Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: 'Regression modelling using priors depending on Fisher information covariance kernels'
29 October 2018 (`v1.2@dae7107`)

Chapter 3

Fisher information and the I-prior

We are interested in calculating the Fisher information for our unknown regression function f (the parameter to be estimated) in (1.1), subject to (1.2) and $f \in \mathcal{F}$, a reproducing kernel Kreĭn space (RKKS). Usually, the Fisher information pertains to finite-dimensional parameters, but as $\mathcal{F}$ may be infinite dimensional, care must be taken when computing derivatives with respect to f. For function spaces that possess an orthonormal basis, which all Hilbert spaces do, then one could define the derivative of the functional $\rho : \mathcal{F} \to \mathbb{R}$ componentwise with respect to the orthonormal basis, as in the finite-dimensional case. This is analogous to the usual concept of partial derivatives.

However, the notion of partial derivatives does not generalise to arbitrary topological vector spaces for two reasons. Firstly, general spaces may not have an orthonormal basis (Tapia, 1971, Sec. 5, p. 76). Secondly, componentwise derivatives, which are in essence limits taken componentwise using the usual definition of derivatives, may not coincide with the overall limit taken with respect to the topology of the vector space. For these reasons, there is a need to consider the rigorous concepts of differentiation suitable for infinite-dimensional vector spaces provided by Fréchet and Gâteaux derivatives. These concepts are introduced in Section 3.2, prior to the actual derivation of the Fisher information of the regression function in Section 3.3.

In the remaining sections, we 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, 1957b, 2003): in the absence of any prior knowledge, a prior distribution which maximises entropy should be used. As it turns out, the entropy maximising prior for f is Gaussian with mean chosen a priori and covariance kernel proportional to the Fisher information. We call such a distribution on f an *I-prior distribution* for f. The I-prior has a simple, intuitive appeal: much information about f corresponds to a larger prior covariance, and thus less influence of the prior mean, and more of the data, in informing the posterior, and vice versa.

3.1 The traditional Fisher information

It was Fisher (1922) who introduced the method of maximum likelihood (ML) as an objective way of conducting statistical inference. This method of inference is distinguished from the Bayesian school of thought in that only the data may inform deductive reasoning, but not any sort of prior probabilities. Towards the later stages of his career[1], his work reflected the view that the likelihood is to be more than simply a device to obtain parameter estimates; it is also a vessel that carries uncertainty about estimation. In this light and in the absence of the possibility of making probabilistic statements, one should look to the likelihood in order to make rational conclusions about an inference problem. Specifically, we may ask two things of the likelihood function: where is the maximum and what does the graph around the maximum look like? The first of these two problems is of course ML estimation, while the second concerns the Fisher information.

In simple terms, the Fisher information measures the amount of information that an observable random variable Y carries about an unknown parameter θ of the statistical model that models Y. To make this concrete, let Y have the density function $p(\cdot|\theta)$ which depends on 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*) with respect to θ as $S(\theta) = \partial L(\theta)/\partial\theta$. The *Fisher information* about the parameter θ is defined to be the 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 texts provide a proof of this fact—see, for example, Wasserman (2004, Sec. 9.7).

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[2] of a function, measures how quickly the function changes with changes in its input values.

[1] The introductory chapter of Pawitan (2001) and the citations therein give a delightful account of the evolution of the Fisherian view regarding statistical inference.

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

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 of a small variance (less uncertainty), while low Fisher information is indicative of a shallow maxima for which many θ share similar log-likelihood values.

3.2 Fisher information in Hilbert space

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 RKKSs later. The score and Fisher information is derived in a familiar manner, but extra care is required when taking derivatives with respect to elements in Hilbert spaces. We discuss a generalisation of the concept of differentiability from real-valued functions of a single, real variable, as is common in calculus, to functions between Hilbert spaces.

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

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

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

Remark 3.1. Since $\mathrm{d}\rho(x)$ is a bounded, linear operator, by Lemma 2.1 (p. 47), it is also continuous.

Remark 3.2. While the Fréchet derivative is most commonly defined as the derivative of functions between Banach spaces, the definition itself also applies to Hilbert spaces, since complete inner product spaces are also complete normed spaces. Since our main focus are RKHSs and RKKSs, i.e. spaces with Hilbertian topology (recall that RKKSs are endowed with the topology of its associated Hilbert space), it is beneficial to present the material using Hilbert spaces. We appeal to the works of Balakrishnan (1981, Def. 3.6.5) and Bouboulis and Theodoridis (2011, Sec. 6) in this regard.

Remark 3.3. The use of the open subset $\mathcal{U}$ in the definition above for the domain of the function ρ is so that the notion of ρ being differentiable is possible even without having it defined on the entire space $\mathcal{V}$.

The intuition here is similar to that of regular differentiability, in that the linear operator T well approximates the change in ρ 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 $\rho(x+v) = \rho(v) + \mathrm{d}\rho(x)(v) + o(v)$, that is, the derivative of ρ at x gives the best linear approximation to ρ near x. Note that the limit in the definition is meant in the usual sense of convergence of functions with respect to the norms of $\mathcal{V}$ and $\mathcal{W}$.

For the avoidance of doubt, $\mathrm{d}\rho(x)$ 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 $\rho : \mathcal{U} \to \mathcal{W}$ is a differentiable function at all points in $\mathcal{U} \subseteq \mathcal{V}$, then its derivative is a linear map

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

It follows that this function may also have a derivative, which by definition will be a linear map as well. This is the *second Fréchet derivative* of ρ, defined by

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

To make sense of the space on the right-hand side, consider the following argument.

