To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
January 30, 2018

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 provides an outline of the thesis, in addition to introducing the statistical model of interest.

1.1 Regression models

For subject $i \in \{1, \dots, n\}$, assume a real-valued response y_i has been observed, as well as a row vector of p covariates $x_i = (x_{i1}, \dots, x_{ip})$, where each x_{ik} belongs to some set $\mathcal{X}_k$, for $k = 1, \dots, p$. Let $\mathcal{S} = \{(y_1, x_1), \dots, (y_n, x_n)\}$ 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

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

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 $j \in \{1, \dots, m\}$. 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 from which we may choose a regression function appropriately. This is given to us by assuming that our regression function for the normal model lies in some reproducing kernel Hilbert space (RKHS) $\mathcal{F}$, with 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.

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

Tikhonov regularisation, Bayesian interpretation, priors

Having decided on a functional structure for f, we now turn to the task of estimating f.

Remark on dimensionality.

1.4 Regression using I-priors

The definition of an RKHS entails that any function in $\mathcal{F}$ can be approximated arbitrarily well by functions of the form

$$f(x) = f_0(x) + \sum_{i=1}^{n} h_\eta(x, x_i) w_i \tag{1.3}$$

where $w_1, \dots, w_n$ are real-valued[1]. Here, $f_0 \in \mathcal{F}$ is some function chosen a priori which represents the 'best guess' of the regression function. The *I-prior* for our regression function f in (1.1) subject to (1.2) is defined as the distribution of a random function of the form (1.3) when the w_i are distributed according to

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

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

$$\mathbf{f} := \big(f(x_1), \dots, f(x_n)\big)^\top \sim \mathrm{N}_n(\mathbf{f}_0, \mathbf{H}_\eta \mathbf{\Psi} \mathbf{H}_\eta) \tag{1.4}$$

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 (Bergsma, 2017), 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.

2. which are?

[1] That is to say, $\mathcal{F}$ is spanned by the functions $h(\cdot, x)$. More precisely, $\mathcal{F}$ is the completion of the space $\mathcal{G} = \operatorname{span}\{h(\cdot, x) | x \in \mathcal{X}\}$ endowed with the squared norm $\|f\|^2_{\mathcal{G}} = \sum_{i=1}^n \sum_{i=1}^n w_i w_j h(x_i, x_j)$ for f of the form (1.3). See, for example, Berlinet and Thomas-Agnan, 2011 for details.

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

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}}, \tag{1.5}$$

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

$$\mathrm{E}\left[f(x)\big|\mathbf{y}\right] = f_0(x) + \mathbf{h}_\eta^\top(x) \cdot \overbrace{\boldsymbol{\Psi}\mathbf{H}_\eta\big(\mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}\big)^{-1}(\mathbf{y} - \mathbf{f}_0)}^{\tilde{\mathbf{w}}} \tag{1.6}$$

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 (1.6) are the posterior means of the random variables w_i in the formulation (1.3). The point-evaluation posterior variance for f is given by

$$\mathrm{Var}\left[f(x)\big|\mathbf{y}\right] = \mathbf{h}_\eta^\top(x)\big(\mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}\big)^{-1}\mathbf{h}_\eta^\top(x). \tag{1.7}$$

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[2] as in (1.6), 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. In our case, θ 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

[2] 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.

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.

   3. Why/How is it simple?

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

   When considering functions that minimise the squared loss function[3]

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

   over-fitting is a genuine concern. In fact, any function f that passes through the data points minimises Λ. Regularising the problem with the use of I-priors prevents over-fitting. Though over-fitting can be solved using Tikhonov regularisation, the posterior solution under an I-prior does not tend to under-smooth as much as Tikhonov regularisation does. 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. Cite or show later?

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

[3] For the normal model, this is in fact twice the negative log-likelihood of f, up to a constant.

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.

