To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
August 15, 2017

Chapter 2

Regression using I-priors

Regression using I-priors. Test equation reference (1.1).

2.1 Some functional analysis

For a cleaner read, we do not use boldface type to denote vectors and matrices in this subsection. We begin by recalling that a vector space is a set endowed with two special operations: addition and scalar multiplication. A normed vector space is a vector space whose vectors have lengths, as induced by its norm.

Definition 2.1 (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:** $||\lambda f||_{\mathcal{F}} = |\lambda| ||f||_{\mathcal{F}}$, $\forall \lambda \in \mathbb{R}$, $\forall f \in \mathcal{F}$
- **Triangle inequality:** $||f + g||_{\mathcal{F}} \leq ||f||_{\mathcal{F}} + ||g||_{\mathcal{F}}$, $\forall f, g \in \mathcal{F}$
- **Separates points:** $||f||_{\mathcal{F}} = 0 \Leftrightarrow f = 0$

The norm $||\cdot||_{\mathcal{F}}$ induces a metric (a notion of distance) on $\mathcal{F}$: $d(f, g) = ||f - g||_{\mathcal{F}}$.

We can then define a Cauchy sequence.

Definition 2.2 (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 $\forall n, m > N$, $||f_n - f_m||_{\mathcal{F}} < \epsilon$.

Another important structure of vector spaces is the inner product, which allows us to study various geometrical notions such as orthogonality, among other things.

Definition 2.3 (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 $||f||_{\mathcal{F}} = \sqrt{\langle f, f\rangle_{\mathcal{F}}}$. Additionally, an inner product is said to be positive definite if $\langle f, f\rangle_{\mathcal{F}} \geq 0$, $\forall f \in \mathcal{F}$, and we can always define a norm on $\mathcal{F}$ using this inner product as $||f||_{\mathcal{F}} = \sqrt{\langle f, f\rangle_{\mathcal{F}}}$. Conversely, an inner product is said to be negative definite if $\langle f, f\rangle_{\mathcal{F}} \leq 0$, $\forall f \in \mathcal{F}$. An inner product is said to be indefinite it is neither positive nor negative definite.

A vector space equipped with a positive definite inner product that is also complete (contains the limits of all Cauchy sequences) is known as a Hilbert space. We now define a reproducing kernel Hilbert space.

A generalisation of a Hilbert space, one which is equipped with an indefinite inner product, is known as a Krein space.

Definition 2.4 (Krein space)**.** A vector space $\mathcal{F}$ for which an inner product $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ is defined is called a Krein space if there are two Hilbert spaces $\mathcal{F}_+$ and $\mathcal{F}_-$ spanning $\mathcal{F}$ such that

- All $f \in \mathcal{F}$ can be decomposed as $f = f_+ + f_-$ where $f_+ \in \mathcal{F}_+$ and $f_- \in \mathcal{F}_-$; and
- $\forall f, f' \in \mathcal{F}$, $\langle f, f'\rangle_{\mathcal{F}} = \langle f_+, f'_+\rangle_{\mathcal{F}_+} - \langle f_-, f'_-\rangle_{\mathcal{F}_-}$.

Any Hilbert space can be seen as a Krein space by taking $\mathcal{F}_- = \{0\}$. As we are dealing with function spaces, it might seem unusual in defining functions from $\mathcal{F}$ to $\mathbb{R}$, as the elements of $\mathcal{F}$ are themselves functions. 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 2.5 (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 2.6 (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**: $A(f + g) = A(f) + A(f')$, $\forall f, f' \in \mathcal{F}$.

Definition 2.7 (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 2.8 (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[1]), i.e. $\exists \lambda_x \geq 0$ such that $\forall f \in \mathcal{F}$,

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

The definition is similar for Krein spaces, but there is a slight technical condition regarding strong topologies (see **Ong2004**). Interestingly, the definition above has no mention of what a reproducing kernel is. Let us define it below.

Definition 2.9 (Reproducing kernel Krein space)**.** A Krein space of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ on a non-empty set $\mathcal{X}$ is called a reproducing kernel Krein space (RKKS) if the evaluation functional δ_x is a bounded linear operator $\forall x \in \mathcal{X}$, endowed with its strong topology.

