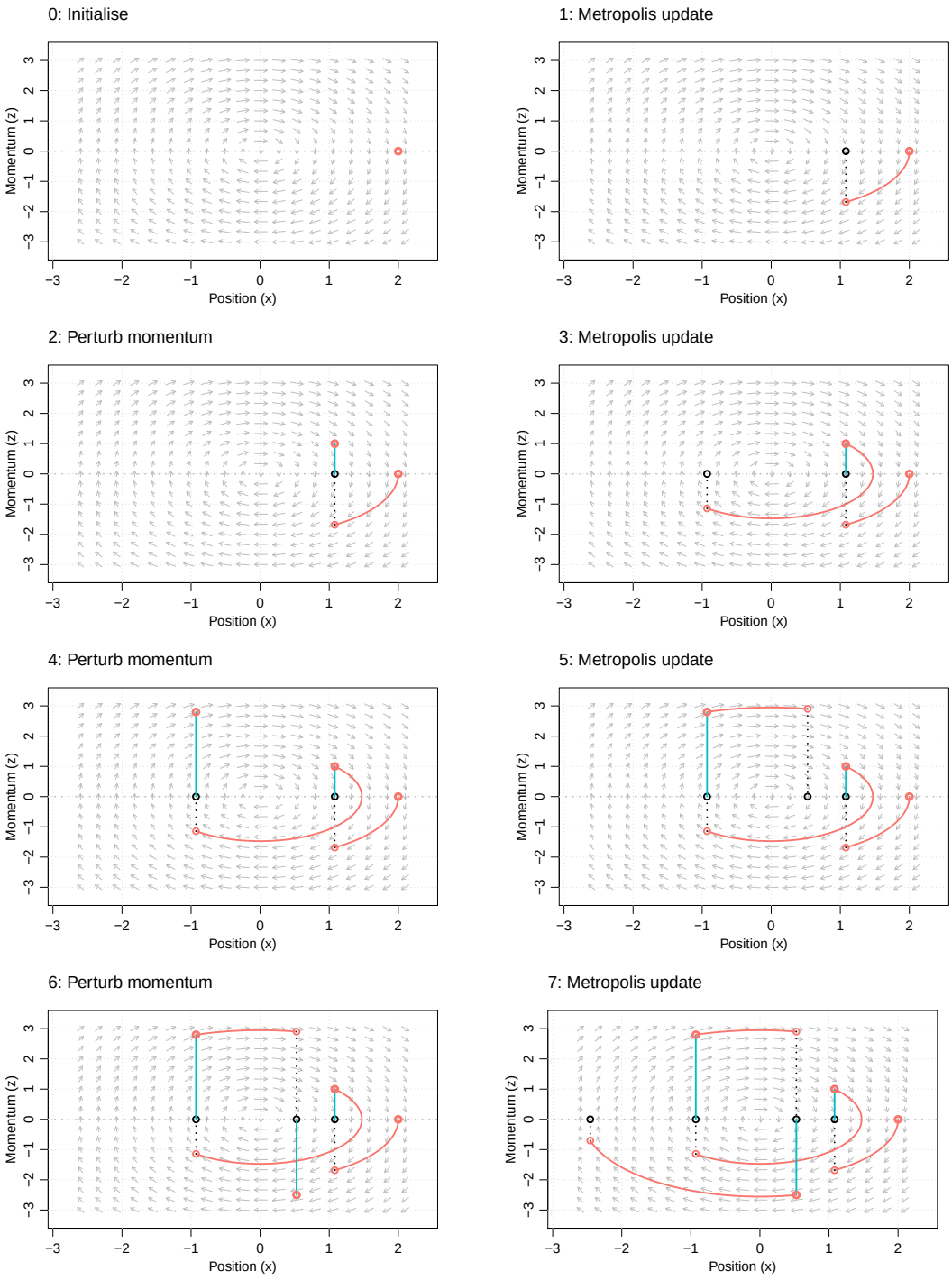


Figure S1: A phase diagram schematic of HMC sampling in one-dimension. At step 0, initialise values for momentum and position. At step 1, simulate movement using Hamiltonian dynamics, accept position and discard momentum. At step 2, perturb momentum using a normal density, then repeat.

Bibliography

Agresti, Alan and Jonathan Hartzel (2000). "Tutorial in biostatistics: Strategies comparing treatment on binary response with multi-centre data". In: *Statistics in Medicine* 19, pp. 1115–1139.

Akaike, Hirotogu (1973). "Information theory and an extension of the maximum likelihood principle". In: *2nd International Symposium on Information Theory*. Akadémiai Kiadó, pp. 267–281.

Albert, James H. and Siddhartha Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data". In: *Journal of the American Statistical Association* 88.422, pp. 669–679. DOI: 10.2307/2290350.

Alpay, Daniel (1991). "Some Remarks on Reproducing Kernel Krein Spaces". In: *The Rocky Mountain Journal of Mathematics* 21.4, pp. 1189–1205. DOI: 10.1216/rmjm/1181072903.

Balakrishnan, Alampallam V. (1981). *Applied Functional Analysis*. 2nd ed. Springer-Verlag. ISBN: 978-1-4612-5867-4. DOI: 10.1007/978-1-4612-5865-0.

Barbieri, Maria Maddalena and James O. Berger (2004). "Optimal predictive model selection". In: *Annals of Statistics* 32.3, pp. 870–897. DOI: 10.1214/009053604000000238.

Beal, Matthew James (2003). "Variational algorithms for approximate Bayesian inference". PhD thesis. Gatsby Computational Neuroscience Unit, University College London.

Beal, Matthew James and Zoubin Ghahramani (2003). "The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures". In: *Bayesian Statistics 7*. Proceedings of the Seventh Valencia International Meeting. Ed. by José M. Bernardo, A. Philip Dawid, James O. Berger, Mike West, David Heckerman, M. J. Bayarri, and Adrian F. M. Smith. Oxford University Press, pp. 453–464.

Berger, James O. (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. New York: Springer-Verlag. ISBN: 978-0-387-96098-2. DOI: 10.1007/978-1-4757-4286-2.

Bergsma, Wicher (2018). *Regression and classification with I-priors*. Manuscript in submission. arXiv: 1707.00274 [math.ST].

Berlinet, Alain and Christine Thomas-Agnan (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer. ISBN: 978-1-4613-4792-7. DOI: 10.1007/978-1-4419-9096-9.

Bickel, Peter, David Choi, Xiangyu Chang, and Hai Zhang (2013). "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels". In: *The Annals of Statistics* 41.4, pp. 1922–1943. DOI: 10.1214/13-AOS1124.

Bishop, Christopher (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag. ISBN: 978-0-387-31073-2.

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

Bouboulis, Pantelis and Sergios Theodoridis (2011). "Extension of Wirtinger's Calculus to Reproducing Kernel Hilbert Spaces and the Complex Kernel LMS". In: *IEEE Transactions on Signal Processing* 59.3, pp. 964–978. DOI: 10.1109/TSP.2010.2096420.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Breiman, Leo and Jerome H. Friedman (1985). "Estimating Optimal Transformations for Multiple Regression and Correlation". In: *Journal of the American Statistical Association* 80.391, pp. 590–598. DOI: 10.1080/01621459.1985.10478157.

Breslow, Norman E. and David G. Clayton (1993). "Approximate Inference in Generalized Linear Mixed Models". In: *Journal of the American Statistical Association* 88.421, pp. 9–25. DOI: 10.2307/2290687.

Bunch, David S. (1991). "Estimability in the multinomial probit model". In: *Transportation Research Part B: Methodological* 25.1, pp. 1–12. DOI: 10.1016/0191-2615(91)90009-8.

Cannings, Timothy I. and Richard J. Samworth (2017). "Random-projection ensemble classification". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4, pp. 959–1035. DOI: 10.1111/rssb.12228.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017). "Stan: A Probabilistic Programming Language". In: *Journal of Statistical Software, Articles* 76.1, pp. 1–32. DOI: 10.18637/jss.v076.i01.

Casella, George (1985). "An Introduction to Empirical Bayes Data Analysis". In: *The American Statistician* 39.2, pp. 83–87. DOI: 10.2307/2682801.

Casella, George and Roger L. Berger (2002). *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury. ISBN: 978-0-534-24312-8.

Casella, George, F. Javier Girón, M. Lina Martínez, and Elías Moreno (2009). "Consistency of Bayesian procedures for variable selection". In: *The Annals of Statistics* 37.3, pp. 1207–1228. DOI: 10.1214/08-AOS606.

Casella, George and Elías Moreno (2006). "Objective Bayesian Variable Selection". In: *Journal of the American Statistical Association* 101.473, pp. 157–167. DOI: 10.1198/016214505000000646.

Chen, Dong, Peter Hall, and Hans-Georg Müller (2011). "Single and Multiple Index Functional Regression Models with Nonparametric Link". In: *The Annals of Statistics* 39.3, pp. 1720–1747. DOI: 10.1214/11-AOS882.

Chen, Yen-Chi, Y. Samuel Wang, and Elena A. Erosheva (2018). "On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example". In: *Annals of Applied Statistics* to appear. arXiv: 1711.11057 [stat.ME].

Cheng, Ching-An and Byron Boots (2017). "Variational Inference for Gaussian Process Models with Linear Complexity". In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Ed. by Isabelle Guyon, Ulrike Von Luxburg, Samy Bengio, Hanna Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, pp. 5184–5194.

Chipman, Hugh, Edward I. George, and Robert E. McCulloch (2001). "The Practical Implementation of Bayesian Model Selection". In: *Model Selection*. Ed. by P. Lahiri. Vol. 38. Institute of Mathematical Statistics, pp. 65–134. DOI: 10.1214/lnms/1215540964.

Chopin, Nicolas (2011). "Fast simulation of truncated Gaussian distributions". In: *Statistics and Computing* 21.2, pp. 275–288. DOI: 10.1007/s11222-009-9168-1.

Cohen, Serge (2002). "Champs localement auto-similaires". In: *Lois d'échelle, fractales et ondelettes*. Ed. by Patrice Abry, Paulo Gonçalves, and Jacques Lévy Véhel. Vol. 1. Hermès Sciences Publications.

Damien, Paul and Stephen G. Walker (2001). "Sampling Truncated Normal, Beta, and Gamma Densities". In: *Journal of Computational and Graphical Statistics* 10.2, pp. 206–215.

Dansie, Brenton R. (1985). "Parameter estimability in the multinomial probit model". In: *Transportation Research Part B: Methodological* 19.6, pp. 526–528. DOI: 10.1016/0191-2615(85)90047-5.

Davidian, Marie and David M. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall/CRC. ISBN: 978-0-412-98341-2.

Davison, Anthony Christopher (2003). *Statistical Models*. Cambridge University Press. ISBN: 978-0-511-81585-0. DOI: 10.1017/CBO9780511815850.

Dean, Angela and Daniel Voss (1999). *Design and Analysis of Experiments*. Springer. ISBN: 978-0-387-98561-9. DOI: 10.1007/978-3-319-52250-0.

Dellaportas, Petros, Jonathan J. Forster, and Ioannis Ntzoufras (2002). "On Bayesian model and variable selection using MCMC". In: *Statistics and Computing* 12.1, pp. 27–36.

Dempster, Arthur P, Nan M Laird, and Donald B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 1–38.

Denwood, Matthew (2016). "**runjags**: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS". In: *Journal of Statistical Software* 71.9, pp. 1–25. DOI: 10.18637/jss.v071.i09.

Deterding, David Henry (1990). "Speaker Normalization for Automatic Speech Recognition". PhD thesis. University of Cambridge.

Diggle, Peter, Paula Moraga, Barry Rowlingson, and Benjamin Taylor (2013). "Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm". In: *Statistical Science* 28.4, pp. 542–563. DOI: 10.1214/13-STS441.

Diggle, Peter, Pingping Zheng, and Peter Durr (2005). "Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54.3, pp. 645–658. DOI: 10.1111/j.1467-9876.2005.05373.x.

Duane, Simon, Anthony D Kennedy, Brian J. Pendleton, and Duncan Roweth (1987). "Hybrid Monte Carlo". In: *Physics Letters B* 195.2, pp. 216–222. DOI: 10.1016/0370-2693(87)91197-X.

Durrande, Nicolas, David Ginsbourger, Olivier Roustant, and Laurent Carraro (2013). "ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis". In: *Journal of Multivariate Analysis* 115, pp. 57–67. DOI: 10.1016/j.jmva.2012.08.016.

Duvenaud, David (2014). "Automatic Model Construction with Gaussian Processes". PhD thesis. University of Cambridge.

Embrechts, Paul and Makoto Maejima (2002). *Selfsimilar Processes*. Princeton, NJ: Princeton University Press. ISBN: 978-0-691-09627-8.

Fernández, Carmen, Eduardo Ley, and Mark F. J. Steel (2001). "Benchmark priors for Bayesian model averaging". In: *Journal of Econometrics* 100.2, pp. 381–427. DOI: 10.1016/S0304-4076(00)00076-2.

Ferraty, Frédéric and Philippe Vieu (2006). *Nonparametric Functional Data Analysis*. Springer-Verlag. ISBN: 978-0-387-30369-7. DOI: 10.1007/0-387-36620-2.

Fisher, Ronald Aylmer (1922). "On the mathematical foundations of theoretical statistics". In: *Philosophical Transactions of the Royal Society A* 222.594-604, pp. 309–368. DOI: 10.1098/rsta.1922.0009.

Fouskakis, Dimitris and David Draper (2008). "Comparing Stochastic Optimization Methods for Variable Selection in Binary Outcome Prediction, With Application to Health Policy". In: *Journal of the American Statistical Association* 103.484, pp. 1367–1381. DOI: 10.1198/016214508000001048.

Fowlkes, Charless, Serge Belongie, Fan Chung, and Jitendra Malik (2004). "Spectral grouping using the Nyström method". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.2, pp. 214–225. DOI: 10.1109/TPAMI.2004.1262185.

Fowlkes, Charless, Serge Belongie, and Jitendra Malik (Dec. 2001). "Efficient Spatiotemporal Grouping Using the Nyström Method". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*. Vol. 1. Kauai, HI, pp. 231–238. DOI: 10.1109/CVPR.2001.990481.

Friedman, Jerome H., Trevor Hastie, and Robert Tibshirani (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

George, Edward I. and Robert E. McCulloch (1993). "Variable Selection Via Gibbs Sampling". In: *Journal of the American Statistical Association* 88.423, pp. 881–889. DOI: 10.2307/2290777.

Geweke, John (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration". In: *Econometrica* 57.6, pp. 1317–1339. DOI: 10.2307/1913710.

— (1996). "Variable Selection and Model Comparison in Regression". In: *Bayesian Statistics 5*. Proceedings of the Fifth Valencia International Meeting. Ed. by José M. Bernardo, James O. Berger, A. Philip Dawid, and Adrian F. M. Smith. Oxford University Press. ISBN: 978-0-19-852356-7.

Geweke, John, Michael Keane, and David Runkle (1994). "Alternative Computational Approaches to Inference in the Multinomial Probit Model". In: *The Review of Economics and Statistics* 76.4, pp. 609–632. DOI: 10.2307/2109766.

Girolami, Mark and Simon Rogers (2006). "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors". In: *Neural Computation* 18.8, pp. 1790–1817. DOI: 10.1162/neco.2006.18.8.1790.

Gu, Chong (2013). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag. ISBN: 978-1-4614-5368-0. DOI: 10.1007/978-1-4614-5369-7.

Guvenir, H. Altay, Burak Acar, Gulsen Demiroz, and Ayhan Cekin (1997). "A supervised machine learning algorithm for arrhythmia analysis". In: *Computers in Cardiology 1997*. Lund, Sweden, pp. 433–436. DOI: 10.1109/CIC.1997.647926.

Hajivassiliou, Vassilis and Daniel McFadden (1998). "The Method of Simulated Scores for the Estimation of LDV Models". In: *Econometrica* 66.4, pp. 863–896. DOI: 10.2307/2999576.

Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud (1996). "Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results". In: *Journal of Econometrics* 72.1–2, pp. 85–134. DOI: 10.1016/0304-4076(94)01716-6.

Hall, Peter, Tung Pham, Matt P. Wand, and Shen S. J. Wang (2011). "Asymptotic normality and valid inference for Gaussian variational approximation". In: *The Annals of Statistics* 39.5, pp. 2502–2532. DOI: 10.1214/11-AOS908.

Hastie, Trevor and Robert Tibshirani (1986). "Generalized Additive Models". In: *Statistical Science* 1.3, pp. 297–310. DOI: 10.1214/ss/1177013604.

Hein, Matthias and Olivier Bousquet (2004). *Kernels, Associated Structures and Generalizations*. Tech. rep. Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Hensman, James, Nicolo Fusi, and Neil D. Lawrence (Aug. 2013). "Gaussian Processes for Big Data". In: *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI 2013)*. Ed. by Ann Nicholson and Padhraic Smyth. Bellevue, WA. ISBN: 978-0-9749039-9-6. arXiv: 1309.6835 [cs.LG].

Hoerl, Arthur E. and Robert W. Kennard (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1, pp. 55–67. DOI: 10.2307/1267351.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky (1999). "Bayesian Model Averaging: A Tutorial". In: *Statistical science* 14.4, pp. 382–401. DOI: 10.1214/ss/1009212519.

Hofmann, Thomas, Bernhard Schölkopf, and Alexander J. Smola (2008). "Kernel Methods in Machine Learning". In: *The Annals of Statistics* 36.3, pp. 1171–1220. DOI: 10.1214/009053607000000677.

Itzykson, Claude and Jean-Michel Drouffe (1991). *Statistical Field Theory*. Vol. 2: Strong Coupling, Monte Carlo Methods, Conformal Field Theory and Random Systems. Cambridge University Press. ISBN: 978-0-511-62278-6. DOI: 10.1017/CBO9780511622786.

Jamil, Haziq (2017). ***iprior**: Regression Modelling using I-Priors*. R package version 0.7.1. URL: https://cran.r-project.org/web/packages/iprior.

— (2018). ***ipriorBVS**: Bayesian Variable Selection using I-priors*. R package version 0.1.1. URL: https://github.com/haziqj/ipriorBVS.

Jaynes, Edwin Thompson (1957a). "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4, p. 620. DOI: 10.1103/PhysRev.106.620.

— (1957b). "Information Theory and Statistical Mechanics II". In: *Physical Review* 108.2, p. 171. DOI: 10.1103/PhysRev.108.171.

— (2003). *Probability Theory: The Logic of Science*. Ed. by G. Larry Bretthorst. Cambridge University Press. ISBN: 978-0-521-59271-0.

