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

Haziq Jamil

Department of Statistics

London School of Economics and Political Science

February 9, 2018

Abstract

Keywords: some, keywords, go, here

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).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of XX,XXX words.

I confirm that Chapter X is jointly co-authored with Wicher Bergsma.

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, and we dedicate a full chapter (Chapter 2) 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 3 describes computational methods relating to the estimation of I-prior models. Chapter 4 extends the I-prior methodology to fit discrete outcome models. Chapter 5 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.

1.1 Regression models

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

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

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

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

where $\mathbf{\Psi} = (\psi_{ij})_{i,j=1}^n$ 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 $f(x_i) = x_i\beta$, $\beta \in \mathbb{R}^p$, 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 $f_{12}(x_{ij}, j) = x_i\beta_j$, with $\beta_j \in \mathbb{R}^p$, and f_1 representing the fixed-effects linear component $x_i\beta$ 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

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

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 $\text{ER}[f]$ 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 $\text{R}[f] = \int \Lambda(y, f(x)) \, \text{dP}(y, x)$, 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

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

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

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

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

Tikhonov regularisation also has a well-known Bayesian interpretation, whereby the regularisation term encodes prior information about the function f. For the normal regression model with $f \in \mathcal{F}$, an RKHS, it can be shown that f_{reg} is the posterior mean of f given a *Gaussian process prior* with mean f_0 and covariance kernel $\operatorname{Cov}\big(f(x_i), f(x_j)\big) = \lambda h(x_i, x_j)$. The exact solution for the coefficients $\mathbf{w} = (w_1, \ldots, w_n)^\top$ 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 $(\mathbf{y} - \mathbf{f}_0) = (y_1 - f_0(x_1), \ldots, y_n - f_0(x_n))^\top$.

1.4 Regression using I-priors

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

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

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

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

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 $\mathbf{y} = (y_1, \dots, y_n)$,

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

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

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

where we have defined $\mathbf{h}_\eta(x)$ to be the vector of length n with entries $h_\eta(x, x_i)$ for $i = 1, \dots, n$. 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

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

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

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

where we had defined $f_{\text{new}} := f(x_{\text{new}})$. This is again a normal distribution in the case of the normal model, with 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 $\mathbf{\Psi}$. 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 $\mathcal{F}$.

3. 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. Deviating from the normality assumption would require approximation techniques to be implemented in order to obtain the posterior distributions of interest.

1.6 Outline of thesis

This thesis is structured as follows:

- Following this introductory chapter, **Chapter 2** gives a brief overview of functional analysis. This allows us to better explain and derive the I-prior for the normal regression model. Note that the reader does not require any in-depth knowledge of RKHSs nor functional analysis to perform I-prior modelling.

- The aforementioned computational methods relating to the estimation of I-prior models are explored in **Chapter 3**, 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). This chapter has also been submitted for publication to Computational Statistics and Data Analysis.

- 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 certainly a more proper way to account for non-normality of errors. **Chapter 4** 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 5** attempts to contribute to the field of model 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 3–5 contain computer implementations of the statistical methodologies described therein, and the code for replication are made available at `http://myphdcode.`

haziqj.ml.

Chapter 2

Reproducing kernel Hilbert and Krein spaces

This chapter provides a concise review of functional analysis, especially on topic of reproducing kernel Hilbert and Krein spaces. In addition, this chapter also describes several RKHSs of interest for the purpose of I-prior modelling. Choosing the appropriate RKHS allows us to fit various models of interest. In I-prior modelling, the kernel defining the RKHS turn out to be negative. In such a case, it is necessary to consider *Krein spaces*, in order to give us the required mathematical platform for I-prior modelling. Krein spaces are simply a generalisation of Hilbert spaces for which the kernels allowed to be non-positive definite it its reproducing kernel space. It is emphasised that a deep knowledge of functional analysis is not necessary for I-prior modelling; the advanced reader may wish to skip Sections 2.1–2.3. Section 2.4 describes the RKHSs and RKKSs of interest.

2.1 Preliminaries

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 it is known, are sets for which its elements adhere to a set of rules (axioms) relating to additivity and multiplication by a constant. Additionally, vector spaces are endowed with some kind of structure so as to allow ideas such as closeness and limits to be conceived. Of particular

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

interest to us is the structure brought about by *inner products*, which allow the rigorous mathematical study of various geometrical concepts such as lengths, directions, and orthogonality, among other things. We begin with the definition of an inner product.

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

- **Symmetry:** $\langle f, g\rangle_{\mathcal{F}} = \langle g, f\rangle_{\mathcal{F}}$, $\forall f, g \in \mathcal{F}$
- **Linearity:** $\langle af_1 + bf_2, g\rangle_{\mathcal{F}} = a\langle f_1, g\rangle_{\mathcal{F}} + b\langle f_2, g\rangle_{\mathcal{F}}$, $\forall f_1, f_2, g \in \mathcal{F}$ and $\forall a, b \in \mathbb{R}$
- **Non-degeneracy:** $\langle f, f\rangle_{\mathcal{F}} = 0 \Leftrightarrow f = 0$
- **Positive-definiteness:** $\langle f, f\rangle_{\mathcal{F}} \geq 0$, $\forall f \in \mathcal{F}$

We can always define a *norm* on $\mathcal{F}$ using the inner product as $\|f\|_{\mathcal{F}} = \sqrt{\langle f, f\rangle_{\mathcal{F}}}$. Norms are another form of structure that specifically describes the notion of length. This is defined below.