Definition 2.10 (Kernels)**.** Let $\mathcal{X}$ be a non-empty set. A function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if there exists a real Hilbert space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}$ such that $\forall x, x' \in \mathcal{X}$,

$$h(x, x') = \langle \phi(x), \phi(x') \rangle.$$

Such a map $\phi : \mathcal{X} \to \mathcal{F}$ is known as the *feature map*, and the space $\mathcal{F}$ as the *feature space*. Out of interest, a given kernel may correspond to more than one feature map.

Definition 2.11 (Reproducing kernels)**.** Let $\mathcal{F}$ be a Hilbert space of functions over a non-empty set $\mathcal{X}$. A function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{F}$ if h satisfies

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

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

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

The connection between the definition of a RKHS and reproducing kernels is this: $\mathcal{F}$ is a RKHS space if and only if $\mathcal{F}$ has a reproducing kernel. It can also be proven that if this kernel exists, it is unique. We now turn to the one of the most important properties of the kernel function: positive-definiteness.

[1]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 $\mathcal{F}$.

By the definition of symmetry and positive definiteness of inner products on Hilbert spaces, it follows that kernel functions are symmetric and positive definite, and the following lemma is easily proven.

Lemma 2.1 (Positive-definiteness)**.** *Let $\mathcal{F}$ be a Hilbert space (not necessarily a RKHS), $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{F}$. Then $h(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$ is a symmetric and positive definite function, where a symmetric function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be positive definite if*

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

$\forall n \geq 1$, $\forall a_1, \ldots, a_n \in \mathbb{R}$, and $\forall x_1, \ldots, x_n \in \mathcal{X}$.

Proof.

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

□

Corollary 2.1.1. *Reproducing kernels of a RKHS are positive definite. For an RKKS, the reproducing kernel can be shown to be the difference between two positive definite kernels, but need not be itself positive definite.*

Proof. Take $\phi : x \mapsto h(\cdot, x)$. By Lemma 2.1, one has $h(x, x') = \langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}}$, which is the reproducing property of the kernel in a RKHS. The second statement follows by a similar argument and by definition of a RKKS (see Definition 2.4). □

Remarkably, the reverse direction also holds: a positive definite function is guaranteed to be the inner product in a Hilbert space between features $\phi(x)$ (Theorem 4.16 pp.118, Steinward and Christman, 2008). This proof is a bit technical so will not be shown here. What's important though, is By Definition 2.11 and Lemma 2.1 above, we can see how a reproducing kernel Hilbert space defines a reproducing kernel function that is both symmetric and positive definite. The celebrated Moore-Aronszajn theorem goes the other direction by stating that every symmetric, positive-definite function is a reproducing kernel[2] and defines a unique RKHS, thus establishing a bijection between the set of all positive definite functions on $\mathcal{X} \times \mathcal{X}$ and the set of all reproducing kernel Hilbert spaces. For Krein spaces it is slightly different: 1) The reproducing kernel of a RKKS

can be shown to be the difference between two positive definite kernels, so need not be positive definite itself; and 2) Every RKKS has a unique reproducing kernel, but a given reproducing kernel may have more than one RKKS associated with it.

Thus far, we have seen that given a RKHS $\mathcal{F}$, we may define a unique reproducing kernel associated with $\mathcal{F}$ which is symmetric and positive definite. The celebrated Moore-Aronszajn theorem goes the other direction by stating that every symmetric, positive-definite function is a reproducing kernel and defines a unique RKHS, thus establishing a bijection between the set of all positive definite functions on $\mathcal{X} \times \mathcal{X}$ and the set of all reproducing kernel Hilbert spaces. In other words, the kernel completely determines the function space. It is not quite the same with Krein spaces, however. Every RKKS has a unique reproducing kernel, but a given reproducing kernel may have more than one RKKS associated with it.

So why the fascination with reproducing kernel Hilbert/Krein spaces? In our case, it is the possibility of representing a regression analysis as functions in a RKKS. This greatly helps facilitate interpretation of models.

Lemma 2.2 (Regression functions in a RKKS)**.** *$\mathcal{F}$ is an RKKS if and only if there exists a feature space $\mathcal{B}$ for which a feature map of $\mathcal{F}$ maps onto.*

Proof. We first define a feature space and a feature map of $\mathcal{F}$.

