يونيبرسيتي بروني دارالسلام
UNIVERSITI BRUNEI DARUSSALAM

Regression modelling using I-priors

NUS Department of Statistics & Data Science Seminar

Haziq Jamil

Mathematical Sciences, Faculty of Science, UBD

https://haziqj.ml

Wednesday, 16 November 2022

Overview

Reproducing kernel Hilbert spaces

Assumption: $f \in \mathcal{F}$ *where* $\mathcal{F}$ *is an RKHS with kernel* h *over* $\mathcal{X}$.

Definition 1 (Hilbert spaces)

A *Hilbert space* $\mathcal{F}$ is a vector space equipped with a positive definite inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$.

Definition 2 (Reproducing kernels)

A symmetric, bivariate function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *kernel*, and it is a *reproducing kernel* of $\mathcal{F}$ if h satisfies

i. $\forall x \in \mathcal{X}$, $h(\cdot, x) \in \mathcal{F}$;
ii. $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{F}$, $\langle f, h(\cdot, x) \rangle_{\mathcal{F}} = f(x)$.

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

Reproducing kernel Hilbert spaces (cont.)

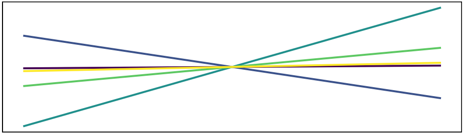
Theorem 3 (Moore-Aronszajn, etc.)

There is a bijection between

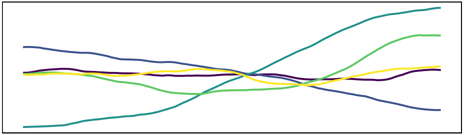
i. the set of positive semidefinite functions; and
ii. the set of RKHSs.

$h(x, x') = 1$ (constant)

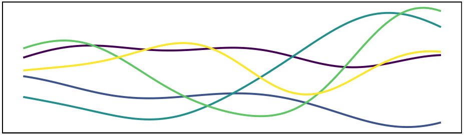
$h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$ (linear)



$h(x, x') = -\frac{1}{2}\left(\|x - x'\|_{\mathcal{X}}^{2\gamma} - \|x\|_{\mathcal{X}}^{2\gamma} - \|x'\|_{\mathcal{X}}^{2\gamma}\right)$ (fBm)



$h(x, x') = \exp\left(-\frac{\|x - x'\|_{\mathcal{X}}^{2\gamma}}{2s^2}\right)$ (Gaussian)



Building more complex RKHSs

We can build complex RKHSs by adding and multiplying kernels:

- $\mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2$ is an RKHS defined by $h = h_1 + h_2$.
- $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$ is an RKHS defined by $h = h_1 h_2$.

Example 4 (ANOVA RKHS)

Consider RKHSs $\mathcal{F}_k$ with kernel h_k, $k = 1, \dots, p$. The ANOVA kernel over the set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ defining the ANOVA RKHS $\mathcal{F}$ is

$$h(x, x') = \prod_{k=1}^{p} \big(1 + h_k(x, x')\big).$$

For $p = 2$ let $\mathcal{F}_k$ be linear RKHS of functions over $\mathbb{R}$. Then $f \in \mathcal{F}$ where $\mathcal{F} = \mathcal{F}_\emptyset \oplus \mathcal{F}_1 \oplus \mathcal{F}_2 \oplus \mathcal{F}_1 \otimes \mathcal{F}_2$ are of the form

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

The Fisher information

For the regression model (**??**), the log-likelihood of f is given by

$$\ell(f|y) = \text{const.} - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \psi_{ij}\big(y_i - \langle f, h(\cdot, x_i)\rangle_{\mathcal{F}}\big)\big(y_j - \langle f, h(\cdot, x_j)\rangle_{\mathcal{F}}\big)$$

Lemma 5 (Fisher information for regression function)

The Fisher information for f is

$$\mathcal{I}_f = -\operatorname{E}\nabla^2 \ell(f|y) = \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(\cdot, x_i) \otimes h(\cdot, x_j)$$

where '$\otimes$' is the tensor product of two vectors in $\mathcal{F}$.

The Fisher information (cont.)

It's helpful to think of $\mathcal{I}_f$ as a bilinear form $\mathcal{I}_f : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$, making it possible to compute the Fisher information on linear functionals $f_g = \langle f, g \rangle_{\mathcal{F}}, \forall g \in \mathcal{F}$ as $\mathcal{I}_{f_g} = \langle \mathcal{I}_f, g \otimes g \rangle_{\mathcal{F} \otimes \mathcal{F}}$.

In particular, between two points $f_x := f(x)$ and $f_{x'} := f(x')$ [since $f_x = \langle f, h(\cdot, x) \rangle_{\mathcal{F}}$] we have:

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

The I-prior

Lemma 6

The kernel (1) induces a finite-dimensional RKHS $\mathcal{F}_n < \mathcal{F}$, consisting of functions of the form $\tilde{f}(x) = \sum_{i=1}^n h(x, x_i) w_i$ (for some real-valued w_is) equipped with the squared norm

$$\|\tilde{f}\|^2_{\mathcal{F}_n} = \sum_{i,j=1}^n \psi_{ij}^- w_i w_j,$$

where ψ_{ij}^- is the (i, j)th entry of Ψ^{-1}.

- Let $\mathcal{R}$ be the orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$. Then $\mathcal{F} = \mathcal{F}_n \oplus \mathcal{R}$, and any $f \in \mathcal{F}$ can be uniquely decomposed as $f = \tilde{f} + r$, with $\tilde{f} \in \mathcal{F}_n$ and $r \in \mathcal{R}$.
- The Fisher information for g is zero iff $g \in \mathcal{R}$. The data only allows us to estimate $f \in \mathcal{F}$ by considering functions in $\tilde{f} \in \mathcal{F}_n$.

The I-prior (cont.)

Theorem 7 (I-prior)

Let ν be a volume measure induced by the norm above, and let

$$\tilde{p} = \arg\max_{p} \left\{ -\int_{\mathcal{F}_n} p(f) \log p(f) \, \nu(\mathrm{d}f) \right\}$$

subject to the constraint

$$\mathrm{E}_{f \sim p} \| f - f_0 \|^2_{\mathcal{F}_n} = \text{constant}, \qquad f_0 \in \mathcal{F}.$$

Then $\tilde{p}$ is the Gaussian with mean f_0 and covariance function $k(x, x')$.

Equivalently, under the I-prior, f can be written in the form

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

References