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

- **Absolute homogeneity:** $||\lambda f||_{\mathcal{F}} = |\lambda| \, ||f||_{\mathcal{F}}$, $\forall \lambda \in \mathbb{R}$, $\forall f \in \mathcal{F}$
- **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 norm $|| \cdot ||_{\mathcal{F}}$ induces a metric (a notion of distance) on $\mathcal{F}$: $d(f, g) = ||f - g||_{\mathcal{F}}$. 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 $\|f\|_{\mathcal{F}} + \|-f\|_{\mathcal{F}} \geq \|f - f\|_{\mathcal{F}} = \|0\|_{\mathcal{F}} = 0$, which implies non-negativity of norms.

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, but not always the other way around. With these notions of distances we can then define *Cauchy sequences.* A sequence is said to be Cauchy if the elements of the sequence become arbitrarily close to one another as the sequence progresses.

Definition 2.3 (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 $\forall n, m > N$, $||f_n - f_m||_{\mathcal{F}} < \epsilon$.

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*.

A vector space equipped with a (positive definite) inner product that is also complete is known as a *Hilbert space*. Out of interest, an incomplete inner product space is known as a *pre-Hilbert space*, since its completion with respect to the norm induced by the inner product is a Hilbert space. A complete normed space is called a *Banach space*.

The next few definitions are introduced as a necessary precursor to defining a reproducing kernel Hilbert space. Firstly,

For a space of functions $\mathcal{F}$ on $\mathcal{X}$, we define the evaluation functional that assigns a value to $f \in \mathcal{F}$ for each $x \in \mathcal{X}$.

Definition 2.4 (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 function $\delta_x : \mathcal{F} \to \mathbb{R}$ as defined by $\delta_x(f) = f(x)$ is called the (Dirac) evaluation functional at x. Evaluation functionals are always linear.

There are two more concepts that we need to cover before defining a reproducing kernel Hilbert/Krein space.

Definition 2.5 (Linear operator)**.** A function $A : \mathcal{F} \to \mathcal{G}$, where $\mathcal{F}$ and $\mathcal{G}$ are both normed vector spaces over $\mathbb{R}$, is called a linear operator if and only if it satisfies the following properties:

- **Homogeneity**: $A(af) = aA(f)$, $\forall a \in \mathbb{R}$, $\forall f \in \mathcal{F}$
- **Additivity**: $A(f + g) = A(f) + A(g)$, $\forall f \in \mathcal{F}, g \in \mathcal{G}$.

Definition 2.6 (Bounder operator)**.** The linear operator $A : \mathcal{F} \to \mathcal{G}$ between two normed spaces $(\mathcal{F}, ||\cdot||_\mathcal{F})$ and $(\mathcal{G}, ||\cdot||_\mathcal{G})$ is said to be a bounded operator if $\exists \lambda \in [0, \infty)$ such that

$$||A(f)||_\mathcal{G} < \lambda ||f||_\mathcal{F}.$$

Now we define a reproducing kernel Hilbert space.

Definition 2.7 (Reproducing kernel Hilbert space)**.** A Hilbert space 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 bounded (equivalently, continuous[2]), i.e. $\exists \lambda_x \geq 0$

such that $\forall f \in \mathcal{F}$,

$$|f(x)| = |\delta_x(f)| \leq \lambda_x ||f||_{\mathcal{F}}.$$

Theorem 2.1 (Representation theorem)**.** *Every continuous linear functional f on a Hilbert space $\mathcal{H}$ has the form*

$$f(x) = \langle x, y \rangle$$

with a unique $y \in \mathcal{M}$ and $\|f\| = \|y\|_{\mathcal{H}}$.

Theorem 2.2 (Orthogonal decomposition)**.** *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{M} \subset \mathcal{H}$ be a closed subspace. For every $x \in \mathcal{H}$, we can write*

$$x = y + z$$

where $y \in \mathcal{M}$ and $z \in \mathcal{M}^{\perp}$, and y and z are uniquely determined by x.

Corollary 2.2.1. *Let $\mathcal{M}$ be a subspace of a Hilbert space $\mathcal{H}$. Then, $\mathcal{M}^{\perp} = \{0\}$ if and only if $\mathcal{M}$ is dense in $\mathcal{H}$.*

`https://en.wikibooks.org/wiki/Functional_Analysis/Hilbert_spaces`

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 $\mathcal{H}$ is the closure of $\mathcal{H}$, i.e. $(\mathcal{H}^{\perp})^{\perp} = \overline{\mathcal{H}}$. If $\mathcal{M}$ is a closed linear subspace of $\mathcal{H}$, then $\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^{\perp}$.

