To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
February 28, 2018

Chapter 3

Fisher information and the I-prior

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

Following this, we shall discuss the notion of prior distributions for regression functions, and how one might assign a suitable prior. In our case, we choose an objective prior following (Jaynes, 1957a; Jaynes, 1957b)—in the absence of any prior knowledge, a prior distribution which maximises entropy should be used. It turns out, the entropy maximising prior for f is Gaussian with mean chosen a priori and covariance kernel proportional to the Fisher information. Such a distribution on f is called the I-prior distribution.

3.1 The traditional Fisher information

3.2 Fisher information for Hilbert space objects

We extend the idea beyond thinking about parameters as merely numbers in the usual sense, to abstract objects in Hilbert spaces. This generalisation allows us to extend the concept of Fisher information to regression functions in RKHSs later. The score and Fisher information is derived in a familiar manner, but extra care is required when taking derivatives with respect to Hilbert space objects.

Let Y be a random variable with density in the parametric family $\{p(\cdot|\theta) \mid \theta \in \Theta\}$, where Θ is now assumed to be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_\Theta$. If $p(Y|\theta) > 0$, the log-likelihood function of θ is the real-valued function $L(\cdot|Y) : \Theta \to \mathbb{R}$ defined by $\theta \mapsto \log p(Y|\theta)$. To discuss derivatives of the log-likelihood function for $\theta \in \Theta$, we require a generalisation of the concept of differentiability from real-valued functions of a single, real variable, as is common in calculus, to functions between Banach spaces.

1. Why wouldn't it be >0 ?

Definition 3.1 (Fréchet derivative)**.** Let $\mathcal{V}$ and $\mathcal{W}$ be two normed spaces, and $\mathcal{U} \subseteq \mathcal{V}$ be an open subset. A function $f : \mathcal{U} \to \mathcal{W}$ is called *Fréchet differentiable* at $x \in \mathcal{U}$ if there exists a bounded, linear operator $T : \mathcal{V} \to \mathcal{W}$ such that

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

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

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

The intuition here is similar to that of regular differentiability, that the linear operator T well approximates the change in f at x (the numerator), relative to the change in x (the denominator)—the fact that the limit exists and is zero, it must mean that the numerator converges faster to zero than the denominator does. In Landau notation, we have the familiar expression $f(x+h) = f(x) + \mathrm{d}f(x)h + o(h)$, that is, the tangent line to f at x gives the best linear approximation to f near x. The limit in the definition is meant in the usual sense of convergence of functions with respect to the norms of $\mathcal{V}$ and $\mathcal{W}$. Of course, we may use Fréchet derivatives in Hilbert spaces too by using the inner product norm of the space.

For the avoidance of doubt, $\mathrm{d}f(x)$ is not a vector in $\mathcal{W}$, but is an element of the set of bounded, linear operators from $\mathcal{V}$ to $\mathcal{W}$, denoted $\mathrm{L}(\mathcal{V}, \mathcal{W})$. That is, if $f : \mathcal{U} \to \mathcal{W}$ is a differentiable function at all points in $\mathcal{U} \subseteq \mathcal{V}$, then its derivative is a linear map

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

It follows that this function may also have a derivative, which by definition will be a linear map as well:

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

The space on the righthand side is identified with the Banach space $\mathrm{L}(\mathcal{V} \times \mathcal{V}, \mathcal{W})$ of all continuous bilinear maps from $\mathcal{V}$ to $\mathcal{W}$. In other words, an element $\phi \in \mathrm{L}\big(\mathcal{V}, \mathrm{L}(\mathcal{V}, \mathcal{W})\big)$ is identified with $\psi \in \mathrm{L}(\mathcal{V} \times \mathcal{V}, \mathcal{W})$ such that for all $x, y \in \mathcal{V}$, $\phi(x)(y) = \psi(x, y)$. Simply put, a function ϕ linear in x with $\phi(x)$ linear in y is the same as a bilinear function ψ in x and y. The second derivative $\mathrm{d}^2 f(x)$ is therefore a bounded, bilinear operator from $\mathcal{V} \times \mathcal{V}$ to $\mathcal{W}$.