- Take any $\phi(\cdot) \in \mathrm{L}\big(\mathcal{V};\mathrm{L}(\mathcal{V};\mathcal{W})\big)$. For all $v \in \mathcal{V}$, $\phi(v) \in \mathrm{L}(\mathcal{V};\mathcal{W})$, and $\phi(v)$ is linear in v.
- Since $\phi(v) \in \mathrm{L}(\mathcal{V};\mathcal{W})$, it is itself a linear operator taking elements from $\mathcal{V}$ to $\mathcal{W}$. We can write it as $\phi(v)(\cdot)$ for clarity.
- So, for any $v' \in \mathcal{V}$, $\phi(v)(v') \in \mathcal{W}$, and it depends linearly on v' too. Thus, given any two $v, v' \in \mathcal{V}$, we obtain an element $\phi(v)(v') \in \mathcal{W}$ which depends linearly on both v and v'.
- It is therefore possible to identify $\phi \in \mathrm{L}\big(\mathcal{V};\mathrm{L}(\mathcal{V};\mathcal{W})\big)$ with an element $\xi \in \mathrm{L}(\mathcal{V} \times \mathcal{V}, \mathcal{W})$ such that for all $v, v' \in \mathcal{V}$, $\phi(v)(v') = \xi(v, v')$.

To summarise, there is an isomorphism between the space on the right-hand side and the space $\mathrm{L}(\mathcal{V} \times \mathcal{V}, \mathcal{W})$ of all continuous, bilinear maps from $\mathcal{V}$ to $\mathcal{W}$. The second derivative $\mathrm{d}^2\rho(x)$ is therefore a bounded, symmetric, 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 $\rho : \mathcal{U} \to \mathcal{W}$.

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

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

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

Remark 3.4. For Gâteaux derivatives, $\mathcal{V}$ need only be a vector space, while $\mathcal{W}$ a topological space. Tapia (1971, p. 55) wrote that for quite some time analysis was simply done using the topology of the real line when dealing with functionals. As a result, important concepts such as convergence could not be adequately discussed.

Remark 3.5. Tapia (1971, p. 52) goes on to remark that 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.

Unlike the Fréchet derivative, which is by definition a linear operator, the Gâteaux derivative may fail to satisfy the additive condition of linearity[3]. Even if it is linear, it may fail to depend continuously on some $v' \in \mathcal{V}$ if $\mathcal{V}$ and $\mathcal{W}$ are infinite dimensional. In this sense, Fréchet derivatives are more demanding than Gâteaux derivatives. Nevertheless, the reasons we bring up Gâteaux derivatives is because it is usually simpler to calculate Gâteaux derivatives than Fréchet derivatives, and the two concepts are connected by the lemma below.

Lemma 3.1 (Fréchet differentiability implies Gâteaux differentiability)**.** *If ρ is Fréchet differentiable at $x \in \mathcal{U}$, then $\rho : \mathcal{U} \to \mathcal{W}$ is Gâteaux differentiable at that point too, and* $\mathrm{d}\rho(x) = \partial\rho(x)$.

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

$$\begin{aligned} \lim_{t \to 0} \left\| \frac{\rho(x + tv) - \rho(x)}{t} - \mathrm{d}\rho(x)(v) \right\|_{\mathcal{W}} &= \lim_{t \to 0} \frac{1}{t} \left\| \rho(x + tv) - \rho(x) - \mathrm{d}\rho(x)(tv) \right\|_{\mathcal{W}} \\ &= \lim_{t \to 0} \frac{\left\| \rho(x + tv) - \rho(x) - \mathrm{d}\rho(x)(tv) \right\|_{\mathcal{W}}}{\|tv\|_{\mathcal{V}}} \|v\|_{\mathcal{V}} \end{aligned} \tag{3.1}$$

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

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

On the other hand, Gâteaux differentiability does not necessarily imply Fréchet differentiability. A sufficient condition for Fréchet differentiability is that the Gâteaux derivative is continuous at the point of differentiation, i.e. the map $\partial\rho : \mathcal{U} \to (\mathcal{V}; \mathcal{W})$ is continuous at $x \in \mathcal{U}$. In other words, if $\partial\rho(x)$ is a bounded linear operator and the convergence in (3.1) is uniform with respect to all v such that $\|v\|_{\mathcal{V}} = 1$, then $\mathrm{d}\rho(x)$ exists and $\mathrm{d}\rho(x) = \partial\rho(x)$ (Tapia, 1971, p. 57 & 66).

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

$$\left\| \frac{\mathrm{d}\rho(x + tv)(v') - \mathrm{d}\rho(x)(v')}{t} - \mathrm{d}^2\rho(x)(v, v') \right\|_{\mathcal{W}} \to 0 \text{ as } t \to 0, \tag{3.2}$$

by a similar argument in the proof of Lemma 3.1 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, as a consequence of the Riesz-Fréchet theorem, the mapping $U : \mathcal{V} \to \mathcal{V}^*$ from the Hilbert space $\mathcal{V}$ to its continuous dual space $\mathcal{V}^*$ defined by $U : v \mapsto \langle \cdot, v \rangle_{\mathcal{V}}$ is an isometric isomorphism. Again, let $\mathcal{U} \subseteq \mathcal{V}$ be an open subset, and let $\rho : \mathcal{U} \to \mathbb{R}$ be a Fréchet differentiable function with derivative $\mathrm{d}\rho : \mathcal{U} \to \mathrm{L}(\mathcal{V}; \mathbb{R}) \equiv \mathcal{V}^*$. We define the gradient as follows.

