To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
February 11, 2018

Chapter 3

Fisher information and the I-prior

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

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

3.1 The traditional Fisher information

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

Suppose Y is a random variable whose density function $p(\cdot|\theta)$ depends on the parameter θ. Write the log-likelihood function of θ as $L(\theta) = \log p(Y|\theta)$, and the gradient function of the log-likelihood (the *score function*) as $S(\theta) = \partial L(\theta)/\partial\theta$. The *Fisher information* about a the parameter θ is defined to be expectation of the second moment

of the score function,

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log p(Y|\theta)\right)^2\right].$$

Here, expectation is taken with respect to the random variable Y under its true distribution. Under certain regularity conditions, it can be shown that $\mathrm{E}[S(\theta)] = 0$, and thus the Fisher information is in fact the variance of the score function, since $\mathrm{Var}[S(\theta)] = \mathrm{E}[S(\theta)^2] - \mathrm{E}^2[S(\theta)]$. Further, if $\log p(Y|\theta)$ is twice differentiable with respect to θ, then it can be shown that under certain regularity conditions,

$$\mathcal{I}(\theta) = \mathrm{E}\left[-\frac{\partial^2}{\partial\theta^2}\log p(Y|\theta)\right].$$

Many textbooks provides a proof of this fact—see, for example, Wasserman (2013).

From the last equation above, we see that the Fisher information is related to the curvature or concavity of the graph of the log-likelihood function, averaged over the random variable Y. The curvature, defined as the second derivative on the graph[1] of a function, measures how quickly the function changes with changes in its input values. This then gives an intuition regarding the uncertainty surrounding θ at its maximal value; high Fisher information is indicative of a sharp peak at the maxima and therefore small variance, while low Fisher information is indicative of a shallow maxima for which many θ share similar log-likelihood values.

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

3.2 Fisher information for Hilbert space objects

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

Let Y be a random variable with density in the parametric family $\{p(\cdot|\theta)\,|\,\theta \in \Theta\}$, where Θ is assumed to be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_\Theta$. If $p(Y|\theta) > 0$, the log-likelihood function of θ is denoted $L(\theta) = \log p(Y|\theta)$. The score and Fisher

2. Why wouldn't it be >0 ?

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

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

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

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

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

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

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

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

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

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

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

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

which corresponds to the idea of directational derivatives.

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

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

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

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

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

or equivalently,

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

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

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

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

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

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

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

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

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

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

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

Differentiating again gives

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

Note that by the bilinear property of tensor products,

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

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

3.3 Fisher information for regression functions

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

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

subject to

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

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 $\boldsymbol{\Psi}$ of the normally distributed model errors. More generally, suppose that $\mathcal{F}$ has a feature space $\mathcal{V}$ such that the mapping $\phi : \mathcal{X} \to \mathcal{V}$ is its feature map, and if $f(x) = \langle\phi(x), v\rangle_{\mathcal{V}}$, then the Fisher information

$I(v) \in \mathcal{V} \otimes \mathcal{V}$ *for* v *is*

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

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

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

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

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

and the score by

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

Differentiating again gives

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

We can then calculate the Fisher information to be

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

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

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

Example 3.1 (Fisher information for linear regression)**.** As before, suppose model (1.1) subject to its assumptions hold. For simplicity, we assume iid errors, i.e. $\mathbf{\Psi} = \psi \mathbf{I}_n$. Let $\mathcal{X} = \mathbb{R}^p$, and the feature space $\mathcal{V} = \mathbb{R}^p$ be equipped with the usual dot product $\langle \cdot, \cdot \rangle_{\mathcal{V}} : \mathcal{V} \otimes \mathcal{V} \to \mathbb{R}$ defined by $v^\top v$. Consider also the feature map $\phi : \mathcal{X} \to \mathcal{V}$ defined by $\phi(x) = x$. For some $\beta \in \mathcal{V}$, the linear regression model is such that $f(x) = x^\top \beta = \langle \phi(x), \beta \rangle_{\mathcal{V}}$. Therefore, according to Lemma (3.2), the Fisher information for β is

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

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

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

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

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

Proof. In a RKHS $\mathcal{F}$, the reproducing property gives $f(x) = \langle f, h(\cdot, x) \rangle_{\mathcal{F}}$ and in partic-

ular, $\langle h(\cdot,x), h(\cdot,x')\rangle_{\mathcal{F}} = h(x,x')$. By Lemma 3.1, we have that

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

□

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

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

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

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

Corollary 3.2.2. *With $g \in \mathcal{F}$, the Fisher information for g is zero if and only if $g \in \mathcal{F}_n^{\perp}$,*

i.e. if and only if $g(x_1) = \cdots = g(x_n) = 0$.

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

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

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

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

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?

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

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

1

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

3.5 The I-prior

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

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

The set $\mathcal{F}$ is potentially "too big" for the purpose of estimating f, that is, for certain pairs of functions $\mathcal{F}$, the data do not allow an assessment of whether one is closer to the truth than the other. In particular, the data do not contain information to distinguish betwen any f and f' for which $f(x_i) = f'(x_i), i = 1, \ldots, n$. 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 3.3 (Entropy)**.** The entropy of a distribution π over $\mathcal{F}$ relative to a measure ν is defined as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

subject to the constraints

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

is the density given by

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

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

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

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

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

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

Set this to zero, and solve for p:

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

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*

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

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

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

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

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

The norm on $\mathcal{F}_n$ induces the metric $d(f, f') = \|f - f'\|_{\mathcal{F}_n}$. Thus, for $f \in \mathcal{F}$ of the

form $f = \sum_{i=1}^n h(\cdot, x_i) w_i$ (i.e., $f \in \mathcal{F}_n$) and provided $f_0 \in \mathcal{F}_n \subset \mathcal{F}$,

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

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

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

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

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

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

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

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,

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

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

3.6 Rate of convergence

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

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

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

Bibliography

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

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

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

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.
$\delta_{xx'}$	The Kronecker delta: $\delta_{xx'} = 1$ if $x = x'$, and 0 otherwise.
$\otimes$	The tensor product.

Index