Statistics 860 Lecture 5

Let $\mathcal{T}$ be an index set and let $\mathcal{H}$ be an RKHS of real-valued functions on $\mathcal{T}$.

This means that all the point evaluations

$$Lf \to f(t_*)$$

are bounded. Equivalently, if $f \in \mathcal{H}$, then for each $t_* \in \mathcal{T}$, $\exists M_{t_*}$ such that

$$|f(t_*)| \leq M_{t_*} \|f\| \ \ \forall f$$

Thus we may associate with $\mathcal{H}$ the unique positive definite function $K(s,t)$ on $\mathcal{T} \times \mathcal{T}$ given by

$$K(s,t) = \langle \xi_s, \xi_t \rangle$$

where ξ_t is the representer of evaluation at t:

$$f(t_*) \equiv \langle f, \xi_{t_*} \rangle \ \ \forall f \in \mathcal{H}.$$

$K(s,t)$ is called the "reproducing kernel" for $\mathcal{H}$.

Conversely [Moore-Aronszajn Theorem], let $K(s,t)$ be a positive definite function on $\mathcal{T} \times \mathcal{T}$. We may identify with K a unique RKHS with K as its reproducing kernel.

PKOOF: We construct the space.
Let $K_t(\cdot)$ be the function given by

$$K_t(s) = K(s,t).$$

Let all functions of the form

$$f(s) = \sum_{l=1}^{L} c_l \, K_{t_l}(s)$$

be in $\mathcal{H}$ for any $L = 1, 2, \ldots,$
$t_1, \ldots \; t_L \in \mathcal{T}$, and $c_1, \ldots, c_L$.

Define the inner product by

$$\langle K_{t_l}, K_{t_k} \rangle = K(t_l, t_k)$$

and extend by linearity:

If

$$f_1(s) = \sum_{l=1}^{L} c_l \, K_{t_l}(s)$$

$$f_2(s) = \sum_{k=1}^{K} d_k \, K_{s_k}(s)$$

then

$$\langle f_1, f_2 \rangle = \sum_{l=1}^{L} \sum_{k=1}^{K} c_l d_k \, K(t_l, s_k).$$

It can be verified that the positive-definiteness of K is enough to insure that this has all the properties of an inner product.

$\mathcal{H}_K$ will be the RKHS with reproducing kernel K.

This collection of functions is a linear manifold–it will be a Hilbert space if we add to it all the limits of all the Cauchy sequences:

Let

$$f_n = \sum_{l=1}^{n} c_{ln} K_{t_{ln}} \quad n = 1, 2, \ldots$$

be a Cauchy sequence–that means

$$\|f_n - f_m\|^2 \to 0 \quad \text{as } n, m \to \infty$$

Now

$$|f_n(t) - f_m(t)| = |\langle f_n - f_m, K_t \rangle|$$

(by the inner product we have defined) and so

$$\leq \|f_n - f_m\| \|K_t\|.$$

Let $f(t) = \lim_n f_n(t)$ –add to space.

Important Remark:

Note that K_t is the representer of evaluation in $\mathcal{H}_K$.

$$K_t \equiv \xi_t$$

Reproducing kernel:

$$\langle K_s, f \rangle = f(s)$$

In particular

$$\langle K_s, K_t \rangle = K_t(s) = K(s,t) \quad !$$

"Reproducing" Property

IMPORTANT REMARK

All this works with no assumptions on the nature of $\mathcal{T}$!!

Mercer–Hilbert–Schmidt Theorems:
(see `/pdf1/riesz.nagy.pdf`)

Suppose $K(s,t)$ is a symmetric, positive definite function on $\mathcal{T} \times \mathcal{T}$ with

$$\int\int_{\mathcal{T}\times\mathcal{T}} K^2(s,t)\; ds\; dt < \infty$$

Then $\exists$ a sequence of eigenvalues $\{\lambda_\nu\}$, and orthonormal eigenfunctions $\{\phi_\nu\}$ in $\mathcal{L}_2(\mathcal{T})$ such that

$$\int_{\mathcal{T}} K(s,t)\phi_\nu(t)\; dt = \lambda_\nu\; \phi_\nu(s)$$

and

$$K(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu\; \phi_\nu(s)\; \phi_\nu(t)$$

Then,

$$\int\int_{\mathcal{T}\times\mathcal{T}} K^2(s,t)\,ds\,dt = \sum_{\nu=1}^{\infty} \lambda_\nu^2$$

Why?

