Linear Spaces

For elements $f, g, h, \ldots$, define the operation of **addition** satisfying (i) $f + g = g + f$, (ii) $(f + g) + h = f + (g + h)$, and (iii) For every $f, g$, $\exists h$ such that $f + h = g$; (iii) implies the existence of an element 0 satisfying $f + 0 = f, \forall f$.

Further, define the operation of **scalar multiplication** satisfying $\alpha(f + g) = \alpha f + \alpha g$, $(\alpha + \beta)f = \alpha f + \beta f$, $1f = f$, and $0f = 0$, where $\alpha, \beta$ are real.

A set $\mathcal{L}$ of such elements form a **linear space** if $f, g \in \mathcal{L}$ implies that $f + g \in \mathcal{L}$ and $\alpha f \in \mathcal{L}$, for $\alpha$ real.

A set of elements $f_i \in \mathcal{L}$ are **linearly independent** if $\sum_i \alpha_i f_i = 0$ holds only for $\alpha_i = 0, \forall i$. The **dimension** of $\mathcal{L}$ is the maximum number of elements that can be linearly independent.

Functional and Bilinear Form

A **functional** in $\mathcal{L}$ operates on an element $f \in \mathcal{L}$ and returns a real value. A **linear functional** $L$ in $\mathcal{L}$ satisfies $L(f + g) = Lf + Lg$, $L(\alpha f) = \alpha Lf, \forall f, g \in \mathcal{L}, \forall \alpha$ real.

A **bilinear form** $J(f, g)$ in $\mathcal{L}$ takes $f, g \in \mathcal{L}$ as arguments and returns a real value, satisfying $J(\alpha f + \beta g, h) = \alpha J(f, h) + \beta J(g, h)$, $J(f, \alpha g + \beta h) = \alpha J(f, g) + \beta J(f, h), \forall f, g, h \in \mathcal{L}, \forall \alpha, \beta$ real.

Fixing $f$, a bilinear form $J(f, g)$ is a linear functional in $g$. A bilinear form $J(\cdot, \cdot)$ is **symmetric** if $J(f, g) = J(g, f)$, **non-negative definite** if $J(f, f) \geq 0, \forall f \in \mathcal{L}$, and **positive definite** if $J(f, f) = 0$ holds only for $f = 0$. For $J(\cdot, \cdot)$ n.n.d., $J(f) = J(f, f)$ is a **quadratic functional**.
Inner Product, Norm, Distance

A linear space is often equipped with an **inner product**, a p.d. bilinear form $(\cdot, \cdot)$. An inner product defines a **norm** in the linear space, $\|f\| = \sqrt{(f, f)}$, which induces a metric to measure the **distance** between elements in the space, $D[f, g] = \|f - g\|$.

There hold the **Cauchy-Schwarz inequality**,  
$$ |(f, g)| \leq \|f\| \|g\|, $$
with equality if and only if $f = \alpha g$, and the **triangle inequality**,  
$$ \|f + g\| \leq \|f\| + \|g\|, $$
with equality if and only if $f = \alpha g$ for some $\alpha > 0$.

Convergence, Continuity, Hilbert Spaces

A sequence $\{f_n\}$ **converges** to its limit point $f$, $\lim_{n \to \infty} f_n = f$ or $f_n \to f$, if $\lim_{n \to \infty} \|f_n - f\| = 0$.

A functional $L$ is **continuous** if $\lim_{n \to \infty} Lf_n = Lf$ whenever $\lim_{n \to \infty} f_n = f$; $(f, g)$ is continuous in $f$ or $g$.

A **Cauchy sequence** $\{f_n\}$ satisfies $\lim_{n, m \to \infty} \|f_n - f_m\| = 0$. A linear space $\mathcal{L}$ is **complete** if every Cauchy sequence in $\mathcal{L}$ converges to an element in $\mathcal{L}$.

An element is a **limit point of a set** $A$ if it is the limit point of a sequence in $A$. A set $A$ is **closed** if it contains its own limit points.

A **Hilbert space** $\mathcal{H}$ is a complete inner product linear space. A closed linear subspace of $\mathcal{H}$ is itself a Hilbert space.
The distance between $f \in \mathcal{H}$ and a closed linear subspace $\mathcal{G} \subset \mathcal{H}$ is $D[f, \mathcal{G}] = \inf_{g \in \mathcal{G}} \|f - g\|$. There exists $f_\mathcal{G} \in \mathcal{G}$, the projection of $f$ in $\mathcal{G}$, such that $\|f - f_\mathcal{G}\| = D[f, \mathcal{G}]$.

