To-do list

1. Initially distinguished observed vs expected information, but realised it doesn't contribute to later discussion. . . . 3
2. Why wouldn't it be >0 ? . . . 8
3. Is it really between two things? . . . 10
4. REASONS? linear isometry? . . . 10
5. Shouldn't f and f_0 be in $\mathcal{F}$? . . . 20
6. Should I say something about this? Rates can be better than GPR? . . . 23

Contents

3 Fisher information and the I-prior **1**

3.1 The traditional Fisher information . . . 1
3.2 Fisher information for Hilbert space objects . . . 3
3.3 Fisher information for regression functions . . . 11
3.4 The induced Fisher information RKHS . . . 14
3.5 The I-prior . . . 17
3.6 Rate of convergence . . . 23
3.7 Conclusion . . . 23
3.7.1 Total Fisher information . . . 24
3.7.2 Functional derivatives . . . 25
3.7.3 Data dependent priors . . . 26

Bibliography **27**

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

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

It was Fisher (1922) who introduced the method of maximum likelihood 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

[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.

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 maxima and what does the graph around the maxima look like? The first of these two problems is maximum likelihood 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, Y has 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 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, Section 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. 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

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

θ share similar log-likelihood values. Fisher information may be added much in the same way as log-likelihood may be added—the *total Fisher information* from n independent and identically distributed random variables $Y_1, \ldots, Y_n$ is simply the sum of the n *unit Fisher information*, i.e. $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$.

1. Initially distinguished observed vs expected information, but realised it doesn't contribute to later discussion.

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. 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 $f : \mathcal{U} \to \mathcal{W}$ is called *Fréchet differentiable* at $x \in \mathcal{U}$ if there exists a bounded, linear operator $T : \mathcal{V} \to \mathcal{W}$ such that

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

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 *(Fréchet) differentiable* on $\mathcal{U}$.

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

Remark 3.2. While the Fréchet derivative is most commonly defined as derivatives of functions between Banach spaces, the definition itself also applies to Hilbert spaces. Since our main focus are RKHSs, it is presented as such, and we follow the definitions supplied in Balakrishnan (1981, Definition 3.6.5) and Bouboulis and Theodoridis (2011, Section 6).

Remark 3.3. The use of the open subset $\mathcal{U}$ in the definition above for the domain of the function f is so that the notion of f 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 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 $f(x+v) = f(v) + \mathrm{d}f(x)(v) + o(v)$, that is, the derivative of f at x gives the best linear approximation to f 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, $\mathrm{d}f(x)$ 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. This is the *second Fréchet derivative* of f, defined by

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

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 $\psi \in \mathrm{L}(\mathcal{V} \times \mathcal{V}, \mathcal{W})$ such that for all $v, v' \in \mathcal{V}$, $\phi(v)(v') = \psi(v, v')$.

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 $\mathrm{d}^2 f(x)$ 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 the functional derivative 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 $\partial f : \mathcal{U} \to (\mathcal{V}; \mathcal{W}) = \{A \,|\, A : \mathcal{V} \to \mathcal{W}\}$ which assigns $x \mapsto \partial f(x)$ simply the *Gâteaux derivative* of f.

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). 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 f is Fréchet differentiable at $x \in \mathcal{U}$, then $f : \mathcal{U} \to \mathcal{W}$ 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 $f(x+v) \approx f(x) + \mathrm{d}f(x)(v)$

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

for some $v \in \mathcal{V}$. Then,

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

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. □

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 $\partial f : \mathcal{U} \to (\mathcal{V}; \mathcal{W})$ is continuous at $x \in \mathcal{U}$. In other words, if $\partial f(x)$ 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 $\mathrm{d}f(x)$ exists and $\mathrm{d}f(x) = \partial f(x)$ (Tapia, 1971, p. 57 & 66).

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

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

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.

Remark 3.6. Alternatively, the gradient can be motivated using the Riesz representation theorem in Definition 3.13 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 f at x.

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 (3.2), 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 3.7. 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.

With the differentiation tools above, we can now derive the Fisher information that we set out to derive 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 $\mathrm{d}^2 L(\cdot|Y) : \Theta \to \mathrm{L}(\Theta \times \Theta, \mathbb{R})$. We now prove the following proposition.