Definition 2.12 (Features)**.** Consider a Krein space $\mathcal{F}$ of real functions over $\mathcal{X}$ with reproducing kernel h. Let $\mathcal{B}$ be a real Krein space over $\mathcal{X}$, and ϕ a map from $\mathcal{X}$ to $\mathcal{B}$, such that for every $f \in \mathcal{F}$, $\exists \beta \in \mathcal{B}$ such that

$$f(x) = \langle \phi(x), \beta \rangle_{\mathcal{B}}, \forall x \in \mathcal{X} \tag{2.1}$$

and

$$\langle f, f' \rangle_{\mathcal{F}} = \langle \beta, \beta' \rangle_{\mathcal{B}}. \tag{2.2}$$

Then $\mathcal{B}$ is called a *feature space* and ϕ a *feature map* of $\mathcal{F}$.

Now suppose $\mathcal{F}$ is a Krein space of real functions over $\mathcal{X}$ with a feature space $\mathcal{B}$ and a feature map ϕ. Then by defining the kernel function as $h(x, x') = \langle \phi(x), \beta \rangle_{\mathcal{B}}$, we show the reproducing property

$$\langle f, h(\cdot, x) \rangle_{\mathcal{F}} = \langle \phi(x), \beta \rangle_{\mathcal{B}} = f(x),$$

where the first equality is by (2.2) and the second by (2.1). Hence h is a reproducing kernel of $\mathcal{F}$ and $\mathcal{F}$ is a RKKS. The other direction is proven by Definition 8. □

[2]Basically every positive definite function is a reproducing kernel, and every reproducing kernel is a kernel, and every kernel is positive definite, so all three notions are exactly the same.

A consequence of the proof of the Moore-Aronszajn theorem (**Hein2004**) is that we can show that any function f in a RKHS $\mathcal{F}$ with kernel h can be written in the form $f(x) = \sum_{i=1}^{n} h(x, x_i)w_i$ for some $n \in \mathbb{N}$ (i.e. $\mathcal{F}$ is spanned by the functions $h(\cdot, x)$). More precisely, $\mathcal{F}$ is the completion of the space $\mathcal{G} = \text{span}\{h(\cdot, x) \,|\, x \in \mathcal{X}\}$ endowed with the inner product

$$\left\langle \sum_{i=1}^{n} w_i h(\cdot, x_i), \sum_{j=1}^{n} w_j h(\cdot, x_j) \right\rangle_{\mathcal{G}} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j h(x_i, x_j).$$

2.2 The Fisher information

Let Y be a random variable with density in the parameteric family $\{p(\cdot\,; f) | f \in \mathcal{F}\}$ with f belonging to a Hilbert space $\mathcal{F}$. If $p(Y; f) > 0$, the log-likelihood function of f is denoted $l(f|Y) = \log p(Y; f)$. Assuming existence, the score is defined as the gradient[3] $\nabla l(f|Y)$. The Fisher information $I[f] \in \mathcal{F} \otimes \mathcal{F}$ for $f \in \mathcal{F}$ is

$$I[f] = -\,\mathrm{E}[\nabla^2 l(f|Y)|f].$$

Specifically for our regression function as defined in (**??**) subject to f belonging to a RKHS, we can derive the Fisher information for f to be

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

where ψ_{ij} are the (i, j)-th entries of the precision matrix Ψ.

Proof. For $x \in \mathcal{X}$, let $k_x : \mathcal{F} \to \mathbb{R}$ be defined by $k_x(f) = \langle h(\cdot, x), f\rangle_{\mathcal{F}}$. By the reproducing property, $k_x(f) = f(x)$. The directional derivative of $k_x(f)$ in the direction g is

$$\begin{aligned}
\nabla_g k_x(f) &= \lim_{\delta \to 0} \frac{k(f + \delta g) - k(f)}{\delta} \\
&= \lim_{\delta \to 0} \frac{\langle h(\cdot, x), f + \delta g\rangle_{\mathcal{F}} - \langle h(\cdot, x), f\rangle_{\mathcal{F}}}{\delta} \\
&= \lim_{\delta \to 0} \frac{\delta \langle h(\cdot, x), g\rangle_{\mathcal{F}}}{\delta} = \langle h(\cdot, x), g\rangle_{\mathcal{F}}.
\end{aligned}$$