One has $(f - f_\mathcal{G}, g) = 0$, $\forall g \in \mathcal{G}$. The closed linear subspace $\mathcal{G}^c = \{f : (f, g) = 0, \forall g \in \mathcal{G}\}$ is the orthogonal complement of $\mathcal{G}$. The unique decomposition $f = f_\mathcal{G} + f_{\mathcal{G}^c}$, $\forall f \in \mathcal{H}$, forms a tensor sum decomposition $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}^c$.

A n.n.d. $J(f, g)$ defines a semi-inner-product inducing a square seminorm $J(f) = J(f, f)$, with null space $\mathcal{N}_J = \{f : J(f) = 0\}$. One may define $\tilde{J}(f, g)$, satisfying (i) p.d. in $\mathcal{N}_J$ and (ii) for every $f$, $\exists g \in \mathcal{N}_J$ such that $\tilde{J}(f - g) = 0$, to make $(J + \tilde{J})(f, g)$ p.d.; $\mathcal{N}_J \oplus \mathcal{N}_J$ forms a tensor sum decomposition.

### Hilbert Space Example

**Vector Space**: Functions on $\{1, \ldots, K\}$ are vectors of length $K$.

Consider the Euclidean inner product $(f, g) = \sum_{x=1}^{K} f(x)g(x) = f^T g$.

The space $\mathcal{G} = \{f : f(1) = \cdots = f(K)\}$ is a closed linear subspace with an orthogonal complement $\mathcal{G}^c = \{f : \sum_{x=1}^{K} f(x) = 0\}$.

Write $\bar{f} = \sum_{x=1}^{K} f(x)/K$. The bilinear form

$$J(f, g) = \sum_{x=1}^{K} (f(x) - \bar{f})(g(x) - \bar{g}) = f^T (I - \bar{1} \bar{1}^T / K) g$$

defines a semi-inner-product in the vector space with a null space $\mathcal{N}_J = \{f : f(1) = \cdots = f(K)\}$. Define $\tilde{J}(f, g) = Cf^T (\bar{1} \bar{1}^T / K) g \propto \bar{f} \bar{g}$.

One has an inner product in the vector space,

$$(f, g) = (J + \tilde{J})(f, g) = f^T (I + (C - 1) \bar{1} \bar{1}^T / K) g,$$

which reduces to the Euclidean inner product when $C = 1$. On $\mathcal{N}_J^c = \{f : \sum_{x=1}^{K} f(x) = 0\}$, $J(f, g)$ is a full inner product.
Hilbert Space Example

\[ L_2 \text{ Space:} \] Consider \( L^2[0,1] = \{ f : \int_0^1 f^2 dx < \infty \} \) with an inner product \( (f,g) = \int_0^1 fg dx \). The space
\[ G = \{ f : f = g1_{x \leq 0.5}, g \in L^2[0,1] \} \]
is a closed linear subspace with an orthogonal complement
\[ G^\perp = \{ f : f = g1_{x \geq 0.5}, g \in L^2[0,1] \}; \] elements in \( L^2[0,1] \) are defined by equivalent classes.

The bilinear form \( J(f,g) = \int_0^{0.5} fg dx \) defines a semi-inner-product in \( L^2[0,1] \), with null space \( N_J = \{ f : f = g1_{x \geq 0.5}, g \in L^2[0,1] \} \). Define \( \tilde{J}(f,g) = C \int_{0.5}^1 fg dx \). One has an inner product on \( L^2[0,1] \),
\[ (f,g) = (J + \tilde{J})(f,g) = \int_0^{0.5} fg dx + C \int_{0.5}^1 fg dx. \]
On \( N_J^\perp = L^2 \ominus N_J \), \( J(f,g) \) is a full inner product.

Riesz Representation, Reproducing Kernel

\[ \diamond \text{ Riesz Representation Theorem:} \] For every continuous linear functional \( L \) in a Hilbert space \( H \), there exists a unique \( g_L \in H \), the representer of \( L \), such that \( Lf = (g_L,f), \forall f \in H \).

Consider a Hilbert space \( H \) of functions on domain \( X \). If the evaluation functional \( [x]f = f(x) \) is continuous in \( H \), \( \forall x \in X \), then \( H \) is a reproducing kernel Hilbert space. [The likelihood part of penalized likelihood functional typically involves evaluations.]

