STA547: Functional Data Analysis*

Zhiling Gu

Contents

*Instructor: Xiongtao Dai, Iowa State University

1 Functional Principal Component Analysis

Consider data $X(t) \in L^2(t), t \in \mathcal{T}$

FPCA is defined analogously to a finite dimensional PCA. Below let $\mathbb{H} = L^2$. The total variation of X is defined as

$$\begin{aligned} E\|X-\mu\|^2 &= E\int_{\mathcal{T}}(X(t)-\mu(t))^2dt \\ &\overset{Funibi}{=} \int_{\mathcal{T}} E(X(t)-\mu(t))^2dt \\ &= \int_{\mathcal{T}} G(t,t)dt \\ \sum_{k=1}^{\infty}\lambda_k &= \sum_{k=1}^{\infty}\int_{\mathcal{T}}\lambda_k\phi_k(t)\phi_k(t)dt \\ &= \int_{\mathcal{T}}\sum_{k=1}^{\infty}\lambda_k\phi_k(t)\phi_k(t)dt \end{aligned}$$

We want to find a good k-dim representation for X.

Let $e_1, e_2, \ldots$ be a fixed orthonormal basis, then we have the expansion

$$X-\mu = \sum_{k=1}^{\infty}\langle X-\mu, e_k\rangle \cdot e_k$$

Or informally we write in poitwise version.

$$X(t) = \mu(t) + \sum_{k=1}^{\infty}\langle X-\mu, e_k\rangle \cdot e_k(t) \tag{1}$$

Convergence in the mean behaviour does not imply pointwise convergence for $\forall t \in \mathcal{T}$

The approximation by the first k terms is

$$\mu(t) + \sum_{k=1}^{K}\langle X-\mu, e_k\rangle \cdot e_k(t)$$

with the total variation

$$\begin{aligned} E\|\sum_{k=1}^{K}\langle X-\mu, e_k\rangle e_k\|^2 &= \sum_{k=1}^{K} E\langle X-\mu, e_k\rangle^2 \\ &= \sum_{k=1}^{K}\int_{\mathcal{T}}\int_{\mathcal{T}} G(t,s)e_k(t)e_k(s)dtds \end{aligned}$$

Theorem: $\max \sum_{k=1}^{K}\int_{\mathcal{T}}\int_{\mathcal{T}} G(t,s)e_k(t)e_k(s)dtds = \sum_{k=1}^{K}\lambda_k$. where the maximum is taken over all orthonormal basis. Maximum achieved at $e_i = \phi_i$. (λ_k, ϕ_k) is the k-th eigen value and eigen function of the covariance function G.

The FPCA takes $\phi_1, \ldots, \phi_k$ as the first k basis functions (or projection directions). We call basis coefficients

$$\xi_k := \int_{\mathcal{T}}(X(t)-\mu(t))\phi_k(t)dt$$

the k-th functional principal component (projection score).

We can use $(\xi_1, \ldots, \xi_k) \in \mathbb{R}^k$ to represent X. Alternatively, the k-dim process $X_k(t) = \mu(t) + \sum_{k=1}^{K}\xi_k\phi_k(t), t \in \mathcal{T}$ to represent $X(t) = \mu(t) + \sum_{k=1}^{\infty}\xi_k\phi_k(t)$

- Can be used to simulate random functions

- FPCA procedure is nested

Theorem: (1) $E\xi_k = 0$, (2) $\text{Var}(\xi_k) = \lambda_k$, (3) $\text{Cov}(\xi_k, \xi_k') = \lambda_k \xi_{kk'}$, (4) $\text{Cov}(\xi_k, X(t) = \lambda_k \phi_k(t))$

The explained variance by the first k FPC is $E\|X_k - \mu\|^2 = \sum_{k=1}^{K} \lambda_k$.