[3] Let $k : \mathcal{F} \to \mathbb{R}$. Denote the directional derivate of k in the direction g by $\nabla_g k$, that is,

$$\nabla_g k(f) = \lim_{\delta \to 0} \frac{k(f + \delta g) - k(f)}{\delta}.$$

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

$$\langle \nabla k(f), g\rangle_{\mathcal{F}} = \nabla_g k(f), \quad \forall f, g \in \mathcal{F}.$$

Thus, the gradient is $\nabla k_x(f) = h(\cdot, x)$ by definition. The log-likelihood of f is given by

$$l(f|y, \alpha, \Psi) = C - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi_{ij}\big(y_i - \alpha - k_{x_i}(f)\big)\big(y_j - \alpha - k_{x_j}(f)\big)$$

for some constant C, and the score by

$$\nabla l(f|y, \alpha, \Psi) = \sum_{i=1}^{n}\sum_{j=1}^{n}\psi_{ij}\big(y_i - \alpha - k_{x_i}(f)\big)\nabla k_{x_j}(f).$$

We can then calculate the Fisher information as

$$\begin{aligned} I[f] = -\operatorname{E}[\nabla^2 l(f|Y)|f] &= \operatorname{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\psi_{ij}\nabla k_{x_i}(f)\otimes\nabla k_{x_j}(f) \;\middle|\; f\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\psi_{ij}h(\cdot, x_i)\otimes h(\cdot, x_j). \end{aligned}$$

by substituting $\nabla k_x(f) = h(\cdot, x)$, the expectation is free of f

□

We can also compute the Fisher information for a linear functional of f, or between two linear functionals of f. We quote the following lemma (**Bergsma2014**):

Lemma 2.3 (Fisher information for linear functionals of elements in a Hilbert space)**.** *Let $\mathcal{F}$ be a Hilbert space. Denote the Fisher information for $f \in \mathcal{F}$ as $I[f]$. The Fisher information for $\langle f, g\rangle$ is given as*

$$I[\langle f, g\rangle_{\mathcal{F}}] = \langle I[f], g\otimes g\rangle_{\mathcal{F}\otimes\mathcal{F}}$$

and more generally, the Fisher information between $\langle f, g\rangle_{\mathcal{F}}$ and $\langle f, g'\rangle_{\mathcal{F}}$ is given as

$$I[\langle f, g\rangle_{\mathcal{F}}, \langle f, g'\rangle_{\mathcal{F}}] = \langle I[f], g\otimes g'\rangle_{\mathcal{F}\otimes\mathcal{F}}$$

The proof of Lemma 2.3 will not be shown here, but in involves the use of Parseval's identity in an inner product space. Using Lemma 2.3, we can derive the Fisher information for our regression function as defined in (**??**) subject to f belonging to a RKHS.

Corollary 2.3.1 (Fisher information for regression function)**.** *For our regression model as defined in* (**??**) *subject to f belonging to a RKHS $\mathcal{F}$, the Fisher information $I[f(x), f(x')]$*

is given by

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

Proof. Note that 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 2.3, we have

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

□

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 $||f||^2_{\mathcal{F}_n} = w^\top \Psi^{-1} w$.

2.3 The I-prior

For our linear model in (**??**) with f belonging to a RKHS $\mathcal{F}$ with kernel h over the set $\mathcal{X}$, define the subspace

$$\mathcal{F}_n = \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; f(x) = \sum_{i=1}^{n} h(x, x_i), \text{ for some } w_1, \ldots, w_n \in \mathbb{R}, \text{ and } x \in \mathcal{X} \right\}$$

for which the Fisher information exists. Effectively, our functions $f \in \mathcal{F}_n$ are parameterized by $w = (w_1, \ldots, w_n)^\top \in \mathbb{R}^n$, so we need only consider priors over $\mathbb{R}^n$. The entropy

of a prior π relative to a Lebesgue measure over $\mathbb{R}^n$ is defined as

$$\mathrm{H}(\pi) = -\int_{\mathbb{R}^n} \pi(w) \log \pi(w) \, \mathrm{d}w.$$

Maximising this entropy subject to a suitable constraint gives us the I-prior definition.