By the Riesz representation theorem, there exists \( R_x \in H \), the representer of \([x](\cdot)\), such that \( (R_x,f) = f(x) \), \( \forall f \in H \). The reproducing kernel \( R(x,y) = R_x(y) = (R_x,R_y) \) has the reproducing property \( (R(x,\cdot),f(\cdot)) = f(x) \).
Reproducing Kernel Hilbert Spaces

The $L_2[0,1]$ space is not an RKHS. Since elements in $L_2[0,1]$ are defined not by individual functions but only by equivalent classes, evaluation is not even well defined.

Consider the vector space with the Euclidean inner product $(f,g) = f^T g$. The vectors are functions on $X = \{1, \ldots, K\}$, and the evaluation $[x]f = f(x)$ is coordinate extraction. Since $f(x) = e_x^T f$, with $e_x$ the $x$th unit vector, one has $R_x(y) = I_{[x=y]}$. A bivariate function on $\{1, \ldots, K\}$ can be written as a square matrix, and the RK in the Euclidean space is simply $I$.

• A finite-dimensional Hilbert space is always a reproducing kernel Hilbert space, as all linear functionals are continuous.

Non-Negative Definite Function and RK

A bivariate function $F(x, y)$ on $X$ is non-negative definite if $\sum_{i,j} \alpha_i \alpha_j F(x_i, x_j) \geq 0$, $\forall x_i \in X$, $\forall \alpha_i$ real.

For $R(x, y) = R_x(y)$ an RK, $\| \sum_i \alpha_i R_x \|^2 = \sum_{i,j} \alpha_i \alpha_j R(x_i, x_j)$.

◊ Theorem: For every RKHS $\mathcal{H}$ of functions on $X$, there corresponds an unique RK $R(x, y)$, which is n.n.d. Conversely, for every n.n.d. function $R(x, y)$ on $X$, there corresponds a unique RKHS $\mathcal{H}$ that has $R(x, y)$ as its RK.

◊ Theorem: If RK $R$ of $\mathcal{H}$ on $X$ decomposes into $R = R_0 + R_1$ with $R_0$ and $R_1$ n.n.d., $R_0(x, \cdot), R_1(x, \cdot) \in \mathcal{H}$, $\forall x \in X$, and $(R_0(x, \cdot), R_1(y, \cdot)) = 0$, $\forall x, y \in X$, then $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_i$ corresponds to $R_i$. Conversely, if $R_0$ and $R_1$ are n.n.d. and $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$, then $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ has an RK $R = R_0 + R_1$. 
Elements of an RKHS $\mathcal{H}$ of functions on domain $\mathcal{X}$ with an RK $R(x, y) = R_x(y)$ are linear combinations $\sum \alpha_i R_{x_i}$ and their limits. Much like a vector space as the column space of some matrix, an RKHS is “generated” from the “columns” $R_x = R(x, \cdot)$ of the RK, for which any n.n.d. function on $\mathcal{X}$ qualifies.

Recall the penalized least squares functional

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda J(f).$$

$J(f)$ shall be taken as a quadratic functional with a finite-dimensional null space $\mathcal{N}_J = \{ f : J(f) = 0 \}$. $J(f)$ is typically a square seminorm in an RKHS.

Consider the column space of a $K \times K$ n.n.d. matrix $B$, $\mathcal{H}_B = \{ f : f = Bc = \sum_j c_j B(\cdot, j) \}$, equipped with the inner product $(f, g) = f^T B g$.

Since $B^+ B = B B^+$ is the projection matrix onto $\mathcal{H}_B$, $B^+ B f = f$, $\forall f \in \mathcal{H}_B$. One has $[x] f = f(x) = e_x^T f = e_x^T B^+ B f = (B^+ e_x)^T B f$, $\forall f \in \mathcal{H}_B$, thus the representer of $[x](\cdot)$ is the $x$th column of $B^+$, and hence the RK of $\mathcal{H}_B$ is $R(x, y) = B^+(x, y)$.

Consider a decomposition of RK in the Euclidean space, $I_{[x=y]} = 1/K + (I_{[x=y]} - 1/K)$, or $I = (11^T/K) + (I - 11^T/K)$. This defines a tensor sum decomposition $\mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 = \{ f : f(1) = \cdots = f(K) \}$ and $\mathcal{H}_1 = \{ f : \sum_{x=1}^{K} f(x) = 0 \}$. A one-way ANOVA is built in with $Af = \sum_{x=1}^{K} f(x)/K$. 

C. Gu Spring 2006
Discrete Splines: Shrinkage Estimates

A spline on $\mathcal{X} = \{1, \ldots, K\}$ can be defined as the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \eta(x_i))^2 + \lambda \eta^T B \eta,$$

which is a shrinkage estimate.

