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

Md. Haziq Md. Jamil
Department of Statistics
London School of Economics and Political Science

April 16, 2018

To my parents.

Abstract

I-priors are a class of objective priors on regression functions which makes use of its Fisher information in a function space framework. We present firstly some methodology and computational work on estimating regression functions by working in the appropriate reproducing kernel Hilbert space of functions and assuming an I-prior on the function of interest. Secondly, work on extending the I-prior methodology to categorical responses for classification is presented, in which estimation is performed using a variational approximation to the likelihood. Finally, a fully Bayes approach is considered where we use I-priors for variable selection. http://phd.haziqj.ml and http://myphdcode.haziqj.ml

Keywords: Gaussian process, regression, binary, multinomial, variational, 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 51,309 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

List of Figures

List of Tables

List of Theorems

List of Definitions

List of Abbreviations

RKHS Reproducing kernel Hilbert space.

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, 2017), 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 computational methods relating to the estimation of I-prior models. Chapter 5 extends the I-prior methodology to fit discrete outcome models. Chapter 6 discusses the use of I-priors for model selection. This short chapter ultimately provides an outline of the thesis, in addition to introducing the statistical model of interest.

sec:introregmod

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:

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

eq:model1

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

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

eq:model1ass

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.

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 that the data 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 ith observation for group $j \in \{1, \ldots, m\}$. Again, assuming a linear parameterisation, this is recognisable as the multilevel or random-effects linear model, 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, and the more popular ones include LOcal regrESSion (LOESS), kernel regression, and smoothing splines. 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 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 1) we may choose a regression function appropriately, and 2) this function will generalise well to unseen data (prediction). This needed structure is given to us by assuming that our regression function for the normal model lies in some reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ equipped with the reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Often, the reproducing kernel (or simply kernel, for short) is indexed by one or more parameters which we shall denote as η. Correspondingly, the kernel is rightfully denoted as h_η to indicate the dependence of the parameters on the kernels, though where this is seemingly obvious, might be omitted. Throughout this thesis we shall make the assumption that our regression function lies in a reproducing kernel Hilbert space $\mathcal{F}$.

RKHSs provides a geometrical advantage to learning algorithms: Projections of the inputs to a richer and more informative (and higher dimensional) feature space, where learning is more likely to be successful, need not be figured out explicitly. Instead, the feature maps are implicitly calculated by the use of kernel functions. This is known as the "kernel trick" in the machine learning literature, 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 a regression function is equivalent to choosing a kernel function, and this is chosen according to the desired effects of the covariates on the regression function. An in-depth discussion on kernels and RKHSs will be provided later in Chapter 2, but for now, it suffices to say that kernels which invoke a linear, smooth and categorical dependence, are of interest. This would allow us to fit the various models described earlier within this RKHS framework.

1.3 Estimating the regression function

Having decided on a functional structure for 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[1] defined by

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

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 *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 $\mathrm{ER}[f]$ would be twice the negative log-likelihood of f, up to a constant.

The ML estimator of f 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 RKHS norm $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}$ as the regularisation term[2]. Therefore, the solution to the regularised least squares problem—call this f_{reg}—is the

[1] 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.

minimiser of the function from $\mathcal{F}$ to $\mathbb{R}$ defined by the mapping

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

{eq:penfunctional}

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, and is not usually known a priori and must be estimated from the data.

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

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

{eq:repform}

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* 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

sec:introregiprior

Building upon the Bayesian interpretation of regularisation, Bergsma (2017) proposes a 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) is defined as the distribution of a random function of the form

[2] 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).

(1.4) when the w_i are distributed according to

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

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

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

{eq:iprior}

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. More on the I-prior in Chapter 2.

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

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

{eq:postmean}

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

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

{eq:postvar}

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 the same mean[3] as in (1.7), but a slightly different variance. These are of course well-known results in Gaussian process literature—see, for example, Rasmussen and Williams (2006) 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 $\boldsymbol{\Psi}$. 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. In a fully Bayesian setting on the other hand, Markov chain Monte Carlo methods 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 RKHS to which the regression function belongs. As such, it can be seen as a unifying methodology for various regression models.

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. This encourages parsimony, as the I-prior allows complex models to be specified by just a handful of model parameters.

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

   As alluded to earlier, the process of inferring f from data is an "ill-posed" problem. In fact, any function f that passes through the data points is a solution. Regularising the problem with the use of I-priors prevents over-fitting, with the

[3] The fact that it is the same is inconsequential. It happens to be that the mean of the predictive distribution $\mathrm{E}[y_{\text{new}}|\mathbf{y}]$ for a normal model is the same as *prediction of the mean at the posterior*, $\mathrm{E}[f(x_{\text{new}})|\mathbf{y}]$. Rasmussen and Williams, 2006 points out that this is due to symmetries in the model and the posterior.

added advantage that the posterior solution under an I-prior does not tend to under-smooth as much as Tikhonov regularisation does (see Chapter 2 for details). Under-smoothing can adversely impact the estimate of f, and in real terms might even show features and artefacts that are not really there.

4. **Better prediction.**

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

5. **Straightforward inference.**

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

6. **Proper prior and posterior**

   Both the I-prior for f and the posterior solution lies in $\mathcal{F}$.

2. Is this an advantage?

The main drawback of using I-prior models 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 initialisation may ultimately lead to a global maximum, although some difficulties may be faced when numerical instabilities occur.

Lastly, a remark on model assumptions, which are twofold: 1) Assumption of $f \in \mathcal{F}$, some RKHS; 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. De-

viating 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 a stable and efficient algorithm for estimating I-prior models. The R package **iprior** is the culmination of the effort put in towards completing this chapter, which has been made publicly available on the Comprehensive R Archive Network (CRAN).
- Many models of interest involve response variables of a categorical nature. A naïve implementation of the I-prior model is certainly possible, but there is a more proper way to account for 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** attempts to contribute to the area of variable selection. The use of I-priors in the normal model, like Gaussian process priors, allow model comparison to be done easily. 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. 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.

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

Chapter 2

Vector space of functions

One of the main assumptions for regression modelling with I-priors is 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 RKHS and RKKS 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, RKHS 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. RKHS are widely studied in the applied statistical and machine learning literature, but perhaps RKKS are less so. To provide an early insight, RKKS are simply a generalisation of RKHS, and are defined as the difference between two RKHSs. The flexibility provided by RKKS will prove both useful and necessary, especially when considering the 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 2.1–2.3. Section 2.4 describes the fundamental RKHS of interest for I-prior regression, which we refer to as the "building block" RKHS/RKKS. The reason for this is that it is possible to construct new RKKS from existing ones, and this is described in Section 2.5.

A remark 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 function spaces, and $\mathcal{X}$ denotes 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 $f(\cdot)$, or 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.

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.

def:innerprod

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:** $\langle af_1 + bf_2, g \rangle_{\mathcal{F}} = a\langle f_1, g \rangle_{\mathcal{F}} + b\langle f_2, g \rangle_{\mathcal{F}}$, $\forall f_1, f_2, g \in \mathcal{F}$ and $\forall a, b \in \mathbb{R}$.
- **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 Krein spaces. However, for the purposes of the discussion moving forward, 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

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

eq:normip

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

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:** $||\lambda f||_{\mathcal{F}} = |\lambda| \cdot ||f||_{\mathcal{F}}$, $\forall \lambda \in \mathbb{R}$, $\forall f \in \mathcal{F}$
- **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*

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

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*

$$\|f + g\|_{\mathcal{F}}^2 + \|f + g\|_{\mathcal{F}}^2 = 4\langle f, g\rangle_{\mathcal{F}},$$

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

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 $\mathcal{F}$: $D(f, g) = ||f - g||_{\mathcal{F}}$, 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. If the vector space contains the limits of all Cauchy sequences (or in other words, if every Cauchy sequence converges), then it is said to be *complete*.

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 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 the vectors onto a different space, or in essence, having a function acted upon them. To establish terminology, we define linear functionals, bilinear form, and linear operators.

Definition 2.5 (Linear functional)**.** Let $\mathcal{F}$ be a Hilbert space. A *functional* L is a map from $\mathcal{F}$ to $\mathbb{R}$, and we denote its action on a function f as $L(f)$. A functional is called *linear* if it satisfies $L(f + g) = L(f) + L(g)$ and $L(\lambda f) = \lambda L(f)$, for all $f, g \in \mathcal{F}$ and $\lambda \in \mathbb{R}$.

Definition 2.6 (Bilinear form)**.** Let $\mathcal{F}$ be a Hilbert space. A *bilinear form* B takes inputs $f, g \in \mathcal{F}$ and returns a real value. It is linear in each argument separately, i.e.

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

for all $f, g, h \in \mathcal{F}$ and $\lambda_1, \lambda_2 \in \mathbb{R}$.

Definition 2.7 (Linear 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 $Af \in \mathcal{G}$. A *linear operator* satisfies $A(f+g) = A(f) + A(g)$ and $A(\lambda f) = \lambda A(f)$, for all $f, g \in \mathcal{F}$ and $\lambda \in \mathbb{R}$.

The term 'functional' is classically used in calculus of variations to denote 'a function of a function', i.e. a function having another function as its input, and outputs a real number. Really, from a function space perspective, it is simply a mapping of functions onto another vector space (the reals in this case). More generally, if the output space is another Hilbert space, then it is an operator. An interesting property of these operators to look at, besides linearity, is whether or not they are *continuous*.

def:continuity

Definition 2.8 (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 special case of linear operator, the evaluation functional, this means that a function in $\mathcal{F}$ is continuous if the evaluation functional is continuous—more on this later in Section 2.2. There is an even stronger notion of continuity called the *Lipschitz continuity*.

Definition 2.9 (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 2.8. A continuous, linear operator is also one that is bounded:

def:boundedop

Definition 2.10 (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 $\|A\| := \sup_{f \in \mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$.

thm:boundcont

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 a bounded if and only if it is continuous.*

Proof. Suppose that L is bounded. Then, $\forall f, f' \in \mathcal{F}$, there exists some $M > 0$ such that $\|A(f-g)\|_{\mathcal{G}} \leq M\|f-g\|_{\mathcal{G}}$. 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 $f \in \mathcal{F}$,

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

and thus 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 2.11 (Dual spaces)**.** Let $\mathcal{F}$ be a Hilbert space. The space $\mathcal{F}^*$ 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 $L \in \mathcal{F}'$; since L is finite-dimensional, it is bounded, and therefore continuous (see Lemma 2.1) so $L \in \mathcal{F}'$ and $\mathcal{F}^* \subseteq \mathcal{F}'$; but $\mathcal{F}' \subseteq \mathcal{F}^*$ trivially, so $\mathcal{F}^* \equiv \mathcal{F}'$. 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 just inner products.

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

Proof. Omitted—see Rudin (1987, Theorem 4.12) for a proof. □

The notion of isometry (transformation that preserves distance) is usually associated with metric spaces—two metric spaces being isometric means that they identical in as far as their metric properties are concerned. For Hilbert spaces (or normed spaces in general), 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 2.12 (Isometric isomorphism)**.** Two Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ are said to be *isometrically isomorphic* if there is a linear bijective map $A : \mathcal{F} \to \mathcal{G}$ which preserves the inner product, i.e.

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

A consequence of the Riesz representation theorem is that it gives us a canonical isometric isomorphism $A : f \mapsto \langle \cdot, f \rangle_{\mathcal{F}}$ between $\mathcal{F}$ and its continuous dual $\mathcal{F}'$, whereby $\|Af\|_{\mathcal{F}'} = \|f\|_{\mathcal{F}}$. 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 $\langle f - Pf, g \rangle_{\mathcal{G}} = 0$ for any $g \in \mathcal{G}$. The remainder $f - Pf$ belongs to the *orthogonal complement* of $\mathcal{G}$.

Definition 2.13 (Orthogonal complement)**.** Let $\mathcal{F}$ be a Hilbert space and $\mathcal{G} \subset \mathcal{F}$ be a closed subspace. The linear subspace $\mathcal{G}^{\perp} = \{f \,|\, \langle f, g \rangle_{\mathcal{G}} = 0, \forall g \in \mathcal{G}\}$ is called the orthogonal complement of $\mathcal{G}$.

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 $g = 0$.

thm:orthdecomp2

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 2.14 (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 2.15 (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 $A : \mathcal{H}_1 \to \mathcal{H}_2$ in the following way:

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

For some $y_1 \in \mathcal{H}_1$ and $y_2 \in \mathcal{H}_2$, we have that

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

It is seen that the tensor product $x_1 \otimes x_2$ is is associated with the rank one operator $B : \mathcal{H}_1' \to \mathcal{H}_2$ defined by $z \mapsto z(x_1)x_2$ with $z = \langle x_1, \cdot\rangle_{\mathcal{H}_1}$. We write $B = x_1 \otimes x_2$. Therefore, this extends a linear identification between $\mathcal{H}_1 \otimes \mathcal{H}_2$ and the space of finite-rank operators from $\mathcal{H}_1'$ to $\mathcal{H}_2$. We now have three distinct interpretations of the tensor product:

3. From Wikipedia. But don't really get it, although it might explain the Fisher information between linear functionals.

- **Bilinear form** (as defined in Definition 3.5).

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

  for $x_1, y_1 \in \mathcal{H}_1$ and $x_2, y_2 \in \mathcal{H}_2$.

- **Operator**.

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

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

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

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 Σ. 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 2.16 (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

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

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

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

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

Definition 2.17 (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 2.18 (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 $C \in \mathcal{H} \otimes \mathcal{H}$.

2.2 Reproducing kernel Hilbert space theory

sec:rkhstheory

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 RKHS, is to convince 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 $x \in \mathcal{X}$.

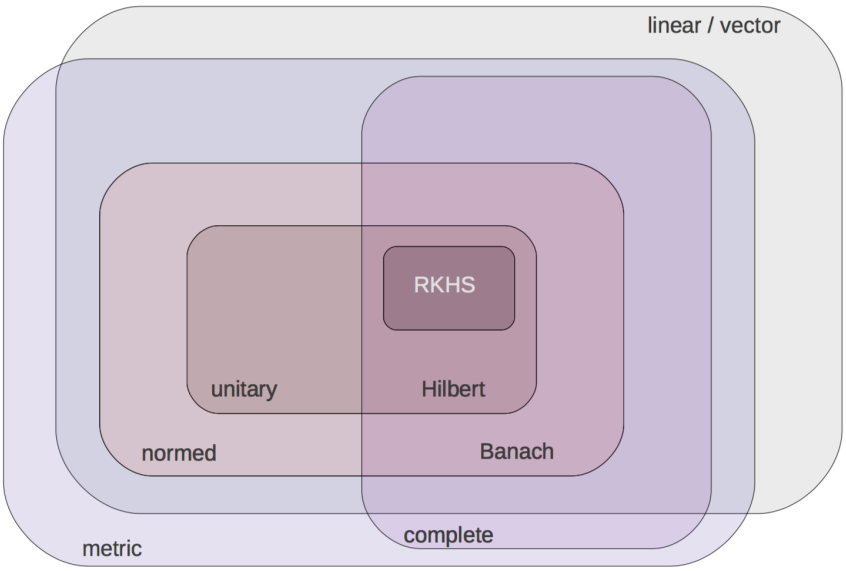


Figure 2.1: A hierarchy of vector spaces[2].

Definition 2.19 (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: $\delta_x(\lambda f + g) = (\lambda f + g)(x) = \lambda f(x) + g(x) = \lambda\delta_x(f) + \delta_x(g)$. This is in fact the linearity that was implied earlier on at the beginning of Chapter 2 when introducing the notion of functions behaving like vectors. As a remark, the calculation of the (penalised) likelihood functional involves evaluations. It is therefore important for the evaluation functional to be continuous. It turns out, this is exactly what RKHS provide.

def:rkhs

Definition 2.20 (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 of evaluation functionals in an RKHS means that functions that are close in RKHS norm imply that they are also close pointwise, but the converse is not neces-

[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.

sarily true. This gives some reassurance when trying to estimate f from $\mathcal{F}$ using the norm of $\mathcal{F}$ as a criterion for selection. More formally,

thm:normpointconv

Corollary 2.3.2 (Norm convergence implies pointwise convergence in RKHS)**.** *Let $\mathcal{F}$ be an RKHS of real functions over $\mathcal{X}$, and let f_n be a sequence of points in $\mathcal{F}$. Then, for some $f \in \mathcal{F}$,*

$$\lim_{n\to\infty} \|f_n - f\|_{\mathcal{F}} = 0 \quad \Rightarrow \quad \lim_{n\to\infty} |f_n(x) - f(x)| = 0.$$

Proof. Suppose $\mathcal{F}$ is an RKHS with reproducing kernel h. Then,

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

□

Insert figure squiggly line and smooth line.

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 inspired its name.

def:repkern

Definition 2.21 (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, $\forall \lambda_1, \ldots, \lambda_n \in \mathbb{R}$ and $\forall x_1, \ldots, x_n \in \mathcal{X}$,

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

thm:posdef

Claim 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.

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

□

Remark 2.1. 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 $x, x' \in \mathcal{X}$. $\mathcal{V}$ is known as the *feature space* and ϕ the *feature map*. In many

mathematical models involving feature space mappings, elucidation of the feature map and feature space is not necessary, and computation is made simpler by the use of kernels (known as the *kernel trick*). Note that kernels defined in this manner are positive definite, while in this thesis, we opt for a more general definition allowing for non-positive kernels.

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

Definition 2.22 (Kernel matrix)**.** Let $\{x_1, \dots, x_n\}$ 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 $h(x_i, x_j)$.

Now, one might ask what the relationship between a reproducing kernel and a RKHS is. We assert the following:

- **RKHS ⇔ reproducing kernel**. 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.
- **P.d. function ⇒ RKHS**. 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.

In essence, there is a bijection between the set of positive-definite kernels and the set of reproducing kernel Hilbert spaces. The rest of this subsection is a discussion of this assertion, which is proven by the two theorems that follow.

thm:rkhsuniqueness

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

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

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 2.21.

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, $h_1 = h_2$. □

thm:moorea

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 an RKHS. For a complete proof, see Berlinet and Thomas-Agnan (2011). 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 indeed a valid inner-product satisfying the conditions laid in Definition 2.1. At this point, the reproducing property is already had:

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

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, 2011, 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 $\mathcal{G} = \text{span}\{h(\cdot, x) \,|\, x \in \mathcal{X}\}$ endowed with the inner product as stated in Section 2.2.

2.3 Reproducing kernel Kreĭn space theory

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

Definition 2.23 (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 2.24 (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.
- $\forall f, f' \in \mathcal{F}$, $\langle f, f' \rangle_{\mathcal{F}} = \langle f_+, f'_+ \rangle_{\mathcal{F}_+} - \langle f_-, f'_- \rangle_{\mathcal{F}_-}$.

Let P be the projection of the Kreĭn space $\mathcal{F}$ onto $\mathcal{F}_+$, and $Q = I - P$ the projection onto $\mathcal{F}_-$. 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 2.25 (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 $f \in \mathcal{F}$, and 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).

Definition 2.26 (Reproducing kernel Krein 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 2.2. 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" RKHS of functions. In the context of regression modelling, we may assume that the regression function lies in any one of these single RKHS, although it may be more appropriate to consider function spaces built upon these RKHS for more complex models. Construction of new function spaces from these building block RKHS will be discussed in the next section.

5. Update graphics.

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 lemma.

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, $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f \rangle_{\mathcal{F}}} = |c_f|$. □

Figure 2.2: Sample 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 $f_\beta \in \mathcal{X}'$,

$$\begin{aligned} f_\beta(x) &= \langle x, \beta \rangle_{\mathcal{X}} \\ &= \langle f_x, f_\beta \rangle_{\mathcal{X}'} \\ &= \left\langle \langle \cdot, x \rangle_{\mathcal{X}}, f_\beta \right\rangle_{\mathcal{X}'}. \end{aligned}$$

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).

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 2.27 (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

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

Proof. 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 2.3. 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 $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

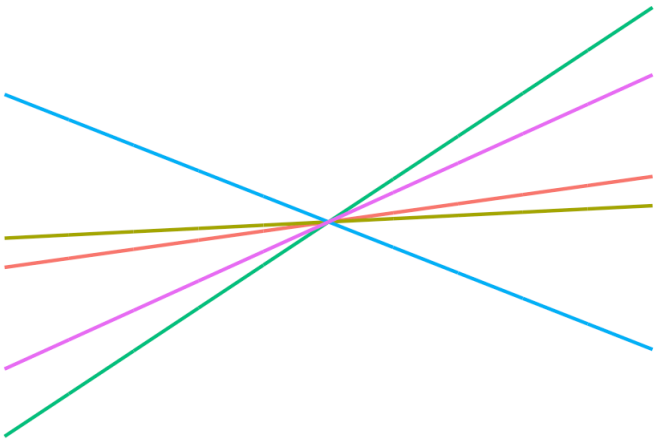


Figure 2.3: Sample paths from the canonical RKHS.

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 influenced 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 $t_1,\dots,t_k$, where each $t_j \in [0,T]$, $\big(B_\gamma(t_1),\dots,B_\gamma(t_k)\big)$ 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

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

where $\gamma \in (0,1)$ is called the *Hurst index*, *Hurst parameter* or even *Hurst coefficient*. Introduced by Mandelbrot and Van 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, then 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 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.

def:fbmrkhs

Definition 2.28 (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 $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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 2.4. 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 (2011, 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 2.29 (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.

Claim 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 $x, x' \in \mathcal{X}$,

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

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 claim. □

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.

6. This is the same for any RKHS?

Definition 2.30 (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 $\bar{h} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

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

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

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

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

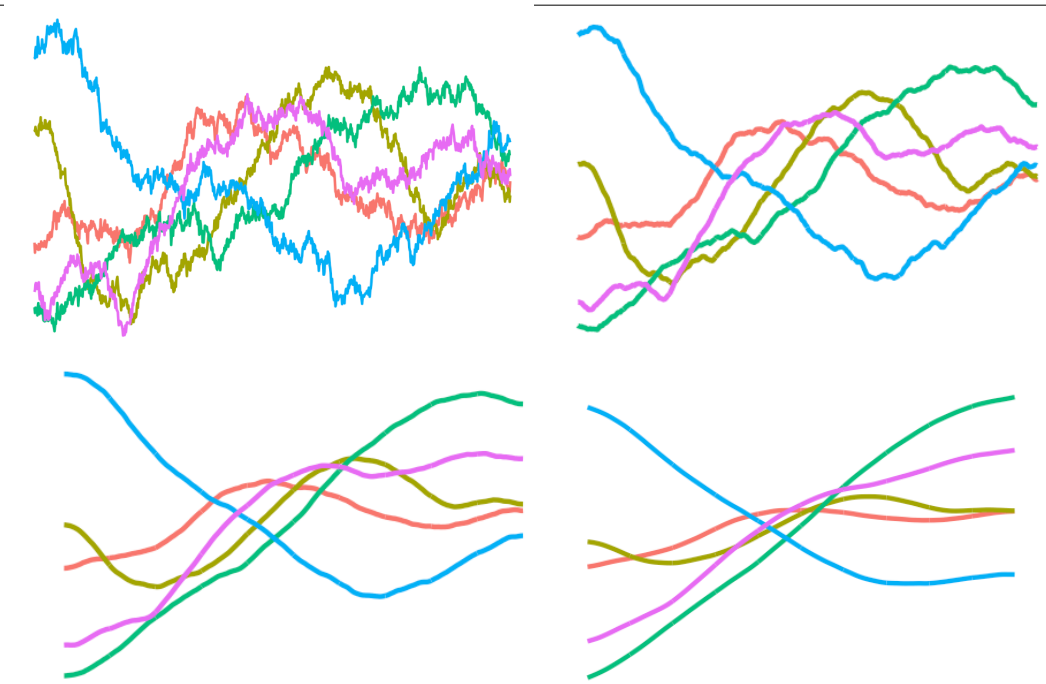


Figure 2.4: Sample paths from the fBm RKHS with varying Hurst coefficients.

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 2.31 (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 $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

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

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 parameteri-

sation $\gamma = 1/2l^2$ is utilised. Duvenaud (2014) remarks that "exponentiated quadratic" is a better fitting and 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, Hush, 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 2.32 (Universal kernel)**.** Let $\mathrm{C}(\mathcal{X})$ is the space of all continuous, complex-valued functions $f : \mathcal{X} \to \mathbb{C}$ equipped with the maximum norm $\|\cdot\|_\infty$, and denote $\mathcal{K}(\mathcal{X})$ 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 some very convenient computational advantages that the SE kernel brings (more on this in a later chapter), 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 2.33 (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 $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

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

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.

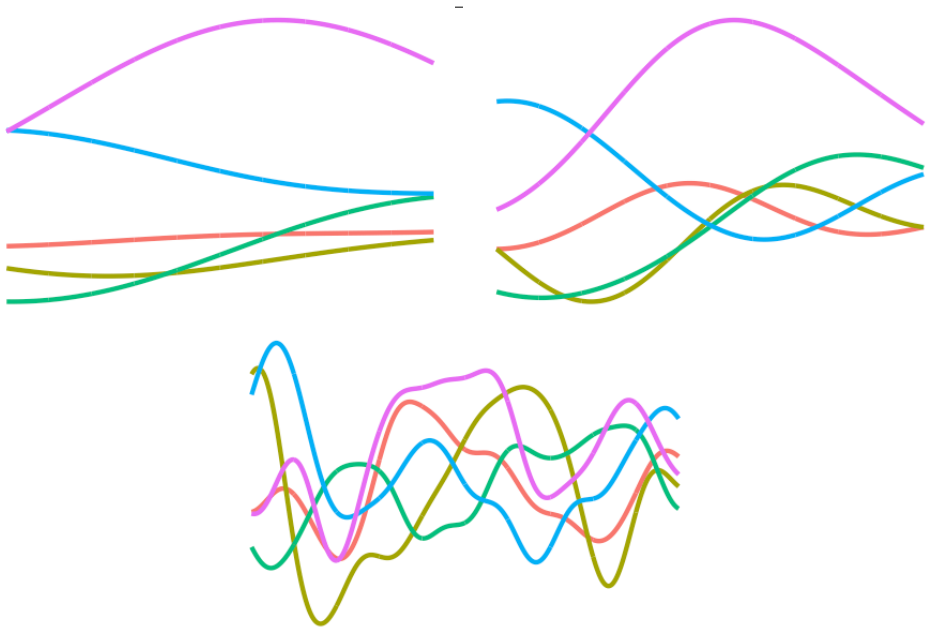


Figure 2.5: Sample paths from the SE RKHS with varying values for the lengthscale.

2.4.5 The Pearson RKHS

In all of the previous RKHS 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.

def:pearson

Definition 2.34 (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 $\mathrm{P}(X = x)$.

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

$$\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R} | \, \mathrm{E}\, f(X) = 0\}$$

with

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

Proof. Write $p_x = \mathrm{P}(X = x)$. The set of functions $\{h(\cdot, x) | x \in \mathcal{X}\}$ 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

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

and thus $h(\cdot, x)$ spans $\mathcal{F}$ so $f \in \mathcal{F}$. To provide the second part, 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 $f(X)$. □

2.5 Constructing RKKS from existing RKHS

sec:constru
ctrkks

The previous section outlined all of the basic RKHSs of functions that will form the building blocks when constructing more complex function spaces. As previously mentioned in the preliminaries, sums of kernels are kernels and products of kernels are also

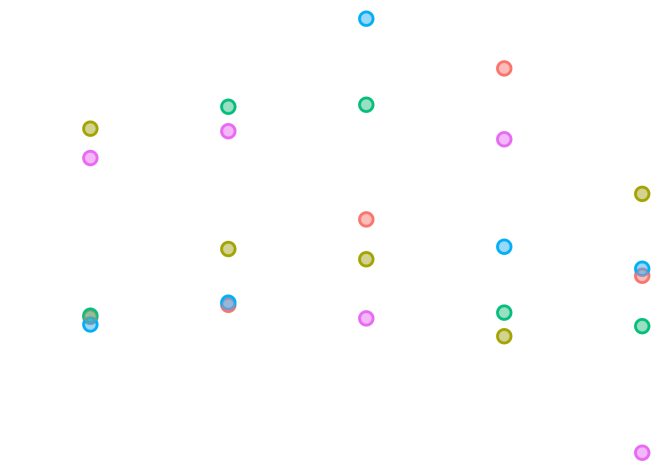


Figure 2.6: Sample "paths" from the Pearson RKHS. These are represented as points over a finite set.

kernels. This provides us a platform for constructing new RKHS 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 RKKS.

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.

thm:sumkernels

Lemma 2.12 (Sum of kernels)**.** *If h_1 and h_2 are kernels on $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively, then $h = h_1 + h_2$ is a 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 an 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 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 (2011, Theorem 5). □

thm:prodkernels

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 an 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

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

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 (2011, Theorem 13). □

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

The scale of an 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 RKKS instead.

Remark 2.6. 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 $\mathcal{F}_+ \oplus \mathcal{F}_-$. If it is important to note that both $\mathcal{F}$ and $\mathcal{F}_\mathcal{H}$ contain identical functions over $\mathcal{X}$, but their topologies are different. 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 $\mathcal{F}_\mathcal{H}$.

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 2.35 (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

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

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 defined the *polynomial RKKS* of degree d.

[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, $(\mathbf{A} \circ \mathbf{B})_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$.

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 an 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.

Claim 2.14. *The polynomial RKKS of functions over* $\mathbb{R}$*, denoted* $\mathcal{F}$*, contains polynomial functions of the form* $f(x) = \sum_{k=0}^{d} \beta_k x^k$*.*

Proof. By construction, $\mathcal{F} = \mathcal{F}_0 \oplus \bigoplus_{i=1}^{d} \bigotimes_{j=1}^{i} \mathcal{F}_j$, where each $\mathcal{F}_j, j \neq 0$ is the canonical RKHS, and $\mathcal{F}_0$ is the RKHS of constant functions. Each $g \in \mathcal{F}$ can therefore be written as $g = \beta_0 + \sum_{i=1}^{d} \prod_{j=1}^{i} f_j$, and $f_j(x) = b_j x$ from before, where b_j is a constant. Therefore, $g(x) = \sum_{k=0}^{d} \beta_k x^k$. □

Remark 2.7. 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 the identifiability 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:

- $Af(x) = f(1) = \mu_1$. 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 $\mu_1 = 0$.
- $Af(x) = \sum_{x=1}^m f(x)/m =: \bar{\alpha}$. 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 [Side note: This idempotent property of the linear operator A on f speaks to the possibility of it being an *orthogonal projection*, and indeed this is so—we shall return to this point later when we describe functional ANOVA decomposition]. 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 $Af_c = 0$.

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 $\prod_{k=1}^p \mathcal{X}_j$ to $\mathbb{R}$. Let A_j be an averaging

operator on $\mathcal{X}_k$ that averages the kth 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

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

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 $A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p$.

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 is

$$f = f_0 + f_1 + f_2 + f_{12}.$$

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,

$$\begin{aligned} f_0 &= A_1 A_2 f && = f(1,1) \\ f_1 &= (I - A_1) A_2 f && = f(x_1, 1) - f(1,1) \\ f_2 &= A_1 (I - A2) f && = f(1, x_2) - f(1,1) \\ f_{12} &= (I - A_1)(I - A2) f && = f(x_1, x_2) - f(x_1, 1) - f(1, x_2) + f(1,1). \end{aligned}$$

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

$$\begin{aligned} f_0 &= A_1 A_2 f && = f_{\cdot\cdot} \\ f_1 &= (I - A_1) A_2 f && = f_{x_1 \cdot} - f_{\cdot\cdot} \\ f_2 &= A_1 (I - A_2) f && = f_{\cdot x_2} - f_{\cdot\cdot} \\ f_{12} &= (I - A_1)(I - A_2) f && = f - f_{x_1 \cdot} - f_{\cdot x_2} + f_{\cdot\cdot}, \end{aligned}$$

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

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 $\mathcal{F}$. Specifically, we shall consider the (Hilbert) space of square integrable functions over $\mathcal{X}$ with measure ν, $\mathcal{F} \equiv \mathrm{L}^2(\mathcal{X}, \nu)$. 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 $\nu = \nu_1 \times \cdots \times \nu_d$ is the product measure on $\mathcal{X}$. 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

$$A_k f = \int_{\mathcal{X}_k} f(x_1, \ldots, x_p) \, \mathrm{d}\nu(x_k). \tag{2.2}$$

{eq:avgoper}

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

$$f = \overbrace{\int_{\mathcal{X}} f(x) \mathrm{d}\nu(x)}^{f_0} + \overbrace{\left(f - \int_{\mathcal{X}} f(x) \mathrm{d}\nu(x) \right)}^{f_1} \tag{2.3}$$

{eq:functionalanova1}

and for the two-way ANOVA ($p = 2$), we have $f = f_0 + f_1 + f_2 + f_{12}$, with

$$\begin{aligned}
f_0 &= \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu(x_1) \, \mathrm{d}\nu(x_2) \\
f_1 &= \int_{\mathcal{X}_2} \left(f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2) \, \mathrm{d}\nu(x_1) \right) \mathrm{d}\nu(x_2) \\
f_2 &= \int_{\mathcal{X}_1} \left(f(x_1, x_2) - \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu(x_2) \right) \mathrm{d}\nu(x_1) \\
f_{12} &= f(x_1, x_2) - \int_{\mathcal{X}_1} f(x_1, x_2) \mathrm{d}\nu(x_1) - \int_{\mathcal{X}_2} f(x_1, x_2) \, \mathrm{d}\nu(x_2) \\
&\quad + \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f(x_1, x_2) \mathrm{d}\nu(x_1) \, \mathrm{d}\nu(x_2).
\end{aligned}$$

As a remark, the averaging operator A_k defined in (2.2) is indeed true to its name, in that it calculates the mean function of f over the kth coordinate. For comparison, this is identical to the second type of restriction we considered in the classical ANOVA previously (i.e., setting $\sum_j \mu_j = 0$). We must also have, as before, that

$A_k f_{\mathcal{K}} = 0, \forall k \in \mathcal{K} \in \mathcal{P}_p$. For the one-way functional ANOVA decomposition in (2.3), it must be that f_1 is a zero-mean function. As for the two-way ANOVA, it is the case that $\int_{\mathcal{X}_k} f_1(x_1, x_2)\mathrm{d}\nu(x_k) = 0, k = 1, 2$, and $\int_{\mathcal{X}_1}\int_{\mathcal{X}_2} f_{12}(x_1, x_2)\mathrm{d}\nu(x_1)\mathrm{d}\nu(x_1) = 0$.

We notice that the decomposition in (2.3) is orthogonal:

Claim 2.15. *For the ANOVA decomposition in (2.3), f_0 and f_1 are orthogonal for the usual L^2 inner product.*

Proof. Note that f_0 is a constant function, and that $f_1 = f - f_0$. Thus,

$$\begin{aligned}\langle f_0, f_1\rangle &= \int f_0 f_1 \mathrm{d}\nu \\ &= f_0 \int (f - f_0)\,\mathrm{d}\nu \\ &= f_0(f_0 - f_0) = 0.\end{aligned}$$

□

In fact, for $k = 1$, any $f \in \mathcal{F}$ can be decomposed as a sum of a constant plus a zero mean function, so we have the geometric decomposition of the vector space $\mathcal{F} = \mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_1$, where $\mathcal{F}_0$ is a vector space of constant functions, and $\bar{\mathcal{F}}_1$ a vector space of zero-mean functions over $\mathcal{X}_1$. For $k \geq 2$ we can argue something similar. The space $\mathcal{F}$ has the tensor product structure[4] $\mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_p$, and considered individually, each $\mathcal{F}_k$ can be decomposed orthogonally $\mathcal{F}_k = \mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_k$. Note that $\mathcal{F}_k$ consists of functions $f : \mathcal{X}_k \to \mathbb{R}$. Expanding out under the distributivity rule of tensor products and rearranging slightly, we obtain

$$\begin{aligned}\mathcal{F} &= \left(\mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_1\right) \otimes \cdots \otimes \left(\mathcal{F}_0 \overset{\perp}{\oplus} \bar{\mathcal{F}}_1\right) \\ &= \mathcal{F}_0^{\otimes p} \overset{\perp}{\oplus} \bigoplus_{j=1}^{p}{}^{\perp} \left(\mathcal{F}_0^{\otimes(p-1)} \otimes \bar{\mathcal{F}}_j\right) \overset{\perp}{\oplus} \bigoplus_{\substack{j,k=1 \\ j<k}}^{p}{}^{\perp} \left(\mathcal{F}_0^{\otimes(p-2)} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k\right) \\ &\quad \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} \left(\bar{\mathcal{F}}_1 \otimes \cdots \otimes \bar{\mathcal{F}}_p\right).\end{aligned} \tag{2.4}$$

{eq:funcanovaspace}

To clarify,

- $\mathcal{F}_0^{\otimes p}$ is the space of constant functions $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$.

- $\left(\mathcal{F}_0^{\otimes(p-1)} \otimes \bar{\mathcal{F}}_j\right)$ is the space of functions that are constant on all coordinates except the jth coordinate of x. Further, the functions are centred on the jth coordinate.
- $\left(\mathcal{F}_0^{\otimes(p-2)} \otimes \bar{\mathcal{F}}_j \otimes \bar{\mathcal{F}}_k\right)$ is the space of functions that are constant on all coordinates except the jth and kth coordinate of x. Further, 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 $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \to \mathbb{R}$.
- Similarly 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 orthogonal spaces in (2.4) leads to the *functional ANOVA representation*

$$f(x) = \mu + \sum_{j=1}^{p} f_j(x_j) + \sum_{\substack{j,k=1 \\ j<k}}^{p} f_{jk}(x_j, x_k) + \cdots + f_{1\cdots p}(x). \tag{2.5}$$ {eq:functionalanova2}

Definition 2.36 (Functional ANOVA representation)**.** Let $\mathcal{P}_d = \mathcal{P}(\{1, \ldots, d\})$, the power set of $\{1, \ldots, d\}$. For any function $f \in \mathcal{F} \equiv \mathrm{L}^2(\mathcal{X}_1 \times \cdots \times \mathcal{X}_d, \nu_1 \otimes \cdots \otimes \nu_d)$, the formula for f in (2.5) is known as the *functional ANOVA representation* of f if $\forall k \in \mathcal{K} \in \mathcal{P}_p$,

$$A_k f_{\mathcal{K}} = \int_{\mathcal{X}_{\mathcal{K}}} f_{\mathcal{K}}(x_{\mathcal{K}}) \mathrm{d}\nu_k(x_k) = 0, \tag{2.6}$$ {eq:funcanovaorth}

where $\mathcal{X}_{\mathcal{K}} = \prod_{k \in \mathcal{K}} \mathcal{X}_k$, and $x_{\mathcal{K}} = \{x_k, k \in \mathcal{K}\}$ is an element of this space. In other words, the integral of $f_{\mathcal{K}}$ with respect to any of the variables indexed by the elements in $\mathcal{K}$ (itself an element of the power set), is zero. The requirement (2.6) ensures orthogonality of the summands in (2.5).

[4] There is an isomorphism $\mathrm{L}^2(\mathcal{X}_1 \times \cdots \times \mathcal{X}_d, \nu_1 \otimes \cdots \otimes \nu_d) \cong \mathrm{L}^2(\mathcal{X}_1, \nu_1) \otimes \cdots \otimes \mathrm{L}^2(\mathcal{X}_d, \nu_d)$. See, for example, Reed and Simon (1972) and Krée (1974).

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

$$f_0 = \int f \mathrm{d}\nu$$
$$f_j = \int f \prod_{i \neq j} \mathrm{d}\nu_i - f_0$$
$$f_{jk} = \int f \prod_{i \neq j,k} \mathrm{d}\nu_i - f_j - f_k - f_0.$$

Remark 2.8. Not all of the higher order terms need to be included. There may even be a model motivated reason for dropping certain main effects or interaction effects.

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 be not in vain. The main idea is to construct an RKKS such that the functions that lie in them will have the ANOVA representation in (2.5). 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 (2.4). The ANOVA RKKS will be constructed by a similar manipulation of the individual kernels representing the RKHS building blocks.

def:anovarkks

Definition 2.37 (The ANOVA RKKS)**.** For $k = 1, \ldots, p$, let $\mathcal{F}_k$ be a centred RKHS 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

$$h_\lambda(x, x') = \prod_{k=1}^{p} \big(1 + \lambda_k h_k(x_k, x'_k)\big). \tag{2.7}$$

{eq:anovarkks}

The construction an ANOVA RKKS is very very simple in through multiplication of univariate kernels. Expanding out equations (2.7), 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 the expansion that the ANOVA RKKS yields functions that resemble those with the ANOVA representation in (2.5): The mean value of the function stems from the '1', i.e. it lies in an 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 remarked in the previous subsection, 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.

2.6 Summary

The brief notes on functional analysis allow us to describe the theory of reproducing kernel Hilbert and Kreĭn spaces. These are of great interest to us because the topology

endowed on such spaces gives great assurances—in particular, all evaluation functionals are continuous in these spaces. Moreover, RKHS and RKKS 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 statistical models (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 the material described. Additionally, Rudin (1987) 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 (2011), 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)
- **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 RKKS.
- **RKHS building blocks**. The main building block RKHS, i.e. the canonical RKHS, the fBm RKHS and the Pearson RKHS are described in the manuscript of Bergsma (2017).
- **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. Kuo et al., 2010), and sensitivity analysis (e.g. Sobol, 2001; Durrande et al., 2013).

2.7 Miscellanea

2.7.1 A vector space... of 'functions'?

At first glance, this may seem strange, that the notion of functions (as mappings from input to output space) and vector spaces are somehow equatable. Upon further thought, one realises that firstly, two functions of a similar, particular form may be added together (in some meaningful way) resulting in a function in that same form. Secondly, multiplication of a function by a scalar c can be thought of as c times the output of that function. Indeed, running through the checklist of what constitutes a vector space, we find that a "space of functions" satisfies them all. In modern linear algebra texts, this checklist is the eight axioms of vector spaces over a field $\mathbb{F}$: The vectors forms an abelian group under addition, and this group has an $\mathbb{F}$-module structure.