Definition 2.13 (I-prior)**. [Bergsma2014]**. Let $\mathcal{F}$ be a Krein space and let $Y \in \mathbb{R}^n$ be a random variable whose distribution depends on $f \in \mathcal{F}$. Denote the Fisher information for f by $I[f]$, and suppose it exists. For a given $f_0 \in \mathcal{F}$, let π be a probability distribution independent of Y such that $\mathrm{Cov}_\pi(f) = I_{f_0}[f]$. Then π is called an I-prior for f with hyperparameter f_0.

Definition 2.14 (I-prior)**.** A prior π for f for the linear model in (**??**) with f belonging to a RKHS $\mathcal{F}$ is called an I-prior if $\pi(\mathcal{F}_n) = 1$, and conditionally on $f \in \mathcal{F}_n$,

$$\pi = \arg\max \mathrm{H}(\pi) \quad \text{subject to} \quad \mathrm{E}_\pi \, ||f||^2_{\mathcal{F}_n} = 1.$$

The following theorem associates I-priors with the Fisher information.

Theorem 2.4 (I-prior for linear models is Gaussian with mean f_0 and covariance matrix the Fisher information)**. [Bergsma2014]**. *Consider the linear model in* (**??**) *with f belonging to a RKHS $\mathcal{F}$ with kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then an I-prior π for f is Gaussian with a hyperparameter f_0 (the prior mean) and covariance matrix as defined by*

$$\mathrm{Cov}_\pi(f(x), f(x')) = I[f(x), f(x')]$$

where

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

is the Fisher information for f, and ψ_{ij} is the (i, j)-th entry of the precision matrix Ψ of the errors. An I-prior for f will then have the random effect representation

$$f(x) = f_0(x) + \sum_{i=1}^{n} h(x, x_i) w_i$$
$$(w_1, \ldots, w_n) \sim \mathrm{N}(0, \Psi).$$

For convenience, we can write the I-prior for f in the more compact matrix notation

$$\mathbf{f} = \mathbf{f}_0 + \mathbf{H}\mathbf{w}$$
$$\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Psi})$$

where $\mathbf{H}$ is the $n \times n$ symmetric kernel matrix whose (i, j)-th entries contain $h(x_i, x_j)$, for $i, j = 1, \ldots, n$.

For the model defined in (**??**), an I-prior on f is a Gaussian distribution with prior mean f_0 and covariance matrix equal to the Fisher information for f. For this model, the Fisher information does not depend on f_0 and can be simply written as $I[f]$. We can also write the I-prior for f in a random effect representation, given by the following theorem:

Theorem 2.5 (I-prior for linear models)**. [Bergsma2014]**. *For the linear regression model stated in* (**??**)*, let $\mathcal{F}$ be the RKKS over $\mathcal{X}$ with kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The Fisher information $I[f] \in \mathcal{F} \otimes \mathcal{F}$ for f is given by*

$$I[f](\mathbf{x}_i, \mathbf{x}_i') = \sum_{k=1}^{n} \sum_{l=1}^{n} \psi_{kl} h(\mathbf{x}_i, \mathbf{x}_k) h(\mathbf{x}_i', \mathbf{x}_l)$$

where ψ_{kl} is the (k, l)-th entry of the precision matrix $\mathbf{\Psi}$ of the errors. Denote by π be the Gaussian distribution mean f_0 and covariance kernel $I[f]$. Then by definition, π is an I-prior for f. Thus, a random vector $f \sim \pi$ will have the covariance matrix as defined by $\mathrm{Cov}_\pi(f(\mathbf{x}_i), f(\mathbf{x}_i')) = I[f](\mathbf{x}_i, \mathbf{x}_i')$, and that the I-prior for f will have the random effect representation

$$f(\mathbf{x}_i) = f_0(\mathbf{x}_i) + \sum_{k=1}^{n} h(\mathbf{x}_i, \mathbf{x}_k) w_k$$
$$(w_1, \ldots, w_n) \sim N(\mathbf{0}, \mathbf{\Psi}).$$

For convenience, we can write the I-prior for f in the more compact matrix notation

$$\mathbf{f} = \mathbf{f}_0 + \mathbf{H}\mathbf{w}$$
$$\mathbf{w} \sim N(\mathbf{0}, \mathbf{\Psi})$$

where the boldface $\mathbf{f}$ represents the vector of functional evaluations $(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$, and $\mathbf{H}$ is the symmetric kernel matrix whose (i, j)-th entries contain $h(\mathbf{x}_i, \mathbf{x}_j)$.