2. Why wouldn't it be >0 ?

Proposition 3.2 (Fisher information in Hilbert space)**.** *Assume that $p(Y|\cdot)$ and $\log p(Y|\cdot)$ are both 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*

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

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

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

The interchange of the integration and the Fréchet differential is allowed under certain conditions (Kammar, 2016). 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 $\mathrm{E}[\mathrm{d}L(\theta|Y)]$ is a zero mean function, we have that

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

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 differentiable at $\theta \in \Theta$:

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

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

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

Thus, we see that from the first result obtained,

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

while

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

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

There are three equivalent interpretations of the Fisher information $\mathcal{I}(\theta)$ for θ, much like the covariance operator, which are

1. As its general form, i.e. an element in $\Theta \otimes \Theta$;
2. As an operator $\mathcal{I}(\theta) : \Theta \to \Theta$ defined by $\mathcal{I}(\theta)(b) = \mathrm{E}[-\nabla^2 L(\theta|Y)](b)$; 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. REASONS? linear isometry? For brevity, we denote this $\mathcal{I}(\theta_b, \theta_{b'})$. The natural isometry between Θ and 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. Is it really between two things?

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

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

Differentiating again gives

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

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

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

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 β 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 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 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 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 (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 $i \in \{1,\dots,n\}$. 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.

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 RKHS $\mathcal{F}$. 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):

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

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

Let $\mathcal{F}_n^{\perp}$ be the orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$. By the orthogonal decomposition

theorem, 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 earlier in 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 we have that $\mathcal{F}_n^\perp = \{0\}$. Alternatively, we could have argued the following: 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 seems to suggest the statement in 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 Fisher information about f.

As it turns out, $\mathcal{F}_n$ can be identified as a 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*

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

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. Since $\mathcal{F}_n$ is a finite subspace of $\mathcal{F}$, it is complete, and thus a Hilbert space. What remains 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 $h(\cdot, x_j) = \sum_{k=1}^n \delta_{jk} h(\cdot, x_k)$, we see that $h(\cdot, x_j)$ is also an element of $\mathcal{F}_n$, and in particular,

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

where δ is the Kronecker delta. Denote by ψ_{ij}^- the (i,j)th element of $\boldsymbol{\Psi}^{-1}$. A fact we will use later is $\sum_{k=1}^n \psi_{jk} \psi_{ik}^- = (\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1})_{ji} = (\mathbf{I}_n)_{ji} = \delta_{ji}$. Then,

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

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

3.5 The I-prior

In the introductory chapter, 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, 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

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

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 of functional derivatives. Since we have already introduced concepts of Fréchet and Gâteaux derivatives earlier, we shall use those instead. Assume that the entropy H is Fréchet differentiable at p, and that the probability densities p under consideration belong to the Hilbert space of square integrable functions $\mathrm{L}^2(\Theta, \nu)$ with inner product $\langle \theta, \theta' \rangle_\Theta = \int \theta\theta' \, \mathrm{d}\nu$. Now since the Fréchet derivative of H at p is assumed to exist, it is equal to

the Gâteaux derivative, which can be computed as follows:

$$\begin{aligned}
\partial_q H(p) &= \frac{\mathrm{d}}{\mathrm{d}t} H(p+tq)\bigg|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\left\{ -\int_\Theta \big(p(\theta)+tq(\theta)\big)\log\big(p(\theta)+tq(\theta)\big)\,\mathrm{d}\nu(\theta)\right\}\bigg|_{t=0} \\
&= -\int_\Theta \left\{ \frac{\mathrm{d}}{\mathrm{d}t}\big(p(\theta)+tq(\theta)\big)\log\big(p(\theta)+tq(\theta)\big)\bigg|_{t=0}\right\}\mathrm{d}\nu(\theta) \\
&= -\int_\Theta \left(\frac{p(\theta)q(\theta)}{p(\theta)+tq(\theta)} + \frac{tq(\theta)^2}{p(\theta)+tq(\theta)} + q(\theta)\log\big(p(\theta)+tq(\theta)\big)\right)\bigg|_{t=0}\mathrm{d}\nu(\theta) \\
&= -\int_\Theta q(\theta)\big(1+\log p(\theta)\big)\,\mathrm{d}\nu(\theta) \\
&= \big\langle -\big(1+\log p\big), q\big\rangle_\Theta \\
&= \mathrm{d}H(p)(q).
\end{aligned}$$

By definition, the gradient of H at p, denoted $\nabla H(p)$, is equal to $-1-\log p$. This agrees with the usual functional derivative of the entropy obtained via standard calculus of variations, which is usually denoted $\partial H/\partial p$. 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*

$$\operatorname*{arg\,max}_{p\in L^2(\Theta,\nu)} H(p) = -\int_\Theta \tilde{p}(\theta)\log\tilde{p}(\theta)\,\mathrm{d}\nu(\theta),$$

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

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

From the above illustration preceding the lemma, taking derivatives with respect to p 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 RKHS $\mathcal{F}$. 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 a linear functional of a non-zero $f \in \mathcal{F}_n$ is non-zero, there is information to allow a comparison between any pair of functions in $f_0 + \mathcal{F}_n := \{f_0 + f \,|\, f_0 \in \mathcal{F}, f \in \mathcal{F}_n\}$. 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. We now state and prove the I-prior theorem.