$$\int\int_{\mathcal{T}\times\mathcal{T}} K^2(s,t)\,ds\,dt =$$

$$\sum_{\mu}\sum_{\nu}\int\int\, ds\,dt\,\lambda_\nu\lambda_\mu\,\phi_\nu(s)\phi_\nu(t)\phi_\mu(s)\phi_\mu(t)$$

Use

$$\int \phi_\nu(s)\phi_\mu(s)\,ds = \begin{cases} 0 & \mu \neq \nu \\ 1 & \mu = \nu \end{cases}$$

$$\begin{aligned} \int_{\mathcal{T}} K(t,t)\,dt &= \sum_{\nu=1}^{\infty}\int \lambda_\nu\,\phi_\nu^2(t)\,dt \\ &= \sum \lambda_\nu \text{ if this is finite} \end{aligned}$$

(K of 'trace class')

Remark: $\mathcal{T}$ can be “anything” on which you can define a positive definite function.

An RK that does not satisfy $\int\int K^2(s,t)dsdt < \infty$ is

$$K(s,t) = e^{-\alpha\|s-t\|^\beta}$$

where s,t in the real line or E^d, $\|\cdot\|$ is the Euclidean norm, and β satisfies $0 < \beta \leq 2$.

These are perfectly good RK’s even though they don’t have a countable sequence of eigenvalues and eigen functions.

The $\beta = 2$ case (‘Gaussian’) is a very popular choice for large d. This RK depends only on the difference $s - t$ and the representers of evaluation are so-called radial basis functions. Other RK’s which depend only on the difference $s - t$ and generate radial basis functions will be discussed later.

Remark: We noted that sines and cosines provide a complete orthonormal sequence for $\mathcal{L}_2[0,1]$. $\mathcal{L}_2(-\infty,\infty)$ does not posess a (countable) complete orthonormal sequence. Instead, we have the Fourier transform

$$f(\nu) = \int_{-\infty}^{\infty} e^{i\nu t} f(t) dt$$

instead of Fourier coefficients a_ν and b_ν, and the family $e^{i\nu t}, \nu \in (-\infty,\infty)$ instead of the countable set of sines and cosines.

Similarly

$$K(s,t) \equiv e^{-|s-t|^2/2} = c \int_{-\infty}^{\infty} e^{i\nu(s-t)} e^{-2\nu^2} d\nu.$$

The $\{e^{-2\nu^2}\}$ play the role of the eigenvalues and $e^{i\nu t}$ the role of eigenfunctions, $\nu \in (-\infty,\infty)$. See the elegant book "Introduction to Hilbert Space and the Theory of Spectral Multiplicity", Paul K. Halmos.

POSITIVE DEFINITE MATRICES:

Let $A_{n\times n}$ and $B_{k\times k}$ be positive definite. Then the tensor product (Kronecker Product) is positive definite. Let $A = \{a_{ij}\}$.

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix}$$

$\boxed{\text{This is part of HW 2}}$

If $k = n$, the Shur product C with ij-th entry

$$C_{ij} = a_{ij} b_{ij}$$

is also positive definite.

$\boxed{\text{This is part of HW 2.}}$

Let K be any matrix. Then KK' and $K'K$ are positive definite matrices and

$$K(s,t) = \int_{\mathcal{U}} G(s,u) G(t,u)\, du$$

is positive definite for any $G(\cdot,\cdot)$ such that $G(s,\cdot) \in \mathcal{L}_2(\mathcal{U})\ \forall s$.

Moore-Aronszajn Theorem:
Let $\mathcal{T}$ be an index set, let $K(s,t)$ be a positive definite function on $\mathcal{T} \times \mathcal{T}$. To every such $K(\cdot,\cdot)$ there corresponds a unique RKHS with$K$ as its RK, and conversely.

Remark:
$K_t(\cdot)$ is the representer of evaluation at t, in $\mathcal{H}_K$,where
$K_t(s) \equiv K(t,s)$

$W_m(per)$, periodic functions with $f^{(m)} \in \mathcal{L}_2[0,1]$:

$$||f||^2 = \sum_{v=1}^{\infty} \frac{(f,\phi_v)^2}{\lambda_v} < \infty$$

$$\langle f, g \rangle = \sum_{v=1}^{\infty} \frac{(f,\phi_v)\,(g,\phi_v)}{\lambda_v}$$

$$\begin{aligned} \phi_0 &= 1 \\ \phi_v &= \sqrt{2}\sin 2\pi v t \\ &= \sqrt{2}\cos 2\pi v t \qquad\qquad v = 1, 2, ... \end{aligned}$$

$$\begin{aligned} \mathcal{H} \ : \ & \lambda_0 = 1 \\ & \lambda_v = (2\pi v)^{-2m} \qquad\qquad v = 1, 2, ... \end{aligned}$$