The fraction of explained variance (FVE) by the first k components is

$$FVE(k) = \frac{\sum_{k=1}^{K} \lambda_k}{\sum_{k=1}^{\infty} \lambda_k}$$

1.1 Gaussiality (TBC)

1.2 Plug-in Estimation

Let $X_1, \ldots, X_n$ be a sample of IID realization of X. The mean and covariance functions are estimated by

$$\hat{\mu}(t) = \sum_i X_i(t)/n \quad \hat{G}(t,s) = \frac{1}{n-1}(X_i(t) - \hat{\mu}(t))(X_i(s) - \hat{\mu}(s)), t, s \in \mathcal{T}$$

The (eigen value, eigen function) pairs are estimated by those of $\hat{G}$, which satisfies

$$\int_{\mathcal{T}} \hat{G}(t,s)\hat{\phi}_k(t)ds = \hat{\lambda}_k \hat{\phi}_k(t)$$

with $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and the $\hat{\phi}_k$ are the associated orthonormal eigen functions.

The sample FPCs are $\hat{\xi}_{ik} = \int_{\mathcal{T}} (X_i(t) - \hat{\mu}(t))\hat{\phi}_k(t)dt, i = 1, \ldots, n, k = 1, 2, \ldots$. The truncated representation of the i-th subject is estimated by $\hat{X}_{ik}(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_k(t)$.

Theorem: Let $\bar{\hat{\xi}_k} := \frac{1}{n}\sum_{i=1}^{n} \hat{\xi_{ik}}$, we have:
(1) $\bar{\hat{\xi}_k} = 0$
(2) $\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\xi}_{ik} - \bar{\hat{\xi}_k})^2 = \hat{\lambda}_k$
(3) $\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\xi}_{ik} - \bar{\hat{\xi}_k})(\hat{\xi_{ik'}} - \bar{\hat{\xi}_{k'}}) = \lambda_k \hat{\xi}_{kk'}$

1.3 Implementation for Dense Obsevation

Suppose we have dense and regular observations on grid points $a = t_1 < \ldots < t_n = b$ on $\mathcal{T} = [a, b]$. Let $\mathcal{T}_m = \{t_1, \ldots, t_m\}$ denote the discretized points. The sample mean $\hat{\mu}$ and covariance $\hat{G}$ are obtained on $\mathcal{T}_m$ instead of $\mathcal{T}$. For the eigen analysis, the integral is approximated by Riemann Sums.

If $t_1, \ldots, t_m$ are equally spaced, one can apply the midpoint rule and approximate the integral of $f \in L^2$ by $\int_a^b f(t)dt = \frac{b-a}{m}\sum_{j=1}^{m} f(t_j)$. The integral operator associated with $\hat{G}$ is

$$\int_a^b \hat{G}(t,s)\hat{\phi}(s)ds = \frac{b-a}{m}\sum_{j=1}^{m} \hat{G}(t,t_j)\hat{\phi}(t_j), t \in \mathcal{T}_m$$

Solve

$$\frac{b-a}{m}\sum_{j=1}^{m} \hat{G}(t,t_j)\hat{\phi}_k(t_j) = \hat{\lambda}_k \hat{\phi}_k(t_l), l = 1, \ldots, m$$

such that

$$\int_a^b \hat{\phi}_k(t)\hat{\phi}_{k'}(t)dt \approx \frac{b-a}{m}\sum_{j=1}^{m} \phi_k(t_j)\phi_{k'}(t_j) = \xi_{kk'}$$

Let $\hat{G}_{m\times m} = \left[\hat{G}(t_k, t_j)\right]_{k,j=1}^{m}$, $\hat{\phi}_k = \left[\hat{\phi}_k(t_1), \ldots, \hat{\phi}_k(t_m)\right]'$.

$$\frac{b-a}{m}\hat{G}\hat{\phi}_k = \hat{\lambda}_k \hat{\phi}_k$$

such that

$$\frac{b-a}{m}\hat{\phi}_k^T\hat{\phi}_k = \xi_{kk'}, k, k' = 1, \ldots, m$$

Write $\hat{G} = \sum_{k=1}^{K}\sigma_j\lambda_k v_k^T$, the eigen decomposition of $\hat{G}$, where (σ_k, v_k) is the k-th eigen value eigen vector pair. Then the solution to the above linear system is given by

$$\hat{\lambda}_k = \frac{b-a}{m}\sigma_k \quad \hat{\phi}_k = v_k/\sqrt{\frac{b-a}{m}}$$