Definition 3.3 (Gradient)**.** The *gradient* of ρ is the operator $\nabla\rho : \mathcal{U} \to \mathcal{V}$ defined by $\nabla\rho = U^{-1} \circ \mathrm{d}\rho$. Thus, for $x \in \mathcal{U}$, the gradient of ρ at x, denoted $\nabla\rho(x)$, is the unique element of $\mathcal{V}$ satisfying

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

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

Remark 3.6. Alternatively, the gradient can be motivated using the Riesz representation theorem in Definition 3.1 of the Fréchet derivative. Since $\mathcal{V}^* \ni T : \mathcal{V} \to \mathbb{R}$, there is a unique element $v^* \in \mathcal{V}$ such that $T(v) = \langle v^*, v \rangle_{\mathcal{V}}$ for any $v \in \mathcal{V}$. The element $v^* \in \mathcal{V}$ is called the gradient of ρ at x.

Since the gradient of ρ is an operator on $\mathcal{U}$ to $\mathcal{V}$, it may itself have a Fréchet derivative. Assuming existence, i.e. ρ is twice Fréchet differentiable at $x \in \mathcal{U}$, we call this derivative

the *Hessian* of ρ. From (3.2), it must be that

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

The second line follows from the definition of gradients, the third line by linearity of inner products, and the final line by definition of Gâteaux derivatives and continuity of inner products[4]. Since $\nabla\rho$ is continuous, its Fréchet and Gâteaux differentials coincide, and we have that $\partial_v \nabla\rho(x) = \mathrm{d}\nabla\rho(x)(v)$. Letting $\mathcal{V}$, $\mathcal{W}$ and $\mathcal{U}$ be as before, we now define the Hessian for the function $\rho : \mathcal{U} \to \mathcal{W}$.

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

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

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

Remark 3.7. Since $\mathrm{d}^2\rho(x)$ is a bilinear form in $\mathcal{V}$, we can equivalently write

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

following the correspondence between bilinear forms and tensor product spaces.

With the differentiation tools above, we can now derive the Fisher information that we set out to obtain at the beginning of this section. Let Y be a random variable with density in the parametric family $\{p(\cdot|\theta) \,|\, \theta \in \Theta\}$, 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 $\theta \mapsto \log p(Y|\theta)$. The score S, assuming existence, is defined to be the (Fréchet) derivative of $L(\cdot|Y)$ at θ, 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 $L(\cdot|Y)$ at θ is then $\mathrm{d}^2L(\cdot|Y) : \Theta \to \mathrm{L}(\Theta\times\Theta;\mathbb{R})$. We now prove the following proposition.

Proposition 3.2 (Fisher information in Hilbert spaces)**.** *Assume that both $p(Y|\cdot)$ and $\log p(Y|\cdot)$ are Fréchet differentiable at θ. Then, the Fisher information for $\theta \in \Theta$ is the element in the tensor product space $\Theta \otimes \Theta$ defined by*

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

[4]For any continuous function $g : \mathbb{R} \to \mathbb{R}$, $\lim_{x\to a} g(x) = g(\lim_{x\to a} x) = g(a)$.

Equivalently, assuming further that $\log p(Y|\cdot)$ *is twice Fréchet differentiable at* θ*, the Fisher information can be written as*

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

Note that both expectations are taken under the true distribution of random variable Y*.*

Proof. The Gâteaux derivative of $L(\cdot|Y) = \log p(Y|\cdot)$ at $\theta \in \Theta$ in the direction $b \in \Theta$, which is also its Fréchet derivative, is

$$\begin{aligned}
\partial_b L(\theta|Y) &= \frac{\mathrm{d}}{\mathrm{d}t} \log p(Y|\theta + tb)\bigg|_{t=0} \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}t} p(Y|\theta + tb)\big|_{t=0}}{p(Y|\theta)} \\
&= \frac{\partial_b p(Y|\theta)}{p(Y|\theta)}.
\end{aligned}$$

Since it assumed that $p(Y|\cdot)$ is Fréchet differentiable at θ, $\mathrm{d}p(Y|\theta)(b) = \partial_b p(Y|\theta)$. The expectation of the score for any $b \in \Theta$ is shown to be

$$\begin{aligned}
\mathrm{E}[\mathrm{d}L(\theta|Y)(b)] &= \mathrm{E}\left[\frac{\mathrm{d}p(Y|\theta)(b)}{p(Y|\theta)}\right] \\
&= \int \frac{\mathrm{d}p(Y|\theta)(b)}{\cancel{p(Y|\theta)}} \cancel{p(Y|\theta)}\, \mathrm{d}Y \\
&= \mathrm{d}\left(\int p(Y|\theta)\, \mathrm{d}Y\right)(b) \\
&= 0.
\end{aligned}$$

The interchange of Lebesgue integrals and Fréchet differentials is allowed under certain conditions[5], which are assumed to be satisfied here. The derivative of $\int p(Y|\cdot)\, \mathrm{d}Y$ at any value of $\theta \in \Theta$ is the zero vector, as it is the derivative of a constant (i.e. 1).

Using the classical notion that the Fisher information is the variance of the score function, then, for fixed $b, b' \in \Theta$, combined with the fact that $\mathrm{d}L(\theta|Y)(\cdot)$ is a zero-

[5] Following Kammar (2016), the conditions are:

1. $L(\cdot|Y)$ is Frechét differentiable on $\mathcal{U} \subseteq \Theta$ for almost every $Y \in \mathbb{R}$.
2. $L(\theta|Y)$ and $\mathrm{d}L(\theta|Y)(b)$ are both integrable with respect to Y, for any $\theta \in \mathcal{U} \subseteq \Theta$ and $b \in \Theta$.
3. There is an integrable function $g(Y)$ such that $L(\theta|Y) \leq g(Y)$ for all $\theta \in \Theta$ and almost every $Y \in \mathbb{R}$.