The proof for this theorem can be found in **Bergsma2014** The prior mean f_0 is a hyperparameter of the I-prior model, and can be given a fixed value such as 0. **Bergsma2014** derives the closed form expression for the posterior distribution of the I-prior regression function f, for which the posterior mean is used as an estimate. An EM algorithm can be employed to find the maximum likelihood estimators of the hyperparameters of the I-prior model, or alternatively the random effects can be integrated out and the marginal likelihood maximised directly. These consist of the intercept α, the error precision Ψ, and any other parameters that the kernel may depend on (more on this in Section 2.4.1). While the intercept can be viewed as being part of the regression function f (technically, it would be a function in the RKHS of constant functions), practically it is much easier to treat it as a separate fixed parameter to be estimated. Hence the reason for segregating the intercept from the regression function in our models thus far.

2.3.1 Example of I-prior modelling: Multiple regression

Now let us take a look at an example of regression modelling with I-priors on the familiar standard linear model as described in (**??**). For this model, we can compute the Fisher information for the regression coefficients $\boldsymbol{\beta}$, by twice differentiating the log-likelihood function and taking negative expectations. This is found to be

$$I[\boldsymbol{\beta}] = \psi \mathbf{X}^\top \mathbf{X}.$$

Thus, an I-prior for $\boldsymbol{\beta}$ with prior mean $\boldsymbol{\beta}_0$ is

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\beta}_0, \psi \mathbf{X}^\top \mathbf{X}).$$

An equivalent way of writing this I-prior would be

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{\beta}_0 + \mathbf{X}^\top \mathbf{w} \\ \mathbf{w} &\sim \mathrm{N}(\mathbf{0}, \psi \mathbf{I}_n) \end{aligned}$$

where $\mathbf{w} = (w_1, \ldots, w_n)$ are the so called I-prior random effects as described in the second part of Theorem 2.5 above. Substituting the above back into model (**??**) we arrive at the I-prior random effects representation

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\alpha} + \overbrace{\mathbf{X}\boldsymbol{\beta}_0}^{\mathbf{f}_0} + \overbrace{\mathbf{X}\mathbf{X}^\top \mathbf{w}}^{\mathbf{Hw}} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \mathrm{N}(\mathbf{0}, \psi^{-1} \mathbf{I}_n) \\ \mathbf{w} &\sim \mathrm{N}(\mathbf{0}, \psi \mathbf{I}_n). \end{aligned} \tag{2.3}$$

Remark 1. The multiple regression model relates to the I-prior methodology by considering the regression function $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$, for some $\boldsymbol{\beta} \in \mathbb{R}^p$. Lemma **??** tells us the form of the Fisher information for f, while Theorem 2.5 sets the I-prior for f as Gaussian with prior mean f_0 and covariance matrix the Fisher information. Deriving the I-prior this way gives similar results to the above.

2.4 Kernel functions

2.4.1 Other commonly used models: A toolbox of kernels

In the above multiple regression example, the regression function are straight line functions over the set of reals. This is a reproducing kernel space with the Euclidean space inner product/dot product as its kernel, i.e. $h(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i \cdot \mathbf{x}_j$, which is what makes up the entries of $\mathbf{H}$ in (2.3). This is known as the Canonical kernel (**Bergsma2014**). As it turns out, exchanging this canonical kernel with a different kernel, hence a different RKHS of functions, we can perform various types of modelling.

Some commonly used models can be achieved using the following kernels:

Type	Description of $\mathcal{X}$=$\{x_i\}$	Name of space	Kernel $h(x_i, x_j)$
Nominal	1) Categorical covariates; 2) In a multilevel setting, x_i = group no. of unit i.	Pearson	$\frac{\mathbb{1}[x_i=x_j]}{p_i} - 1$ where $p_i = ¶[X = x_i]$
Real	In a classical regression setting, x_i = covariate associated with unit i.	Canonical / Centred Canonical	$x_i x_j$ / $x_i x_j - \bar{x}$
Real	In 1-dim smoothing, x_i = data point associated with observation y_i.	Fractional Brownian Motion (FBM)	$\lvert x_i\rvert^{2\gamma} + \lvert x_j\rvert^{2\gamma} - \lvert x_i - x_j\rvert^{2\gamma}$ with Hurst index $\gamma \in (0, 1)$

Table 2.1: A toolbox of kernels - Names and descriptions of some useful RKHS of functions.