The FPCs are obtained by

$$\hat{\xi}_{jk} = \frac{b-a}{m}\sum_{i=1}^{n}(X_i(t_j) - \hat{\mu}(t_j))\hat{\phi}_k(t_j)$$

Demonstration in R:

- Finding the eigen value is equivalent to ???max rotation?
- FPCs can be used for classification/ prediction
- Eigen values have to have faster converfence than the harmonic sequence because $\sum_i \lambda_i < \infty$
- Eigen value decays faster, smaller explained variance for later FPCs leading to smaller weights.

2 Hilbert Space

c.f. HE Chapter 2-4

2.1 Linear Operators and Functionals

(17 Sep 2019)

Let $\mathbb{X}_1, \mathbb{X}_2$ be normed linear spaces with norm $\|\cdot\|_1, \|\cdot\|_2$. Let $\mathcal{T} : \mathbb{X}_1 \to \mathbb{X}_2$ be a linear transformation s.t. $\mathcal{T}(ax+b) = a\mathcal{T}(x) + b, x \in \mathbb{X}, a, b \in \mathbb{R}$. Denote the image(or range) of $\mathcal{T}$ as

$$Im(\mathcal{T}) = \{\mathcal{T}x : x \in \mathbb{X}_1\} \subset \mathbb{X}_2$$

the kernel as

$$Ker(\mathcal{T}) = \{x \in \mathbb{X}_1 : \mathcal{T}x = 0\} \subset \mathbb{X}_1$$

the rank of $\mathcal{T}$ be

$$rank(\mathcal{T}) = \dim(Im(\mathcal{T}))$$

The linear transformation $\mathcal{T}$ is bounded if there exists a finite $c > 0$ s.t.

$$\|\mathcal{T}x\|_2 \leq c\|x\|_1$$

for all $\forall x \in \mathbb{X}_1$

Let $B(\mathbb{X}_1, \mathbb{X}_2)$ be the set of bounded linear transformations from $\mathbb{X}_1$ to $\mathbb{X}_2$. This is a vector space under $a\mathcal{T}_1 + b\mathcal{T}_2 + c)(x) := a\mathcal{T}_1 x + b\mathcal{T}_2 x + c$. Write $B(\mathbb{X}) = B(\mathbb{X}, \mathbb{X})$. Define zero element to be an operator mapping everything to 0.

Inner product: : Let $\mathbb{H}$ ne a Hilbert space, $h \in \mathbb{H}, \mathcal{T}h = \langle \cdot, h\rangle$. It is an element of $\mathcal{B}(\mathbb{H}, \mathbb{R})$ since $\langle x, h\rangle \leq \|x\|\|h\|$ by taking $c = \|h\| < \infty$.

When equipped with the operator norm

$$\|\mathcal{T}\| = \sup_{x\in\mathbb{X}_1, \|x\|_1=1}\|\mathcal{T}x\|_2$$

, $B(\mathbb{X}_1, \mathbb{X}_2)$ becomes a normed linear space. For $\forall x \in \mathbb{X}_1, \|\mathcal{T}x\|_2 \leq \|\mathcal{T}\|\|x\|_1$, and the elements of $B(\mathbb{X}_1, \mathbb{X}_2)$ is called bounded linear operators (BLO)

Inner product: $\|\mathcal{T}_h\| = \sup_{\|X\|=1}\langle x, h\rangle \leq_{CS} \sup_{\|X\|=1}\|X\|\|h\| = \|h\|$. The equality is achieved at $x = h/\|h\|$ so that $\|\mathcal{T}_h\| = \|h\|$

Identity Operator: Let $T : \mathbb{X} \to \mathbb{X}$ be $Ix = x$, then $\|I\| = \sup_{\|X\|=1}\|IX\| = \sup_{\|X\|=1}\|X\| = 1$