Chapter 3

Fisher information and the I-prior

Traditionally, Fisher information is calculated for unknown parameters θ of probability distribution from observable random variables. In a similar light, we can treat the regression function f in the model stated in (1.1), subject to (1.2), as the unknown "parameter" for which we would like information regarding. In this chapter, we extend the notion of Fisher information to abstract objects in Hilbert spaces, and also to linear functionals of these objects. This will allow us to achieve our aim of deriving the Fisher information for our regression function.

Following this, we shall 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; Jaynes, 1957b)—in the absence of any prior knowledge, a prior distribution which maximises entropy should be used. It turns out, the entropy maximising prior for f is Gaussian with mean chosen a priori and covariance kernel proportional to the Fisher information. Such a distribution on f is called the I-prior distribution.

3.1 The traditional Fisher information

It was Fisher (1922) who introduced the method of maximum likelihood 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 (2013, Section 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. Fisher information may be added much in the same way as log-likelihood may be added—the *total Fisher information* from n independent and identically distributed random variables $Y_1, \ldots, Y_n$ is simply the sum of the n *unit Fisher information*, i.e. $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$.

7. Check if total Fisher information is relevant.

3.2 Fisher information for Hilbert space objects

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 RKHSs later. The score and Fisher information is derived in a familiar manner, but extra care is required when taking derivatives with respect to Hilbert space objects. 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.

def:frechet

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 $f : \mathcal{U} \to \mathcal{W}$ is called *Fréchet differentiable* at $x \in \mathcal{U}$ if there exists a bounded, linear operator $T : \mathcal{V} \to \mathcal{W}$ such that

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

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

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

Remark 3.2. While the Fréchet derivative is most commonly defined as derivatives of functions between Banach spaces, the definition itself also applies to Hilbert spaces. Since our main focus are RKHSs, it is presented as such, and we follow the definitions supplied in Balakrishnan (1981, Definition 3.6.5) and Bouboulis and Theodoridis (2011, Section 6).

Remark 3.3. The use of the open subset $\mathcal{U}$ in the definition above for the domain of the function f is so that the notion of f 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 f 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 $f(x+v) = f(v) + \mathrm{d}f(x)(v) + o(v)$, that is, the derivative of f at x gives the best linear approximation to f 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, $\mathrm{d}f(x)$ 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 $f : \mathcal{U} \to \mathcal{W}$ is a differentiable function at all points in $\mathcal{U} \subseteq \mathcal{V}$, then its derivative is a linear map

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

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 f, defined by

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

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 $\psi \in \mathrm{L}(\mathcal{V} \times \mathcal{V}, \mathcal{W})$ such that for all $v, v' \in \mathcal{V}$, $\phi(v)(v') = \psi(v, v')$.

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 $\mathrm{d}^2 f(x)$ is therefore a bounded, 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 the functional derivative in calculus of variations. Let $\mathcal{V}$, $\mathcal{W}$ and $\mathcal{U}$ be as before, and consider the function $f : \mathcal{U} \to \mathcal{W}$.

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

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

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

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). 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.

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

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

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

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

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

{eq:frecimplygat}

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 $\partial f : \mathcal{U} \to (\mathcal{V}; \mathcal{W})$ is continuous at $x \in \mathcal{U}$. In other words, if $\partial f(x)$ 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 $\mathrm{d}f(x)$ exists and $\mathrm{d}f(x) = \partial f(x)$ (Tapia, 1971, p. 57 & 66).

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

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

{eq:frech2gat}

by a similar argument in the proof 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 the Riesz representation theorem says that the mapping $A : \mathcal{V} \to \mathcal{V}'$ from the Hilbert space $\mathcal{V}$ to its continuous dual space $\mathcal{V}'$ defined by $A = \langle \cdot, v \rangle_{\mathcal{V}}$ for some $v \in \mathcal{V}$ is an isometric isomorphism. Again, let $\mathcal{U} \subseteq \mathcal{V}$ be an open subset, and let $f : \mathcal{U} \to \mathbb{R}$ be a (Fréchet)

differentiable function with derivative $\mathrm{d}f : \mathcal{U} \to \mathrm{L}(\mathcal{V}, \mathbb{R}) \equiv \mathcal{V}'$. We define the gradient as follows.

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

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

for any $v \in \mathcal{V}$. Note that ∇f 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 f at x.

Since the gradient of f is an operator on $\mathcal{U}$ to $\mathcal{V}$, it may itself have a (Fréchet) derivative. Assuming existence, i.e., f is twice Fréchet differentiable at $x \in \mathcal{U}$, we call this derivative the *Hessian* of f. From (3.2), it must be that

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

The second line follows from the definition of gradients, and the third line follows by linearity of inner products. Note that since the Fréchet and Gâteaux differentials coincide, we have that $\partial_v \nabla f(x) = \mathrm{d}\nabla f(x)(v)$. Letting $\mathcal{V}$, $\mathcal{W}$ and $\mathcal{U}$ be as before, we now define the Hessian for the function $f : \mathcal{U} \to \mathcal{W}$.

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

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

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

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

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

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 derive 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 $\mathrm{d}^2 L(\cdot|Y) : \Theta \to \mathrm{L}(\Theta \times \Theta, \mathbb{R})$. We now prove the following proposition.

thm:fisherinfohilbert

Proposition 3.2 (Fisher information in Hilbert space)**.** *Assume that $p(Y|\cdot)$ and $\log p(Y|\cdot)$ are both 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*

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

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

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

The interchange of Lebesgue integrals and Fréchet differentials is allowed under certain conditions[4] (Kammar, 2016). 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 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 $\mathrm{E}[\nabla L(\theta|Y) \otimes \nabla L(\theta|Y)] \in \Theta \otimes \Theta$.

[4] 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, like the normal density that we will be working with, this isn't an issue.

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

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

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

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

Thus, we see that from the first result obtained,

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

while

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

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

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

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

{eq:fisher-linear-functional}

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

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

where $\alpha \in \mathbb{R}$ is an intercept and f is in an RKHS $\mathcal{F}$ with kernel $h_\eta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

thm:fisherregf

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 an RKHS 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 $\boldsymbol{\theta} = \{\alpha, \boldsymbol{\Psi}, \eta\}$. Without loss of generality, assume $\alpha = 0$; even if not, we can always add back α to the y_i's later. Regardless, both α and $\mathbf{y}$ are constant in the differential of $L(v|\mathbf{y}, \boldsymbol{\theta})$. The log-likelihood of v is given by

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

and the score by

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

Differentiating again gives

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

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

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

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 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 $f\in\mathcal{F}$, an 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 Lemma 3.3, the Fisher information for β 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 $f(x')$.

thm:fisherreglinfunc

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 RKHS $\mathcal{F}$ 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 RKHS $\mathcal{F}$, 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.

sec:inducedFisherRKHS

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 RKHS $\mathcal{F}$. 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):

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

{eq:subspaceFn}

Since $h(\cdot, x_i) \in \mathcal{F}$, then 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 in Theorem 2.6, this is a finite subspace with dimension n.

Let $\mathcal{F}_n^\perp$ be the orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$. By the orthogonal decomposition theorem, 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 earlier in 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 we have that $\mathcal{F}_n^\perp = \{0\}$. Alternatively, we could have argued the following: 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 seems to suggest the statement in 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 Fisher 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*

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

{eq:Fninner prod}

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. Since $\mathcal{F}_n$ is a finite subspace of $\mathcal{F}$, it is complete, and thus a Hilbert space. What remains 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 $h(\cdot, x_j) = \sum_{k=1}^n \delta_{jk} h(\cdot, x_k)$, we see that $h(\cdot, x_j)$ is also an element of $\mathcal{F}_n$, and in particular,

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

where δ is the Kronecker delta. Denote by ψ_{ij}^- the (i,j)th element of $\boldsymbol{\Psi}^{-1}$. A fact we will use later is $\sum_{k=1}^n \psi_{jk}\psi_{ik}^- = (\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1})_{ji} = (\mathbf{I}_n)_{ji} = \delta_{ji}$. Then,

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

Therefore, h_n is a reproducing kernel for $\mathcal{F}_n$. □

3.5 The I-prior

In the introductory chapter, 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, 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.

def:entropy

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

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

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 of functional derivatives. 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, which is typically denoted $\partial H / \partial p$. We now present another well known result from information theory, regarding the form of the maximum entropy distribution.

thm:maxentr

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*

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

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

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

From the above illustration preceding the lemma, taking derivatives with respect to p 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 RKHS $\mathcal{F}$. 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 $f(x_i) = g(x_i), i = 1, \dots, n$. 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 $f_0 + \mathcal{F}_n := \{f_0 + f \,|\, f_0 \in \mathcal{F}, f \in \mathcal{F}_n\}$. 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 I-prior theorem.

Theorem 3.6 (The I-prior)**.** *Let $\mathcal{F}$ be an RKHS with kernel h, and consider the finite dimensional subspace $\mathcal{F}_n$ of $\mathcal{F}$ equipped with an inner product as in Lemma 2.5. 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 $\mathcal{P}_0$ be the class of distributions p such that*

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

Denote by $\tilde{p}$ the density of the entropy maximising distribution among the class of distributions within $\mathcal{P}_0$. 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 $D(f, f') = \|f - f'\|_{\mathcal{F}_n}$. For $f, f_0 \in \mathcal{F}$, take the orthogonal projections of these vectors onto $\mathcal{F}_n$

8. Double check this proof.

$$f = \sum_{i=1}^{n} h(\cdot, x_i) w_i + r_f \quad \text{and} \quad f_0 = \sum_{i=1}^{n} h(\cdot, x_i) w_{i0} + r_0$$

and compute the squared distance between them:

$$\begin{aligned} D(f, f_0)^2 &= \|f - f_0\|_{\mathcal{F}_n}^2 \\ &= \left\| \sum_{i=1}^n h(\cdot, x_i) w_i - \sum_{i=1}^n h(\cdot, x_i) w_{i0} \right\|_{\mathcal{F}_n}^2 \\ &= \left\| \sum_{i=1}^n h(\cdot, x_i)(w_i - w_{i0}) \right\|_{\mathcal{F}_n}^2 \\ &= (\mathbf{w} - \mathbf{w}_0)^\top \boldsymbol{\Psi}^{-1} (\mathbf{w} - \mathbf{w}_0). \end{aligned}$$

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

$$(w_1, \ldots, w_n)^\top \sim \mathrm{N}_n(\mathbf{w}_0, \boldsymbol{\Psi}).$$

This implies that f is Gaussian, since

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

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

$$\begin{aligned} \mathrm{E}\langle f, f' \rangle_{\mathcal{F}} &= \mathrm{E} \left\langle \sum_{i=1}^n h(\cdot, x_i) w_i, f' \right\rangle_{\mathcal{F}} \\ &= \mathrm{E} \left[\sum_{i=1}^n w_i \left\langle h(\cdot, x_i), f' \right\rangle_{\mathcal{F}} \right] \\ &= \sum_{i=1}^n w_{i0} \left\langle h(\cdot, x_i), f' \right\rangle_{\mathcal{F}} \\ &= \left\langle \sum_{i=1}^n h(\cdot, x_i) w_{i0}, f' \right\rangle_{\mathcal{F}} \\ &= \langle f_0, f' \rangle_{\mathcal{F}}, \end{aligned}$$

so $\mu \equiv f_0 = \sum_{i=1}^n h(\cdot, x_i) w_{i0}$.

The covariance between two evaluation functionals of f is shown to satisfy

$$\begin{aligned}\operatorname{Cov}\big(f(x), f(x')\big) &= \operatorname{Cov}\big(\langle f, h(\cdot,x)\rangle_{\mathcal{F}}, \langle f, h(\cdot,x')\rangle_{\mathcal{F}}\big) \\ &= \mathrm{E}\big(\langle f - f_0, h(\cdot,x)\rangle_{\mathcal{F}} \langle f - f_0, h(\cdot,x')\rangle_{\mathcal{F}}\big) \\ &= \big\langle C, h(\cdot,x) \otimes h(\cdot,x')\big\rangle_{\mathcal{F}\otimes\mathcal{F}},\end{aligned}$$

where $C \in \mathcal{F} \otimes \mathcal{F}$ is the covariance element of f. Write $h_x := \langle h(\cdot,x), f\rangle_{\mathcal{F}}$. Then, by the usual definition of covariances, we have that

$$\operatorname{Cov}(h_x, h_{x'}) = \mathrm{E}[h_x h_{x'}] - \mathrm{E}[h_x]\,\mathrm{E}[h_{x'}],$$

where, making use of the reproducing property of h for $\mathcal{F}$, the first term on the right-hand side is

$$\begin{aligned}\mathrm{E}[h_x h_{x'}] &= \mathrm{E}\left[\left\langle h(\cdot,x), \sum_{i=1}^n h(\cdot,x_i) w_i \right\rangle_{\mathcal{F}} \left\langle h(\cdot,x'), \sum_{j=1}^n h(\cdot,x_j) w_j \right\rangle_{\mathcal{F}}\right] \\ &= \mathrm{E}\left[\sum_{i=1}^n \sum_{j=1}^n w_i w_j \left\langle h(\cdot,x), h(\cdot,x_i)\right\rangle_{\mathcal{F}} \left\langle h(\cdot,x'), h(\cdot,x_j)\right\rangle_{\mathcal{F}}\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n (\psi_{ij} + w_{i0} w_{j0}) h(x,x_i) h(x',x_j),\end{aligned}$$

while the second term on the right-hand side is

$$\begin{aligned}\mathrm{E}[h_x]\,\mathrm{E}[h_{x'}] &= \left(\sum_{i=1}^n w_{i0} \left\langle h(\cdot,x), h(\cdot,x_i)\right\rangle_{\mathcal{F}}\right) \left(\sum_{j=1}^n w_{j0} \left\langle h(\cdot,x'), h(\cdot,x_j)\right\rangle_{\mathcal{F}}\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n w_{i0} w_{j0} h(x,x_i) h(x',x_j).\end{aligned}$$

Thus,

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

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 RKHS $\mathcal{F}$ has the simple representation

$$f(x_i) = f_0(x_i) + \sum_{k=1}^{n} h(x_i, x_k) w_k$$
$$(w_1, \ldots, w_n)^\top \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}).$$

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

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

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 $\mathbf{H}$ 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 an RKHS $\mathcal{F}$, we established that the entropy maximising prior distribution for f is Gaussian with some prior mean f_0 that needs to be chosen, and covariance function equal 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 $\mathcal{F}$. 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.

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 equal to the Fisher information for f. We also saw how new function spaces can be constructed via the polynomial and ANOVA RKKS. 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.

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 and Bergsma, 2017), which is a publicly available R package that has been published to 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.

sec:various-regression

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

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

{eq:ipriorcanonical}

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 scaled similarly, then the values of f are incoherent—if x_1 measures weight in kilograms and x_2 height in centimetres, what measurement does $\beta_1 x_1 + \beta_2 x_2$ represent? To overcome this, 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,

9. Can't I just standardise x?

one for each of the RKKSs 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

$$\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, \lambda^2 \mathbf{X}^\top \boldsymbol{\Lambda} \boldsymbol{\Psi} \boldsymbol{\Lambda} \mathbf{X}),$$

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. 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 Miscellanea 4.7.1 for a discussion.

4.1.2 Multilevel linear modelling

sec:multilevelmodels

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:

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

{eq:anovamultilevel}

To mimic the 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,

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

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 Miscellanea 4.7.2. 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

- $f(\mathbf{x}_i^{(j)}, j) = f_0 + f_1(\mathbf{x}_i^{(j)}) + f_2(j)$ (random intercept model);
- $f(\mathbf{x}_i^{(j)}, j) = f_0 + f_1(\mathbf{x}_i^{(j)})$ (linear regression model);
- $f(\mathbf{x}_i^{(j)}, j) = f_0 + f_2(j)$ (ANOVA model);
- $f(\mathbf{x}_i^{(j)}, j) = f_0$ (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] 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.

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 $j \in \{1, \dots, m_1\}$ and $k \in \{1, \dots, m_2\}$ denote the two levels, and simply decompose the regression function accordingly:

$$f(\mathbf{x}_i^{(j,k)}, j, k) = f_0 + f_1(\mathbf{x}_i^{(j,k)}) + f_2(j) + f_3(k) + f_{12}(\mathbf{x}_i^{(j,k)}, j) + f_{13}(\mathbf{x}_i^{(j,k)}, k) + f_{23}(j, k) + f_{123}(\mathbf{x}_i^{(j,k)}, j, k).$$

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:

$$f(\mathbf{x}_i^{(j)}, j, \mathbf{z}_j) = f_0 + f_1(\mathbf{x}_i^{(j)}) + f_2(j) + f_3(\mathbf{z}_i^{(j)}) + f_{12}(\mathbf{x}_i^{(j)}, j).$$

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

Remark 4.2. Indexing can be tricky, but we find the following helpful. Supposing $m = 2$, and $n_1 = n_2 = 3$, then a typical panel data set looks like this:

y	x	z	i	j	k
y_{11}	x_{11}	z_1	1	1	1
y_{21}	x_{21}	z_1	2	1	2
y_{31}	x_{31}	z_1	3	1	3
y_{12}	x_{12}	z_2	1	2	4
y_{22}	x_{22}	z_2	2	2	5
y_{32}	x_{32}	z_2	3	2	6

The y's are the responses, x's covariates, and z's group-level covariates. If $\iota : (i, j) \mapsto k$ is a function which maps the dual index set (i, j) to the single index set $k \in \{1, \dots, n\}$, then the multilevel regression function can be expressed as the regression function in model (1.1).

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

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

where the distribution of $\boldsymbol{\epsilon}_i = \big(\epsilon_i(t_{i1}), \ldots, \epsilon_i(t_{iT_i})\big)^\top$ 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 $i, i' \in \{1, \ldots, n\}$.

Using the functional ANOVA decomposition on the regression function, we obtain

$$f(x_i, t) = f_0 + f_1(x_i) + f_2(t) + f_{12}(x_i, t), \tag{4.3}$$

{eq:longitudinalanova}

where f_0 is an overall constant, $f_1 \in \mathcal{F}_1$, $f_2 \in \mathcal{F}_2$, and $f_{12} \in \mathcal{F}_1 \otimes \mathcal{F}_2$. Choices for $\mathcal{F}_1$ and $\mathcal{F}_2$ 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.

Remark 4.3. Although (4.3) is a special case of the multilevel model decomposition (4.2) for which $x_i = x_i(t)$ (time-varying covariates), it is different to how longitudinal models are normally treated using a mixed effects model. As a multilevel model, longitudinal models treat the individuals as the groups or clusters (level two), and the time points as the various measurements within the clusters (level one).

4.1.4 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

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

According to 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, then Bergsma (2017) shows that f can be represented as

$$f(x) = \int_{-\infty}^{x} \beta(t) \, \mathrm{d}t \tag{4.4}$$

{eq:ipriorbrownianbridge}

where

$$\beta(t) = \sum_{i : x_i \leq t} w_i = \frac{f(x_{i_t+1}) - f(x_{i_t})}{x_{i_t+1} - x_{i_t}} \tag{4.5}$$

with $i_t = \max_{x_i \leq t} i$. 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 (2017), it is stated 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 $\mathrm{Var}\, \beta(x) = \mathrm{P}_n(X_n < x)\big(1 - \mathrm{P}_n(X_n < x)\big)$, 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 4.4. From (4.4), 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.

4.1.5 Regression with functional covariates

sec:regfunctionalcov

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,

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

sec:ipriorestimation

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

$$\begin{aligned} y_i &= \alpha + f_0(x_i) + \overbrace{\sum_{k=1}^{n} h_\eta(x_i, x_k) w_k}^{f(x_i)} + \epsilon_i \\ (\epsilon_1, \dots, \epsilon_n)^\top &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}^{-1}) \\ (w_1, \dots, w_n)^\top &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}), \end{aligned} \tag{4.6}$$

eq:model2}

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 $\mathbf{w} := (w_1, \dots, w_n)^\top$ is $\mathbf{w}|\mathbf{y} \sim \mathrm{N}_n(\tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$, where

$$\tilde{\mathbf{w}} = \boldsymbol{\Psi}\mathbf{H}_\eta \mathbf{V}_y^{-1}(\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{f}_0) \quad \text{and} \quad \tilde{\mathbf{V}}_w = \left(\mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}\right)^{-1} = \mathbf{V}_y^{-1}, \tag{4.7}$$

eq:posteriorw}

using the familiar notation that we introduced in Section 1.4. For a derivation, see Appendix A.1. 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 $\mathcal{F}_1, \dots, \mathcal{F}_p$, then at the very least, there are p scale parameters $\lambda_1, \dots, \lambda_p$ 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 (4.6):

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

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

ass:A3 A3 **Iid errors**. Assume identical and independent (iid) errors 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

sec:intercept

In most statistical models, an intercept is a necessary inclusion which aids interpretation. In the context of the I-prior model (4.6), 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 addition 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 $\mathcal{F}_0$, and thereby tensor summing this space to $\mathcal{F}$. 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. The second is to simply treat the intercept as a parameter of the model to be estimated. In any of the other RKHSs described in Chapter 2, an intercept would need to be added separately.

These two methods convey the same mathematical model, and there is very little difference in the way of interpretation, although estimation is entirely different. 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 $h + 1$. The intercept will then be implicitly taken care of without having dealt with it explicitly. However, it can be obtained by realising that for $\alpha \in \mathcal{F}_0$ 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 (4.6), $\mathrm{E}\, y_i = \alpha + f_0(x_i)$ for all $i = 1, \ldots, n$, so the maximum likelihood 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 $\hat{\alpha} = \bar{y} - \frac{1}{n}\sum_{i=1}^{n} f_0(x_i)$. Alternatively, the estimation of α under a fully Bayesian treatment is possible by assuming an appropriate hyperprior on it, such as a conjugate normal prior $\mathrm{N}(a, A^{-1})$. If so, the conditional posterior of α given $\mathbf{w}$, η, $\boldsymbol{\Psi}$ and f_0 is also normal with mean $\tilde{a}$ and variance $\tilde{A}$, where

$$\tilde{A} = \sum_{i,j=1}^{n} \psi_{ij} + A \quad \text{and} \quad \tilde{a} = \tilde{A}^{-1}\left(\sum_{i=1}^{n}[(\mathbf{y} - \mathbf{f}_0 - \mathbf{H}_\eta \mathbf{w})\boldsymbol{\Psi}]_i + Aa\right).$$

This fact can be used, say, in conjunction with a Gibbs sampling procedure treating the rest of the unknowns as random. Note that the posterior mean for α is

$$\mathrm{E}[\alpha|\mathbf{y}] = \mathrm{E}_{\mathbf{w}}\left[\mathrm{E}[\alpha|\mathbf{y}, \mathbf{w}]\right] = \frac{\sum_{i,j=1}^{n} \psi_{ij}(y_i - f_0(x_i)) + Aa}{\sum_{i,j=1}^{n} \psi_{ij} + A},$$

which, in the iid errors case, is seen to be a weighted sum of the ML estimate $\hat{\alpha}$ and the prior mean a. Unless there is a strong reason to add prior information to the intercept, the ML estimate seems to be the simplest approach. Assumption A1 implies a ML estimation of the intercept parameter.

Now, a note on the prior mean f_0. For kernels with the property that $h(x, x^*) \to 0$ as $D(x, x^*) \to \infty$ for $x \in \mathcal{X}_{\text{train}}$ and $x^* \in \mathcal{X}_{\text{new}}$ such as the SE kernel, this means that predictions outside the training set will be zero and thus rely on the prior mean f_0. However, all of the other kernels in this thesis, namely the fBm, canonical, and polynomial kernels, do not have this property—they instead use information provided by the training data to extrapolate predictions far away from the data set. A prior mean of zero seems reasonable and safe in the absence of any prior information, so long as the global and local properties of the regression function are understood with respect to the kernel chosen. $f_0 = 0$ also implies a complete reliance on the data rather than subjective prior belief of a suitable choice for f.

Of course, should it be felt appropriate, a non-zero function f_0 may be imposed as the prior mean. If $f_0(x) = \mu_0 \in \mathbb{R}$ for all $x \in \mathcal{X}$, then this basically implies another intercept in the model, if it is not already present. Note that when treating μ_0 as a

hyperparameter to be estimated, then this does not yield a fully identified model, and only $\alpha + \mu_0$ may be estimated.

4.2.2 Direct optimisation

Under assumptions A1 and A2, a direct optimisation of the parameters $\theta = \{\eta, \boldsymbol{\Psi}\}$ using the log-likelihood of θ is straightforward to implement. Denote $\boldsymbol{\Sigma}_\theta := \mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta + \boldsymbol{\Psi}^{-1} = \mathbf{V}_y$. From (4.6), the (marginal) log-likelihood of θ is given by

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y}|\mathbf{w}) p(\mathbf{w}) \, \mathrm{d}\mathbf{w} \\ &= -\frac{n}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_\theta| - \frac{1}{2} \tilde{\mathbf{y}}^\top \boldsymbol{\Sigma}_\theta^{-1} \tilde{\mathbf{y}}. \end{aligned} \tag{4.8}$$

{eq:marglog liky}

The term marginal refers to the fact that we are averaging out the random function represented by $\mathbf{w}$. Direct optimisation is typically done using conjugate gradients with a Cholesky decomposition on the covariance kernel to maintain stability, but we opt for an eigendecomposition of the kernel matrix $\mathbf{H}_\eta = \mathbf{V} \cdot \mathrm{diag}(u_1, \ldots, u_n) \cdot \mathbf{V}^\top$ instead. Further, under assumption A3 and since $\mathbf{H}_\eta$ is a symmetric matrix, we have that $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_n$, and thus

$$\mathbf{V}_y = \mathbf{V} \cdot \mathrm{diag}(\psi u_1^2 + \psi^{-1}, \ldots, \psi u_n^2 + \psi^{-1}) \cdot \mathbf{V}^\top$$

for which the inverse and log-determinant is easily obtainable. This method is relatively robust to numerical instabilities and is better at ensuring positive definiteness of the covariance kernel. The eigendecomposition is performed using the **Eigen** C++ template library and linked to **iprior** using **Rcpp** (Eddelbuettel and Francois, 2011). The hyperparameters are transformed by the **iprior** package so that an unrestricted optimisation using the quasi-Newton L-BFGS algorithm provided by `optim()` in R. Note that minimisation is done on the deviance scale, i.e., minus twice the log-likelihood. 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.

10. Show ridge in the log-likelihood plot.

Let $\mathbf{U}$ be the Fisher information matrix for $\theta \in \mathbb{R}^q$. Standard calculations (Appendix A.4) 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

$$u_{ij} = \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}_\theta^{-1} \frac{\partial \boldsymbol{\Sigma}_\theta}{\partial \theta_i} \boldsymbol{\Sigma}_\theta^{-1} \frac{\partial \boldsymbol{\Sigma}_\theta}{\partial \theta_j} \right)$$

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 errors for θ_k are the diagonal elements of $\mathbf{U}^{-1/2}$.

4.2.3 Expectation-maximisation algorithm

Evidently, (4.6) lends itself to resembling a random-effects model, for which the EM algorithm can easily be employed to estimate its hyperparameters. Assume A1 and A2 holds. By treating the complete data as $\{\mathbf{y}, \mathbf{w}\}$ and the w_i's as "missing", the tth iteration of the E-step entails computing

{eq:QfnEstep}

$$\begin{aligned} Q(\theta) &= \mathrm{E}_{\mathbf{w}}\left[\log p(\mathbf{y}, \mathbf{w}|\theta) \middle| \mathbf{y}, \theta^{(t)}\right] \\ &= \mathrm{E}_{\mathbf{w}}\left[\text{const.} - \frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{H}_\eta \mathbf{w})^\top \boldsymbol{\Psi}(\tilde{\mathbf{y}} - \mathbf{H}_\eta \mathbf{w}) - \frac{1}{2}\mathbf{w}^\top \boldsymbol{\Psi}^{-1}\mathbf{w} \,\middle|\, \mathbf{y}, \theta^{(t)}\right] \\ &= \text{const.} - \frac{1}{2}\tilde{\mathbf{y}}^\top \boldsymbol{\Psi}\tilde{\mathbf{y}} - \frac{1}{2}\operatorname{tr}\left(\overbrace{(\mathbf{H}_\eta \boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1})}^{\boldsymbol{\Sigma}_\theta}\tilde{\mathbf{W}}^{(t)}\right) + \tilde{\mathbf{y}}^\top \boldsymbol{\Psi}\mathbf{H}_\eta \tilde{\mathbf{w}}^{(t)}, \end{aligned} \tag{4.9}$$

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 (4.7), substituting for $\theta^{(t)} = \{\eta^{(t)}, \boldsymbol{\Psi}^{(t)}\}$ as appropriate. Note that (4.9) follows as a direct consequence of the results in Appendix A.1.

Now, assume that A3 holds. The M-step then assigns $\theta^{(t+1)}$ the value of θ which maximises the Q function above. This boils down to solving the first order conditions

{eq:emtheta}

$$\frac{\partial Q}{\partial \eta} = -\frac{1}{2}\operatorname{tr}\left(\frac{\partial \boldsymbol{\Sigma}_\theta}{\partial \eta}\tilde{\mathbf{W}}^{(t)}\right) + \psi \cdot \tilde{\mathbf{y}}^\top \frac{\partial \mathbf{H}_\eta}{\partial \eta}\tilde{\mathbf{w}}^{(t)} \tag{4.10}$$

{eq:empsi}

$$\frac{\partial Q}{\partial \psi} = -\frac{1}{2}\tilde{\mathbf{y}}^\top \tilde{\mathbf{y}} - \operatorname{tr}\left(\frac{\partial \boldsymbol{\Sigma}_\theta}{\partial \psi}\tilde{\mathbf{W}}^{(t)}\right) + \tilde{\mathbf{y}}^\top \mathbf{H}_\eta \tilde{\mathbf{w}}^{(t)} \tag{4.11}$$

equated to zero. As $\partial \boldsymbol{\Sigma}_\theta / \partial \psi = \mathbf{H}_\eta^2 - \psi^{-2}\mathbf{I}_n$, the solution to (4.11) for ψ is separable in η. Meaning, given values for η, the solution $\psi^{(t+1)}$ emits a closed form

{eq:closedformpsi}

$$\psi^{(t+1)} = \left\{\frac{\operatorname{tr}\tilde{\mathbf{W}}^{(t)}}{\tilde{\mathbf{y}}^\top \tilde{\mathbf{y}} + \operatorname{tr}(\mathbf{H}_\eta^2 \tilde{\mathbf{W}}^{(t)}) - 2\tilde{\mathbf{y}}^\top \mathbf{H}_\eta \tilde{\mathbf{w}}^{(t)}}\right\}^{1/2}. \tag{4.12}$$

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). The solution to (4.10) 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 L-BFGS algorithm.

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. The reason for this is that the Q function is generally convex in the parameters (at the very least, it is convex in each coordinate of θ, in most cases anyway). 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 (4.8), rather than the hierarchical specification (4.6). 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. The posterior sampling for the w_i's in (equivalently, the posterior Gaussian process $f(x) = \sum_{i=1}^{n} h_\lambda(x, x_i) w_i$) is performed using the normal posterior distribution in (4.7).

sec:compare estimation

4.2.5 Comparison of estimation methods

Consider a one-dimensional smoothing example. $n = 150$ data pairs (y_i, x_i) have been randomly sampled according to the true relationship

$$r_i = \text{const.} + \overbrace{0.35 \cdot \phi(x_i | 1, 0.8^2) + 0.65 \cdot \phi(x_i | 4, 1.5^2) + \mathbb{1}(x_i > 4.5) \cdot e^{1.25(x_i - 4.5)}}^{f_{\text{true}}(x_i)}, \tag{4.13}$$

{eq:example smoothingdata}

where $\phi(\cdot | \mu, \sigma^2)$ is the probability density function of the normal distribution with mean μ and variance σ^2. The observed y_i's are thought to be noisy versions of the true points, i.e. $y_i = r_i + \epsilon_i$, with ϵ_i following 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 4.1 plots the sampled points and the true regression function.

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 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

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

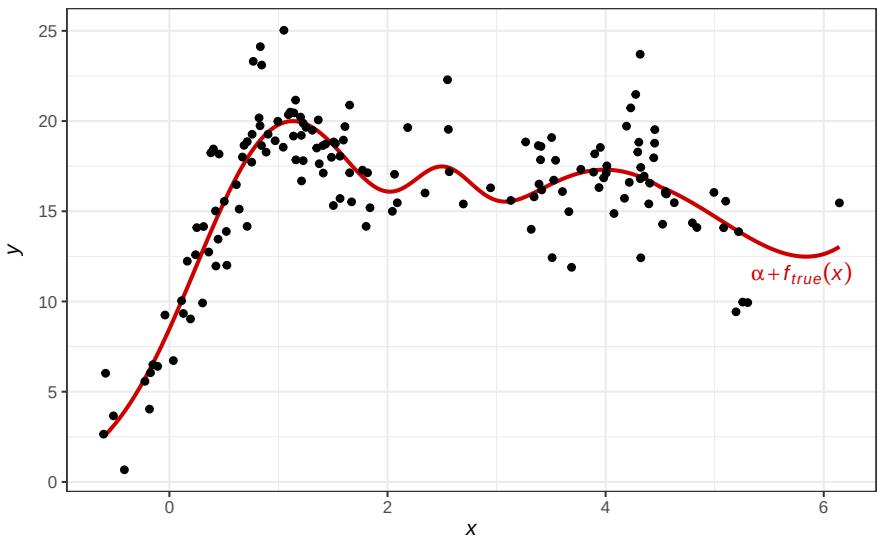


Figure 4.1: A plot of the sampled data points according to equation (4.13), with the true regression function superimposed.

fig:example smoothingdata

where $\mathrm{N}_+(\mu, \sigma^2)$ represents the *half-normal* distribution[2]. 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.

tab:compare methodsestimate

	Direct optimisation	EM algorithm	Hamiltonian MC
Intercept (α)	16.1 (NA)	16.1 (NA)	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

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

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.

4.3 Computational considerations

sec:ipriorcompcons

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 $\boldsymbol{\Sigma}_\theta = \mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}$, 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 $O(n^2)$. In what follows, assumptions A1–A3 hold.

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 $\mathbf{Q}\mathbf{Q}^\top$ 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. 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 $m \ll n$ of the data points.

Let $\mathbf{H}_\eta = \mathbf{V}\mathbf{U}\mathbf{V}^\top = \sum_{i=1}^n u_i \mathbf{v}_i \mathbf{v}_i^\top$ 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 $\mathcal{M}$ of size m of the n data points $\{1, \dots, n\}$, so that $\mathbf{H}_\eta$ may be approximated using the rank m matrix $\mathbf{H}_\eta \approx \sum_{i \in \mathcal{M}} \tilde{u}_i \tilde{\mathbf{v}}_i \tilde{\mathbf{v}}_i^\top$. Without loss of generality, reorder the rows and columns of $\mathbf{H}_\eta$ so that the data points indexed by $\mathcal{M}$ are used first:

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

In other words, the data points indexed by $\mathcal{M}$ forms the smaller $m \times m$ kernel matrix $\mathbf{A}$. Let $\mathbf{A} = \mathbf{V}_m \mathbf{U}_m \mathbf{V}_m^\top = \sum_{i=1}^m u_i^{(m)} \mathbf{v}_i^{(m)} \mathbf{v}_i^{(m)\top}$ 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

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

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

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

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. (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 $\hat{\mathbf{V}}\hat{\mathbf{V}}^\top = \mathbf{I}_n$. Estimating I-prior models with the Nyström method including the orthogonalisation step takes roughly $O(nm^2)$ time and $O(nm)$ storage.

11. Attempt to prove this.

The issue of selecting the subset $\mathcal{M}$ remains. The simplest method, and that which is implemented in the **iprior** package, would be to uniformly sample a subset of size m 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 m points, so as to reduce some error criterion relating to the quality of approximation. For a brief review of more sophisticated methods of selecting $\mathcal{M}$, see Rasmussen and Williams (2006, §8.1, pp. 173–174).

4.3.2 An efficient EM algorithm

sec:efficientEM1

The evaluation of the Q function in (4.9) 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. 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.

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 $h_1, \ldots, h_p$. 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 as simple as a list containing the indices to multiply out. For the four scenarios above, the list $\mathcal{Q}$ is

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 $\mathbf{A}_1, \ldots, \mathbf{A}_q$ from the application of instruction set 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'. In the **iprior** package, kernel loading is performed using the `kernL()` command.

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.

As a remark, we have achieved efficiency at the expense of storage and a potentially long initialisation phase of kernel loading. The storing of the kernel matrices a can be very expensive, especially if the sample size is very large. On the bright side, once the kernel matrices are stored in 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.

sec:expfamEM

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* (see Appendix A.2 for details). Certain I-prior models emit this property, namely regression functions belonging to the full or limited ANOVA RKKS, and we describe its estimation 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 is

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

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 applying the instructions of $\mathcal{S}_k$. Now, we have

$$
r_k = \frac{1}{\lambda_k} \sum_{i=1}^{|\mathcal{R}_k|} \zeta_{ik} e_{ik} \quad \text{and} \quad s_k = \sum_{i=1}^{|\mathcal{S}_k|} \xi_{ik} a_{ik}.
$$

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, 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 ψ and $\lambda_{-k} = \{\lambda_1, \ldots, \lambda_p\} \backslash \{\lambda_k\}$:

{eq:loglikl ambdak}

$$\begin{aligned} L(\lambda_k | \mathbf{y}, \mathbf{w}, \lambda_{-k}, \psi) &= \text{const.} - \frac{1}{2} \operatorname{tr}\left((\psi \mathbf{H}_\eta^2 + \psi^{-1} \mathbf{I}_n) \mathbf{w}\mathbf{w}^\top \right) + \psi \tilde{\mathbf{y}}^\top \mathbf{H}_\eta \mathbf{w} \qquad (4.14) \\ &= \text{const.} - \lambda_k^2 \cdot \frac{\psi}{2} \operatorname{tr}(\mathbf{R}_k^2 \mathbf{w}\mathbf{w}^\top) + \lambda_k \cdot \left(\psi \tilde{\mathbf{y}}^\top \mathbf{R}_k \mathbf{w} - \frac{\psi}{2} \operatorname{tr}(\mathbf{U}_k \mathbf{w}\mathbf{w}^\top) \right). \end{aligned}$$

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 (4.7). That the λ_k's have closed-form expressions, together with the closed-form expression for ψ in (4.12), 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 algorithm is summarised in Algorithm 1. The exponential family EM for ANOVA-type I-prior models require $O(n^3)$ 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. 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.

While the exponential family EM algorithm takes similar computational time as the efficient EM algorithm described in Section 4.3.2, there is one compelling reason to consider Algorithm 1: 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

alg:EM2

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**
      $T_{1k} \leftarrow \frac{1}{2} \operatorname{tr}(\mathbf{R}_k^2 \tilde{\mathbf{W}})$
      $T_{2k} \leftarrow \tilde{\mathbf{y}}^\top \mathbf{R}_k \tilde{\mathbf{w}} - \frac{1}{2} \operatorname{tr}(\mathbf{U}_k^2 \tilde{\mathbf{W}}^\top)$
      $\lambda_k^{(t+1)} \leftarrow T_{2k} / 2 T_{1k}$
    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

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 log-likelihood prior involving λ_k added on. Importantly, viewed as a posterior log-density for λ_k, the posterior density for λ_k would also be a normal distribution. The EM as a whole would then generate maximum a posteriori (MAP) estimates for the parameters. Although not shown here, similar conjugacy benefits for the ψ parameter can be argues, whereby the gamma distribution is the density in question. The usual EM algorithm without using any priors can be viewed as using improper priors for the parameters, i.e. $p(\lambda_k) \propto$ const. and $p(\psi) \propto$ const..

In the next chapter on binary and multinomial regression using I-priors, the exponential family EM algorithm described here is especially relevant, as it is connected to the variational Bayesian algorithm (Bernardo et al., 2003) that will be used for estimating the models described therein.

Remark 4.5. Earlier, we restricted attention to ANOVA RKKS. Hopefully, it is now apparent that ANOVA kernels are a requirement for Algorithm 1 to work easily. As soon as higher degrees of the λ_k's come into play, e.g. using the polynomial kernel, then the ML estimate for λ_k involve solving a polynomial of degree $2d-1$ the 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 $\lambda_k's$, is lost.

4.3.4 Accelerating the EM algorithm

A criticism of the EM algorithm is that it may take many iterations to converge. Several novel ideas have been looked at in a bid to 'accelerate the EM algorithm', as it were. One such approach, which does not require any amendment to the particular EM algorithm at hand, is called the *monotonically over-relaxed EM algorithm* (MOEM) by Yu (2012).

The idea of MOEM is as follows. At every iteration of the MOEM, perform as usual the E-step and M-step to obtain an updated parameter value $\theta_{\text{EM}}^{(t+1)}$. Instead of using this update value of the parameter, modify it instead, and use

$$\theta^{(t+1)} = (1+\omega)\theta_{\text{EM}}^{(t+1)} - \omega\theta^{(t)},$$

where ω is an *over-relaxation* parameter. Under mild conditions, among them that $Q(\theta^{(t+1)}) > Q(\theta^{(t)})$, the MOEM estimate does not decrease the log-likelihood at each step. This condition is a slight inconvenience to check under the usual EM algorithm, but is a great companion to exponential family EM algorithm. From (4.14), we see that $Q(\lambda_k) = \mathrm{E}_{\mathbf{w}}\left[L(\lambda_k|\theta\backslash\{\lambda_k\})|\mathbf{y}, \theta^{(t)}\right]$ is quadratic in λ_k, therefore any $\omega \in [0,1]$ will maintain monotonicity of the EM algorithm.

4.4 Post-estimation

sec:ipriorpostest

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. The plots that are shown in this subsection is a continuation of the example from subsection 4.2.5.