Remark 2. The origin of a Hilbert space over a set $\mathcal{X}$ may be arbitrary, in which case a centering may be appropriate. Hence, the centred Canonical kernel.

New reproducing kernel spaces can be constructed from existing ones. An example is the so-called ANOVA kernel constructed from a Canonical and Pearson kernel applied to the two-dimensional vector $\mathbf{x}_i = (x_{i1}, x_{i2})$, where the first component is real-valued, and the second component consists of nominal values. The ANOVA kernel is constructed as

$$h(\mathbf{x}_i, \mathbf{x}_j) = h_1(x_{i1}, x_{j1}) + h_2(x_{i2}, x_{j2}) + h_1(x_{i1}, x_{j1})h_2(x_{i2}, x_{j2}).$$

This kernel is particularly useful to model interaction effects. Take for example a random slope model. The effect of a covariate is assumed to be different for each group. This can be thought of as having an interaction present between the real-valued covariate x_{i1} and the grouping x_{i2}, which is captured by the product of the two kernels $h_1 h_2$.

Remark 3. We are able to circumvent the positive definite restriction of inner products (and kernels which define them in the reproducing kernel space) by working in a Krein space, and hence a reproducing kernel Krein space (RKKS). Krein spaces generalise Hilbert spaces by dropping the positive-definiteness requirement of inner products. Inner products may turn out to be not positive definite when scale parameters for the space are considered (which may be negative) and new kernels are constructed by way of adding and multiplying kernels together, as in the ANOVA kernel above. For a review of RKKSs, see **alpay1991** and **Ong2004** RKKSs are actively being researched, and is out of the scope of this paper for now.

2.4.2 The RKHS scale parameter

The scale of an RKHS $\mathcal{F}$ over a set $\mathcal{X}$ with kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ may be arbitrary. To resolve this, a scale parameter $\lambda \in \mathbb{R}$ is introduced, resulting in the RKHS denoted

by $\mathcal{F}_\lambda$ with kernel $h_\lambda = \lambda h$. This results in at most p scale parameters $\lambda_1, \ldots, \lambda_p$ - one for each of the function space over the set of p covariates. If there are several covariates which are known to be measured on the same scale, e.g. repeated measures of weight, then these may share the same scale parameter (technically, the same RKHS $\mathcal{F}_\lambda$).

Remark 4. For the ANOVA kernel described above, there are two possible ways of introducing scale parameters. Since the ANOVA kernel is constructed from two existing kernels, the Canonical kernel h_1 and Pearson kernel h_2, each with their own scale parameter λ_1 and λ_2 respectively, then the interaction effect or the product between the two kernels has the scale parameter equal to the product of the two scale parameters $\lambda_1\lambda_2$. This is the more parsimonious method. Another valid way is to introduce a separate scale parameter for the interactions, λ_{12} say. This is the less parsimonious method, and in this case, there will be at most $p(p-1)/2$ scale parameters when a model with p covariates and all its two-way interactions are considered.

In the example of multiple regression in Section 2.3.1, the canonical kernel with scale parameters $\lambda_1, \ldots, \lambda_p$ can be written as

$$\mathbf{H} = \mathbf{H}_{\boldsymbol{\lambda}} := \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^\top$$

where $\boldsymbol{\Lambda} = \operatorname{diag}[\lambda_1, \ldots, \lambda_p]$. This corresponds to the scaled Canonical RKHS with kernel $h_\lambda(\mathbf{x}_i, \mathbf{x}_j) = \lambda_1 x_{i1}x_{j1} + \cdots + \lambda_p x_{ip}x_{jp}$, and the covariance matrix for the I-prior on $\boldsymbol{\beta}$ is adjusted to be $\psi\boldsymbol{\Lambda}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Lambda}$.