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 (1.6) and (1.7), 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

1.7 Regression and regularised least squares

The task of regression modelling is to choose the most appropriate regression function $f \in \mathcal{F}$. It would be helpful if we had a measure of the quality of our choice of f. Define the risk functional $R : \mathcal{F} \to \mathbb{R}$ as

$$R[f] = \mathrm{E}[L\big(y, f(x)\big)] = \int L\big(y, f(x)\big) \, \mathrm{dP}(y, x), \tag{1.8}$$

where $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is some loss function, and $\mathrm{P}(y, x)$ is the probability measure of the observed sample. In most cases, this probability measure is unknown, and an empirical risk measure is used instead:

$$R[f] = \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big). \tag{1.9}$$

The squared loss function given by

$$L\big(y_i, f(x_i)\big) = \sum_{j=1}^{n} \psi_{ij} \big(y_i - f(x_i)\big)\big(y_j - f(x_j)\big), \tag{1.10}$$

when used, defines the least squares regression. It is worthwhile noting that for the normal regression model, a solution obtained by minimising the squared loss function is equivalent to the maximum likelihood estimator.

This problem may be ill-posed, in the sense that if the space of functions is relatively unconstrained, then there is likely to be more than one solution to the regression problem. In fact, any function which passes through all the data points is an acceptable solution. This clearly leads to overfitting and poor generalisations.

The most common method to overcome this issue is Tikhonov regularisation. A regularisation term is added to the risk function, with the aim of imposing a penalty on the complexity of f. Concrete notions of complexity penalties can be introduced if $\mathcal{F}$ is a normed space, though reproducing kernel Hilbert space (RKHS) are typically used as it gives great conveniences (see Section 2). In particular, smoothness assumptions on f encoded by a suitable RKHS can be represented by using the RKHS norm $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}$ as the regularisation term. Therefore, the solution to the regularised least squares problem f_{reg} is the minimiser of the function from $\mathcal{F}$ to $\mathbb{R}$ defined by the mapping

$$f \mapsto \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} \big(y_i - f(x_i)\big)\big(y_j - f(x_j)\big) + n\lambda \|f\|_{\mathcal{F}}^2, \tag{1.11}$$

which also happens to be the penalised maximum likelihood solution. The $\lambda > 0$ 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.

Tikhonov regularisation also has a well-known Bayesian interpretation, whereby the regularisation term encodes prior information about the function f. For the regression model stated earlier in (1.1) subject to the assumption in (1.2), let $\mathcal{F}$ be an RKHS equipped with the positive definite kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then, it can be shown that f_{reg} is the posterior mean of f given a Gaussian process prior with zero mean and covariance kernel $H_\lambda = \big(\frac{1}{n\lambda} h(x_i, x_j)\big)_{i,j=1}^n$. This

There are two drawbakcs to Tikhonov regularisation, and the first is that it can systematically undersmooth. From a Bayesian viewpoint, undersmoothing can be said to occur if the support of the prior consists of functions that are rougher than those in $\mathcal{F}$. In particular, at least for certain RKHSs, the sample paths of the Gaussian process with the reproducing kernel of the RKHS as the covariance kernel are rougher (by some margin) than the roughest functions in the RKHS. Undersmoothing can then adversely impact the estimation of f, and in real terms might even show features and artefacts that are not really there.

5. Is it fair to say that most processes that we want to estimate hardly come about from realisations of smooth functions? I.e. Brownian motion paths (rough) are more likely to occur in "nature" than very, very smooth paths (say squared exponential paths).

The second drawback is in regards to estimation of the regularisation parameters. Estimation of these parameters requires either minimisation of some cross-validation error criterion, or a direct minimisation of the penalised functional (1.11). The latter of these two methods can be seen as obtaining an empirical Bayes estimate by maximising the marginal likelihood in the Bayesian interpretation of regularisation. Either way, estimation can prove difficult when there are a lot of regularisation parameters to estimate.

The I-prior methodology does not suffer from these two drawbacks. According to Bergsma (2017), ...

6. Complete this

1.8 The I-prior and its advantages

As alluded to earlier, RKHSs has many desirable properties. For the regression model (1.1) subject to (1.2), let $\mathcal{F}$ be an RKHS. Every RKHS defines a reproducing kernel function that is both symmetric and positive definite, and the converse is also true.