Recall that for the I-prior model (4.6), the 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 multivariate-normal mean and variance of the $\tilde{w}_i$'s given in (4.7). 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

$$p\big(f(x)|\mathbf{y}\big) \sim \mathrm{N}\left(\mathbf{h}_{\hat{\eta}}(x)\hat{\mathbf{w}}, \mathbf{h}_{\hat{\eta}}(x)^\top \big(\mathbf{H}_{\hat{\eta}}\hat{\mathbf{\Psi}}\mathbf{H}_{\hat{\eta}} + \hat{\mathbf{\Psi}}^{-1}\big)^{-1}\mathbf{h}_{\hat{\eta}}(x)\right) \tag{4.15}$$

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 also of interest. 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 $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ 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 $i \in \{1, \ldots, n\}$, 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

$$\mathrm{E}[y_{\text{new}}|\mathbf{y}] = \hat{\alpha} + \mathrm{E}\left[f(x_{\text{new}})|\mathbf{y}\right] + \text{correction term} \tag{4.16}$$

and

$$\operatorname{Var}[y_{\text{new}}|\mathbf{y}] = \operatorname{Var}\left[f(x_{\text{new}})|\mathbf{y}\right] + \psi_{\text{new}}^{-1} + \text{correction term}. \tag{4.17}$$

A derivation is presented in section A.3. 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 the (mean) predicted value y_{new} using

$$\mathrm{E}[y_{\text{new}}|\mathbf{y}] \pm \Phi^{-1}(1-\alpha/2) \cdot \operatorname{Var}[y_{\text{new}}|\mathbf{y}]^{\frac{1}{2}},$$

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 (4.6) is a normal distribution. Suppose that we are interested in comparing two candidate models M_1 and M_2, each with the parameter

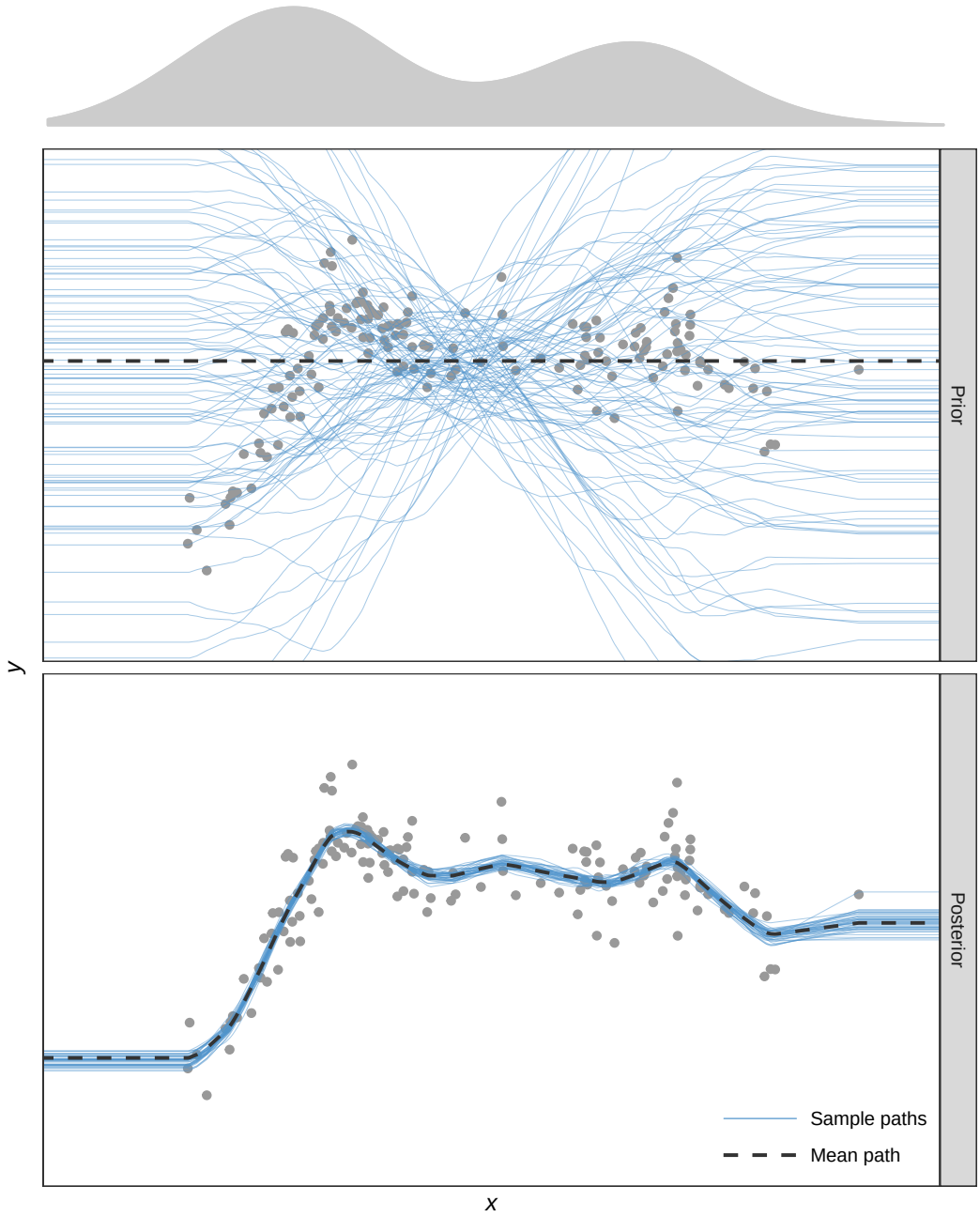


Figure 4.2: Prior (top) and posterior (bottom) sample path realisations of regression functions drawn from their respective distributions when $\mathcal{F}$ is a fBm-0.5 RKHS. At the very top of the figure, a smoothed density estimate of the x's is overlaid. In regions with few data points (near the centre), there is little Fisher information, and hence a conservative prior closer to zero, the prior mean, for this region.

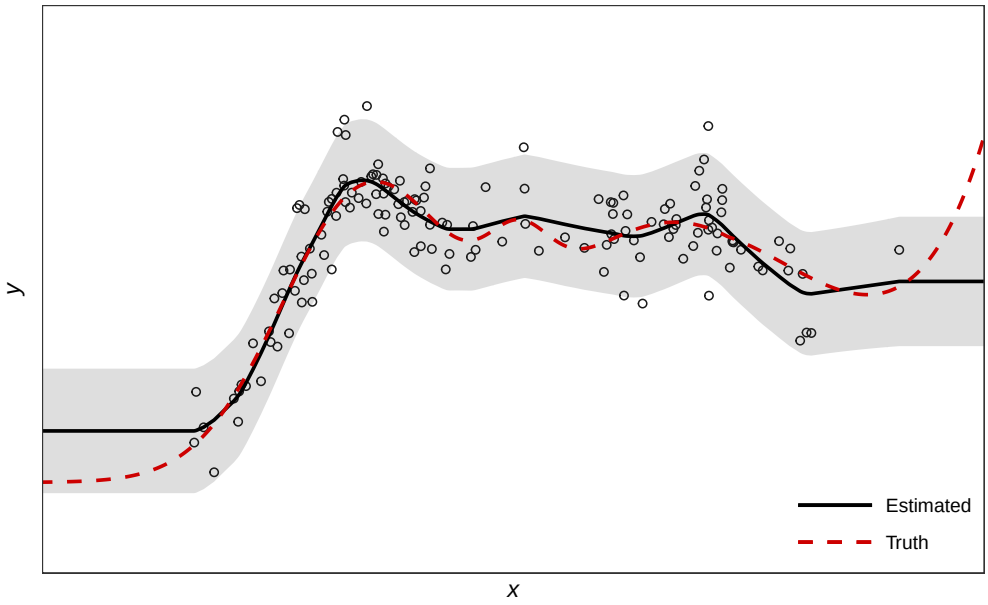


Figure 4.3: The estimated regression line (solid black) is the posterior mean estimate of the regression function (shifted by the intercept), which also gives the posterior mean estimate for the responses y. The shaded region is the 95% credibility interval for predictions. The true regression line (dashed red) is shown for comparison.

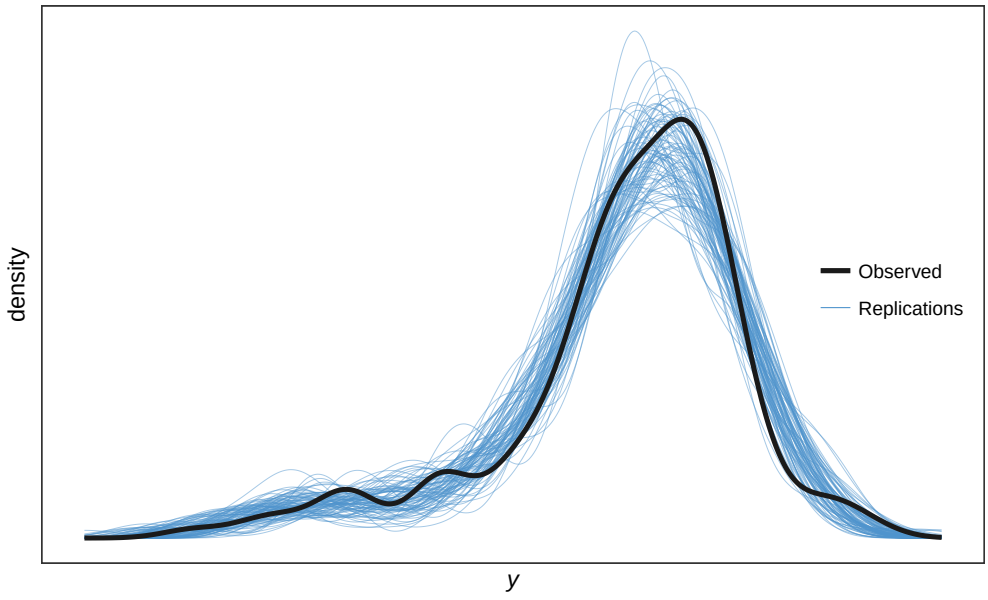


Figure 4.4: Posterior predictive density checks of the responses: repeated sampling from the posterior density of the y_i's and plotting their densities allows us to compare model predictions against observed samples.

set θ_1 and θ_2. 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 θ_2 and θ_1. This is assuming model M_1 is nested within M_2, 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 4.6. 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 to model M_2 is defined as

$$\mathrm{BF}(M_1, M_2) = \frac{\int p(\mathbf{y}|\theta_1, \mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}}{\int p(\mathbf{y}|\theta_2, \mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}}.$$

The word 'empirical' stems from the fact that the parameters are estimated via an empirical Bayes approach (maximum marginal likelihood). This approach is fine when the number of comparisons to be made is small, but can be computationally unfeasible when many marginal likelihoods need to be pairwise compared. In Chapter 6, we explore a fully Bayesian approach to explore the entire model space for the special case of linear models.

4.5 Examples

sec:ipriorexamples

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. In all of these examples, A1–A3 were assumed.

4.5.1 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 (4.13) 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 goes 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 above 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
```

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 approx-

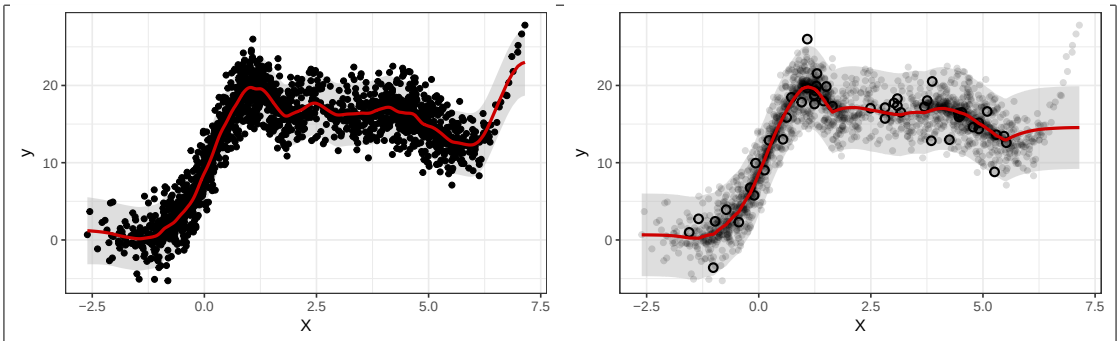


Figure 4.5: Plot of predicted regression function for the full model (left) and the Nyström approximated method (right). For the Nyström plot, the data points that were active are shown by circles with bold outlines.

fig:nystrom.plot

imation procedure. Nevertheless, we see from Figure 4.5 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 under 15 minutes to converge, while the Nyström method took just seconds. Storage savings is significantly higher with the Nyström method as well.

```
R> get_time(mod.full); get_size(mod.full, units = "MB")

## 14.63474 mins
## 128.2 MB

R> get_time(mod.nys); get_size(mod.nys)

## 1.324355 secs
## 965.2 kB
```

4.5.2 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[3] 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 a 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. J. C. 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 gives the intended effect of a linear mixed-effects model, which is thought to be suitable

[3] This data is available in the R package **nlme** (J. Pinheiro et al., 2017).

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 `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: 3.043089 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 4.6 does show an almost horizontal line for each `Lot`.

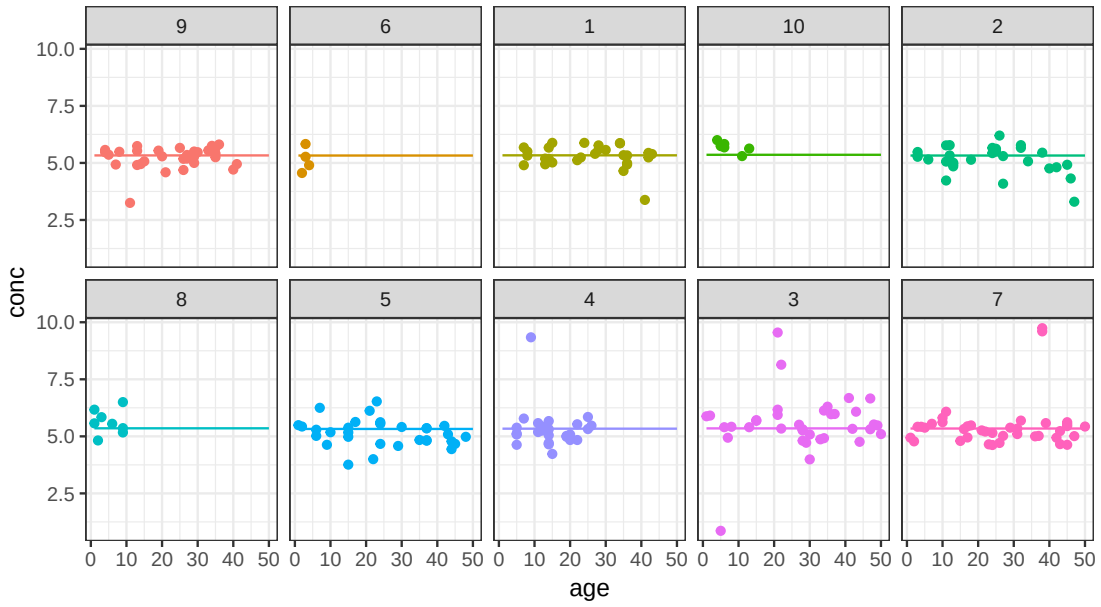


Figure 4.6: Plot of fitted regression line for the I-prior model on the IGF data set, separated into each of the 10 lots.

fig:IGF.mod.iprior.plot

The standard random effects model, as explored by Davidian and Giltinan (1995) and J. C. Pinheiro and Bates (2000), is

$$\texttt{conc}_{ij} = \beta_{0j} + \beta_{1j}\texttt{age}_{ij} + \epsilon_{ij}$$
$$\begin{pmatrix}\beta_{0j}\\ \beta_{1j}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{pmatrix}\right)$$
$$\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$$

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 $= 0.616$), and arrive at the same conclusion as in the I-prior model. 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.

```
R> (mod.lmer <- lmer(conc ~ age + (age | Lot), IGF))

## Linear mixed model fit by REML ['lmerMod']
## 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> 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. 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 less number of parameters to fit a simi-

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.

tab:igf

Parameter	`iprior`	`lmer`
σ_0	0.012	0.083
σ_1	0.000	0.008
ρ_{01}	0.690	-1.000

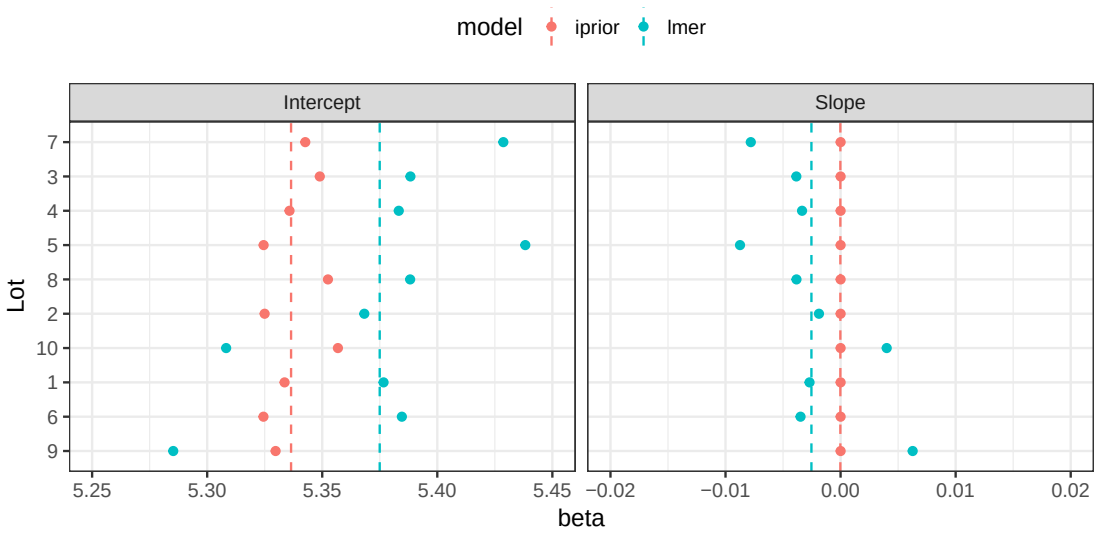


Figure 4.7: A comparison of the estimates for random intercepts and slopes (denoted as points) using the I-prior model and the standard random effects model. The dashed vertical lines indicate the fixed effect values.

fig:IGF.plot.beta

lar 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 subsection 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 $\hat{f}^j(x_{ij})$, $i = 1, \ldots, n_j$. 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 4.7 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.

4.5.3 Longitudinal data analysis

sec:cows

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, 2016) 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> 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 "A","B": 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.

tab:cowmodel

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`

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

$$\texttt{weight} = \alpha + f(\texttt{id}, \texttt{group}, \texttt{time}) + \epsilon$$
$$\epsilon \sim \mathrm{N}(0, \psi^{-1}).$$

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 0.5. 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

tab:cowresults

Table 4.4: Summary of the five I-prior models fitted to the cow data set.

Model	Formula (`weight ~ ...`)	Log-likelihood	Error S.D.	Number of parameters
1	`time`	-2789.23	16.33	1
2	`id * time`	-2789.60	16.35	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

algorithm in the `mixed` routine 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, kernel = "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
```

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 $D = -2(-2789.60-(-2270.85)) = 1037.49$ 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 than Model 2.

We can next investigate whether the treatment effect differs among cows by comparing Model 5 against Model 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

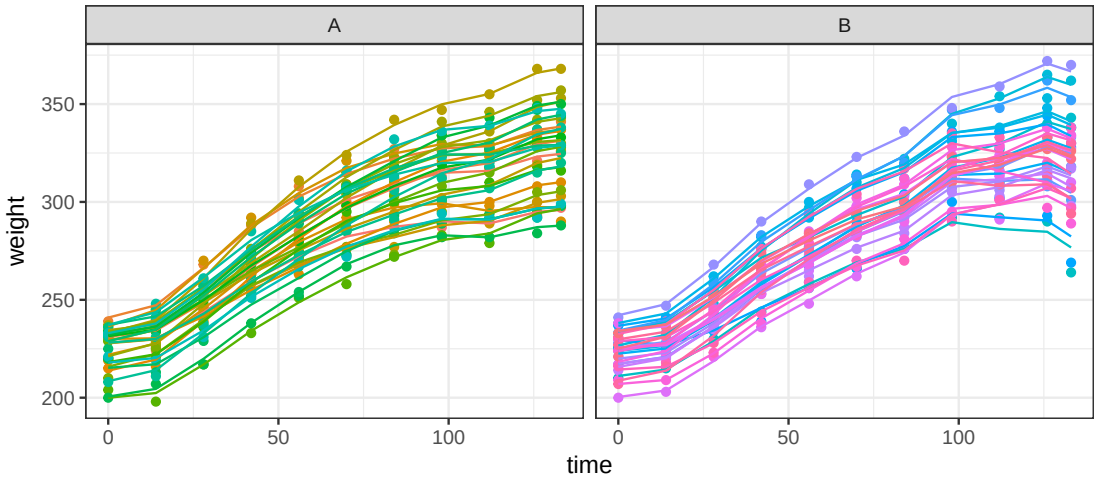


Figure 4.8: A plot of the I-prior fitted regression curves from Model 5. In this model, growth curves differ among cows and by treatment effect (with an interaction between cows and treatment effect), thus producing these 60 individual lines, one for each cow, split between their respective treatment groups (A or B).

fig:cows.plot

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 4.8.

4.5.4 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[4] 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 4.9 shows a sample of 10 such spectrometric curves.

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.

[4]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).

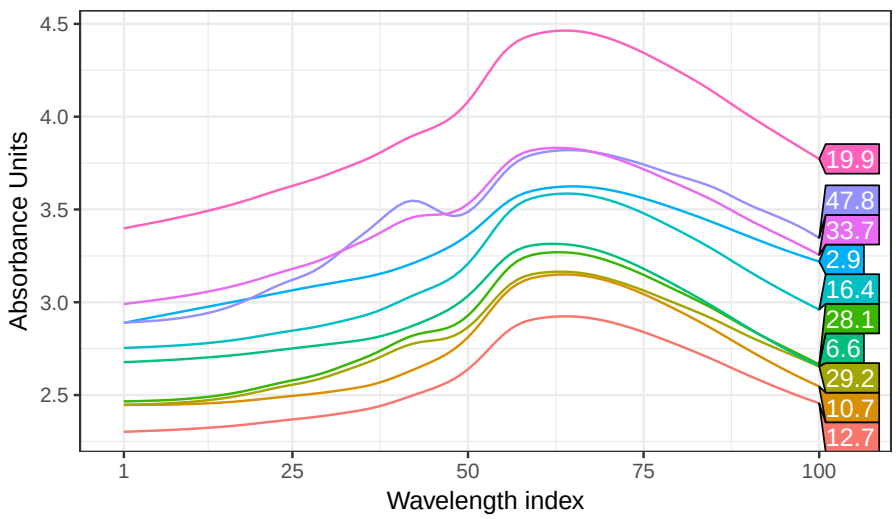


Figure 4.9: Sample of spectrometric curves used to predict fat content of meat. For each meat sample the data consists of a 100 channel spectrum of absorbances and the contents of moisture, fat (numbers shown in boxes) and protein measured in percent. The absorbance is $-\log 10$ of the transmittance measured by the spectrometer. The three contents, measured in percent, are determined by analytic chemistry.

fig:tecator.data

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 (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 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 results of all of these methods from the various references.

Assuming a regression model as in (4.6), 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 4.9, 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 Section 4.1.5. 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`). Investigating further, it seems that estimates at these large values give poor training and test error rates, though likelihood values here are high (local optima). To get around 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
## -680.4637 -231.5440 -231.5440 -231.5440 -231.5440 -680.4637 -680.4637
##     Run 8
## -231.5440
## Continuing on Run 3
## final  value 115.771932
## converged
## Log-likelihood value: -231.544
##
##      lambda lengthscale         psi
##    96.10718     0.09269     6.15429
```