Another closely related type of differentiability is the concept of *Gâteaux differentials*, which is the formalism of functional derivatives in calculus of variations. Let $\mathcal{V}$, $\mathcal{W}$ and $\mathcal{U}$ be as before, and consider the function $f : \mathcal{U} \to \mathcal{W}$.

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

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

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

Remark 3.2. The space $(\mathcal{V}, \mathcal{W})$ of operators from $\mathcal{V}$ to $\mathcal{W}$ is not a topological space, and there is no obvious way to define a topology on it. Consequently, we cannot consider the Gâteaux derivative of the Gâteaux derivative. Furthermore, unlike the Fréchet derivative, which is by definition a linear operator, the Gâteaux derivative may fail to satisfy the additive condition of linearity[1]. Finally, even if it is linear, it may fail to depend continuously on v if $\mathcal{V}$ and $\mathcal{W}$ are infinite dimensional.

Nevertheless, the reason we bring up Gâteaux differentials is that it may motivate higher-order Fréchet differentials. First note the connection between the two, by again

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

considering the function $f : \mathcal{U} \to \mathcal{W}$.

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

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

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

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

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

$$\left\| \frac{\mathrm{d}f(x+tv)(v') - \mathrm{d}f(x)(v')}{t} - \mathrm{d}^2 f(x)(v, v') \right\|_{\mathcal{W}} \to 0 \text{ as } t \to 0, \tag{3.1}$$

by a similar argument in the proof above. We will use this fact when we describe the Hessian in a little while.

There is also the concept of *gradients* in Hilbert space. Recall that the Riesz representation theorem says that the mapping $A : \mathcal{V} \to \mathcal{V}'$ from the Hilbert space $\mathcal{V}$ to its continuous dual space $\mathcal{V}'$ defined by $A = \langle \cdot, v \rangle_{\mathcal{V}}$ for some $v \in \mathcal{V}$ is an isometric isomorphism. Again, let $\mathcal{U} \subseteq \mathcal{V}$ be an open subset, and let $f : \mathcal{U} \to \mathbb{R}$ be a (Fréchet) differentiable function with derivative $\mathrm{d}f : \mathcal{U} \to \mathrm{L}(\mathcal{V}, \mathbb{R}) \equiv \mathcal{V}'$. We define the gradient as follows.

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

$\nabla f(x)$, is the unique element of $\mathcal{V}$ satisfying

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

for any $v \in \mathcal{V}$. Note that ∇f being a composition of two continuous functions, is itself continuous.

Since the gradient of f is an operator on $\mathcal{U}$ to $\mathcal{V}$, it may itself have a (Fréchet) derivative. Assuming existence, i.e., f is twice Fréchet differentiable at $x \in \mathcal{U}$, we call this derivative the *Hessian* of f. From (3.1), it must be that

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

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

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

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

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

Remark 3.3. Since $\mathrm{d}^2 f(x)$ is a bilinear form in $\mathcal{V}$, we can equivalently write

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

following the correspondence between bilinear forms and tensor product spaces.

We can now define the score S, assuming existence, as the (Fréchet) derivative of $L(\cdot|Y)$, i.e. $S : \Theta \to \mathrm{L}(\Theta, \mathbb{R}) \equiv \Theta'$ defined by $S = \mathrm{d}L(\cdot|Y)$. The second (Fréchet) derivative of $L(\cdot|Y)$ is then $\mathrm{d}^2 L(\cdot|Y) : \Theta \to \mathrm{L}(\Theta \times \Theta, \mathbb{R})$. The Fisher information $\mathcal{I}(\theta)$

at $\theta \in \Theta$ is defined to be

$$\mathcal{I}(\theta) = -\operatorname{E}[\mathrm{d}^2 L(\theta|Y)] \in \Theta \otimes \Theta.$$

or alternatively

2. Is this required?

$$\begin{aligned}
\mathcal{I}(\theta) &= \operatorname{E}[\mathrm{d}L(\theta|Y) \otimes \mathrm{d}L(\theta|Y)] \\
&= \operatorname{E}[\langle \nabla L(\theta|Y), \cdot \rangle_\Theta \otimes \langle \nabla L(\theta|Y), \cdot \rangle_\Theta] \\
&= \operatorname{E}\langle \nabla L(\theta|Y) \otimes \nabla L(\theta|Y), \cdot \rangle_{\Theta\otimes\Theta} \\
&= \langle \operatorname{E}[\nabla L(\theta|Y) \otimes \nabla L(\theta|Y)], \cdot \rangle_{\Theta\otimes\Theta}.
\end{aligned}$$