Jeffreys, Harold (1946). "An invariant form for the prior probability in estimation problems". In: *Proceedings of the Royal Society A* 186.1007, pp. 453–461. DOI: 10.1098/rspa.1946.0056.

Jordan, Michael I., Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul (1999). "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2, pp. 183–233. DOI: 10.1023/A:1007665907178.

Kadane, Joseph B. (2011). *Principles of Uncertainty*. Chapman & Hall/CRC. ISBN: 978-1-4398-6161-5.

Kammar, Ohad (2016). *A note on Fréchet diffrentiation under Lebesgue integrals*. URL: https://www.cs.ox.ac.uk/people/ohad.kammar/notes/kammar-a-note-on-frechet-differentiation-under-lebesgue-integrals.pdf.

Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors". In: *Journal of the American Statistical Association* 90.430, pp. 773–795. DOI: 10.2307/2291091.

Keane, Michael (1992). "A Note on Identification in the Multinomial Probit Model". In: *Journal of Business & Economic Statistics* 10.2, pp. 193–200. DOI: 10.2307/1391677.

Keane, Michael and Kenneth Wolpin (1994). "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence". In: *The Review of Economics and Statistics* 76.4, pp. 648–672. DOI: 10.2307/2109768.

Kenward, Michael G. (1987). "A Method for Comparing Profiles of Repeated Measurements". In: *Journal of the Royal Statistical Society C (Applied Statistics)* 36.3, pp. 296–308. DOI: 10.2307/2347788.

Kimeldorf, George S and Grace Wahba (1970). "A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines". In: *The Annals of Mathematical Statistics* 41.2, pp. 495–502. DOI: 10.1214/aoms/1177697089.

Kokoszka, Piotr and Matthew Reimherr (2017). *Introduction to Functional Data Analysis*. Chapman & Hall/CRC. ISBN: 978-1-4987-4634-2.

Kuhn, Max et al. (2017). ***caret**: Classification and Regression Training*. R package version 6.0–77. URL: https://CRAN.R-project.org/package=caret.

Kuo, Frances Y., Ian H. Sloan, Grzegorz Wasilkowski, and Henryk Woźniakowski (2010). "On decompositions of multivariate functions". In: *Mathematics of Computation* 79.270, pp. 953–966. DOI: 10.1090/S0025-5718-09-02319-9.

Kuo, Lynn and Bani Mallick (1998). "Variable selection for regression models". In: *Sankhyā: The Indian Journal of Statistics, Series B* 60.1, pp. 65–81.

Kuss, Malte and Carl Edward Rasmussen (2005). "Assessing Approximate Inference for Binary Gaussian Process Classification". In: *Journal of Machine Learning Research* 6, pp. 1679–1704.

Kyung, Minjung, Jeff Gill, Malay Ghosh, and George Casella (2010). "Penalized regression, standard errors, and Bayesian lassos". In: *Bayesian Analysis* 5.2, pp. 369–411. DOI: 10.1214/10-BA607.

Lange, Kenneth (1995). "A quasi-Newton acceleration of the EM algorithm". In: *Statistica Sinica* 5.1, pp. 1–18.

Lee, Kyeong Eun, Naijun Sha, Edward R. Dougherty, Marina Vannucci, and Bani Mallick (2003). "Gene selection: a Bayesian variable selection approach". In: *Bioinformatics* 19.1, pp. 90–97. DOI: 10.1093/bioinformatics/19.1.90.

Lian, Heng and Gaorong Li (2014). "Series Expansion for Functional Sufficient Dimension Reduction". In: *Journal of Multivariate Analysis* 124, pp. 150–165. DOI: 10.1016/j.jmva.2013.10.019.

Liang, Feng, Rui Paulo, German Molina, Merlise A. Clyde, and James O. Berger (2008). "Mixtures of g Priors for Bayesian Variable Selection". In: *Journal of the American Statistical Association* 103.481, pp. 410–423. DOI: 10.1198/016214507000001337.

Liu, Chuanhai, Donald B. Rubin, and Ying Nian Wu (1998). "Parameter expansion to accelerate EM: The PX-EM algorithm". In: *Biometrika* 85.4, pp. 755–770. DOI: 10.1093/biomet/85.4.755.

Louppe, Gilles (Oct. 2014). "Understanding Random Forests: From Theory to Practice". PhD thesis. University of Liege, Belgium. arXiv: 1407.7502 [stat.ML].

Lunn, David J., Andrew Thomas, Nicky Best, and David Spiegelhalter (Oct. 2000). "WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility". In: *Statistics and Computing* 10.4, pp. 325–337. DOI: 10.1023/A:1008929526011.

Madigan, David and Adrian E. Raftery (1994). "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window". In: *Journal of the American Statistical Association* 89.428, pp. 1535–1546. DOI: 10.2307/2291017.

Mallows, Colin L. (1973). "Some comments on C_p". In: *Technometrics* 15.4, pp. 661–675. DOI: 10.2307/1267380.

Mandelbrot, Benoit B. and John W. Van Ness (1968). "Fractional Brownian Motions, Fractional Noises and Applications". In: *SIAM Review* 10.4, pp. 422–437.

Marsaglia, George and Wai Wan Tsang (2000). "The Ziggurat Method for Generating Random Variables". In: *Journal of Statistical Software* 5.8, pp. 1–7. DOI: 10.18637/jss.v005.i08.

Mary, Xavier (2003). "Hilbertian subspaces, subdualities and applications". PhD thesis. INSA de Rouen.

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models*. 2nd ed. Chapman & Hall/CRC. ISBN: 978-0-412-31760-6.

McCulloch, Robert E., Nicholas G. Polson, and Peter E. Rossi (2000). "A Bayesian analysis of the multinomial probit model with fully identified parameters". In: *Journal of Econometrics* 99.1, pp. 173–193. DOI: 10.1016/S0304-4076(00)00034-8.

McDonald, Gary C. and Richard C. Schwing (1973). "Instabilities of Regression Estimates Relating Air Pollution to Mortality". In: *Technometrics* 15.3, pp. 463–481. DOI: 10.2307/1266852.

McLachlan, Geoffrey and Thriyambakam Krishnan (2007). *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons.

Meng, Xiao-Li and Donald B. Rubin (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework". In: *Biometrika* 80.2, pp. 267–278. DOI: 10.1093/biomet/80.2.267.

Meng, Xiao-Li and David Van Dyk (1997). "The EM Algorithm—an Old Folk-song Sung to a Fast New Tune". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.3, pp. 511–567. DOI: 10.1111/1467-9868.00082.

Micchelli, Charles A., Yuesheng Xu, and Haizhang Zhang (2006). "Universal Kernels". In: *Journal of Machine Learning Research* 7, pp. 2651–2667.

Miller, Alan (2002). *Subset Selection in Regression*. Chapman & Hall/CRC. ISBN: 978-1-58488-171-1.

Minka, Thomas P. (Aug. 2001). "Expectation propagation for approximate Bayesian inference". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*. Ed. by Daphne Koller John Breese. San Francisco, CA, pp. 362–369. ISBN: 1-55860-800-1. arXiv: 1301.2294 [cs.AI].

Mitchell, Toby J. and John J. Beauchamp (1988). "Bayesian Variable Selection in Linear Regression". In: *Journal of the American Statistical Association* 83.404, pp. 1023–1032. DOI: 10.2307/2290129.

Neal, Radford M. (1999). "Regression and Classification using Gaussian Process Priors". In: *Bayesian Statistics 6*. Proceedings of the Sixth Valencia International Meeting. Ed. by José M. Bernardo, James O. Berger, A. Philip Dawid, and Adrian F. M. Smith. Oxford University Press, pp. 475–501. ISBN: 978-0-19-850485-6.

— (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo*. Ed. by Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. Chapman & Hall/CRC. ISBN: 978-1-4200-7941-8. arXiv: 1206.1901 [stat.CO].

Nobile, Agostino (1998). "A hybrid Markov chain for the Bayesian analysis of the multinomial probit model". In: *Statistics and Computing* 8.3, pp. 229–242. DOI: 10.1023/A:10089053.

Ntzoufras, Ioannis (2011). *Bayesian Modeling Using WinBUGS*. Wiley. ISBN: 978-0-470-14114-4. DOI: 10.1002/9780470434567.

O'Hara, Robert B. and Mikko J. Sillanpää (2009). "A Review of Bayesian Variable Selection Methods: What, How and Which". In: *Bayesian Analysis* 4.1, pp. 85–117. DOI: 10.1214/09-BA403.

Ong, Cheng Soon, Xavier Mary, Stéphane Canu, and Alexander J. Smola (July 2004). "Learning with non-positive kernels". In: *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004)*. Ed. by Russ Greiner and Dale Schuurmans. Banff, Alberta, Canada.

Ormerod, John T., Chong You, and Samuel Müller (2017). "A variational Bayes approach to variable selection". In: *Electronic Journal of Statistics* 11.2, pp. 3549–3594. DOI: 10.1214/17-EJS1332.

Pan, Jianxin and Yi Pan (2017). "**jmcm**: An R Package for Joint Mean-Covariance Modeling of Longitudinal Data". In: *Journal of Statistical Software* 82.1, pp. 1–29. DOI: 10.18637/jss.v082.i09.

Park, Trevor and George Casella (2008). "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482, pp. 681–686. DOI: 10.1198/016214508000000337.

Pawitan, Yudi (2001). *In All Likelihood*. Statistical Modelling and Inference Using Likelihood. Oxford University Press. ISBN: 978-0-19-850765-9.

Petersen, Kaare Brandt and Michael Syskind Pedersen (2012). *The Matrix Cookbook*. Technical University of Denmark. URL: http://www2.imm.dtu.dk/pubdb/p.php?3274.

Pinheiro, José C. and Douglas M. Bates (2000). *Mixed-Effects Models in S and S-plus*. Springer-Verlag. ISBN: 978-0-387-98957-0. DOI: 10.1007/b98882.

Pinheiro, José C., Douglas M. Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team (2017). ***nlme**: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-131. URL: https://CRAN.R-project.org/package=nlme.

Plummer, Martyn (Mar. 2003). "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling". In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Ed. by Kurt Hornik, Friedrich Leisch, and Achim Zeileis. Vienna, Austria.

Quiñonero-Candela, Joaquin and Carl Edward Rasmussen (Dec. 2005). "A Unifying View of Sparse Approximate Gaussian Process Regression". In: *Journal of Machine Learning Research* 6, pp. 1939–1959.

Rabe-Hesketh, Sophia and Anders Skrondal (2012). *Multilevel and Longitudinal Modeling Using Stata*. 3rd ed. Stata Press. ISBN: 978-1-59718-108-2.

Raftery, Adrian E., David Madigan, and Jennifer A Hoeting (1997). "Bayesian Model Averaging for Linear Regression Models". In: *Journal of the American Statistical Association* 92.437, pp. 179–191. DOI: 10.1080/01621459.1997.10473615.

Ramsay, James and Bernard W. Silverman (2005). *Functional Data Analysis*. New York: Springer-Verlag. ISBN: 978-1-4757-7107-7. DOI: 10.1007/978-1-4757-7107-7.

Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press. ISBN: 0-262-18253-X. URL: http://www.gaussianprocess.org/gpml/.

Raykar, Vikas C and Ramani Duraiswami (Mar. 2007). "Fast large scale Gaussian process regression using approximate matrix-vector products". In: *Learning Workshop 2007*. San Juan, Peurto Rico.

Robbins, Herbert (1956). "An Empirical Bayes Approach to Statistics". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by Jerzy Neyman. Vol. 1: Contributions to the Theory of Statistics. Berkeley, CA: University of California Press, pp. 157–163.

Robert, Christian (1995). "Simulation of truncated normal variables". In: *Statistics and Computing* 5.2, pp. 121–125. DOI: 10.1007/BF00143942.

— (2007). *The Bayesian Choice*. From Decision-Theoretic Foundations to Computational Implementation. New York: Springer-Verlag. ISBN: 978-0-387-95231-4. DOI: 10.1007/0-387-71599-1.

Robinson, Anthony John (1989). "Dynamic error propagation networks". PhD thesis. University of Cambridge.

Rudin, Walter (1987). *Real and Complex Analysis*. 3rd ed. McGraw-Hill Education. ISBN: 978-0-07-100276-9.

SAS Institute Inc. (2008). *SAS/STAT(R) 9.2 User's Guide*. 2nd ed. Cary, NC: SAS Institute Inc. ISBN: 978-1-60764-566-5.

Schoenberg, Isaac J. (1937). "On Certain Metric Spaces Arising From Euclidean Spaces by a Change of Metric and Their Imbedding in Hilbert Space". In: *Annals of Mathematics* 38.4, pp. 787–793. DOI: 10.2307/1968835.

Schölkopf, Bernhard and Alexander J. Smola (2002). *Learning with Kernels*. Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press. ISBN: 978-0-262-19475-4.

Schwarz, Gideon (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464. DOI: 10.1214/aos/1176344136.

Scott, Steven L. and Hal R. Varian (2014). "Predicting the present with Bayesian structural time series". In: *International Journal of Mathematical Modelling and Numerical Optimisation* 5.1-2, pp. 4–23. DOI: 10.1504/IJMMNO.2014.059942.

Sejdinovic, Dino and Arthur Gretton (2012). *What is an RKHS?* COMPGI13 Advanced Topics in Machine Learning: Lectures conducted at University College London. URL: http://www.gatsby.ucl.ac.uk/%7B~%7Dgretton/coursefiles/RKHS%7B%5C_%7DNotes1.pdf.

Skrondal, Anders and Sophia Rabe-Hesketh (2004). *Generalized Latent Variable Modeling*. Multilevel, Longitudinal, and Structural Equation Models. Chapman & Hall/CRC. ISBN: 978-1-58488-000-4.

Sobol, Ilya M (2001). "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates". In: *Mathematics and Computers in Simulation* 55.1–3, pp. 271–280. DOI: 10.1016/S0378-4754(00)00270-6.

Sriperumbudur, Bharath, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar (2017). "Density Estimation in Infinite Dimensional Exponential Families". In: *Journal of Machine Learning Research* 18.57, pp. 1–59. arXiv: 1312.3516 [math.ST].

Stan Development Team (2016). ***RStan**: The R Interface to Stan*. R package version 2.14.1. URL: http://mc-stan.org/.

Steinwart, Ingo and Andreas Christmann (2008). *Support Vector Machines*. New York: Springer-Verlag. ISBN: 978-0-387-77241-7. DOI: 10.1007/978-0-387-77242-4.

Steinwart, Ingo, Don Hush, and Clint Scovel (2006). "An Explicit Description of the Reproducing Kernel Hilbert Spaces of Gaussian RBF Kernels". In: *IEEE Transactions on Information Theory* 52.10, pp. 4635–4643. DOI: 10.1109/TIT.2006.881713.

Sturtz, Sibylle, Uwe Ligges, and Andrew Gelman (2005). "**R2WinBUGS**: A Package for Running WinBUGS from R". In: *Journal of Statistical Software* 12.3, pp. 1–16. DOI: 10.18637/jss.v012.i03.

Tapia, Richard A. (1971). *The Differentiation and Integration of Nonlinear Operators*. Ed. by Louis B. Rall. DOI: 10.1016/C2013-0-11348-7.

Taylor, Benjamin, Tilman Davies, Barry Rowlingson, and Peter Diggle (2013). "**lgcp**: An R Package for Inference with Spatial and Spatio-Temporal Log-Gaussian Cox Processes". In: *Journal of Statistical Software* 52.4, pp. 1–40. DOI: 10.18637/jss.v052.i04.

Thodberg, Hans Henrik (1996). "A review of Bayesian neural networks with an application to near infrared spectroscopy". In: *IEEE Transactions on Neural Networks* 7.1, pp. 56–72. DOI: 10.1109/72.478392.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58.1, pp. 267–288. DOI: 10.1111/j.1467-9868.2011.00771.x.

Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu (May 2002). "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression". In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS 2002)*. Vol. 99. 10, pp. 6567–6572. DOI: 10.1073/pnas.082099299.

Titsias, Michalis (Apr. 2009). "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In: *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*. Ed. by David van Dyk and Max Welling. Clearwater Beach, FL, pp. 567–574.

Train, Kenneth (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press. ISBN: 978-0-511-80527-1. DOI: 10.1017/CBO9780511805271.

van der Vaart, Aad W. and J. Harry van Zanten (2008). "Reproducing kernel Hilbert spaces of Gaussian priors". In: *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*. Ed. by Bertrand Clarke and Subhashis Ghosal. Beachwood, OH: Institute of Mathematical Statistics, pp. 200–222. DOI: 10.1214/074921708000000156.

Wahba, Grace (1990). *Spline Models for Observational Data*. SIAM. ISBN: 978-0-89871-244-5. DOI: 10.1137/1.9781611970128.

Wasserman, Larry (2004). *All of Statistics. A Concise Course in Statistical Inference*. New York: Springer-Verlag. ISBN: 978-0-387-40272-7. DOI: 10.1007/978-0-387-21736-9.

Wassermann, Larry (2006). *All of Nonparametric Statistics*. New York: Springer-Verlag. ISBN: 978-0-387-25145-5. DOI: 10.1007/0-387-30623-4.

Wei, Greg C. G. and Martin A. Tanner (1990). "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms". In: *Journal of the American statistical Association* 85.411, pp. 699–704. DOI: 10.2307/2290005.

Williams, Christopher K. I. and Matthias Seeger (2001). "Using the Nyström Method to Speed Up Kernel Machines". In: *Advances in Neural Information Processing Systems 13 (NIPS 2000)*. Ed. by Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, pp. 682–688.

Yamamoto, Yutaka (2012). *From Vector Spaces to Function Spaces: Introduction to Functional Analysis with Applications*. Vol. 127. SIAM. ISBN: 978-1-61197-230-6. DOI: 10.1137/9781611972313.

Zafeiriou, Stefanos (Oct. 2012). "Subspace Learning in Krein Spaces: Complete Kernel Fisher Discriminant Analysis with Indefinite Kernels". In: *Proceedings of the Twelfth European Conference on Computer Vision (ECCV 2012)*. Ed. by Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid. Springer. Florence, Italy, pp. 488–501. ISBN: 978-3-642-33764-2. DOI: 10.1007/978-3-642-33765-9_35.

Zellner, Arnold (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions". In: *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. New York: Elsevier, pp. 233–243.

Zhang, Haizhang, Yuesheng Xu, and Jun Zhang (2009). "Reproducing Kernel Banach Spaces for Machine Learning". In: *Journal of Machine Learning Research* 10, pp. 2741–2775.