2.2 Reproducing kernel Hilbert spaces

2.3 Reproducing kernel Krein spaces

2.4 RKHS building blocks

In what follows, each of the kernel functions will have its associated scale parameter denoted by λ. Further, to make the distinction between centred and non-centred versions

[2] For any two function $f, g \in \mathcal{F}$, $|f(x) - g(x)| = |\delta_x(f) - \delta_x(g)| = |\delta_x(f - g)| \leq \lambda_x ||f - g||_{\mathcal{F}}$ for some $\lambda_x \geq 0$, thus is said to be Lipschitz continuous, which implies uniform continuity. This property implies pointwise convergence from norm convergence in $\mathcal{F}$.

of the kernels, we use the notation h to denote the uncentred version, and $\bar{h}$ to denote the centred version.

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

In I-prior modelling, one need not consider any scale parameter on reproducing kernel, as the scale parameter would not be identified otherwise. See later chapter for details. I think the scale parameter λ would just be absorbed by the norm, which is a single value of interest and that is what is "observed", and the decomposition $\lambda \cdot c_f$ is not so interesting.

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 $f_\beta = \langle \cdot, \beta \rangle_\mathcal{X}$. We can show that the continuous dual space of $\mathcal{X}$ is a RKHS which consists of these linear functions.

Proposition 2.4 (The canonical RKHS)**.** *The continuous dual space a Hilbert space $\mathcal{X}$, denoted by $\mathcal{X}'$, is a RKHS of linear functions over $\mathcal{X}$ of the form $\langle \cdot, \beta \rangle_{\mathcal{X}}$, $\beta \in \mathcal{X}$. Its reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by*

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

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

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

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

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

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

In many other literature, the kernel $h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$ is also known as the *linear kernel*. The use of the term 'canonical' is fitting not just due to the relation between a Hilbert space and its continuous dual space. Let $\phi : \mathcal{X} \to \mathcal{V}$ be the feature map from the space of covariates (inputs) to some feature space $\mathcal{V}$. Suppose both $\mathcal{X}$ and $\mathcal{V}$ is a Hilbert space, 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 2.8 (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 (i.e. $\operatorname{E}\langle x, x' \rangle_{\mathcal{X}} = \langle \mu, x' \rangle_{\mathcal{X}}$ for all $x' \in \mathcal{X}$) with respect to this probability measure. 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 $\operatorname{E} f_\beta(x) = 0$. The reproducing kernel of $(\mathcal{X} - \mu)'$ is

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

Proof. Proof of the claim $\operatorname{E} f_\beta(x) = 0$:

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

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

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

2.4.3 The fractional Brownian motion RKHS

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 $\operatorname{E} B_\gamma(t) = 0$ for all $t \in [0, T]$ and

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

where $\gamma \in (0, 1)$ is called the Hurst index or Hurst parameter. 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 Brownian motion and its increments are independent; when $\gamma > 1/2$ ($\gamma < 1/2$) its increments are positively (negatively) correlated.

Let $\mathcal{X}$ be a Hilbert space. Defining a kernel function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ identical to the fBm covariance kernel yields the so-called *fractional Brownian motion RKHS*.

Definition 2.9 (Fractional Brownian motion (fBm) RKHS)**.** The fractional Brownian motion 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

$$h(x, x') = \frac{1}{2}\big(\|x\|_{\mathcal{X}}^{2\gamma} + \|x'\|_{\mathcal{X}}^{2\gamma} + \|x - x'\|_{\mathcal{X}}^{2\gamma}\big),$$

which depends on the Hurst coefficient $\gamma \in (0, 1)$.

It is clear that the fBm kernel is positive definite by construction, and thus induces 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).

2.4.4 The squared exponential RKHS

Are SE smoother than fBm? Mainly used because of the "universal approximation" property of SE kernels.

2.4.5 The Pearson RKHS

2.5 Constructing RKKS from existing RKHS

2.5.1 Scale of an RKHS

2.5.2 The polynomial RKHS

2.5.3 The ANOVA RKKS

2.6 The Sobolev-Hilbert inner product

2.7 Discussion

Chapter 3

Fisher information and the I-prior

The main aim of this chapter is to derive the I-prior for the normal regression model stated earlier in (1.1).

In this section, we derive the Fisher information for the regression function f in the model stated in (1.1) subject to (1.2). Traditionally, Fisher information are calculated for unknown parameters θ of probability distribution from observable random variables. In a similar light, we can treat the regression function f as the unknown quantity for which we would like information to be measured from the random variables for which f is assumed to model.

3.1 The traditional Fisher information

It was Fisher (1922) who introduced the method of maximum likelihood as an objective way of obtaining parameter estimates of a statistical model. The purpose of maximum likelihood estimation extended to include this view of capturing uncertainty about parameter estimates, especially through the likelihood function as a whole and also a derivative of it known as the Fisher information.

Suppose Y is a random variable whose density function $p(\cdot|\theta)$ depends on the parameter 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*) as $S(\theta) = \partial L(\theta)/\partial\theta$. The *Fisher information* about a 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).

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[1] of a function, measures how quickly the function changes with changes in its input values. 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.

Initially, I wrote about observed vs expected Fisher information and total vs unit Fisher information, but realised it does not contribute to the later discussion. We need the true and total Fisher information.

3.2 Fisher information for Hilbert space objects

We extend the idea beyond thinking about parameters as merely numbers, to abstract objects in Hilbert spaces. This generalisation allows us to extend the concept of Fisher information to regression functions in RKHSs later.