Predicted values of the test data set can be obtained using the `predict()` function. 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
```

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.

tab:tecator

	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		
Linear	0.18	2.36
Smooth (SE-7.04)	0.17	2.10
Others		
Neural network[a]		0.36
Kernel smoothing[b]		1.49
Single/multiple indices model[c]		1.55
Sliced inverse regression		0.90
Sliced average variance estimation		1.70
MARS[d]		0.88
Partial least squares[d]		1.01
CSEFAM[d]		0.85

[a] Neural network best results with automatic relevance determination (ARD) quoted.
[b] Data set used was a 160/55 training/test split.
[c] These are results of a leave-one-out cross-validation scheme.
[d] Data set used was an extended version with $n = 240$, and a random 185/55 training/test split.

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[5].

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

[5] GPR models were fit using `gausspr()` in **kernlab**.

Boots (2017) also suggests a variational algorithm with linear complexity for GPR-type models.

2. **Accelerating the EM algorithm further**. 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.

4.7 Miscellanea

4.7.1 Similarity to the g-prior

misc:gprior

The I-prior for $\boldsymbol{\beta}$ 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 argues 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. 80), the Fisher information for $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathcal{I}_g(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij}(\mathbf{X}^\top\boldsymbol{\Psi}\mathbf{X})^{-1}\mathbf{x}_i \otimes (\mathbf{X}^\top\boldsymbol{\Psi}\mathbf{X})^{-1}\mathbf{x}_j \\ &= (\mathbf{X}^\top\boldsymbol{\Psi}\mathbf{X})^{-1}\cancel{(\mathbf{X}^\top\boldsymbol{\Psi}\mathbf{X})\,(\mathbf{X}^\top\boldsymbol{\Psi}\mathbf{X})^{-1}} \\ &= (\mathbf{X}^\top\boldsymbol{\Psi}\mathbf{X})^{-1},\end{aligned}$$

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 commonly 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.

misc:multilevelmodels

4.7.2 Multilevel models

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 $w_{i'j'} \sim \text{N}(0, \psi^{-1})$.

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:

$$\begin{aligned}
\operatorname{Cov}(\beta_{0j}, \boldsymbol{\beta}_{1j}) &= \lambda_1 \lambda_2^2 \operatorname{Cov} \left(\frac{1}{p_j} \mathbf{1}_{n_j}^\top \mathbf{w}_j - \mathbf{1}_n^\top \mathbf{w}, \frac{1}{p_j} \tilde{\mathbf{X}}_j^\top \mathbf{w}_j - \tilde{\mathbf{X}}^\top \mathbf{w} \right) \\
&= \psi \lambda_1 \lambda_2^2 \left(\cancel{\mathbf{1}_n^\top \tilde{\mathbf{X}}}^{0} + \frac{1}{p_j^2} \mathbf{1}_{n_j}^\top \tilde{\mathbf{X}}_j - \frac{2}{p_j} \mathbf{1}_{n_j}^\top \tilde{\mathbf{X}}_j \right) \\
&= n\psi \lambda_1 \lambda_2^2 \left(\left(\frac{1}{p_j} - 2 \right) \frac{1}{n_j} \sum_{i=1}^{n_j} (\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}) \right) \\
&= n\psi \lambda_1 \lambda_2^2 \left(\frac{1}{p_j} - 2 \right) (\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}})
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{Cov}(\beta_{0j}, \boldsymbol{\beta}_{1j'}) &= \lambda_1\lambda_2^2 \operatorname{Cov}\left(\frac{1}{p_j}\mathbf{1}_{n_j}^\top \mathbf{w}_j - \mathbf{1}_n^\top \mathbf{w}, \frac{1}{p_{j'}}\tilde{\mathbf{X}}_{j'}^\top \mathbf{w}_{j'} - \tilde{\mathbf{X}}^\top \mathbf{w}\right) \\
&= \psi\lambda_1\lambda_2^2 \left(\cancelto{0}{\mathbf{1}_n^\top \tilde{\mathbf{X}}} + \frac{1}{p_j p_{j'}}\mathbf{1}_{n_j}^\top \cancelto{0}{\operatorname{Cov}(\mathbf{w}_j, \mathbf{w}_{j'})} \tilde{\mathbf{X}}_{j'} - \frac{1}{p_j}\mathbf{1}_{n_j}^\top \tilde{\mathbf{X}}_j - \frac{1}{p_{j'}}\mathbf{1}_{n_{j'}}^\top \tilde{\mathbf{X}}_{j'}\right) \\
&= n\psi\lambda_1\lambda_2^2 \left(-\frac{1}{n_j}\sum_{i=1}^{n_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}) - \frac{1}{n_{j'}}\sum_{i=1}^{n_{j'}}(\mathbf{x}_i^{(j')} - \bar{\mathbf{x}})\right) \\
&= n\psi\lambda_1\lambda_2^2 \left(2\bar{\mathbf{x}} - \bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}}^{(j')}\right).
\end{aligned}$$

4.7.3 A recap on the exponential family EM algorithm

apx:expem

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

$$p(\mathbf{z}|\boldsymbol{\eta}) = B(\mathbf{z}) \exp\big(\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z})\rangle - A^*(\boldsymbol{\eta})\big) \tag{4.18}$$

{eq:pdfexpfamnat}

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

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{z}|\boldsymbol{\eta}) = \mathbf{T}(\mathbf{z}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = 0. \tag{4.19}$$

{eq:expEM1}

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

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}|\boldsymbol{\eta}) &= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}|\boldsymbol{\eta}) \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} \left(\int p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \, \mathrm{d}\mathbf{w} \right) \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \int \left(\frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) \mathrm{d}\mathbf{w} \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \int \left(p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) \mathrm{d}\mathbf{w} \\
&= \int \left(\mathbf{T}(\mathbf{y}, \mathbf{w}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) \right) p(\mathbf{w}|\mathbf{y}, \boldsymbol{\eta}) \, \mathrm{d}\mathbf{w} \\
&= \mathrm{E}_{\mathbf{w}} \left[\mathbf{T}(\mathbf{y}, \mathbf{w}) | \mathbf{y} \right] - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) \qquad (4.20)
\end{aligned}$$

{eq:expEM2}

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 (A.2) and that of (A.3) 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 (A.3) 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 6.

Algorithm 2 Exponential family EM

alg:EM3

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 6 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 (A.1), 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 (A.3) when obtaining ML estimate of $\boldsymbol{\eta}$. Thus, Q_t is maximised by the solution to line 4 in Algorithm Algorithm 6.

4.7.4 A brief introduction to Hamiltonian Monte Carlo

misc:hmc

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 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 Radford M Neal et al. (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 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

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial H}{\partial z} \quad \text{and} \quad \frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{\partial H}{\partial x}, \tag{4.21}$$ {eq:hamilton1}

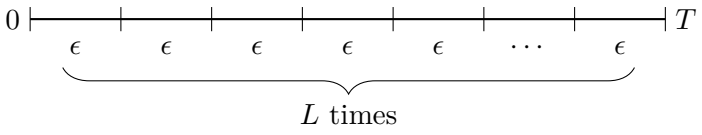
where $H = H(x, z)$ is called the Hamiltonian of the system. The Hamiltonian is an operator which encapsulates the total energy of the system. In a closed system, one can express the sum of operators corresponding to the kinetic energy $K(p)$ and the potential energy $U(z)$ of the system

$$H(x, z) = K(z) + U(x). \tag{4.22}$$ {eq:hamilton2}

Substituing (4.22) into (4.21), we get the system of partial differential equations (PDEs)

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial}{\partial z} K(z) \quad \text{and} \quad \frac{\mathrm{d}z}{\mathrm{d}t} = -\frac{\partial}{\partial x} U(x). \tag{4.23}$$ {eq:hamilton3}

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 (4.22) 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 $K(z) = z^\top M^{-1} z/2$[6], 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 3.

Algorithm 3 Hamiltonian Monte Carlo

alg:hmc

1: **initialise** $x^{(0)}$, $z^{(0)}$ and choose values for L, ϵ and M
2: **while** not converged **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

$$A = \frac{p(x^*, z^*)}{(x^{(t)}, z^{(t)})} = \exp\Big(H(x, z) - H(x^{(t)}, z^{(t)})\Big)$$

6: **end while**
7: **return** Samples $\big\{x^{(t)} \,|\, t = 1, 2, \ldots\big\}$

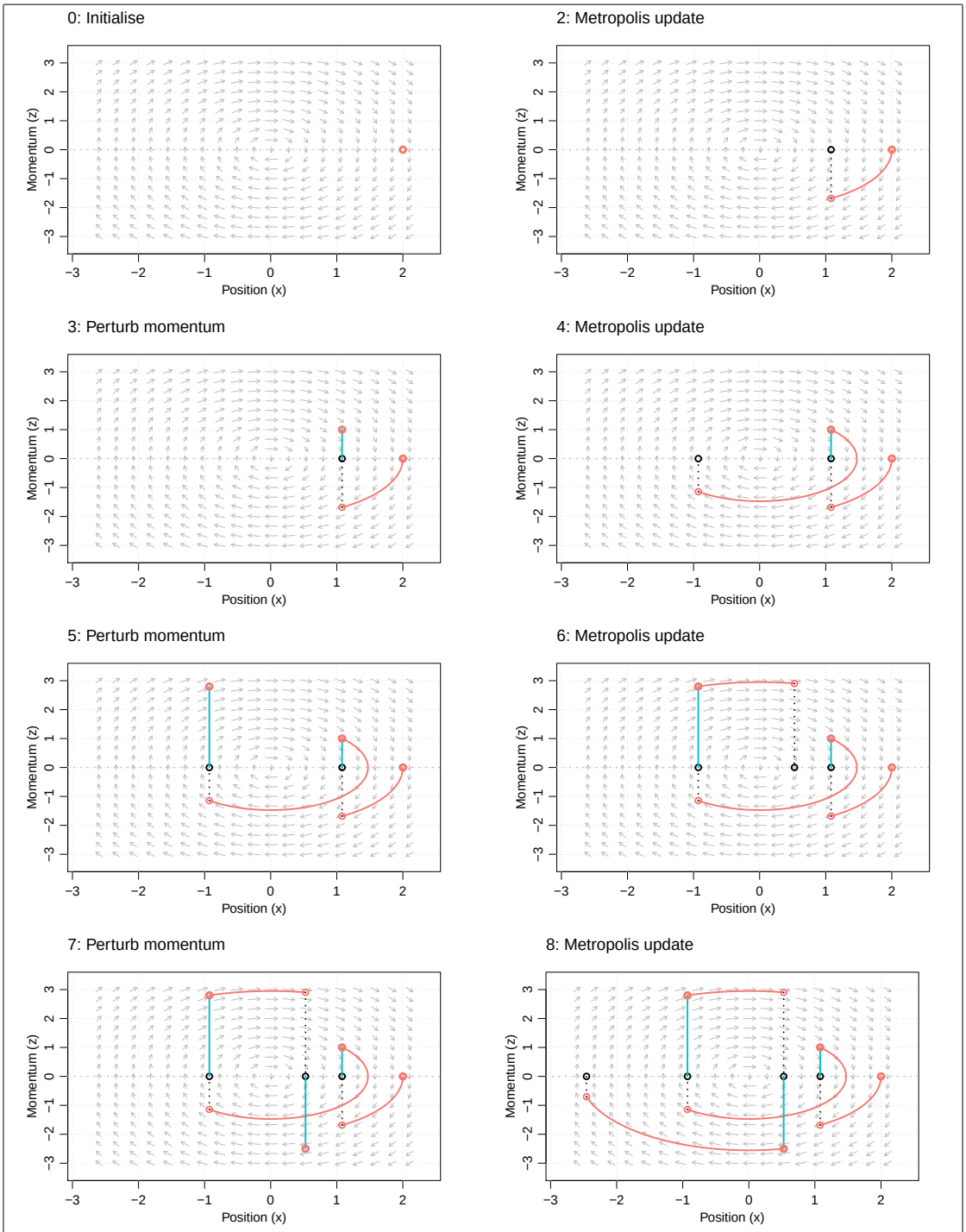


Figure 4.10: A phase diagram schematic of HMC sampling in one-dimension. At step 0, initialise values for momentum and position. At step 1, simulate movement using Hamiltonian dynamics, accept position and discard momentum. At step 2, perturb momentum using a normal density, then repeat.

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.

[6]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.

Chapter 5

I-priors for categorical responses

In a regression setting such as (1.1), consider polytomous response variables $y_1, \ldots, y_n$, where each y_i takes on exactly one of the values from the set of m possible choices $\mathcal{M} = \{1, \ldots, m\}$. 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 often interested in discrete choice models to explain and predict choices between several alternatives, such as consumer choice 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 normality assumption (1.2) is not entirely appropriate anymore. 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, 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, a 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, 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. We explore a fully Bayesian approach to estimate I-probit models using *variational inference*. The main idea is to replace the difficult posterior distribution with an approximation that is tractable. Working in a Bayesian setting together with variational inference allows us to estimate the model much faster than traditional MCMC sampling methods, 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 relevant posteriors, which are typically made up of densities which are familiar and readily available in software.

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 5.8. 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.

sec:iprobit naive

5.1 A naïve model

A naïve application of the normal I-prior methodology to fit categorical data is insightful for the upcoming sections. Suppose, as before, we observe data $\{(y_1, x_1), \dots, (y_n, x_n)\}$ where each $x_i \in \mathcal{X}$, for $i = 1, \dots, n$. Here, the responses are categorical $\mathbf{y}_i \in \{1, \dots, m\} =: \mathcal{M}$, and additionally, write $y_{i\cdot} = (y_{i1}, \dots, y_{im})^\top$ 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

$$\begin{aligned} y_{ij} &= \alpha + \alpha_j + f_j(x_i) + \epsilon_{ij} \\ (\epsilon_{i1}, \dots, \epsilon_{im})^\top &\overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}). \end{aligned} \tag{5.1}$$

{eq:naivecl assmod}

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 (5.1) on the set $\mathcal{X} \times \mathcal{M}$ as $f(x_i, j) = \alpha_j + f_j(x_i)$. This regression function can be seen as 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 $b : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then, the ANOVA RKKS $\mathcal{F}$ possesses the reproducing kernel $h : (\mathcal{X} \times \mathcal{M})^2 \to \mathbb{R}$ as defined by

$$b_\eta\big((x, j), (x', j')\big) = a(j, j') + a(j, j')h_\eta(x, x'). \tag{5.2}$$

{eq:anovacl ass}

The kernel h_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 2.34). With $J \sim \mathrm{P}$, a probability measure over $\mathcal{M}$,
$$a(j, j') = \frac{\delta_{jj'}}{\mathrm{P}(J = j)} - 1.$$

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

$$a(j, j') = \delta_{jj'}.$$

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. We have a slight preference for the identity kernel, which lends itself as being easy to handle computationally. 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.

As a remark, 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 2.37). 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,

$$\begin{aligned}\sum_{j=1}^m y_{ij} &= \sum_{j=1}^m (\alpha + \alpha_j + f_j(x_i)) \\ &= m\alpha + \cancelto{0}{\sum_{j=1}^m \alpha_j} + \cancelto{0}{\sum_{j=1}^m f_j(x_i)}\end{aligned}$$

and since $\sum_{j=1}^m y_{ij} = 1$, we have that $\alpha = 1/m$ and can thus be fixed to resolve identification. The Pearson RKHS will contain zero mean functions, but the RKHS of constant functions induced by the identity kernel may not. If this is the case, then it should be ensured that $\sum_{j=1}^m \alpha_j = 0$ in other ways; perhaps, as a requirement during estimation.

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

$$\begin{aligned} f(x_i, j) &= \sum_{j'=1}^m \sum_{i'=1}^n a(j, j')\big(1 + h_\eta(x_i, x_{i'})\big) w_{i'j'} \\ (w_{i'1}, \ldots, w_{i'm})^\top &\overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}) \end{aligned} \tag{5.3}$$

{eq:naiveclassiprior}

assuming a zero prior mean $f_0(x, j) = 0$. 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 $\mathbf{H}_\eta$ the $n \times n$ kernel matrix with entries supplied by $1 + h_\eta$, and $\mathbf{A}$ the $m \times m$ matrix with entries supplied by a. From (5.3), we have that

$$\mathbf{f} = \mathbf{H}_\eta \mathbf{w} \mathbf{A} \in \mathbb{R}^{n \times m},$$

and thus $\operatorname{vec} \mathbf{f} \sim \mathrm{N}_{nm}(\mathbf{0}, \mathbf{A}\boldsymbol{\Psi}\mathbf{A} \otimes \mathbf{H}_\eta^2)$. 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

$$\operatorname{vec} \mathbf{y} \sim \mathrm{N}_{nm}\left(\operatorname{vec} \boldsymbol{\alpha}, (\mathbf{A}\boldsymbol{\Psi}\mathbf{A} \otimes \mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \right) \tag{5.4}$$

and

$$\operatorname{vec} \mathbf{y} | \operatorname{vec} \mathbf{w} \sim \mathrm{N}_{nm}\left(\operatorname{vec}(\boldsymbol{\alpha} + \mathbf{H}_\eta \mathbf{w} \mathbf{A}), (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \right). \tag{5.5}$$

which can then be estimated using the methods described in Chapter 4.

When using the identity kernel 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

$$\begin{aligned} \operatorname{Var}(\operatorname{vec} \mathbf{y}) &= (\psi \mathbf{I}_m \otimes \mathbf{H}_\eta^2) + (\psi^{-1} \mathbf{I}_m \otimes \mathbf{I}_n) \\ &= (\mathbf{I}_m \otimes \psi \mathbf{H}_\eta^2) + (\mathbf{I}_m \otimes \psi^{-1} \mathbf{I}_n) \\ &= \mathbf{I}_m \otimes (\overbrace{\psi \mathbf{H}_\eta^2 + \psi^{-1} \mathbf{I}_n}^{\mathbf{V}_y}). \end{aligned}$$

which implies independence and identical variances $\mathbf{V}_y$ for the vectors $(y_{1j}, \dots, y_{nj})^\top$ for each class $j = 1, \dots, m$. Evidently, this stems from the implied independence structure of the prior on f too, since now $\operatorname{Var}(\operatorname{vec} \mathbf{f}) = \operatorname{diag}(\psi \mathbf{H}_\eta^2, \dots, \psi \mathbf{H}_\eta^2)$, which could be interpreted as having independent and identical I-priors on the regression functions for each class $\mathbf{f}_{\cdot j} = \big(f(x_1, j), \dots, f(x_n, j)\big)^\top$.

There are several downfalls 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 the next section, we propose an improvement to this naïve I-prior classification model by considering a probit-like transformation of the regression functions.

5.2 A latent variable motivation: the I-probit model

Let y_i, $\mathbf{y}_{i\cdot} = (y_{i1}, \ldots, y_{im})^\top$ and $x_i \in \mathcal{X}$ be as described in Section 5.1, and additionally, for $i = 1, \ldots, n$, let $y_i \sim \text{Cat}(p_{i1}, \ldots, p_{im})$. In this formulation, each y_{ij} is distributed as Bernoulli with probability p_{ij}. Now, assume that, for each $y_{i1}, \ldots, y_{im}$, there exists corresponding *continuous, underlying, latent variables* $y^*_{i1}, \ldots, y^*_{im}$ such that

$$y_i = \begin{cases} 1 & \text{if } y^*_{i1} \geq y^*_{i2}, y^*_{i3}, \ldots, y^*_{im} \\ 2 & \text{if } y^*_{i2} \geq y^*_{i1}, y^*_{i3}, \ldots, y^*_{im} \\ \vdots & \\ m & \text{if } y^*_{im} \geq y^*_{i2}, y^*_{i3}, \ldots, y^*_{i\,m-1}. \end{cases} \tag{5.6}$$

{eq:latentmodel}

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 $j = 1, \ldots, m$ according to the regression problem

$$\begin{aligned} y^*_{ij} &= \alpha + \alpha_j + f_j(x_i) + \epsilon_{ij} \\ (\epsilon_{i1}, \ldots, \epsilon_{im})^\top &\overset{\text{iid}}{\sim} \text{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}). \end{aligned} \tag{5.7}$$

{eq:multinomial-latent}

We can see some semblance of this model with the one in (5.3), 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 $\mathbf{y}_{i\cdot} = (y^*_{i1}, \ldots, y^*_{im})^\top$ has the distribution $\text{N}_m(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1})$, conditional on the data x_i, the intercepts $\alpha, \alpha_1, \ldots, \alpha_m$, the evaluations of the functions at x_i for each class $f_1(x_i), \ldots, f_m(x_i)$, and the error covariance matrix $\boldsymbol{\Psi}^{-1}$.

The probability p_{ij} of observation i belonging to class j is calculated as

$$\begin{aligned} p_{ij} &= \text{P}(y_i = j) \\ &= \text{P}\left(\{y^*_{ij} > y^*_{ik} \,|\, \forall k \neq j\}\right) \\ &= \int_{\{y^*_{ij} > y^*_{ik} \,|\, \forall k \neq j\}} \cdots \int \phi(y^*_{i1}, \ldots, y^*_{im} | \boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}) \, \text{d}y^*_{i1} \cdots \text{d}y^*_{im}, \end{aligned} \tag{5.8}$$

{eq:pij}

where $\phi(\cdot|\mu, \Sigma)$ is the density of the multivariate normal with mean μ and variance Σ. This is the probability that the normal random variable $\mathbf{y}_{i\cdot}^*$ belongs to the set $\{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}$, 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 for the classes. 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 5.7.1 for a note regarding this matter.

Now, we'll see how to specify an I-prior on the regression problem (5.7). In the naïve I-prior model, 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 (5.2). Instead of doing the same, we take a different approach. Treat the α_j's in (5.7) 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

$$f_j(x_i) = f_0(x_i) + \sum_{k=1}^n h_\eta(x_i, x_k) w_{ik}$$

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 (5.3), except that the intercept have been treated as parameters rather than accounting for them using an RKHS of constant functions. Importantly, the overall regression relationship still satisfies the ANOVA functional decomposition. We find that this approach bodes well down the line computationally.

We call the multinomial probit regression model of (5.6) subject to (5.7) and I-priors on $f_j \in \mathcal{F}$, the *I-probit model*. For completeness, this is stated again: for $i = 1, \ldots, n$, $y_i = \arg\max_{k=1}^m y_{ik}^* \in \{1, \ldots, m\}$, where, for $j = 1, \ldots, m$,

$$\begin{aligned} y_{ij}^* &= \alpha + \alpha_j + \overbrace{f_0(x_i) + \sum_{k=1}^n h_\eta(x_i, x_k) w_{ik}}^{f_j(x_i)} + \epsilon_{ij} \\ \boldsymbol{\epsilon}_{i\cdot} &:= (\epsilon_{i1}, \ldots, \epsilon_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}) \\ \mathbf{w}_{i\cdot} &:= (w_{i1}, \ldots, w_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}). \end{aligned} \tag{5.9}$$

{eq:iprobit mod}

The parameters of the I-probit model are denoted by $\theta = \{\alpha_1, \ldots, \alpha_m, \eta, \boldsymbol{\Psi}\}$. To establish notation, let

- $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries ϵ_{ij}, whose rows are $\boldsymbol{\epsilon}_{i\cdot}$ and columns are $\boldsymbol{\epsilon}_{\cdot j}$. Its distribution is $\boldsymbol{\epsilon} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1})$;
- $\mathbf{w} \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries w_{ij}, whose rows are $\mathbf{w}_{i\cdot}$ and columns are $\mathbf{w}_{\cdot j}$. Its distribution is $\mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$;
- $\mathbf{f} \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries $f_j(x_i)$, and $\mathbf{f}_0$ a vector equal to $\big(f_0(x_1), \ldots, f_0(x_n)\big)^\top$. We then have $\mathbf{f} = \mathbf{1}_n \mathbf{f}_0^\top + \mathbf{H}_\eta \mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{1}_n \mathbf{f}_0^\top, \mathbf{H}_\eta^2, \boldsymbol{\Psi})$;
- $\boldsymbol{\alpha} = (\alpha + \alpha_1, \ldots, \alpha + \alpha_m)^\top \in \mathbb{R}^m$ 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 $\operatorname{vec} \mathbf{y}^* \sim \mathrm{N}_{nm}\big(\operatorname{vec}(\mathbf{1}_n \boldsymbol{\alpha}^\top), \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n\big)$. The marginal distribution of $\mathbf{y}^*$ cannot be written as a matrix normal, except when $\boldsymbol{\Psi} = \mathbf{I}_m$.

Before proceeding with estimating the I-probit model (5.9), we lay out several standing assumptions:

ass:A4 A4 **Centred responses**. Set $\alpha = 0$.

ass:A5 A5 **Zero prior mean**. Assume a zero prior mean $f_0(x) = 0$ for all $x \in \mathcal{X}$.

ass:A6 A6 **Fixed error precision**. Assume $\boldsymbol{\Psi}$ is fixed.

Assumption A4 is a requirement for identifiability. Assumption A5 is motivated by a similar argument to assumption A2 in the normal I-prior model. As for assumption A6, we do not consider estimation of the error precision in this thesis mainly due to time limitations. More on this in Section 5.7.3.

5.3 Identifiability and IIA

sec:iia

The parameters in a linear multinomial probit model is well known to be unidentified (Michael P. Keane, 1992; Train, 2009), and the reason for this is two-fold. Firstly, an addition of a constant to the latent variables y_{ij}^*'s in (5.6) will not change which latent variable is maximal, and therefore leaves the model unchanged. Secondly, all

latent variables can be scaled by some positive constant without changing which latent variable is largest. Therefore, a *linear parameterisation* for the multinomial probit model is not identified as there can be more than one set of parameters for which the class probabilities are the same. To fix this issue, constraints are imposed on location and scale of the latent variables.

However, for the I-probit model, this is not the case, because the model is not related to the parameters $\theta = \{\alpha_1, \dots, \alpha_m, \eta, \boldsymbol{\Psi}\}$ linearly. One cannot simply add to or multiply θ by a constant and expect the model to be left unchanged. Thus, the I-probit model is identified in the parameter set θ without having to impose any restrictions, particularly on the precision matrix $\boldsymbol{\Psi}$ (if this is to be estimated).

To see how the I-probit model is location identified, suppose assumptions A4 and A5 hold, and consider a constant a added to the latent propensities. This would then imply the relationship

$$a + y_{ij}^* = \overbrace{a + \alpha_j}^{\alpha_j^*} + f_j(x_i) + \epsilon_{ij},$$

which is similar to adding the constant a to all of the intercept parameters α_j—denote these new intercepts by α_j^*. As a requirement of the functional ANOVA decomposition, the α_j^*'s need to sum to zero, but we already have that $\sum_{j=1}^m \alpha_j = 0$, so it must be that $a = 0$. This also highlights the reason behind assumption A4 and A5 for fixing the grand intercept α to zero.

As for identification in scale, consider multiplying the latent variables by $c > 0$. Denote by $\mathbf{V}_y^*(\omega) \in \mathbb{R}^{nm \times nm}$ the marginal covariance matrix of the latent propensities, which depends on the scale parameters $\omega = \{\eta, \boldsymbol{\Psi}\}$. The scaled latent variables $\{c^{1/2} y_{ij}^* \,|\, \forall i, j = 1, \dots\}$, which collectively has (marginal) variance and covariances given by the matrix $c\mathbf{V}_y^*(\omega)$, is expected to have been generated from the model with parameters $c\omega$. However, we have that

$$\begin{aligned} c\mathbf{V}_y^*(\omega) &= c(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + c(\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \\ &= (c\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + (c\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \\ &\neq \mathbf{V}_y^*(c\omega). \end{aligned}$$

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 $\mathbf{\Psi}$. Specifically, the off-diagonal elements $\mathbf{\Psi}_{jk}$ capture the correlation between alternatives j and k. Allowing all $m(m+1)/2$ covariance elements of $\mathbf{\Psi}$ to be non-zero leads to the *full I-probit model*, and would not assume an IIA position.

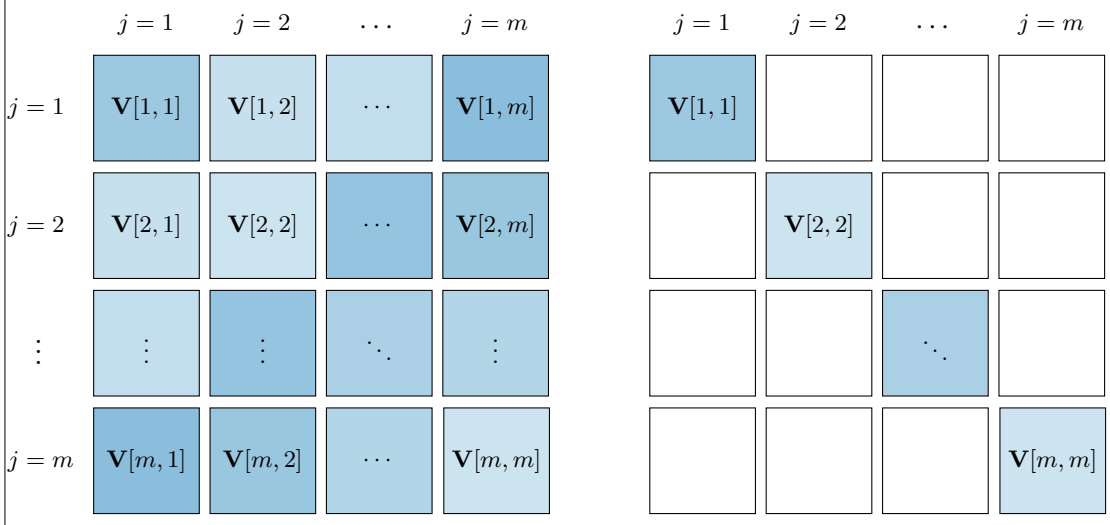


Figure 5.1: Illustration of the covariance structure of the full I-probit model (left) and the independent I-probit model (right). The full model has m^2 blocks of $n \times n$ symmetric matrices, and the blocks themselves are arranged symmetrically about the diagonal. The independent model, on the other hand, has a block diagonal structure, and its sparsity induces simpler computational methods for estimation.

fig:iprobcovstr

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 the IIA assumption in the I-probit model. We refer to this model as the *independent I-probit model*.

The independence assumption causes the distribution of the latent variables to be $y_{ij}^* \sim \text{N}(\mu_k(x_i), \sigma_j^2)$ for $j = 1, \ldots, m$, where $\sigma_j^2 = \psi_j^{-1}$. As a continuation of line (5.8), we can show the class probability p_{ij} to be

$$\begin{aligned} p_{ij} &= \int_{\{y_{ij}^* > y_{ik}^* | \forall k \neq j\}} \cdots \int \prod_{k=1}^{m} \left\{ \phi(y_{ik}^* | \mu_k(x_i), \sigma_k^2) \, \mathrm{d}y_{ik}^* \right\} \\ &= \int \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{y_{ij}^* - \mu_k(x_i)}{\sigma_k} \right) \cdot \phi(y_{ij}^* | \mu_j(x_i), \sigma_j^2) \, \mathrm{d}y_{ij}^* \\ &= \text{E}_Z \left[\prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{\sigma_j}{\sigma_k} Z + \frac{\mu_j(x_i) - \mu_k(x_i)}{\sigma_k} \right) \right] \end{aligned} \tag{5.10}$$ {eq:pij2}

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 (5.8) is thus simplified to a unidimensional integral involving the Gaussian pdf and cdf, which can be computed fairly efficiently using quadrature methods. The probit link function is evidently seen in the above equation.

5.4 Estimation

As with the normal I-prior model, an estimate of the posterior regression function with optimised hyperparameters is sought. The log likelihood function $L(\cdot)$ for θ using all n observations $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ is obtained by integrating out the I-prior from the

categorical likelihood, as follows:

$$
\begin{aligned}
L(\theta) &= \log \int p(\mathbf{y}|\mathbf{w},\theta)\, p(\mathbf{w}|\theta)\, \mathrm{d}\mathbf{w} \\
&= \log \int \prod_{i=1}^{n} \prod_{j=1}^{m} \Big(g_j^{-1}\big(\alpha_k + \overbrace{f_k(x_i)}^{\sum_{i'=1}^{n} h_\eta(x_i,x_{i'}) w_{i'k}}\big)_{k=1}^{m} \Big)^{[y_i=j]} \cdot \mathrm{MN}_{n,m}(\mathbf{w}|\mathbf{0},\mathbf{I}_n,\boldsymbol{\Psi})\, \mathrm{d}\mathbf{w} \qquad (5.11)
\end{aligned}
$$

{eq:intractablelikelihood}

where we have denoted the probit relationship from (5.8) using the function $g_j^{-1} : \mathbb{R}^m \to [0,1]$. Unlike in the continuous response models, the integral does not present itself in closed form due to the conditional categorical PMF of the y_i's, which they themselves involve integrals of multivariate normal densities. For binary response models, g^{-1} is simply the probit function, but for multinomial responses, this can be quite challenging to evaluate—more on this in Section 5.7.1.

Furthermore, the posterior distribution of the regression function, which requires the density of $\mathbf{w}|\mathbf{y}$, depends on the marginalisation provided by (5.11). The challenge of estimation is then to first overcome this intractability by means of a suitable approximation of the integral. We present three possible avenues to achieve this aim, namely the Laplace approximation, variational Bayes, and Markov chain Monte Carlo (MCMC) methods.

5.4.1 Laplace approximation

To compute the posterior density $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w}) =: e^{Q(\mathbf{w})}$ with normalising constant equal to the marginal density of $\mathbf{y}$, $p(\mathbf{y}) = \int e^{Q(\mathbf{w})}\, \mathrm{d}\mathbf{w}$, we have established that this is intractable. Laplace's method (Kass and Raftery, 1995, §4.1.1, pp. 777–778) entails expanding a Taylor series for Q about its posterior mode $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$, which gives the relationship

$$
\begin{aligned}
Q(\mathbf{w}) &= Q(\hat{\mathbf{w}}) + \cancelto{0}{(\mathbf{w}-\hat{\mathbf{w}})^\top \nabla Q(\hat{\mathbf{w}})} - \frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \boldsymbol{\Omega} (\mathbf{w}-\hat{\mathbf{w}}) + \cdots \\
&\approx Q(\hat{\mathbf{w}}) + -\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \boldsymbol{\Omega} (\mathbf{w}-\hat{\mathbf{w}}),
\end{aligned}
$$

because, assuming that Q has a unique maxima, ∇Q 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, $\boldsymbol{\Omega} = -\nabla^2 Q(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}}$ is the negative Hessian of Q evaluated at the

posterior mode, and is typically obtained as a byproduct of the maximisation routine of Q using gradient or quasi-gradient based methods.

The marginal distribution is then approximated by

$$\begin{aligned} p(\mathbf{y}) &\approx \int \exp \overbrace{Q(\mathbf{w})}^{Q(\hat{\mathbf{w}})-\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w}-\hat{\mathbf{w}})} \mathrm{d}\mathbf{w} \\ &= (2\pi)^{n/2}|\boldsymbol{\Omega}|^{-1/2} e^{Q(\hat{\mathbf{w}})} \int (2\pi)^{-n/2}|\boldsymbol{\Omega}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w}-\hat{\mathbf{w}})\right) \mathrm{d}\mathbf{w} \\ &= (2\pi)^{n/2}|\boldsymbol{\Omega}|^{-1/2} p(\mathbf{y}|\hat{\mathbf{w}})p(\hat{\mathbf{w}}). \end{aligned}$$

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 g^{-1} is efficient—because the dimension of this integral is exactly the sample size. Furthermore, obtaining standard errors for the parameters are cumbersome, and it is likely that a computationally burdensome bootstrapping approach is needed. Lastly, as a comment, Laplace's method only approximates the true marginal likelihood well if the true function is small far away from the mode.

5.4.2 Variational inference

We turn to variational inference as a method of estimation. Variational methods are widely discussed in the machine learning literature, but there have been efforts to popularise it in statistics (Blei et al., 2017). In a fully Bayesian setting, one obtains an approximation to the intractable posterior distribution of interest, which is then used for inferential purposes in lieu of the actual posterior distribution.

In addition to the I-probit model, suppose that prior distributions are assigned on the hyperparameters of the model, $\theta \sim p(\theta)$. By appending the latent variables $\{\mathbf{y}^*, \mathbf{w}\}$ to the hyperparameters θ, we seek an approximation

$$p(\mathbf{y}^*, \mathbf{w}, \theta|\mathbf{y}) \approx \tilde{q}(\mathbf{y}^*, \mathbf{w}, \theta),$$

where $\tilde{q}$ satisfies $\tilde{q} = \arg\min_q \mathrm{KL}(q\|p)$, subject to certain constraints. The constraint considered by us in this thesis is that q satisfies a *mean-field* factorisation

$$q(\mathbf{y}^*, \mathbf{w}, \theta) = q(\mathbf{y}^*)q(\mathbf{w})q(\theta).$$

Under this scheme, the posterior for $\mathbf{y}^*$ is found to be a *conically truncated multivariate normal* distribution, and for $\mathbf{w}$, a multivariate normal distribution. The posterior density $q(\theta)$ is often of a recognisable form, and usually one of the exponential family densities (normal, Wishart or gamma). This is useful, because point estimates of the hyperparameters can be taken to be either the mean or mode of these well-known distributions. In cases where $q(\theta)$ does not conform to an exponential family type density, then inference can still be done by sampling methods.

It can be shown that, for some variational density q, the marginal log-likelihood is an upper-bound for the quantity $\mathcal{L}$

$$\log p(\mathbf{y}) \geq \mathrm{E}_q \log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta) - \mathrm{E}_q \log \tilde{q}(\mathbf{y}^*, \mathbf{w}, \theta) =: \mathcal{L},$$

a quantity often referred to as the *evidence lower bound* (ELBO). It turns out that minimising $\mathrm{KL}(q\|p)$ is equivalent to maximising the ELBO, a quantity that is more practical to work with than the KL divergence. That is, if $\tilde{q}$ approximates the true posterior well, then the ELBO is a suitable proxy for the maximised marginal log-likelihood.

The algorithm to obtain $\tilde{q}$ which maximises the ELBO is known as the *coordinate ascent variational inference* (CAVI) algorithm. The algorithm itself typically condenses to that of a simple, sequential updating scheme, akin to the expectation-maximisation (EM) algorithm for exponential families we saw in Chapter 4, which is very fast to implement compared to the other methods described in the previous subsections. A full derivation of the variational algorithm used by us will be described in Section 5.5.

5.4.3 Markov chain Monte Carlo methods

As an alternative to the deterministic Bayesian approach of variational inference, it is possible to use Markov chain Monte Carlo sampling methods as an approach to stochastically approximate the intractable posterior distribution.

Albert and Chib (1993) showed that 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. On the other hand, this data augmentation scheme enlarges the variable space to $n + q$ dimensions, where q is the number of parameters to estimate, which is inefficient and computationally challenging especially when n is large. It is no longer possible to marginalise the normal latent variables from the model, as this is intractable, as discussed previously.

Hamiltonian Monte Carlo is another possibility, since it does not require conjugacy. For binary models, this is a feasible approach because the class probabilities are a function of the normal CDF, 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.

5.4.4 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 5.2. The predictor variables are X_1 and X_2, each of which are scaled similarly.

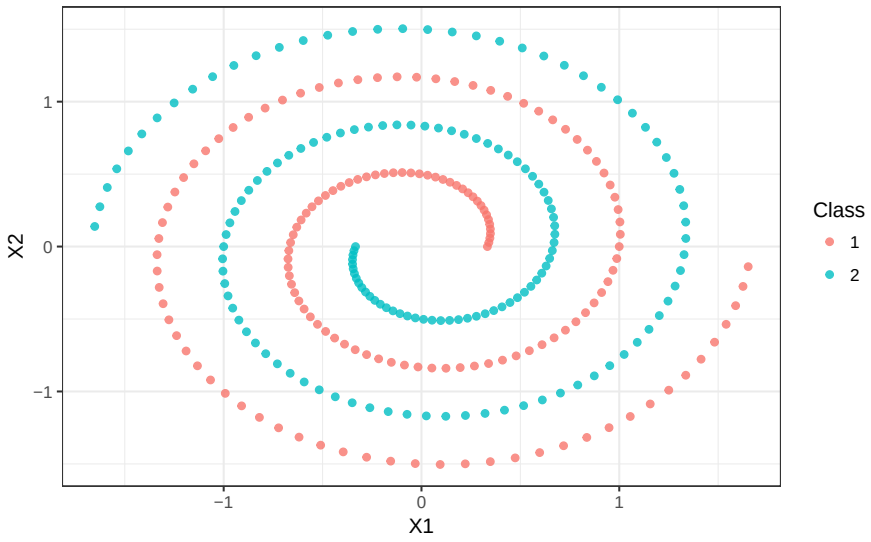


Figure 5.2: A plot of simulated spiral data set.

fig:example iprobit

The I-probit model that is fitted is

$$y_i \sim \text{Bern}(p_i)$$
$$\Phi^{-1}(p_i) = \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k$$
$$w_1, \ldots, w_n \overset{\text{iid}}{\sim} \text{N}(0, 1).$$