$$f(t) \sim a_0 + \sqrt{2}\sum a_v \sin 2\pi v t + \sqrt{2}\sum b_v \cos 2\pi v t$$

$$
\begin{aligned}
\sum \frac{(f, \phi_v)^2}{\lambda_v} &< \infty \\
&= \left[\int_0^1 f(t)\,dt\right]^2 + \underbrace{\sum_{v=1}^{\infty} (2\pi v)^{2m} \left(a_v^2 + b_v^2\right)}_{if\ finite} \\
&= \left[\int_0^1 f(t)dt\right]^2 + \int_0^1 \left(f^{(m)}(t)\right)^2 dt
\end{aligned}
$$

$f \in W_m(per)$ if its Fourier coefficients with respect to the sines and cosines satisfies

$$\sum_{v=1}^{\infty} (2\pi v)^{2m}(a_v^2 + b_v^2) < \infty.$$

If $f \in W_m(per)$ it is legitimate to differentiate f $m-1$ times and the series will converge pointwise. So

$$f^{(v)}(1) - f^{(v)}(0) = 0, \qquad v = 0, 1, \cdots, m-1.$$

$W_m(per)$, periodic functions with $f^{(m)} \in \mathcal{L}_2[0,1]$.

Periodic: $f^{\nu-1}(1) - f^{\nu-1}(0) = 0, \nu = 1, 2, \cdots, m$, with
$\|f\|^2 = [\int_0^1 f(u)du]^2 + \int_0^1 (f^{(m)}(u))^2 du$.

Claim:

$$\begin{aligned} K(s,t) &= 1 + 2\sum_{\nu=1}^{\infty} \frac{1}{(2\pi\nu)^{2m}} [\cos(2\pi\nu s)\cos(2\pi\nu t) \\ &\quad + \sin(2\pi\nu s)\sin(2\pi\nu t)] \\ &= 1 + 2\sum_{\nu=1}^{\infty} \frac{1}{(2\pi\nu)^{2m}} \cos 2\pi\nu(s-t) \end{aligned}$$

A closed form for $K(s,t)$ is available using the Bernoulli Polynomials:

..
$$\begin{aligned}
B_0(x) &= 1 \\
B_1(x) &= x - \tfrac{1}{2} \\
B_2(x) &= x^2 - x + \tfrac{1}{6} \\
B_3(x) &= x^3 - \tfrac{3}{2}x^2 + \tfrac{1}{2}x \\
B_4(x) &= x^4 - 2x^3 + x^2 - \tfrac{1}{30} \\
&\vdots
\end{aligned}$$

Abramowitz and Stegun, 1972.

Note that $B_r(1) - B_r(0) = 0, r = 2, 3, ...$

Abramowitz-Stegun: Bernoulli polynomials of even degree B_{2m}:

$$B_{2m}(x) = (-1)^m 2(2m!) \sum_{\nu=1}^{\infty} \frac{1}{(2\pi\nu)^{2m}} \cos 2\pi\nu x$$

so

$$K(s,t) = 1 + \frac{(-1)^{m-1}}{(2m)!} B_{2m}([s-t])$$

where $[s-t]$ is the fractional part of $s-t$. (If $s-t = 1.2$, then $[s-t] = 0.2$) (See `http://mathworld.wolfram.com/BernoulliPolynomial.html`) for formulas.

The RK squared norm for the RKHS with this RK is:

$$\|f\|^2 = [\int_0^1 f(u)du]^2 + \int_0^1 (f^{(m)}(u))^2 du.$$

Convince Yourself.

WHAT DOES THE RK DO FOR YOU?

Let $\mathcal{H}_K$ be the RKHS with RK K. Then $\langle K_t, f\rangle = f(t)$, for all $f \in \mathcal{H}_{\mathcal{K}}$ and each fixed t. You can find the representer of any other bounded linear functional, if you know $K(\cdot,\cdot)$. We illustrate this by an example. We examine the derivative at t as a possible bounded linear functional: Fix t,
$\langle K_{t+\delta} - K_t, f\rangle = f(t+\delta) - f(t)$, each $t, t+\delta \in \mathcal{T}$.

$\frac{1}{\delta}\langle K_{t+\delta} - K_t, f\rangle = \frac{1}{\delta}[f(t+\delta) - f(t)]$

(The divided difference.)