Zhang, Haizhang and Jun Zhang (2012). "Regularized learning in Banach spaces as an optimization problem: representer theorems". In: *Journal of Global Optimization* 54.2, pp. 235–250. DOI: 10.1007/s10898-010-9575-z.

Zhang, Huamin and Feng Ding (2013). "On the Kronecker Products and Their Applications". In: *Journal of Applied Mathematics* 296185. DOI: 10.1155/2013/296185.

Zhu, Hongxiao, Fang Yao, and Hao Helen Zhang (2014). "Structured functional additive regression in reproducing kernel Hilbert spaces". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.3, pp. 581–603. DOI: 10.1111/rssb.12036.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

Appendix A

Functional derivative of the entropy

We present the functional derivative of the entropy $H(p)$ in equation 3.6 (p. 90). Typically, this is tackled using calculus of variations, but it can also be obtained using the Fréchet and Gâteaux differentials. Both methods are presented.

A.1 The usual functional derivative

The functional derivative is defined as follows.

Definition A.1 (Functional derivative)**.** Given a manifold M representing continuous/smooth functions ρ with certain boundary conditions, and a functional $F : M \to \mathbb{R}$, the functional derivative of $F(\rho)$ with respect to ρ, denoted $\partial F/\partial\rho$, is defined by

$$\begin{aligned}\int \frac{\partial F}{\partial \rho}(x)\phi(x)\,\mathrm{d}x &= \lim_{\epsilon\to 0}\frac{F(\rho+\epsilon\phi)-F(\rho)}{\epsilon}\\ &= \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon}F(\rho+\epsilon\phi)\right]_{\epsilon=0},\end{aligned}$$

where ϕ is an arbitrary function. The function $\partial F/\partial\rho$ as the gradient of F at the point ρ, and

$$\partial F(\rho,\phi) = \int \frac{\partial F}{\partial \rho}(x)\phi(x)\,\mathrm{d}x$$

as the directional derivative at point ρ in the direction of ϕ. Analogous to vector calculus, the inner product with the gradient gives the directional derivative.

Now let X be a discrete random variable with probability mass function $p(x) \geq 0$, for $\forall x \in \Omega$, a finite set. The entropy is a functional of p, namely

$$H(p) = -\sum_{x\in\Omega} p(x)\log p(x).$$

Equivalently, using the counting measure ν on Ω, we can write

$$H(p) = -\int_\Omega p(x) \log p(x) \, \mathrm{d}\nu(x).$$

Using the definition of functional derivatives, we find that

$$\begin{aligned}
\int_\Omega \frac{\partial H}{\partial p}(x)\phi(x) \, \mathrm{d}x &= \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon} H(p + \epsilon\phi)\right]_{\epsilon=0} \\
&= \left[-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\big(p(x) + \epsilon\phi(x)\big) \log\big(p(x) + \epsilon\phi(x)\big)\right]_{\epsilon=0} \\
&= -\int_\Omega \left(\frac{p(x)\phi(x)}{p(x) + \epsilon\phi(x)} + \frac{\epsilon\phi(x)}{p(x) + \epsilon\phi(x)} + \phi(x) \log\big(p(x) + \epsilon\phi(x)\big)\right) \mathrm{d}x \\
&= -\int_\Omega \big(1 + \log p(x)\big)\, \phi(x) \, \mathrm{d}x.
\end{aligned}$$

Thus, $(\partial H / \partial p)(x) = -1 - \log p(x)$.

A.2 Fréchet differential of the entropy

Since we have already introduced concepts of Fréchet and Gâteaux derivatives earlier, we shall use those instead. Assume that the entropy H is Fréchet differentiable at p, and that the probability densities p under consideration belong to the Hilbert space of square integrable functions $\mathrm{L}^2(\Theta, \nu)$ with inner product $\langle p, p' \rangle_{L^2(\Theta,\nu)} = \int pp' \, \mathrm{d}\nu$. Now since the Fréchet derivative of H at p is assumed to exist, it is equal to the Gâteaux derivative, which can be computed as follows:

$$\begin{aligned}
\partial_q H(p) &= \frac{\mathrm{d}}{\mathrm{d}t} H(p + tq)\bigg|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\left\{-\int_\Theta \big(p(\theta) + tq(\theta)\big) \log\big(p(\theta) + tq(\theta)\big) \, \mathrm{d}\nu(\theta)\right\}\bigg|_{t=0} \\
&= -\int_\Theta \left\{\frac{\mathrm{d}}{\mathrm{d}t}\big(p(\theta) + tq(\theta)\big) \log\big(p(\theta) + tq(\theta)\big)\bigg|_{t=0}\right\} \mathrm{d}\nu(\theta) \\
&= -\int_\Theta \left(\frac{p(\theta)q(\theta)}{p(\theta) + tq(\theta)} + \frac{tq(\theta)^2}{p(\theta) + tq(\theta)} + q(\theta) \log\big(p(\theta) + tq(\theta)\big)\right)\bigg|_{t=0} \mathrm{d}\nu(\theta) \\
&= -\int_\Theta q(\theta)\big(1 + \log p(\theta)\big) \, \mathrm{d}\nu(\theta) \\
&= \big\langle -\big(1 + \log p\big), q\big\rangle_\Theta \\
&= \mathrm{d}H(p)(q).
\end{aligned}$$

By definition, the gradient of H at p, denoted $\nabla H(p)$, is equal to $-1 - \log p$. This agrees with the usual functional derivative of the entropy obtained via standard calculus of variations.

Appendix B

Kronecker product and vectorisation

The Kronecker product crops up in the definition of matrix normal distributions, which is used in Chapter 5 for the I-probit model.

Definition B.1 (Kronecker product)**.** The Kronecker matrix product, denoted by $\otimes$, for two matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{p\times q}$ is defined by

$$A \otimes B = \begin{pmatrix} A_{11}B & A_{12}B & \cdots & A_{1m}B \\ A_{21}B & A_{22}B & \cdots & A_{2m}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1}B & A_{n2}B & \cdots & A_{nm}B \end{pmatrix} \in \mathbb{R}^{np\times mq}.$$

The Kronecker product is a generalisation of the outer product for vectors to matrices. Of use will be these properties of the Kronecker product (Huamin Zhang and Ding, 2013):

- **Bilinearity and associativity**. For appropriately sized matrices A, B and C, and a scalar λ,

$$\begin{aligned} A \otimes (B + C) &= A \otimes B + A \otimes C \\ (A + B) \otimes C &= A \otimes C + B \otimes C \\ \lambda A \otimes B &= A \otimes \lambda B = \lambda(A \otimes B) \\ (A \otimes B) \otimes C &= A \otimes (B \otimes C) \end{aligned}$$

- **Non-commutative**. In general, $A \otimes B \neq B \otimes A$, but they are *permutation equivalent*, i.e. $A \otimes B \neq P(B \otimes A)Q$ for some permutation matrices P and Q.
- **The mixed product property**. $(A \otimes B)(C \otimes D) = AC \otimes BD$.
- **Inverse**. $A \otimes B$ is invertible if and only if A and B are both invertible, and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

- **Transpose**. $(A \otimes B)^\top = A^\top \otimes B^\top$.
- **Determinant**. If A is $n \times n$ and B is $m \times m$, then $|A \otimes B| = |A|^m |B|^n$. Note that the exponent of $|A|$ is the order of B and vice versa.
- **Trace**. Suppose A and B are square matrices. Then $\operatorname{tr}(A \otimes B) = \operatorname{tr} A \operatorname{tr} B$.
- **Rank**. $\operatorname{rank}(A \otimes B) = \operatorname{rank} A \operatorname{rank} B$.
- **Matrix equations**. $AXB = C \Leftrightarrow (B^\top \otimes A) \operatorname{vec} X = \operatorname{vec}(AXB) = \operatorname{vec} C$.

The equivalence between matrix normal and multivariate normal distributions are established making use of vectorisation for matrices. This is defined below.

Definition B.2 (Vectorisation)**.** The vectorisation operation 'vec' stacks the columns of the matrices into one long vector, for instance, for the matrix $A \in \mathbb{R}^{n \times m}$

$$\operatorname{vec} A = (A_{11}, \ldots, A_{n1}, A_{12}, \ldots, A_{n2}, \ldots, A_{1m}, \ldots, A_{nm})^\top \in \mathbb{R}^{nm}.$$

Appendix C

Statistical distributions and their properties

This is a reference relating to the multivariate normal, matrix normal, truncated univariate and multivariate normal, which are collated from various sources for convenience. Of interest are probability density functions, first and second moments, and entropy (Definition 3.5, page 90). Note that in this part of the appendix, boldface notation for matrix and vectors are not used.

C.1 Multivariate normal distribution

Let $X \in \mathbb{R}^d$ be distributed according to a multivariate normal (Gaussian) distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^d$ (a square, symmetric, positive-definite matrix). We say that $X \sim \mathrm{N}_d(\mu, \Sigma)$. Then,

- **Pdf**. $p(X|\mu, \Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\big(-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)\big)$.
- **Moments**. $\mathrm{E}\, X = \mu$, $\mathrm{E}[XX^\top] = \Sigma + \mu\mu^\top$.
- **Entropy**. $H(p) = \frac{1}{2}\log|2\pi e\Sigma| = \frac{d}{2}(1 + \log 2\pi) + \frac{1}{2}\log|\Sigma|$.

For $d = 1$, i.e. X is univariate, then its pdf is $p(X|\mu, \sigma^2) = \frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)$, and its cdf is $F(X|\mu, \sigma^2) = \Phi\left(\frac{X-\mu}{\sigma}\right)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of a univariate standard normal distribution. In the special case that $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, then the components of $X = (X_1, \ldots, X_d)^\top$ are independently distributed according to $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$.

Lemma C.1 (Properties of multivariate normal)**.** *Assume that* $X \sim \mathrm{N}_d(\mu, \Sigma)$ *and* $Y \sim \mathrm{N}_d(\nu, \Psi)$*, where*

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \textit{and} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_b \end{pmatrix}.$$

Then,

- ***Marginal distributions.***

$$X_a \sim \mathrm{N}_{\dim X_a}(\mu_a, \Sigma_a) \quad \textit{and} \quad X_b \sim \mathrm{N}_{\dim X_b}(\mu_b, \Sigma_b).$$

- ***Conditional distributions.***

$$X_a | X_b \sim \mathrm{N}_{\dim X_a}(\tilde{\mu}_a, \tilde{\Sigma}_a) \quad \textit{and} \quad X_b \sim \mathrm{N}_{\dim X_b}(\tilde{\mu}_b, \tilde{\Sigma}_b),$$

  where

$$\begin{aligned} \tilde{\mu}_a &= \mu_a + \Sigma_{ab}\Sigma_b^{-1}(X_b - \mu_b) & \tilde{\mu}_b &= \mu_b + \Sigma_{ab}^\top \Sigma_a^{-1}(X_a - \mu_a) \\ \tilde{\Sigma}_a &= \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ab}^\top & \tilde{\Sigma}_b &= \Sigma_b - \Sigma_{ab}^\top \Sigma_a^{-1}\Sigma_{ab} \end{aligned}$$

- ***Linear combinations.***

$$AX + BY + C \sim \mathrm{N}_d(A\mu + B\nu + C, A\Sigma A^\top + B\Psi B^\top)$$

  where A and B are appropriately sized matrices, and $C \in \mathbb{R}^d$.

- ***Product of Gaussian densities.***

$$p(X|\mu, \Sigma)p(Y|\nu, \Psi) \propto p(Z|m, S)$$

  where $p(Z)$ is a Gaussian density, $m = S(\Sigma^{-1}\mu + \Psi^{-1}\nu)$ and $S = (\Sigma^{-1} + \Psi^{-1})^{-1}$. The normalising constant is equal to the density of $\mu \sim \mathrm{N}(\nu, \Sigma + \Psi)$.

Proof. Omitted—see Petersen and Pedersen (2012, §8). ■

Frequently, in Bayesian statistics especially, the following identities will be useful in deriving posterior distributions involving multivariate normals.

Lemma C.2. *Let $x, b \in \mathbb{R}^d$ be a vector, $X, B \in \mathbb{R}^{n\times d}$ a matrix, and $A \in \mathbb{R}^{d\times d}$ a symmetric, invertible matrix. Then,*

$$\begin{aligned} -\frac{1}{2}x^\top Ax + b^\top x &= -\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) + \frac{1}{2}b^\top A^{-1}b \\ -\frac{1}{2}\operatorname{tr}(X^\top AX) + \operatorname{tr}(B^\top X) &= -\frac{1}{2}\operatorname{tr}\left((X - A^{-1}B)^\top A(X - A^{-1}B)\right) + \frac{1}{2}\operatorname{tr}(B^\top A^{-1}B). \end{aligned}$$

Proof. Omitted—see Petersen and Pedersen (2012, §8.1.6). ■

Lemma C.3. *Let $X \sim \mathrm{N}_p(\mu_\theta, \Sigma_\theta)$, that is, the mean vector μ_θ and covariance matrix Σ_θ depends on a real, q-dimensional vector θ. The Fisher information matrix $U \in \mathbb{R}^{q\times q}$ for θ has (i, j) entries given by*

$$U_{ij} = \frac{\partial \mu_\theta^\top}{\partial \theta_i}\Sigma_\theta^{-1}\frac{\partial \mu_\theta}{\partial \theta_j} + \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_j}\right) \tag{C.1}$$

for $i, j = 1, \ldots, q$.

Proof. Define the derivative of a matrix $\Sigma \in \mathbb{R}^{p\times p}$ with respect to a scalar z, denoted $\partial\Sigma/\partial z \in \mathbb{R}^{p\times p}$, by $(\partial\Sigma/\partial z)_{ij} = \partial\Sigma_{ij}/\partial z$, i.e. derivatives are taken element-wise. The two identities below

are useful:

$$\frac{\partial}{\partial z}\operatorname{tr}\Sigma = \operatorname{tr}\frac{\partial \Sigma}{\partial z} \tag{C.2}$$

$$\frac{\partial}{\partial z}\log|\Sigma| = \operatorname{tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial z}\right) \tag{C.3}$$

$$\frac{\partial \Sigma^{-1}}{\partial z} = -\Sigma^{-1}\frac{\partial \Sigma}{\partial z}\Sigma^{-1} \tag{C.4}$$

A useful reference for these identities is Petersen and Pedersen (2012).

Differentiating the log-likelihood for θ with respect to the i'th component of θ yields

$$\begin{aligned}
\frac{\partial}{\partial \theta_i}L(\theta|X) =& -\frac{1}{2}\frac{\partial}{\partial \theta_i}\log|\Sigma_\theta| - \frac{1}{2}\frac{\partial}{\partial \theta_i}\operatorname{tr}(\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top) \\
=& -\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\right) - \frac{1}{2}\operatorname{tr}\left(\frac{\partial \Sigma_\theta^{-1}}{\partial \theta_i}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
& -\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial}{\partial \theta_i}\big((X-\mu_\theta)(X-\mu_\theta)^\top\big)\right) \\
=& \overbrace{-\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\right)}^{(A)} \overbrace{-\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top\right)}^{(B)} \\
=& \overbrace{+\operatorname{tr}\left(\Sigma_\theta^{-1}(X-\mu_\theta)\frac{\partial \mu_\theta^\top}{\partial \theta_i}\right)}^{(C)}.
\end{aligned}$$

Taking derivatives again, this time with respect to θ_j, of the three parts (A), (B) and (C) above, we get:

- **(A)**

$$\frac{1}{2}\frac{\partial}{\partial \theta_j}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\right) = \frac{1}{2}\operatorname{tr}\left(\frac{\partial \Sigma_\theta^{-1}}{\partial \theta_j}\frac{\partial \Sigma_\theta}{\partial \theta_i} + \Sigma_\theta^{-1}\frac{\partial^2 \Sigma_\theta}{\partial \theta_i \theta_j}\right)$$

- **(B)**

$$\begin{aligned}
&\frac{1}{2}\frac{\partial}{\partial \theta_j}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
&\quad= \frac{1}{2}\operatorname{tr}\left(\frac{\partial \Sigma_\theta^{-1}}{\partial \theta_j}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
&\qquad+ \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial^2 \Sigma_\theta}{\partial \theta_i \theta_j}\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
&\qquad+ \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\frac{\partial \Sigma_\theta^{-1}}{\partial \theta_j}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
&\qquad- \operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1}\frac{\partial \mu_\theta}{\partial \theta_j}(X-\mu_\theta)^\top\right)
\end{aligned}$$

- **(C)**

$$\frac{\partial}{\partial\theta_j}\operatorname{tr}\left(\Sigma_\theta^{-1}(X-\mu_\theta)\frac{\partial\mu_\theta^\top}{\partial\theta_i}\right)=\operatorname{tr}\left(\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}(X-\mu_\theta)\frac{\partial\mu_\theta^\top}{\partial\theta_i}-\Sigma_\theta^{-1}\frac{\partial\mu_\theta}{\partial\theta_j}\frac{\partial\mu_\theta^\top}{\partial\theta_i}\right.$$
$$\left.-\Sigma_\theta^{-1}(X-\mu_\theta)\frac{\partial^2\mu_\theta}{\partial\theta_i\partial\theta_j}\right)$$

The Fisher information matrix U contains (i,j) entries equal to the expectation of $-\frac{\partial^2}{\partial\theta_i\theta_j}L(\theta|X)$. Using the fact that 1) $\mathrm{E}[X-\mu_\theta]=0$; 2) $\mathrm{E}[\operatorname{tr}\Sigma]=\operatorname{tr}(\mathrm{E}\,\Sigma)$; 3) $\mathrm{E}[XX^\top]=\Sigma_\theta$; and 4) the trace is invariant under cyclic permutations, we get

$$\begin{aligned}U_{ij}=&\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\mu_\theta}{\partial\theta_j}\frac{\partial\mu_\theta^\top}{\partial\theta_i}\right)\\&+\frac{1}{2}\operatorname{tr}\left(\cancel{\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i}}+\cancel{\Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j}}-\cancel{\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i}}-\cancel{\Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j}}-\frac{\partial\Sigma_\theta}{\partial\theta_i}\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\right)\\=&\frac{\partial\mu_\theta^\top}{\partial\theta_i}\Sigma_\theta^{-1}\frac{\partial\mu_\theta}{\partial\theta_j}+\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_j}\right)\end{aligned}$$

as required. ■

C.2 Matrix normal distribution

The matrix normal distribution is an extension of the Gaussian distribution to matrices. Let $X\in\mathbb{R}^{n\times m}$ matrix, and let X follow a matrix normal distribution with mean $\mu\in\mathbb{R}^{n\times m}$ and row and column variances $\Sigma\in\mathbb{R}^{n\times n}$ and $\Psi\in\mathbb{R}^{m\times m}$ respectively, which we denote by $X\sim\mathrm{MN}_{n,m}(\mu,\Sigma,\Psi)$. Then,

- **Pdf**. $p(X|\mu,\Sigma,\Psi)=(2\pi)^{-nm/2}|\Sigma|^{-m/2}|\Psi|^{-n/2}e^{-\frac{1}{2}\operatorname{tr}\left(\Psi^{-1}(X-\mu)^\top\Sigma^{-1}(X-\mu)\right)}$.
- **Moments**. $\mathrm{E}\,X=\mu$, $\operatorname{Var}(X_{i\cdot})=\Psi$ for $i=1,\dots,n$, and $\operatorname{Var}(X_{\cdot j})=\Sigma$ for $j=1,\dots,m$.
- **Entropy**. $H(p)=\frac{1}{2}\log|2\pi e(\Psi\otimes\Sigma)|=\frac{nm}{2}(1+\log 2\pi)+\frac{1}{2}\log|\Sigma|^m|\Psi|^n$.