Integral Operator: Consider $\mathbb{X} = L^2[0,1]$ and the linear mapping defined as

$$(\mathcal{T}f)(\cdot) = \int_0^1 k(\cdot, u)f(u)du$$

for $f \in L^2[0,1]$ and some $K \in L^2([0,1] \times [0,1])$. Operators of this type is called integral operators. The function K is called kernel. Now

$$|(\mathcal{T}f)(t)|^2 = (\int_0^1 K(t,u)f(u)du)^2 \tag{2}$$

$$\leq \int_0^1 K(t,u)^2 du \int_0^1 f^2(u)du \tag{3}$$

$$= \int_0^1 K(t,u)^2 \|f\| \tag{4}$$

$$\|\mathcal{T}f\|^2 \leq \|f\|^2 \int_0^1 \int_0^1 K(t,u)^2 dudt = \|f\|^2 \cdot c \tag{5}$$

Therefore $\|\mathcal{T}\| < \infty$, so $\mathcal{T} \in B(L^2[0,1], L^2[0,1])$

Theorem: (HE Thm 3.1.2) Linear transformation between two normed space is uniformaly continuous $\iff$ it is bounded.

Theorem: (HE Thm 3.1.3) Let $\mathbb{X}_1, \mathbb{X}_2$ be normed linear spaces. If $\mathbb{X}_2$ is complete, then $B(\mathbb{X}_1, \mathbb{X}_2)$ with the operator norm $\|\mathcal{T}\| = \sup_{x \in \mathbb{X}_1, \|x\|_1 = 1} \|\mathcal{T}x\|_2$ is a Banach space.

Given a normed space X, $B(X, R)$ is called the dual space of X and its elements are called bounded linear functionals (BLF). BLF of a Hilbert space has a particularly simple form.

Theorem: (HE Thm 3.2.1) Suppose $\mathbb{H}$ is a Hilbert space and $T \in B(\mathbb{H}, \mathbb{R})$. There exists a unique element $e_{\mathcal{T}} \in \mathbb{H}$ called the represeter of $\mathcal{T}$ with the property that

$$\mathcal{T}(x) = \langle x, e_{\mathcal{T}} \rangle$$

for all $x \in \mathbb{H}$ and $\|\mathcal{T}\| = \|e_{\mathcal{T}}\|$

left: operator norm, right: Hilbert norm

2.2 Compact Operator (TBC)

c.f HE Chap 4.

A linear transfomation $T : \mathbb{X}_1 \to \mathbb{X}_2$ is compact if for all bounded sequence $x_n \in \{\mathbb{X}_1\}$, $\{\mathcal{T}x_n\}$ contains a convergent subsequence in $\mathbb{X}_2$. Matrices as operators: on $\mathbb{R}^d$ is compact.
Infinite dimensional Hilbert space: is NOT compact. $B = \{h \in \mathbb{H}, \|h\| = 1\}$ is not a compact set in $\mathbb{H}$ which has infinite dimension.
Identity Operator I: on $\mathbb{H}$ is not compact.
Prove by considering a CONS $\{e_n\}_n^\infty$. Any subsequence e_{n_k} of $\{e_n\}$ is not converging, so B is not compact.

Let x_1, x_2 be elements of Hilbert spaces $\mathbb{H}_1, \mathbb{H}_2$. The tensor product operator $(x_1 \otimes_1 x_2) : \mathbb{H}_1 \to \mathbb{H}_2$ is defined by

$$(x_1 \underset{1}{\otimes} x_2)y = \langle x_1, y \rangle_1 x_2 \tag{6}$$

for $y \in \mathbb{H}$. If $\mathbb{H}_1 = \mathbb{H}_2$, we use $\otimes$ instead of $\otimes_1$

(skip)

2.3 Adjoint Operators (TBC)

(skip)

2.4 Eigen Decomposition(TBC)

2.5 Mercer's Theorem(TBC)

2.6 Hilbert Schimidt Operator(TBC)

3 Mean-square continuous process and Karhunen-Leove Theorem

(26 Sep 2019)

Theorem: (c.f. HE Ch7.3)
Let $\{X(t), t \in \mathcal{T}\}$ be a second order stochastic process, i.e. $E(X^2(t) < \infty, t \in \mathcal{T}$, where $\mathcal{T} \subset R$ is a compact interval. Let $L^2(\Omega)$ to denote the space of real-valued r.v. with a finite second moment. Define the inner product $\langle Z, W\rangle = E(ZW)$, for $a, w \in L^2(\Omega)$, then $L^2(\Omega)$ is a Hilbert space with this norm.