This binary model follows from the more general multinomial I-probit model by fixing all latent propensities in one of the classes to zero, and setting $\boldsymbol{\Psi} = \mathbf{I}_m$. This is possible because only differences in latent propensities are of interest, and not the actual values themselves, and thus only $m - 1$ sets of posterior regression functions need to be estimated—see Section 5.7.1 for further details.

The three estimation methods described aim to overcome the intractable integral by means of either a deterministic approximation (Laplace and variational inference) or a stochastic approximation (Hamiltonian MC). For the Bayesian methods, i.e. variational inference and Hamiltonian MC, vague priors were used on α and λ, namely $\text{N}(0, 100)$ and $\text{N}_+(0, 100)$ respectively. Restriction of λ to the positive orthant is required for identifiability. The Laplace and variational methods were performed in the **iprobit** package, while Stan was used to code the Hamiltonian MC sampler. The results are presented in Table 5.1.

Table 5.1: Table comparing the estimated parameter values, (marginal) log-likelihood values, and also time taken for the three estimation methods.

tab:comprei probit

	Laplace approximation	Variational inference	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	-202.7	-140.7	-163.8
Error rate (%)	44.7	0.00	2.24
Brier score	0.20	0.02	0.01
Iterations	20	56	2000
Time taken (s)	>3600	5.32	>3600

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 at the optima 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.

A plot of the log-likelihood surface for three methods in Figure 5.3 reveals some insight. The variational likelihood reveals 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 to probit-type likelihoods, as studied by Kuss and Rasmussen (2005). On the other hand, the log-likelihood using the posterior distribution of the Hamiltonian MC sampler (treating parameters as fixed values) yields a completely different shape compared to the other two methods.

In terms of predictive abilities, both the variational and Hamiltonian MC methods, even though the posteriors are differently estimated, has good predictive performance as indicated by their error rates and Brier scores. Figure 5.3 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 stronger than VI). 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 $\{w_1, \ldots, w_n, \alpha, \lambda\}$), and unlike in the normal model, we are not able to easily marginalise out the I-prior. As for Laplace, each Newton step involves obtaining posterior modes of the w'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 hours to. The predictive performance, while not as good as HMC, is certainly an acceptable compromise in favour of speed.

5.5 A variational algorithm

sec:iprobit
var

We present a variational inference algorithm to estimate the I-probit latent variables $\mathbf{y}^*$ and $\mathbf{w}$, together with the parameters $\theta = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top, \eta\}$ with *fixed error*

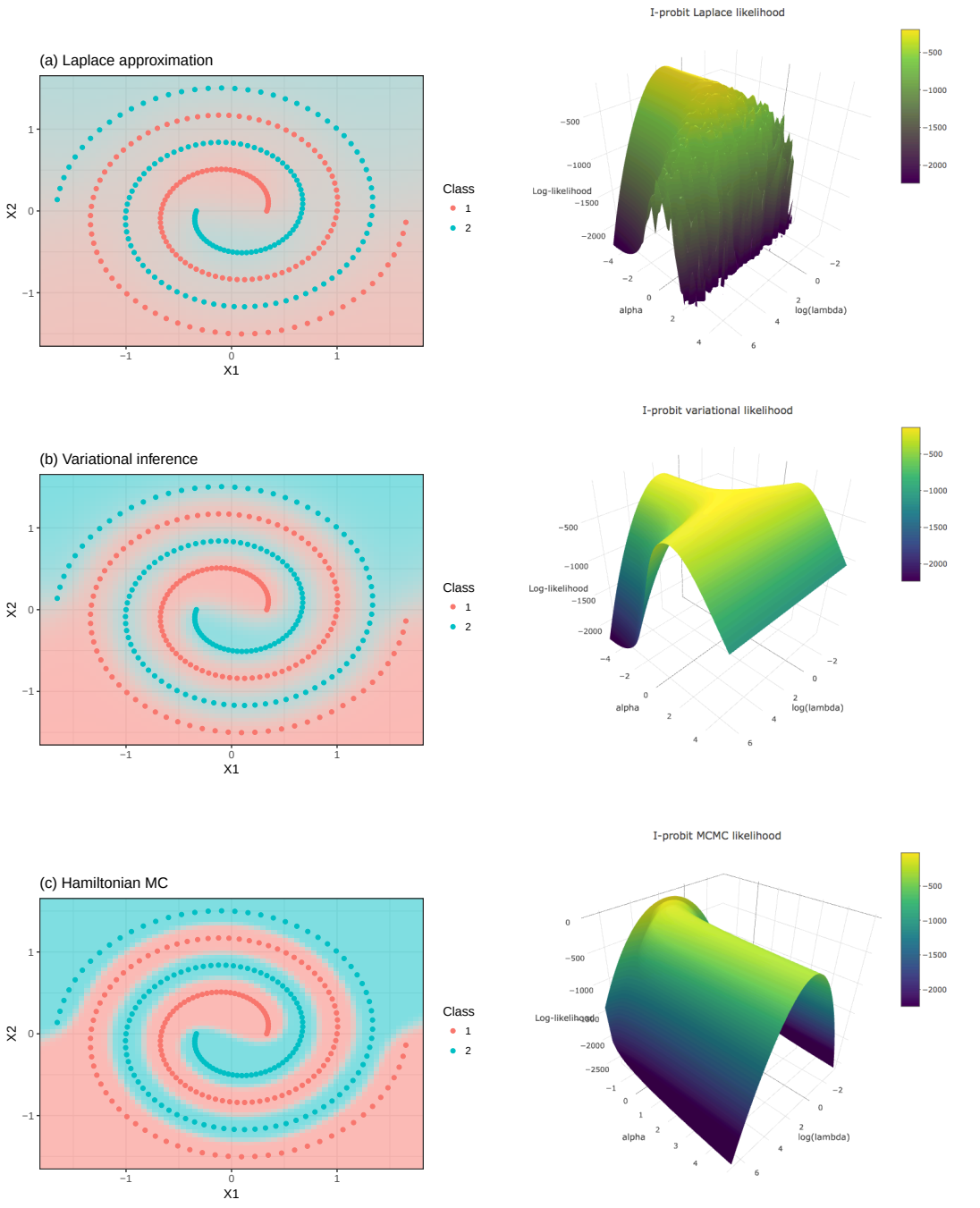


Figure 5.3: Plots showing predicted probabilities (shaded region) for belonging to class 1 or 2 indicated by colour and intensity, and likelihood surface plots for (a) Laplace's method, (b) variational inference, and (c) Hamiltonian MC.

fig:example iprobitfit

precision $\boldsymbol{\Psi}$[2]. Begin by choosing prior distributions on the parameters, $p(\theta) = p(\boldsymbol{\alpha})p(\eta)$. The following flat, uninformative priors are suggested:

- **Kernel parameters** η. This may include parameters such as the Hurst index, lengthscale and offset parameters, in addition to the RKHS scale parameters $\lambda_1, \ldots, \lambda_p$, and each with their own support. For the scale parameters, assign each λ_k the vague prior

$$\lambda_k \overset{\text{iid}}{\sim} \text{N}(0, v_\lambda = 0.001^{-1}),\ k = 1, \ldots, p.$$

  As $v_k^{-1} \to 0$, the prior becomes $p(\lambda_k) \propto \text{const.}$, an improper prior. The default choice for the rest of the kernel parameters is an improper prior $p(\eta) \propto \text{const.}$

- **Intercepts** $\alpha_1, \ldots, \alpha_m$. Assign independent, vague normal priors for each intercept

$$\alpha_j \overset{\text{iid}}{\sim} \text{N}(0, v_\alpha = 0.001^{-1}).$$

Although one may devote more attention to the prior specification of these parameters, for our purposes it suffices that they are independent component-wise, and that they are conjugate priors for the complete conditional density $p(\theta|\mathbf{y}, \mathbf{y}^*, \mathbf{w})$.

The posterior density of $\mathcal{Z} = \{\mathbf{y}^*, \mathbf{w}, \theta\}$ is approximated by a mean-field variational density q, i.e.

$$p(\mathbf{y}^*, \mathbf{w}, \theta|\mathbf{y}) = q(\mathbf{y}^*)q(\mathbf{w})q(\theta).$$

Additionally, we assume independence among the components of θ so that $q(\theta) = \prod_k q(\theta_k)$. We now present the mean-field variational distributions for each of unknowns in $\mathcal{Z}$. On notation: we will typically refer to posterior means of the parameters $\mathbf{y}^*$, $\mathbf{w}$, θ and so on by the use of a tilde. For instance, we write $\tilde{\mathbf{w}}$ to mean $\text{E}_{\mathbf{w}\sim q}[\mathbf{w}]$, the expected value of $\mathbf{w}$ under the pdf $q(\mathbf{w})$. The distributions are simply stated, but a full derivation is given in the appendix.

[2] It turns out that the variational algorithm as presented is not suited to estimate $\boldsymbol{\Psi}$. This issue is discussed further in Section X.

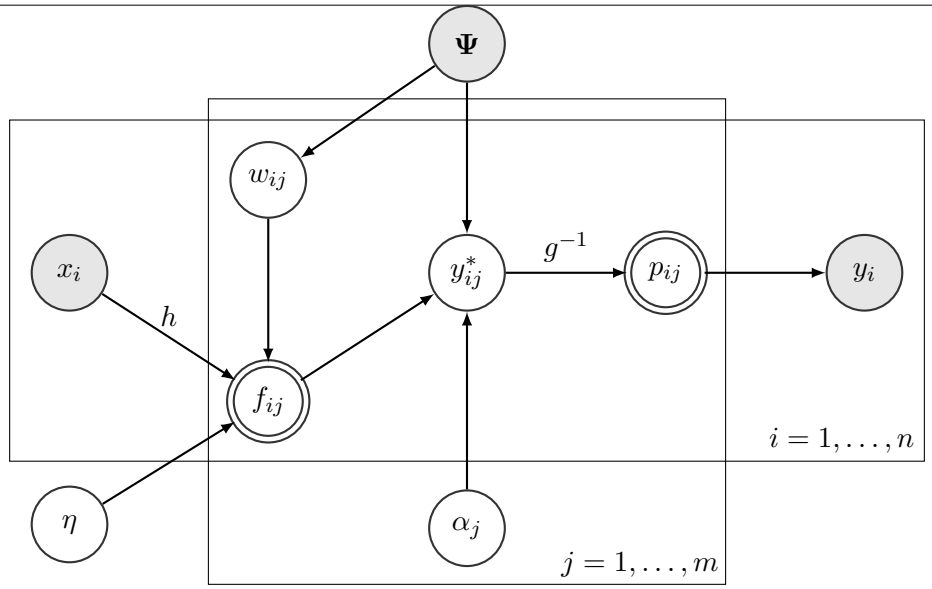


Figure 5.4: A DAG of the I-probit model. Observed/fixed nodes are shaded, while double-lined nodes represents calculable quantities.

5.5.1 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, which yields an 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 $\tilde{\boldsymbol{\mu}}_{i\cdot} = \tilde{\boldsymbol{\alpha}} + \tilde{\mathbf{w}}^\top \tilde{\mathbf{h}}_\eta(x_i)$, and variance $\boldsymbol{\Psi}^{-1}$ subject to the truncation of its components to the set $\mathcal{C}_{y_i}$. That is, for each $i = 1, \ldots, n$ and noting the observed value $y_i \in \{1, \ldots, m\}$, the $\mathbf{y}^*_i$'s are distributed according to

$$\mathbf{y}^*_{i\cdot} \overset{\text{iid}}{\sim} \begin{cases} \mathrm{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}) & \text{if } y^*_{iy_i} > y^*_{ik}, \forall k \neq y_i \\ 0 & \text{otherwise.} \end{cases} \tag{5.12}$$

{eq:ystardist}

We denote this by $\mathbf{y}^*_{i\cdot} \overset{\text{iid}}{\sim} {}^{\text{t}}\mathrm{N}(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, and the important properties of this distribution are explored in the appendix.

The required expectations $\mathrm{E}\,\mathbf{y}^*_{i\cdot} = \mathrm{E}(y^*_{i1}, \ldots, y^*_{im})^\top$ are tricky to compute. One strategy might be 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 fast, Gibbs based approach

to estimating the mean or any other quantity $\mathrm{E}[r(\mathbf{y}_{i\cdot}^*)]$ can be implemented, and this is detailed in the appendix.

If the independent I-probit model is considered, where the covariance matrix has the independent structure $\boldsymbol{\Psi} = \operatorname{diag}(\sigma_1^{-2}, \dots, \sigma_m^{-2})$, then the expected value can be considered component-wise, and each component of this expectation is given by

$$\tilde{y}_{ik}^* = \begin{cases} \tilde{\mu}_{ik} - \sigma_k C_i^{-1} \int \phi_{ik}(z) \prod_{l \neq k, j} \Phi_{il}(z) \phi(z) \, \mathrm{d}z & \text{if } k \neq y_i \\ \tilde{\mu}_{iy_i} - \sigma_{y_i} \sum_{k \neq y_i} \left(\tilde{y}_{ik}^* - \tilde{\mu}_{ik} \right) & \text{if } k = y_i \end{cases} \tag{5.13}$$

{eq:ystarupdate}

with

$$\begin{aligned} \phi_{ik}(Z) &= \phi\left(\frac{\sigma_{y_i}}{\sigma_k} Z + \frac{\tilde{\mu}_{iy_i} - \tilde{\mu}_{ik}}{\sigma_k} \right) \\ \Phi_{ik}(Z) &= \Phi\left(\frac{\sigma_{y_i}}{\sigma_k} Z + \frac{\tilde{\mu}_{iy_i} - \tilde{\mu}_{ik}}{\sigma_k} \right) \\ C_i &= \int \prod_{l \neq j} \Phi_{il}(z) \phi(z) \, \mathrm{d}z \end{aligned}$$

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.

5.5.2 I-prior random effects w

Given that both $\operatorname{vec} \mathbf{y}^* | \operatorname{vec} \mathbf{w}$ and $\operatorname{vec} \mathbf{w}$ are normally distributed, we find that the conditional posterior distribution $p(\mathbf{w} | \mathcal{Z}_{-\mathbf{w}}, \mathbf{y})$ is also normal, and therefore the approximate posterior density q for $\operatorname{vec} \mathbf{w} \in \mathbb{R}^{nm}$ is also normal with mean and precision given by

$$\operatorname{vec} \tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w (\boldsymbol{\Psi} \otimes \tilde{\mathbf{H}}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \tilde{\boldsymbol{\alpha}}^\top) \quad \text{and} \quad \tilde{\mathbf{V}}_w^{-1} = (\boldsymbol{\Psi} \otimes \tilde{\mathbf{H}}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n). \tag{5.14}$$

{eq:varipostw}

We note the similarity between (5.14) above and the posterior distribution for the I-prior random effects in a normal model (4.7) seen in the previous chapter. Naïvely computing the inverse $\tilde{\mathbf{V}}_w^{-1}$ presents a computational challenge, as this takes $O(n^3 m^3)$ time. By exploiting the Kronecker product structure in $\tilde{\mathbf{V}}_w$, we are able to efficiently compute the required inverse in roughly $O(n^3 m)$ time—see the appendix for details.

If the independent I-probit model is assumed, i.e. $\tilde{\boldsymbol{\Psi}} = \operatorname{diag}(\tilde{\psi}_1, \ldots, \tilde{\psi}_m)$, then the posterior covariance matrix $\tilde{\mathbf{V}}_w$ has a simpler structure: random matrix $\mathbf{w}$ will have columns which are independent of each other. 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

$$\tilde{\mathbf{w}}_{\cdot j} = \psi_j \tilde{\mathbf{V}}_{w_j} \tilde{\mathbf{H}}_\eta (\tilde{\mathbf{y}}_j^* - \tilde{\alpha}_j \mathbf{1}_n) \quad \text{and} \quad \tilde{\mathbf{V}}_{w_j} = \left(\psi_j \tilde{\mathbf{H}}_\eta^2 + \psi_j^{-1} \mathbf{I}_n \right)^{-1}.$$

The consequence of this is that the posterior regression functions are class independent, the exact intended effect by specifying a diagonal precision matrix $\boldsymbol{\Psi}$.

5.5.3 Kernel parameters η

sec:varupdeta

The posterior density q involving the kernel parameters is of the form

$$\log q(\eta) = -\frac{1}{2} \operatorname{tr} \mathrm{E}_{\mathcal{Z} \setminus \{\eta\} \sim q} \left[(\mathbf{y}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top - \mathbf{H}_\eta \mathbf{w}) \boldsymbol{\Psi} (\mathbf{y}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top - \mathbf{H}_\eta \mathbf{w})^\top \right] + \log p(\eta) + \text{const.}$$

where $p(\eta)$ is an appropriate prior density for η. Generally, samples $\eta^{(1)}, \ldots, \eta^{(T)}$ from $\tilde{q}(\eta)$ may be obtained using a Metropolis algorithm, so that quantities such as $\tilde{\mathbf{H}}_\eta = \mathrm{E}_{\eta \sim q} \mathbf{H}_\eta$ and the like may be approximated using $\frac{1}{T} \sum_{t=1}^{T} \mathbf{H}_{\eta^{(t)}}$. Details of the Metropolis sampler is available in the appendix.

When only RKHS scale parameters are involved, then the distribution q can be found in closed-form, much like in the exponential family EM algorithm described in Section 4.3.3. 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 $\mathbf{H}_\eta^2 = \lambda_k^2 \mathbf{R}_k^2 + \lambda_k \mathbf{U}_k + \mathbf{S}_k^2$. Additionally, we impose a further mean-field restriction on $q(\eta)$, i.e., $q(\eta) = \prod_{k=1}^{p} p(\lambda_k)$. Then, by using independent and identical normal priors on the λ_k's, such as the one listed at the beginning of this section, we find that $q(\lambda_k)$ is the density of a normal distribution with

mean $d_k c_k^{-1}$ and variance c_k^{-1}, where

$$c_k = \operatorname{tr}\left(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top \mathbf{R}_k^2 \mathbf{w}]\right) + v_\lambda^{-2}$$

and

$$d_k = \operatorname{tr}\left(\boldsymbol{\Psi}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top)^\top \mathbf{R}_k\tilde{\mathbf{w}} - \frac{1}{2}\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top \mathbf{U}_k \mathbf{w}]\right).$$

For a method of evaluating quantities such as $\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}])$ for suitably sized matrices $\mathbf{C}$ and $\mathbf{D}$, refer to the appendix.

5.5.4 Intercepts $\boldsymbol{\alpha}$

Finally, the posterior distribution for the intercepts follow a normal distribution with the normal priors specified earlier. The posterior mean and variance for the intercepts are given by $\tilde{\boldsymbol{\alpha}} = \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{a}}$ and $\tilde{\mathbf{A}}^{-1}$ respectively, where

$$\tilde{\mathbf{a}} = \sum_{i=1}^{n} \boldsymbol{\Psi}\left(\tilde{\mathbf{y}}_{i\cdot}^* - \tilde{\mathbf{w}}^\top \tilde{\mathbf{h}}_\eta(x_i)\right) \quad \text{and} \quad \tilde{\mathbf{A}} = n\boldsymbol{\Psi} + v_\alpha \mathbf{I}_m.$$

If $\boldsymbol{\Psi}$ is diagonal, the components of $\boldsymbol{\alpha}$ would be independent, and each would be distributed according to

$$\mathrm{N}\left(\frac{\psi_j \sum_{i=1}^n (\tilde{y}_{ij}^* - \tilde{f}_{ij})^2}{n\psi_j + v_\alpha^{-1}}, \frac{1}{n\psi_j + v_\alpha^{-1}}\right).$$

Here, we used the notation $\tilde{f}_{ij}$ to mean the (i,j)'th element of $\mathrm{E}[\mathbf{H}_\eta\mathbf{w}] = \tilde{\mathbf{H}}_\eta\tilde{\mathbf{w}} \in \mathbb{R}^{n\times m}$. Note that it is necessary, as discussed earlier, that $\sum_{j=1}^m \alpha_j = 0$ for identifiability.

5.5.5 The CAVI algorithm

One will have noticed that the evaluation of each component of the posterior depends on knowing the posterior distribution of the rest of the components. This circular dependence is dealt with by way of an iterative updating scheme of the components. Using an arbitrary starting value, each component is updated in turn according to the above derivations, until a maximum number of iterations is reached, or ideally, until a convergence criterion is met. In variational inference, the ELBO is used to asses convergence.

The expression for the ELBO for the I-probit model is derived in the appendix. The CAVI algorithm for the I-probit model is summarised in Algorithm 4.

Similar to the EM algorithm, each iteration of the algorithm increases the ELBO to a stationary point (Blei et al., 2017)—hence the name coordinate ascent variational inference (CAVI). Unlike the EM algorithm though, the CAVI 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.

5.6 Post-estimation

Working within a variational Bayesian framework means that we are able to perform inferences on any quantity of interest using the (approximate) posterior distributions obtained. Any of the post estimation procedures explained in the previous chapter when dealing with normal I-prior models can be extended here.

Prediction of a new data point x_{new} is described. Step one is to determine the distribution of the posterior regression functions in each class, $\mathbf{f}(x_{\text{new}}) = \mathbf{w}^\top \mathbf{h}_\eta(x_{\text{new}})$, given values for the parameters θ of the I-probit model. To this end, we use the posterior mean estimate for θ, and denote them with tildes, as we have done so far in this chapter. As we know, $\operatorname{vec} \mathbf{w}$ is normally distributed with mean and variance according to (5.14). By writing $\operatorname{vec} \tilde{\mathbf{w}} = (\tilde{\mathbf{w}}_{\cdot 1}, \dots, \tilde{\mathbf{w}}_{\cdot m})^\top$ to separate out the I-prior random effects per class, we have that $\mathbf{w}_{\cdot j} | \tilde{\theta} \sim \mathrm{N}_n(\tilde{\mathbf{w}}_{\cdot 1}, \tilde{\mathbf{V}}_w[j,j])$, 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 $j = 1, \dots, m$ and any $x \in \mathcal{X}$,

$$f_j(x) | \mathbf{y}, \tilde{\theta} \sim \mathrm{N}\left(\tilde{\mathbf{h}}_\eta(x)^\top \mathbf{w}_{\cdot j},\ \tilde{\mathbf{h}}_\eta(x)^\top \tilde{\mathbf{V}}_w[j,j] \tilde{\mathbf{h}}_\eta(x) \right),$$

and the covariance between the regression functions in two different classes is

$$\operatorname{Cov}\left[f_j(x), f_k(x) | \mathbf{y}, \tilde{\theta} \right] = \tilde{\mathbf{h}}_\eta(x)^\top \tilde{\mathbf{V}}_w[j,k] \tilde{\mathbf{h}}_\eta(x).$$

alg:caviipr
obit

Algorithm 4 CAVI for the I-probit model

procedure INITIALISATION
  Initialise $\tilde{\mathbf{y}}^{*(0)}, \tilde{\mathbf{w}}^{(0)}, \tilde{\boldsymbol{\alpha}}^{(0)}, \tilde{\mathbf{H}}_{\eta^{(0)}}, \boldsymbol{\Psi}$
  $t \leftarrow 0$
end procedure

while not converged **do**
  for $i = 1, \ldots, n$ **do** ▷ Update $\mathbf{y}^*$
    $q^{(t+1)}(\mathbf{y}_{i\cdot}^*) \leftarrow {}^{\text{t}}\text{N}_m\left(\tilde{\boldsymbol{\alpha}}^{(t)} + \tilde{\mathbf{w}}^{(t)\top}\tilde{\mathbf{h}}_{\eta^{(t)}}(x_i), \boldsymbol{\Psi}, \mathcal{C}_{y_i}\right)$
    $\tilde{\mathbf{y}}_{i\cdot}^{*(t+1)} \leftarrow \text{E}_{q^{(t+1)}}[\mathbf{y}_{i\cdot}^*]$
  end for

  $\mathbf{V}_w^{(t+1)} \leftarrow \left((\boldsymbol{\Psi} \otimes \tilde{\mathbf{H}}_{\eta^{(t)}}^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right)^{-1}$ ▷ Update $\mathbf{w}$
  $\tilde{\mathbf{w}}^{(t+1)} \leftarrow \tilde{\mathbf{V}}_w^{(t+1)}(\boldsymbol{\Psi} \otimes \tilde{\mathbf{H}}_{\eta^{(t)}})\,\text{vec}(\tilde{\mathbf{y}}^{*(t+1)} - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^{(t)\top})$
  $q^{(t+1)}(\mathbf{w}) \leftarrow \text{N}_{nm}\left(\tilde{\mathbf{w}}^{(t+1)}, \mathbf{V}_w^{(t+1)}\right)$

  Update $q^{(t+1)}(\eta)$ as per Section 5.5.3 ▷ Update η
  Sample $\eta^{[1]}, \ldots, \eta^{[T]} \sim q^{(t+1)}(\eta)$
  $\tilde{\mathbf{H}}_{\eta^{(t+1)}} \leftarrow \frac{1}{T}\sum_{i=1}^T \mathbf{H}_{\eta^{[i]}}$ and $\tilde{\mathbf{H}}_{\eta^{(t+1)}}^2 \leftarrow \frac{1}{T}\sum_{i=1}^T \mathbf{H}_{\eta^{[i]}}^2$

  $\tilde{\boldsymbol{\alpha}}^{(t+1)} \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\tilde{\mathbf{y}}_{i\cdot}^{*(t+1)} - \tilde{\mathbf{w}}^{(t+1)\top}\tilde{\mathbf{h}}_{\eta^{(t+1)}}(x_i)\right)$ ▷ Update $\boldsymbol{\alpha}$
  $q^{(t+1)}(\boldsymbol{\alpha}) \leftarrow \text{N}_m\left(\tilde{\boldsymbol{\alpha}}^{(t+1)}, \frac{1}{n}\boldsymbol{\Psi}^{-1}\right)$

  Calculate ELBO $\mathcal{L}^{(t+1)}$
  $t \leftarrow t + 1$
end while

Then, in step two, using the results obtained in the previous chapter in Section 4.4, we have that the latent propensities $y_{\text{new},j}^*$ for each class are normally distributed with mean, variance, and covariances

$$\begin{aligned}
\text{E}[y_{\text{new},j}^*|\mathbf{y}, \tilde{\theta}] &= \tilde{\alpha}_j + \text{E}\left[f_j(x_{\text{new}})|\mathbf{y}, \tilde{\theta}\right] &&=: \hat{\mu}_j(x_{\text{new}}) \\
\text{Var}[y_{\text{new},j}^*|\mathbf{y}, \tilde{\theta}] &= \text{Var}\left[f(x_{\text{new}})|\mathbf{y}, \tilde{\theta}\right] + \boldsymbol{\Psi}_{jj}^{-1} &&=: \hat{\sigma}_j^2(x_{\text{new}}) \\
\text{Cov}[y_{\text{new},j}^*, y_{\text{new},k}^*|\mathbf{y}, \tilde{\theta}] &= \text{Cov}\left[f_j(x), f_k(x)|\mathbf{y}, \tilde{\theta}\right] + \boldsymbol{\Psi}_{jk}^{-1} &&=: \hat{\sigma}_{jk}(x_{\text{new}}).
\end{aligned}$$

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

$$\hat{\boldsymbol{\mu}}_{\text{new}} = \big(\mu_1(x_{\text{new}}), \ldots, \mu_m(x_{\text{new}})\big)^\top \quad \text{and} \quad \hat{\mathbf{V}}_{\text{new},jk} = \begin{cases} \hat{\sigma}_j^2(x_{\text{new}}) & \text{if } i = j \\ \hat{\sigma}_{jk}(x_{\text{new}}) & \text{if } i \neq j. \end{cases}$$

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 (5.8), and restated here for convenience:

$$\hat{p}_{\text{new},j} = \int_{\{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}} \cdots \int \phi(y_{i1}^*, \ldots, y_{im}^* | \hat{\boldsymbol{\mu}}_{\text{new}}, \hat{\mathbf{V}}_{\text{new}}) \, \mathrm{d}y_{i1}^* \cdots \mathrm{d}y_{im}^*.$$

For the independent I-probit model, class probabilities are obtained in a more compact manner via

$$\hat{p}_{\text{new},j} = \mathrm{E}_Z \left[\prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{\hat{\sigma}_j(x_{\text{new}})}{\hat{\sigma}_k(x_{\text{new}})} Z + \frac{\hat{\mu}_j(x_{\text{new}}) - \hat{\mu}_k(x_{\text{new}})}{\hat{\sigma}_k^2(x_{\text{new}})} \right) \right],$$

as per (5.10), 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).

In this Bayesian setting, the analogue of standard errors for the parameters are their posterior standard deviations, which explain the uncertainty surrounding parameters. For the most part, these are easy to come by, and their posterior densities are easy to sample from. This allows us to conduct inference on transformed parameters, such as log odds ratios, quite easily. The procedure would be like this: first obtain samples of $\theta^{(1)}, \ldots, \theta^{(T)}$ from their respective distributions, then sample $\mathbf{w}^{(i)} \sim p(\mathbf{w}|\theta^{(i)})$ for $i = 1, \ldots, T$, and finally obtain samples of class probabilities $\hat{p}_{xj}^{(1)}, \ldots, \hat{p}_{xj}^{(T)}$, $j = 1, \ldots, m$, for a given data point $x \in \mathcal{X}$. 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 this transformed sample. In this manner, all aspects of uncertainty, from the parameters to the latent variables of the generative model, are accounted for.

It is possible to perform model comparison by comparing the maximised ELBO quantity of several candidate models (Beal and Ghahramani, 2003), and the justification for this is that it supposedly gives a tight lower bound to the marginal likelihood (model evidence), especially if the variational density is close in the KL divergence sense to the true posterior density. This would allow model selection using Bayes factor as a model selection criterion. Kass and Raftery (1995) suggest the following interpretation of observed Bayes factor values for comparing model M_1 against model M_0.

Table 5.2: Guidelines for interpreting Bayes factors.

tab:bf

$2\log \mathrm{BF}(M_1, M_0)$	$\mathrm{BF}(M_1, M_0)$	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

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).

5.7 Computational consideration

Computational challenges for the I-probit model stems from two sources: 1) calculation of the class probabilities (5.8); and 2) storage and time requirements for the CAVI. We also discuss issues faced with the estimation of the error precision $\boldsymbol{\Psi}$, and suggest ways to overcome this for future work.

5.7.1 Efficient computation of class probabilities

sec:mnint

As an opening remark, note that the dimension of the integral (5.8) is $m-1$, since the j'th coordinates is fixed relative to the others. An alternative specification of the I-probit model can be made in terms of *relative differences* of the latent propensities. Choosing the first category as the reference category, define new random variables $z_{ij} = y_{ij}^* - y_{i1}^*$, for $j = 2, \ldots, m$. The model (5.6) is equivalently represented by

$$y_i = \begin{cases} 1 & \text{if } \max(z_{i2}, \ldots, z_{im}) < 0 \\ j & \text{if } \max(z_{i2}, \ldots, z_{im}) = z_{ij} \geq 0. \end{cases} \tag{5.15}$$

Write $\mathbf{z}_{i\cdot} = (z_{i2}, \ldots, z_{im})^\top \in \mathbb{R}^{m-1}$. Then $\mathbf{z}_{i\cdot} = \mathbf{Q}\mathbf{y}^*_{i\cdot}$, where $\mathbf{Q} \in \mathbb{R}^{(m-1)\times m}$ is the $(m-1)$ identity matrix pre-augmented with a column vector of minus ones. We have that $\mathbf{z}_{i\cdot} \overset{\text{iid}}{\sim} \mathrm{N}_{m-1}(\boldsymbol{\nu}, \boldsymbol{\Omega})$, where $\boldsymbol{\nu}_{i\cdot} = \mathbf{Q}\boldsymbol{\mu}_{i\cdot}$ and $\boldsymbol{\Omega} = \mathbf{Q}\boldsymbol{\Psi}^{-1}\mathbf{Q}^\top$. Note that if $\boldsymbol{\Psi}$ is diagonal, then the transformation to $\boldsymbol{\Omega}$ will not retain diagonality—indeed, each component will undoubtedly be correlated with one another as they are all anchored on the same latent variable.

Now, the class probabilities for $j = 2, \ldots, m$ are

$$p_{ij} = \int_{\{z_{ik}<0 \,|\, \forall k \neq 1, j\}} \mathbb{1}[z_{ij} \geq 0]\, \phi(\mathbf{z}_{i\cdot} | \boldsymbol{\nu}_{i\cdot}, \boldsymbol{\Omega})\, \mathrm{d}\mathbf{z}_{i\cdot} \,. \tag{5.16}$$ {eq:pij3}

The class probability p_{i1} is simply $p_{i1} = 1 - \sum_{k\neq 1} p_{ik}$. From this representation of the model, with $m = 2$ (binary outcomes) we see that

$$p_{i1} = \Phi\left(\frac{z_{i2} - \nu}{\omega^{1/2}}\right) \quad \text{and} \quad p_{i2} = 1 - \Phi\left(\frac{z_{i2} - \nu}{\omega^{1/2}}\right),$$

where $\Phi(\cdot)$ is the CDF of the standard normal univariate distribution, and ν and ω are the mean and variance of the univariate random variable $\mathbf{z}_{i\cdot} = z_{i2}$. The probit link function involving the cdf of a standard normal is clearly seen here, especially if the error precision is treated as fixed such that $\omega = 1$.

The issue at hand here is that for $m > 4$, the evaluation of the class probabilities in (5.8) is computationally burdensome using classical methods such as quadrature methods Geweke et al. (1994).

The simplest strategy to overcome this is a frequency simulator (otherwise known as Monte Carlo integration): obtain random samples from $\mathrm{N}_{m-1}(\boldsymbol{\nu}_{i\cdot}, \boldsymbol{\Omega})$, 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, accuracy of this method is questionable when the mean $\boldsymbol{\nu}_{i\cdot}$ of the multivariate normal is many standard deviations away from zero (the cutoff region as per (5.16)).

A more reliable method is the probability simulator of Geweke-Hajivassiliou-Keane (GHK) (Geweke, 1991; Hajivassiliou et al., 1996; Michael P Keane, 1994), which we describe now. For clarity, we drop the subscript i denoting individuals, and write $\mathbf{z} = (z_1, \ldots, z_m)$, remembering that $z_1 = 0$. Suppose that an observation $y = j$ has been

made. Rewrite the model by anchoring on the j'th latent variable z_j as follows:

$$\tilde{\mathbf{z}} := (\overbrace{z_1 - z_j}^{\tilde{z}_1}, \ldots, \overbrace{z_{j-1} - z_j}^{\tilde{z}_{j-1}}, \overbrace{z_{j+1} - z_j}^{\tilde{z}_{j+1}}, \ldots, \overbrace{z_m - z_j}^{\tilde{z}_m},)^\top \in \mathbb{R}^{m-1}.$$

Let $\boldsymbol{\nu}_{(j)}$ and $\boldsymbol{\Omega}_{(j)}$ be the appropriately transformed mean vector and covariance matrix for $\tilde{\mathbf{z}}$. These are indexed by '(j)' because the transformation is dependent on which latent variable the $\mathbf{z}$'s are anchored on. Since this transformation is linear, $\tilde{\mathbf{z}} \sim \mathrm{N}_{m-1}(\boldsymbol{\nu}_{(j)}, \boldsymbol{\Omega}_{(j)})$. For the symmetric and positive definite matrix $\boldsymbol{\Psi}^{-1}$, obtain its Cholesky decomposition as $\boldsymbol{\Omega}_{(j)} = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a lower triangular matrix. Then, $\tilde{\mathbf{z}} = \boldsymbol{\nu}_{(j)} + \mathbf{L}\boldsymbol{\zeta}$, where $\boldsymbol{\zeta} \sim \mathrm{N}_{m-1}(\mathbf{0}, \mathbf{I}_{m-1})$. That is,

$$\begin{pmatrix} \tilde{z}_1 \\ \tilde{z}_2 \\ \vdots \\ \tilde{z}_m \end{pmatrix} = \begin{pmatrix} \nu_{(j)1} \\ \nu_{(j)2} \\ \vdots \\ \nu_{(j)m} \end{pmatrix} + \begin{pmatrix} L_{11} & & & \\ L_{21} & L_{22} & & \\ \vdots & \vdots & \ddots & \\ L_{m1} & L_{m2} & \cdots & L_{mm} \end{pmatrix} \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_m \end{pmatrix} = \begin{pmatrix} \nu_{(j)1} + L_{11}\zeta_1 \\ \nu_{(j)2} + \sum_{k=1}^{2} L_{k2}\zeta_k \\ \vdots \\ \nu_{(j)m} + \sum_{k=1}^{m} L_{km}\zeta_k \end{pmatrix}.$$

With this setup, we can calculate p_j, the probability of class j, which is equivalent to the probability that each $\tilde{z}_k = z_k - z_j < 0$, as follows

$$\begin{aligned} p_j &= \mathrm{P}(\tilde{z}_1 < 0, \ldots, \tilde{z}_{j-1} < 0, \tilde{z}_{j+1} < 0, \ldots, \tilde{z}_m < 0) \\ &= \mathrm{P}(\zeta_1 < u_1, \ldots, \zeta_{j-1} < u_{j-1}, \zeta_{j+1} < u_{j+1}, \ldots, \zeta_m < u_m) \\ &= \mathrm{P}(\zeta_1 < u_1)\, \mathrm{P}(\zeta_2 < u_2 | \zeta_1 < u_1) \cdots \\ &\qquad \cdots \mathrm{P}(\zeta_m < u_m | \zeta_1 < u_1, \ldots, \zeta_{j-1} < u_{j-1}, \zeta_{j+1} < u_{j+1}, \ldots, \zeta_{m-1} < u_{m-1}), \end{aligned}$$

where $u_i = u_i(\zeta_1, \ldots, \zeta_{i-1}) = -(\nu_{(j)i} + \sum_{k=1}^{i-1} L_{ki}\zeta_k)/L_{ii}$. Thus, the integral involving a $(m-1)$-variate normal density (5.16) is turned into a product of $m-1$ univariate normal cdfs, which can be computed fairly efficiently in modern computer systems.

As an aside, the GHK probability simulator, can be used to sample from a truncated multivariate normal distribution:

- 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 $\tilde{u}_2 = u_2(\tilde{\zeta}_1)$.
- $\cdots$
- Draw $\tilde{\zeta}_{j-1} \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_{j-1})$, where $\tilde{u}_{j-1} = u_{j-1}(\tilde{\zeta}_1, \ldots, \tilde{\zeta}_{j-2})$.

- Draw $\tilde{\zeta}_{j+1} \sim {}^{\text{t}}\text{N}(0, 1, -\infty, \tilde{u}_{j+1})$, where $\tilde{u}_{j+1} = u_{j+1}(\tilde{\zeta}_1, \ldots, \tilde{\zeta}_{j-1})$.
- $\cdots$
- Draw $\tilde{\zeta}_m \sim {}^{\text{t}}\text{N}(0, 1, -\infty, \tilde{u}_m)$, where $\tilde{u}_m = u_m(\tilde{\zeta}_1, \ldots, \tilde{\zeta}_{j-1}, \tilde{\zeta}_{j+1}, \ldots, \tilde{\zeta}_{m-1})$.

Then, $\tilde{z} = \boldsymbol{\nu}_{(j)}\mathbf{L}\tilde{\zeta}$ will be distributed according to $\text{N}_{m-1}(\boldsymbol{\nu}_{(j)}, \boldsymbol{\Omega}_{(j)})$. Any quantity of interest, e.g. $\text{E}\, r(\tilde{z})$, can then be estimated by the sample mean. In the variational algorithm, we require quantities such as first and second moments and also the entropy of a truncated multivariate normal distribution. Alternative methods are also discussed in the appendix.

Finally, a point on independent probit models. As we alluded to earlier in the chapter, the class probabilities condense to a unidimensional integral involving products of normal cdfs (see (5.10)) if $\boldsymbol{\Psi}$ is diagonal. While this represents a massive simplification, care should be taken when dealing with the formula in (5.10). When at least one of the normal cdfs in the product is extremely small, this can cause loss of significance due to floating-point errors. In the **iprobit** package, the product of normal cdfs is handled as a sum on the log scale to avoid this issue.

5.7.2 Computational complexity of the CAVI algorithm

sec:complxiprobit

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

$$\mathbf{V}_w = \left[(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right]^{-1}. \qquad \text{(from 5.14)}$$

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 CAVI down to a similar requirement as if $\boldsymbol{\Psi}$ was diagonal.

Storage requirements are still $O(n^2)$, and methods described in the previous chapter are applicable, particularly the discussion surrounding exponential family EM algorithm. Prediction of a new data point is $O(n^2 m)$, because there are essentially m 'separate' normal I-prior regressions, and each take $O(n^2)$ to evaluate.

5.7.3 Difficulties faced with estimating $\boldsymbol{\Psi}$

sec:difficultPsi

Suppose that, alongside the $\mathbf{y}^*$, $\mathbf{w}$, η and $\boldsymbol{\alpha}$ in the CAVI algorithm described in Section 5.5, $\boldsymbol{\Psi}$ is a free parameter to be estimated. If so, we find that the variational density q for $\boldsymbol{\Psi}$ satisfies

$$q(\boldsymbol{\Psi}) \propto \exp\left[-\frac{1}{2} \operatorname{tr}\left(\overbrace{\left(\mathrm{E}[(\mathbf{y}^* - \boldsymbol{\mu})^\top (\mathbf{y}^* - \boldsymbol{\mu})]\right)}^{\mathbf{G}_1} \boldsymbol{\Psi} + \overbrace{\mathrm{E}[\mathbf{w}^\top \mathbf{w}]}^{\mathbf{G}_2} \boldsymbol{\Psi}^{-1} \right) \right] \times p(\boldsymbol{\Psi})$$

where $p(\boldsymbol{\Psi})$ is a prior density chosen for $\boldsymbol{\Psi}$. Unfortunately, this does not resemble any known distribution, regardless of the prior choice for $\boldsymbol{\Psi}$. One can resort to sampling techniques to obtain quantities such as the mean or entropy, which are needed, but this has not been studied for this project due to time limitations. Even if this was possible, this requires, among other things, second moments of a truncated multivariate normal density $\mathbf{y}^*$, and also of $\mathbf{H}_\eta \mathbf{w} \in \mathbb{R}^{n \times m}$—of which both are a bit awkward to obtain.

What we realise, however, is that the *posterior mode* is relatively easy to obtain, especially with an improper prior $p(\boldsymbol{\Psi}) \propto \text{const.}$ To see this, we look specifically at the case where $\boldsymbol{\Psi}$ is diagonal. On the log scale,

$$\log q(\psi_j) = \text{const.} - \frac{1}{2}\sum_{j=1}^{m} \psi_j \, \mathrm{E}\|\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j}\|^2 - \frac{1}{2}\sum_{j=1}^{m} \psi_j^{-1} \, \mathrm{E}\|\mathbf{w}_{\cdot j}\|^2$$

is maximised, for $j = 1, \ldots, m$, at

$$\hat{\psi}_j = \sqrt{\frac{\mathrm{E}\|\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j}\|^2}{\mathrm{E}\|\mathbf{w}_{\cdot j}\|^2}}.$$

Perhaps, if the posterior mean is close to the mode, and not withstanding the involved calculations of the require second moments, then this quantity can be used instead in the CAVI algorithm. This ties with the idea of *variational Bayes EM algorithm*, which is an alternative to a fully Bayesian treatment of variational inference. This is discussed in Sections 5.10.2 and 5.10.3, but unfortunately, time constraints had made it impossible for us to examine this within the scope of this thesis.

5.8 Examples

sec:iprobit
eg

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 the R package **iprobit** developed by us. All of these examples had assumed a fixed error precision $\boldsymbol{\Psi} = \mathbf{I}_m$.

5.8.1 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 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).

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. The experiment is set up as follows:

[3] Data is made publicly available at https://archive.ics.uci.edu/ml/datasets/arrhythmia.

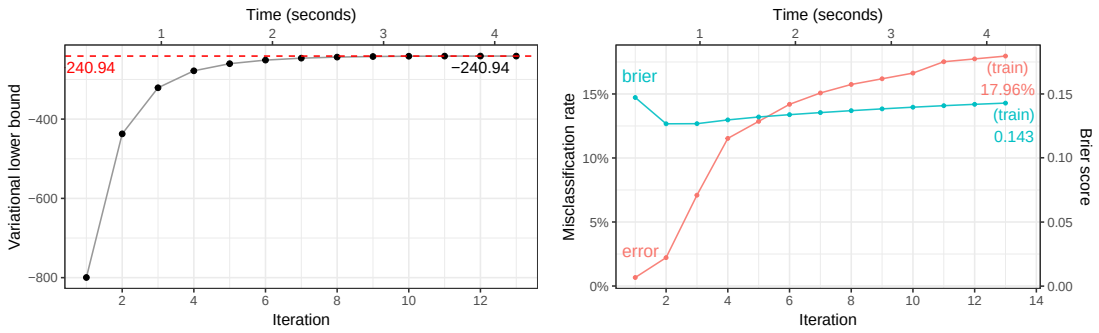


Figure 5.5: Plot of variational lower bound over time (left), and plot of training error rate and Brier scores over time (right).

fig:cardiac.mod.full.plot

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.