These conditions as stated are analogous to the measure theoretic requirements for Leibniz's integral rule to hold (differentiation under the integral sign). For nice and well-behaved probability densities, such as the normal density that we will be working with, there aren't issues with interchanging integrals and derivatives.

meaned function, we have that

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

Hence, $\mathcal{I}(\theta)$ as a bilinear form corresponds to the element $\mathrm{E}[\nabla L(\theta|Y)\otimes\nabla L(\theta|Y)] \in \Theta\otimes\Theta$.

The Gâteaux derivative of the Fréchet differential is the second Fréchet derivative, since $L(\cdot|Y)$ is assumed to be twice Fréchet differentiable at $\theta\in\Theta$:

$$\begin{aligned}\mathrm{d}^2L(\theta|Y)(b,b') &= \partial_{b'}\mathrm{d}L(\theta|Y)(b) \\ &= \partial_{b'}\left(\frac{\mathrm{d}p(Y|\theta)(b)}{p(Y|\theta)}\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}p(Y|\theta+tb')(b)}{p(Y|\theta+tb')}\right)\Bigg|_{t=0} \\ &= \frac{p(Y|\theta)\mathrm{d}^2p(Y|\theta)(b,b') - \mathrm{d}p(Y|\theta)(b)\mathrm{d}p(Y|\theta)(b')}{p(Y|\theta)^2} \\ &= \frac{\mathrm{d}^2p(Y|\theta)(b,b')}{p(Y|\theta)} - \mathrm{d}L(\theta|Y)(b)\,\mathrm{d}L(\theta|Y)(b').\end{aligned}$$

Taking expectations of the first term in the right-hand side, we get that

$$\begin{aligned}\mathrm{E}\left[\frac{\mathrm{d}^2p(Y|\theta)(b,b')}{p(Y|\theta)}\right] &= \int \frac{\mathrm{d}\big(\mathrm{d}p(Y|\theta)\big)(b,b')}{\cancel{p(Y|\theta)}}\cancel{p(Y|\theta)}\,\mathrm{d}Y \\ &= \mathrm{d}^2\left(\int p(Y|\theta)\,\mathrm{d}Y\right)(b,b') \\ &= 0.\end{aligned}$$

Thus, we see that from the first result obtained,

$$\begin{aligned}\mathrm{E}[-\mathrm{d}^2L(\theta|Y)(b,b')] &= \mathrm{E}[\mathrm{d}L(\theta|Y)(b)\,\mathrm{d}L(\theta|Y)(b')] \\ &= \mathcal{I}(\theta)(b,b'),\end{aligned}$$

while

$$\begin{aligned}\mathrm{E}[-\mathrm{d}^2L(\theta|Y)(b,b')] &= -\,\mathrm{E}\langle\nabla^2L(\theta|Y)(b),b'\rangle_\Theta \\ &= \langle -\,\mathrm{E}\,\nabla^2L(\theta|Y)(b),b'\rangle_\Theta.\end{aligned}$$

It would seem that $\mathrm{E}[-\nabla^2L(\theta|Y)(b)]$ is an operator from Θ onto itself which also induces a bilinear form equivalent to $\mathrm{E}[-\mathrm{d}^2L(\theta|Y)]$. Therefore, $\mathcal{I}(\theta) = \mathrm{E}[-\nabla^2L(\theta|Y)]$. ■

The Fisher information $\mathcal{I}(\theta)$ for θ, much like the covariance operator, can be viewed in one of three ways:

1. As its general form, i.e. an element in the tensor product space $\Theta \otimes \Theta$;
2. As an operator $\mathcal{I}(\theta) : \Theta \to \Theta$ defined by $\mathcal{I}(\theta) : b \mapsto \mathrm{E}[-\nabla^2 L(\theta|Y)](b)$; and finally
3. As a bilinear form $\mathcal{I}(\theta) : \Theta \times \Theta \to \mathbb{R}$ defined by $\mathcal{I}(\theta)(b, b') = \langle -\mathrm{E}\,\nabla^2 L(\theta|Y)(b), b' \rangle_\Theta$ $= \mathrm{E}[-\mathrm{d}^2 L(\theta|Y)(b, b')]$.

In particular, viewed as a bilinear form, the evaluation of the Fisher information for θ at two points b and b' in Θ is seen as the Fisher information between two continuous, linear functionals of θ. For brevity, we denote this $\mathcal{I}(\theta_b, \theta_{b'})$, where $\theta_b = \langle \theta, b \rangle_\theta$ for some $b \in \Theta$. The natural isometry between Θ and its continuous dual Θ^* then allows us to write

$$\mathcal{I}(\theta_b, \theta_{b'}) = \langle \mathcal{I}(\theta), b \otimes b' \rangle_{\Theta \otimes \Theta} = \big\langle \mathcal{I}(\theta), \langle \cdot, b \rangle_\Theta \otimes \langle \cdot, b' \rangle_\Theta \big\rangle_{\Theta^* \otimes \Theta^*}. \tag{3.3}$$

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 $x_i \in \mathcal{X}$, for $i = 1, \ldots, n$ is

$$y_i = \alpha + f(x_i) + \epsilon_i \qquad \text{(from 1.1)}$$
$$(\epsilon_1, \ldots, \epsilon_n)^\top \sim \mathrm{N}_n(0, \boldsymbol{\Psi}^{-1}) \qquad \text{(from 1.2)}$$

where $\alpha \in \mathbb{R}$ is an intercept and f is in an RKKS $\mathcal{F}$ with kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Note that the dependence of the kernel on parameters η is implicitly assumed.