Let Y be a random variable with density in the parametric family $\{p(\cdot|\theta) \,|\, \theta \in \Theta\}$, where Θ is 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 denoted $L(\theta) = \log p(Y|\theta)$. The score and Fisher

7. Why wouldn't it be >0 ?

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

information is derived in a familiar manner, but a extra care is required when taking derivatives with respect to Hilbert space objects. In particular, we require *directional derivatives* and *gradients* concerning inner product space objects.

Definition 3.1 (Directional derivative and gradient)**.** Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be an inner product space, and consider a function $g : \mathcal{H} \to \mathbb{R}$. Denote the directional derivate of g in the direction z by $\nabla_z g$, that is,

$$\nabla_z g(x) = \lim_{\delta\to 0} \frac{g(x+\delta z) - g(x)}{\delta}.$$

The gradient of g, denoted by ∇g, is the unique vector field satisfying

$$\langle \nabla g(x), z\rangle_{\mathcal{H}} = \nabla_z g(x), \quad \forall x, z \in \mathcal{H}.$$

Definition 3.2. Let $\mathcal{X}$ and $\mathcal{Y}$ be two Hilbert spaces. A functional ϕ is a map from $\mathcal{X}$ to $\mathbb{R}$, and we denote its action on a function f as $\phi(f)$. An operator F is a map from $\mathcal{X}$ to $\mathcal{Y}$, and we denote its action on a function f as Ff. We say that a functional ϕ is Fréchet differentiable at $f \in \mathcal{X}$ when there exists a linear functional $A : \mathcal{X} \to \mathbb{R}$ such that

$$\lim_{h\to 0} \frac{\left|\phi(f+h) - \phi(f) - A(h)\right|}{\|h\|_{\mathcal{X}}} = 0$$

If this relation holds, we say that A is the functional derivative, or Fréchet derivative, of ϕ at f, and we denote it as

$$A = \frac{\partial \phi}{\partial f}[f].$$

The differential ratio formula $\partial\phi/\partial f$ is called the Gâteaux derivative

$$\frac{\partial \phi}{\partial f}[f](h) = A(h) = \lim_{t\to 0} \frac{\phi(f+th) - \phi(f)}{t}$$

which corresponds to the idea of directational derivatives.

So $\frac{\partial\phi}{\partial f}[f](h) \equiv A(h) \equiv \nabla_h \phi(f)$ and the gradient $\nabla\phi$ satisfies

$$\langle \nabla\phi(f), h\rangle = \nabla_h \phi(f) = A(h) = \langle A, h\rangle$$

thus $\nabla\phi(f) = A$.

We can now define the score, assuming existence, as the gradient of $L(\theta)$, i.e. $S(\theta) = \nabla L(\theta)$. The Fisher information $\mathcal{I}(\theta) \in \mathcal{H} \otimes \mathcal{H}$ for $\theta \in \Theta$ is

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

or equivalently,

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

where again, stated for clarity, expectations are taken with respect to the random variable Y under the true distribution $p(\cdot|\theta)$. In the above definitions, ∇^2 is the second-order gradient, and the operation $\otimes : \mathcal{H} \times \mathcal{H} \to \mathcal{H} \otimes \mathcal{H}$ is the tensor product, mapping elements from $\mathcal{H}^2$ to the tensor product space $\mathcal{H} \otimes \mathcal{H}$.

Taking this concept further, we can also define the Fisher information for a linear functional of θ, or between two linear functionals of θ. This is essence of the next lemma.

Lemma 3.1 (Fisher information for linear functionals)**.** *Following the above definitions, suppose that the Fisher information for $\theta \in \Theta$ is $\mathcal{I}(\theta)$, with Θ a Hilbert space with inner product $\langle \cdot, \cdot \rangle_\Theta$. For some $b \in \Theta$, denote $\theta_b = \langle \theta, b \rangle_\Theta$. Then, the Fisher information for θ_b is given as*

$$\mathcal{I}(\theta_b) = \langle \mathcal{I}(\theta), b \otimes b \rangle_{\Theta \otimes \Theta},$$

and, more generally, the Fisher information between θ_b and $\theta_{b'}$ is given as

$$\mathcal{I}(\theta_b, \theta_{b'}) = \langle \mathcal{I}(\theta), b \otimes b' \rangle_{\Theta \otimes \Theta}$$

Proof. Let $\mathcal{B}$ be a set containing an orthonormal sequence of points in Θ, i.e. $\mathcal{B}$ is a Hilbert basis for the Hilbert space Θ. Then, by definition, every $\theta \in \Theta$ can be written as

$$\theta = \sum_{\beta \in \mathcal{B}} \langle \theta, \beta \rangle_\Theta \beta.$$

Now, the score function with respect to the linear functional $\theta_b = \langle \theta, b \rangle_\Theta$ is

$$\begin{aligned} \frac{\partial}{\partial \theta_b} L(\theta) &= \dots \\ &= \nabla_b L(\theta) \\ &= \langle \nabla L(\theta), b \rangle_\Theta \end{aligned}$$