Since $\mathcal{I}(\theta) \in \Theta \otimes \Theta$ we may view it also as a bilinear form. That is, for any $b, b' \in \Theta$, we have

$$\mathcal{I}(\theta)(b, b') = \langle \mathcal{I}(\theta), b \otimes b' \rangle_{\Theta\otimes\Theta}. \tag{3.2}$$

We call this the Fisher information for θ evaluated at two points b and b' in Θ. Setting $\theta_b = \langle \theta, b \rangle_\Theta$ for some $b \in \Theta$, we may view this also as the Fisher information between two continuous, linear functionals of θ.

3. This part seems sketchy.

3.3 Fisher information for regression functions

We are now equipped to derive the Fisher information for our regression function. For convenience, we restate the regression model and its assumptions. The regression model relating response variables $y_i \in \mathbb{R}$ and the covariates $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$, for $i = 1, \dots, n$ is

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

subject to

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

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

Lemma 3.2 (Fisher information for regression function)**.** *For the regression model stated in* (1.1) *subject to* (1.2) *and $f \in \mathcal{F}$ where $\mathcal{F}$ is an RKHS with kernel h, the Fisher* *information for f is given by*

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

where ψ_{ij} are the (i,j)-th entries of the precision matrix $\mathbf{\Psi}$ of the normally distributed model errors. More generally, suppose that $\mathcal{F}$ has a feature space $\mathcal{V}$ such that the mapping $\phi : \mathcal{X} \to \mathcal{V}$ is its feature map, and if $f(x) = \langle \phi(x), v \rangle_{\mathcal{V}}$, then the Fisher information $I(v) \in \mathcal{V} \otimes \mathcal{V}$ for v is

$$\mathcal{I}(v) = \sum_{i=1}^{n}\sum_{j=1}^{n} \psi_{ij} \phi(x_i) \otimes \phi(x_j).$$

Proof. For $x \in \mathcal{X}$, let $k_x : \mathcal{V} \to \mathbb{R}$ be defined by $k_x(v) = \langle \phi(x), v \rangle_{\mathcal{V}}$. Clearly, k_x is linear and continuous. Hence, the Gâteaux derivative of $k_x(v)$ in the direction u is

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

Since clearly $\partial_u k_x(v)$ is a continuous linear operator for any $u \in \mathcal{V}$, it is bounded, so the Fréchet derivative exists and $\mathrm{d}k_x(v) = \partial k_x(v)$. Thus, the gradient is $\nabla k_x(v) = \phi(x)$ by definition. Let $\mathbf{y} = \{y_1, \dots, y_n\}$, and denote the hyperparameters of the regression model by $\boldsymbol{\theta} = \{\alpha, \mathbf{\Psi}, \eta\}$. Without loss of generality, assume $\alpha = 0$; otherwise we can always add back α to y later. The log-likelihood of v is given by

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

and the score by

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

Differentiating again gives

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

since the derivative of $\mathrm{d}k_x$ is zero as it is the derivative of a constant. We can then calculate the Fisher information to be

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

Here, we had treated $\phi(x_i) \otimes \phi(x_j)$ as a bilinear operator, since $\mathcal{I}(v) \in \mathcal{V} \otimes \mathcal{V}$ as well.

By taking the canonical feature $\phi(x) = h(\cdot, x)$, we have that $\phi \equiv h(\cdot, x) : \mathcal{X} \to \mathcal{F} \equiv \mathcal{V}$ and therefore for $f \in \mathcal{F}$, the reproducing property gives us $f(x) = \langle h(\cdot, x), f \rangle_{\mathcal{F}}$, so the formula for $\mathcal{I}(f) \in \mathcal{F} \otimes \mathcal{F}$ follows. $\square$

The above lemma gives the form of the Fisher information for f in a rather abstract fashion. Consider the following example of applying Lemma (3.2) to obtain the Fisher information for a standard linear regression model.