Proof. In the I-prior method, our model is $\mathbf{y} = \boldsymbol{\alpha} + \mathbf{H}\mathbf{w} + \boldsymbol{\epsilon}$ where f_0 has been assumed to be zero for simplicity. Replacing the canonical kernel matrix $\mathbf{H}$ with the scaled canonical kernel matrix, we have

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\overbrace{\boldsymbol{\Lambda}\mathbf{X}^\top\mathbf{w}}^{\boldsymbol{\beta}} + \boldsymbol{\epsilon}$$

with $\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \psi\mathbf{I}_n)$. Equivalently, $\boldsymbol{\beta}$ is normally distributed with mean and variance

$$\mathrm{E}\,\boldsymbol{\beta} = \mathrm{E}[\boldsymbol{\Lambda}\mathbf{X}^\top\mathbf{w}] = \mathbf{0}$$
and
$$\operatorname{Var}\boldsymbol{\beta} = \operatorname{Var}[\boldsymbol{\Lambda}\mathbf{X}^\top\mathbf{w}] = \psi\boldsymbol{\Lambda}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Lambda}.$$

□

2.5 Comparison to Gaussian process priors

Key differences:

1. Typically no scale parameter is estimated for the kernels in GPR. Instead, the x and y variables are centred *and* scaled before estimating. New data points are then centred and scaled on the mean and s.d. of the training points.
2. GPR not usually interested in estimating the error precision ψ.
3. The "go-to" kernel is the squared exponential kernel or Gaussian radial basis function defined as
$$k(x, x') = \exp(-\sigma\|x - x'\|^2)$$
σ usually chosen by cross-validation or grid-search methods.

Why do we need to estimate scale parameters and error precision in I-prior models?

2.5.1 The Bayesian connection

The I-prior methodology is less of a fully Bayesian approach and more of an empirical-Bayes approach, whereby an objective using the Fisher information as the covariance matrix of the prior is used to estimate the parameters of the model through maximisation of the likelihood, set up in a RKHS paradigm. However, the I-prior methodology is still this notion of priors and posteriors, something which is arguably Bayesian. Recall the standard linear regression model with independent errors:

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \psi^{-1}\mathbf{I}_n).$$

The I-prior method transformed this model into the random effect representation with kernels that we saw earlier in Section 2.3.1. However, by simply taking the fundamental idea of I-priors, which is a prior with the covariance matrix equal to the Fisher information, nothing is really stopping us from estimating this model fully Bayes. We simply need to assign further priors on the intercept and precision, such as

$$\underline{\text{Priors}}$$
$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \psi\boldsymbol{\Lambda}\mathbf{X}^\top\mathbf{X}\boldsymbol{\Lambda})$$
$$\alpha \sim \mathrm{N}(0, 1000)$$
$$\psi, \lambda_1^{-2}, \ldots, \lambda_p^{-2} \sim \Gamma(0.001, 0.001).$$

Here, an I-prior with mean zero is chosen. The choices of normal for α, and gamma for the scale parameters ψ and a reparameterization of the λs is chosen for conjugacy convenience. In the absence of any prior knowledge about the parameters, it is reasonable to choose such hyperparameters to make the priors quite flat and uninformative. Another choice of uninformative prior for the scale parameters would be the **Jeffreys1946**' prior, which is in fact the limit of the gamma distribution as both hyperparameters approach zero. An MCMC approach such as Gibbs or Metropolis-Hastings sampling is then able

to estimate this model, and software such as WinBUGS or JAGS are then able to be used.

The main motivation behind I-priors was to guard against over-fitting in cases where model dimensionality is very large relative to sample size. A prior is devised based on an objective principle (of maximum entropy) which brings about simpler estimation while requiring minimal assumptions, as well as model parsimony. A maximum likelihood approach is used to fit I-prior models, which give promising results in terms of predictive abilities from the simulations conducted. In the next section, I-priors will be discussed with a more Bayesian connotation, applied to Bayesian variable selection.

List of Figures

List of Tables

List of Theorems

List of Definitions

List of Symbols

$\mathrm{N}_p(\mu, \Sigma)$	p-dimensional multivariate normal distribution with mean vector μ and covariance Σ.
$\sim$	Is distributed as.
$\otimes$	The tensor product.

Index