Lemma C.4 (Equivalence between matrix and multivariate normal)**.** $X\sim\mathrm{MN}_{n,m}(\mu,\Sigma,\Psi)$ *if and only if* $\operatorname{vec}X\sim\mathrm{N}_{nm}(\operatorname{vec}\mu,\Psi\otimes\Sigma)$.

Proof. In the exponent of the matrix normal pdf, we have

$$\begin{aligned}-\frac{1}{2}\operatorname{tr}\left(\Psi^{-1}(X-\mu)^\top\Sigma^{-1}(X-\mu)\right)&\\=-\frac{1}{2}\operatorname{vec}(X-\mu)^\top\operatorname{vec}(\Sigma^{-1}(X-\mu)\Psi^{-1})&\\=-\frac{1}{2}\operatorname{vec}(X-\mu)^\top(\Psi^{-1}\otimes\Sigma^{-1})\operatorname{vec}(X-\mu)&\\=-\frac{1}{2}(\operatorname{vec}X-\operatorname{vec}\mu)^\top(\Psi\otimes\Sigma)^{-1}(\operatorname{vec}X-\operatorname{vec}\mu).&\end{aligned}$$

Also, $|\Sigma|^{-m/2}|\Psi|^{-n/2}=|\Psi\otimes\Sigma|^{-1/2}$. This converts the matrix normal pdf to that of a multivariate normal pdf. ■

Some useful properties of the matrix normal distribution are listed:

- **Expected values**.

$$\begin{aligned}
\mathrm{E}(X-\mu)(X-\mu)^\top &= \mathrm{tr}(\Psi)\Sigma \in \mathbb{R}^{n\times n} \\
\mathrm{E}(X-\mu)^\top(X-\mu) &= \mathrm{tr}(\Sigma)\Psi \in \mathbb{R}^{m\times m} \\
\mathrm{E}[XAX^\top] &= \mathrm{tr}(A^\top\Psi)\Sigma + \mu A\mu^\top \\
\mathrm{E}[X^\top BX] &= \mathrm{tr}(\Sigma B^\top)\Psi + \mu^\top B\mu \\
\mathrm{E}[XCX] &= \Sigma C^\top\Psi + \mu C\mu
\end{aligned}$$

- **Transpose**. $X^\top \sim \mathrm{MN}_{m,n}(\mu^\top, \Psi, \Sigma)$.
- **Linear transformation**. Let $A \in \mathbb{R}^{a\times n}$ be of full-rank $a \leq n$ and $B \in \mathbb{R}^{m\times b}$ be of full-rank $b \leq m$. Then $AXB \sim \mathrm{MN}_{a,b}(\mu^\top, A\Sigma A^\top, B^\top\Psi B)$.
- **Iid**. If $X_i \overset{\text{iid}}{\sim} \mathrm{N}_m(\mu, \Psi)$ for $i = 1, \ldots, n$, and we arranged these vectors row-wise into the matrix $X = (X_1^\top, \ldots, X_n^\top)^\top \in \mathbb{R}^{n\times m}$, then $X \sim \mathrm{MN}(1_n\mu^\top, I_n, \Psi)$.

C.3 Truncated univariate normal distribution

Let $X \sim \mathrm{N}(\mu, \sigma^2)$ with the random variable X restricted to the interval $(a, b) \subset \mathbb{R}$. Then we say that X follows a truncated normal distribution, and we denote this by $X \sim {}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, a, b)$. Let $\alpha = (a-\mu)/\sigma$, $\beta = (b-\mu)/\sigma$, and $C = \Phi(\beta) - \Phi(\alpha)$. Then,

- **Pdf**. $p(X|\mu, \sigma, a, b) = C^{-1}(2\pi\sigma^2)^{-1/2}e^{-\frac{1}{2\sigma^2}(X-\mu)^2} = \sigma C^{-1}\phi(\frac{X-\mu}{\sigma})$.
- **Moments**.

$$\begin{aligned}
\mathrm{E}\,X &= \mu + \sigma\frac{\phi(\alpha)-\phi(\beta)}{C} \\
\mathrm{E}\,X^2 &= \sigma^2 + \mu^2 + \sigma^2\frac{\alpha\phi(\alpha)-\beta\phi(\beta)}{C} + 2\mu\sigma\frac{\phi(\alpha)-\phi(\beta)}{C} \\
\mathrm{Var}\,X &= \sigma^2\left[1 + \frac{\alpha\phi(\alpha)-\beta\phi(\beta)}{C} - \left(\frac{\phi(\alpha)-\phi(\beta)}{C}\right)^2\right]
\end{aligned}$$

- **Entropy**.

$$\begin{aligned}
H(p) &= \frac{1}{2}\log 2\pi e\sigma^2 + \log C + \frac{\alpha\phi(\alpha)-\beta\phi(\beta)}{2C} \\
&= \frac{1}{2}\log 2\pi e\sigma^2 + \log C + \frac{1}{2\sigma^2}\cdot\overbrace{\sigma^2\frac{\alpha\phi(\alpha)-\beta\phi(\beta)}{C}}^{\mathrm{Var}\,X-\sigma^2+(\mathrm{E}\,X-\mu)^2} \\
&= \frac{1}{2}\log 2\pi\sigma^2 + \log C + \frac{1}{2\sigma^2}\,\mathrm{E}[X-\mu]^2
\end{aligned}$$

because $\mathrm{Var}\,X + (\mathrm{E}\,X-\mu)^2 = \mathrm{E}\,X^2 - \cancel{(\mathrm{E}\,X)^2} + \cancel{(\mathrm{E}\,X)^2} + \mu^2 - 2\mu\,\mathrm{E}\,X$.

For binary probit models, the distributions that come up are one-sided truncations at zero, i.e. ${}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, 0, +\infty)$ (upper tail/positive part) and ${}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, -\infty, 0)$ (lower tail/negative part), for which their moments are of interest. As an aside, if $\mu = 0$ then the truncation ${}^{\mathrm{t}}\mathrm{N}(0, \sigma^2, 0, +\infty) \equiv \mathrm{N}_+(0, \sigma^2)$ is called the *folded-normal* distribution. For the positive one-sided truncation at zero, $C = \Phi(+\infty) - \Phi(-\mu/\sigma) = 1 - \Phi(-\mu/\sigma) = \Phi(\mu/\sigma)$, and for the negative one-sided truncation at zero, $C = \Phi(-\mu/\sigma) - \Phi(-\infty) = 1 - \Phi(\mu/\sigma)$. Additionally, if $\sigma = 1$, then ${}^{\mathrm{t}}\mathrm{N}(0, 1, 0, +\infty) \equiv \mathrm{N}_+(0, 1)$ is called the *half-normal* distribution.

One may simulate random draws from a truncated normal distribution by drawing from $\mathrm{N}(\mu, \sigma^2)$ and discarding samples that fall outside (a, b). Alternatively, the inverse-transform method using

$$X = \mu + \sigma\Phi^{-1}\left(\Phi(\alpha) + UC\right)$$

with $U \sim \mathrm{Unif}(0, 1)$ will work too. Either of these methods will work reasonably well as long as the truncation region is not too far away from μ, but neither is particularly fast. Efficient algorithms have been explored which are along the lines of either accept/reject algorithms (Robert, 1995), Gibbs sampling (Damien and Walker, 2001), or pseudo-random number generation algorithms (Chopin, 2011). The latter algorithm is inspired by the Ziggurat algorithm (Marsaglia and Tsang, 2000) which is considered to be the fastest Gaussian random number generator.

C.4 Truncated multivariate normal distribution

Consider the restriction of $X \sim \mathrm{N}_d(\mu, \Sigma)$ to a convex subset[1] $\mathcal{A} \subset \mathbb{R}^d$. Call this distribution the truncated multivariate normal distribution, and denote it $X \sim {}^{\mathrm{t}}\mathrm{N}_d(\mu, \Sigma, \mathcal{A})$. The pdf is $p(X|\mu, \Sigma, \mathcal{A}) = C^{-1}\phi(X|\mu, \Sigma)\,\mathbb{1}[X \in \mathcal{A}]$, where

$$C = \int_{\mathcal{A}} \phi(x|\mu, \Sigma)\,\mathrm{d}x = \mathrm{P}(X \in \mathcal{A}).$$

Generally speaking, there are no closed-form expressions for $\mathrm{E}\,g(X)$ for any well-defined functions g on X. One strategy to obtain values such as $\mathrm{E}\,X$ (mean), $\mathrm{E}\,X^2$ (second moment) and $E \log p(X)$ (entropy) would be Monte Carlo integration. If $X^{(1)}, \ldots, X^{(T)}$ are samples from $X \sim {}^{\mathrm{t}}\mathrm{N}_d(\mu, \Sigma, \mathcal{A})$, then $\widehat{\mathrm{E}\,g(X)} = \frac{1}{T}\sum_{i=1}^{T} g(X^{(i)})$.

Sampling from a truncated multivariate normal distribution is described by Robert (1995), who used a Gibbs-based approach, which we now describe. Assume that the one-dimensional slices of $\mathcal{A}$

$$\mathcal{A}_k(X_{-j}) = \{X_j | (X_1, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_d) \in \mathcal{A}\}$$

are readily available so that the bounds or anti-truncation region of X_j given the rest of the components X_{-j} are known to be (x_j^-, x_j^+). Using properties of the normal distribution, the full

[1] A convex subset is a subset of a space that is closed under convex combinations. In Euclidean space, for every pair of points in a convex set, all the points that lie on the straight line segment which joins the pair of points are also in the set.

conditionals of X_j given X_{-j} is

$$\begin{aligned} X_j|X_{-j} &\sim {}^{\mathrm{t}}\mathrm{N}(\tilde{\mu}_j, \tilde{\sigma}_j^2, x_j^-, x_j^+) \\ \tilde{\mu}_j &= \mu_j + \Sigma_{j,-j}^\top \Sigma_{-j,-j}(x_{-j} - \mu_{-j}) \\ \tilde{\sigma}_j^2 &= \Sigma_{11} - \Sigma_{j,-j}^\top \Sigma_{-j,-j} \Sigma_{j,-j}. \end{aligned}$$

According to Robert (1995), if $\Psi = \Sigma^{-1}$, then

$$\Sigma_{-j,-j}^{-1} = \Psi_{-j,-j} - \Psi_{j,-j}\Psi_{-j,-j}^\top / \Psi_{jj}$$

which means that we need only compute one global inverse Σ^{-1}. Therefore, the Gibbs sampler makes draws from truncated normal distributions in the following sequence, given initial values $X^{(0)}$:

- Draw $X_1^{(t)}|X_2^{(t)}, \ldots, X_d^{(t)} \sim {}^{\mathrm{t}}\mathrm{N}(\tilde{\mu}_1, \tilde{\sigma}_1^2, x_1^-, x_1^+)$.
- Draw $X_2^{(t)}|X_1^{(t+1)}, X_3^{(t)}, \ldots, X_d^{(t)} \sim {}^{\mathrm{t}}\mathrm{N}(\tilde{\mu}_2, \tilde{\sigma}_1^2, x_2^-, x_2^+)$.
- $\cdots$
- Draw $X_d^{(t)}|X_1^{(t+1)}, \ldots, X_{d-1}^{(t+1)} \sim {}^{\mathrm{t}}\mathrm{N}(\tilde{\mu}_d, \tilde{\sigma}_d^2, x_d^-, x_d^+)$.

In a later work, Damien and Walker (2001) introduce a latent variable $Y \in \mathbb{R}$ such that the joint pdf of X and Y is

$$p(X_1, \ldots, X_d, Y) \propto \exp(-Y/2)\, \mathbb{1}\left(Y > (X-\mu)^\top \Sigma^{-1}(X-\mu)\right) \mathbb{1}(X \in \mathcal{A}).$$

Now, the Gibbs conditional densities for the X_k's are given by

$$p(X_j|X_{-j}, Y) \propto \mathbb{1}(X_j \in \mathcal{B}_j)$$

where

$$\mathcal{B}_j \in (x_j^-, x_j^+) \cap \{X_j \,|\, (X-\mu)^\top \Sigma^{-1}(X-\mu) < Y\}.$$

Thus, given values for X_{-j} and Y, the bounds for X_j involves solving a quadratic equation in X_j. The Gibbs conditional density for $Y|X$ is a shifted exponential distribution, which can be sampled using the inverse-transform method. Thus, both X and Y can be sampled directly from uniform variates.

For probit models, we are interested in the conical truncations $\mathcal{C}_j = \{X_j > X_k | k \neq j, \text{and } k = 1, \ldots, m\}$ for which the j'th component of X is largest. These truncations form cones in d-dimensional space such that $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_d = \mathbb{R}^d$, and hence the name.

In the case where Σ is a diagonal matrix, the conically truncated multivariate normal distributions are easier to deal with due to the independence structure in the covariance matrix. In particular, most calculations of interest involve only a one dimensional integral of products of normal cdfs. We present some results that we have not previously seen before elsewhere.

Lemma C.5. *Let* $X \sim {}^{\mathrm{t}}\mathrm{N}_d(\mu, \Sigma, \mathcal{C}_j)$, *with* $\mu = (\mu_1, \ldots, \mu_d)^\top$ *and* $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, *and* $\mathcal{C}_j = \{X_j > X_k | k \neq j, \text{and } k = 1, \ldots, m\}$ *a conical truncation of* $\mathbb{R}^d$ *such that the* j*'th component is largest. Then,*

(i) ***Pdf***. *The pdf of X has the following functional form:*

$$p(X) = \frac{C^{-1}}{\sigma_1 \cdots \sigma_d (2\pi)^{d/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right]$$

where ϕ is the pdf of a standard normal distribution and

$$C = \mathrm{E}_Z\left[\prod_{\substack{i=1\\ i\neq j}}^{d} \Phi\left(\frac{\sigma_j}{\sigma_i} Z + \frac{\mu_j - \mu_i}{\sigma_i}\right)\right]$$

where $Z \sim \mathrm{N}(0,1)$.

(ii) ***Moments***. *The expectation $\mathrm{E}\, X = \left(\mathrm{E}\, X_1, \ldots, \mathrm{E}\, X_d\right)^\top$ is given by*

$$\mathrm{E}\, X_i = \begin{cases} \mu_i - \sigma_i C^{-1}\, \mathrm{E}_Z\left[\phi_i \prod_{k\neq i,j} \Phi_k\right] & \text{if } i \neq j \\ \mu_j - \sigma_j \sum_{i\neq j}\left(\mathrm{E}\, X_i - \mu_i\right) & \text{if } i = j \end{cases}$$

and the second moments $\mathrm{E}[X - \mu]^2$ are given by

$$\mathrm{E}[X_i - \mu_i]^2 = \begin{cases} \sigma_i^2 + (\mu_j - \mu_i)(\mathrm{E}\, X_i - \mu_i) + \sigma_i \sigma_j C^{-1}\, \mathrm{E}_Z\left[Z\phi_i \prod_{k\neq i,j} \Phi_k\right] & \text{if } i \neq j \\ C^{-1}\sigma_j^2\, \mathrm{E}_Z\left[Z^2 \prod_{k\neq j} \Phi_k\right] & \text{if } i = j \end{cases}$$

where we had defined

$$\phi_i = \phi_i(Z) = \phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i}\right), \text{ and}$$
$$\Phi_i = \Phi_i(Z) = \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i}\right).$$

(iii) ***Entropy***. *The entropy is given by*

$$H(p) = \log C + \frac{d}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{d}\frac{1}{\sigma_i^2}\,\mathrm{E}[x_i - \mu_i]^2.$$

Proof. See Appendix D for the proof. ■

C.5 Gamma distribution

For $X \in \mathbb{R}_{\geq 0}$, let X be distributed according to the gamma distribution with shape s and rate r, denoted $X \sim \Gamma(s, r)$. Then,

- **Pdf**. $p(X) = \Gamma(s)^{-1} r^s X^{s-1} e^{-rX}$.
- **Moments**. $\mathrm{E}\, X = s/r$, $\mathrm{Var}\, X = s/r^2$.
- **Entropy**. $H(p) = s - \log r + \log\Gamma(s) + (1-s)\psi(s)$.

In the above, $\Gamma(\cdot) = \Gamma_1(\cdot)$ and $\psi(\cdot) = \psi_1(\cdot)$ are the gamma and digamma functions, defined by

$$\Gamma(a) = \begin{cases} (a-1)! & \text{if } a \in \mathbb{Z}^+ \\ \int_0^\infty u^{a-1} e^{-u} \, \mathrm{d}u & \text{otherwise} \end{cases}$$

and

$$\psi(a) = \frac{\partial}{\partial a} \log \Gamma(a) = \frac{\partial \Gamma(a)/\partial a}{\Gamma(a)}.$$

Often, the gamma distribution is parameterised according to shape s and scale $\sigma = 1/r$ parameters, $X \sim \Gamma(s, \sigma)$.

C.6 Inverse gamma distribution

For $X \in \mathbb{R}_{\geq 0}$, a random variable X distributed according to an inverse gamma distribution with parameters s (shape) and σ (scale) is denoted by $X \sim \Gamma^{-1}(s, \sigma)$. Then,

- **Pdf**. $p(X) = \Gamma(s)^{-1} \sigma^s X^{-(s+1)} e^{-\sigma/X}$.
- **Moments**. $\mathrm{E}\, X = \sigma/(s-1)$, $\mathrm{Var}\, X = \sigma^2 \big((s-1)^2(s-2)\big)^{-1}$.
- **Entropy**. $H(p) = s + \log\big(\sigma \Gamma(s)\big) - (1+s)\psi(s)$.

with $\Gamma(\cdot)$ and $\psi(\cdot)$ representing the gamma and digamma functions respectively, as defined in Appendix C.5.