Example 3.1 (Fisher information for linear regression)**.** As before, suppose model (1.1) subject to its assumptions hold. For simplicity, we assume iid errors, i.e. $\boldsymbol{\Psi} = \psi \mathbf{I}_n$. Let $\mathcal{X} = \mathbb{R}^p$, and the feature space $\mathcal{V} = \mathbb{R}^p$ be equipped with the usual dot product $\langle \cdot, \cdot \rangle_{\mathcal{V}} : \mathcal{V} \otimes \mathcal{V} \to \mathbb{R}$ defined by $v^\top v$. Consider also the feature map $\phi : \mathcal{X} \to \mathcal{V}$ defined by $\phi(x) = x$. For some $\beta \in \mathcal{V}$, the linear regression model is such that $f(x) = x^\top \beta =$

$\langle\phi(x),\beta\rangle_{\mathcal{V}}$. Therefore, according to Lemma (3.2), the Fisher information for β is

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

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

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

Corollary 3.2.1 (Fisher information between two linear functionals of the regression function)**.** *For our regression model as defined in* (1.1) *subject to* (1.2) *and f belonging to a RKHS $\mathcal{F}$ with kernel h, the Fisher information between two points $f(x)$ and $f(x')$ is given by*

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

Proof. In a RKHS $\mathcal{F}$, the reproducing property gives $f(x) = \langle f, h(\cdot, x)\rangle_{\mathcal{F}}$ and in particular, $\langle h(\cdot, x), h(\cdot, x')\rangle_{\mathcal{F}} = h(x, x')$. By (3.2), we have that

$$
\begin{aligned}
\mathcal{I}\big(f(x), f(x')\big) &= \mathcal{I}\big(\langle f, h(\cdot, x)\rangle_{\mathcal{F}}, \langle f, h(\cdot, x')\rangle_{\mathcal{F}}\big) \\
&= \big\langle \mathcal{I}(f), h(\cdot, x) \otimes h(\cdot, x') \big\rangle_{\mathcal{F}\otimes\mathcal{F}} \\
&= \left\langle \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(\cdot, x_i) \otimes h(\cdot, x_j) \ , \ h(\cdot, x) \otimes h(\cdot, x') \right\rangle_{\mathcal{F}\otimes\mathcal{F}} \\
&= \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} \big\langle h(\cdot, x_i), h(\cdot, x) \big\rangle_{\mathcal{F}} \big\langle h(\cdot, x_j), h(\cdot, x') \big\rangle_{\mathcal{F}} \\
&= \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(x, x_i) h(x', x_j).
\end{aligned}
$$

The second to last line follows from the definition of the usual inner product for tensor spaces, and the last line follows by the reproducing property. □

An inspection of the formula in Corollary (3.2.1) reveals the fact that the Fisher information for $f(x)$ is positive if and only if $h(x, x_i) \neq 0$ for at least one $i \in \{1, \ldots, n\}$. In practice, this condition is often satisfied for all x, so this result might be considered both remarkable and reassuring, because it suggests we can estimate f over its entire domain, no matter how big, even though we only have a finite amount of data points.

4. Rewrite

3.4 The induced Fisher information RKHS

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

5. Rewrite

$$\mathcal{F}_n = \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; f(x) = \sum_{i=1}^{n} h(x, x_i) w_i, \; w_i \in \mathbb{R}, \; i = 1, \ldots, n \right\}. \tag{3.3}$$

Since $h(\cdot, x_i) \in \mathcal{F}$, then any $f \in \mathcal{F}_n$ is also in $\mathcal{F}$ by linearity, and thus $\mathcal{F}_n$ is a subset of $\mathcal{F}$. Further, $\mathcal{F}_n$ is closed under addition and multiplication by a scalar, and is therefore a subspace of $\mathcal{F}$. Let $\mathcal{F}_n^\perp$ be the orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$. Then, any $r \in \mathcal{F}_n^\perp$ is orthogonal to each of the $h(\cdot, x_i)$, so by the reproducing property of h, $r(x_i) = \langle r, h(\cdot, x_i) \rangle_{\mathcal{F}} = 0$.

Corollary 3.2.2. *With $g \in \mathcal{F}$, the Fisher information for g is zero if and only if $g \in \mathcal{F}_n^\perp$, i.e. if and only if $g(x_1) = \cdots = g(x_n) = 0$.*

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

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

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

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

where $\mathbf{w} = (w_1, \ldots, w_n)$ *and* $f_w(x) = \sum_{i=1}^n h(x, x_i) w_i$*. Then,* h_n *defined by*

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

is the reproducing kernel of $\mathcal{F}_n$*.*

Proof. Prove $\mathcal{F}_n$ is a Hilbert space?