There are three main types of RKHS studied in this thesis, allowing linear and smooth effects of Euclidean covariates as well as the incorporation of categorical covariates: the *canonical* RKHS, consisting of linear functions of the covariates; the fractional Brownian motion (fBm) RKHS, consisting of smooth functions of the covariates; and the *Pearson* RKHS for nominal or categorical covariates. The fBm RKHS has smoothness parameter $\gamma \in (0,1)$, called the Hurst coefficient. The most common value for this parameter is $1/2$, which for a real covariate gives a fitted function close to the familiar cubic spline smoother, although this could be treated as an unknown parameter to be estimated. More on these kernels later.

We can build upon these kernels by adding or multiplying them, and the result is still a positive definite kernel which induces a new RKHS. This is particularly useful because we can think of our regression function as being decomposed of functions belonging to different RKHS, depending on the effect of the covariate desired. As an example, suppose that each $x \in \mathcal{X}$ is 2-dimensional, so that $x = (x_1, x_2)$. We can assume that the regression function decomposes as follows:

$$f(x) = f(x_1, x_2) = f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2)$$

This is possible because $\mathcal{F}$ is a vector space over $\mathbb{R}$. Here, we have assumed that the function f partitions into two main effects f_1 and f_2 and an *interaction effect* f_{12}. Each of the main effects are in some RKHS, depending on the effect of the corresponding covariate (linear, smooth, or nominal), and thus would have a kernel $h_j : \mathcal{X}_j \times \mathcal{X}_j \to \mathbb{R}$. As the scale of an RKHS over a set $\mathcal{X}$ may be arbitrary, each of the kernels are multiplied by a scale parameter λ_j. The space of functions for the interaction effects are then assumed to be in the so-called tensor product space of the corresponding main effect functions. In our case, $f_{12} \in \mathcal{F}_{12}$, where $\mathcal{F}_{12}$ is an RKHS with kernel equal to the product of kernels $h_{12}\big((x_1, x_2), (x_1', x_2')\big) = \lambda_1\lambda_2 \cdot h_1(x_1, x_1')h_2(x_2, x_2')$.

Suppose that we have a multilevel data set, where x_1 is real-valued, x_2 is nominal-valued indicating the level to which the observation belongs to. We can model these data by choosing the canonical kernel on x_1 and the Pearson kernel for the x_2, and the interaction effect represents the varying effect of x_1 in each level x_2. If instead we had a time covariate x_1 and a categorical covariate x_2 representing treatment effect say, then we can build a longitudinal model with either a linear or smooth effect of time. Again, the interaction effect will convey x_2 as time-varying. Of course, we can partition the function as is necessary, such as excluding the interaction effect or including additional terms such as three-way interactions. Now suppose that we have functional data, i.e. the set $\mathcal{X}$ consists of functions. If we assume that the xs lie in some Hilbert space (not necessarily an RKHS) then we have an inner-product (and also a norm) defined on $\mathcal{X}$. As we will see later, the canonical and fBm kernels make use of the inner-product on $\mathcal{X}$ so we are able to proceed as we did before. We can see that this framework provides a unifying approach to various regression models.

We discussed earlier that regularisation has a Bayesian interpretation, whereby a prior distribution is assigned to the regression function. Specifically, it is a Gaussian process prior with mean zero and covariance kernel equal to the reproducing kernel of the RKHS that f belongs to. The I-prior for f is a also a zero mean Gaussian prior but has a different covariance kernel, namely the Fisher information for f. If h is the reproducing kernel for the RKHS, then the Fisher information between $f(x)$ and $f(x')$ is given as

$$\mathcal{I}[f(x), f(x')] = \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(x, x_i) h(x', x_j). \tag{1.12}$$

Hence, f follows an I-prior distribution if it can be written in the form

$$f(x) = \sum_{i=1}^n h(x, x_j) w_j, \tag{1.13}$$

where

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

Note that, of course, a non-zero value or function for the prior mean could be taken as well. The I-prior is a class of objective priors - it is the distribution for which entropy is maximised (subject to certain constraints). In this sense, it is considered the prior which gives the least amount of information a priori - then perhaps the term I-prior is somewhat of a misnomer, since the 'I' stands for (Fisher) information.

An intuitively attractive property of the I-prior is that if much information about a linear functional of f (e.g. a regression coefficient) is available, its prior variance is large, and the data have a relatively large influence on the posterior, while if little information about a linear functional is available, the posterior will be largely determined by the prior mean, which serves as a 'best guess' of f.