Let $\quad \xi_\delta = \frac{1}{\delta}[K_{t+\delta} - K_t], \quad$ t fixed and $\xi_\delta \in \mathcal{H}$ for any $\delta > 0$

then
$\langle \xi_\delta, f \rangle = \frac{1}{\delta}[f(t+\delta) - f(t)]$

Suppose for any sequence

$$\delta_1, \delta_2, \cdots \to 0,$$

that

$$\xi_{\delta_1}, \xi_{\delta_2}, \cdots$$

is a Cauchy sequence in $\mathcal{H}$. That means that

$$\lim_{\substack{m \to \infty \\ n \to \infty}} \|\xi_{\delta_m} - \xi_{\delta_n}\| \to 0$$

and, for each fixed s,

$$|\xi_{\delta_m}(s) - \xi_{\delta_n}(s)| \leq \|\xi_{\delta_m} - \xi_{\delta_n}\| \|K_s\| \to 0$$

so that, $\lim_{n\to\infty} \xi_{\delta_n}(s) = \xi(s)$ exists.

Recalling,

$$\begin{aligned}\xi_{\delta_n}(s) &= \frac{1}{\delta_n}[K_{t+\delta_n}(s) - K_t(s)] \\ &= \frac{1}{\delta_n}[K(t+\delta_n, s) - K(t, s)]\end{aligned}$$

we have

$$\begin{aligned}\lim_{n\to\infty} \xi_{\delta_n}(s) &= \frac{\partial}{\partial u}K(u, s)|_{u=t} \quad (if \quad it \quad exists) \\ &\equiv \xi_t(s)\end{aligned}$$

Suppose this $\xi_t \in \mathcal{H}$, then

$$\langle \xi_t, f \rangle = \lim_{n\to\infty} \frac{1}{\delta_n}[f(t+\delta_n) - f(t)] = \frac{\partial}{\partial u} f(u)|_{u=t}$$

Lemma: Let $\xi_t(s) = \frac{\partial}{\partial u}K(u, s)|_{u=t} \in \mathcal{H}$, then

$$Lf = f^{'}(t)$$

is a **bounded linear functional** in $\mathcal{H}$ with representer $\xi_t(s) = \frac{\partial}{\partial u}K(u, s)|_{u=t}$

MORE GENERALLY:

Let L be a bounded linear functional in $\mathcal{H}$, then $Lf = \langle \eta, f \rangle$ for some $\eta \in \mathcal{H}$.

What is η??

$$\eta(s) = \langle \eta, K_s \rangle$$

equivalently,

$$\begin{aligned} \eta(s) &= LK_s \\ &\equiv L_{(u)} K(s,u) \end{aligned}$$

where $L_{(u)}$ means L applied to what follows considered as a function of u,(u is the dummy variable).

To check whether a particular L is a bouned linear functional. If η given by the formula :

$$\eta(s) = L_{(u)} K(s, u)$$

is an element of $\mathcal{H}$, then L is a bounded linear functional in $\mathcal{H}$ with representer $\eta(\cdot)$

Another example:
Let

$$Lf = \int w(u) f(u) du$$

If

$$\xi(s) = \int w(u) K_s(u) du = \int w(u) K(s, u) du$$

is in $\mathcal{H}$, then L is a **bounded linear functional in** $\mathcal{H}$ with representer ξ.

Consider $W_m(per)$:

Convince yourself that

$$Lf = f^{(\nu)}(t_*)$$

is a bounded linear functional for fixed t_*in $[0, 1]$ and $\nu = 0, 1, \cdots, m-1$.
That is,

$$L_{(u)}K_s(u) = \frac{\partial^{(\nu)}}{\partial u^{(\nu)}}K(s, u)|_{u=t_*}$$

considered as a function of s, is in $\mathcal{H}$, for fixed t_*.

The first variational problem:

The statistical model is

$$f \in \mathcal{H}$$

$L_i, i = 1, \cdots, n$ are bounded linear functionals in $\mathcal{H}$. One observes

$$y_i = L_i f + \epsilon_i, \quad i = 1, \cdots, n$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{pmatrix} \sim N(0, \sigma^2 I)$$

f_λ, the 'penalized least squares' estimate of f, is the minimizer in $\mathcal{H}$ of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - L_i f)^2 + \lambda \|f\|^2, \text{ where } \lambda > 0$$

The first variational problem:
Find $f \in \mathcal{H}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - L_i f)^2 + \lambda \|f\|^2, \text{ where } \lambda > 0$$

Let η_i be the representer of L_i,

$$L_i f \equiv \langle \eta_i, f \rangle.$$

Let $\sigma_{ij} = \langle \eta_i, \eta_j \rangle \equiv L_{i(u)} L_{j(v)} K(u, v)$

$$\Sigma = \{\sigma_{ij}\}$$

$\Sigma \succeq 0$, since it is a Gram matrix,

$$a' \Sigma a = \| \sum a_i \eta_i \|^2 \geq 0$$

Theorem:

$$f_\lambda = \sum_{i=1}^{n} c_i \eta_i$$

with $c = (\Sigma + n\lambda I)^{-1} y$.