- With $B = I - 11^T/K$, or $f^T B f = \sum_{x=1}^{K} (f(x) - \bar{f})^2$, $f(x)$ are shrunk towards the mean $\bar{f}$; the penalty appears natural for a nominal $x$.

- With $f^T B f = \sum_{x=2}^{K} (f(x) - f(x-1))^2$, $f(x)$ at adjacent levels are shrunk towards each other; the penalty appears natural for an ordinal $x$. The null space of $f^T B f$ is still $\{f : f(1) = \cdots = f(K)\}$, but the internal “scaling” of $\mathcal{H}_B$ is different.

Polynomial Smoothing Splines

Consider the space $C^{(m)}[0, 1] = \{f : f^{(m)} \in L_2[0, 1]\}$ on $\mathcal{X} = [0, 1]$. A polynomial smoothing spline is the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \eta(x_i))^2 + \lambda \int_0^1 (\eta^{(m)})^2 dx.$$

It is a piecewise polynomial of order $2m - 1$ ($m - 1$ beyond the first and last knots), with up to the $(2m - 2)$nd derivatives continuous.

$J(f) = \int_0^1 (f^{(m)})^2 dx$ is a square seminorm in $C^{(m)}[0, 1]$, with polynomials of orders up to $m - 1$ as its null space $\mathcal{N}_j$.

- With $m = 2$, one has the cubic splines.

- With $m = 1$, one has the linear splines, broken lines that are flat beyond the first and last knots.
A Reproducing Kernel in $C^{(m)}[0,1]$

For $f \in C^{(m)}[0,1]$, the Taylor expansion gives

$$f(x) = \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!} f^{(\nu)}(0) + \int_0^1 \frac{(x-u)^{m-1}}{(m-1)!} f^{(m)}(u) du,$$

where $(\cdot)_+ = \max(0, \cdot)$. With an inner product

$$(f,g) = \sum_{\nu=0}^{m-1} f^{(\nu)}(0) g^{(\nu)}(0) + \int_0^1 f^{(m)} g^{(m)} dx,$$

the representer of evaluation $[x](\cdot)$ is seen to be

$$R_x(y) = \sum_{\nu=0}^{m-1} \frac{x^\nu y^{\nu}}{\nu! \nu!} + \int_0^1 \frac{(x-u)^{m-1}}{(m-1)!} \frac{(y-u)^{m-1}}{(m-1)!} du.$$

The RK decomposes naturally, and $J(f) = \int_0^1 (f^{(m)})^2 dx$ is p.d. in $\mathcal{H}_J = \{ f : f^{(\nu)}(0) = 0, \nu = 0, \ldots, m-1, \int_0^1 (f^{(m)})^2 dx < \infty \}$.

Linear and Cubic Splines

Setting $m = 1$, the RK becomes

$$R(x,y) = R_0 + R_1 = 1 + \int_0^1 (x-u)_+ (y-u)_+ du = 1 + x \wedge y.$$  

A one-way ANOVA is built in the tensor sum decomposition with $Af = f(0)$.

Setting $m = 2$, the RK becomes

$$R(x,y) = R_{00} + R_{01} + R_1 = 1 + xy + \int_0^1 (x-u)_+ (y-u)_+ du;$$

the RK in $\mathcal{N}_J = \{ f : f = \beta_0 + \beta_1 x \}$ further decomposes into two terms. A one-way ANOVA is built in the tensor sum decomposition with $Af = f(0)$; $R_{01}$ generates the “parametric contrast” and $R_1$ generates the “nonparametric contrast.”
Bernoulli Polynomials

Define periodic, real valued functions

\[ k_r(x) = -\left( \sum_{\mu = -\infty}^{-1} + \sum_{\mu = 1}^{\infty} \right) \frac{\exp(2\pi i \mu x)}{(2\pi i \mu)^r}, \quad r = 1, 2, \ldots, \]

where \( i = \sqrt{-1} \). One has \( k_r^{(p)} = k_{r-p}, \) \( p = 1, \ldots, r - 2 \) and \( k_r^{(r-1)}(x) = k_1(x) \) for \( x \) not an integer; \( k_1(x) = x - 0.5 \) on \((0, 1)\). Set \( k_0(x) = 1 \). These are scaled Bernoulli polynomials \( k_r = B_r/r! \).