Differentiating again gives

$$\frac{\partial^2}{\partial\theta_b \partial\theta_{b'}} L(\theta) = \langle \nabla L^2(\theta), b \otimes b' \rangle_{\Theta\otimes\Theta}$$

Note that by the bilinear property of tensor products,

$$\begin{aligned} -\frac{\partial^2}{\partial\theta_b \partial\theta_{b'}} L(\theta) &= (-1) \cdot \langle \nabla L^2(\theta), b \otimes b' \rangle_{\Theta\otimes\Theta} \\ &= \langle -\nabla L^2(\theta), b \otimes b' \rangle_{\Theta\otimes\Theta}. \end{aligned}$$

Provided $\mathrm{E}\|\nabla L^2(\theta)\|_{\Theta\otimes\Theta} < \infty$, taking expectations of both sides gives the desired result, since $b \otimes b'$ is free of Y and is therefore constant under the expectation. Not really convinced of this proof. □

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 $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$, for $i = 1, \dots, n$ is

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

subject to

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

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

Lemma 3.2 (Fisher information for regression function)**.** *For the regression model stated in* (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 $\mathbf{\Psi}$ 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 directional derivative of $k_x(v)$ in the direction u is

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

Thus, the gradient is $\nabla k_x(f) = \phi(x)$ by definition. Let $\mathbf{y} = \{y_1, \dots, y_n\}$, and denote the hyperparameters of the regression model by $\boldsymbol{\theta} = \{\alpha, \boldsymbol{\Psi}, \eta\}$. The log-likelihood of v is given by

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

and the score by

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

Differentiating again gives

$$\nabla^2 L(v|\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} \nabla k_{x_i}(v) \otimes \nabla k_{x_j}(v).$$

We can then calculate the Fisher information to be

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

where we had made the substitution $\nabla k_x(v) = \phi(x)$. By taking the canonical feature $\phi(x) = h(\cdot, x)$, the formula for $\mathcal{I}(f)$ follows. □

The above lemma gives the form of the Fisher information for f in a rather abstract fashion. Consider the following example of applying Lemma (3.2) 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 its assumptions hold. For simplicity, we assume iid errors, i.e. $\mathbf{\Psi} = \psi \mathbf{I}_n$. 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 feature map $\phi : \mathcal{X} \to \mathcal{V}$ defined by $\phi(x) = x$. For some $\beta \in \mathcal{V}$, the linear regression model is such that $f(x) = x^\top \beta = \langle \phi(x), \beta \rangle_{\mathcal{V}}$. Therefore, according to Lemma (3.2), the Fisher information for β is

$$\begin{aligned}
\mathcal{I}(\beta) &= \sum_{i=1}^n \sum_{j=1}^n \psi \phi(x_i) \otimes \phi(x_j) \\
&= \psi \sum_{i=1}^n \sum_{j=1}^n x_i \otimes x_j \\
&= \psi \mathbf{X}^\top \mathbf{X},
\end{aligned}$$

where $\mathbf{X}$ is a $n \times p$ matrix containing the entries $x_1^\top, \dots, x_n^\top$ row-wise. This is of course recognised as the Fisher information for the regression coefficients in the standard linear regression model.

Lemma 3.1 enables us to also compute the Fisher information for a linear functionals of f, and in particular for point evaluation functionals of f, thereby allowing us to compute the Fisher information between two points $f(x)$ and $f(x')$.

Corollary 3.2.1 (Fisher information between two linear functionals of the regression function)**.** *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 between 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 Lemma 3.1, we have that

$$\begin{aligned}
\mathcal{I}\big(f(x), f(x')\big) &= \mathcal{I}\big(\langle f, h(\cdot,x)\rangle_{\mathcal{F}}, \langle f, h(\cdot,x')\rangle_{\mathcal{F}}\big) \\
&= \big\langle \mathcal{I}(f), h(\cdot,x) \otimes h(\cdot,x') \big\rangle_{\mathcal{F}\otimes\mathcal{F}} \\
&= \left\langle \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij} h(\cdot,x_i) \otimes h(\cdot,x_j) \ , \ h(\cdot,x) \otimes h(\cdot,x') \right\rangle_{\mathcal{F}\otimes\mathcal{F}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij} \big\langle h(\cdot,x_i), h(\cdot,x) \big\rangle_{\mathcal{F}} \big\langle h(\cdot,x_j), h(\cdot,x') \big\rangle_{\mathcal{F}} \\
& \text{(by using the fact that inner products are linear, and that } \forall a_1, a_2 \in \mathcal{A} \\
& \text{and } \forall b_1, b_2 \in \mathcal{B},\ \langle a_1 \otimes b_1, a_2 \otimes b_2 \rangle_{\mathcal{A}\otimes\mathcal{B}} = \langle a_1, a_2 \rangle_{\mathcal{A}} \langle b_1, b_2 \rangle_{\mathcal{B}}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij} h(x,x_i) h(x',x_j). \quad \text{(by the reproducing property)}
\end{aligned}$$

□

An inspection of the formula in Corollary (3.2.1) reveals the fact that the Fisher information for $f(x)$ is positive if and only if $h(x,x_i) \neq 0$ for at least one $i \in \{1,\dots,n\}$. 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.