Theorem 3.6 (The I-prior)**.** *Let $\mathcal{F}$ be an RKHS with kernel h, and consider the finite dimensional 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* $\mathcal{P}_0$ *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* $\mathcal{P}_0$. *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 $D(f, f') = \|f - f'\|_{\mathcal{F}_n}$. Thus, for $f, f_0 \in \mathcal{F}$ of the forms $f = \sum_{i=1}^n h(\cdot, x_i) w_i$ and $f_0 = \sum_{i=1}^n h(\cdot, x_i) w_i$ (i.e., $f, f_0 \in \mathcal{F}_n$),

5. Shouldn't f and f_0 be in $\mathcal{F}$?

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

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

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

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 between two evaluation functionals of f is shown to satisfy

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

where $C \in \mathcal{F} \otimes \mathcal{F}$ is the covariance element of f. 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 of h for $\mathcal{F}$, the first term on the right-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 right-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,

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

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

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

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

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

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

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 $\mathbf{H}$ is the kernel matrix.

3.6 Rate of convergence

Should I say something about this? Rates can be better than GPR?

3.7 Conclusion

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

The dimension of the function space $\mathcal{F}$ could be huge, infinite-dimensional even, while the task of estimating $f \in \mathcal{F}$ only relies on a finite amount of data point. 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 a RKHS with reproducing kernel equal to the Fisher information for f. Importantly, functions in the subspace $\mathcal{F}_n$ are structurally similar to the functions 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.

Miscellanea

3.7.1 Total Fisher information

For many applications, it is of interest to evaluate the (total) Fisher information at the maximum likelihood estimate under a sampling scenario. However, the expectation required to calculate the Fisher information above cannot be done without knowing the true value of θ. As a point of clarification, we ought to make the distinction between the *expected* Fisher information and the *observed* Fisher information under a sampling scenario. There are two quantities that are typically used as an approximation, and these are explained below. Let $y = \{y_1, \ldots, y_n\}$ represent an independent and identically distributed observed sample from $p(\cdot|\theta)$. The maximum likelihood (ML) estimator $\hat{\theta} = \arg\max_\theta L(\theta)$ for θ satisfies the first order conditions $S(\hat{\theta}) = 0$, where the log-likelihood function and the score function makes use of all of the observed samples, i.e. $L(\theta) = \sum_{i=1}^n \log p(y_i|\theta)$. In a sampling experiment, the total Fisher information (denoted $\mathcal{I}_n(\theta)$) is just n times the unit Fisher information, i.e. $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$. Following Efron and Hinkley (1978), the expected Fisher information is defined to be $\mathcal{I}_n(\hat{\theta})$, while the observed Fisher information is

$$\hat{\mathcal{I}}_n = -\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} \log p(y_i|\theta)\bigg|_{\theta=\hat{\theta}}.$$

which is also by definition the negative Hessian. Note that

$$\mathcal{J}(\theta) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} \log p(y_i|\theta)$$

$\frac{1}{n}\hat{\mathcal{I}}_n \to \mathcal{I}(\hat{\theta})$ in probability as $n \to \infty$ by the weak law of large numbers. Both of these quantities are used as replacements of the actual Fisher information about the "true" parameter. In the context of measuring curvatures, the expected Fisher information would be used (Pawitan, 2001), but in the context of efficient variance for ML estimates, the observed Fisher information is favoured (Efron and Hinkley, 1978). which by the law of large numbers, converges in probability to the expected Fisher information $\mathcal{I}(\theta)$ as defined above. In practice, one would not be able to calculate $\mathcal{I}$ without knowing the true value for θ, so replacing occurrences of θ with (the MLE)

In particular, near the MLE, low Fisher information indicates a shallow maxima, while high observed information indicates a "sharp" maxima. A shallow maxima is an indication that many nearby values have similar log-likelihood, but a sharp maxima is

indicative of a high confidence surrounding the MLE.

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-th

3.7.2 Functional derivatives

Definition 3.6 (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 3.7 (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

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

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

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

3.7.3 Data dependent priors

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).

Bibliography

Balakrishnan, Alampallam V (1981). *Applied Functional Analysis*. 2nd ed. Vol. 3. Springer Science & Business Media. 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.

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

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

Jaynes, Edwin 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.

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.

Tapia, R A (1971). *The differentiation and integration of nonlinear operators*. Ed. by Louis B Rall.

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