Lemma 3.3 (Fisher information for regression function)**.** *For the regression model (1.1) subject to (1.2) and $f \in \mathcal{F}$ where $\mathcal{F}$ is an RKKS 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 Gâteaux derivative of $k_x(v)$ in the direction u is

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

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 $\mathrm{d}k_x(v) = \partial k_x(v)$. Let $\mathbf{y} = \{y_1, \ldots, y_n\}$, and denote the hyperparameters of the regression model by $\theta = \{\alpha, \mathbf{\Psi}, \eta\}$. Without loss of generality, assume $\alpha = 0$, and even if this is not so, we can always add back α to the y_i's later. Regardless, both α and $\mathbf{y}$ are constant in the differential of $L(v|\mathbf{y}, \theta)$. The log-likelihood of v is given by

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

and the score by

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

Differentiating again gives

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

since the derivative of $\mathrm{d}k_x = \langle \phi(x), \cdot \rangle_{\mathcal{V}}$ is zero (it is the derivative of a constant). We can then calculate the Fisher information to be

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

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. Also, the expectation is free of the random variable under expectation (i.e. $\mathbf{y}$), which makes the second line possible.

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. ■

The above lemma gives the form of the Fisher information for f in a rather abstract fashion. Consider the following example of applying Lemma 3.3 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 (1.2) and $f \in \mathcal{F}$, an RKHS. 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 identity feature map $\phi : \mathcal{X} \to \mathcal{V}$ defined by $\phi(\mathbf{x}) = \mathbf{x}$. For some $\boldsymbol{\beta} \in \mathcal{V}$, the linear regression model is such that $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta} = \langle \phi(\mathbf{x}), \boldsymbol{\beta} \rangle_{\mathcal{V}}$. Therefore, according to Lemma 3.3, the Fisher information for $\boldsymbol{\beta}$ is

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta}) &= \sum_{i=1}^n \sum_{j=1}^n \psi \cdot \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_j) \\ &= \psi \sum_{i=1}^n \sum_{j=1}^n \mathbf{x}_i \otimes \mathbf{x}_j \\ &= \psi \mathbf{X}^\top \mathbf{X}.\end{aligned}$$

Note that the operation '$\otimes$' on two vectors in Euclidean space is simply their outer product. The resulting $\mathbf{X}$ is a $n \times p$ matrix containing the entries $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$ 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 linear functionals of f, and in particular, for point evaluation functionals of f, thereby allowing us to compute the Fisher information at two points $f(x)$ and $f(x')$.

Corollary 3.3.1 (Fisher information between two linear functionals of f)**.** *For our regression model as defined in (1.1) subject to (1.2) and f belonging to an RKKS $\mathcal{F}$ with kernel h, the Fisher information at 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 an RKKS $\mathcal{F}$, 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 (3.3), we have that

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

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 3.3.1 reveals the fact that the Fisher information for $f(x)$, $\mathcal{I}\big(f(x), f(x)\big)$, 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.

3.4 The induced Fisher information RKHS

From Lemma 3.3, the formula for the Fisher information uses n points of the observed data $x_i \in \mathcal{X}$. This seems to suggest that the Fisher information only exists for a finite subspace of the RKKS $\mathcal{F}$. Indeed, this is the case, and we will be specific about the subspace for which there is Fisher information. Consider the following set, a similar one

considered in the proof of the Moore-Aronszajn theorem (Theorem 2.6, p. 57):

$$\mathcal{F}_n = \left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, f(x) = \sum_{i=1}^{n} h(x, x_i) w_i,\ w_i \in \mathbb{R},\ i = 1, \dots, n \right\}. \tag{3.4}$$

Since $h(\cdot, x_i) \in \mathcal{F}$, 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 $\mathcal{F}$. Unlike Theorem 2.6, $\mathcal{F}_n$ defined here is a finite subspace of dimension n.

Let $\mathcal{F}_n^\perp$ be the orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$. By the orthogonal decomposition theorem (Theorem 2.3, p. 49), any regression function $f \in \mathcal{F}$ can be uniquely decomposed as $f = f_n + r$, with $f_n \in \mathcal{F}_n$ and $r \in \mathcal{F}_n^\perp$, where $\mathcal{F} = \mathcal{F}_n \oplus \mathcal{F}_n^\perp$. We saw in the proof of Theorem 2.6 that $\mathcal{F}$ is the closure of $\mathcal{F}_n$, so therefore $\mathcal{F}$ is dense in $\mathcal{F}_n$, and hence by Corollary 2.3.1 (p. 49) we have that $\mathcal{F}_n^\perp = \{0\}$. Alternatively, we could have argued that any $r \in \mathcal{F}_n^\perp$ is orthogonal to each of the $h(\cdot, x_i) \in \mathcal{F}$, so by the reproducing property of h, $r(x_i) = \langle r, h(\cdot, x_i) \rangle_{\mathcal{F}} = 0$. This suggests the following corollary.