9.
Rewrite

3.4 The induced Fisher information RKHS

Next, let us see for which linear functionals of f there is Fisher information. Let

10.
Rewrite

$$\mathcal{F}_n = \left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, f(x) = \sum_{i=1}^{n} h(x,x_i) w_i, \ w_i \in \mathbb{R}, \ i = 1,\dots,n \right\}. \tag{3.1}$$

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 $\mathcal{F}$. Let $\mathcal{F}_n^{\perp}$ be the orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$. Then, any $r \in \mathcal{F}_n^{\perp}$ is orthogonal to each of the $h(\cdot,x_i)$, so by the reproducing property of h, $r(x_i) = \langle r, h(\cdot,x_i)\rangle_{\mathcal{F}} = 0$.

Corollary 3.2.2. *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* $g(x_1) = \cdots = g(x_n) = 0$.

Hence, r cannot be estimated from the data and has to be estimated by a prior guess.

[OLD], but some stuff relevant here., but some stuff relevant here.] Note that any regression function $f \in \mathcal{F}$ can be decomposed into $f = f_n + r$, with $f \in \mathcal{F}_n$ and $r \in \mathcal{R}$ where $\mathcal{F} = \mathcal{F}_n + \mathcal{R}$ and $\mathcal{F}_n \perp \mathcal{R}$. Fisher information exists only on the n-dimensional subspace $\mathcal{F}_n$, while there is no information for $\mathcal{R}$. Thus, we will only ever consider the RKHS $\mathcal{F}_n \subset \mathcal{F}$ where there is Fisher information. Let h be a real symmetric and positive definite function over $\mathcal{X}$ defined by $h(x, x') = I[f(x), f(x')]$. As we saw earlier, h defines a RKHS, and it can be shown that the RKHS induced is in fact $\mathcal{F}_n$ spanned by the reproducing kernel on the dataset with the squared norm $||f||^2_{\mathcal{F}_n} = w^\top \Psi^{-1} w$.

Lemma 3.3. *Let* $\mathcal{F}_n$ *be equipped with the inner product*

$$\langle f_w, f_{w'} \rangle_{\mathcal{F}_n} = \mathbf{w}^\top \boldsymbol{\Psi}^{-1} \mathbf{w}',$$

where $\mathbf{w} = (w_1, \dots, w_n)$ *and* $f_w(x) = \sum_{i=1}^n h(x, x_i) w_i$. *Then,* h_n *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. Prove $\mathcal{F}_n$ is a Hilbert space?

$$f_j = \sum h(\cdot, x_i) w_{ij}$$

$$\begin{aligned} ||f_j - f||^2_{\mathcal{F}_n} &= \langle f_j - f, f_j - f \rangle \\ &\leq \langle f_j, f_j \rangle + \langle f, f \rangle \\ &= w_j \Psi w_j + w \Psi w \\ &= \Psi(w_j w_j^\top + w w^\top) \end{aligned}$$

Note that by defining $w_j(x) = \sum_{k=1}^{n} \psi_{jk} h(x, x_k)$, we see that

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

is an element of $\mathcal{F}_n$. Now, we just need to prove the reproducing property. Denote by ψ_{ij}^{-} the (i, j)th element of $\mathbf{\Psi}^{-1}$. Since $\langle h(\cdot, x_i), h(\cdot, x_j) \rangle_{\mathcal{F}_n} = \psi_{ij}^{-}$, we have

13. How?

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

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

1

Is the Fisher information metric and semi-norm over $\mathcal{F}$ useful?

3.5 The I-prior

Here we consider data dependent priors—seemingly data dependent (i.e. dependent on X) but the whole model is conditional on X implicitly, so there is no issue. If prior depended on y then there is a problem, at least, violates Bayesian first principles (using the data twice such that a priori and a posteriori same amount of information). Rather, more of a principled prior. One that is based on objectivity of maximum entropy—if one does not know anything, best to choose prior which maximises uncertainty. We see that it coincides with the Fisher information induced RKHS.

Goal is always to estimate $f \in \mathcal{F}$ based on finite amount of data points. We know MLE is not so good, so want regularise by some prior. Unfortunately, $\mathcal{F}$ might be huge such that data don't provide enough information for f to be estimated sufficiently well. We ask: What is the smallest subset for which there is full information coming from the data? Intuitively, it must be of n-dimensions, the sample size of the data. Rather separately, we found out what the Fisher information for f looks like, and deduced that there is Fisher information only on an orthogonal projection of $\mathcal{F}$ on to $\mathcal{F}_n$. There is this flavour of dimension reduction—no need to consider the entire space, because this is futile, but just consider functions in the smaller subspace, as this is the best we can do anyway. Therefore, we just look in this subspace $\mathcal{F}_n$ for an appropriate approximation to f. In particular, what prior should I use? On the basis of maximum entropy principle, I figure out that the form of our I-prior. The connection of $\mathcal{F}_n$ to Fisher information is this: $\mathcal{F}_n$ is the subspace of $\mathcal{F}$ for which Fisher information exists. Equipping this space with a particular inner product reveals that $\mathcal{F}_n$ is a RKHS with reproducing kernel equal to the Fisher information for f.