- From \( k_1(x) \), the \( k_r(x) \) functions can be obtained by successive integrations; note that \( \int_0^1 k_r(x) \, dx = 0 \). One has
  \[ k_2(x) = \frac{1}{2} \left( k_1^2(x) - \frac{1}{12} \right), \]
  \[ k_4(x) = \frac{1}{24} \left( k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240} \right). \]

Another Reproducing Kernel in \( C^{(m)}[0, 1] \)

For \( f \in C^{(m)}[0, 1] \), it can be shown that

\[ f(x) = \sum_{\nu=0}^{m-1} k_\nu(x) \int_0^1 f^{(\nu)}(y) \, dy + \int_0^1 (k_m(x) - k_m(x-y)) f^{(m)}(y) \, dy. \]

With an alternative inner product

\[ (f, g) = \sum_{\nu=0}^{m-1} (\int_0^1 f^{(\nu)}(x) \, dx) (\int_0^1 g^{(\nu)}(x) \, dx) + \int_0^1 f^{(m)}(x) g^{(m)}(x) \, dx, \]

the RK is given by

\[ R_x(y) = [\sum_{\nu=1}^{m-1} k_\nu(x) k_\nu(y)] + [k_m(x) k_m(y) + (-1)^{m-1} k_{2m}(x-y)]. \]

- The different norm in \( N_J \) changes the composition of \( H_J = C^{(m)}[0, 1] \cap N_J \); the side conditions are now \( \int_0^1 f^{(\nu)}(x) \, dx = 0, \) \( \nu = 0, \ldots, m - 1. \)
Linear and Cubic Splines

Setting $m = 1$, the RK becomes

$$R(x, y) = R_0 + R_1 = 1 + [k_1(x)k_1(y) + k_2(x - y)]$$

A one-way ANOVA is built in the tensor sum decomposition with $Af = \int_0^1 f dx$.

Setting $m = 2$, the RK becomes

$$R(x, y) = R_{00} + R_{01} + R_1 = 1 + k_1(x)k_1(y) + [k_2(x)k_2(y) - k_4(x - y)]$$

the RK in $\mathcal{N}_J = \{f : f = \beta_0 + \beta_1 x\}$ further decomposes into two terms. A one-way ANOVA is built in the tensor sum decomposition with $Af = \int_0^1 f dx$; $R_{01}$ generates the “parametric contrast” and $R_1$ generates the “nonparametric contrast.”

Solution Expression, Computation

Write $\eta \in \mathcal{C}^{(m)}[0, 1]$ as

$$\eta(x) = \sum_{\nu=1}^m d_\nu \phi_\nu(x) + \sum_{i=1}^n c_i R_J(x_i, x) + \rho(x),$$

where $\mathcal{N}_J = \text{span}\{\phi_\nu\}$, $R_J$ is the RK in $\mathcal{H}_J = \mathcal{C}^{(m)}[0, 1] \ominus \mathcal{N}_J$, and $J(R_J(x_i, \cdot), \rho) = \rho(x_i) = 0$, $i = 1, \ldots, n$. The penalized least squares problem reduces to

$$(Y - Sd - Qc)^T(Y - Sd - Qc) + n\lambda c^TQc + n\lambda J(\rho),$$

where $S_{i,\nu} = \phi_\nu(x_i)$ and $Q_{i,j} = R_J(x_i, x_j)$; at the minimum $\rho = 0$.

- One would need a basis of $\mathcal{N}_J$ and the RK in $\mathcal{H}_J$, but nothing else. In particular, an explicit $J(f)$ is *not* needed.
Tensor Product RKHS

\[ \textbf{Theorem:} \text{ For } R_{(1)}(x_{(1)},y_{(1)}) \text{ n.n.d. on } \mathcal{X}_1 \text{ and } R_{(2)}(x_{(2)},y_{(2)}) \text{ n.n.d. on } \mathcal{X}_2, R(x,y) = R_{(1)}(x_{(1)},y_{(1)})R_{(2)}(x_{(2)},y_{(2)}) \text{ is n.n.d. on } \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2. \]

Given \( \mathcal{H}_{(1)} \) on \( \mathcal{X}_1 \) with RK \( R_{(1)} \) and \( \mathcal{H}_{(2)} \) on \( \mathcal{X}_2 \) with RK \( R_{(2)} \), the space corresponding to \( R = R_{(1)}R_{(2)} \) on \( \mathcal{X}_1 \times \mathcal{X}_2 \) is the tensor product space \( \mathcal{H}_{(1)} \otimes \mathcal{H}_{(2)}. \)