Corollary 3.3.2. *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 $g(x_1) = \cdots = g(x_n) = 0$.*

Proof. Let $\mathcal{I}(f)$ be the Fisher information for f. The Fisher information for $\langle f, r \rangle_{\mathcal{F}}$ is

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

So if $r \in \mathcal{F}_n^\perp$, then $r(x_1) = \cdots = r(x_n) = 0$, and thus the Fisher information at $r \in \mathcal{F}_n^\perp$ is zero. Conversely, if the Fisher information is zero, it must necessarily mean that $r(x_1) = \cdots = r(x_n) = 0$ since $\psi_{ij} > 0$, and thus $r \in \mathcal{F}_n^\perp$. ∎

The above corollary implies that the Fisher information for our regression function $f \in \mathcal{F}$ exists only on the n-dimensional subspace $\mathcal{F}_n$. More subtly, as there is no Fisher information for $r \in \mathcal{F}_n^\perp$, r cannot be estimated from the data. Thus, in estimating f, we will only ever consider the finite subspace $\mathcal{F}_n \subset \mathcal{F}$ where there is information about f.

As it turns out, $\mathcal{F}_n$ can be identified as an RKHS with reproducing kernel equal to the Fisher information for f. That is, the real, symmetric, and positive-definite function h_n over $\mathcal{X} \times \mathcal{X}$ defined by $h_n(x, x') = \mathcal{I}\big(f(x), f(x')\big)$ is associated to the RKHS which is $\mathcal{F}_n$, equipped with the squared norm $\|f\|_{\mathcal{F}_n}^2 = \sum_{i,j=1}^{n} w_i (\boldsymbol{\Psi}^{-1})_{ij} w_j$. This is stated in the next lemma.

Lemma 3.4. *Let $\mathcal{F}_n$ as in (3.4) be equipped with the inner product*

$$\langle f, f' \rangle_{\mathcal{F}_n} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i (\mathbf{\Psi}^{-1})_{ij} w'_j = \mathbf{w}^\top \mathbf{\Psi} \mathbf{w}' \tag{3.5}$$

for any two $f = \sum_{i=1}^{n} h(\cdot, x_i) w_i$ and $f' = \sum_{j=1}^{n} h(\cdot, x_j) w'_j$ in $\mathcal{F}_n$. Then, $h_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as 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. What needs to be proven is the reproducing property of h_n for $\mathcal{F}_n$. First note that by defining $w_j(x) = \sum_{k=1}^{n} \psi_{jk} h(x, x_k)$, we see that

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

Furthermore, writing $h(\cdot, x_j) = \sum_{k=1}^{n} \delta_{jk} h(\cdot, x_k)$, with δ being the Kronecker delta, we see that $h(\cdot, x_j)$ is also an element of $\mathcal{F}_n$, and in particular,

$$\big\langle h(\cdot, x_i), h(\cdot, x_k) \big\rangle_{\mathcal{F}_n} = \sum_{j=1}^{n} \sum_{l=1}^{n} \delta_{ij} (\mathbf{\Psi}^{-1})_{jl} \delta_{lk} = (\mathbf{\Psi}^{-1})_{ik}.$$

Denote by ψ^-_{ij} the (i, j)'th element of $\mathbf{\Psi}^{-1}$. A fact we will use later is $\sum_{k=1}^{n} \psi_{jk} \psi^-_{ik} = (\mathbf{\Psi}\mathbf{\Psi}^{-1})_{ji} = (\mathbf{I}_n)_{ji} = \delta_{ji}$. In the mean time,

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

Therefore, h_n is a reproducing kernel for $\mathcal{F}_n$. Obviously, h_n is positive definite (it is a squared kernel), and hence it defines the RKHS $\mathcal{F}_n$. ∎

3.5 The I-prior

In the introductory chapter (Chapter 1), we discussed that unless the regression function f is regularised (for instance, using some prior information), the ML estimator of f is likely to be inadequate. In choosing a prior distribution for f, we appeal to the principle of maximum entropy (Jaynes, 1957a, 1957b, 2003), which states that the probability distribution which best represents the current state of knowledge is the one with largest entropy. In this section, we aim to show the relationship between the Fisher information for f and its maximum entropy prior distribution. Before doing this, we recall the definition of entropy and derive the maximum entropy prior distribution for a parameter which has unrestricted support. Let (Θ, D) be a metric space and let $\nu = \nu_D$ be a volume measure induced by D (e.g. Hausdorff measure). In addition, assume ν is a probability measure over Θ so that $(\Theta, \mathcal{B}(\Theta), \nu)$ is a Borel probability space.

Definition 3.5 (Entropy)**.** Denote by p a probability density over Θ relative to ν. Suppose that $\int p \log p \, \mathrm{d}\nu < \infty$, i.e. $p \log p$ is Lebesgue integrable and belongs to the space $\mathrm{L}^1(\Theta, \nu)$. The entropy of a distribution p over Θ relative to a measure ν is defined as

$$H(p) = -\int_\Theta p(\theta) \log p(\theta) \, \mathrm{d}\nu(\theta). \tag{3.6}$$

In deriving the maximum entropy distribution, we will need to maximise the functional H with respect to p. Typically, this is done using calculus of variations techniques, and standard calculations (Appendix A, p. 269) reveal that the functional derivative of $H(p)$ with respect to p, denoted $\partial H / \partial p$, is equal to $-1 - \log p$. We now present another well known result from information theory, regarding the form of the maximum entropy distribution.