Def: A stochastic process is said to be mean square continuous if

$$\lim_{n\to\infty} E(X(t_n) - X(t))^2 \to 0 \qquad (*)$$

for any $t \in \mathcal{T}$ and $t_n \to t$
Let $\mu(t)$ and $G(t, s)$ denote the mean and covariance functions.

Theorem: Let X be a second order process, then X is mean square continuous iff μ an G are continuous functions.

proof. ($\Longrightarrow$)

$$\begin{aligned} E(X(s) - X(t))^2 &= E((X_s - \mu_s) + (\mu_s - \mu_t) + (\mu_t - X_t))^2 \\ &= G(s, s) + (\mu_s - \mu_t)^2 + G(t, t) - 2G(s, t) \end{aligned}$$

So the continuity of μ and G implies when $s \to t$ the RHS of the last equation goes to zero.
($\Longleftarrow$)

$$|\mu_s - \mu_t| = |E(X_s - X_t)| \leq_{CS} (E(X_s - S_t)^2 E(1))^{1/2}$$

Since (*) holds, we have RHS $\to 0$ as $s \to r$ so μ is continous.
Now $G(s, t) = EX_sX_t - \mu_x\mu_t$, WLOG assume $\mu = 0$

$$|G(t, s) - G(t', s')| \leq |G(t, s) - G(t', s)| + |G(t', s) - G(t', s')| \qquad (**)$$

The first of RHS is

$$|G(t, s) - G(t', s)| = |E[(X_t - X_{t'})X_s]| \leq_{CS} (E(X_t - X_{t'})^2 E(X_s)^2)^{1/2}$$

So as $t' \to t$, RHS above $\to 0$.

Similarly, the second term in (**) is

$$|G(t', s) - G(t', s')| \leq_{CS} [E(X_s - X_{s'})^2 EX_{t'}^2]^{1/2} \qquad (***)$$

Since $EX^2(t') = E(X_t + X_{t'} - X_t)^2 \leq 2E(X_t^2 + E(X(t') - X(t)^2)) \to 2EX(t)^2$ as $t \to t'$, we have the RHS of (***) $\to 0$ as $(s', t') \to (s, t)$ Similarly,

$$\lim_{t'\to t, s'\to s} |G(t, s) - G(t', s')| = 0$$

so G is continuous at (t, s).

We now define a L^2-stochastic integral for a mean square continous process $X(t), t \in \mathcal{T}$, without assuming it is a random element, by the last theorem, μ and G are continuous.

Let $P = \{E_i, t_i : 1 \leq i \leq m(n)\}$ where $\{E_i\}_{i=1}^{m(n)}$ is a partition of T s.t. $\cup E_i = T$ and the diameter of each E_i is less than $1/n$. and each t_i is an arbitraty point in E_i .

Want to define a integral operator in the form $I_X(f) = \int X(t)f(t)dt$, need to prove the uniqueness

For any function $f \in L^2(T)$ define

$$I_X(f,P) = \sum_{i=1}^{m(n)} X(t_i) \int_{E_i} f(u)du \tag{7}$$

which is an r.v. in $L^2(\Omega)$

Now let $Q = \{E_i', t_i' : 1 \le i \le m(n')\}$ be another partition with the diameter of each E_i' smaller than $1/n'$

$$E(I_X(f,P) - I_X(f,Q))^2 = \sum_{i=1}^{m(n)} \sum_{j=1}^{m(n)} G(t_i,t_j) \int_{E_i} f(u)du \int_{E_j} f(v)dv \tag{8}$$

$$+ \sum_{i=1}^{m(n')} \sum_{j=1}^{m(n')} G(t_{i'}',t_{j'}') \int_{E_{i'}'} f(u)du \int_{E_{j'}'} f(v)dv \tag{9}$$

$$- 2 \sum_{i=1}^{m(n)} \sum_{j'=1}^{m(n')} G(t_i,t_{j'}') \int_{E_i} f(u)du \int_{E_{j'}'} f(v)dv \tag{10}$$