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 any f and f' for which $f(x_i) = f'(x_i), i = 1,\dots,n$. 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. Since the Fisher information for $\langle g, f\rangle_\mathcal{F}$ is non-zero for any non-zero $g \in \mathcal{F}_n$, there is information to allow a comparison between any pair of functions in $f_0 + \mathcal{F}_n$.

15. If data do not provide enough information, isn't the purpose of the prior to provide the missing information?

Key questions:

- What does it mean to say that the measure space $(\mathcal{F}, \nu)$ has a probability density function π? A probability density function p on $(\mathcal{F}, \nu)$ is a ν-measurable function from $\mathcal{F}$ to $[0, \infty)$ such that $p \, \mathrm{d}\nu$ is a probability measure on $\mathcal{F}$.

- What does it mean for $f \in \mathcal{F}$ to be Gaussian?

Let (Θ, D) be a metric space and let $\nu = \nu_D$ be a volume measure induced by D (e.g. Hausdorff measure). Denote by π a density of Θ relative to ν, i.e. if θ is a random variable with density π, then for any measurable subset $A \subset \Theta$, $\mathrm{P}(\theta \in A) = \int_A \pi(t)\nu(\mathrm{d}t)$.

Definition 3.3 (Entropy)**.** The entropy of a distribution π over $\mathcal{F}$ relative to a measure ν is defined as

$$\mathcal{E}(\pi) = -\int_{\mathcal{F}} \pi(f) \log \pi(f) \, \mathrm{d}\nu(f).$$

This converges if $\pi \log \pi$ is Lebesgue integrable, i.e. $\pi \log \pi \in \mathrm{L}^1(\mathcal{F}, \nu)$.

Definition 3.4 (Functional derivative)**.** Given a manifold M representing continuous/smooth functions ρ with certain boundary conditions, and a functional $F : M \to \mathbb{R}$, the functional derivative of $F[\rho]$ with respect to ρ, denoted $\partial F / \partial \rho$, is defined by

$$\begin{aligned}\int \frac{\partial F}{\partial \rho}(x)\phi(x) \, \mathrm{d}x &= \lim_{\epsilon \to 0} \frac{F[\rho + \epsilon\phi] - F[\rho]}{\epsilon} \\ &= \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon} F[\rho + \epsilon\phi]\right]_{\epsilon=0},\end{aligned}$$

where ϕ is an arbitrary function. The function $\partial F / \partial \rho$ as the gradient of F at the point ρ, and

$$\partial F(\rho, \phi) = \int \frac{\partial F}{\partial \rho}(x)\phi(x) \, \mathrm{d}x$$

as the directional derivative at point ρ in the direction of ϕ. Analogous to vector calculus, the inner product with the gradient gives the directional derivative.

Example 3.2 (Functional derivative of entropy)**.** Let X be a discrete random variable with probability mass function $p(x) \geq 0$, for $\forall x \in \Omega$, a finite set. The entropy is a functional of p, namely

$$\mathcal{E}[p] = -\sum_{x \in \Omega} p(x) \log p(x).$$

Equivalently, using the counting measure ν on Ω, we can write

$$\mathcal{E}[p] = -\int_{\Omega} p(x) \log p(x) \, \mathrm{d}\nu(x).$$

$$
\begin{aligned}
\int_\Omega \frac{\partial \mathcal{E}}{\partial p}(x)\phi(x) &= \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathcal{E}[p+\epsilon\phi]\right]_{\epsilon=0} \\
&= \left[-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\big(p(x)+\epsilon\phi(x)\big)\log\big(p(x)+\epsilon\phi(x)\big)\right]_{\epsilon=0} \\
&= -\int_\Omega \left(\frac{p(x)\phi(x)}{p(x)+\epsilon\phi(x)} + \frac{\epsilon\phi(x)}{p(x)+\epsilon\phi(x)} + \phi(x)\log\big(p(x)+\epsilon\phi(x)\big)\right)\mathrm{d}x \\
&= -\int_\Omega \left(1+\log p(x)\right)\phi(x)\,\mathrm{d}x.
\end{aligned}
$$

Thus, $(\partial\mathcal{E}/\partial p)(x) = -1 - \log p(x)$.

Lemma 3.4 (Maximum entropy distribution)**.** *Let $(\mathcal{X}, d)$ be a metric space and let $\nu = \nu_d$ be a volume measure induced by d. Let p be a probability density function on $(\mathcal{X}, d)$. The entropy maximising density, which satisfies*

$$\arg\max_p \mathcal{E}(p) = -\int_\mathcal{X} p(x)\log p(x)\ \mathrm{d}\nu(x),$$

subject to the constraints

$$\mathrm{E}\left[d(x,x_0)^2\right] = \int_\mathcal{X} d(x,x_0)^2 p(x)\,\mathrm{d}\nu(x) = \textit{const.}, \qquad \int_\mathcal{X} p(x)\,\mathrm{d}\nu(x) = 1,$$
$$\textit{and} \quad p(x) \geq 0,$$

is the density given by

$$\tilde{p}(x) \propto \exp\left(-\frac{1}{2}d(x,x_0)^2\right),$$

for some $x_0 \in \mathcal{X}$. If $(\mathcal{X}, d)$ is a Euclidean space and ν a flat (Lebesgue) measure then $\tilde{p}$ represent a (multivariate) normal density.