Lemma 3.5 (Maximum entropy distribution)**.** *Let (Θ, D) be a metric space, $\nu = \nu_D$ be a volume measure induced by D, and p be a probability density function on Θ. The entropy maximising density $\tilde{p}$, which satisfies*

$$\underset{p \in L^2(\Theta, \nu)}{\arg\max} \left\{ H(p) = -\int_\Theta p(\theta) \log p(\theta) \, \mathrm{d}\nu(\theta) \right\},$$

subject to the constraints

$$\mathrm{E}\left[D(\theta, \theta_0)^2\right] = \int_\Theta D(\theta, \theta_0)^2 p(\theta) \, \mathrm{d}\nu(\theta) = \textit{const.}, \qquad \int_\Theta p(\theta) \, \mathrm{d}\nu(\theta) = 1,$$
$$\textit{and} \quad p(\theta) \geq 0, \forall \theta \in \Theta,$$

is the density given by

$$\tilde{p}(\theta) \propto \exp\left(-\frac{1}{2} D(\theta, \theta_0)^2\right),$$

for some fixed $\theta_0 \in \Theta$. If (Θ, D) is a Euclidean space and ν a flat (Lebesgue) measure then $\tilde{p}$ represents a (multivariate) normal density.

Sketch proof. This follows from standard calculus of variations, though we provide a sketch proof here. Set up the Langrangian

$$\begin{aligned}\mathcal{L}(p, \gamma_1, \gamma_2) = &- \int_\Theta p(\theta) \log p(\theta) \, \mathrm{d}\nu(\theta) + \gamma_1 \left(\int_\Theta D(\theta, \theta_0)^2 p(\theta) \, \mathrm{d}\nu(\theta) - \text{const.} \right) \\ &+ \gamma_2 \left(\int_\Theta p(\theta) \, \mathrm{d}\nu(\theta) - 1 \right).\end{aligned}$$

Taking derivatives with respect to p (see Appendix A, p. 269 for definition of functional derivatives) yields

$$\frac{\partial}{\partial p} \mathcal{L}(p, \gamma_1, \gamma_2)(\theta) = -1 - \log p(\theta) + \gamma_1 D(\theta, \theta_0)^2 + \gamma_2.$$

Set this to zero, and solve for $p(\theta)$:

$$\begin{aligned}p(\theta) &= \exp\left(\gamma_1 D(\theta, \theta_0)^2 + \gamma_2 - 1\right) \\ &\propto \exp\left(\gamma_1 D(\theta, \theta_0)^2\right).\end{aligned}$$

This density is positive for any values of γ_1 (and γ_2), and it normalises to one if $\gamma_1 < 0$. As γ_1 can take any value less than zero, we choose $\gamma_1 = -1/2$.

Now, if $\Theta \equiv \mathbb{R}^m$ and ν is the Lebesgue measure, then $D(\theta, \theta_0)^2 = \|\theta - \theta_0\|^2_{\mathbb{R}^m}$, so $\tilde{p}$ is recognised as a multivariate normal density centred at θ_0 with identity covariance matrix. ■

Returning to the normal regression model of (1.1) subject to (1.2), we shall now derive the maximum entropy prior for f in some RKKS $\mathcal{F}$. One issue that we have is that 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 between two functions f and g in $\mathcal{F}$ for which $f(x_i) = g(x_i), i = 1, \ldots, n$ since the Fisher information for the difference between f and g would be zero. Since the Fisher information for a linear functional of a non-zero $f_n \in \mathcal{F}_n$ is non-zero, there is information to allow a comparison between any pair of functions in $f_0 + \mathcal{F}_n := \{f_0 + f_n \,|\, f_n \in \mathcal{F}_n\}$ for some $f_0 \in \mathcal{F}$. 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. We now state and prove the main I-prior theorem.

Theorem 3.6 (I-prior)**.** *Let $\mathcal{F}$ be an RKKS with kernel h, and consider the finite-dimensional subspace $\mathcal{F}_n$ of $\mathcal{F}$ equipped with an inner product as per (3.5). 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 $\mathcal{D}_0$ be the class of distributions p such that*

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

Denote by $\tilde{p}$ the density of the entropy maximising distribution among the class of distributions within $\mathcal{D}_0$. Then, $\tilde{p}$ is Gaussian over $\mathcal{F}$ with mean f_0 and covariance function 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 $f = f_n + r$, with $f_n \in \mathcal{F}_n$ and $r \in \mathcal{F}_n^\perp$. Also recall that there is no Fisher information about any $r \in \mathcal{R}_n$, and therefore it is not possible to estimate r from the data. Therefore, $p(r) = 0$, 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 $D(f, f') = \|f - f'\|_{\mathcal{F}_n}$. Consider functions in the set $f_0 + \mathcal{F}_n$, i.e. functions of the form

$$f = f_0 + \sum_{i=1}^n h(\cdot, x_i) w_i,$$

such that $(f - f_0) \in \mathcal{F}_n$. Compute the squared distance between f and f_0:

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

Thus, by Lemma 3.5, the maximum entropy distribution for $f - f_0 = \sum_{i=1}^n h(\cdot, x_i) w_i$ is

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

This implies that f is Gaussian, since

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

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

$$\begin{aligned}
\mathrm{E}\langle f, f' \rangle_{\mathcal{F}} &= \langle f_0, f' \rangle_{\mathcal{F}} + \mathrm{E}\left[\sum_{i=1}^n w_i \left\langle h(\cdot, x_i), f' \right\rangle_{\mathcal{F}} \right] \\
&= \langle f_0, f' \rangle_{\mathcal{F}} + \sum_{i=1}^n \mathrm{E}\, \cancelto{0}{w_i} \left\langle h(\cdot, x_i), f' \right\rangle_{\mathcal{F}} \\
&= \langle f_0, f' \rangle_{\mathcal{F}},
\end{aligned}$$

so $\mu \equiv f_0$.