Each term is an approximation of $\int_{\mathcal{T}} \int_{\mathcal{T}} G(s,t)f(t)f(s)dtds$ as $n, n' \to \infty$. Since $L^2(\Omega$ is complete, $I_X(fP)$ and $I_X(f,Q)$ cannot have different limits. By the completeness of $L^2(\Omega)$, the Cauchy sequence $I_n(f,P)$ converges to a r.v. denoted as $I_n(f) \in L^2(\Omega)$ as $n \to \infty$. This limit does nor depend on the partition.

Theorem: Let $\{X_t, t \in \mathcal{T}\}$ be a mean square continuous process with mean $\mu = 0$, then

1. $E(I_X(f)) = 0$
2. $E(I_X(f)X(t)) = \int_{\mathcal{T}} G(t,s)f(s)ds, t \in \mathcal{T}$
3. $E(I_X(f)I_X(g)) = \int_{\mathcal{T}} \int_{\mathcal{T}} G(t,s)f(t)g(s)dtds$

Proof of 2 for example

$$\begin{aligned}
|E(I_X(f)X(t)) - \int_{\mathcal{T}} G(t,s)f(s)ds| &= \lim_{n\to\infty} |E(I_X(f)X(t)) - \sum_{i=1}^{m(n)} G(t,t_i) \int_{E_i} f(s)ds| \\
&= \lim_{n\to\infty} |E(I_X(f)X(t)) - (\sum_{i=1}^{m(n)} X(t_i) \int_{E_i} f(s)ds)X(t)| \\
&= \lim_{n\to\infty} |E(I_X(f)X(t)) - I_X(f,P)X(t)| \\
&\le_{CS} \lim_{n\to\infty} (EX(t)^2 E(I_X(f) - I_X(f,P))^2)^{1/2} = 0
\end{aligned}$$

because by definition $I_X(f)$ is the limit of $I_X(f,P)$ in $L^2(\Omega)$. □

Recall that the Mercers theorem implies that the continuous covariance function

$$G(t,s) = \sum_{i=1}^{\infty} \lambda_j e_j(t) t_j(s)$$

where $\lambda_1 \ge \lambda_2 \ge \ldots \ge 0$ are eigen values. and the e_j are the orthonormal eigenfunctions. The series converges obsolutely and uniformaly for $s, t, \in \mathcal{T}$.

Corollary: (Here we still assume $\mu = 0$)

1. $E(I_X(e_j)) = 0$
2. $E(I_X(e_j)I_X(e_k)) = \lambda_j \delta_{jk}$

$\delta_j = I_{X-\mu}(e_j) = \int (X(t) - \mu(t))e_j(t)dt$

Theorem: (Karhune Loeve)
Let $\{X(t), t \in \mathcal{T}\}$ be a mean sq cont process. Then

$$\lim_{k\to\infty} \sup_{t\in\mathcal{T}} E(X(t) - X_k(t))^2 = 0$$

where

$$X_K(t) := \mu(t) + \sum_{j=1}^{K} I_{x-\mu}(e_j)e_j(t)$$

alternatively, we say

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} I_{x-\mu}(e_j)e_j(t)$$

where the sum converges in $L^2(\Omega)$ uniformly over $t \in \mathcal{T}$.