Results for the methods listed above were extracted from the in-depth study by Cannings and Samworth (2017), who also conducted an identical experiment using their random projection ensemble classification method (RP). The results are 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 ensemble learning method, which depending on the number of random decisions trees generated simultaneously, might be slow. The time complexity of a random forest algorithm is

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.

tab:cardiac

	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)

$O(pqn \log(n))$, 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.

5.8.2 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 to quit 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.

A summary of the data is displayed by the box-plot in Figure 5.6. On the whole, there are a total of 5908 patients, and they are distributed roughly equally among the control and treatment groups (46.33% and 53.67% 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 the 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

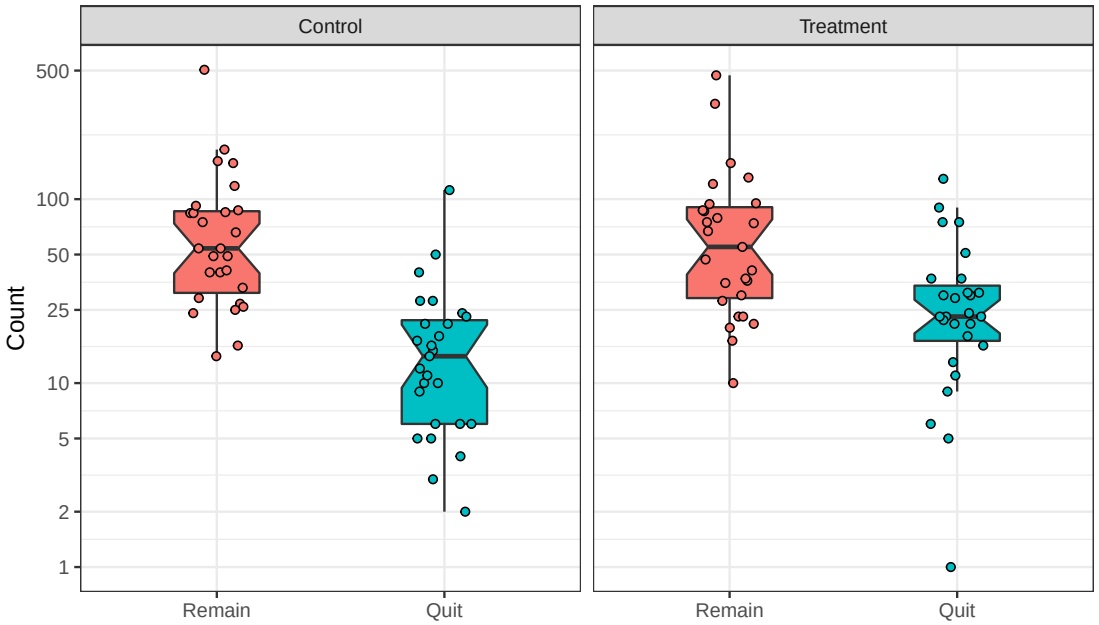


Figure 5.6: Comparative box-plots of the distribution of patients who successfully quit smoking and those who remained smokers, in the two treatment groups.

fig:plot.data.smoke

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 $i = 1, \ldots, n_j$ index the patients in study group $j \in \{1, \ldots, 27\}$. For patient i in study j, p_{ij} denotes the probability that the patient has successfully quit smoking. Additionally, x_{ij} is the centred dummy variable indicating patient i's treatment group in study j. 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_{ij}, the total variance $\mathrm{Var}(\beta_{0j} + \beta_{1j}x_{ij})$ would be

constant in both treatment groups. The overall log odds ratio is represented by β_1, and this is estimated as $0.57 = \log 1.76$.

In an I-prior model, the Bernoulli probabilities p_{ij} are regressed against the treatment group indicators x_{ij} and also the patients' study group j via the regression function f and a probit link:

$$\begin{aligned}\Phi^{-1}(p_{ij}) &= f(x_{ij}, j) \\ &= f_1(x_{ij}) + f_2(j) + f_{12}(x_{ij}, j).\end{aligned}$$

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_{ij} and j 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_{ij} 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 I-prior modelling, 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-prior model fit for three models.

tab:mod.compare.smoke

Model	Log-likelihood	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 j - call these $p_j(\text{treatment})$ and $p_j(\text{control})$. That is,

$$p_j(\text{treatment}) = \Phi\big(\tilde{\mu}(\text{treatment}, j)\big)$$
$$p_j(\text{control}) = \Phi\big(\tilde{\mu}(\text{control}, j)\big),$$

where $\tilde{\mu}$ represents the posterior mean estimate for the regression function given in Which section?. These 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 $0.55 = \log 1.73$, slightly lower than both the raw log odds ratio and the log odds ratio estimated by the 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 in Figure 5.7 for the log odds ratios under an I-prior are also noticeably narrower compared to the multilevel model fitted.

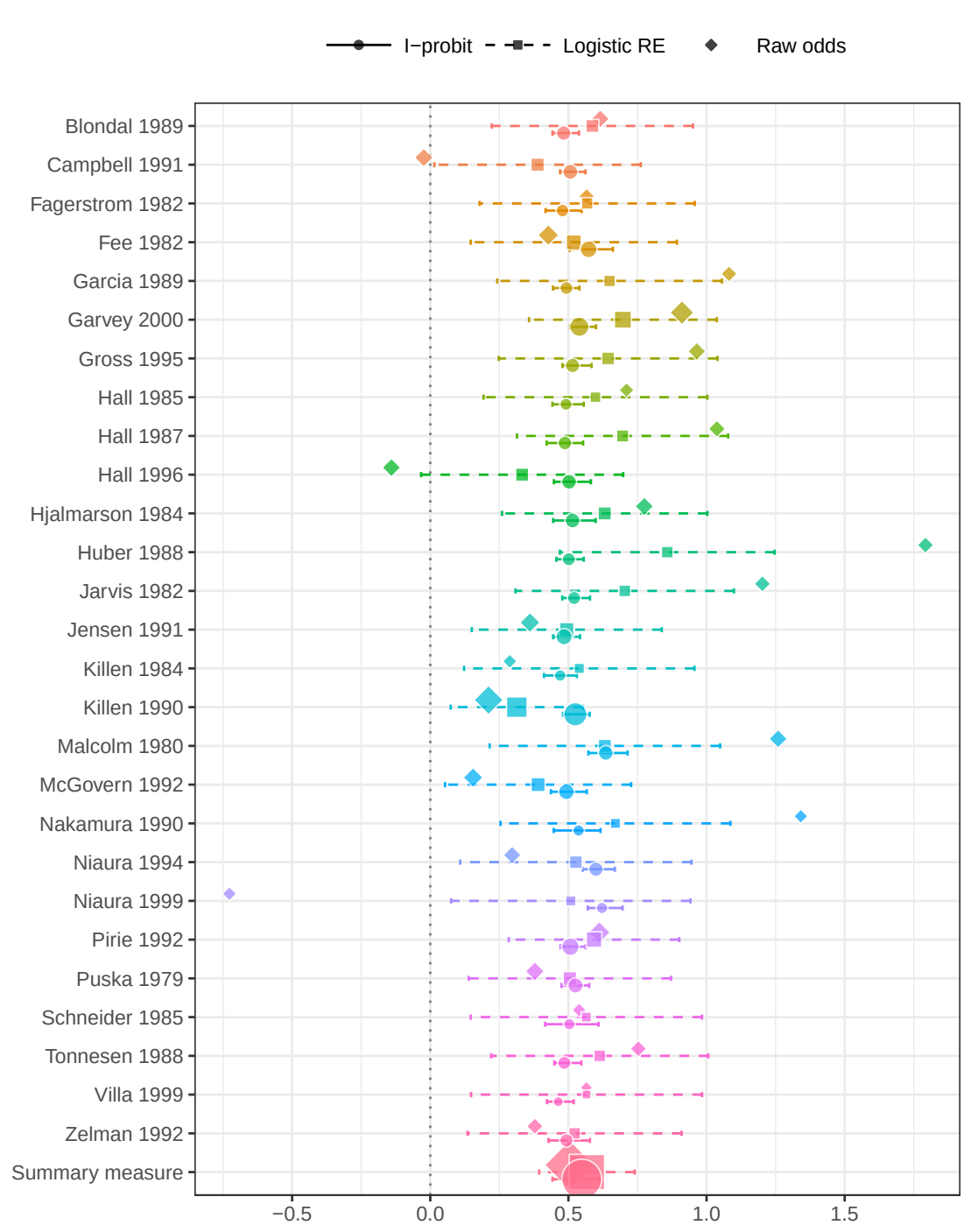


Figure 5.7: Forest plot of effect sizes (log odds ratios) in each group as well as the overall effect size together with their 95% confidence bands. The plot compares the raw log odds ratios, the logistic random-effect estimates, and the I-prior estimates. Sizes of the points indicate the relative sample sizes per study group.

fig:smoke.forest.plot

5.8.3 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 528, while 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, 1989). 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.

tab:vowel

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 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.

[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).

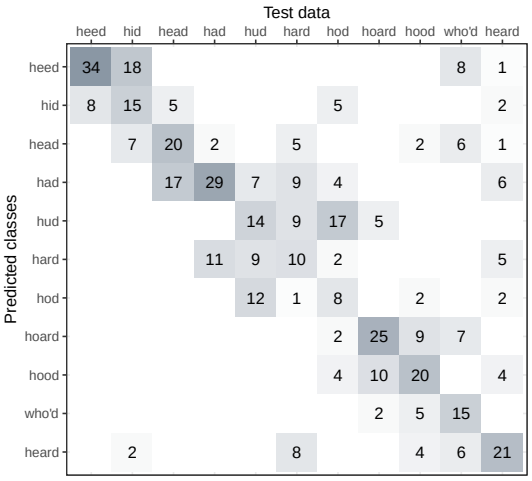
tab:vowel.tab

Table 5.6: Results of various classification methods for the vowel data set.

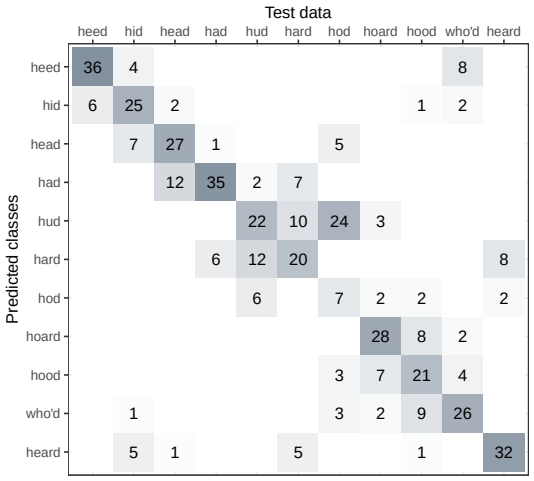
	Error rate (%)	
Model	Train	Test
I-probit		
Linear	29	54
Smooth (fBm-0.5)	22	40
Smooth (SE-1.0)	7	34
Others		
Linear regression	48	67
Logistic regression	22	51
Linear discriminant analysis	32	56
Quadratic discriminant analysis	1	53
Decision trees	5	54
Neural networks		45
k-nearest neighbours		44
FDA/BRUTO	6	44
FDA/MARS	13	39

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.

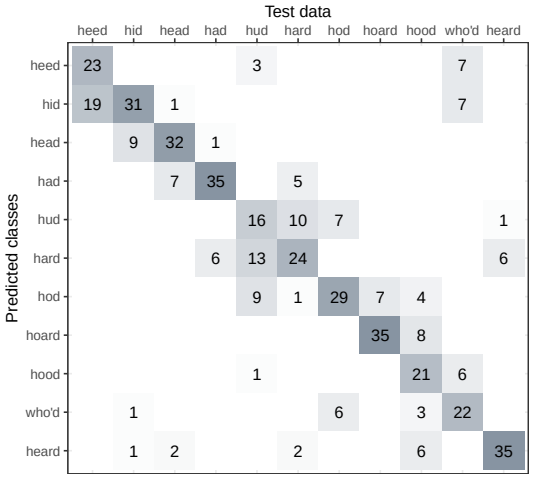
Comparisons to other methods that had been used to analyse this data set is given in Table 5.6. In particular, the I-probit model is compared against 1) linear regression; 2) logistic linear regression; 3) linear and quadratic discriminant analysis; 4) decision trees; 5) neural networks; 6) k-nearest 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.



(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 one cell is 42. Blank cells indicate nil values.

5.8.4 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 between the herds, and whether there is a time-element to presence or absence of this spatial segregation. There have been previous work done to analyse this data set: P. 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 work includes P. J. Diggle et al. (2013), who used a fully Bayes scheme for spatio-temporal multivariate log-Gaussian Cox processes, and implemented in the R package **lgcp** (Taylor et al., 2013).

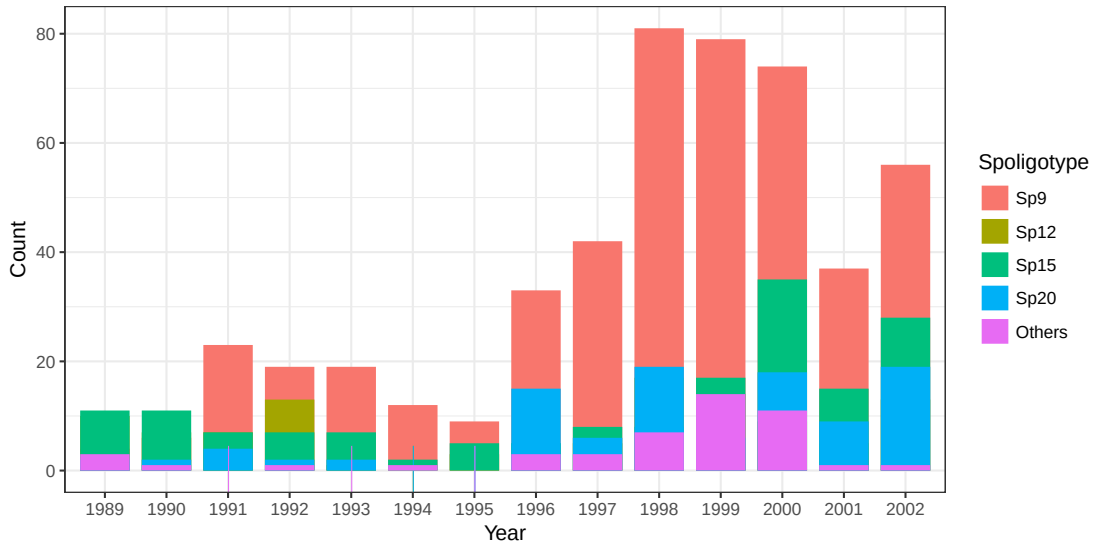


Figure 5.9: Distribution of the different types (Spoligotypes) of bovine tubercolosis affecting herds in Cornwall over the period 1989 to 2002.

fig:plot.cow

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.

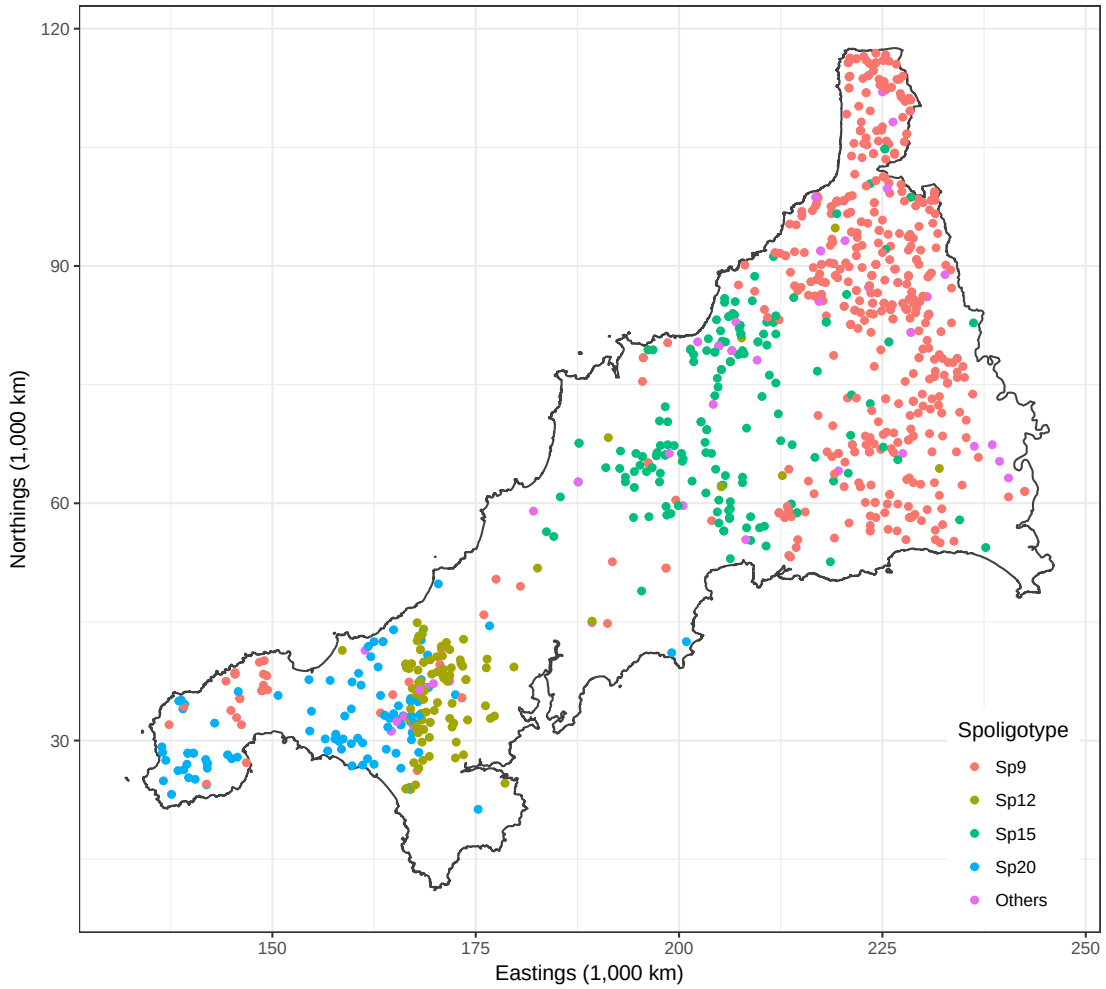


Figure 5.10: Spatial distribution of all cases over the 14 years.

fig:plot.cornwall

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 $g(p_{ij})$ (as described in the categorical response chapter) as following a smooth function f which takes two covariates: the spatial data $x_1 \in \mathbb{R}^2$, and the temporal data x_2 (year

of infection):

$$\begin{aligned} g(p_{ij}) &= f_j(x_1, x_2) \\ &= f_{1j}(x_1) + f_{2j}(x_2) + f_{12j}(x_1, x_2) \end{aligned}$$

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 P. 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**.
$$p_{ij} = g_j^{-1}(\alpha_k)_{k=1}^m$$
- $\boldsymbol{M_1}$**: 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 $f_{12k} \in \mathcal{F}_1 \otimes \mathcal{F}_2$

where g^{-1} is the link function described earlier. Model M_0 corresponds to a model which ignores any spatial or temporal effects (the baseline intercept only model). Model M_1

tab:table.btb

Table 5.7: Results of the fitted I-probit models.

	M_0: Intercepts only		M_1: Spatial only		M_2: Spatio-temporal		M_3: Spatio-period	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Intercept (Sp9)	0.948	0.033	1.364	0.033	1.401	0.033	1.395	0.033
Intercept (Sp12)	-0.173	0.033	-0.435	0.033	-0.506	0.033	-0.463	0.033
Intercept (Sp15)	0.103	0.033	-0.020	0.033	-0.008	0.033	-0.010	0.033
Intercept (Sp20)	-0.202	0.033	-0.775	0.033	-0.795	0.033	-0.783	0.033
Intercept (Others)	-0.676	0.033	-0.134	0.033	-0.091	0.033	-0.139	0.033
Scale (spatial)			0.194	0.003	-0.176	0.003	0.172	0.003
Scale (temporal)					-0.006	0.000	-0.004	0.000
Log-likelihood	-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	

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.

Alternatively, spatio-temporal effects of the BTB breakdowns can easily be inferred through the RKHS scale parameters. Let h_k, $k \in \{1, 2\}$ denote the reproducing kernel of the spatial and temporal RKHSs respectively. Then, an I-prior on $f_j = f_{1j} + f_{2j} + f_{12j}$, $j = 1, \ldots, 5$, takes the form

$$f_j(x_1, x_2) = \sum_{i=1}^{n} \big(\lambda_1 h_1(x_1, x_{i1}) + \lambda_2 h_2(x_2, x_{i2}) + \lambda_1\lambda_2 h_1(x_1, x_{i1}) h_2(x_2, x_{i2})\big) w_{ij}$$

where it is assumed $(w_{i1}, \ldots, w_{i5})^\top \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{I}_5)$. The hypothesis of temporal significance is the same as testing the significance of the λ_2 parameter, while the test of both spatial and temporal effects are conducted on λ_1 and λ_2 simultaneously. From Chapter X, we know that these scale parameters follow a normal posterior distribution, so we can calculate the Z-scores by dividing the mean by its corresponding standard deviation. Absolute values greater than three would satisfy a Bayesian hypothesis test of significance at the 0.01 level. The conclusion from Table 5.7 is that the data supports a hypothesis for a spatio-temporal or spatio-period 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 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 model 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. For

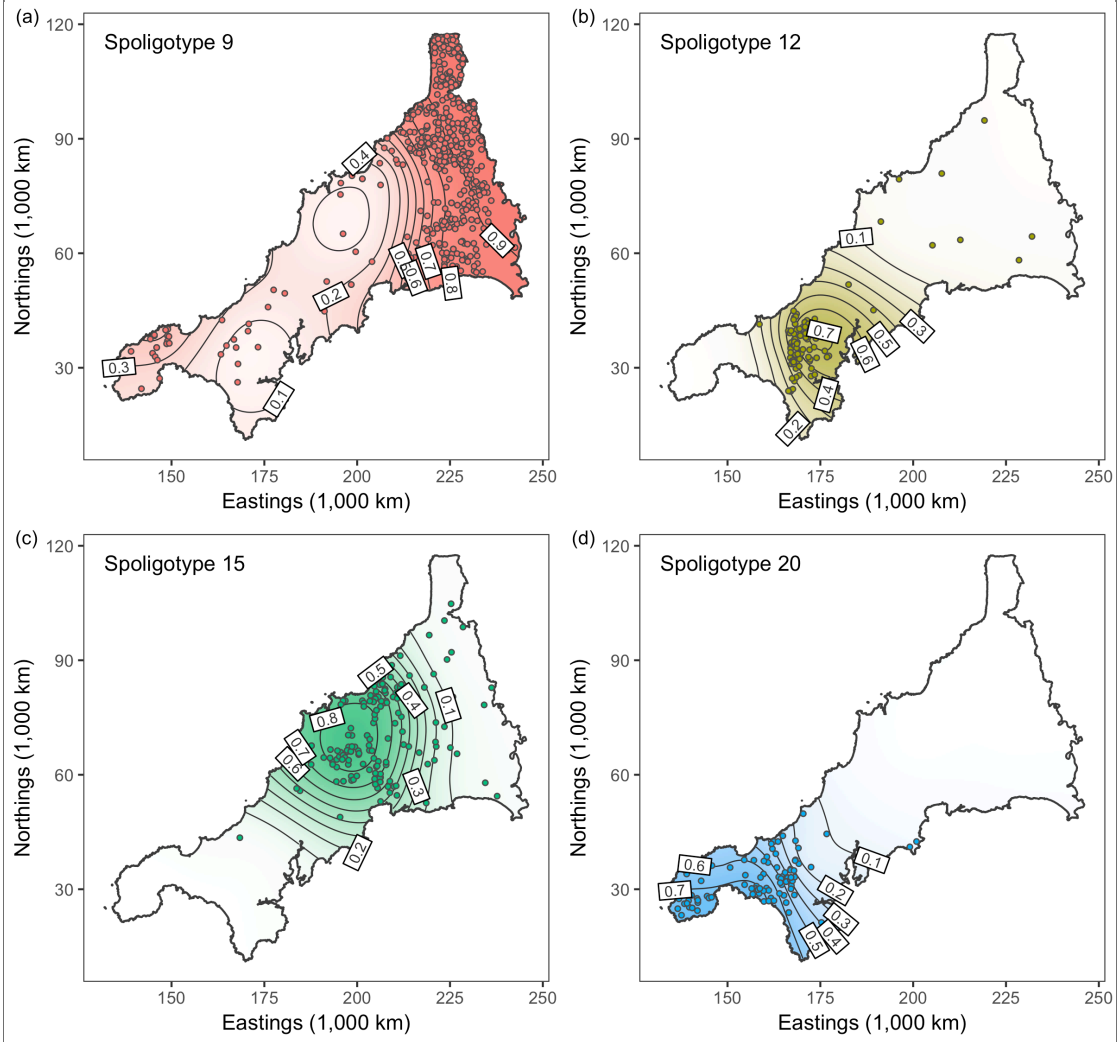


Figure 5.11: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium Mycobacterium bovis over the entire time period using model M_1.

fig:plot.btb

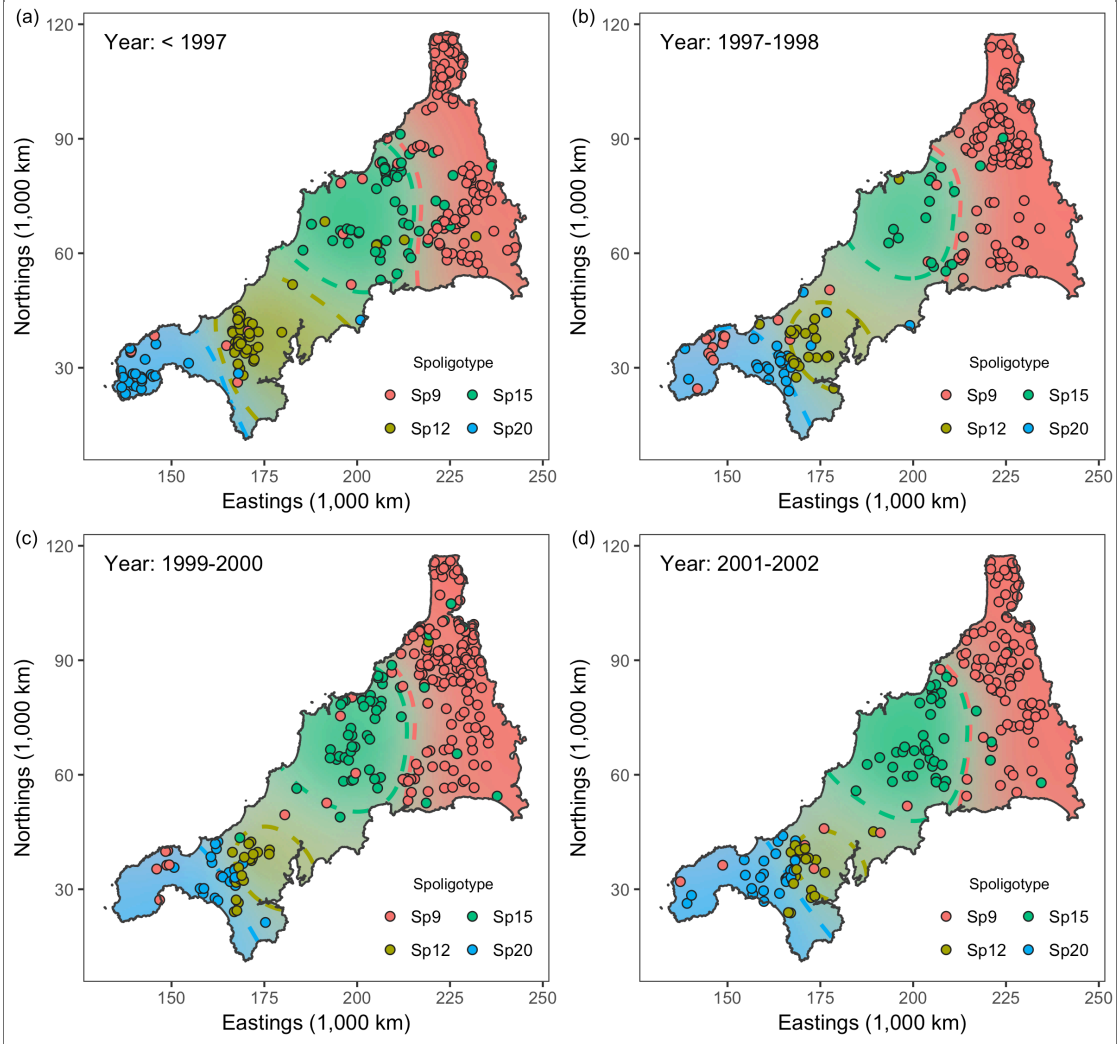


Figure 5.12: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium Mycobacterium bovis over four different time periods using model M_3.

fig:plot.temporal.btb

each time period, we superimpose the actual observations onto the predicted surface probabilities. In addition, coloured dotted lines are displayed to indicate the "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. This is supported also by the spatio-period model results in Table 5.7, where the test of nullity for the scale parameters of these two spoligotypes are not rejected.

5.9 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 (5.11). 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 (Radford M. 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. This limitation was in part due to the non-conjugate nature of the variational density for $\boldsymbol{\Psi}$. We believe the variational Bayes EM algorithm, which estimates maximum a posteriori values for the parameters, could alleviate this issue. This would bring the estimation procedure on par with the frequentist objective of maximum likelihood via the EM algorithm, albeit with the use of approximate posterior densities (see Sections 5.10.2 and 5.10.3 for further discussions).

2. **Inclusion of class-specific covariates**. Throughout the chapter, we assumed that covariates were unit-specific, rather than class-specific. One such example is modelling the choice of travel mode between two destinations (car, coach, train or aeroplane) as a function of travel time. Clearly, travel time depends on the mode of transport. This would require a careful rethink of the appropriate RKHS/RKKS to which the regression function belongs: the regression on the latent propensities could be extended as such:

$$y_{ij}^* = \alpha_j + f_j(x_i) + e(z_{ij})$$

and $f_j \in \mathcal{F}_\mathcal{X}$, the RKHS with kernel $h : (\mathcal{X} \times \mathcal{M})^2 \to \mathbb{R}$ defined by $\delta_{jj'}h(x, x')$, and $e \in \mathcal{F}_\mathcal{Z}$, the RKHS of functions of the form $e : \{z_{ij} | i = 1, \ldots, n,\ j = 1, \ldots, m\} \to \mathbb{R}$. An I-prior would then be applied as usual, but the implications on the estimation would need to be considered as well.

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 compu-

tational efficiency could lie in sparse variational methods, a suggestion that was made to improve normal I-prior models as well.

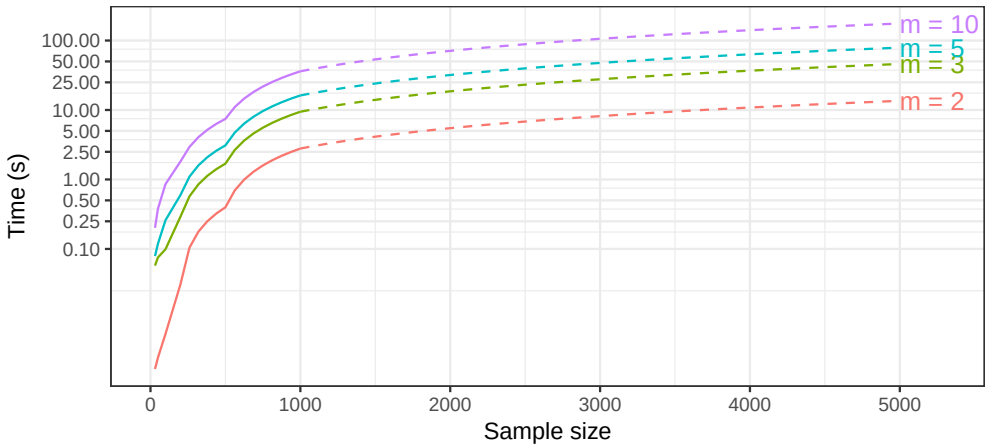


Figure 5.13: Time taken to complete a single variational inference iteration for varying sample sizes and number of classes m. The solid line represents actual timings, while the dotted lines are linear extrapolations.

5.10 Miscellanea

5.10.1 A brief introduction to variational inference

sec:varintro

Consider a statistical model for which we have observations $\mathbf{y} := \{y_1, \dots, y_n\}$, but also some latent variables $\mathbf{z} := \{z_1, \dots, z_n\}$. Typically, in such models, there is a want to to evaluate the integral

$$\mathcal{I} = \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z}. \tag{5.17}$$

eq:varint

Models that include latent variables are plenty, for example: Gaussian mixture models, latent class analysis, factor models, random coefficient models, and so on. Marginalising out the latent variables in (5.17) is usually a precursor to obtaining a log-likelihood function to be maximised, in a frequentist setting. In Bayesian analysis, the $\mathbf{z}$'s are parameters which are treated as random, and the integral corresponds to the marginal density for $\mathbf{y}$, on which the posterior depends.

In many instances, for one reason or another, evaluation of $\mathcal{I}$ is difficult, in which case inference is halted unless a way of overcoming the intractable integral (5.17) is

found. Here, we discus *variational inference* (VI), a fully Bayesian treatment of the statistical model with a deterministic algorithm, i.e. does not involve sampling from posteriors. The crux of variational inference is this: find a suitably close distribution function $q(z)$ that approximates the true posterior $p(\mathbf{z}|\mathbf{y})$, where closeness here is defined in the Kullback-Leibler divergence sense,

$$\mathrm{KL}(q\|p) = \int \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{y})} q(\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Posterior inference is then conducted using $q(\mathbf{z})$ in lieu of $p(\mathbf{z}|\mathbf{y})$. 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 $\mathbf{z}$. One may show that log marginal density (the log of the intractable integral (5.17)) holds the following bound:

$$\begin{aligned}
\log p(y) &= \log p(\mathbf{y},\mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \quad \text{(Bayes' theorem)} \\
&= \int \left\{ \log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z})\,\mathrm{d}z \quad \text{(expectations both sides)} \\
&= \mathcal{L}(q) + \mathrm{KL}(q\|p) \\
&\geq \mathcal{L}(q) \qquad (5.18)
\end{aligned}$$

{eq:varboun d}

since the KL divergence is a non-negative quantity. The functional $\mathcal{L}(q)$ given by

$$\begin{aligned}
\mathcal{L}(q) &= \int \log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z})} q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&= \mathrm{E}_{\mathbf{z}\sim q} \log p(\mathbf{y},\mathbf{z}) + H(q), \qquad (5.19)
\end{aligned}$$

{eq:elbo1}

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 (5.18) achieves equality if and only if $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y})$, but of course the true form of the posterior is unknown to us—see Section 5.10.2 for a discussion. Maximising $\mathcal{L}(q)$ or minimising $\mathrm{KL}(q\|p)$ 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 $\mathrm{KL}(q||p)$ is impossible to compute, since one does not know the true distribution $p(\mathbf{z}|\mathbf{y})$. 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 $p(\mathbf{z}|\mathbf{y})$ in terms of KL divergence. In this case, the task is to find optimal mean and variance parameters of a normal distribution.

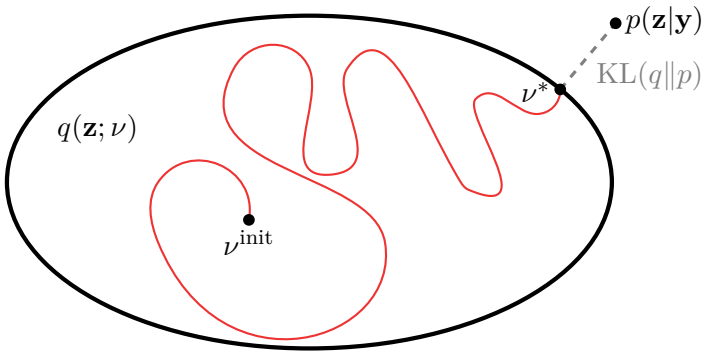


Figure 5.14: Schematic view of variational inference[5]. The aim is to find the closest distribution q (parameterised by a variational parameter ν) to p in terms of KL divergence within the set of variational distributions, represented by the ellipse.

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 $\mathbf{z}$ into M disjoint groups $\mathbf{z} = (z_{[1]}, \ldots, z_{[M]})$. Note that each factor $z_{[k]}$ may be multidimensional. Then, the structure

$$q(\mathbf{z}) = \prod_{k=1}^{M} q_k(z_{[k]})$$

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).

Let us denote the distributions which minimise the Kullback-Leibler divergence (maximise the variational lower bound) by the use of tildes. By appealing to Bishop (2006,

[5]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.

equation 10.9, p. 466), we find that for each $z_{[k]}$, $k = 1, \ldots, M$, $\tilde{q}_k$ satisfies

$$\log \tilde{q}_k(z_{[k]}) = \mathrm{E}_{-k} \log p(\mathbf{y}, \mathbf{z}) + \text{const.} \tag{5.20}$$ {eq:qtilde}

where expectation of the joint log density of $\mathbf{y}$ and $\mathbf{z}$ is taken with respect to all of the unknowns $\mathbf{z}$, except the one currently in consideration $z_{[k]}$, under their respective $\tilde{q}_k$ densities.

In practice, rather than an explicit calculation of the normalising constant, one simply needs to inspect (5.20) 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 $p(z_{[k]}|\mathbf{z}_{-k}, \mathbf{y})$, where $\mathbf{z}_{-k} = \{z_{[i]}|i \neq k\}$, follows an exponential family distribution

$$p(z_{[k]}|\mathbf{z}_{-k}, \mathbf{y}) = B(z_{[k]}) \exp\big(\langle \zeta_k(\mathbf{z}_{-k}, \mathbf{y}), z_{[k]} \rangle - A(\zeta_k)\big).$$

Then, from (5.20),

$$\begin{aligned} \tilde{q}(z_{[k]}) &\propto \exp\big(\mathrm{E}_{-k} \log p(z_{[k]}|\mathbf{z}_{-k}, \mathbf{y})\big) \\ &= \exp\Big(\log B(z_{[k]}) + \mathrm{E}\langle \zeta_k(\mathbf{z}_{-k}, \mathbf{y}), z_{[k]} \rangle - \mathrm{E}[A(\zeta_k)]\Big) \\ &\propto B(z_{[k]}) \exp \mathrm{E}\langle \zeta_\xi(\mathbf{z}_{-k}, \mathbf{y}), z_{[k]} \rangle \end{aligned}$$

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 $\mathbf{z}_{-k}$. 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.

alg:cavi

Algorithm 5 The CAVI algorithm

1: **initialise** Variational factors $q_k(z_{[k]})$
2: **while** ELBO $\mathcal{L}(q)$ not converged **do**
3: **for** $k = 1, \ldots, M$ **do**
4: $\tilde{q}_k(z_{[k]}) \leftarrow \text{const.} \times \exp \mathrm{E}_{-k} \log p(\mathbf{y}, \mathbf{z})$ ▷ from (5.20)
5: **end for**
6: $\mathcal{L}(q) \leftarrow \mathrm{E}_{\mathbf{z} \sim \prod_k \tilde{q}_k} \log p(\mathbf{y}, \mathbf{z}) + \sum_{k=1}^{m} H\left[q_k(z_{[k]})\right]$ ▷ Update ELBO
7: **end while**
8: **return** $\tilde{q}(\mathbf{z}) = \prod_{k=1}^{M} \tilde{q}_j(z_{[k]})$

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.

5.10.2 Variational methods and the EM algorithm

sec:varEM

Consider again the latent variable setup described in Section 5.10.1, but suppose the goal now is to maximise the (marginal) log-likelihood of the parameters θ of the model. We will see how the EM algorithm relates to minimising the KL divergence between a density $q(\mathbf{z})$ and the posterior of $\mathbf{z}$, and connect this idea to variational methods.

As we did in deriving (5.18), we decompose the marginal log-likelihood as

$$\log p(y|\theta) = \mathrm{E}\left[\log \frac{p(\mathbf{y}, \mathbf{z}|\theta)}{q(\mathbf{z})}\right] - \mathrm{E}\left[\log \frac{p(\mathbf{z}|\mathbf{y}, \theta)}{q(\mathbf{z})}\right] = \mathcal{L}(q) + \mathrm{KL}(q\|p).$$

This decomposition is shown in Figure 5.15. We realise that the KL divergence non-negative, and is zero exactly when $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y}, \theta)$. Substituting this into the above equation yields the relationship

$$\begin{aligned}\log p(y|\theta) &= \mathrm{E}\left[\log \frac{p(\mathbf{y}, \mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{y}, \theta)}\right] - \cancel{\mathrm{E}\left[\log \frac{p(\mathbf{z}|\mathbf{y}, \theta)}{p(\mathbf{z}|\mathbf{y}, \theta)}\right]} \\ &= \mathrm{E} \log p(\mathbf{y}, \mathbf{z}|\theta) - \mathrm{E}\, p(\mathbf{z}|\mathbf{y}, \theta).\end{aligned}$$

[6] Reproduced from Bishop (2006, Figure 9.11).

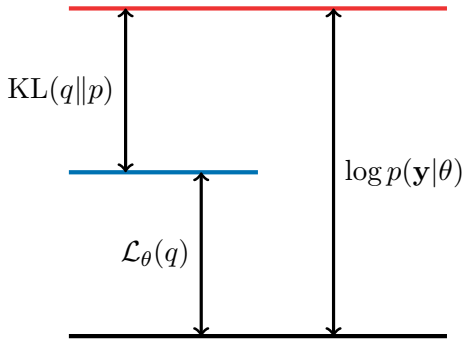


Figure 5.15: Illustration[6] of the decomposition of the log-likelihood into $\mathcal{L}_\theta(q)$ and $\mathrm{KL}[q(\mathbf{z})\|p(\mathbf{z}|\mathbf{y})]$. The quantity $\mathcal{L}_\theta(q)$ is a lower bound for the log-likelihood.

fig:loglikdecomp

By taking expectations under the posterior distribution with known parameter values $\theta^{(t)}$, the term on the left becomes the Q function of the E-step

$$Q(\theta) = Q(\theta|\theta^{(t)}) = \mathrm{E}_{\mathbf{z}}\left[\log p(\mathbf{y}, \mathbf{z}|\theta) \,\middle|\, \mathbf{y}, \theta^{(t)}\right],$$

while the term on the left is an entropy term. Thus, minimising the KL divergence corresponds to the E-step in the EM algorithm. As a side fact, 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, the EM algorithm is seen as

- **E-step**. Maximise $\mathcal{L}_\theta\big[q(\mathbf{z})\big]$ with respect to q, keeping θ fixed. This is equivalent to minimising $\mathrm{KL}(q\|p)$.
- **M-step**. Maximise $\mathcal{L}\big[q(\mathbf{z}|\theta)\big]$ with respect to θ, keeping q fixed.

When the true posterior distribution $p(\mathbf{z}|\mathbf{y})$ 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 5.10.1, we saw that constraining q to be of a factorised form, then $\tilde{q}$ is a mean-field density. This form of the EM is known as *variational Bayes EM algorithm* (VB-EM) (Beal and Ghahramani, 2003).

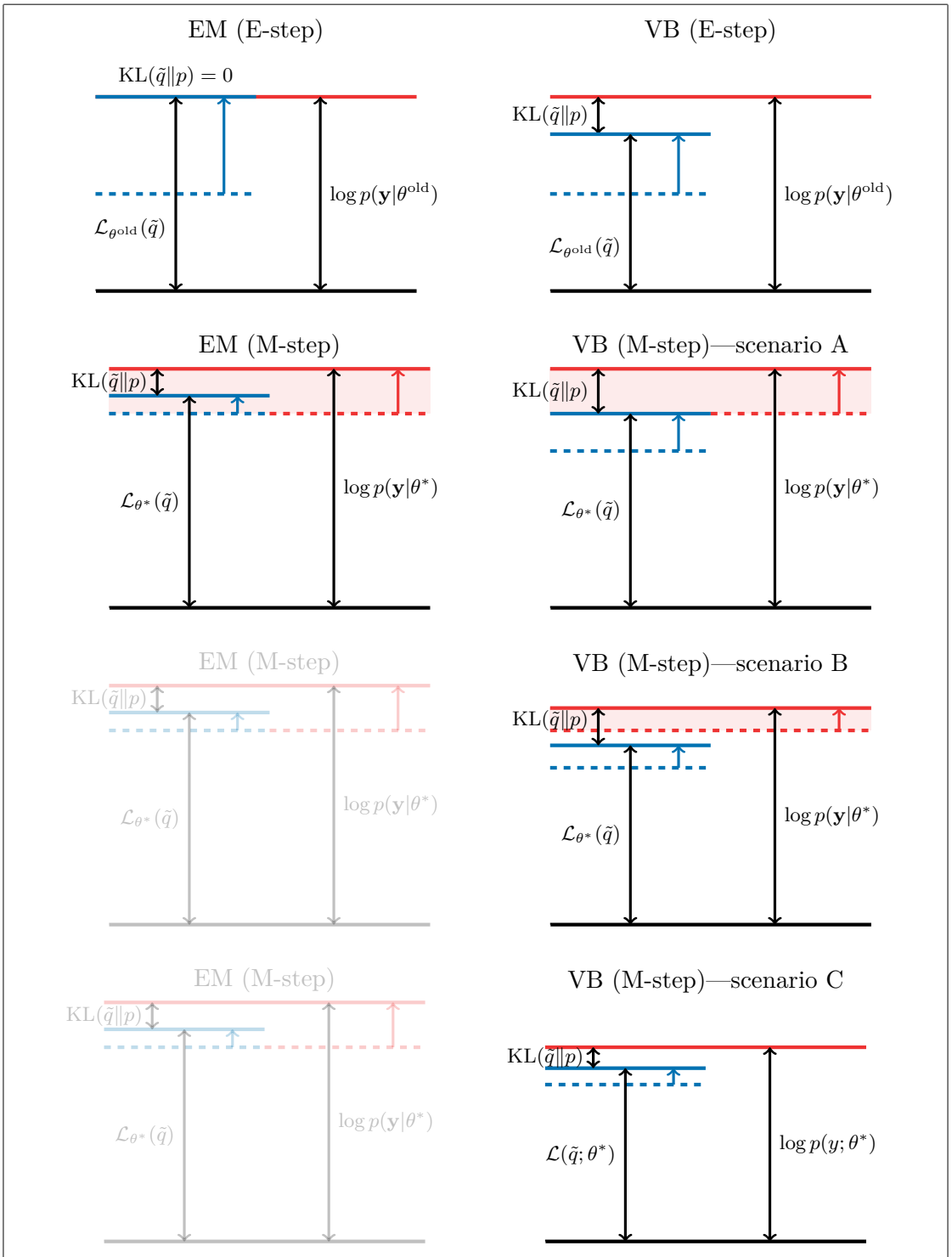


Figure 5.16: Illustration of EM vs VB-EM. Whereas the EM guarantees an increase in log-likelihood value (red shaded region), the VB-EM does not.

In variational inference, a fully Bayesian treatment of the parameters is considered, with the aim of obtaining approximation to their posterior distributions. In VB-EM, the variational approximation is only performed on the latent, or 'missing' variables, to use the EM nomenclature. After a variational E-step, the M-step proceeds as usual, and as such, all of the material relating to the EM in the previous chapter is applicable. The VB-EM can also be seen as obtaining (approximate) maximum a posteriori estimates with diffuse priors on the parameters.

5.10.3 The EM algorithm for I-probit models is intractable—variational Bayes EM?

sec:vbemipr obit

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

$$
\begin{aligned}
&\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}) \\
&\quad = \sum_{i=1}^{n} \Big\{ \log p(y_i | \mathbf{y}_{i\cdot}^*) + \log p(\mathbf{y}_{i\cdot}^* | \mathbf{w}_{i\cdot}) + \log p(\mathbf{w}_{i\cdot}) \Big\} \\
&\quad = -\frac{1}{2} \sum_{i=1}^{n} \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*] \Big[(\mathbf{y}_{i\cdot}^* - \boldsymbol{\alpha} - \mathbf{w}_{i\cdot}^\top \mathbf{h}_\eta(x_i))^\top \boldsymbol{\Psi} (\mathbf{y}_{i\cdot}^* - \boldsymbol{\alpha} - \mathbf{w}_{i\cdot}^\top \mathbf{h}_\eta(x_i)) \\
&\qquad\qquad + \mathbf{w}_{i\cdot}^\top \boldsymbol{\Psi}^{-1} \mathbf{w}_{i\cdot} \Big] + \text{const.}
\end{aligned}
$$

which looks like the complete data log-likelihood seen previously in (4.9), except that here, together with the $\mathbf{w}_{i\cdot}$'s, the $\mathbf{y}_{i\cdot}^*$'s are never observed.

For the E-step, it is of interest to determine the posterior density $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}) = p(\mathbf{y}^*|\mathbf{w}, \mathbf{y})p(\mathbf{w}|\mathbf{y})$, which apparently is hard to obtain. We can go as far as determining that the full conditional of the latent propensities is multivariate subject to a conical truncation $\mathcal{C}_j = \{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}$, i.e. $\mathbf{y}_{i\cdot}^* | \mathbf{w}_{i\cdot}, \{y_i = j\} \overset{\text{iid}}{\sim} {}^{\text{t}}\text{N}_m(\boldsymbol{\alpha} + \mathbf{w}_{i\cdot}^\top \mathbf{h}_\eta(x_i), \boldsymbol{\Psi}^{-1}, \mathcal{C}_j)$, for each $i = 1, \dots, n$, and that $\text{vec}\, \mathbf{w} | \mathbf{y}^* \sim \text{N}(\tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$ is found to be similar to the distribution in (5.14). To obtain the first and second posterior moments for the I-prior

random effects, we can use the law of total expectations:

$$\mathrm{E}[\operatorname{vec}\mathbf{w}|\mathbf{y}] = \mathrm{E}_{\mathbf{y}^*}\left[\mathrm{E}[\operatorname{vec}\mathbf{w}|\mathbf{y}^*]\middle|\mathbf{y}\right] =: \hat{\mathbf{w}}$$

and

$$\mathrm{E}[\operatorname{vec}\mathbf{w}(\operatorname{vec}\mathbf{w})^\top|\mathbf{y}] = \mathrm{E}_{\mathbf{y}^*}\left[\mathrm{E}[\operatorname{vec}\mathbf{w}(\operatorname{vec}\mathbf{w})^\top|\mathbf{y}^*]\middle|\mathbf{y}\right] =: \hat{\mathbf{W}},$$

but this requires $p(\mathbf{y}^*|\mathbf{y})$ which does not come by easily. A similar problem has been faced by Chan and Kuk (1997), who analysed binary linear probit models with random effects. The authors ultimately resort to Monte Carlo sampling within an EM framework to overcome the difficult distributions of interest.

Suppose that, instead of the true posterior distribution $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y})$ being used, a mean-field variational approximation $q(\mathbf{y}^*, \mathbf{w}) = q(\mathbf{y}^*)q(\mathbf{w})$ is used instead. As we know from Section 5.5, $q(\mathbf{y}^*)$ is a truncated multivariate normal distribution, and $q(\mathbf{w})$ is multivariate normal, whose means and second moments can be computed with some effort. Let $\bar{\mathbf{y}}^* = \mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top$. The (approximate) E-step then entails computing

$$\begin{aligned} Q(\theta) &= \mathrm{E}_{\mathbf{y}^*,\mathbf{w}\sim q} \log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}|\theta) \\ &= \text{const.} - \frac{1}{2}\operatorname{tr}\mathrm{E}_{\mathbf{y}^*,\mathbf{w}\sim q}\left[\boldsymbol{\Psi}(\bar{\mathbf{y}}^{*\top}\bar{\mathbf{y}}^* + \mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w} - 2\bar{\mathbf{y}}^*\boldsymbol{\Psi}\mathbf{w}^\top\mathbf{H}_\eta) + \boldsymbol{\Psi}^{-1}\mathbf{w}^\top\mathbf{w}\right]. \end{aligned}$$

In the M-step, this is maximised with respect to θ. This is the VB-EM algorithm described in Section 5.10.2. As per the discussion in Section 5.7.3, this alleviates the problem of non-conjugacy of the complete conditional for $\boldsymbol{\Psi}$. One downside to VB-EM is that it is not entirely certain how one could obtain standard errors for the parameters, other than by bootstrapping, which for the I-probit model, is likely to be computationally intensive.

Bibliography

agresti2000tutorial 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.

albert1993bayesian 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.

alpay1991some Alpay, Daniel (1991). "Some remarks on reproducing kernel Krein spaces". In: *The Rocky Mountain Journal of Mathematics*, pp. 1189–1205.

balakrishnan1981applied Balakrishnan, Alampallam V (1981). *Applied Functional Analysis*. 2nd ed. Vol. 3. Springer Science & Business Media. DOI: 10.1007/978-1-4612-5865-0.

beal2003 Beal, M. J. and Z. 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: Oxford University Press, pp. 453–464.

bergsma2017 Bergsma, Wicher (2017). "Regression with I-priors". In: *Unpublished manuscript*.

berlinet2011reproducing Berlinet, Alain and Christine Thomas-Agnan (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer-Verlag. DOI: 10.1007/978-1-4419-9096-9.

bernardo2003variational Bernardo, JM, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, M West, et al. (2003). "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures". In: *Bayesian statistics* 7, pp. 453–464.

bishop2006pattern Bishop, Christopher (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.

blei2017variational
Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* just-accepted.

bouboulis2011extension
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.

breiman2001random
Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

cannings2017random
Cannings, Timothy I and Richard J Samworth (2017). "Random-projection ensemble classification". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology) with discussion* 79.4, pp. 959–1035.

carpenter2016stan
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.

casella2002statistical
Casella, George and Roger L Berger (2002). *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA.

chan1997maximum
Chan, Jennifer SK and Anthony YC Kuk (1997). "Maximum likelihood estimation for probit-linear mixed models with correlated random effects". In: *Biometrics*, pp. 86–97.

chen2011single
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.

cheng2017variational
Cheng, Ching-An and Byron Boots (2017). "Variational Inference for Gaussian Process Models with Linear Complexity". In: *Advances in Neural Information Processing Systems*, pp. 5190–5200.

chopin2011fast
Chopin, Nicolas (2011). "Fast simulation of truncated Gaussian distributions". In: *Statistics and Computing* 21.2, pp. 275–288.

cohen2002
Cohen, S (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.

damien2001sampling 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.

davidian1995nonlinear Davidian, Marie and David M Giltinan (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall/CRC.

dean1999design Dean, Angela and Daniel Voss (1999). *Design and analysis of experiments*. Vol. 1. Springer.

dempster1977maximum 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 (methodological)*, pp. 1–38.

denwood2016runjags 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.

deterding1989speaker Deterding, David Henry (1989). "Speaker normalization for automatic speech recognition". PhD thesis. University of Cambridge.

diggle2013spatial Diggle, Peter J, Paula Moraga, Barry Rowlingson, and Benjamin M Taylor (2013). "Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm". In: *Statistical Science*, pp. 542–563.

diggle2005nonparametric 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.

duane1987hybrid Duane, Simon, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth (1987). "Hybrid monte carlo". In: *Physics letters B* 195.2, pp. 216–222.

durrande2013anova 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.

duvenaud2014automatic Duvenaud, David (2014). "Automatic model construction with Gaussian processes". PhD thesis. University of Cambridge.

eddelbuettel2011rcpp
Eddelbuettel, Dirk and Romain Francois (2011). "**Rcpp**: Seamless R and C++ Integration". In: *Journal of Statistical Software* 40.8, pp. 1–18. DOI: 10.18637/jss.v040.i08.

embrechts2002selfsimilar
Embrechts, Paul and Makoto Maejima (2002). *Selfsimilar Processes. Princeton series in applied mathematics*. Princeton University Press, Princeton, NJ.

ferraty2006nonparametric
Ferraty, Frédéric and Philippe Vieu (2006). *Nonparametric Functional Data Analysis*. 1st. Springer-Verlag. DOI: 10.1007/0-387-36620-2.

ra1922mathematical
Fisher, RA (1922). "On the mathematical foundations of theoretical statistics". In: *Phil. Trans. R. Soc. Lond. A* 222.594-604, pp. 309–368.

fowlkes2001efficient
Fowlkes, C, S Belongie, and J Malik (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, pp. 231–238. DOI: 10.1109/CVPR.2001.990481.

friedman2001elements
Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

geweke1991efficient
Geweke, John (1991). *Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities*.

geweke1994alternative
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*, pp. 609–632.

girolami2006variational
Girolami, Mark and Simon Rogers (2006). "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors". In: *Neural Computation* 18.8, pp. 1790–1817.

groves1969note
Groves, Theodore and Thomas Rothenberg (1969). "A note on the expected value of an inverse matrix". In: *Biometrika* 56.3, pp. 690–691.

gu2013smoothing
Gu, Chong (2013). *Smoothing spline ANOVA models*. Vol. 297. Springer Science & Business Media.

guvenir1997supervised
Guvenir, H Altay, Burak Acar, Gulsen Demiroz, and Ayhan Cekin (1997). "A supervised machine learning algorithm for arrhythmia analysis". In: *Computers in Cardiology 1997*. IEEE, pp. 433–436.

hajivassiliou1996simulation
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.

hastie1986
Hastie, Trevor and Robert Tibshirani (Aug. 1986). "Generalized Additive Models". In: *Statist. Sci.* 1.3, pp. 297–310. DOI: 10.1214/ss/1177013604. URL: https://doi.org/10.1214/ss/1177013604.

hein2004kernels
Hein, Matthias and Olivier Bousquet (2004). "Kernels, associated structures and generalizations". In: *Max-Planck-Institut fuer biologische Kybernetik, Technical Report*.

hensman2013gaussian
Hensman, James, Nicolo Fusi, and Neil D Lawrence (2013). "Gaussian processes for big data". In: *arXiv preprint arXiv:1309.6835*.

itzykson1991statistical
Itzykson, Claude and Jean Michel Drouffe (1991). *Statistical Field Theory: Volume 2, Strong Coupling, Monte Carlo Methods, Conformal Field Theory and Random Systems*. Cambridge University Press.

jamil2017
Jamil, Haziq and Wicher Bergsma (2017). "iprior: An R Package for Regression Modelling using I-priors". In: *Manuscript in submission*.

jaynes1957a
Jaynes, Edwin T (1957a). "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4, p. 620.

jaynes1957b
— (1957b). "Information Theory and Statistical Mechanics II". In: *Physical Review* 108.2, p. 171.

kammar2016
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.

kass1995bayes
Kass, Robert E and Adrian E Raftery (1995). "Bayes factors". In: *Journal of the american statistical association* 90.430, pp. 773–795.

keane1994computationally
Keane, Michael P (1994). "A computationally practical simulation estimator for panel data". In: *Econometrica: Journal of the Econometric Society*, pp. 95–116.

Keane1992
Keane, Michael P. (1992). "A Note on Identification in the Multinomial Probit Model". In: *Journal of Business & Economic Statistics* 10.2, pp. 193–200. ISSN: 0735-0015. DOI: 10.2307/1391677. URL: http://www.jstor.org/stable/1391677%5Cnhttp://www.jstor.org/stable/pdfplus/1391677.pdf?acceptTC=true.

kenward1987method — 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.

kimeldorf1970correspondence — 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.

kree1974produits — Krée, Paul (1974). "Produits tensoriels complétés d'espaces de Hilbert". In: *Séminaire Paul Krée* 1.7, pp. 1974–1975.

caret — Kuhn, Max et al. (2017). ***caret**: Classification and Regression Training*. R package version 6.0–77. URL: https://CRAN.R-project.org/package=caret.

kuo2010decompositions — Kuo, F, I Sloan, G Wasilkowski, and Henryk Woźniakowski (2010). "On decompositions of multivariate functions". In: *Mathematics of computation* 79.270, pp. 953–966.

kuss2005assessing — Kuss, Malte and Carl Edward Rasmussen (2005). "Assessing approximate inference for binary Gaussian process classification". In: *Journal of machine learning research* 6.Oct, pp. 1679–1704.