Lemma C.6. *If X has a Gamma distribution with shape and rate s and r, then $1/X \sim \Gamma^{-1}(s, r)$.*

Proof. Let $Y = 1/X$. Then the pdf of Y is

$$\begin{aligned} p_Y(Y) &= p_X(1/Y) \left| \frac{\partial}{\partial Y}(1/Y) \right| \\ &= \Gamma(s)^{-1} r^s (1/Y)^{s-1} e^{-r/Y} (1/Y^2) \\ &= \Gamma(s)^{-1} r^s Y^{-(s+1)} e^{-r/Y} \end{aligned}$$

which is the pdf of an inverse gamma with shape s and scale r. ∎

Appendix D

Proofs related to the conically truncated independent multivariate normal distribution

We present the proof for Lemma C.5 related to the conically truncated independent multivariate normal distribution, which we had not encountered in the literature.

D.1 Proof of Lemma C.5: Pdf

Using the fact that $\int p(x)\,\mathrm{d}x = 1$, and that

$$
\begin{aligned}
&\int \cdots \int [x_i < x_j, \forall i \neq j] \cdot \prod_{i=1}^{d} \phi(x_i|\mu_i, \sigma_i^2)\,\mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&\quad = \int \cdots \int \mathbb{1}[x_i < x_j, \forall i \neq j] \prod_{i=1}^{d} \left[\frac{1}{\sigma_i}\phi\left(\frac{x_i - \mu_i}{\sigma_i}\right)\right] \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&\quad = \int \cdots \int \mathbb{1}[x_i < x_j, \forall i \neq j] \frac{1}{\sigma_j}\phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \prod_{\substack{i=1 \\ i\neq j}}^{d} \left[\frac{1}{\sigma_i}\phi\left(\frac{x_i - \mu_i}{\sigma_i}\right)\right] \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&\quad = \int \prod_{\substack{i=1 \\ i\neq j}}^{d} \Phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \frac{1}{\sigma_j}\phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \mathrm{d}x_j \\
&\quad = \int \prod_{\substack{i=1 \\ i\neq j}}^{d} \Phi\left(\frac{\sigma_j z + \mu_j - \mu_i}{\sigma_i}\right) \phi(z)\,\mathrm{d}z \\
&\qquad \text{(by using the standardisation } z = (x_j - \mu_j)/\sigma_j)
\end{aligned}
$$

$$= \mathrm{E}_Z \left[\prod_{\substack{i=1 \\ i \neq j}}^{d} \Phi\left(\frac{\sigma_j}{\sigma_i} Z + \frac{\mu_j - \mu_i}{\sigma_i} \right) \right]$$

the proof follows directly.

D.2 Proof of Lemma C.5: Moments

Recall that for $Y \sim {}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, -\infty, b)$, for some function g of Y, we have that

$$\mathrm{E}\, g(Y) = \Phi(\beta)^{-1} \int [y < b] \cdot g(y) \phi(y|\mu, \sigma^2)\, \mathrm{d}y,$$

and in particular, we have

$$\mathrm{E}[Y - \mu] = -\sigma \frac{\phi(\beta)}{\Phi(\beta)} \tag{D.1}$$

$$\mathrm{E}[Y - \mu]^2 - \sigma^2 = -\sigma^2 \frac{\beta \phi(\beta)}{\Phi(\beta)} \tag{D.2}$$

where $\beta = (b - \mu)/\sigma$. For the conically truncated multivariate normal distribution $X \sim {}^{\mathrm{t}}\mathrm{N}_d(\mu, \Sigma, \mathcal{A}_j)$, where $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, the independence structure of Σ makes it possible to consider the expectations of each of the components separately by marginalising out the rest of the components. For simplicity, denote $p(x_k) = \phi(x_k|\mu_k, \sigma_k) = \sigma_k^{-1} \phi(\frac{x_k - \mu_k}{\sigma_k})$. For $i \neq j$, we have

$$\begin{aligned}
\mathrm{E}\, g(X_i) &= C^{-1} \int \cdots \int [x_k < x_j, \forall k \neq j] \cdot g(x_i) \prod_{k=1}^{d} p(x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \frac{\Phi((x_j - \mu_j)/\sigma_j)}{\Phi((x_j - \mu_j)/\sigma_j)} \iint [x_i < x_j] \cdot g(x_i) p(x_i) p(x_j) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k} \right) \mathrm{d}x_i\, \mathrm{d}x_j \\
&= C^{-1} \int \mathrm{E}_{X_i \sim {}^{\mathrm{t}}\mathrm{N}(\mu_i, \sigma_i^2, -\infty, x_j)}[g(X_i)] \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k} \right) p(x_j)\, \mathrm{d}x_j
\end{aligned} \tag{D.3}$$

where C is the normalising constant for X, while for the j'the component we have

$$\begin{aligned}
\mathrm{E}\, g(X_j) &= C^{-1} \int \cdots \int [x_k < x_j, \forall k \neq j] \cdot g(x_j) \prod_{k=1}^{d} p(x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \int g(x_j) \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k} \right) p(x_j)\, \mathrm{d}x_d.
\end{aligned} \tag{D.4}$$

Plugging in (D.1) for $g(X_i) = X_i - \mu_i$ in (D.3) we get

$$\begin{aligned}
\mathrm{E}\, X_i - \mu_i &= -C^{-1} \int \left(\sigma_i \phi\left(\frac{x_j - \mu_i}{\sigma_i} \right) \Big/ \Phi\left(\frac{x_j - \mu_i}{\sigma_i} \right) \right) \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k} \right) p(x_j)\, \mathrm{d}x_j \\
&= -\sigma_i C^{-1} \int \phi\left(\frac{x_j - \mu_i}{\sigma_i} \right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k} \right) p(x_j)\, \mathrm{d}x_j \\
&= -\sigma_i C^{-1} \int \phi\left(\frac{\sigma_j z + \mu_j - \mu_i}{\sigma_i} \right) \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{\sigma_j z + \mu_j - \mu_k}{\sigma_k} \right) \phi(z)\, \mathrm{d}z \\
&= -\sigma_i C^{-1}\, \mathrm{E}_Z \left[\phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i} \right) \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_k}{\sigma_k} \right) \right]
\end{aligned}$$

where Z is the distribution of $\mathrm{N}(0,1)$, and we had used a change of variable $x_j = \sigma_j z + \mu_j$, so that $p(x_j) = \sigma_j^{-1}\phi(z)$ and $\mathrm{d}x_j = \sigma_j \mathrm{d}z$. For the j'th component, substitute $g(x_j) = x_j - \mu_j$ in (D.4) to get

$$\begin{aligned}
\mathrm{E}\, X_j - \mu_j &= C^{-1} \int (x_j - \mu_j) \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k} \right) p(x_j)\, \mathrm{d}x_j \\
&= C^{-1} \sigma_j \int z \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{\sigma_j z + \mu_j - \mu_k}{\sigma_k} \right) \phi(z)\, \mathrm{d}z \\
&= \sigma_j \sum_{\substack{i=1 \\ i \neq j}}^{d} \sigma_i C^{-1}\, \mathrm{E} \left[\phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i} \right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_k}{\sigma_k} \right) \right] \\
&= -\sigma_j \sum_{\substack{i=1 \\ i \neq j}}^{d} \big(\mathrm{E}\, X_i - \mu_i \big),
\end{aligned}$$

where we have made use of Lemma D.1 in the second last step.

For the second moments, plug in (D.2) for $g(X_i) = (X_i - \mu_i)^2 - \sigma_i^2$ in (D.3) to get

$$\begin{aligned}
\mathrm{E}[X_i - \mu_i]^2 - \sigma_i^2 &= -\sigma_i^{\cancel{2}} C^{-1} \int \overbrace{\frac{x_j - \mu_i}{\sigma_i^{\cancel{i}}}}^{x_j - \mu_i - \mu_j + \mu_j} \cdot \frac{\phi\big((x_j - \mu_i)/\sigma_i\big)}{\Phi\big((x_j - \mu_i)/\sigma_i\big)} \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j \\
&= -\sigma_i C^{-1} \int (x_j - \mu_j)\phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j \\
&\quad + (\mu_j - \mu_i) \cdot \overbrace{-\sigma_i C^{-1} \int \phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j}^{\mathrm{E}\, X_i - \mu_i} \\
&= (\mu_j - \mu_i)(\mathrm{E}\, X_i - \mu_i) \\
&\quad + \sigma_i C^{-1} \int \sigma_j z \phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j z + \mu_j - \mu_k}{\sigma_k}\right) \phi(z)\,\mathrm{d}z \\
&= (\mu_j - \mu_i)(\mathrm{E}\, X_i - \mu_i) \\
&\quad + \sigma_i \sigma_j C^{-1}\, \mathrm{E}\left[Z\phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_k}{\sigma_k}\right) \right]
\end{aligned}$$

And similarly, for the j'th component

$$\begin{aligned}
\mathrm{E}[X_j - \mu_j]^2 &= C^{-1} \int (x_j - \mu_j)^2 \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j \\
&= C^{-1}\sigma_j^2 \int z^2 \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{z\sigma_j + \mu_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}z \\
&= C^{-1}\sigma_j^2\, \mathrm{E}_Z\left[Z^2 \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{Z\sigma_j + \mu_j - \mu_k}{\sigma_k}\right) \right].
\end{aligned}$$

Lastly, we use the following result in the derivation above.

Lemma D.1. *Let $Z \sim \mathrm{N}(0,1)$. Then for all $m \in \{\mathbb{N} \mid m > 1\}$ and $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,*

$$\mathrm{E}\left[Z \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi(\sigma_k Z + \mu_k) \right] = \sum_{\substack{i=1 \\ i \neq j}}^{m} \mathrm{E}\left[\sigma_i \phi(\sigma_i Z + \mu_i) \prod_{\substack{k=1 \\ k \neq i,j}}^{m} \Phi(\sigma_k Z + \mu_k) \right]$$

for some $j \in \{1, \ldots, m\}$.

Proof. Use the fact that for any differentiable function g, $\mathrm{E}[Zg(Z)] = \mathrm{E}[g'(Z)]$, and apply the result with the function $g_m : z \mapsto \prod_{k \neq j} \Phi(\sigma_k z + \mu_k)$. All that is left is to derive the derivative of

g, and we use an inductive proof to do this. Introduce the following notation for convenience:

$$\phi_i = \phi(\sigma_i z + \mu_i)$$
$$\Phi_i = \Phi(\sigma_i z + \mu_i)$$

The simplest case is when $m = 2$, which can be trivially shown to be true. Without loss of generality, let $j = 1$. Then

$$\begin{aligned} & g_2(z) = \Phi_2 \\ \Rightarrow\ & \dot{g}_2(z) = \sigma_2\phi_2 = \sum_{\substack{i=1 \\ i\neq 1}}^{2} \left[\sigma_i\phi_i \sum_{\substack{k=1 \\ k\neq 1,2}}^{2} \Phi_k \right]. \end{aligned}$$

Now assume that the inductive hypothesis holds for some $m \in \{\mathbb{N} \,|\, m > 1\}$. That is, the derivative of $g_m(z) = \prod_{k\neq j} \Phi_k$,

$$\dot{g}_m(z) = \sum_{\substack{i=1 \\ i\neq j}}^{m} \left[\sigma_i\phi_i \prod_{\substack{k=1 \\ k\neq i,j}}^{m} \Phi_k \right],$$

is assumed to be true. Also assume that, without loss of generality, $j \neq m + 1$. Then, the derivative of

$$g_{m+1}(z) = \prod_{\substack{k=1 \\ k\neq j}}^{m+1} \Phi_k = g_m(z)\Phi_{m+1}$$

is found to be

$$\begin{aligned} \dot{g}_{m+1}(z) &= \sigma_{m+1}\phi_{m+1}g_m(z) + \dot{g}_m(z)\Phi_{m+1} \\ &= \sigma_{m+1}\phi_{m+1} \prod_{\substack{k=1 \\ k\neq j}}^{m} \Phi_k + \sum_{\substack{i=1 \\ i\neq j}}^{m} \left[\sigma_i\phi_i \prod_{\substack{k=1 \\ k\neq i,j}}^{m} \Phi_k \right] \Phi_{m+1} \\ &= \sigma_{m+1}\phi_{m+1} \prod_{\substack{k=1 \\ k\neq j,m+1}}^{m+1} \Phi_k + \sum_{\substack{i=1 \\ i\neq j}}^{m} \left[\sigma_i\phi_i \prod_{\substack{k=1 \\ k\neq i,j}}^{m+1} \Phi_k \right] \\ &= \sum_{\substack{i=1 \\ i\neq j}}^{m+1} \left[\sigma_i\phi_i \prod_{\substack{k=1 \\ k\neq i,j}}^{m+1} \Phi_k \right], \end{aligned}$$

as required for the inductive proof. Using linearity of expectations, the proof is complete. ■

D.3 Proof of Lemma C.5: Entropy

As a direct consequence of the definition of entropy,

$$\begin{aligned} H(p) &= -\operatorname{E}[\log p(X)] \\ &= -\operatorname{E}\left[-\log C - \frac{d}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right] \\ &= \log C + \frac{d}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{d}\frac{1}{\sigma_i^2}\operatorname{E}[x_i-\mu_i]^2. \end{aligned}$$

Appendix E

I-prior interpretation of the g-prior

The I-prior for $\boldsymbol{\beta}$ in a standard linear model resembles the objective g-prior (Zellner, 1986) for regression coefficients,

$$\boldsymbol{\beta} \sim \mathrm{N}_p\left(\mathbf{0}, g(\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1}\right),$$

although they are quite different objects. The g-prior for $\boldsymbol{\beta}$ has the *inverse* (scaled) Fisher information matrix as its covariance matrix. This, in itself, has a much different and arguably counterintuitive meaning: large amounts of Fisher information about $\boldsymbol{\beta}$ corresponds to a small prior variance, and hence less deviation away from the prior mean of zero in estimating $\boldsymbol{\beta}$. The choice of the hyperparameter g has been the subject of much debate, with choices ranging from fixing $g = n$ (corresponding to the concept of *unit Fisher information*), to fully Bayesian and empirical Bayesian methods of estimating g from the data.

On the other hand, we note that the g-prior has an I-prior interpretation when argued as follows. Assume that the regression function f lies in the continual dual space of $\mathbb{R}^p$ equipped with the inner product $\langle \mathbf{x}, \mathbf{x}' \rangle_{\mathcal{X}} = \mathbf{x}^\top (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \mathbf{x}$. With this inner product and from (3.3) (p. 84), the Fisher information for $\boldsymbol{\beta}$ is

$$\begin{aligned}
\mathcal{I}_g(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \mathbf{x}_i \otimes (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \mathbf{x}_j \\
&= (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \cancel{(\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})} \cancel{(\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1}} \\
&= (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1},
\end{aligned}$$

and this, rather than the usual $\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X}$ as the prior covariance matrix for $\boldsymbol{\beta}$, means that the I-prior is in fact the standard g-prior.

The metric induced by the inner product is actually the *Mahalanobis distance*, a scale-invariant natural distance if the covariates are measured on different scales. To expand on this idea, circle back to the regression function and write it as $f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{X}}$. In usual least squares regression, the choice of inner product is irrelevant, so the usual dot product is com-

monly used (however, as we have seen above, the choice of inner product determines the form of the Fisher information for $\boldsymbol{\beta}$). In particular, suppose that all the x_{ik}'s, $k = 1, \ldots, p$ for each unit $i = 1, \ldots, n$ are measured on the same scale; for instance, these could be measurements in centimetres. In this case, the dot product is reasonable, because $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \sum_{k=1}^{p} x_{ik} x_{jk}$ and the inner product has a coherent unit, namely the squared unit of the x_{ik}'s. However, if they were a mix of various scaled measurements, then obviously the inner product's unit is incoherent—one would be resorted to adding measurements in different units, for example, cm^2 and kg^2 and so on. In such a case, a unitless inner product is appropriate, like the Mahalonobis inner product, which technically rescales the x_{ik}'s to unity. In summary, if the covariates are all measured on the same scale, then the I-prior is appropriate, and if not, the g-prior is appropriate.

Appendix F

Additional details for various I-prior regression models

These are additional details relating to discussion on various I-prior regression models in Section 4.1 of Chapter 4 (p. 96). These details relate to the standard linear multilevel model and the naïve classification model.

F.1 The I-prior for standard multilevel models

We show the corresponding I-prior for the regression coefficients of the standard linear multilevel model (4.3). Write $\alpha = \beta_0$, and for simplicity, assume iid errors, i.e., $\boldsymbol{\Psi} = \psi\mathbf{I}_n$. The form of $f \in \mathcal{F}$ is now $f(\mathbf{x}_i^{(j)}, j) = \sum_{i'=1}^{n_{j'}} \sum_{j'=1}^{m} h_\lambda\big((\mathbf{x}_i^{(j)}, j), (\mathbf{x}_{i'}^{(j')}, j')\big) w_{i'j'}$, where each $w_{i'j'} \sim \mathrm{N}(0, \psi^{-1})$.