1.9 Estimation

The I-prior methodology consists of estimation of the regression function by its posterior distribution under the I-prior, where we take the posterior mean as the summary measure. Write $\mathbf{f} = \big(f(x_1), \dots, f(x_n)\big)^\top$. From (1.12), the Fisher covariance kernel for f is $H_\lambda \Psi H_\lambda$, where $H_\lambda = \big(h_\lambda(x_i, x_j)\big)_{i,j=1}^n$ and h_λ is the (scaled) reproducing kernel of $\mathcal{F}$. The I-prior on f for model (1.1) subject to (1.2) is

$$\mathbf{f} \sim \mathrm{N}_n(\mathbf{f}_0, H_\lambda \Psi H_\lambda)$$

where $\mathbf{f}_0 = \big(f_0(x_1), \dots, f_0(x_n)\big)^\top$ is some prior mean typically set to zero. We are then interested in two main things:

1. The posterior distribution for the regression function

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

2. The posterior predictive distribution for new data x_{new}

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

   where $f_{\text{new}} = f(x_{\text{new}})$.

It can be shown that for any $x \in \mathcal{X}$, the posterior distribution of f is normal with mean and variance given by

$$\mathrm{E}[f(x)|\mathbf{y}] = f_0(x) + \mathbf{h}_\lambda(x) \Psi H_\lambda (H_\lambda \Psi H_\lambda + \Psi^{-1})^{-1} \big(y - f_0(x)\big)$$

and

$$\mathrm{Var}[f(x)|\mathbf{y}] = \mathbf{h}_\lambda(x)^\top (H_\lambda \Psi H_\lambda + \Psi^{-1})^{-1} \mathbf{h}_\lambda(x),$$

where $\mathbf{h}_\lambda(x) = \big(h_\lambda(x, x_1), \dots, h_\lambda(x, x_n)\big)^\top$.

There is the matter of estimating the model (hyper-)parameters - the error precision

Ψ, the RKHS scale parameters λ, and any other parameters that might be associated with the kernel (e.g. the smoothing parameter in an fBm kernel). These may be estimated in a variety of ways. The first is by maximum marginal likelihood, which is also known as the empirical Bayes approach. The marginal distribution $p(y) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}) \, \mathrm{d}\mathbf{f}$ is easily obtained in the case of the normal model, and it is

$$\mathbf{y} = (y_1, \ldots, y_n)^\top \sim \mathrm{N}_n \big(f_0(x), H_\lambda \Psi H_\lambda + \Psi^{-1} \big).$$

The marginal likelihood can be maximised in the usual way (e.g. Newton-type methods) with respect to the model parameters, but for complex models involving a lot of parameters, this may be challenging. Instead, a better approach is the expectation-maximisation (EM) algorithm. The I-prior model emits a simple E- and M-step which makes this method favourable. For most models, a closed-form solution to the M-step is available thereby reducing the EM algorithm to an iterative updating scheme of the parameters. Finally, a fully Bayesian approach may be taken as well, whereby prior distributions are assigned to the model parameters and posterior samples obtained via Markov chain Monte Carlo (MCMC) methods.

Regardless of the estimation procedure, computational complexity is dominated by the inversion of the $n \times n$ matrix $V_y = H_\lambda \Psi H_\lambda + \Psi^{-1}$ as a function of the model parameters. In the case of Newton-type approaches to likelihood maximisation, V_y^{-1} appears in the kernel of the marginal Gaussian density for $\mathbf{y}$. In the case of the EM algorithm, every update cycle also involves a similar calculation, and this is quite similar to the calculations required from a Gibbs-sampling approach for stochastic MCMC sampling.

I-priors, while being philosophically different from Gaussian process priors, do share the same computational hurdle. As such, several methods exist in the machine learning literature to overcome this issue. Amongst others, is a method to approximate the covariance kernel by a low-rank matrix, so that the most expensive operation of inverting a $n \times n$ matrix is greatly reduced. Our approach is to apply the Nyström method of low-rank matrix approximation, and we find that this works reasonably well for the fBm RKHS.

Another computational hurdle is to ensure numerical stability. We find that due to the structure of the marginal covariance V_y, numerical instabilities can and are likely to occur - which may give rise to embarrassments such as negative covariances. We employ a stable eigendecomposition regime which allows us to efficiently calculate matrix squares and inverses by making use of the spectral theorem.