lange1995quasi — Lange, Kenneth (1995). "A quasi-Newton acceleration of the EM algorithm". In: *Statistica sinica*, pp. 1–18.

lian2014series — Lian, Heng and Gaorong Li (2014). "Series Expansion for Functional Sufficient Dimension Reduction". In: *Journal of Multivariate Analysis* 124.C, pp. 150–165. DOI: 10.1016/j.jmva.2013.10.019.

liu1998parameter — 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.

lunn2000winbugs — 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.

mandelbrot1968fractional — Mandelbrot, Benoit B and John W Van Ness (1968). "Fractional Brownian motions, fractional noises and applications". In: *SIAM review* 10.4, pp. 422–437.

marsaglia2000ziggurat — Marsaglia, George and Wai Wan Tsang (2000). "The ziggurat method for generating random variables". In: *Journal of statistical software* 5.8, pp. 1–7.

mary2003hilbertian — Mary, Xavier (2003). "Hilbertian subspaces, subdualities and applications". PhD thesis. INSA de Rouen.

mccullagh1989 McCullagh, P. and John A. Nelder (1989). *Generalized Linear Models*. 2nd. Chapman & Hall/CRC Press.

meng1993maximum 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.

meng1997algorithm 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.

micchelli2006universal Micchelli, Charles A, Yuesheng Xu, and Haizhang Zhang (2006). "Universal kernels". In: *Journal of Machine Learning Research* 7.Dec, pp. 2651–2667.

minka2001expectation Minka, Thomas P (2001). "Expectation propagation for approximate Bayesian inference". In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 362–369.

neal1999 Neal, Radford M. (1999). "Regression and Classification using Gaussian Process Priors". In: *Bayesian Statistics*. Ed. by J M Bernardo, J O Berger, A P Dawid, and A F M Smith. Vol. 6. Oxford University Press. (with discussion), pp. 475–501.

neal2011mcmc Neal, Radford M et al. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of Markov Chain Monte Carlo* 2.11.

ong2004learning Ong, Cheng Soon, Xavier Mary, Stéphane Canu, and Alexander J Smola (2004). "Learning with non-positive kernels". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM, p. 81.

jmcm Pan, Jianxin and Yi Pan (2016). ***jmcm**: Joint Mean-Covariance Models using Armadillo and S4*. R package version 0.1.7.0. URL: https://CRAN.R-project.org/package=jmcm.

pawitan2001all Pawitan, Yudi (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

petersen2008matrix Petersen, Kaare Brandt and Michael Syskind Pedersen (2008). "The matrix cookbook". In: *Technical University of Denmark* 7.15, p. 510.

pinheiro2000mixed Pinheiro, José C and Douglas M Bates (2000). *Mixed-Effects Models in S and S-plus*. Springer-Verlag. DOI: 10.1007/b98882.

nlme
Pinheiro, Joséo, Douglas 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.

plummer2003jags
Plummer, Martyn (2003). "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling". In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vol. 124. Vienna, Austria, p. 125.

quinonero2005unifying
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.

rasmussen2006gaussian
Rasmussen, Carl Edward and Christopher K I Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

reed1972methods
Reed, Michael and Barry Simon (1972). *Methods of mathematical physics I: Functional analysis*.

robert1995simulation
Robert, Christian P (1995). "Simulation of truncated normal variables". In: *Statistics and computing* 5.2, pp. 121–125.

robinson1989dynamic
Robinson, Anthony John (1989). "Dynamic error propagation networks". PhD thesis. University of Cambridge.

rudin1987real
Rudin, Walter (1987). *Real and complex analysis*. Tata McGraw-Hill Education.

schoenberg1937
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*, pp. 787–793.

scholkopf2002learning
Schölkopf, Bernhard and Alexander J Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

sejdinovic2012
Sejdinovic, Dino and Arthur Gretton (2012). "Lecture notes: What is an RKHS?" In: *COMPGI13 Advanced Topics in Machine Learning. Lecture conducted at University College London*, pp. 1–24. URL: http://www.gatsby.ucl.ac.uk/%7B~%7Dgretton/coursefiles/RKHS%7B%5C_%7DNotes1.pdf.

skrondal2004generalized
Skrondal, Anders and Sophia Rabe-Hesketh (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

sobol2001global
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.

rstan
Stan Development Team (2016). ***RStan**: The R Interface to Stan*. R package version 2.14.1. URL: http://mc-stan.org/.

steinwart2008support
Steinwart, Ingo and Andreas Christmann (2008). *Support vector machines*. Springer Science & Business Media.

steinwart2006explicit
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.

sturtz2005r2winbugs
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.

tapia1971diff
Tapia, R A (1971). *The differentiation and integration of nonlinear operators*. Ed. by Louis B Rall.

taylor2013lgcp
Taylor, Benjamin M, Tilman M Davies, Barry S Rowlingson, Peter J Diggle, et al. (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.

thodberg1996review
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.

tibshirani2002diagnosis
Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression". In: *Proceedings of the National Academy of Sciences* 99.10, pp. 6567–6572.

titsias2009variational
Titsias, Michalis (2009). "Variational learning of inducing variables in sparse Gaussian processes". In: *Artificial Intelligence and Statistics*, pp. 567–574.

train2009discrete
Train, Kenneth E (2009). *Discrete choice methods with simulation*. Cambridge university press.

van2008reproducing
van der Vaart, Aad W and van Zanten (2008). "Reproducing kernel Hilbert spaces of Gaussian priors". In: *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*. Institute of Mathematical Statistics, pp. 200–222.

wahba1990spline Wahba, Grace (1990). *Spline models for observational data*. Vol. 59. Siam.

wasserman2013all Wasserman, Larry (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

williams2001using 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*. The MIT Press, pp. 682–688.

yu2012monotonically Yu, Yaming (2012). "Monotonically overrelaxed EM algorithms". In: *Journal of Computational and Graphical Statistics* 21.2, pp. 518–537.

zafeiriou2012subspace Zafeiriou, Stefanos (2012). "Subspace learning in krein spaces: Complete kernel fisher discriminant analysis with indefinite kernels". In: *European Conference on Computer Vision*. Springer, pp. 488–501.

zellner1986assessing Zellner, Arnold (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions". In: *Bayesian inference and decision techniques*.

zhang2013kronecker Zhang, Huamin and Feng Ding (2013). "On the Kronecker products and their applications". In: *Journal of Applied Mathematics* 2013.

zhu2014structured 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 B (Statistical Methodology)* 76.3, pp. 581–603. DOI: 10.1111/rssb.12036.

Appendix A

Regression modelling using I-priors

A.1 Deriving the posterior distribution for w

apx:posteriorw

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

$$\begin{pmatrix}\mathbf{y} \\ \mathbf{w}\end{pmatrix} \sim \mathrm{N}_{n+n}\left(\begin{pmatrix}\boldsymbol{\alpha} + \mathbf{f}_0 \\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{V}_y & \mathbf{H}_\eta\boldsymbol{\Psi} \\ \mathbf{H}_\eta\boldsymbol{\Psi} & \boldsymbol{\Psi}\end{pmatrix}\right).$$

Thus,

$$\begin{aligned}\mathrm{E}[\mathbf{w}|\mathbf{y}] &= \mathrm{E}\,\mathbf{w} + \operatorname{Cov}(\mathbf{w}, \mathbf{y})(\operatorname{Var}\mathbf{y})^{-1}(\mathbf{y} - \mathrm{E}\,\mathbf{y}) \\ &= \mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{V}_y^{-1}(\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0),\end{aligned}$$

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.

A.2 A recap on the exponential family EM algorithm

apx:expem

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 $p(\mathbf{z}|\boldsymbol{\theta}) = B(\mathbf{z})\exp\left(\langle\boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{T}(\mathbf{z})\rangle - A(\boldsymbol{\theta})\right)$, where $\boldsymbol{\eta} : \mathbb{R}^s \mapsto \mathbb{R}$ is a link function, $\mathbf{T}(\mathbf{z}) = \left(T_1(\mathbf{z}), \ldots, T_s(\mathbf{z})\right)^\top \in \mathbb{R}^s$ 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

$$p(\mathbf{z}|\boldsymbol{\eta}) = B(\mathbf{z}) \exp\left(\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z})\rangle - A^*(\boldsymbol{\eta})\right) \tag{A.1}$$

{eq:pdfexpfamnat}

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

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{z}|\boldsymbol{\eta}) = \mathbf{T}(\mathbf{z}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = 0. \tag{A.2}$$

{eq:expEM1}

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

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}|\boldsymbol{\eta}) &= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}|\boldsymbol{\eta}) \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} \left(\int p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \, \mathrm{d}\mathbf{w} \right) \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \int \left(\frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) \mathrm{d}\mathbf{w} \\
&= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \int \left(p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) \mathrm{d}\mathbf{w} \\
&= \int \left(\mathbf{T}(\mathbf{y}, \mathbf{w}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) \right) p(\mathbf{w}|\mathbf{y}, \boldsymbol{\eta}) \, \mathrm{d}\mathbf{w} \\
&= \mathrm{E}_{\mathbf{w}} \left[\mathbf{T}(\mathbf{y}, \mathbf{w}) | \mathbf{y} \right] - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta})
\end{aligned} \tag{A.3}$$

{eq:expEM2}

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 (A.2) and that of (A.3) 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 (A.3) 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 6.

Algorithm 6 Exponential family EM

alg:EM3

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 6 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 (A.1), 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 (A.3) when obtaining ML estimate of $\boldsymbol{\eta}$. Thus, Q_t is maximised by the solution to line 4 in Algorithm Algorithm 6.

A.3 Deriving the posterior predictive distribution

apx:postpred

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

$$\begin{pmatrix} y_{\text{new}} \\ \mathbf{y} \end{pmatrix} \sim \mathrm{N}_{n+1}\left(\begin{pmatrix} \hat{\alpha} \\ \hat{\alpha}\mathbf{1}_n \end{pmatrix}, \begin{pmatrix} v_{\text{new}} & \mathrm{Cov}(y_{\text{new}}, \mathbf{y}) \\ \mathrm{Cov}(y_{\text{new}}, \mathbf{y})^\top & \tilde{\mathbf{V}}_y \end{pmatrix} \right),$$

where

$$\begin{aligned} \mathrm{Cov}(y_{\text{new}}, \mathbf{y}) &= \mathrm{Cov}(f_{\text{new}} + \epsilon_{\text{new}}, \mathbf{f} + \boldsymbol{\epsilon}) \\ &= \mathrm{Cov}(f_{\text{new}}, \mathbf{f}) + \mathrm{Cov}(\epsilon_{\text{new}}, \boldsymbol{\epsilon}) \\ &= \mathrm{Cov}\left(\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \tilde{\mathbf{w}}, \mathbf{H}_{\tilde{\eta}} \tilde{\mathbf{w}} \right) + (\sigma_{\text{new},1}, \ldots, \sigma_{\text{new},n}) \\ &= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\tilde{\eta}} + \boldsymbol{\sigma}_{\text{new}}. \end{aligned}$$

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. Assuming $\boldsymbol{\sigma}_{\text{new}} = \mathbf{0}$ would be acceptable, especially under an iid assumption the error precisions. In any case, using standard multivariate normal results, we get that $y_{\text{new}}|\mathbf{y}$ is also normally distributed with mean

$$\begin{aligned} E[y_{\text{new}}|\mathbf{y}] &= \hat{\alpha} + (\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\tilde{\eta}} + \boldsymbol{\sigma}_{\text{new}}) \tilde{\mathbf{V}}_y^{-1} \tilde{\mathbf{y}} \\ &= \hat{\alpha} + \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \overbrace{\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\tilde{\eta}} \tilde{\mathbf{V}}_y^{-1} \tilde{\mathbf{y}}}^{\hat{\mathbf{w}}} + \boldsymbol{\sigma}_{\text{new}} \tilde{\mathbf{V}}_y^{-1} \tilde{\mathbf{y}} \\ &= \hat{\alpha} + \mathrm{E}\left[f(x_{\text{new}}) | \mathbf{y} \right] + \text{mean correction term} \end{aligned}$$

and variance

$$\begin{aligned} \mathrm{Var}[y_{\text{new}}|\mathbf{y}] &= v_{\text{new}} - (\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\tilde{\eta}} + \boldsymbol{\sigma}_{\text{new}}) \tilde{\mathbf{V}}_y^{-1} (\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\tilde{\eta}} + \boldsymbol{\sigma}_{\text{new}})^\top \\ &= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \hat{\mathbf{h}}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1} - \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\tilde{\eta}} \tilde{\mathbf{V}}_y^{-1} \mathbf{H}_{\tilde{\eta}} \hat{\boldsymbol{\Psi}} \mathbf{h}_{\hat{\eta}}(x_{\text{new}}) \\ &\quad + \text{variance correction term} \\ &= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \left(\hat{\boldsymbol{\Psi}} - \hat{\boldsymbol{\Psi}} \mathbf{H}_{\tilde{\eta}} \tilde{\mathbf{V}}_y^{-1} \mathbf{H}_{\tilde{\eta}} \hat{\boldsymbol{\Psi}} \right) \mathbf{h}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1} \\ &\quad + \text{variance correction term} \\ &= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\mathbf{V}}_y^{-1} \mathbf{h}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1} + \text{variance correction term} \\ &= \mathrm{Var}\left[f(x_{\text{new}}) | \mathbf{y} \right] + \psi_{\text{new}}^{-1} + \text{variance correction term}. \end{aligned}$$

A.4 Derivation of the Fisher information for multivariate normal distributions

apx:fisherm ultinormal

Let $X \sim \mathrm{N}_p(0, \Sigma_\theta)$, that is, the covariance matrix Σ_θ depends on a real, q-dimensional vector θ. 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 elementwise. The two identities below are useful:

$$\frac{\partial}{\partial z}\operatorname{tr}\Sigma = \operatorname{tr}\frac{\partial\Sigma}{\partial z} \tag{A.4}$$

$$\frac{\partial}{\partial z}\log|\Sigma| = \operatorname{tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial z}\right) \tag{A.5}$$

$$\frac{\partial\Sigma^{-1}}{\partial z} = -\Sigma^{-1}\frac{\partial\Sigma}{\partial z}\Sigma^{-1} \tag{A.6}$$

A useful reference for these identities is Petersen and Pedersen (2008).

Taking derivative of the log-likelihood for θ with respect to the i'th component yields

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}L(\theta|X) &= -\frac{1}{2}\frac{\partial}{\partial\theta_i}\log|\Sigma_\theta| - \frac{1}{2}\frac{\partial}{\partial\theta_i}\operatorname{tr}(\Sigma_\theta^{-1}XX^\top) \\
&= -\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\right) - \frac{1}{2}\operatorname{tr}\left(\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_i}XX^\top\right) \\
&= -\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i} - \Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}XX^\top\right).
\end{aligned}$$

Taking derivatives again, this time with respect to θ_j, we get

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_i\theta_j}L(\theta|X) &= -\frac{1}{2}\frac{\partial}{\partial\theta_j}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i} - \Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}XX^\top\right) \\
&= -\frac{1}{2}\operatorname{tr}\left(\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i} + \Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j} - \frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}XX^\top\right. \\
&\qquad \left. - \Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j}\Sigma_\theta^{-1}XX^\top - \Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}XX^\top\right).
\end{aligned}$$

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) $\mathrm{E}[\operatorname{tr}\Sigma] = \operatorname{tr}(\mathrm{E}\,\Sigma)$, 2) $\mathrm{E}[XX^\top] = \Sigma_\theta$; and 3) the

trace is invariant under cyclic permutations, we get

$$\begin{aligned} U_{ij} &= \frac{1}{2}\operatorname{tr}\left(\cancel{\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i}} + \cancel{\Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j}} - \cancel{\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i}} - \cancel{\Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j}} - \frac{\partial\Sigma_\theta}{\partial\theta_i}\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\right) \\ &= \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_j}\right) \end{aligned}$$

as required.

Appendix B

I-priors for categorical responses

B.1 Some distributions and their properties

This is a reference relating to the multivariate normal, matrix normal, truncated univariate and multivariate normal, Wishart, and gamma distributions which are collated from various sources for convenience. Of interest are probability density functions, first and second moments, and entropy (Definition 3.5, page 86).

B.1.1 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**. $H(p) = \frac{1}{2}\log|2\pi e\Sigma| = \frac{d}{2}(1 + \log 2\pi) + \frac{1}{2}\log|\Sigma|$.

Lemma B.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 (2008, §8). □

Frequently, in Bayesian statistics especially, the following identities will be useful in deriving posterior distributions involving multivariate normals.

Lemma B.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,*

$$-\frac{1}{2}x^\top Ax + b^\top x = -\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) + \frac{1}{2}b^\top A^{-1}b$$
$$-\frac{1}{2}\operatorname{tr}(X^\top AX) + \operatorname{tr}(B^\top X) = -\frac{1}{2}\operatorname{tr}\left((X - A^{-1}B)^\top A(X - A^{-1}B)\right) + \frac{1}{2}\operatorname{tr}(B^\top A^{-1}B).$$

Proof. Omitted—see Petersen and Pedersen (2008, §8.1.6). □

B.1.2 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**. $\mathrm{E}\,X = \mu$, $\mathrm{Var}(X_{i\cdot}) = \Psi$ for $i = 1,\dots,n$, and $\mathrm{Var}(X_{\cdot j}) = \Sigma$ for $j = 1,\dots,m$.
- **Entropy**. $H(p) = \frac{1}{2}\log|2\pi e(\Psi\otimes\Sigma)| = \frac{nm}{2}(1+\log 2\pi) + \frac{1}{2}\log|\Sigma|^m|\Psi|^n$.

In the above, '$\otimes$' denotes the Kronecker matrix product defined by

$$\Psi\otimes\Sigma = \begin{pmatrix} \Psi_{11}\Sigma & \Psi_{12}\Sigma & \cdots & \Psi_{1m}\Sigma \\ \Psi_{21}\Sigma & \Psi_{22}\Sigma & \cdots & \Psi_{2m}\Sigma \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{m1}\Sigma & \Psi_{m2}\Sigma & \cdots & \Psi_{mm}\Sigma \end{pmatrix} \in \mathbb{R}^{nm\times nm}.$$

Of use will be these properties of the Kronecker product (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 vectorisation operation 'vec' stacks the columns of the matrices into one long vector, for instance,

$$\operatorname{vec} \Psi = (\Psi_{11}, \ldots, \Psi_{m1}, \Psi_{12}, \ldots, \Psi_{m2}, \ldots, \Psi_{1m}, \ldots, \Psi_{mm})^\top \in \mathbb{R}^{m \times m}.$$

Lemma B.3 (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 $X \sim \mathrm{MN}(1_n\mu^\top, I_n, \Psi)$.

B.1.3 Truncated univariate normal distribution

Let $X \sim \mathrm{N}(\mu, \sigma^2)$ with X lying in the interval (a, b). 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**.

$$
\begin{aligned}
H(p) &= \frac{1}{2}\log 2\pi e\sigma^2 + \log C + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{2C} \\
&= \frac{1}{2}\log 2\pi e\sigma^2 + \log C + \frac{1}{2\sigma^2}\cdot \overbrace{\sigma^2\frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{C}}^{\text{Var}\,X - \sigma^2 + (\text{E}\,X-\mu)^2} \\
&= \frac{1}{2}\log 2\pi\sigma^2 + \log C + \frac{1}{2\sigma^2}\,\text{E}[X-\mu]^2
\end{aligned}
$$

because $\text{Var}\,X + (\text{E}\,X - \mu)^2 = \text{E}\,X^2 - \cancel{(\text{E}\,X)^2} + \cancel{(\text{E}\,X)^2} + \mu^2 - 2\mu\,\text{E}\,X$.

For binary probit models, the distributions that come up are one-sided truncations at zero, i.e. ${}^{\text{t}}\text{N}(\mu, \sigma^2, 0, +\infty)$ (upper tail/positive part) and ${}^{\text{t}}\text{N}(\mu, \sigma^2, -\infty, 0)$ (lower tail/negative part), for which their moments are of interest. As an aside, if $\mu = 0$ then the truncation ${}^{\text{t}}\text{N}(0, \sigma^2, 0, +\infty)$ is called the *half-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, $C = \Phi(-\mu/\sigma) - \Phi(-\infty) = 1 - \Phi(\mu/\sigma)$.

One may simulate random draws from a truncated normal distribution by drawing from $\text{N}(\mu, \sigma^2)$ 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 $U \sim \text{Unif}(0, 1)$ 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.

B.1.4 Truncated multivariate normal distribution

Consider the restriction of $X \sim \text{N}_d(\mu, \Sigma)$ to a convex subset[1] $\mathcal{A} \subset \mathbb{R}^d$. Call this distribution the truncated multivariate normal distribution, and denote it $X \sim {}^{\text{t}}\text{N}_d(\mu, \Sigma, \mathcal{A})$.

[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.

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) and Damien and Walker (2001). In the latter, the authors explore a simple Gibbs-based approach that is easy to implement in practice. 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 conditionals of X_j given X_{-j} is

$$\begin{aligned} X_j &\sim {}^{\mathrm{t}}\mathrm{N}(\tilde{\mu}_j,\tilde{\sigma}_j^2,x_j^-,x_j^+)\\ \tilde{\mu}_j &= \mu_j + \Sigma_{j,-j}^\top\Sigma_{-j,-j}(x_{-j}-\mu_{-j})\\ \tilde{\sigma}_j^2 &= \Sigma_{11} - \Sigma_{j,-j}^\top\Sigma_{-j,-j}\Sigma_{j,-j}. \end{aligned}$$

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}. Introduce a latent variable $Y\in\mathbb{R}$ such that the joint pdf of X and Y is

$$p(X_1,\ldots,X_d,Y)\propto\exp(-Y/2)\,\mathbb{1}[y>(x-\mu)^\top\Sigma^{-1}(x-\mu)]\,\mathbb{1}[X\in\mathcal{A}].$$

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.

thm:contruncn

Lemma B.4. *Let $X \sim {}^{\text{t}}\text{N}_d(\mu, \Sigma, \mathcal{C}_j)$, with $\mu = (\mu_1, \ldots, \mu_d)^\top$ and $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, 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

$$C = \text{E}_Z \left[\prod_{\substack{i=1 \\ i \neq j}}^{d} \Phi\left(\frac{\sigma_j}{\sigma_i} Z + \frac{\mu_j - \mu_i}{\sigma_i} \right) \right]$$

where $Z \sim \text{N}(0, 1)$.

(ii) ***Moments****. The expectation $\text{E}\, X = \left(\text{E}\, X_1, \ldots, \text{E}\, X_d \right)^\top$ is given by*

$$\text{E}\, X_i = \begin{cases} \mu_i - \sigma_i C^{-1} \text{E}_Z \left[\phi_i \prod_{k \neq i,j} \Phi_k \right] & \text{if } i \neq j \\ \mu_j - \sigma_j \sum_{i \neq j} \left(\text{E}\, X_i - \mu_i \right) & \text{if } i = j \end{cases}$$

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 B.2 for the proof. □

B.2 Proofs related to conically truncated multivariate normal distribution

apx:contrun proof

B.2.1 Proof of Lemma B.4: Pdf

Using the fact that $\int p(x)\,\mathrm{d}x = 1$, and that

$$
\begin{aligned}
&\int \cdots \int \mathbb{1}[x_i < x_j, \forall i \neq j] \prod_{i=1}^{d} \phi(x_i|\mu_i, \sigma_i^2)\,\mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&\quad = \int \cdots \int \mathbb{1}[x_i < x_j, \forall i \neq j] \prod_{i=1}^{d} \left[\frac{1}{\sigma_i}\phi\left(\frac{x_i - \mu_i}{\sigma_i}\right)\right] \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&\quad = \int \cdots \int \mathbb{1}[x_i < x_j, \forall i \neq j] \frac{1}{\sigma_j}\phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \prod_{\substack{i=1 \\ i\neq j}}^{d} \left[\frac{1}{\sigma_i}\phi\left(\frac{x_i - \mu_i}{\sigma_i}\right)\right] \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&\quad = \int \prod_{\substack{i=1 \\ i\neq j}}^{d} \Phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \frac{1}{\sigma_j}\phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \mathrm{d}x_j \\
&\quad = \int \prod_{\substack{i=1 \\ i\neq j}}^{d} \Phi\left(\frac{\sigma_j z + \mu_j - \mu_i}{\sigma_i}\right) \phi(z)\,\mathrm{d}z \\
&\qquad\qquad \text{(by using the standardisation } z = (x_j - \mu_j)/\sigma_j\text{)} \\
&\quad = \mathrm{E}_Z\left[\prod_{\substack{i=1 \\ i\neq j}}^{d} \Phi\left(\frac{\sigma_j}{\sigma_i} Z + \frac{\mu_j - \mu_i}{\sigma_i}\right)\right]
\end{aligned}
$$

the proof follows directly.

B.2.2 Proof of Lemma B.4: Moments

Recall that for $Y \sim {}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, -\infty, b)$, for some function g of Y, we have that

$$
\mathrm{E}\, g(Y) = \Phi(\beta)^{-1} \int g(y)\, \mathbb{1}[y < b] \phi(y|\mu, \sigma^2)\,\mathrm{d}y,
$$

and in particular, we have

$$\mathrm{E}[Y-\mu] = -\sigma\frac{\phi(\beta)}{\Phi(\beta)} \tag{B.1}$$ {eq:tnXminMu}

$$\mathrm{E}[Y-\mu]^2 - \sigma^2 = -\sigma^2\frac{\beta\phi(\beta)}{\Phi(\beta)} \tag{B.2}$$ {eq:tnXminMusq}

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

$$\begin{aligned}
\mathrm{E}\, g(X_i) &= C^{-1}\int\cdots\int g(x_i)\,\mathbb{1}[x_k < x_j, \forall k \neq j]\prod_{k=1}^{d} p(x_k)\,\mathrm{d}x_1\cdots\mathrm{d}x_d \\
&= C^{-1}\frac{\Phi((x_j-\mu_j)/\sigma_j)}{\Phi((x_j-\mu_j)/\sigma_j)}\iint g(x_i)\,\mathbb{1}[x_i < x_j]p(x_i)p(x_j)\prod_{\substack{k=1\\k\neq i,j}}^{d}\Phi\left(\frac{x_j-\mu_k}{\sigma_k}\right)\mathrm{d}x_i\,\mathrm{d}x_j \\
&= C^{-1}\int \mathrm{E}_{X_i\sim{}^{\mathrm{t}}\mathrm{N}(\mu_i,\sigma_i^2,-\infty,x_j)}\, g(X_i)\prod_{\substack{k=1\\k\neq j}}^{d}\Phi\left(\frac{x_j-\mu_k}{\sigma_k}\right)p(x_j)\,\mathrm{d}x_j
\end{aligned} \tag{B.3}$$ {eq:tnproofi}

where C is the normalising constant for X, while for the j'the component we have

$$\begin{aligned}
\mathrm{E}\, g(X_j) &= C^{-1}\int\cdots\int g(x_j)\,\mathbb{1}[x_k < x_j, \forall k \neq j]\prod_{k=1}^{d} p(x_k)\,\mathrm{d}x_1\cdots\mathrm{d}x_d \\
&= C^{-1}\int g(x_j)\prod_{\substack{k=1\\k\neq j}}^{d}\Phi\left(\frac{x_j-\mu_k}{\sigma_k}\right)p(x_j)\,\mathrm{d}x_d.
\end{aligned} \tag{B.4}$$ {eq:tnproofj}

Plugging in (B.1) for $g(X_i) = X_i - \mu_i$ in (B.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 (B.4) to get

$$\begin{aligned}
\mathrm{E}\, X_j - \mu_j &= C^{-1} \int (x_j - \mu_j) \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k} \right) p(x_j)\, \mathrm{d}x_j \\
&= C^{-1} \sigma_j \int z \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{\sigma_j z + \mu_j - \mu_k}{\sigma_k} \right) \phi(z)\, \mathrm{d}z \\
&= \sigma_j \sum_{\substack{i=1 \\ i \neq j}}^{d} \sigma_i C^{-1}\, \mathrm{E} \left[\phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i} \right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_k}{\sigma_k} \right) \right] \\
&= -\sigma_j \sum_{\substack{i=1 \\ i \neq j}}^{d} \left(\mathrm{E}\, X_i - \mu_i \right),
\end{aligned}$$

where we have made use of Lemma B.5 in the second last step.

For the second moments, plug in (B.2) for $g(X_i) = (X_i - \mu_i)^2 - \sigma_i^2$ in (B.3) to get

$$\begin{aligned}
\mathrm{E}[X_i - \mu_i]^2 - \sigma_i^2 &= -\sigma_i^2 C^{-1} \int \overbrace{\frac{x_j - \mu_i}{\sigma_i}}^{x_j - \mu_i - \mu_j + \mu_j} \cdot \frac{\phi\big((x_j - \mu_i)/\sigma_i\big)}{\Phi\big((x_j - \mu_i)/\sigma_i\big)} \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j \\
&= -\sigma_i C^{-1} \int (x_j - \mu_j)\phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j \\
&\quad + (\mu_j - \mu_i) \cdot \overbrace{-\sigma_i C^{-1} \int \phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j}^{\mathrm{E}\, X_i - \mu_i} \\
&= (\mu_j - \mu_i)(\mathrm{E}\, X_i - \mu_i) \\
&\quad + \sigma_i C^{-1} \int \sigma_j z \phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j z + \mu_j - \mu_k}{\sigma_k}\right) \phi(z)\,\mathrm{d}z \\
&= (\mu_j - \mu_i)(\mathrm{E}\, X_i - \mu_i) \\
&\quad + \sigma_i \sigma_j C^{-1}\, \mathrm{E}\left[Z \phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_k}{\sigma_k}\right) \right]
\end{aligned}$$

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.

lem:EZgZ

Lemma B.5. *Let $Z \sim \mathrm{N}(0,1)$. Then for all $m \in \{\mathbb{N} \,|\, m > 1\}$ 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. □

B.2.3 Proof of Lemma B.4: Entropy

As a direct consequence of the definition of entropy,

$$\begin{aligned}
H(p) &= -\operatorname{E}\log p(X) \\
&= -\operatorname{E}\left[-\log C - \frac{d}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right] \\
&= \log C + \frac{d}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{d}\frac{1}{\sigma_i^2}\operatorname{E}[x_i-\mu_i]^2.
\end{aligned}$$

B.3 Derivation of the CAVI algorithm

Let $\mathcal{Z} = \{\mathbf{y}^*, \mathbf{w}, \boldsymbol{\alpha}, \eta, \boldsymbol{\Psi}\}$. Approximate the posterior for $\mathcal{Z}$ by a mean-field variational distribution

$$\begin{aligned}
p(\mathbf{y}^*, \mathbf{w}, \alpha, \eta, \boldsymbol{\Psi}|\mathbf{y}) &\approx q(\mathbf{y}^*)q(\mathbf{w})q(\boldsymbol{\alpha})q(\eta)q(\boldsymbol{\Psi}) \\
&= \prod_{i=1}^{n} q(\mathbf{y}_i^*)q(\mathbf{w})q(\boldsymbol{\alpha})q(\eta)q(\boldsymbol{\Psi}).
\end{aligned}$$

The first line is by assumption, while the second line follows from an induced factorisation on the latent propensities, as we will see later. If needed, we also assume that $q(\eta)$ factorises into its constituents components. Recall that, for each $\xi \in \mathcal{Z}$, the optimal mean-field variational density $\tilde{q}$ for ξ satisfies

$$\log \tilde{q}(\xi) = \mathrm{E}_{-\xi}[\log p(\mathbf{y}, \mathcal{Z})] + \text{const.} \tag{5.20}$$

Write $\mathbf{f} = \mathbf{H}_\eta \mathbf{w} \in \mathbb{R}^{n \times m}$. The joint likelihood $p(\mathbf{y}, \mathcal{Z})$ is given by

$$\begin{aligned} p(\mathbf{y}, \mathcal{Z}) &= p(\mathbf{y}|\mathcal{Z})p(\mathcal{Z}) \\ &= p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi})p(\mathbf{w}|\boldsymbol{\Psi})p(\eta)p(\boldsymbol{\Psi})p(\boldsymbol{\alpha}). \end{aligned}$$

For reference, the relevant distributions are listed below.

- $\boldsymbol{p(\mathbf{y}|\mathbf{y}^*)}$. For each observation $i \in \{1, \dots, n\}$, given the corresponding latent propensities $\mathbf{y}_i^* = (y_{i1}^*, \dots, y_{im}^*)$, 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.

$$p(\mathbf{y}|\mathbf{y}^*) = \prod_{i=1}^{n} \prod_{j=1}^{m} p_{ij}^{[y_i=j]} = \prod_{i=1}^{n} \prod_{j=1}^{m} \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*]^{\mathbb{1}[y_i=j]}.$$

- $\boldsymbol{p(\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi})}$. Given values for the parameters and I-prior random effects, the distribution of the latent propensities is matrix normal

$$\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi} \sim \mathrm{MN}_{n,m}(\mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{H}_\eta \mathbf{w}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1}).$$

  Write $\boldsymbol{\mu} = \mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{H}_\eta \mathbf{w}$. Its pdf is

$$\begin{aligned} p(\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi}) &= \exp\left[-\frac{nm}{2} \log 2\pi + \frac{n}{2} \log|\boldsymbol{\Psi}| - \frac{1}{2} \operatorname{tr}\left((\mathbf{y}^* - \boldsymbol{\mu}) \boldsymbol{\Psi} (\mathbf{y}^* - \boldsymbol{\mu})^\top \right) \right] \\ &= \exp\left[-\frac{nm}{2} \log 2\pi + \frac{n}{2} \log|\boldsymbol{\Psi}| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot})^\top \boldsymbol{\Psi} (\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot}) \right], \end{aligned}$$

  where $\mathbf{y}_i^* \in \mathbb{R}^m$ and $\boldsymbol{\mu}_i \in \mathbb{R}^m$ 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 $\mathbf{y}_i^*$ are independent multivariate normal with mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\Psi}^{-1}$.

- $\boldsymbol{p(\mathbf{w}|\Psi)}$. The $\mathbf{w}$'s are normal random matrices $\mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$ with pdf

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\Psi}) &= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\operatorname{tr}\left(\mathbf{w}\boldsymbol{\Psi}^{-1}\mathbf{w}^\top\right)\right] \\ &= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}\mathbf{w}_{i\cdot}^\top\boldsymbol{\Psi}^{-1}\mathbf{w}_{i\cdot}\right]. \end{aligned}$$

- $\boldsymbol{p(\eta)}$. The most common scenario would be $\eta = \{\lambda_1, \ldots, \lambda_p\}$ only. In this case, choose independent normal priors for each $\lambda_k \sim \mathrm{N}(m_k, v_k)$, $k = 1, \ldots, p$, whose pdf is

$$p(\eta) = \prod_{k=1}^{p} \exp\left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log v_k - \frac{1}{2v_k}(\lambda_k - m_k)^2\right].$$

  An improper prior $p(\eta) \propto$ const. can be used as well, and this is the same as letting $m_k \to 0$ and $v_k \to 0$. The resulting posterior will be proper. If η contains other parameters as well, such as the Hurst coefficient $\gamma \in (0,1)$, SE lengthscale $l > 0$ or polynomial offset $c > 0$, then appropriate priors should be used to match the support of the parameter. Choices include $p(\gamma) = \mathbb{1}\left(\gamma \in (0,1)\right)$ and $l, c \sim \Gamma(a, b)$.

- $\boldsymbol{p(\Psi)}$. Our analysis shows that regardless of prior choice of $\boldsymbol{\Psi}$, be it in the full or independent I-probit model, the posterior for $\boldsymbol{\Psi}$ will not be of a recognisable form. Without giving too much thought, assume an improper prior on $\boldsymbol{\Psi}$, i.e. $p(\boldsymbol{\Psi}) \propto$ const.

- $\boldsymbol{p(\alpha)}$. Choose independent normal priors for the intercept, $\alpha_j \sim \mathrm{N}(a_j, A_j)$ for $j = 1, \ldots, m$. The pdf is

$$p(\boldsymbol{\alpha}) = \prod_{j=1}^{m} \exp\left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log A_j - \frac{1}{2A_j}(\alpha_j - a_j)^2\right].$$

Remark B.1. The priors on the parameters $\{\boldsymbol{\alpha}, \eta\}$ can be set to very vague or even improper priors, and the resulting posterior will still yield a proper distribution. Using improper priors eases the algebra slightly. For the precision matrix $\boldsymbol{\Psi}$, it is best to stick with the Wishart prior to avoid positive-definite issues, unless the independent I-probit model is used, in which case Jeffreys' prior for the precisions $p(\sigma_j^{-2}) \propto \sigma_j^2$ is a convenient choice.

B.3.1 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 $j \in \{1, \ldots, m\}$. The mean-field density $q(\mathbf{y}_i^*)$ for each $i = 1, \ldots, n$ is found to be

$$\begin{aligned}
\log \tilde{q}(\mathbf{y}_i^*) &= \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*] \, \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{y}^*\}\sim q} \left[-\frac{1}{2} (\mathbf{y}_i^* - \boldsymbol{\mu}_i)^\top \boldsymbol{\Psi} (\mathbf{y}_i^* - \boldsymbol{\mu}_i) \right] + \text{const.} \\