Now, functions in the scaled RKHS $\mathcal{F}_2$ have the form

$$\begin{aligned} f_2(j) &= \sum_{i=1}^{n_{j'}} \sum_{j'=1}^{m} \lambda_2 \left(\frac{\delta_{jj'}}{p_j} - 1 \right) w_{ij'} \\ &= \lambda_2 \left(\frac{w_{+j}}{p_j} - w_{++} \right), \end{aligned}$$

where a '+' in the index of w_{ik} indicates a summation over that index, and p_j is the empirical distribution over $\mathcal{M}$, i.e. $p_j = n_j/n$. Clearly $f_2(j)$ is a variable depending on j, so write $f_2(j) = \beta_{0j}$. The distribution of β_{0j} is normal with zero mean and variance

$$\begin{aligned} \operatorname{Var} \beta_{0j} &= \lambda_2^2 \left(\frac{\cancel{n_j}\psi}{\cancel{n_j^2}/n^2} + n\psi \right) \\ &= n\psi\lambda_2^2 \left(\frac{1}{p_j} + 1 \right). \end{aligned}$$

The covariance between any two random intercepts β_{0j} and $\beta_{0j'}$ is

$$\begin{aligned}
\operatorname{Cov}[\beta_{0j}, \beta_{0j'}] &= \operatorname{Cov}\left[\lambda_2\left(\frac{w_{+j}}{p_j} - w_{++}\right), \lambda_2\left[\frac{w_{+j'}}{p_{j'}} - w_{++}\right)\right] \\
&= \frac{\lambda_2^2}{p_j p_{j'}} \cancelto{0}{\operatorname{Cov}(w_{+j}, w_{+j'})} - \frac{\lambda_2^2}{p_j}\operatorname{Cov}[w_{+j}, w_{++}] - \frac{\lambda_2^2}{p_{j'}}\operatorname{Cov}[w_{++}, w_{+j'}] \\
&\quad + \lambda_2^2 \operatorname{Cov}[w_{++}, w_{++}] \\
&= -\frac{\lambda_2^2}{\cancel{n_j/n}}\cancel{n_j}\psi - \frac{\lambda_2^2}{\cancel{n_{j'}/n}}\cancel{n_{j'}}\psi + \lambda_2^2 n\psi \\
&= -n\psi\lambda_2^2.
\end{aligned}$$

Functions in $\mathcal{F}_{12}$, on the other hand, have the form

$$\begin{aligned}
f_{12}(\mathbf{x}_i, j) &= \sum_{i'=1}^{n_{j'}}\sum_{j'=1}^{m} \lambda_1\lambda_2 \cdot \tilde{\mathbf{x}}_i^{(j)\top}\tilde{\mathbf{x}}_{i'}^{(j')} \cdot \left(\frac{\delta_{jj'}}{p_j} - 1\right) w_{i'j'} \\
&= \tilde{\mathbf{x}}_i^{(j)\top} \underbrace{\left(\frac{\lambda_1\lambda_2}{p_j}\sum_{i'=1}^{n_j}\tilde{\mathbf{x}}_{i'}^{(j)} w_{i'j} - \lambda_1\lambda_2\sum_{i'=1}^{n_{j'}}\sum_{j'=1}^{m}\tilde{\mathbf{x}}_{i'}^{(j')} w_{i'j'}\right)}_{\boldsymbol{\beta}_{1j}},
\end{aligned}$$

and this is, as expected, a linear form dependent on cluster j. We can calculate the variance for β_{1j} to be

$$\begin{aligned}
\operatorname{Var}\boldsymbol{\beta}_{1j} &= \lambda_1^2\lambda_2^2 \operatorname{Var}\left[\frac{1}{p_j}\tilde{\mathbf{X}}_j^\top \mathbf{w}_j - \tilde{\mathbf{X}}^\top \mathbf{w}\right] \\
&= \lambda_1^2\lambda_2^2\left(\frac{\psi}{n_j^2/n^2}\tilde{\mathbf{X}}_j^\top\tilde{\mathbf{X}}_j + \psi\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \frac{1}{p_j}\tilde{\mathbf{X}}_j^\top \operatorname{Cov}(\mathbf{w}_j, \mathbf{w})\tilde{\mathbf{X}}^\top\right) \\
&= n\psi\lambda_1^2\lambda_2^2\left(\frac{1}{p_j}\mathbf{S}_j + \mathbf{S} - \mathbf{S}_j\right) \\
&= n\psi\lambda_1^2\lambda_2^2\left(\left(\frac{1}{p_j} - 1\right)\mathbf{S}_j + \mathbf{S}\right)
\end{aligned}$$

where $\mathbf{S}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})^\top(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})$, $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})^\top(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})$, and $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{m}\mathbf{x}_i^{(j)}$. The covariance between two vectors of the random slopes is

$$\begin{aligned}
\operatorname{Cov}[\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{1j'}] &= \lambda_1^2\lambda_2^2 \operatorname{Cov}\left[\frac{1}{p_j}\tilde{\mathbf{X}}_j^\top \mathbf{w}_j - \tilde{\mathbf{X}}^\top \mathbf{w}, \frac{1}{p_{j'}}\tilde{\mathbf{X}}_{j'}^\top \mathbf{w}_{j'} - \tilde{\mathbf{X}}^\top \mathbf{w}\right] \\
&= \psi\lambda_1^2\lambda_2^2\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \frac{1}{p_j}\tilde{\mathbf{X}}_j^\top\tilde{\mathbf{X}}_j - \frac{1}{p_{j'}}\tilde{\mathbf{X}}_{j'}^\top\tilde{\mathbf{X}}_{j'}\right) \\
&= n\psi\lambda_1^2\lambda_2^2\left(\mathbf{S} - \mathbf{S}_j - \mathbf{S}_{j'}\right).
\end{aligned}$$

Another quantity of interest is the covariance between the random intercepts and random slopes:

$$\begin{aligned}
\operatorname{Cov}[\beta_{0j}, \boldsymbol{\beta}_{1j}] &= \lambda_1\lambda_2^2 \operatorname{Cov}\left[\frac{1}{p_j}\mathbf{1}_{n_j}^\top \mathbf{w}_j - \mathbf{1}_n^\top \mathbf{w}, \frac{1}{p_j}\tilde{\mathbf{X}}_j^\top \mathbf{w}_j - \tilde{\mathbf{X}}^\top \mathbf{w}\right] \\
&= \psi\lambda_1\lambda_2^2 \left(\cancelto{0}{\mathbf{1}_n^\top \tilde{\mathbf{X}}} + \frac{1}{p_j^2}\mathbf{1}_{n_j}^\top \tilde{\mathbf{X}}_j - \frac{2}{p_j}\mathbf{1}_{n_j}^\top \tilde{\mathbf{X}}_j\right) \\
&= n\psi\lambda_1\lambda_2^2 \left(\left(\frac{1}{p_j} - 2\right)\frac{1}{n_j}\sum_{i=1}^{n_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})\right) \\
&= n\psi\lambda_1\lambda_2^2 \left(\frac{1}{p_j} - 2\right)(\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}})
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{Cov}[\beta_{0j}, \boldsymbol{\beta}_{1j'}] &= \lambda_1\lambda_2^2 \operatorname{Cov}\left[\frac{1}{p_j}\mathbf{1}_{n_j}^\top \mathbf{w}_j - \mathbf{1}_n^\top \mathbf{w}, \frac{1}{p_{j'}}\tilde{\mathbf{X}}_{j'}^\top \mathbf{w}_{j'} - \tilde{\mathbf{X}}^\top \mathbf{w}\right] \\
&= \psi\lambda_1\lambda_2^2 \left(\cancelto{0}{\mathbf{1}_n^\top \tilde{\mathbf{X}}} + \frac{1}{p_j p_{j'}}\mathbf{1}_{n_j}^\top \cancelto{0}{\operatorname{Cov}(\mathbf{w}_j, \mathbf{w}_{j'})}\tilde{\mathbf{X}}_{j'} - \frac{1}{p_j}\mathbf{1}_{n_j}^\top \tilde{\mathbf{X}}_j - \frac{1}{p_{j'}}\mathbf{1}_{n_{j'}}^\top \tilde{\mathbf{X}}_{j'}\right) \\
&= n\psi\lambda_1\lambda_2^2 \left(-\frac{1}{n_j}\sum_{i=1}^{n_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}) - \frac{1}{n_{j'}}\sum_{i=1}^{n_{j'}}(\mathbf{x}_i^{(j')} - \bar{\mathbf{x}})\right) \\
&= n\psi\lambda_1\lambda_2^2 \left(2\bar{\mathbf{x}} - \bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}}^{(j')}\right).
\end{aligned}$$

F.2 The I-prior for naïve classification

For the naïve I-prior classification model (4.7), the I-prior is derived as follows. Firstly, the functions in $\mathcal{F}_\mathcal{M}$ and $\mathcal{F}_\mathcal{X}$ need necessarily be zero-mean functions (as per the functional ANOVA definition in Definition 2.36, but also, as per the definition of the Pearson RKHS and centred identity kernel RKHS). What this means is that $\sum_{j=1}^m \alpha_j = 0$, $\sum_{j=1}^m f_j(x_i) = 0$, and $\sum_{i=1}^n f_j(x_i) = 0$. In particular,

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^m y_{ij}\right] &= \sum_{j=1}^m (\alpha + \alpha_j + f_j(x_i)) \\
&= m\alpha + \cancelto{0}{\sum_{j=1}^m \alpha_j} + \cancelto{0}{\sum_{j=1}^m f_j(x_i)}
\end{aligned}$$

and since $\sum_{j=1}^m y_{ij} = 1$, we get the ML estimate $\hat{\alpha} = 1/m$, and thus the grand intercept can be fixed to resolve identification.

It is much more convenient to work in vector and matrix form, so let us introduce some notation. Let $\mathbf{w}$ (c.f. $\mathbf{y}$, $\mathbf{f}$ and $\boldsymbol{\epsilon}$) be an $n \times m$ matrix whose (i, j) entries contain w_{ij} (c.f. y_{ij}, $f(x_i, j)$, and ϵ_{ij}). The row-wise entries of $\mathbf{w}$ are independent of each other (independence assumption of the n observations), while any two of their columns have covariance as specified in $\boldsymbol{\Psi}$. This means that $\mathbf{w}$ follows a matrix normal distribution $\mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$, which implies $\operatorname{vec}\mathbf{w} \sim \mathrm{N}_{nm}(\mathbf{0}, \boldsymbol{\Psi} \otimes \mathbf{I}_n)$, and similarly, $\boldsymbol{\epsilon} \sim \mathrm{N}_{nm}(\mathbf{0}, \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)$. Denote by $\mathbf{B}_\eta$ the $n \times n$ kernel

matrix with entries supplied by kernel $1 + b_\eta$ over $\mathcal{X} \times \mathcal{X}$, and $\mathbf{A}$ the $m \times m$ matrix with entries supplied by a over $\mathcal{M} \times \mathcal{M}$. From (4.7), we have that

$$\mathbf{f} = \mathbf{B}_\eta \mathbf{w} \mathbf{A} \in \mathbb{R}^{n \times m},$$

and thus $\operatorname{vec} \mathbf{f} \sim \mathrm{N}_{nm}(\mathbf{0}, \mathbf{A}\boldsymbol{\Psi}\mathbf{A} \otimes \mathbf{B}_\eta^2)$. As $\mathbf{y} = \mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{f} + \boldsymbol{\epsilon}$, where $\boldsymbol{\alpha} \in \mathbb{R}^m$ with j'th component $\alpha + \alpha_j = 1/m + \alpha_j$, by linearity we have that

$$\operatorname{vec} \mathbf{y} \sim \mathrm{N}_{nm}\left(\operatorname{vec} \boldsymbol{\alpha}, \mathbf{A}\boldsymbol{\Psi}\mathbf{A} \otimes \mathbf{B}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n\right) \tag{F.1}$$

and

$$\operatorname{vec} \mathbf{y} | \mathbf{w} \sim \mathrm{N}_{nm}\left(\operatorname{vec}(\boldsymbol{\alpha} + \mathbf{B}_\eta \mathbf{w} \mathbf{A}), \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n\right). \tag{F.2}$$

By the results of Chapter 4, the posterior distribution of the I-prior random effects is $\operatorname{vec} \mathbf{w} | \mathbf{y} \sim \mathrm{N}(\operatorname{vec} \tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$, where

$$\operatorname{vec} \tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w (\boldsymbol{\Psi} \otimes \mathbf{H}_\eta) \operatorname{vec}(\mathbf{y} - \mathbf{1}_n \boldsymbol{\alpha}^\top) \quad \text{and} \quad \tilde{\mathbf{V}}_w^{-1} = \mathbf{A}\boldsymbol{\Psi}\mathbf{A} \otimes \mathbf{B}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n = \mathbf{V}_y. \tag{F.3}$$

Suppose hypothetically, one uses the uncentered identity kernel $a(j, j') = \delta_{jj'}$, in which case centring of the intercepts α_j must be handled separately. In conjunction with an assumption of iid errors ($\boldsymbol{\Psi} = \psi \mathbf{I}_n$), the above distributions simplify further. Specifically, the variance in the marginal distribution becomes

$$\begin{aligned}
\operatorname{Var}[\operatorname{vec} \mathbf{y}] &= (\psi \mathbf{I}_m \otimes \mathbf{B}_\eta^2) + (\psi^{-1} \mathbf{I}_m \otimes \mathbf{I}_n) \\
&= (\mathbf{I}_m \otimes \psi \mathbf{B}_\eta^2) + (\mathbf{I}_m \otimes \psi^{-1} \mathbf{I}_n) \\
&= \mathbf{I}_m \otimes (\overbrace{\psi \mathbf{B}_\eta^2 + \psi^{-1} \mathbf{I}_n}^{\tilde{\mathbf{V}}_y}).
\end{aligned}$$

which implies independence and identical variances $\tilde{\mathbf{V}}_y$ for the vectors $(y_{1j}, \ldots, y_{nj})^\top$ for each class $j = 1, \ldots, m$. Evidently, this stems from the implied independence structure of the prior on f too, since now $\operatorname{Var}(\operatorname{vec} \mathbf{f}) = \operatorname{diag}(\psi \mathbf{B}_\eta^2, \ldots, \psi \mathbf{B}_\eta^2)$, which could be interpreted as having independent and identical I-priors on the regression functions for each class $\mathbf{f}_{\cdot j} = \big(f(x_1, j), \ldots, f(x_n, j)\big)^\top$.

Appendix G

Posterior distribution of the I-prior regression function

We derive the posterior distribution for the I-prior random effects $\mathbf{w} = (w_1, \ldots, w_n)^\top$, which is related to the I-prior regression function via $f(x_i) = \sum_{k=1}^n h_\eta(x_i, x_k)w_k$, or in matrix terms, $\mathbf{f} := \big(f(x_1), \ldots, f(x_n)\big)^\top = \mathbf{H}_\eta \mathbf{w}$, and $f \in \mathcal{F}$ an RKHS with kernel h_η. A closely related distribution of interest is the posterior predictive distribution of y_{new}, the prediction at a new data point x_{new}. We note the similarity of these results with the posterior distributions of Gaussian process regressions (Rasmussen and Williams, 2006).

G.1 Deriving the posterior distribution for w

In the following derivation, we implicitly assume the dependence on $\mathbf{f}_0$ and θ. The distribution of $\mathbf{y}|\mathbf{w}$ is $\mathrm{N}_n(\boldsymbol{\alpha} + \mathbf{f}_0 + \mathbf{H}_\eta\mathbf{w}, \boldsymbol{\Psi}^{-1})$, where $\boldsymbol{\alpha} = \alpha\mathbf{1}_n$, while the prior distribution for $\mathbf{w}$ is $\mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi})$. Since $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$, we have that

$$
\begin{aligned}
\log p(\mathbf{w}|\mathbf{y}) &= \log p(\mathbf{y}|\mathbf{w}) + \log p(\mathbf{w}) \\
&= \text{const.} + \cancel{\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}(\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0 - \mathbf{H}_\eta\mathbf{w})^\top\boldsymbol{\Psi}(\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0 - \mathbf{H}_\eta\mathbf{w}) \\
&\quad - \cancel{\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}\mathbf{w}^\top\boldsymbol{\Psi}^{-1}\mathbf{w} \\
&= \text{const.} - \frac{1}{2}\mathbf{w}^\top(\mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1})\mathbf{w} + (\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0)^\top\boldsymbol{\Psi}\mathbf{H}_\eta\mathbf{w}.
\end{aligned}
$$

Setting $\mathbf{A} = \mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}$, $\mathbf{a}^\top = (\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0)^\top\boldsymbol{\Psi}\mathbf{H}_\eta$, and using the fact that

$$
\mathbf{w}^\top\mathbf{A}\mathbf{w} - 2\mathbf{a}^\top\mathbf{w} = (\mathbf{w} - \mathbf{A}^{-1}\mathbf{a})^\top\mathbf{A}(\mathbf{w} - \mathbf{A}^{-1}\mathbf{a}),
$$

we have that $\mathbf{w}|\mathbf{y}$ is normally distributed with the required mean and variance.

Alternatively, one could have shown this using standard results of multivariate normal distributions. Noting that the covariance between $\mathbf{y}$ and $\mathbf{w}$ is

$$\begin{aligned}\operatorname{Cov}[\mathbf{y}, \mathbf{w}] &= \operatorname{Cov}[\boldsymbol{\alpha} + \mathbf{f}_0 + \mathbf{H}_\eta \mathbf{w} + \boldsymbol{\epsilon}, \mathbf{w}] \\ &= \mathbf{H}_\eta \operatorname{Cov}[\mathbf{w}, \mathbf{w}] \\ &= \mathbf{H}_\eta \boldsymbol{\Psi}\end{aligned}$$

and that $\operatorname{Cov}[\mathbf{w}, \mathbf{y}] = \boldsymbol{\Psi}\mathbf{H}_\eta = \mathbf{H}_\eta\boldsymbol{\Psi} = \operatorname{Cov}[\mathbf{y}, \mathbf{w}]$ by symmetry, the joint distribution $(\mathbf{y}, \mathbf{w})$ is

$$\begin{pmatrix}\mathbf{y} \\ \mathbf{w}\end{pmatrix} \sim \mathrm{N}_{n+n}\left(\begin{pmatrix}\boldsymbol{\alpha} + \mathbf{f}_0 \\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{V}_y & \mathbf{H}_\eta\boldsymbol{\Psi} \\ \boldsymbol{\Psi}\mathbf{H}_\eta & \boldsymbol{\Psi}\end{pmatrix}\right).$$

Thus,

$$\begin{aligned}\mathrm{E}[\mathbf{w}|\mathbf{y}] &= \mathrm{E}\,\mathbf{w} + \operatorname{Cov}(\mathbf{w}, \mathbf{y})(\operatorname{Var}\mathbf{y})^{-1}(\mathbf{y} - \mathrm{E}\,\mathbf{y}) \\ &= \boldsymbol{\Psi}\mathbf{H}_\eta\mathbf{V}_y^{-1}(\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0),\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}[\mathbf{w}|\mathbf{y}] &= \operatorname{Var}\mathbf{w} - \operatorname{Cov}(\mathbf{w}, \mathbf{y})(\operatorname{Var}\mathbf{y})^{-1}\operatorname{Cov}(\mathbf{y}, \mathbf{w}) \\ &= \boldsymbol{\Psi} - \mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{V}_y^{-1}\mathbf{H}_\eta\boldsymbol{\Psi} \\ &= \boldsymbol{\Psi} - \boldsymbol{\Psi}\mathbf{H}_\eta\left(\boldsymbol{\Psi}^{-1} + \mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta\right)^{-1}\mathbf{H}_\eta\boldsymbol{\Psi} \\ &= \left(\boldsymbol{\Psi}^{-1} + \mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta\right)^{-1} \\ &= \mathbf{V}_y^{-1}\end{aligned}$$

as a direct consequence of the Woodbury matrix identity (Petersen and Pedersen, 2012, eq. 156, §3.2.2).