Following Definition 2.16 (p. 52), the covariance between two evaluation functionals of f is shown to satisfy

$$\begin{aligned}
\mathrm{Cov}\left(f(x), f(x')\right) &= \mathrm{Cov}\left(\langle f, h(\cdot, x) \rangle_{\mathcal{F}}, \langle f, h(\cdot, x') \rangle_{\mathcal{F}}\right) \\
&= \mathrm{E}\left[\langle f - f_0, h(\cdot, x) \rangle_{\mathcal{F}} \langle f - f_0, h(\cdot, x') \rangle_{\mathcal{F}} \right].
\end{aligned}$$

Then, making use of the reproducing property of h for $\mathcal{F}$, we have that

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

which is the reproducing kernel for $\mathcal{F}_n$. ■

In closing, we reiterate the fact that the I-prior for f in the normal regression model subject to f belonging to some RKKS $\mathcal{F}$ with kernel h_η has the simple representation

$$\begin{aligned}
f(x_i) &= f_0(x_i) + \sum_{k=1}^n h_\eta(x_i, x_k) w_k \\
(w_1, \ldots, w_n)^\top &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}).
\end{aligned} \tag{3.7}$$

Equivalently, this may be written as a Gaussian process-like prior

$$\big(f(x_1), \ldots, f(x_n)\big)^\top \sim \mathrm{N}(\mathbf{f}_0, \mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta), \tag{3.8}$$

where $\mathbf{f}_0 = \big(f_0(x_1), \ldots, f_0(x_n)\big)^\top$ is the vector of prior mean functional evaluations, and $\mathbf{H}_\eta = \big(h_\eta(x_i, x_j)\big)_{i,j=1}^n$ is the kernel matrix.

3.6 Conclusion

In estimating the regression function f of the normal model in (1.1) subject to (1.2) and f belonging to an RKKS $\mathcal{F}$, we established that the entropy maximising prior distribution for f is Gaussian with some chosen prior mean f_0, and covariance function proportional[6] to the Fisher information for f. We call this the I-prior for f.

The concept of Fisher information for a regression function f is brought about by viewing the regression model (1.1) subject to (1.2) as being parameterised by f. One caveat is that the dimension of the function space $\mathcal{F}$ to which f belongs is potentially infinite-dimensional, in which case the tools such as Fréchet and Gâteaux differentials are necessary in order to calculate first and second derivatives.

On a related note, should $\mathcal{F}$ be infinite dimensional, the task of estimating $f \in \mathcal{F}$ relies only on a finite amount of data points. However, we are certain that the Fisher information for f exists only for the finite subspace $\mathcal{F}_n$ as defined in (3.4), and it is zero everywhere else. This suggests that the data only allows us to provide an estimation to the function $f \in \mathcal{F}$ by considering functions in an (at most) n-dimensional subspace instead. In other words, it would be futile to consider functions in a space larger than this, and hence there is an element of dimension reduction here, especially when $\dim(\mathcal{F}) \gg n$.

By equipping the subspace $\mathcal{F}_n$ with the inner product (3.5), $\mathcal{F}_n$ is revealed to be an RKHS with reproducing kernel equal to the Fisher information for f. Importantly, since $\mathcal{F}_n$ as in (3.4) is the pre-Hilbert space whose completion as $n \to \infty$ is $\mathcal{F}$, functions in the subspace $\mathcal{F}_n$ contain "similarly shaped" functions as in the parent space $\mathcal{F}$. The problem at hand then boils down to a Gaussian process regression using the kernel of the RKHS $\mathcal{F}_n$, which is the Fisher information for f.

[6] Proportionality, rather than equality, is a consequence of any RKHS scale parameters that $\mathcal{F}$ may have.

Bibliography

Balakrishnan, Alampallam V. (1981). *Applied Functional Analysis*. 2nd ed. Springer-Verlag. ISBN: 978-1-4612-5867-4. DOI: 10.1007/978-1-4612-5865-0.

Bouboulis, Pantelis and Sergios Theodoridis (2011). "Extension of Wirtinger's Calculus to Reproducing Kernel Hilbert Spaces and the Complex Kernel LMS". In: *IEEE Transactions on Signal Processing* 59.3, pp. 964–978. DOI: 10.1109/TSP.2010.2096420.

Fisher, Ronald Aylmer (1922). "On the mathematical foundations of theoretical statistics". In: *Philosophical Transactions of the Royal Society A* 222.594-604, pp. 309–368. DOI: 10.1098/rsta.1922.0009.

Jaynes, Edwin Thompson (1957a). "Information Theory and Statistical Mechanics". In: *Physical Review* 106.4, p. 620. DOI: 10.1103/PhysRev.106.620.

— (1957b). "Information Theory and Statistical Mechanics II". In: *Physical Review* 108.2, p. 171. DOI: 10.1103/PhysRev.108.171.

— (2003). *Probability Theory: The Logic of Science*. Ed. by G. Larry Bretthorst. Cambridge University Press. ISBN: 978-0-521-59271-0.

Kammar, Ohad (2016). *A note on Fréchet diffrentiation under Lebesgue integrals*. URL: https://www.cs.ox.ac.uk/people/ohad.kammar/notes/kammar-a-note-on-frechet-differentiation-under-lebesgue-integrals.pdf.

Pawitan, Yudi (2001). *In All Likelihood*. Statistical Modelling and Inference Using Likelihood. Oxford University Press. ISBN: 978-0-19-850765-9.

Tapia, Richard A. (1971). *The Differentiation and Integration of Nonlinear Operators*. Ed. by Louis B. Rall. DOI: 10.1016/C2013-0-11348-7.

Wasserman, Larry (2004). *All of Statistics. A Concise Course in Statistical Inference*. New York: Springer-Verlag. ISBN: 978-0-387-40272-7. DOI: 10.1007/978-0-387-21736-9.