&= \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*] \left[-\frac{1}{2} (\mathbf{y}_i^* - \tilde{\boldsymbol{\mu}}_i)^\top \tilde{\boldsymbol{\Psi}} (\mathbf{y}_i^* - \tilde{\boldsymbol{\mu}}_i) \right] + \text{const.} \qquad (\star) \\
&\equiv \begin{cases} \phi(\mathbf{y}_i^* | \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Psi}}) & \text{if } y_{ij}^* > y_{ik}^*, \forall k \neq j \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

where $\tilde{\boldsymbol{\mu}}_i = \mathrm{E}\,\boldsymbol{\alpha} + (\mathrm{E}\,\mathbf{H}_\eta \,\mathrm{E}\,\mathbf{w})_i$, and expectations are taken under the optimal mean-field distribution $\tilde{q}$. The distribution $q(\mathbf{y}_i^*)$ is a truncated m-variate normal distribution such that the j'th component is always largest. Unfortunately, the expectation of this distribution cannot be found in closed-form, and must be approximated by techniques such as Monte Carlo integration. If, however, the independent I-probit model is used and $\tilde{\boldsymbol{\Psi}}$ is diagonal, then Lemma B.4 provides a simplification.

Remark B.2. In $(\star)$ above, we needn't consider the second order terms in the expectations because they do not involve $\mathbf{y}^*$ and can be absorbed into the constant. To see this,

$$\begin{aligned}
\mathrm{E}[(\mathbf{y}_i^* - \boldsymbol{\mu}_i)^\top \boldsymbol{\Psi} (\mathbf{y}_i^* - \boldsymbol{\mu}_i)] &= \mathrm{E}[\mathbf{y}_i^{*\top} \boldsymbol{\Psi} \mathbf{y}_i^* + \boldsymbol{\mu}_i^\top \boldsymbol{\Psi} \boldsymbol{\mu}_i - 2\boldsymbol{\mu}_i^\top \boldsymbol{\Psi} \mathbf{y}_i^*] \\
&= \mathbf{y}_i^{*\top} \boldsymbol{\Psi} \mathbf{y}_i^* - 2\,\mathrm{E}[\boldsymbol{\mu}_i^\top]\,\mathrm{E}[\boldsymbol{\Psi}] \mathbf{y}_i^* + \text{const.} \\
&= \mathbf{y}_i^{*\top} \boldsymbol{\Psi} \mathbf{y}_i^* - 2\tilde{\boldsymbol{\mu}}_i^\top \tilde{\boldsymbol{\Psi}} \mathbf{y}_i^* + \text{const.} \\
&= (\mathbf{y}_i^* - \tilde{\boldsymbol{\mu}}_i)^\top \tilde{\boldsymbol{\Psi}} (\mathbf{y}_i^* - \tilde{\boldsymbol{\mu}}_i) + \text{const.}
\end{aligned}$$

We will see this occurring a lot later on and we shall take note of this fact.

B.3.2 Derivation of $\tilde{q}(\mathbf{w})$

The terms involving $\mathbf{w}$ in (5.20) are the $p(\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi})$ and $p(\mathbf{w}|\boldsymbol{\Psi})$ terms, and 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

$$\operatorname{vec} \mathbf{y}^* | \boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi} \sim \mathrm{N}_{nm}\left(\operatorname{vec}(\mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{H}_\eta \mathbf{w}), \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n\right)$$

and

$$\operatorname{vec} \mathbf{w} | \boldsymbol{\Psi} \sim \mathrm{N}_{nm}(\mathbf{0}, \boldsymbol{\Psi} \otimes \mathbf{I}_n)$$

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,

$$\begin{aligned}
\log \tilde{q}(\mathbf{w}) &= \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}\left[-\frac{1}{2}(\bar{\mathbf{y}}^* - \mathbf{M}\operatorname{vec}\mathbf{w})^\top (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)^{-1} (\bar{\mathbf{y}}^* - \mathbf{M}\operatorname{vec}\mathbf{w})\right] \\
&\quad + \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}\left[-\frac{1}{2}(\operatorname{vec}\mathbf{w})^\top (\boldsymbol{\Psi} \otimes \mathbf{I}_n)^{-1} \operatorname{vec}(\mathbf{w})\right] + \text{const.} \\
&= -\frac{1}{2}\mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}\left[(\operatorname{vec}\mathbf{w})^\top \Big(\overbrace{\mathbf{M}^\top(\boldsymbol{\Psi} \otimes \mathbf{I}_n)\mathbf{M} + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)}^{\mathbf{A}}\Big) \operatorname{vec}(\mathbf{w})\right] \\
&\quad + \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}\left[\overbrace{\bar{\mathbf{y}}^{*\top}(\boldsymbol{\Psi} \otimes \mathbf{I}_n)\mathbf{M}}^{\mathbf{a}^\top} \operatorname{vec}(\mathbf{w})\right] + \text{const.} \\
&= -\frac{1}{2}\mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}\left[(\operatorname{vec}\mathbf{w} - \mathbf{A}^{-1}\mathbf{a})^\top \mathbf{A} (\operatorname{vec}\mathbf{w} - \mathbf{A}^{-1}\mathbf{a})\right] + \text{const.}
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{V}_w^{-1} &= \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}[\mathbf{A}] \\
&= \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}\left[(\mathbf{I}_m \otimes \mathbf{H}_\eta)^\top (\boldsymbol{\Psi} \otimes \mathbf{I}_n)(\mathbf{I}_m \otimes \mathbf{H}_\eta) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right] \\
&= \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}\left[(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right] \\
&= (\tilde{\boldsymbol{\Psi}} \otimes \tilde{\mathbf{H}}_\eta^2) + (\tilde{\boldsymbol{\Psi}}^{-1} \otimes \mathbf{I}_n)
\end{aligned}$$

and making a first-order approximation $(\mathrm{E}\,\mathbf{A})^{-1} \approx \mathrm{E}[\mathbf{A}^{-1}]$[2],

$$\begin{aligned}
\operatorname{vec}\tilde{\mathbf{w}} &= \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}[\mathbf{A}^{-1}\mathbf{a}] \\
&= \tilde{\mathbf{V}}_w \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}\left[(\mathbf{I}_m \otimes \mathbf{H}_\eta)(\boldsymbol{\Psi} \otimes \mathbf{I}_n) \operatorname{vec}(\mathbf{y}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top)\right] \\
&= \tilde{\mathbf{V}}_w \mathrm{E}_{\mathcal{Z}\setminus\{\mathbf{w}\}\sim q}\left[(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta) \operatorname{vec}(\mathbf{y}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top)\right] \\
&= \tilde{\mathbf{V}}_w (\tilde{\boldsymbol{\Psi}} \otimes \tilde{\mathbf{H}}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \tilde{\boldsymbol{\alpha}}^\top).
\end{aligned}$$

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 5.7.2 for details.

In the case of the I-probit model, where $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \dots, \psi_m)$, then the covariance matrix takes a simpler form. Specifically, it has the block diagonal structure:

$$\begin{aligned}
\tilde{\mathbf{V}}_w &= \mathrm{E}\left[\operatorname{diag}(\psi_1, \dots, \psi_m) \otimes \mathbf{H}_\eta^2 + \operatorname{diag}(\psi_1, \dots, \psi_m) \otimes \mathbf{I}_n\right]^{-1} \\
&= \operatorname{diag}\left(\mathrm{E}\left(\psi_1 \mathbf{H}_\eta^2 + \psi_1^{-1}\mathbf{I}_n\right)^{-1}, \cdots, \mathrm{E}\left(\psi_m \mathbf{H}_\eta^2 + \psi_m^{-1}\mathbf{I}_n\right)^{-1}\right) \\
&\approx \operatorname{diag}\left(\left(\tilde{\psi}_1 \tilde{\mathbf{H}}_\eta^2 + \tilde{\psi}_1^{-1}\mathbf{I}_n\right)^{-1}, \cdots, \left(\tilde{\psi}_m \tilde{\mathbf{H}}_\eta^2 + \tilde{\psi}_m^{-1}\mathbf{I}_n\right)^{-1}\right) \\
&=: \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \dots, \tilde{\mathbf{V}}_{w_m}).
\end{aligned}$$

The mean $\operatorname{vec} \tilde{\mathbf{w}}$ is

$$\begin{aligned}
\operatorname{vec} \tilde{\mathbf{w}} &= \tilde{\mathbf{V}}_w (\operatorname{diag}(\tilde{\psi}_1, \dots, \tilde{\psi}_m) \otimes \tilde{\mathbf{H}}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \tilde{\boldsymbol{\alpha}}^\top) \\
&= \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \dots, \tilde{\mathbf{V}}_{w_m}) \operatorname{diag}(\tilde{\psi}_1 \tilde{\mathbf{H}}_\eta, \dots, \tilde{\psi}_m \tilde{\mathbf{H}}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \tilde{\boldsymbol{\alpha}}^\top) \\
&= \operatorname{diag}(\tilde{\psi}_1 \tilde{\mathbf{V}}_{w_1} \tilde{\mathbf{H}}_\eta, \dots, \tilde{\psi}_m \tilde{\mathbf{V}}_{w_m} \tilde{\mathbf{H}}_\eta)(\tilde{\mathbf{y}}^* - \mathbf{1}_n \tilde{\boldsymbol{\alpha}}^\top) \\
&\qquad\qquad \tilde{\mathbf{w}}_{\cdot 1} \qquad\quad \cdots \qquad\quad \tilde{\mathbf{w}}_{\cdot m} \\
&= \begin{pmatrix} \tilde{\psi}_1 \tilde{\mathbf{V}}_{w_1} \tilde{\mathbf{H}}_\eta (\tilde{\mathbf{y}}^*_{\cdot 1} - \tilde{\alpha}_1 \mathbf{1}_n) & \cdots & \tilde{\psi}_m \tilde{\mathbf{V}}_{w_m} \tilde{\mathbf{H}}_\eta (\tilde{\mathbf{y}}^*_{\cdot m} - \tilde{\alpha}_m \mathbf{1}_n) \end{pmatrix}^\top.
\end{aligned}$$

Therefore, we can consider the distribution of $\mathbf{w} = (\mathbf{w}_{\cdot 1}, \dots, \mathbf{w}_{\cdot m})$ columnwise, and each are normally distributed with mean and variance

$$\tilde{\mathbf{w}}_{\cdot j} = \tilde{\sigma}_j^{-2} \tilde{\mathbf{V}}_{w_j} \tilde{\mathbf{H}}_\eta (\tilde{\mathbf{y}}^*_{\cdot j} - \tilde{\alpha}_j \mathbf{1}_n) \quad \text{and} \quad \tilde{\mathbf{V}}_{w_j} = \left(\tilde{\sigma}_j^{-2} \tilde{\mathbf{H}}_\eta^2 + \tilde{\sigma}_j^2 \mathbf{I}_n\right)^{-1}.$$

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

$$\begin{aligned}
\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}]) &= \sum_{i,j=1}^{m} \mathbf{C}_{ij}\, \mathrm{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}]_{ij} \\
&= \sum_{i,j=1}^{m} \mathbf{C}_{ij}\, \mathrm{E}[\mathbf{w}_{\cdot i}^\top \mathbf{D} \mathbf{w}_{\cdot j}]. \qquad \text{(B.5)}
\end{aligned}$$

{eq:trCEwDw}

[2] Groves and Rothenberg (1969) show that $\mathrm{E}[\mathbf{A}^{-1}] = (\mathrm{E}\,\mathbf{A})^{-1} + \mathbf{B}$, where $\mathbf{B}$ is a positive-definite matrix. This approximation has been used also by Girolami and Rogers (2006) in their work.

The expectation of the univariate quantity $\mathbf{w}_{\cdot i}^\top \mathbf{D}\mathbf{w}_{\cdot j}$ is inspected below:

$$\begin{aligned}
\mathrm{E}[\mathbf{w}_{\cdot i}^\top \mathbf{D}\mathbf{w}_{\cdot j}] &= \operatorname{tr}(\mathbf{D}\,\mathrm{E}[\mathbf{w}_{\cdot j}\mathbf{w}_{\cdot i}^\top]) \\
&= \operatorname{tr}\left(\mathbf{D}(\operatorname{Cov}(\mathbf{w}_{\cdot j}, \mathbf{w}_{\cdot i}) + \mathrm{E}[\mathbf{w}_{\cdot j}]\,\mathrm{E}[\mathbf{w}_{\cdot i}]^\top)\right) \\
&= \operatorname{tr}\left(\mathbf{D}(\mathbf{V}_w[i,j] + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot i}^\top)\right).
\end{aligned}$$

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 (B.5) leads us to

$$\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]) = \sum_{i,j=1}^{m} \mathbf{C}_{ij}\left(\operatorname{tr}\left(\mathbf{D}(\delta_{ij}\mathbf{V}_{w_j} + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot i}^\top)\right).\right).$$

If $\mathbf{C} = \operatorname{diag}(c_1, \ldots, c_m)$, then

$$\begin{aligned}
\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]) &= \sum_{j=1}^{m} c_j\left(\operatorname{tr}\left(\mathbf{D}\tilde{\mathbf{V}}_{w_j}\right) + \tilde{\mathbf{w}}_{\cdot j}^\top\mathbf{D}\tilde{\mathbf{w}}_{\cdot j}\right) \\
&= \sum_{j=1}^{m} c_j \operatorname{tr}\left(\mathbf{D}(\tilde{\mathbf{V}}_{w_j} + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot j}^\top)\right)
\end{aligned}$$

B.3.3 Derivation of $\tilde{q}(\eta)$

By looking at only the terms involving η in (5.20), we deduce that $\tilde{q}$ for η satisfies

$$\begin{aligned}
\log\tilde{q}(\eta) &= -\frac{1}{2}\operatorname{tr}\mathrm{E}_{\mathcal{Z}\setminus\{\eta\}\sim q}\left[(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})^\top\right] + \log p(\eta) \\
&\quad + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\mathrm{E}_{\mathcal{Z}\setminus\{\eta\}\sim q}\left(\boldsymbol{\Psi}\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w} - 2\boldsymbol{\Psi}\mathbf{w}^\top\mathbf{H}_\eta(\mathbf{y}^* - \boldsymbol{\alpha})\right) + \log p(\eta) + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\left(\tilde{\boldsymbol{\Psi}}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}] - 2\tilde{\boldsymbol{\Psi}}\tilde{\mathbf{w}}^\top\mathbf{H}_\eta(\tilde{\mathbf{y}}^* - \tilde{\boldsymbol{\alpha}})\right) + \log p(\eta) + \text{const.}
\end{aligned}$$

with some appropriate prior $p(\eta)$. In general, this does not have a recognisable form in η, especially when it is not linearly dependent on the kernel matrix. This happens when considering parameters other than the scales of the RKHSs. Our interest would

be to obtain $\tilde{\mathbf{H}}_\eta := \mathrm{E}_{\eta\sim q}\,\mathbf{H}_\eta$ and $\tilde{\mathbf{H}}^2_\eta := \mathrm{E}_{\eta\sim q}\,\mathbf{H}^2_\eta$. We use a Metropolis random-walk algorithm to obtain these quantities, as detailed in the algorithm below.

Algorithm 7 Metropolis random-walk to sample η

1: **inputs** $\tilde{\boldsymbol{\alpha}}$, $\tilde{\mathbf{w}}$, $\tilde{\boldsymbol{\Psi}}$, and s Metropolis sampling s.d.
2: **initialise** $\eta^{(0)} \in \mathbb{R}^q$ and $t \leftarrow 0$
3: **for** $t = 1, \ldots, T$ **do**
4: Draw $\eta^* \sim \mathrm{N}_q(\eta^{(t)}, s^2)$
5: Accept/reject proposal state, i.e.

$$\eta^{(t+1)} \leftarrow \begin{cases} \eta^* & \text{if } u \sim \mathrm{Unif}(0,1) < \pi_{\text{acc}} \\ \eta^{(t)} & \text{otherwise} \end{cases}$$

where

$$\pi_{\text{acc}} = \min\left(1, \exp\left(\log \tilde{q}(\eta^*) - \log \tilde{q}(\eta^{(t)})\right)\right).$$

6: **end for**
7: $\tilde{\mathbf{H}}_\eta \leftarrow \frac{1}{T}\sum_{i=1}^T \mathbf{H}_{\eta^{(t)}}$ and $\tilde{\mathbf{H}}^2_\eta \leftarrow \frac{1}{T}\sum_{i=1}^T \mathbf{H}^2_{\eta^{(t)}}$

Now consider the case where $\eta = \{\lambda_1, \ldots, \lambda_p\}$ (RKHS scale parameters only), and the scenario described in the exponential family EM algorithm of Section 4.3.3 applies. In particular, for $k = 1, \ldots, p$, we can decompose the kernel matrix as $\mathbf{H}_\eta = \lambda_k \mathbf{R}_k + \mathbf{S}_k$ and its square as $\mathbf{H}^2_\eta = \lambda_k^2 \mathbf{R}_k^2 + \lambda_k \mathbf{U}_k + \mathbf{S}_k^2$. Then, for $j = 1, \ldots, m$, assuming each of

the $q(\lambda_k)$ densities are independent of each other, we find that

$$\begin{aligned}
\log \tilde{q}(\lambda_k) &= \mathrm{E}_{\mathcal{Z}\setminus\{\eta\}\sim q}\left[-\frac{1}{2}\operatorname{tr}\left((\mathbf{y}^* - \boldsymbol{\mu})\boldsymbol{\Psi}(\mathbf{y}^* - \boldsymbol{\mu})^\top\right)\right] - \frac{1}{2v_k^2}(\lambda_k - m_k)^2 + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\mathrm{E}_{\mathcal{Z}\setminus\{\eta\}\sim q}\left[\boldsymbol{\Psi}\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w} - 2\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}\boldsymbol{\alpha}^\top)^\top\mathbf{H}_\eta\mathbf{w}\right] \\
&\quad - \frac{1}{2v_k^2}(\lambda_k - m_k)^2 + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\mathrm{E}_{\mathcal{Z}\setminus\{\eta\}\sim q}\left[\boldsymbol{\Psi}\mathbf{w}^\top(\lambda_k^2\mathbf{R}_k^2 + \lambda_k\mathbf{U}_k)\mathbf{w} - 2\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}\boldsymbol{\alpha}^\top)^\top(\lambda_k\mathbf{R}_k)\mathbf{w}\right] \\
&\quad - \frac{1}{2v_k^2}(\lambda_k^2 - 2m_k\lambda_k) + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\mathrm{E}_{\mathcal{Z}\setminus\{\eta\}\sim q}\left[\lambda_k^2\boldsymbol{\Psi}\mathbf{w}^\top\mathbf{R}_k^2\mathbf{w} - 2\lambda_k\Big(\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}\boldsymbol{\alpha}^\top)^\top\mathbf{R}_k\mathbf{w} - \frac{1}{2}\boldsymbol{\Psi}\mathbf{w}^\top\mathbf{U}_k\mathbf{w}\Big)\right] \\
&\quad - \frac{1}{2}\left(\frac{1}{v_k^2}\lambda_k^2 - 2\frac{m_k}{v_k^2}\lambda_k\right) + \text{const.} \\
&= -\frac{1}{2}\Big[\lambda_k^2\overbrace{\left(\operatorname{tr}(\tilde{\boldsymbol{\Psi}}\,\mathrm{E}[\mathbf{w}^\top\mathbf{R}_k^2\mathbf{w}]) + v_k^{-2}\right)}^{c_k} \\
&\qquad\qquad - 2\lambda_k\overbrace{\left(\operatorname{tr}\left(\tilde{\boldsymbol{\Psi}}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top)^\top\mathbf{R}_k\tilde{\mathbf{w}} - \frac{1}{2}\tilde{\boldsymbol{\Psi}}\,\mathrm{E}[\mathbf{w}^\top\mathbf{U}_k\mathbf{w}]\right) + m_k v_k^{-2}\right)}^{d_k}\Big]
\end{aligned}$$

By completing the squares, we recognise this is as the kernel of a univariate normal density. Specifically, $\lambda_k \sim \mathrm{N}(d_k/c_k, 1/c_k)$. The quantity $\tilde{\mathbf{H}}_\eta$ can be obtained by substituting $\lambda_k \mapsto \mathrm{E}_{\lambda_k\sim q}[\lambda_k]$ in the expression $\mathbf{H}_\eta = \lambda_k\mathbf{R}_k + \mathbf{S}_k$. However, in the calculation of $\tilde{\mathbf{H}}_\eta^2$ using $\lambda_k^2\mathbf{R}_k^2 + \lambda_k\mathbf{U}_k + \mathbf{S}_k^2$, we must replace occurrences of λ_k^2 with $\mathrm{E}_{\lambda_k\sim q}[\lambda_k]^2 + \mathrm{Var}_{\lambda_k\sim q}[\lambda_k]$. This can be cumbersome, so if felt necessary, use the approximation $\lambda_k^2 \mapsto \mathrm{E}_{\lambda_k\sim q}[\lambda_k]^2$ instead.

Example B.1. Suppose $k = 1$, and we only have λ to estimate. Then, $\mathbf{H}_\eta = \lambda\mathbf{H}$, $\mathbf{R}_k = \mathbf{H}$, $\mathbf{R}_k^2 = \mathbf{H}^2$, and $\mathbf{U}_k = \mathbf{0}$. Suppose also we use an improper prior $\lambda_k \propto \text{const.}$, which is the same as having $v_k^2 \to 0$ and $m_k v_k^{-2} \to 0$. The mean field distribution for λ is then

$$\lambda \sim \mathrm{N}\left(\frac{\operatorname{tr}\left(\tilde{\boldsymbol{\Psi}}(\tilde{\mathbf{y}}^* - \mathbf{1}\tilde{\boldsymbol{\alpha}}^\top)^\top\mathbf{H}\tilde{\mathbf{w}}\right)}{\operatorname{tr}(\tilde{\boldsymbol{\Psi}}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}^2\mathbf{w}])}, \frac{1}{\operatorname{tr}(\tilde{\boldsymbol{\Psi}}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}^2\mathbf{w}])}\right)$$

Further, if $\tilde{\boldsymbol{\Psi}} = \tilde{\psi}\mathbf{I}_m$, then

$$\lambda \sim \mathrm{N}\left(\frac{\sum_{j=1}^m(\tilde{\mathbf{y}}_{\cdot j}^* - \tilde{\alpha}_j\mathbf{1})^\top\mathbf{H}\tilde{\mathbf{w}}_{\cdot j}}{\sum_{j=1}^m\operatorname{tr}\left(\mathbf{H}^2\,\mathrm{E}[\mathbf{w}_{\cdot j}\mathbf{w}_{\cdot j}^\top]\right)}, \frac{1}{\sum_{j=1}^m\operatorname{tr}\left(\mathbf{H}^2\,\mathrm{E}[\mathbf{w}_{\cdot j}\mathbf{w}_{\cdot j}^\top]\right)}\right)$$

which bears a resemblance to the exponential family EM algorithm solutions described in Chapter 4. Now, $\tilde{\mathbf{H}}_\eta = \mathrm{E}[\lambda \mathbf{H}] = \tilde{\lambda}\mathbf{H}$, and $\tilde{\mathbf{H}}_\eta^2 = \mathrm{E}[\lambda^2 \mathbf{H}^2] = (\operatorname{Var}\lambda + \tilde{\lambda}^2)\mathbf{H}^2$.

B.3.4 Derivation of $\tilde{q}(\boldsymbol{\Psi})$

We find that $q(\boldsymbol{\Psi})$ satisfies

$$\begin{aligned}
\log q(\boldsymbol{\Psi}) &= \mathrm{E}_{\mathcal{Z}\setminus\{\boldsymbol{\Psi}\}\sim q}\Big[-\frac{1}{2}\operatorname{tr}\left((\mathbf{y}^* - \boldsymbol{\mu})^\top (\mathbf{y}^* - \boldsymbol{\mu})\boldsymbol{\Psi}\right) - \frac{1}{2}\operatorname{tr}\left(\mathbf{w}^\top \mathbf{w}\boldsymbol{\Psi}^{-1}\right)\Big] \\
&\quad + \log p(\boldsymbol{\Psi}) + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\Big(\overbrace{\left(\mathrm{E}[(\mathbf{y}^* - \boldsymbol{\mu})^\top (\mathbf{y}^* - \boldsymbol{\mu})]\right)}^{\mathbf{G}_1}\boldsymbol{\Psi} + \overbrace{\mathrm{E}[\mathbf{w}^\top \mathbf{w}]}^{\mathbf{G}_2}\boldsymbol{\Psi}^{-1}\Big) \\
&\quad + \log p(\boldsymbol{\Psi}) + \text{const.}
\end{aligned}$$

This seems to be the pdf of $\mathrm{Wis}(\mathbf{G} + \mathbf{G}_1, g)$ plus the pdf of a distribution which almost resembles an inverse Wishart pdf. Unfortunately, the properties such as its moments and entropy are unknown.

The matrix $\mathbf{G}_1$ is

$$\begin{aligned}
\mathbf{G}_1 &= \mathrm{E}[(\mathbf{y}^* - \boldsymbol{\mu})^\top (\mathbf{y}^* - \boldsymbol{\mu})] \\
&= \mathrm{E}\left[\mathbf{y}^{*\top}\mathbf{y}^* + \boldsymbol{\alpha}\mathbf{1}_n^\top \mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{w}^\top \mathbf{H}_\eta^2 \mathbf{w} - 2\mathbf{y}^{*\top}\mathbf{1}_n\boldsymbol{\alpha}^\top - 2\mathbf{y}^{*\top}\mathbf{H}_\eta\mathbf{w} - 2\boldsymbol{\alpha}\mathbf{1}_n^\top\mathbf{H}_\eta\mathbf{w}\right] \\
&= \mathrm{E}\left[\mathbf{y}^{*\top}\mathbf{y}^*\right] + n\,\mathrm{E}[\boldsymbol{\alpha}\boldsymbol{\alpha}^\top] + \mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta\mathbf{w}] - 2(\tilde{\mathbf{y}}^{*\top}\mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top + \tilde{\mathbf{y}}^{*\top}\tilde{\mathbf{H}}_\eta\tilde{\mathbf{w}} + \tilde{\boldsymbol{\alpha}}\mathbf{1}_n^\top\tilde{\mathbf{H}}_\eta\tilde{\mathbf{w}}),
\end{aligned}$$

and this involves second order moments of a conically truncated multivariate normal distribution, which needs to be obtained via simulation. Meanwhile,

$$\begin{aligned}
\mathbf{G}_{2,ij} &= \mathrm{E}[\mathbf{w}^\top\mathbf{w}]_{ij} \\
&= \mathrm{E}[\mathbf{w}_{\cdot i}^\top \mathbf{w}_{\cdot j}] \\
&= \tilde{\mathbf{V}}_w[i,j] + \tilde{\mathbf{w}}_{\cdot i}^\top \tilde{\mathbf{w}}_{\cdot j}.
\end{aligned}$$

In the case of the independent I-probit model, we use a gamma prior on each of the precisions in the diagonal entries of $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \ldots, \psi_m)$. Then, the variational density

for each ψ_j is found to be

$$\begin{aligned}\log q(\psi_j) &= \mathrm{E}_{\mathcal{Z}\setminus\{\boldsymbol{\Psi}\}\sim q}\left[\frac{n}{2}\log(\psi_1\cdots\psi_m) - \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n}\psi_j(\mathbf{y}^*_{ij}-\boldsymbol{\mu}_{ij})^2\right] \\ &\quad + \mathrm{E}_{\mathcal{Z}\setminus\{\boldsymbol{\Psi}\}\sim q}\left[-\frac{n}{2}\log(\psi_1\cdots\psi_m) - \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n}\psi_j^{-1}\mathbf{w}_{ij}^2\right] \\ &\quad + \sum_{j=1}^{m}\left((s_j-1)\log\psi_j - r_j\psi_j\right) + \text{const.} \\ &= (s_j-1)\log\psi_j - \psi_j\left(\frac{1}{2}\mathrm{E}\|\mathbf{y}^*_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2 + r_j\right) \\ &\quad - \psi_j^{-1}\left(\frac{1}{2}\mathrm{E}\|\mathbf{w}_{\cdot j}\|^2\right) + \text{const.}\end{aligned}$$

which again, is a pdf of an unknown distribution. However, its posterior mode can be computed. Write $a = -\left(\frac{1}{2}\mathrm{E}\|\mathbf{y}^*_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2 + r_j\right)$, $b = s_j - 1$, and $c = \left(\frac{1}{2}\mathrm{E}\|\mathbf{w}_{\cdot j}\|^2\right)$. Then,

$$\frac{\partial}{\partial\psi_j}\log q(\psi_j) = \frac{\partial}{\partial\psi_j}\left(a\psi_j + b\log\psi_j - c\psi_j^{-1}\right) = a + b\psi_j^{-1} + c\psi_j^{-2}$$

equated to zero means solving a quadratic equation in ψ_j. Suppose that $p(\psi_j) \propto \text{const.}$, then $s_j = 1$ and $r_j = 0$ so $\tilde{\psi}_j$ can be solved directly to be

$$\hat{\psi}_j = \sqrt{\frac{\mathrm{E}\|\mathbf{y}^*_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2}{\mathrm{E}\|\mathbf{w}_{\cdot j}\|^2}}.$$

If the posterior mean is close to its mode, then $\hat{\psi}_j$ is a good approximation for $\tilde{\psi}_j$.

To calculate $\mathrm{E}\|y^*_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2 = \mathrm{E}\sum_{i=1}^{n}(\mathbf{y}^*_{ij}-\mu_{ij})^2$, one first needs $\mathrm{E}(y^*_{ij}-\alpha_j-\mathbf{w}_{\cdot j}^\top\mathbf{h}_\eta(x_i))^2$. This, in itself, presents a challenge to compute analytically, because it requires, among other things, the second moments $\mathrm{E}\,y^{*2}_{ij}$ and $\mathrm{E}[\mathbf{w}_{\cdot j}^\top\mathbf{h}_\eta(x_i)\mathbf{h}_\eta(x_i)^\top\mathbf{w}_{\cdot j}]$. Although not entirely accurate, it is simpler to use the approximation

$$\mathrm{E}\|\mathbf{y}^*_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2 \approx \|\tilde{\mathbf{y}}^*_{\cdot j}-\tilde{\boldsymbol{\mu}}_{\cdot j}\|^2.$$

(see note 2 on page 254). Also, we have $\mathbf{w}_{\cdot j} \sim \mathrm{N}_n(\tilde{\mathbf{w}}_{\cdot j}, \tilde{\mathbf{V}}_{w_j})$, and so $\mathrm{E}\|\mathbf{w}_{\cdot j}\|^2 = \mathrm{tr}(\tilde{\mathbf{V}}_{w_j} + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot j}^\top)$.

B.3.5 Derivation of $\tilde{q}(\boldsymbol{\alpha})$

Let $\mathbf{A} = \text{diag}(A_1, \ldots, A_m)$ and $\mathbf{a} = (a_1, \ldots, a_m)^\top$. The terms involving α_j in (5.20) are

$$\begin{aligned}
\log q(\boldsymbol{\alpha}) &= \mathrm{E}_{\mathcal{Z}\setminus\{\boldsymbol{\alpha}\}\sim q}\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_{i\cdot}^* - \boldsymbol{\alpha} - \mathbf{w}^\top \mathbf{h}_\eta(x_i)\right)^\top \boldsymbol{\Psi}\left(\mathbf{y}_{i\cdot}^* - \boldsymbol{\alpha} - \mathbf{w}^\top \mathbf{h}_\eta(x_i)\right)\right] \\
&\quad - \frac{1}{2}(\boldsymbol{\alpha}-\mathbf{a})^\top \mathbf{A}^{-1}(\boldsymbol{\alpha}-\mathbf{a}) + \text{const.} \\
&= -\frac{1}{2}\left[\boldsymbol{\alpha}^\top \overbrace{(n\boldsymbol{\Psi} + \mathbf{A}^{-1})}^{\tilde{\mathbf{A}}}\boldsymbol{\alpha} - 2\overbrace{\left(\sum_{i=1}^{n}\boldsymbol{\Psi}\left(\tilde{\mathbf{y}}_{i\cdot}^* - \tilde{\mathbf{w}}^\top \tilde{\mathbf{h}}_\eta(x_i)\right) + \mathbf{A}^{-1}\mathbf{a}\right)^\top}^{\tilde{\mathbf{a}}}\boldsymbol{\alpha}\right]
\end{aligned}$$

which implies a normal mean-field distribution for $\boldsymbol{\alpha}$ whose mean and variance are $\tilde{\boldsymbol{\alpha}} = \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{a}}$ and $\tilde{\mathbf{A}}^{-1}$ respectively. If $\boldsymbol{\Psi}$ is diagonal, the components of $\boldsymbol{\alpha}$ would be independent.

As a remark, due to identifiability, only $m-1$ of these intercept are estimable. We can either put a constraint that one of the intercepts is fixed at zero, or the sum of the intercepts equals zero. The latter constraint is implemented in this thesis, and this is realised by estimating all the intercepts and then centring them.

B.4 Deriving the ELBO expression

The evidence lower bound (ELBO) expression involves the following calculation:

$$\begin{aligned}
\mathcal{L} &= \int \cdots \int q(\mathbf{y}^*, \mathbf{w}, \theta) \log \frac{p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)}{q(\mathbf{y}^*, \mathbf{w}, \theta)} \,\mathrm{d}\mathbf{y}^* \,\mathrm{d}\mathbf{w} \,\mathrm{d}\theta \\
&= \mathrm{E}\log \overbrace{p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)}^{\text{joint likelihood}} + \overbrace{(-\mathrm{E}\log q(\mathbf{y}^*, \mathbf{w}, \theta))}^{\text{entropy}} \\
&= \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\log p(\cancel{y_i|y_{ij}^*}) + \sum_{i=1}^{n}\log p(\mathbf{y}_{i\cdot}^*|\boldsymbol{\alpha}, \mathbf{w}, \boldsymbol{\Psi}, \eta) + \log p(\mathbf{w}|\boldsymbol{\Psi}) + \log p(\boldsymbol{\Psi}) \right. \\
&\qquad \left. + \log p(\eta) + \log p(\boldsymbol{\alpha})\right] \\
&\quad + \sum_{i=1}^{n} H\left[q(\mathbf{y}_{i\cdot}^*)\right] + H\left[q(\mathbf{w})\right] + H\left[q(\boldsymbol{\Psi})\right] + H\left[q(\eta)\right] + H\left[q(\boldsymbol{\alpha})\right].
\end{aligned}$$

As we saw earlier, the distribution of $q(\boldsymbol{\Psi})$ is not of recognisable form. This makes computation of $\mathrm{E}\log|\boldsymbol{\Psi}|$, $\mathrm{E}\log p(\boldsymbol{\Psi})$, and $H\big[q(\boldsymbol{\Psi})\big]$, which are required in the expression of the ELBO, problematic. For simplicity, we present the ELBO calculations for when $\boldsymbol{\Psi}$ is treated to be fixed.

Remark B.3. As discussed, given the latent propensities $\mathbf{y}^*$, the pdf of $\mathbf{y}$ is degenerate and hence can be disregarded.

Remark B.4. When using improper priors for the hyperparameters, i.e. $p(\eta, \boldsymbol{\alpha}) \propto$ const., then these terms can be disregarded.

B.4.1 Terms involving distributions of $\mathbf{y}^*$

$$
\begin{aligned}
&\sum_{i=1}^{n}\Big(\mathrm{E}\log p(\mathbf{y}_{i\cdot}^*|\boldsymbol{\alpha},\mathbf{w},\boldsymbol{\Psi},\eta) + H\big[q(\mathbf{y}_{i\cdot}^*)\big]\Big) \\
&\quad= -\frac{nm}{2}\log 2\pi + \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\,\mathrm{E}\sum_{i=1}^{n}(\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot})^\top\boldsymbol{\Psi}(\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot}) \\
&\qquad+ \frac{nm}{2}\log 2\pi - \frac{n}{2}\log|\tilde{\boldsymbol{\Psi}}| + \frac{1}{2}\,\mathrm{E}\sum_{i=1}^{n}(\mathbf{y}_{i\cdot}^* - \tilde{\boldsymbol{\mu}}_{i\cdot})^\top\boldsymbol{\Psi}(\mathbf{y}_{i\cdot}^* - \tilde{\boldsymbol{\mu}}_{i\cdot}) + \log C_i \\
&\quad= \text{const.} + \sum_{i=1}^{n}\log C_i
\end{aligned}
$$

where C_i is the normalising constant for the distribution of multivariate truncated normal $\mathbf{y}_{i\cdot}$.

Notes:

1. $p(\mathbf{y}_{i\cdot}^*)$ is the pdf of $\mathrm{N}(\boldsymbol{\mu}_{i\cdot}, \boldsymbol{\Psi}^{-1})$, and $q(\mathbf{y}_{i\cdot}^*)$ is the pdf of ${}^{\mathrm{t}}\mathrm{N}(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, where $\boldsymbol{\mu}_{i\cdot} = \boldsymbol{\alpha} + \mathbf{w}^\top\mathbf{h}_\eta(x_i) \in \mathbb{R}^m$.
2. It is simpler to use the approximation

note2

$$
\mathrm{E}(\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot})^\top\boldsymbol{\Psi}(\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot}) \approx \mathrm{E}(\mathbf{y}_{i\cdot}^* - \tilde{\boldsymbol{\mu}}_{i\cdot})^\top\boldsymbol{\Psi}(\mathbf{y}_{i\cdot}^* - \tilde{\boldsymbol{\mu}}_{i\cdot}). \tag{B.6}
$$

{eq:elboyaprx}

rather than work out the actual quantity, which is

$$
\mathrm{E}(\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot})^\top\boldsymbol{\Psi}(\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot}) = \mathrm{E}(\mathbf{y}_{i\cdot}^* - \tilde{\boldsymbol{\mu}}_{i\cdot})^\top\boldsymbol{\Psi}(\mathbf{y}_{i\cdot}^* - \tilde{\boldsymbol{\mu}}_{i\cdot}) + \mathrm{tr}(\boldsymbol{\Psi}\,\mathrm{Var}\,\boldsymbol{\mu}_{i\cdot}) \tag{B.7}
$$

{eq:elboyact}

where $\operatorname{Var} \boldsymbol{\mu}_{i\cdot} = \operatorname{Var} \boldsymbol{\alpha} + \operatorname{Var} \mathbf{w}^\top \mathbf{h}_\eta(x_i)$, obtained by taking expectations with respect to everything except $\mathbf{y}^*_{i\cdot}$. The first term is a diagonal matrix of the posterior variances of the intercepts. The second term is where things get complicated. Let $\boldsymbol{\Omega}_i = \operatorname{Var} \mathbf{w}^\top \mathbf{h}_\eta(x_i)$. Then $\boldsymbol{\Omega}_{i,kj} \approx \operatorname{Cov}(\mathbf{w}_{\cdot k}^\top \mathbf{h}_\eta(x_i), \mathbf{w}_{\cdot j}^\top \mathbf{h}_\eta(x_i)) = \mathbf{h}_\eta(x_i)^\top \tilde{\mathbf{V}}_w[k,j] \mathbf{h}_\eta(x_i)$. So

$$\operatorname{tr}(\boldsymbol{\Psi}\boldsymbol{\Omega}_i) \approx \sum_{k,j=1}^{m} \boldsymbol{\Psi}_{kj} \mathbf{h}_\eta(x_i)^\top \tilde{\mathbf{V}}_w[k,j] \mathbf{h}_\eta(x_i)$$

However, we know that $\operatorname{Var} XY = \operatorname{E} X^2Y^2 - (\operatorname{E} XY)^2 = \operatorname{Var} X \operatorname{Var} Y + \operatorname{Var} X (\operatorname{E} Y)^2 + \operatorname{Var} Y (\operatorname{E} X)^2$, so there is actually some covariance terms which need to be considered, and these are not so easily computed. In practice, we find that using (B.6) gives satisfactory results as far as determining convergence for the variational algorithm goes.

B.4.2 Terms involving distributions of w

$$\begin{aligned}
\operatorname{E} \log p(\mathbf{w}|\boldsymbol{\Psi}) + H\big[q(\mathbf{w})\big] &= -\frac{nm}{2} \cancel{\log 2\pi} - \frac{n}{2} \log|\boldsymbol{\Psi}| - \frac{1}{2} \operatorname{E} \operatorname{tr}\left(\mathbf{w}\boldsymbol{\Psi}^{-1}\mathbf{w}^\top\right) \\
&\quad + \frac{nm}{2}(1 + \cancel{\log 2\pi}) + \frac{1}{2} \log|\tilde{\mathbf{V}}_w| \\
&= \text{const.} - \frac{1}{2} \sum_{j=1}^{m} \operatorname{tr}\left(\boldsymbol{\Psi}^{-1}(\tilde{\mathbf{V}}_w[j,j] + \tilde{\mathbf{w}}_{\cdot j} \tilde{\mathbf{w}}_{\cdot j}^\top)\right)
\end{aligned}$$

Notes:

1. $p(\mathbf{w})$ is the pdf of $\mathrm{MN}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$, and $q(\mathbf{w})$ is the pdf of $\mathrm{N}(\operatorname{vec} \tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$.
2. $\tilde{\mathbf{V}}_w[j,j]$ are the $n \times n$ sub matrices along the diagonal of $\tilde{\mathbf{V}}_w$.

B.4.3 Terms involving distributions of η

If no closed-form expression for $q(\eta)$ is found, then the expression $\mathrm{E}[\log p(\eta) - q(\eta)]$ must be obtained by sampling methods. Otherwise, consider the case where $\eta = \{\lambda_1, \ldots, \lambda_p\}$.

Then, the contribution to the ELBO is

$$\begin{aligned}
&\mathrm{E}\log p(\lambda_1,\ldots,\lambda_p) + H\big[q(\lambda_1,\ldots,\lambda_p)\big] \\
&\qquad = -\frac{p}{2}\log 2\pi - \frac{1}{2}\log v_1\cdots v_k - \frac{1}{2}\sum_{k=1}^{p}\frac{\mathrm{E}(\lambda_k - m_k)^2}{v_k} \\
&\qquad\quad + \frac{p}{2}(1+\log 2\pi) + \frac{1}{2}\log\tilde{v}_1\cdots\tilde{v}_p \\
&\qquad = \text{const.} + \frac{1}{2}\sum_{k=1}^{p}\log\tilde{v}_k - \frac{1}{2}\sum_{k=1}^{p}\frac{\tilde{v}_k + \tilde{\lambda}_k^2 - 2\tilde{\lambda}_k m_k}{v_k}
\end{aligned}$$

Notes:

1. The priors on the λ_k's are $\mathrm{N}(m_k, v_k)$, and $q(\lambda_k)$ is the density of $\mathrm{N}(\tilde{\lambda}_k, v_{\lambda_k})$.
2. When using improper priors $\lambda_k \propto$ const., then we need only consider the middle term involving the sums of $\log\tilde{v}_{\lambda_k}$.

B.4.4 Terms involving distribution of $\boldsymbol{\alpha}$

For the intercepts, consider only

$$\begin{aligned}
\mathrm{E}\log p(\boldsymbol{\alpha}) + H\big[q(\boldsymbol{\alpha})\big] &= \text{const.} - \frac{1}{2}\mathrm{E}\sum_{j=1}^{m}\frac{(\alpha_j - a_j)^2}{A_j} + \frac{1}{2}\log\tilde{v}_{\alpha_1}\cdots\tilde{v}_{\alpha_m} \\
&= \text{const.} + \frac{1}{2}\sum_{j=1}^{m}\log\tilde{v}_{\alpha_j} - \frac{1}{2}\sum_{j=1}^{m}\frac{v_{\alpha_j} + \tilde{\alpha}_j^2 - 2a_j\tilde{\alpha}_j}{A_j}
\end{aligned}$$

Notes:

1. $p(\boldsymbol{\alpha})$ is $\prod_{j=1}^{m}\phi(\alpha_j|a_j, A_j)$, and $q(\boldsymbol{\alpha})$ $\prod_{j=1}^{m}\phi(\alpha_j|\tilde{\alpha}_j, \tilde{v}_{\alpha_j})$.

B.4.5 ELBO summarised

In the example section of Chapter 5, we considered only 1) the independent I-probit model; 2) fixed $\boldsymbol{\Sigma} = \mathbf{I}_m$; 3) only RKHS scale parameters to estimate; and 4) and improper priors on the hyperparameters. In such situations, the ELBO expression is

simply

$$\mathcal{L} = \text{const.} + \sum_{i=1}^{n} \log C_i - \frac{1}{2} \sum_{j=1}^{m} \text{tr}\left(\tilde{\mathbf{V}}_{w_j} + \tilde{\mathbf{w}}_{\cdot j} \tilde{\mathbf{w}}_{\cdot j}^{\top}\right) + \frac{1}{2} \sum_{k=1}^{p} \log \tilde{v}_k.$$

As a final remark, often times the ELBO is treated as a proxy for the (penalised) marginal likelihood of the model, in which case it must be noted that the ELBO as we had derived is correct up to a constant. We find that keeping track of the constants is slightly tedious, and hence decided not to do so. When comparing ELBOs of two or more models, the comparison is still valid as only differences between the ELBOs matter, in which case the constants would cancel out.

Index