To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: 'Regression modelling using Fisher information covariance kernels (I-priors)'

Chapter 1

Introduction

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

Estimation of the regression function may be done in many ways. This thesis concerns the use of *I-priors* (Bergsma, 2017), in a semi-Bayesian manner, for regression modelling. The I-prior is an objective, entropy-maximising prior for the regression function which makes use of its Fisher information. The essence of regression modelling using I-priors is introduced briefly below, but as the development of I-priors is fairly recent, we dedicate two full chapters (Chapters 2 and 3) to describe the concept fully.

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

sec:introregmod

1.1 Regression models

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

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

{eq:model1}

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

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

{eq:model1ass}

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

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

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

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

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

Moving on from linear models, smoothing models may be of interest as well. A myriad of models exist for this type of problem, with most classed as nonparametric regression, and the more popular ones include LOcal regrESSion (LOESS), kernel regression, and smoothing splines. Semiparametric regression models, on the other hand, combines the linear component of a regression model with a non-parameteric component.

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

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

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

1.2 Vector space of functions

It would be beneficial to prescribe some sort of structure for which 1) we may choose a regression function appropriately, and 2) this function will generalise well to unseen data (prediction). This needed structure is given to us by assuming that our regression function for the normal model lies in some reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ equipped with the reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Often, the reproducing kernel (or simply kernel, for short) is indexed by one or more parameters which we shall denote as η. Correspondingly, the kernel is rightfully denoted as h_η to indicate the dependence of the parameters on the kernels, though where this is seemingly obvious, might be omitted. Throughout this thesis we shall make the assumption that our regression function lies in a reproducing kernel Hilbert space $\mathcal{F}$.

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

Due to the one-to-one mapping between the set of kernel functions and the set of RKHSs, choosing the space in which the regression function lies is equivalent to choosing a particular kernel function, and this is chosen according to the desired effects of the covariates on the regression function. 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 either a linear, smooth or categorical dependence, or any combinations thereof, are of interest. This would allow us to fit the various models described earlier within this RKHS framework.

1.3 Estimating the regression function

Having decided on a functional structure for f, we now turn to the task of choosing the best $f \in \mathcal{F}$ that fits the data sample $\mathcal{S}$. 'Best' here could mean a great deal of things, such as choosing f which minimises an empirical risk measure[1] defined by

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

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

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

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

The ML estimator of f interpolates the data if the dimension of $\mathcal{F}$ is at least n, so is of little use. The most common method to overcome this issue is *Tikhonov regularisation*, whereby a regularisation term is added to the risk function, with the aim of imposing a penalty on the complexity of f. In particular, smoothness assumptions on f can be represented by using its RKHS norm $\|\cdot\|_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}$ as the regularisation term[2].

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

Therefore, the solution to the regularised least squares problem—call this f_{reg}—is the minimiser of the function from $\mathcal{F}$ to $\mathbb{R}$ defined by the mapping

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

{eq:penfunctional}

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

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

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

{eq:repform}

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

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

1.4 Regression using I-priors

sec:introregiprior

Building upon the Bayesian interpretation of regularisation, Bergsma (2017) proposes a prior distribution for the regression function such that its realisations admit the form for the solution given in the representer theorem. The *I-prior* for the regression function f in (1.1) subject to (1.2) is defined as the distribution of a random function of the form

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

(1.4) when the w_i are distributed according to

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

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

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

{eq:iprior}

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

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

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

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

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

{eq:postmean}

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

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

{eq:postvar}

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

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

where we had defined $f_{\text{new}} := f(x_{\text{new}})$. This is again a normal distribution in the case of the normal model, with the same mean[3] as in (1.7), but a slightly different variance. These are of course well-known results in Gaussian process literature—see, for example, Rasmussen and Williams (2006) for details.

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

1.5 Advantages and limitations of I-priors

The I-prior methodology has the following advantages:

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

   The I-prior methodology has the ability to fit a multitude of regression models simply by choosing the RKHS to which the regression function belongs. As such, it can be seen as a unifying methodology for various regression models.

2. **Simple estimation procedure.**

   Estimation of model parameters using the aforementioned methods are very simple to implement, barring any computational and numerical hurdles, which will be discussed. This encourages parsimony, as the I-prior allows complex models to be specified by just a handful of model parameters.

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

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

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

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

4. **Better prediction.**

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

5. **Straightforward inference.**

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

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

Another issue when performing likelihood based methods is that the optimisation objective may be non-convex such that multiple local optima may exist. In such cases, multiple restarts from different 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) the assumption of $f \in \mathcal{F}$ an 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** provides a brief overview of functional analysis, and in particular, descriptions of interesting function spaces for regression. In **Chapter 3**, the concept of the Fisher information is extended to potentially infinite-dimensional parameters. This allows us to define the Fisher information for the regression function which parameterises the normal regression model, and we explain how this relates to the I-prior.
- The aforementioned computational methods relating to the estimation of I-prior models are explored in **Chapter 4**, namely the direct optimisation of the log-likelihood, the EM algorithm, and MCMC methods. The goal is to describe a stable and efficient algorithm for estimating I-prior models. The R package **iprior** is the culmination of the effort put in towards completing this chapter, which has been made publicly available on the Comprehensive R Archive Network (CRAN).
- Many models of interest involve response variables of a categorical nature. A naïve implementation of the I-prior model is certainly possible, but there is a more proper way to account for non-normality of errors. **Chapter 5** extends the I-prior methodology to discrete outcomes. There, the non-Gaussian likelihood that arises in the posteriors are approximated by way of variational inference. The advantages of the I-prior in normal regression models carry over into categorical response models.
- **Chapter 6** attempts to contribute to the area of variable selection. The use of I-priors in the normal model, like Gaussian process priors, allow model comparison to be done easily. Specifically for linear models with p variables to select from, model comparison requires elucidation of 2^p marginal likelihoods, and this becomes infeasible when p is large. We use a stochastic search method to choose models that have high posterior probabilities of occurring, equivalent to choosing models that have large Bayes factors.

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

Bibliography

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

hofmann2008kernel Hofmann, Thomas, Bernhard Schölkopf, and Alexander J Smola (2008). "Kernel methods in machine learning". In: *The annals of statistics*, pp. 1171–1220.

kimeldorf1970correspondence Kimeldorf, George S and Grace Wahba (1970). "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines". In: *The Annals of Mathematical Statistics* 41.2, pp. 495–502.

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