Given \( \mathcal{H}_{(\gamma)} = \mathcal{H}_{0(\gamma)} \oplus \mathcal{H}_{1(\gamma)} \) with built-in one-way ANOVAs, one may construct tensor product space with multi-way ANOVA,

\[
\mathcal{H} = \bigoplus_{\gamma=1}^{\Gamma} (\mathcal{H}_{0(\gamma)} \oplus \mathcal{H}_{1(\gamma)}) = \bigoplus_{S} \left\{ \bigotimes_{\gamma \in S} \mathcal{H}_{1(\gamma)} \otimes \bigotimes_{\gamma \not\in S} \mathcal{H}_{0(\gamma)} \right\} = \bigoplus_{S} \mathcal{H}_{S}.
\]

TPRKHS on Discrete Domain

A function on \( \{1, \ldots, K_1\} \times \{1, \ldots, K_2\} \) can be written as a vector of length \( K_1K_2 \),

\[
f = (f(1,1), \ldots, f(1,K_2), \ldots, f(K_1,1), \ldots, f(K_1,K_2))^T,
\]

and an RK as a \( (K_1K_2) \times (K_1K_2) \) matrix.

Based on \( I = (11/K) + (I - 11/K) \), one has

<table>
<thead>
<tr>
<th>Subspace</th>
<th>Reproducing Kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{H}<em>{0(1)} \otimes \mathcal{H}</em>{0(2)} )</td>
<td>( (1_{K_1}1_{K_1}^T/K_1) \otimes (1_{K_2}1_{K_2}^T/K_2) )</td>
</tr>
<tr>
<td>( \mathcal{H}<em>{0(1)} \otimes \mathcal{H}</em>{1(2)} )</td>
<td>( (1_{K_1}1_{K_1}^T/K_1) \otimes (I_{K_2} - 1_{K_2}1_{K_2}^T/K_2) )</td>
</tr>
<tr>
<td>( \mathcal{H}<em>{1(1)} \otimes \mathcal{H}</em>{0(2)} )</td>
<td>( (I_{K_1} - 1_{K_1}1_{K_1}^T/K_1) \otimes (1_{K_2}1_{K_2}^T/K_2) )</td>
</tr>
<tr>
<td>( \mathcal{H}<em>{1(1)} \otimes \mathcal{H}</em>{1(2)} )</td>
<td>( (I_{K_1} - 1_{K_1}1_{K_1}^T/K_1) \otimes (I_{K_2} - 1_{K_2}1_{K_2}^T/K_2) )</td>
</tr>
</tbody>
</table>
Tensor Product Linear Splines

Based on $R_0 + R_1 = 1 + x \wedge y$ on $[0, 1]$, one has the product RKs

<table>
<thead>
<tr>
<th></th>
<th>$H_{0(2)}$</th>
<th>$H_{1(2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{0(1)}$</td>
<td>1</td>
<td>$x_{(2)} \wedge y_{(2)}$</td>
</tr>
<tr>
<td>$H_{1(1)}$</td>
<td>$x_{(1)} \wedge y_{(1)}$</td>
<td>$(x_{(1)} \wedge y_{(1)}) (x_{(2)} \wedge y_{(2)})$</td>
</tr>
</tbody>
</table>

with $A_1 f = f(0, x_{(2)})$ and $A_2 f = f(x_{(1)}, 0)$ in two-way ANOVA.

Based on $R_0 + R_1 = 1 + [k_1(x)k_1(y) + k_2(x - y)]$, one has a similar construction but the averaging operators become $A_1 f = \int_0^1 f dx_{(1)}$ and $A_2 f = \int_0^1 f dx_{(2)}$.

- One may use different marginal RKs, which may imply different averaging operators on different axes.

Tensor Product Cubic Splines

Based on $R_{00} + R_{01} + R_1 = 1 + k_1(x)k_1(y) + [k_2(x)k_2(y) - k_4(x - y)]$, one has the product RKs

<table>
<thead>
<tr>
<th></th>
<th>$H_{00(2)}$</th>
<th>$H_{01(2)}$</th>
<th>$H_{1(2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{00(1)}$</td>
<td>1</td>
<td>$R_{01(2)}$</td>
<td>$R_{1(2)}$</td>
</tr>
<tr>
<td>$H_{01(1)}$</td>
<td>$R_{01(1)}$</td>
<td>$R_{01(1)}R_{01(2)}$</td>
<td>$R_{01(1)}R_{1(2)}$</td>
</tr>
<tr>
<td>$H_{1(1)}$</td>
<td>$R_{1(1)}$</td>
<td>$R_{1(1)}R_{01(2)}$</td>
<td>$R_{1(1)}R_{1(2)}$</td>
</tr>
</tbody>
</table>

A two-way ANOVA is built in with $A_1 f = \int_0^1 f dx_{(1)}$ and $A_2 f = \int_0^1 f dx_{(2)}$; the main effects each contain two terms and the interaction contains four.