Proof. This follows from standard calculus of variations. We provide a sketch proof here. Set up the Langrangian

$$
\begin{aligned}
\mathcal{L}(p,\gamma_1,\gamma_2) = -\int_\mathcal{X} p(x)\log p(x)\ \mathrm{d}\nu(x) &+ \gamma_1\left(\int_\mathcal{X} d(x,x_0)^2 p(x)\,\mathrm{d}\nu(x) - \text{const.}\right) \\
&+ \gamma_2\left(\int_\mathcal{X} p(x)\,\mathrm{d}\nu(x) - 1\right).
\end{aligned}
$$

From the above lemma and example, taking derivatives with respect to p yields

$$\frac{\partial}{\partial p}\mathcal{L}(p,\gamma_1,\gamma_2)(x) = -1 - \log p(x) + \gamma_1 d(x,x_0)^2 + \gamma_2.$$

Set this to zero, and solve for p:

$$\begin{aligned} p(x) &= \exp\big(\gamma_1 d(x,x_0)^2 + \gamma_2 - 1\big) \\ &\propto \exp\big(\gamma_1 d(x,x_0)^2\big) \end{aligned}$$

which is positive for any values of γ_1 (and γ_2). This density normalises to one if $\gamma_1 < 0$, so we choose $\gamma_1 = -1/2$. If $\mathcal{X} = \mathbb{R}^n$ and that ν is the Lebesgue measure then $d(x,x_0) = \|x - x_0\|_{\mathbb{R}^n}$, so $\tilde{p}$ is recognised as a multivariate normal density centred at x_0 with identity covariance matrix. □

Theorem 3.5 (The I-prior). *Let $\mathcal{F}$ be an RKHS with kernel h, and consider the finite dimensional affine 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 Π_0 be the class of distributions p such that*

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

Denote by $\tilde{p}$ the density of the entropy maximising distribution among the class of distributions within Π_0. Then, $\tilde{p}$ is Gaussian over $\mathcal{F}$ with mean f_0 and covariance kernel equal to the reproducing kernel of $\mathcal{F}_n$, i.e.

$$\mathrm{Cov}\big(f(x), f(x')\big) = h_n(x,x').$$

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

Proof. Recall the fact that any $f \in \mathcal{F}$ can be decomposed into $f = f_n + r_n$, with $f_n \in \mathcal{F}_n$ and $r_n \in \mathcal{R}_n$, the orthogonal complement of $\mathcal{F}_n$. Also recall that there is no Fisher information about any $r \in \mathcal{R}_n$, and therefore it is not possible to estimate r_n from the data. Therefore, $p(r_n) = 0$, 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 $d(f,f') = \|f - f'\|_{\mathcal{F}_n}$. Thus, for $f \in \mathcal{F}$ of the form $f = \sum_{i=1}^n h(\cdot, x_i) w_i$ (i.e., $f \in \mathcal{F}_n$) and provided $f_0 \in \mathcal{F}_n \subset \mathcal{F}$,

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

Thus, by Lemma 3.4, the maximum entropy distribution for $f = \sum_{i=1}^n h(\cdot, x_i) w_i$ is

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

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 kernel Σ is the bilinear form satisfying

$$\begin{aligned}\operatorname{Cov}\big(f(x), f(x')\big) &= \operatorname{Cov}\big(f, \langle h(\cdot,x)\rangle_{\mathcal{F}_n}, \langle f, h(\cdot,x')\rangle_{\mathcal{F}}\big) \\ &= \big\langle \Sigma, h(\cdot,x) \otimes h(\cdot,x')\big\rangle_{\mathcal{F}\otimes\mathcal{F}} .\end{aligned}$$

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

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

where, making use of the reproducing property, the first term on the left hand side is

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

while the second term on the left hand side is

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

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 $\mathcal{F}_n$. □

3.6 Rate of convergence

We used the true Fisher information. Efron and Hinkley (1978) say favour the observed information instead. Does this change if we use MLE $\hat{f}$ instead? Probably not... we don't use MLE anyway!

https://stats.stackexchange.com/questions/179130/gaussian-process-proofs-and-results

https://stats.stackexchange.com/questions/268429/do-gaussian-process-regression-have-the-universal-approximation-property

Bibliography

Bergsma, W. (2017). “Regression with I-priors”. In: *Unpublished manuscript.*

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

Cohen, S. (2002). “Champs localement auto-similaires”. In: *Lois d’échelle, fractales et ondelettes.* Ed. by P. Abry, P. Gonçalves, and J. L. Véhel. Vol. 1. Hermès Sciences Publications.

Efron, B. and D. V. Hinkley (1978). “Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information”. In: *Biometrika* 65.3, pp. 457–483.

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

Kimeldorf, G. S. and G. 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.

Mandelbrot, B. B. and J. W. Van Ness (1968). “Fractional Brownian motions, fractional noises and applications”. In: *SIAM review* 10.4, pp. 422–437.

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

Steinwart, I. and A. Christmann (2008). *Support vector machines.* Springer Science & Business Media.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference.* Springer Science & Business Media.

Appendix A

Appendix for I-probit

Index