1.10 I-priors for classification

Suppose now we are interested in a regression model where the responses are categorical. Assume a categorical distribution on the responses with certain probabilities for each class and for each observation. This is of course a generalisation of the Bernoulli distribution to more than two possible outcomes. The question is how can we relate the effect of the covariates through the function, which has unrestricted range, to the responses, which may only take one of m several outcomes? In the spirit of generalised linear models, we answer this by making use of an appropriate link function, and our case, the probit link function. In the binary case, this amounts to squashing our regression function through the (inverse) probit link function in order to model probabilities which are between zero and one. This idea is then extended to the multinoulli case, giving rise to a multinomial probit I-prior model, which we call I-probit.

The main issue with estimation now is that because our responses no longer follow a Gaussian distribution, the relevant marginal distribution, on which the posterior depends, can no longer be found in closed form. The integral required to perform the calculation is intractable, and the focus now is on methods to adequately approximate the integral.

In the Bayesian literature, the Laplace approximation amounts to approximating the posterior distribution with a normal distribution centred around the mode of the integrand. Additionally, the covariance matrix is equal to the inverse (negative) Hessian. Having approximated the posterior by a Gaussian distribution, one could then proceed to find the marginal easily, which is then maximised. Due to the Newton step in the Laplace step, the whole procedure scales cubicly with both the sample size and the number of outcomes, which makes it undesirable to implement.

MCMC methods such as Gibbs sampling or Hamiltonian Monte Carlo can also provide a stochastic approximation to the integral. Unfortunately, the difficulties faced in the continuous case for MCMC methods also present themselves in the categorical case.

Deterministic approaches such as quadrature methods prove unfeasible. Quadrature methods scale exponentially with the variables of integration, in our case, is the sample size.

We consider a type of approximation based on minimising the Kullback-Leibler (KL) divergence from the approximating density to the true posterior density. This is done without making any distributional assumptions, only that the our approximating density factorises over its components (ie an independence assumption) - this is known as the mean-field approximation, which has its roots in the physics literature. As an aside, the term ‘variational’ stems from the fact that a minimisation of a functional, rather than a function, is involved, and this requires the calculus of variations.

By working in a fully Bayesian setting, we append the model parameters to the list of unknowns in which to estimate, and employ the variational approximation to find a suitable approximation to the required posterior density. The result is an iterative algorithm, similar to the EM.

As this variational EM works harmoniously with exponential family distributions, the probit link is much preferred over the logit. Most of the EM updating cycles which could be found in closed form are also applicable in the variational EM. Unlike Gaussian process priors, the variational method does not typically result in closed-form updates for the RKHS scale parameters. In such cases, an additional step such as importance sampling is required, which arguably reduces efficiency of the whole variational scheme.

The variational EM is implemented in the R package iprobit. This has been shown to work well for several toy examples as well as real world applications. In the binary case, the I-prior outperforms other popular classication methods including k-nearest neighbours, support vector machines, and Gaussian process classification.

1.11 I-priors for Model selection

As mentioned earlier, model selection can easily be done by comparing likelihoods (empirical Bayes factors). However, with a large number of variables, these pairwise comparisons quickly become unfeasible to perform. We suggest a fully Bayesian approach to estimating posterior model probabilities, and selecting models based on these quantities. For example, one may choose the model which gives the largest posterior model probability (maximum probability model). These are done using MCMC methods (Gibbs sampling).

We restrict ourselves to linear models only. We can easily derive an equivalent I-prior representation by working in the feature space of the betas (linear effects). As a side note, if the dimensions of the linear effects is much, much less than the sample size, then it is worth working in this representation.

We believe the I-prior performs superiorly in cases where there is multicollinearity.

This is evidenced by the simulation results that we conducted on a 100-variate experiment, and also in the real-data examples comparing our method with others such as greedy selection, g-priors, and regularisation (ridge and Lasso).

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.

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

List of Figures

List of Tables

List of Theorems

List of Definitions

List of Abbreviations

fBm	Fractional Brownian motion.
Lasso	Least absolute shrinkage and selection operator.
RKHS	Reproducing kernel Hilbert space.

Index