- As with linear splines, one may use different marginal RKs that may imply different averaging operators.
Consider $\mathcal{H} = \oplus_\beta \mathcal{H}_\beta$, with $\mathcal{H}_\beta$ having inner products $(f,g)_\beta$ and RKs $R_\beta$. Scaling and adding $(f,g)_\beta$, an inner product in $\mathcal{H}$ and the associated RK are given by

$$J(f,g) = \sum_\beta \theta_\beta^{-1} (f_\beta,g_\beta)_\beta, \quad R_J = \sum_\beta \theta_\beta R_\beta,$$

where $f_\beta$ and $g_\beta$ are projections of $f$ and $g$ in $\mathcal{H}_\beta$.

When some of the $\theta_\beta$ are set to $\infty$, $J(f,g)$ defines a semi-inner-product in $\mathcal{H} = \oplus_\beta \mathcal{H}_\beta$, which may be used to specify the penalty $J(f) = J(f,f)$.

Subspaces not contributing to $J(f)$ form the null space $\mathcal{N}_J = \{f : J(f) = 0\}$. Subspaces contributing to $J(f)$ form the space $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$.

__Tensor Product Cubic Splines__

Consider $\mathcal{H} = \oplus_{\nu,\mu} \mathcal{H}_{\nu,\mu}$, $\nu,\mu = 00,01,1$, where $\mathcal{H}_{\nu,\mu} = \mathcal{H}_{\nu(1)} \otimes \mathcal{H}_{\mu(2)}$ with inner products $(f,g)_{\nu,\mu}$ and RKs $R_{\nu,\mu} = R_{\nu(1)} R_{\mu(2)}$. One may set

$$J(f,g) = \theta_{1,00}^{-1} (f,g)_{1,00} + \theta_{00,1}^{-1} (f,g)_{00,1}$$

$$+ \theta_{1,01}^{-1} (f,g)_{1,01} + \theta_{01,1}^{-1} (f,g)_{01,1} + \theta_{1,1}^{-1} (f,g)_{1,1},$$

with a null space $\mathcal{N}_J = \text{span}\{1,k_1(x_{(1)}),k_1(x_{(2)}),k_1(x_{(1)})k_1(x_{(2)})\}$.

The minimizer of the penalized LS problem has an expression

$$\eta(x) = \sum_{\nu,\mu=00,01} d_{\nu,\mu} \phi_{\nu,\mu}(x) + \sum_{i=1}^n c_i R_J(x_i, x),$$

where $\phi_{\nu,\mu}$ are given in the expression of $\mathcal{N}_J$ and

$$R_J = \theta_{1,00} R_{1,00} + \theta_{00,1} R_{00,1} + \theta_{1,01} R_{1,01} + \theta_{01,1} R_{01,1} + \theta_{1,1} R_{1,1}.$$
Shrinkage Estimates as Bayes Estimates

On $\mathcal{X} = \{1, \ldots, K\}$, consider $\eta = \alpha 1 + \eta_1$, with independent priors $\alpha \sim N(0, \tau^2)$ for the mean and $\eta_1 \sim N(0, b(I - 11^T/K))$ for the contrast; $\eta_1^T 1 = 0$ a.s. and $\bar{\eta} = \sum_{x=1}^{K} \eta(x)/K = \alpha$. The posterior mean $E[\eta|Y]$ is given by the minimizer of

$$
\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \eta(x_i))^2 + \frac{1}{\tau^2} \bar{\eta}^2 + \frac{1}{b} \sum_{x=1}^{K} (\eta(x) - \bar{\eta})^2.
$$

Letting $\tau^2 \to \infty$ and setting $b = \sigma^2/n\lambda$, one gets the shrinkage estimate (slide 13) with $B = I - 11^T/K = (I - 11^T/K)^+$. 

- In the limit, $\alpha$ is said to have a diffuse prior.
- This setting may be perceived as a mixed-effect model, with $\alpha 1$ being the fixed effect and $\eta_1$ being the random effect.