Proof:(follows Kimeldorf and Wahba, 1971). Any element g in $\mathcal{H}$ can be represented as

$$g = \sum_{j=1}^{n} c_j \eta_j + \rho$$

where $\rho \in \mathcal{H}$,$\rho \perp \eta_j$,$j = 1, 2, \cdots, n$
Set

$$f_\lambda = \sum_{j=1}^{n} c_j \eta_j + \rho$$

and solve for $c = (c_1, \cdots, c_n)'$ and ρ.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \eta_i, \sum_{j=1}^{n} c_j\eta_j + \rho\rangle)^2 +$$
$$\lambda \sum_{i,j} c_i c_j \langle \eta_i, \eta_j \rangle + \lambda \|\rho\|^2$$

(using $\quad \langle \eta_i, \rho \rangle = 0$)

$$\equiv \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n} c_j \langle \eta_j, \eta_i \rangle\right)^2 + \lambda c' \Sigma c + \lambda \|\rho\|^2$$

$$\equiv \frac{1}{n}\|y - \Sigma c\|^2 + \lambda c' \Sigma c + \lambda \|\rho\|^2$$

minimize over c,ρ,$\Rightarrow$ $\|\rho\|^2 = 0$

minimize c

$$\tfrac{1}{n}\|y - \Sigma c\|^2 + \lambda c'\Sigma c \qquad (*)$$

If Σ is of full rank,then c is unique

$$(*) = \tfrac{1}{n}[y'y - 2y'\Sigma c + c'\Sigma\Sigma c] + \lambda c'\Sigma c$$

$$\frac{1}{2}\frac{\partial}{\partial c}(*) = 0 = -\Sigma y + \Sigma^2 c + n\lambda\Sigma c \qquad (**)$$

or $\quad y = (\Sigma + n\lambda I)c$

$$c = (\Sigma + n\lambda I)^{-1} y$$

where we have multiplied $(**)$ by Σ^{-1}

Suppose Σ is not of full rank: since

$$\Sigma = \{\sigma_{ij}\} = \langle \eta_i, \eta_j \rangle$$

that means there exist $(c_1, \cdots, c_n)$ such that

$$\textstyle\sum_{i=1}^{n} c_i \eta_i \equiv 0$$

and we only know

$$(\Sigma^2 + n\lambda\Sigma)c = \Sigma y \qquad (***)$$

any two solutions to $(***)$ differ by u with $\Sigma u = 0$. Therefore if $u = (u_1, \cdots, u_n)'$, then $\| \sum_{i=1}^{n} u_i \eta_i \|^2 = u'\Sigma u = 0$ and $\Sigma u_i \eta_i = 0$.

So if

$$(\Sigma^2 + n\lambda\Sigma)c = \Sigma y$$

AND

$$(\Sigma^2 + n\lambda\Sigma)\tilde{c} = \Sigma y$$

$$\|\Sigma c_i\eta_i - \Sigma\tilde{c}_i\eta_i\|^2 = 0$$

so that $f_\lambda = \sum_{i=1}^n c_i\eta_i$ is the minimizer of

$$\frac{1}{n}\|y_i - L_i f\|^2 + \lambda\|f\|^2$$

and is unique even though its representation may not be.

Where do I get

$$
\begin{aligned}
\sigma_{ij} &= \langle \eta_i, \eta_j \rangle \\
\eta_i(s) &= L_{i(u)} K_s(u) \\
\langle \eta_j, \eta_i \rangle &= L_j \eta_i \\
&= L_{j(t)} L_{i(s)} K(s,t) \qquad \text{!!}
\end{aligned}
$$

Examples:

$$
\begin{aligned}
L_i f &= f(t_i), \qquad i = 1, \cdots, n \\
\eta_i &= K_{t_i} \\
\langle \eta_i, \eta_j \rangle &= \langle K_{t_i}, K_{t_j} \rangle = K(t_i, t_j)
\end{aligned}
$$

$$
\begin{aligned}
L_i f &= \int w_i(u) f(u) du \\
\eta_i(\cdot) &= \int w_i(u) K(u, \cdot) du \\
\langle \eta_i, \eta_j \rangle &= \int \int w_i(u) w_j(v) K(u,v) du dv
\end{aligned}
$$