G.2 Deriving the posterior predictive distribution

The posterior predictive distribution is obtained in an empirical Bayesian manner, in which the parameters of the model are replaced with their ML estimates (denoted with hats).

A priori, assume that $y_{\text{new}} \sim \mathrm{N}(\hat{\alpha}, v_{\text{new}})$, where $v_{\text{new}} = \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top\hat{\boldsymbol{\Psi}}\mathbf{h}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1}$. Consider the joint distribution of $(y_{\text{new}}, \mathbf{y}^\top)^\top$, which is multivariate normal (since both y_{new} and $\mathbf{y}$ are. Write

$$\begin{pmatrix}y_{\text{new}} \\ \mathbf{y}\end{pmatrix} \sim \mathrm{N}_{n+1}\left(\begin{pmatrix}\hat{\alpha} \\ \hat{\alpha}\mathbf{1}_n\end{pmatrix}, \begin{pmatrix}v_{\text{new}} & \operatorname{Cov}[y_{\text{new}}, \mathbf{y}] \\ \operatorname{Cov}[y_{\text{new}}, \mathbf{y}]^\top & \hat{\mathbf{V}}_y\end{pmatrix}\right),$$

where

$$\begin{aligned}
\mathrm{Cov}[y_{\text{new}}, \mathbf{y}] &= \mathrm{Cov}[f_{\text{new}} + \epsilon_{\text{new}}, \mathbf{f} + \boldsymbol{\epsilon}] \\
&= \mathrm{Cov}[f_{\text{new}}, \mathbf{f}) + \mathrm{Cov}(\epsilon_{\text{new}}, \boldsymbol{\epsilon}] \\
&= \mathrm{Cov}\left[\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \tilde{\mathbf{w}}, \mathbf{H}_{\hat{\eta}} \tilde{\mathbf{w}}\right] + (\sigma_{\text{new},1}, \ldots, \sigma_{\text{new},n}) \\
&= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} + \boldsymbol{\sigma}_{\text{new}}.
\end{aligned}$$

The vector of covariances $\boldsymbol{\sigma}_{\text{new}}$ between observations $y_1, \ldots, y_n$ and the predicted point y_{new} would need to be prescribed a priori (treated as extra parameters), or estimated again, which seems excessive. Under an iid assumption of the error precisions, then $\boldsymbol{\sigma}_{\text{new}} = \mathbf{0}$ would be acceptable.

In any case, using standard multivariate normal results, we get that $y_{\text{new}}|\mathbf{y}$ is also normally distributed with mean

$$\begin{aligned}
E[y_{\text{new}}|\mathbf{y}] &= \hat{\alpha} + (\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} + \boldsymbol{\sigma}_{\text{new}}) \hat{\mathbf{V}}_y^{-1} \tilde{\mathbf{y}} \\
&= \hat{\alpha} + \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \overbrace{\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} \hat{\mathbf{V}}_y^{-1} \tilde{\mathbf{y}}}^{\hat{\mathbf{w}}} + \boldsymbol{\sigma}_{\text{new}} \hat{\mathbf{V}}_y^{-1} \tilde{\mathbf{y}} \\
&= \hat{\alpha} + \mathrm{E}\left[f(x_{\text{new}})|\mathbf{y}\right] + \text{mean correction term}
\end{aligned}$$

and variance

$$\begin{aligned}
\mathrm{Var}[y_{\text{new}}|\mathbf{y}] &= v_{\text{new}} - (\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} + \boldsymbol{\sigma}_{\text{new}}) \hat{\mathbf{V}}_y^{-1} (\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} + \boldsymbol{\sigma}_{\text{new}})^\top \\
&= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \hat{\mathbf{h}}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1} - \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} \hat{\mathbf{V}}_y^{-1} \mathbf{H}_{\hat{\eta}} \hat{\boldsymbol{\Psi}} \mathbf{h}_{\hat{\eta}}(x_{\text{new}}) \\
&\quad + \text{variance correction term} \\
&= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \left(\hat{\boldsymbol{\Psi}} - \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} \hat{\mathbf{V}}_y^{-1} \mathbf{H}_{\hat{\eta}} \hat{\boldsymbol{\Psi}}\right) \mathbf{h}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1} \\
&\quad + \text{variance correction term} \\
&= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\mathbf{V}}_y^{-1} \mathbf{h}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1} + \text{variance correction term} \\
&= \mathrm{Var}\left[f(x_{\text{new}})|\mathbf{y}\right] + \psi_{\text{new}}^{-1} + \text{variance correction term.}
\end{aligned}$$

Appendix H

Variational EM algorithm for I-probit models

The two sections that follow detail the derivation of the variational densities used in the E-step of the variational EM algorithm, and also the lower bound (ELBO) used to monitor convergence.

H.1 Derivation of the variational densities

In what follows, the implicit dependence of the densities on the parameters of the model θ are dropped. We derive a mean-field variational approximation of

$$\begin{aligned} p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}) &\approx q(\mathbf{y}^*)q(\mathbf{w}) \\ &= \prod_{i=1}^{n} q(\mathbf{y}_i^*)q(\mathbf{w}). \end{aligned}$$

The first line is by assumption, while the second line follows from an induced factorisation on the latent propensities, as we will see later. Recall that the optimal mean-field variational density $\tilde{q}$ satisfy

$$\log \tilde{q}(\mathbf{y}^*) = \mathrm{E}_{\mathbf{w}\sim\tilde{q}}\left[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w})\right] + \text{const.} \qquad \text{(from 5.13)}$$

$$\log \tilde{q}(\mathbf{w}) = \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}\left[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w})\right] + \text{const.} \qquad \text{(from 5.14)}$$

The joint likelihood is given by

$$p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}) = p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\mathbf{w})p(\mathbf{w}).$$

For reference, the three relevant distributions are listed below.

- $\boldsymbol{p(\mathbf{y}|\mathbf{y}^*)}$. For each observation $i \in \{1, \ldots, n\}$, given the corresponding latent propensities $\mathbf{y}_i^* = (y_{i1}^*, \ldots, y_{im}^*)$, the distribution for y_i is a degenerate distribution which depends on the j'th component of $\mathbf{y}_i^*$ being largest, where the value observed for y_i was j. Since each

of the y_i's are independent, everything is multiplicative.

$$p(\mathbf{y}|\mathbf{y}^*) = \prod_{i=1}^{n}\prod_{j=1}^{m} p_{ij}^{[y_i=j]} = \prod_{i=1}^{n}\prod_{j=1}^{m} \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*]^{\mathbb{1}[y_i=j]}.$$

- $\boldsymbol{p(\mathbf{y}^*|\mathbf{w})}$. Given values for the parameters and I-prior random effects, the distribution of the latent propensities is matrix normal

$$\mathbf{y}^*|\mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1}).$$

Write $\boldsymbol{\mu} = \mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}$. Its pdf is

$$\begin{aligned} p(\mathbf{y}^*|\mathbf{w}) &= \exp\left[-\frac{nm}{2}\log 2\pi + \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\operatorname{tr}\left((\mathbf{y}^*-\boldsymbol{\mu})\boldsymbol{\Psi}(\mathbf{y}^*-\boldsymbol{\mu})^\top\right)\right] \\ &= \exp\left[-\frac{nm}{2}\log 2\pi + \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_{i\cdot}^*-\boldsymbol{\mu}_{i\cdot})^\top\boldsymbol{\Psi}(\mathbf{y}_{i\cdot}^*-\boldsymbol{\mu}_{i\cdot})\right], \end{aligned}$$

where $\mathbf{y}_{i\cdot}^* \in \mathbb{R}^m$ and $\boldsymbol{\mu}_{i\cdot} \in \mathbb{R}^m$ are the rows of $\mathbf{y}^*$ and $\boldsymbol{\mu}$ respectively. The second line follows directly from the definition of the trace, but also emanates from the fact that $\mathbf{y}_{i\cdot}^*$ are independent multivariate normal with mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\Psi}^{-1}$.

- $\boldsymbol{p(\mathbf{w})}$. The $\mathbf{w}$'s are normal random matrices $\mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$ with pdf

$$\begin{aligned} p(\mathbf{w}) &= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\operatorname{tr}\left(\mathbf{w}\boldsymbol{\Psi}^{-1}\mathbf{w}^\top\right)\right] \\ &= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}\mathbf{w}_{i\cdot}^\top\boldsymbol{\Psi}^{-1}\mathbf{w}_{i\cdot}\right]. \end{aligned}$$

H.1.1 Derivation of $\tilde{q}(\mathbf{y}^*)$

The rows of $\mathbf{y}^*$ are independent, and thus we can consider the variational density for each $\mathbf{y}_i^*$ separately. Consider the case where y_i takes one particular value $j \in \{1, \dots, m\}$. In such cases, we have that $y_{ij}^* > y_{ik}$ for all $k \neq j$, and that

$$\begin{aligned} \log\tilde{q}(\mathbf{y}_{i\cdot}^*) &= \mathrm{E}_{\mathbf{w}\sim\tilde{q}}\left[-\frac{1}{2}(\mathbf{y}_i^*-\boldsymbol{\mu}_i)^\top\boldsymbol{\Psi}(\mathbf{y}_i^*-\boldsymbol{\mu}_i)\right] + \text{const.} \\ &= \left[-\frac{1}{2}(\mathbf{y}_i^*-\tilde{\boldsymbol{\mu}}_i)^\top\boldsymbol{\Psi}(\mathbf{y}_i^*-\tilde{\boldsymbol{\mu}}_i)\right] + \text{const.} \qquad (\star) \end{aligned}$$

where $\tilde{\boldsymbol{\mu}}_{i\cdot} = \boldsymbol{\alpha} + \tilde{\mathbf{w}}\mathbf{h}_\eta(x_i)$, $\tilde{\mathbf{w}} = \mathrm{E}_{\mathbf{w}\sim\tilde{q}}[\mathbf{w}]$. This is recognised as the logarithm of a multivariate normal pdf with mean $\tilde{\boldsymbol{\mu}}_{i\cdot}$ and variance $\boldsymbol{\Psi}^{-1}$. On the other hand, when $y_i \neq j$, the pdf is zero. Thus,

$$\tilde{q}(\mathbf{y}_{i\cdot}^*) = \begin{cases} \phi(\mathbf{y}_{i\cdot}^*|\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}) & \text{if } y_{ij}^* > y_{ik}^*, \forall k \neq j \\ 0 & \text{otherwise,} \end{cases}$$

implying a truncated multivariate normal distribution for $\mathbf{y}^*_{i\cdot}$. The required moments from the truncated multivariate normal distribution can be obtained using the methods described in Appendix C.4 (p. 268).

Remark H.1. In the above derivation, we needn't consider the second order terms in the expectations because they do not involve $\mathbf{y}^*_{i\cdot}$, and thus, these terms can be absorbed into the constant. To see this,

$$\begin{aligned}
\mathrm{E}[(\mathbf{y}^*_{i\cdot} - \boldsymbol{\mu}_{i\cdot})^\top \boldsymbol{\Psi}(\mathbf{y}^*_{i\cdot} - \boldsymbol{\mu}_{i\cdot})] &= \mathrm{E}[\mathbf{y}^{*\top}_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} + \boldsymbol{\mu}^\top_{i\cdot}\boldsymbol{\Psi}\boldsymbol{\mu}_{i\cdot} - 2\boldsymbol{\mu}^\top_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot}] \\
&= \mathbf{y}^{*\top}_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} - 2\,\mathrm{E}[\boldsymbol{\mu}^\top_{i\cdot}]\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} + \text{const.} \\
&= \mathbf{y}^{*\top}_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} - 2\tilde{\boldsymbol{\mu}}^\top_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} + \text{const.} \\
&= (\mathbf{y}^*_{i\cdot} - \tilde{\boldsymbol{\mu}}_{i\cdot})^\top \boldsymbol{\Psi}(\mathbf{y}^*_{i\cdot} - \tilde{\boldsymbol{\mu}}_{i\cdot}) + \text{const.}
\end{aligned}$$

The square is then completed to get the final line, which is the expression for the term $(\star)$ multiplied by a half.

H.1.2 Derivation of $\tilde{q}(\mathbf{w})$

The terms involving $\mathbf{w}$ in the joint likelihood (5.14) are the $p(\mathbf{y}^*|\mathbf{w})$ and $p(\mathbf{w})$ terms, so the rest are absorbed into the constant. The easiest way to derive $\tilde{q}(\mathbf{w})$ is to vectorise $\mathbf{y}^*$ and $\mathbf{w}$. We know that

$$\begin{gathered}
\operatorname{vec}\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi} \sim \mathrm{N}_{nm}\left(\operatorname{vec}(\mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}), \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n\right) \\
\text{and} \\
\operatorname{vec}\mathbf{w}|\boldsymbol{\Psi} \sim \mathrm{N}_{nm}(\mathbf{0}, \boldsymbol{\Psi} \otimes \mathbf{I}_n)
\end{gathered}$$

using properties of matrix normal distributions.

We also use the fact that $\operatorname{vec}(\mathbf{H}_\eta\mathbf{w}) = (\mathbf{I}_m \otimes \mathbf{H}_\eta)\operatorname{vec}\mathbf{w}$. For simplicity, write $\bar{\mathbf{y}}^* = \operatorname{vec}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)$, and $\mathbf{M} = (\mathbf{I}_m \otimes \mathbf{H}_\eta)$. Thus,

$$\begin{aligned}
\log\tilde{q}(\mathbf{w}) = {} & \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}\left[-\frac{1}{2}(\bar{\mathbf{y}}^* - \mathbf{M}\operatorname{vec}\mathbf{w})^\top(\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)^{-1}(\bar{\mathbf{y}}^* - \mathbf{M}\operatorname{vec}\mathbf{w})\right] \\
& + \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}\left[-\frac{1}{2}(\operatorname{vec}\mathbf{w})^\top(\boldsymbol{\Psi} \otimes \mathbf{I}_n)^{-1}\operatorname{vec}(\mathbf{w})\right] + \text{const.} \\
= {} & -\frac{1}{2}\mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}\left[(\operatorname{vec}\mathbf{w})^\top\Big(\overbrace{\mathbf{M}^\top(\boldsymbol{\Psi} \otimes \mathbf{I}_n)\mathbf{M} + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)}^{\mathbf{A}}\Big)\operatorname{vec}(\mathbf{w})\right] \\
& + \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}\left[\overbrace{\bar{\mathbf{y}}^{*\top}(\boldsymbol{\Psi} \otimes \mathbf{I}_n)\mathbf{M}}^{\mathbf{a}^\top}\operatorname{vec}(\mathbf{w})\right] + \text{const.} \\
= {} & -\frac{1}{2}\mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}\left[(\operatorname{vec}\mathbf{w} - \mathbf{A}^{-1}\mathbf{a})^\top\mathbf{A}(\operatorname{vec}\mathbf{w} - \mathbf{A}^{-1}\mathbf{a})\right] + \text{const.}
\end{aligned}$$

This is recognised as a multivariate normal of dimension nm with mean and precision given by $\operatorname{vec}\tilde{\mathbf{w}} = \mathrm{E}[\mathbf{A}^{-1}\mathbf{a}]$ and $\tilde{\mathbf{V}}_w^{-1} = \mathrm{E}[\mathbf{A}]$ respectively. With a little algebra, we find that

$$\begin{aligned}\tilde{\mathbf{V}}_w &= \left\{ \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}[\mathbf{A}] \right\}^{-1} \\ &= \left\{ \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}} \left[(\mathbf{I}_m \otimes \mathbf{H}_\eta)^\top (\boldsymbol{\Psi} \otimes \mathbf{I}_n)(\mathbf{I}_m \otimes \mathbf{H}_\eta) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \right] \right\}^{-1} \\ &= \left(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n \right)^{-1}\end{aligned}$$

and

$$\begin{aligned}\operatorname{vec}\tilde{\mathbf{w}} &= \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}[\mathbf{A}^{-1}\mathbf{a}] \\ &= \tilde{\mathbf{V}}_w \, \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}} \left[(\mathbf{I}_m \otimes \mathbf{H}_\eta)(\boldsymbol{\Psi} \otimes \mathbf{I}_n) \operatorname{vec}(\mathbf{y}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top) \right] \\ &= \tilde{\mathbf{V}}_w (\boldsymbol{\Psi} \otimes \mathbf{H}_\eta) \, \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}} \left[\operatorname{vec}(\mathbf{y}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top) \right] \\ &= \tilde{\mathbf{V}}_w (\boldsymbol{\Psi} \otimes \mathbf{H}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top).\end{aligned}$$

We will often refer to $\tilde{\mathbf{w}}$ as the $n \times m$ matrix constructed by filling in its entries with $\operatorname{vec}\tilde{\mathbf{w}}$ column-wise (akin to the opposite of vectorisation). This way, the $\tilde{\mathbf{w}}$ contains posterior mean values arranged by class $j = 1, \ldots, m$ column-wise, and by observations $i = 1, \ldots, n$ row-wise. Ideally, we do not want to work with the $nm \times nm$ matrix $\mathbf{V}_w$, since its inverse is expensive to compute. Refer to Section 5.6.2 for details.