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

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

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

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

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

8. How?

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

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

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

3.5 The I-prior

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

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

The set $\mathcal{F}$ is potentially "too big" for the purpose of estimating f, that is, for certain pairs of functions $\mathcal{F}$, the data do not allow an assessment of whether one is closer to the truth than the other. In particular, the data do not contain information to distinguish betwen any f and f' for which $f(x_i) = f'(x_i), i = 1, \dots, n$. A prior for f therefore need not have support $\mathcal{F}$, instead it is sufficient to consider priors with support $f_0 + \mathcal{F}_n$, where $f_0 \in \mathcal{F}$ is fixed and chosen a priori as a "best guess" of f. Since the Fisher information for $\langle g, f \rangle_{\mathcal{F}}$ is non-zero for any non-zero $g \in \mathcal{F}_n$, there is information to allow a comparison between any pair of functions in $f_0 + \mathcal{F}_n$.

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

Key questions:

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

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

Definition 3.5 (Entropy)**.** The entropy of a distribution π over $\mathcal{F}$ relative to a measure ν is defined as

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

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

Definition 3.6 (Functional derivative)**.** Given a manifold M representing continuous/smooth functions ρ with certain boundary conditions, and a functional $F : M \to \mathbb{R}$, the functional derivative of $F[\rho]$ with respect to ρ, denoted $\partial F/\partial\rho$, is defined by

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

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

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

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

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

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

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

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

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

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

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

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

subject to the constraints

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

is the density given by

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

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

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

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

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

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

Set this to zero, and solve for p:

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

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

Theorem 3.5 (The I-prior)**.** *Let $\mathcal{F}$ be an RKHS with kernel h, and consider the finite dimensional affine subspace $\mathcal{F}_n$ of $\mathcal{F}$ equipped with an inner product as in Lemma 2.5. Let ν be a volume measure induced by the norm $\|\cdot\|_{\mathcal{F}_n} = \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{F}_n}}$. With $f_0 \in \mathcal{F}$, let Π_0*

be the class of distributions p such that

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

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

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

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

Proof. Recall the fact that any $f \in \mathcal{F}$ can be decomposed into $f = f_n + r_n$, with $f_n \in \mathcal{F}_n$ and $r_n \in \mathcal{R}_n$, the orthogonal complement of $\mathcal{F}_n$. Also recall that there is no Fisher information about any $r \in \mathcal{R}_n$, and therefore it is not possible to estimate r_n from the data. Therefore, $p(r_n) = 0$, and one needs only consider distributions over $\mathcal{F}_n$ when building distributions over $\mathcal{F}$.

The norm on $\mathcal{F}_n$ induces the metric $d(f, f') = \|f - f'\|_{\mathcal{F}_n}$. Thus, for $f \in \mathcal{F}$ of the form $f = \sum_{i=1}^n h(\cdot, x_i) w_i$ (i.e., $f \in \mathcal{F}_n$) and provided $f_0 \in \mathcal{F}_n \subset \mathcal{F}$,

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

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

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

This implies that f is Gaussian, since

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

is a sum of normal random variables, and therefore $\langle f, f' \rangle_{\mathcal{F}}$ is normally distributed for any $f' \in \mathcal{F}$. The mean $\mu \in \mathcal{F}$ of this random vector f satisfies $\mathrm{E}\langle f, f' \rangle_{\mathcal{F}} = \langle \mu, f' \rangle_{\mathcal{F}}$ for all $f' \in \mathcal{F}_n$, but

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

so $\mu \equiv f_0 = \sum_{i=1}^{n} h(\cdot, x_i) w_{i0}$. The covariance kernel Σ is the bilinear form satisfying

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

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

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

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

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

while the second term on the left hand side is

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

Thus,

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

the reproducing kernel for $\mathcal{F}_n$. $\square$

3.6 Rate of convergence

3.7 Conclusion

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

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

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

Bibliography

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

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

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

List of Figures

List of Tables

List of Theorems

List of Definitions