Polynomial Splines as Bayes Estimates

Consider $\eta = \eta_0 + \eta_1$ on $\mathcal{X} = [0, 1]$, with $\eta_0$ and $\eta_1$ having independent Gaussian priors with mean 0 and covariance functions

$$
E[\eta_0(x)\eta_0(y)] = \tau^2 \sum_{\nu=1}^{m} \phi_{\nu}(x)\phi_{\nu}(y), \quad E[\eta_1(x)\eta_1(y)] = b R_1(x, y).
$$

Observing $Y_i \sim N(\eta(x_i), \sigma^2)$, it can be shown that

$$
E[\eta(x)|Y] = (b \xi^T + \tau^2 \phi^T S^T)(bQ + \tau^2 SS^T + \sigma^2 I)^{-1} Y,
$$

where $Q_{i,j} = R_1(x_i, x_j)$, $S_{i,\nu} = \phi_{\nu}(x_i)$, and $\xi_i = R_1(x_i, x)$. Setting $n\lambda = \sigma^2/b$ and letting $\tau^2 \to \infty$, one has

$$
E[\eta(x)|Y] = \phi^T d + \xi^T c,
$$

where $c$ and $d$ minimize the penalized LS score (slide 20)

$$
(Y - Sc - Qc)^T (Y - Sc - Qc) + n\lambda c^T Qc.
$$
Smoothing Splines as Bayes Estimates

Consider an RKHS $\mathcal{H} = \oplus_{\beta=0}^p \mathcal{H}_\beta$ on $\mathcal{X}$ with an inner product 
\[(f,g) = \sum_{\beta=0}^p \theta^{-1}_\beta (f,g)_\beta = \sum_{\beta=0}^p \theta^{-1}_\beta (f_\beta,g_\beta)\]
and an RK $R(x,y) = \sum_{\beta=0}^p \theta_\beta R_\beta(x,y)$; assume a finite-dimensional $\mathcal{H}_0$.

Observing $Y_i \sim N(\eta(x_i),\sigma^2)$, a smoothing spline on $\mathcal{X}$ can be defined as the minimizer in $\mathcal{H}$ of the penalized LS functional
\[\frac{1}{n} \sum_{i=1}^n (Y_i - \eta(x_i))^2 + \lambda \sum_{\beta=1}^p \theta^{-1}_\beta (\eta,\eta)_\beta.\]

The solution is the posterior mean of $\eta = \sum_{\beta=0}^p \eta_\beta$, where $\eta_0$ is diffuse in $\mathcal{H}_0$ and $\eta_\beta$, $\beta = 1,\ldots,p$, have independent mean 0 Gaussian priors with covariance functions $E[\eta_\beta(x)\eta_\beta(y)] = b \theta_\beta R_\beta(x,y)$, where $b = \sigma^2/n\lambda$.

- Perceived as a mixed-effect model, $\eta_0$ contains the fixed effects and $\eta_\beta$, $\beta = 1,\ldots,p$, are the random effects.

Minimization of Penalized Functional

A functional $A(f)$ in $\mathcal{L}$ is Fréchet differentiable if $\dot{A}_{f,g}(0)$ exists and is linear in $g$, $\forall f,g \in \mathcal{L}$, where $A_{f,g}(\alpha) = A(f + \alpha g)$ is a function of $\alpha$ real. $A(f)$ is convex if $A(\alpha f + (1-\alpha)g) \leq \alpha A(f) + (1-\alpha)A(g)$, $\forall \alpha \in (0,1)$.

◊ Existence Theorem: Suppose $L(f)$ is a continuous and convex functional in a Hilbert space $\mathcal{H}$ and $J(f)$ is a square (semi) norm in $\mathcal{H}$ with a finite-dimensional null space $\mathcal{N}_J$. If $L(f)$ has a unique minimizer in $\mathcal{N}_J$, then $L(f) + (\lambda/2)J(f)$ has a minimizer in $\mathcal{H}$.

◊ Theorem: Let $L(f)$ be continuous, convex, and Fréchet differentiable in a Hilbert space $\mathcal{H}$ with a square (semi) norm $J(f)$. If $f^*$ minimizes $L(f)$ in $C_\rho = \{ f : J(f) \leq \rho \}$, then $f^*$ minimizes $L(f) + (\lambda/2)J(f)$ in $\mathcal{H}$, where $\lambda = -\rho^{-1}\dot{L}_{f^*,f^*}(0) \geq 0$, with $f^*$ the projection of $f^*$ in $\mathcal{H} \ominus \mathcal{N}_J$.

Conversely, if $f^o$ minimizes $L(f) + (\lambda/2)J(f)$ in $\mathcal{H}$, then $f^o$ minimizes $L(f)$ in $\{ f : J(f) \leq J(f^o) \}$.