In the case of the I-probit model, where $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \ldots, \psi_m)$, then the covariance matrix takes a simpler form. Specifically, it has the block diagonal structure:

$$\begin{aligned}\tilde{\mathbf{V}}_w &= \left(\operatorname{diag}(\psi_1, \ldots, \psi_m) \otimes \mathbf{H}_\eta^2 + \operatorname{diag}(\psi_1, \ldots, \psi_m) \otimes \mathbf{I}_n \right)^{-1} \\ &= \operatorname{diag}\left(\left(\psi_1 \mathbf{H}_\eta^2 + \psi_1^{-1} \mathbf{I}_n \right)^{-1}, \cdots, \left(\psi_m \mathbf{H}_\eta^2 + \psi_m^{-1} \mathbf{I}_n \right)^{-1} \right) \\ &=: \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \ldots, \tilde{\mathbf{V}}_{w_m}).\end{aligned}$$

The mean $\operatorname{vec}\tilde{\mathbf{w}}$ is

$$\begin{aligned}\operatorname{vec}\tilde{\mathbf{w}} &= \tilde{\mathbf{V}}_w (\operatorname{diag}(\psi_1, \ldots, \psi_m) \otimes \tilde{\mathbf{H}}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top) \\ &= \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \ldots, \tilde{\mathbf{V}}_{w_m}) \operatorname{diag}(\psi_1 \mathbf{H}_\eta, \ldots, \psi_m \mathbf{H}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top) \\ &= \operatorname{diag}(\psi_1 \tilde{\mathbf{V}}_{w_1} \mathbf{H}_\eta, \ldots, \psi_m \tilde{\mathbf{V}}_{w_m} \mathbf{H}_\eta)(\tilde{\mathbf{y}}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top) \\ &\qquad\qquad \tilde{\mathbf{w}}_{\cdot 1}^\top \qquad\qquad \cdots \qquad\qquad \tilde{\mathbf{w}}_{\cdot m}^\top \\ &= \begin{pmatrix} \left(\psi_1 \tilde{\mathbf{V}}_{w_1} \mathbf{H}_\eta (\tilde{\mathbf{y}}^*_{\cdot 1} - \alpha_1 \mathbf{1}_n) \right)^\top & \cdots & \left(\psi_m \tilde{\mathbf{V}}_{w_m} \mathbf{H}_\eta (\tilde{\mathbf{y}}^*_{\cdot m} - \alpha_m \mathbf{1}_n) \right)^\top \end{pmatrix}^\top.\end{aligned}$$

Therefore, we can consider the distribution of $\mathbf{w} = (\mathbf{w}_{\cdot 1}, \ldots, \mathbf{w}_{\cdot m})$ columnwise, and each are normally distributed with mean and variance

$$\tilde{\mathbf{w}}_{\cdot j} = \psi_j \tilde{\mathbf{V}}_{w_j} \mathbf{H}_\eta (\tilde{\mathbf{y}}^*_{\cdot j} - \alpha_j \mathbf{1}_n) \quad \text{and} \quad \tilde{\mathbf{V}}_{w_j} = \left(\psi_j \mathbf{H}_\eta^2 + \psi_j^{-1} \mathbf{I}_n \right)^{-1}.$$

A quantity that we will be requiring time and again will be $\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}])$, where $\mathbf{C} \in \mathbb{R}^{m \times m}$ and $\mathbf{D} \in \mathbb{R}^{n \times n}$ are both square and symmetric matrices. Using the definition of the trace directly,

we get

$$\begin{aligned}\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]) &= \sum_{i,j=1}^{m}\mathbf{C}_{ij}\,\mathrm{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]_{ij}\\ &= \sum_{i,j=1}^{m}\mathbf{C}_{ij}\,\mathrm{E}[\mathbf{w}_{\bullet i}^\top\mathbf{D}\mathbf{w}_{\bullet j}].\end{aligned} \tag{H.1}$$

The expectation of the univariate quantity $\mathbf{w}_{\bullet i}^\top\mathbf{D}\mathbf{w}_{\bullet j}$ is inspected below:

$$\begin{aligned}\mathrm{E}[\mathbf{w}_{\bullet i}^\top\mathbf{D}\mathbf{w}_{\bullet j}] &= \operatorname{tr}(\mathbf{D}\,\mathrm{E}[\mathbf{w}_{\bullet j}\mathbf{w}_{\bullet i}^\top])\\ &= \operatorname{tr}\left(\mathbf{D}(\operatorname{Cov}(\mathbf{w}_{\bullet j},\mathbf{w}_{\bullet i})+\mathrm{E}[\mathbf{w}_{\bullet j}]\,\mathrm{E}[\mathbf{w}_{\bullet i}]^\top)\right)\\ &= \operatorname{tr}\left(\mathbf{D}(\mathbf{V}_w[i,j]+\tilde{\mathbf{w}}_{\bullet j}\tilde{\mathbf{w}}_{\bullet i}^\top)\right).\end{aligned}$$

where $\mathbf{V}_w[i,j]\in\mathbb{R}^{n\times n}$ refers to the (i,j)'th submatrix block of $\mathbf{V}_w$. Of course, in the independent the I-probit model, this is equal to

$$\mathbf{V}_w[i,j] = \delta_{ij}(\psi_j\mathbf{H}_\eta^2+\psi_j^{-1}\mathbf{I}_n)^{-1}$$

where δ is the Kronecker delta. Continuing on (H.1) leads us to

$$\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]) = \sum_{i,j=1}^{m}\mathbf{C}_{ij}\left(\operatorname{tr}\left(\mathbf{D}(\delta_{ij}\mathbf{V}_{w_j}+\tilde{\mathbf{w}}_{\bullet j}\tilde{\mathbf{w}}_{\bullet i}^\top)\right).\right).$$

If $\mathbf{C}=\operatorname{diag}(c_1,\ldots,c_m)$, then

$$\begin{aligned}\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]) &= \sum_{j=1}^{m}c_j\left(\operatorname{tr}\left(\mathbf{D}\tilde{\mathbf{V}}_{w_j}\right)+\tilde{\mathbf{w}}_{\bullet j}^\top\mathbf{D}\tilde{\mathbf{w}}_{\bullet j}\right)\\ &= \sum_{j=1}^{m}c_j\operatorname{tr}\left(\mathbf{D}(\tilde{\mathbf{V}}_{w_j}+\tilde{\mathbf{w}}_{\bullet j}\tilde{\mathbf{w}}_{\bullet j}^\top)\right).\end{aligned}$$

H.2 Deriving the ELBO expression

The evidence lower bound (ELBO) expression involves the following calculation:

$$\begin{aligned}\mathcal{L}_q(\theta) &= \int\cdots\int q(\mathbf{y}^*,\mathbf{w})\log\frac{p(\mathbf{y},\mathbf{y}^*,\mathbf{w}|\theta)}{q(\mathbf{y}^*,\mathbf{w})}\,\mathrm{d}\mathbf{y}^*\,\mathrm{d}\mathbf{w}\,\mathrm{d}\theta\\ &= \mathrm{E}\left[\log\overbrace{p(\mathbf{y},\mathbf{y}^*,\mathbf{w}|\theta)}^{\text{joint likelihood}}\right]+\overbrace{(-\mathrm{E}\left[\log q(\mathbf{y}^*,\mathbf{w},)\right])}^{\text{entropy}}\\ &= \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\cancel{\log p(y_i|y_{ij}^*)}+\sum_{i=1}^{n}\log p(\mathbf{y}_{i\bullet}^*|\boldsymbol{\alpha},\mathbf{w},\boldsymbol{\Psi},\eta)+\log p(\mathbf{w}|\boldsymbol{\Psi})\right]\\ &\quad+\sum_{i=1}^{n}H\left[q(\mathbf{y}_{i\bullet}^*)\right]+H\left[q(\mathbf{w})\right].\end{aligned}$$

As discussed, given the latent propensities $\mathbf{y}^*$, the pdf of $\mathbf{y}$ is degenerate and hence can be disregarded.

H.2.1 Terms involving distributions of $\mathbf{y}^*$

$$\begin{aligned}
&\sum_{i=1}^{n}\Big\{\mathrm{E}\left[\log p(\mathbf{y}_{i\cdot}^*|\boldsymbol{\alpha},\mathbf{w},\boldsymbol{\Psi},\eta)\right]+H\left[q(\mathbf{y}_{i\cdot}^*)\right]\Big\}\\
&\quad=-\frac{nm}{2}\log 2\pi+\frac{n}{2}\log|\boldsymbol{\Psi}|-\frac{1}{2}\,\mathrm{E}\left[\sum_{i=1}^{n}(\mathbf{y}_{i\cdot}^*-\tilde{\boldsymbol{\mu}}_{i\cdot})^\top\boldsymbol{\Psi}(\mathbf{y}_{i\cdot}^*-\tilde{\boldsymbol{\mu}}_{i\cdot})\right]\\
&\qquad+\frac{nm}{2}\log 2\pi-\frac{n}{2}\log|\boldsymbol{\Psi}|+\frac{1}{2}\,\mathrm{E}\left[\sum_{i=1}^{n}(\mathbf{y}_{i\cdot}^*-\tilde{\boldsymbol{\mu}}_{i\cdot})^\top\boldsymbol{\Psi}(\mathbf{y}_{i\cdot}^*-\tilde{\boldsymbol{\mu}}_{i\cdot})\right]+\log C_i\\
&\quad=\sum_{i=1}^{n}\log C_i
\end{aligned}$$

where C_i is the normalising constant for the distribution of multivariate truncated normal $\mathbf{y}_{i\cdot}^*\sim{}^{\mathrm{t}}\mathrm{N}(\tilde{\boldsymbol{\mu}}(x_i),\boldsymbol{\Psi}^{-1},\mathcal{C}_{y_i})$, with $\tilde{\boldsymbol{\mu}}(x_i)=\boldsymbol{\alpha}+\tilde{\mathbf{w}}\mathbf{h}_\eta(x_i)$.

H.2.2 Terms involving distributions of $\mathbf{w}$

$$\begin{aligned}
\mathrm{E}\log p(\mathbf{w}|\boldsymbol{\Psi})+H\left[q(\mathbf{w})\right]&=\cancel{-\frac{nm}{2}\log 2\pi}-\frac{n}{2}\log|\boldsymbol{\Psi}|-\frac{1}{2}\,\mathrm{E}\,\mathrm{tr}\left(\mathbf{w}\boldsymbol{\Psi}^{-1}\mathbf{w}^\top\right)\\
&\quad+\frac{nm}{2}(1+\cancel{\log 2\pi})+\frac{1}{2}\log|\tilde{\mathbf{V}}_w|\\
&=\frac{nm}{2}-\frac{n}{2}\log|\boldsymbol{\Psi}|-\frac{1}{2}\sum_{i,j=1}^{m}\boldsymbol{\Psi}_{ij}^{-1}\,\mathrm{tr}\,\mathrm{E}\left[\tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot j}^\top\right]+\frac{1}{2}\log|\tilde{\mathbf{V}}_w|
\end{aligned}$$

Appendix I

The Gibbs sampler for the I-prior Bayesian variable selection model

The I-prior Bayesian variable selection model has the following hierarchical form:

$$\begin{aligned}
\mathbf{y}|\alpha, \boldsymbol{\beta}, \gamma, \sigma^2, \kappa &\sim \mathrm{N}_n(\alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\theta}, \sigma^2 I_n) \\
\boldsymbol{\theta} &= (\gamma_1\beta_1, \ldots, \gamma_p\beta_p)^\top \\
\boldsymbol{\beta}|\sigma^2, \kappa &\sim \mathrm{N}_p(\mathbf{0}, \sigma^2\kappa\mathbf{X}^\top\mathbf{X}) \\
\alpha|\sigma^2 &\sim \mathrm{N}(0, \sigma^2 A) \\
\sigma^2, \kappa &\sim \Gamma^{-1}(c, d) \\
\gamma_j &\sim \mathrm{Bern}(\pi_j) \qquad j = 1, \ldots, p
\end{aligned}$$

In the simulations and real-data examples, we used $\pi_j = 0.5, \forall j$, $A = 100$, and $c = d = 0.001$, and the columns of the matrix $\mathbf{X}$ are standardised.

The first line of the set of equations above is the likelihood, while the joint prior density is given by

$$p(\alpha, \beta, \gamma, \sigma^2, \kappa) = p(\beta|\sigma^2)p(\alpha|\sigma^2)p(\sigma^2)p(\kappa)p(\gamma_1)\cdots p(\gamma_p).$$

For simplicity, in the following subsections we shall denote by Θ the entire set of parameters, while $\Theta_{-\xi}$ implies the set of parameters excluding the parameter ξ.

I.1 Conditional posterior for $\boldsymbol{\beta}$

$$\begin{aligned}
\log p(\boldsymbol{\beta}|\mathbf{y}, \Theta_{-\boldsymbol{\beta}}) &= \text{const.} + \log p(\mathbf{y}|\Theta) + \log p(\boldsymbol{\beta}|\sigma^2) \\
&= \text{const.} - \frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}_\gamma\boldsymbol{\beta}\|^2 - \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top(\kappa\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta} \\
&= \text{const.} - \frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma + (\kappa\mathbf{X}^\top\mathbf{X})^{-1})\boldsymbol{\beta} - 2(\mathbf{y} - \alpha\mathbf{1}_n)^\top\mathbf{X}_\gamma\boldsymbol{\beta}\right) \\
&= \text{const.} - \frac{1}{2\sigma^2}\big(\boldsymbol{\beta} - \tilde{\mathbf{B}}(\mathbf{y} - \alpha\mathbf{1}_n)\big)^\top\tilde{\mathbf{B}}^{-1}\big(\boldsymbol{\beta} - \tilde{\mathbf{B}}(\mathbf{y} - \alpha\mathbf{1}_n)\big)
\end{aligned}$$

where $\tilde{\mathbf{B}} = \mathbf{X}_\gamma^\top\mathbf{X}_\gamma + (\kappa\mathbf{X}^\top\mathbf{X})^{-1}$, and $\mathbf{X}_\gamma = (\gamma_1 X_1 \cdots \gamma_p X_p)$ is the $n \times p$ design matrix $\mathbf{X}$ with each of the p columns multiplied by the indicator variable γ. This is of course recognised as the log density of a p-variate normal distribution with mean and variance

$$\mathrm{E}[\boldsymbol{\beta}|\Theta_{-\boldsymbol{\beta}}] = \tilde{\mathbf{B}}(y - \alpha\mathbf{1}_n) \text{ and } \mathrm{Var}[\boldsymbol{\beta}|\Theta_{-\boldsymbol{\beta}}] = \sigma^2\tilde{\mathbf{B}}.$$

I.2 Conditional posterior for γ

Consider each γ_j in turn. For $j \in \{1, \ldots, p\}$,

$$\begin{aligned}
p(\gamma_j|\mathbf{y}, \Theta_{-\gamma_j}) &\propto p(\mathbf{y}|\Theta)p(\gamma_j) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2\right)\pi_j^{\gamma_j}(1 - \pi_j)^{1-\gamma_j}
\end{aligned}$$

Since the support of γ_j is $\{0, 1\}$, the above is a probability mass function which can be normalised easily. When $\gamma_j = 1$, we have

$$p(\gamma_j|\mathbf{y}, \Theta_{-\gamma_j}) \propto \pi_j \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}_j^{[1]}\|^2\right) := u_j$$

while for $\gamma_j = 0$, we have

$$p(\gamma_j|\mathbf{y}, \Theta_{-\gamma_j}) \propto (1 - \pi_j) \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}_j^{[0]}\|^2\right) := v_j.$$

For $j = 1, \ldots, p$, we have used the notation $\theta_j^{[\omega]}$ to mean

$$\boldsymbol{\theta}_j^{[\omega]} = \begin{cases} (\theta_1, \ldots, \theta_{j-1}, \beta_j, \theta_{j+1}, \ldots, \theta_p) & \omega = 1 \\ (\theta_1, \ldots, \theta_{j-1}, 0, \theta_{j+1}, \ldots, \theta_p) & \omega = 0. \end{cases}$$

Therefore, the conditions distribution for γ_j is Bernoulli with success probability

$$\tilde{\pi}_j = \frac{u_j}{u_j + v_j}.$$

I.3 Conditional posterior for α

We can obtain the conditional posterior for α in a similar fashion we obtained the conditional posterior for $\boldsymbol{\beta}$. That is,

$$\begin{aligned}
\log p(\alpha|\mathbf{y}, \Theta_{-\alpha}) &= \text{const.} + \log p(\mathbf{y}|\Theta) + \log p(\alpha|\sigma^2) \\
&= \text{const.} - \frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2 - \frac{\alpha^2}{2\sigma^2 A} \\
&= \text{const.} - \frac{1}{2\sigma^2}\left((n + A^{-1})\alpha^2 - 2\alpha\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\theta})\right) \\
&= \text{const.} - \frac{1}{2\sigma^2(n + A^{-1})}\left(\alpha - \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\theta})}{n + A^{-1}}\right)^2.
\end{aligned}$$

Thus, the conditional posterior for α is normal with mean and variance which can be easily read off the final line above.

I.4 Conditional posterior for σ^2

The conditional density for σ^2 is

$$\begin{aligned}
\log p(\sigma^2|\mathbf{y}, \Theta_{-\sigma^2}) &= \text{const.} + \log p(\mathbf{y}|\Theta) + \log p(\sigma^2) \\
&= \text{const.} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2 - (c+1)\log\sigma^2 - d/\sigma^2 \\
&= \text{const.} - (n/2 + c + 1)\log\sigma^2 - \frac{\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2/2 + d}{\sigma^2}
\end{aligned}$$

which is an inverse gamma distribution with shape $\tilde{c} = n/2 + c + 1$ and scale $\tilde{d} = \|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2/2 + d$.

I.5 Conditional posterior for κ

Interestingly, since κ is a hyperparameter to be estimated, it does not actually make use of any data, apart from the appearance of $\mathbf{X}$ in the covariance matrix for $\boldsymbol{\beta}$.

$$\begin{aligned}
\log p(\kappa|\mathbf{y}, \Theta_{-\kappa}) &= \text{const.} + \log p(\beta|\sigma^2, \kappa) + \log p(\kappa) \\
&= \text{const.} - \frac{p}{2}\log\kappa - \frac{1}{\kappa}\cdot\frac{1}{2\sigma^2}\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta} - (c+1)\log\kappa - d/\kappa \\
&= \text{const.} - (p/2 + c + 1)\log\kappa - \frac{\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta}/\sigma^2 + d}{\kappa}
\end{aligned}$$

This is an inverse gamma distribution with shape $\tilde{c} = p/2+c+1$ and scale $\tilde{d} = \boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta}/\sigma^2 + d$.

I.6 Computational note

From the above, we see that all of the Gibbs conditionals are of recognisable form, making Gibbs sampling a straightforward MCMC method to implement. We built an R package **ipriorBVS** that uses JAGS (Plummer, 2003), a variation of WinBUGS, internally for the Gibbs sampling, and wrote a wrapper function which takes formula based inputs for convenience. The **ipriorBVS** also performs two-stage BVS, and supported priors are the I-prior, g-prior, and independent prior, as used in this thesis. Although a Gibbs sampler could be coded from scratch, JAGS has the advantage of being tried and tested and has simple controls for tuning (burn-in, adaptation, thinning, etc.). Furthermore, the output from JAGS can be inspected using a myriad of multi-purpose MCMC tools to diagnose convergence problems. The **ipriorBVS** package is available at https://github.com/haziqj/ipriorBVS.

In all examples, a default setting of 4,000 burn-in samples, 1,000 adaptation size, and 10,000 samples with no thinning seemed adequate. There were no major convergence issues encountered.

Computational complexity is dominated by the inversion of a $p \times p$ matrix, and matrix multiplications of order $O(np^2)$. These occur in the conditional posterior for $\boldsymbol{\beta}$. Overall, if $n \gg p$, then time complexity is $O(np^2)$. Storage requirements are $O(np)$.