Proof.

$$\begin{aligned} E(X(t) - X_K(t))^2 =& E(X(t) - \mu(t) + \mu(t) - X_K(t))^2 \\ =& E(X(t) - \mu(t))^2 + E(\mu(t) - X_K(t))^2 - 2E[(X(t) - \mu(t))(X_K(t) - \mu(t))] \end{aligned}$$

$$\begin{aligned} \text{First term} =& G(t,t) \\ \text{Second term} =& E(\sum_{j=1}^{K} I_{X-\mu}(e_j)e_j(t))^2 \\ =& \sum_{j=1}^{K}\sum_{k=1}^{K} E(I_{X-\mu}(e_k)I_{X-\mu}(e_j))e_j^2(t) \\ =& (Cov)\sum_{j=1}^{K} \lambda_j e_j^2(t) \\ \text{Third term} =& -\sum_{j=1}^{K} 2E(I_{X-\mu}(e_j)X(t))e_j(t) \\ =& -2\sum_{j=1}^{K} \int G(t,s)e_j(s)ds \\ =& -2\sum_{j=1}^{K} \lambda_j e_j^2(t) \end{aligned}$$

So $(\Gamma) = G(t,t) - \sum_{j=1}^{K} \lambda_j e_j^2(t), t \in \mathcal{T}$. This difference goes to zero uniformly for all $t \in \mathcal{T}$ as $n \to \infty$ by the Mercer's theorem. Therefore $\lim_{k\to\infty} \sup_{t\in\mathcal{T}} E(X(t) - X_k(t))^2 = 0$

Brownian Motion: The Brownian motion $\{X(t), t \in [0,1]\}$ is a Gaussian stochastic process with mean 0 and covariance $G(t,s) = \min(t,s)$. To simulate Brownian motion efficiently, we can apply KL expansion.

First we find the eigenvalue and eigen function pairs by solving

$$\int_0^1 \min(t,s)e(s)ds = \lambda e(t), t \in [0,1] \tag{11}$$

So $\int_0^t se(s)ds + t\int_t^1 e(s)ds = \lambda e(t)$. By a previous theorem, $e(\cdot)$ is in the image of the integral operator $\mathscr{G}$ associated with a continuous kernel G, so $e(\cdot)$ is a continuous function. Apply the foundamental theorem of calculus, differentiate both sides, we have

$$\int_t^1 e(s)ds = \lambda e'(t)(2)$$

Differentiate again

$$-e(t) = \lambda e''(t) (3)$$

Now observe that if $\lambda = 0$, then $e(\cdot) = 0$, so 0 cannot be an eigenvalue. The solution to (3) is

$$e(s) = A\sin(s/\sqrt{\lambda}) + B\cos(s/\sqrt{\lambda}), \text{ for some } A, B \in \mathbb{R}$$

Observe $B = 0$ since $e(0) = 0$. By (2), $e'(1) = 0$ so $A\cos(1/\sqrt{\lambda})/\sqrt{\lambda} = 0$, so $1/\sqrt{\lambda} = (2j-1)\pi/2$. This shows

$$\lambda_j = 4/((2j-1)^2\pi^2) \quad e_j(t) = \sqrt{2}\sin(\frac{t(2j-1)\pi}{2}), j = 1, 2, \ldots \tag{12}$$

Here $\xi_j := I_X(e_j)$ is a limit of Gaussian rv in $L^2(\Omega)$, so ξ_j is also Gaussian with mean 0. By a previous corollary, $Var(\xi_j) = E(\xi_j\xi_j) = \lambda_j$, so $\xi_j \sim N(0, 4/((2j-1)^2\pi^2))$. The KL expansion tells us that,

$$X(t) = 0 + \sum_{j=1}^{\infty} \xi_j\sqrt{2}\sin(\frac{(2j-1)\pi t}{2})$$

where the convergence holds for each $t \in [0, 1]$ in $L^2(\Omega)$.

4 Smoothing

(1 Oct 2019)

4.1 Local Polynomial

Drawback of global polynomial regression: dont know which order p to choose, not numerical convergent. A better choice: local polynomial regression

Nonparametric regression is developed in the classical framwork, where the predictor X and the response Y are both real-valued r.v.s. We want to estimate the regression line $\mu(t) := E(Y|X = x)$ which is not assumed to have any parametric form.

Let $(X_i, Y_i), i = 1, \ldots, n$ be a sample of paired observations. A nonparametric regression model is $Y_i = \mu(X_i)+\varepsilon_i$ where the ε_i are iid noise with $E(\varepsilon_i) - 0, Var(\varepsilon_i) = \sigma_i^2 < \infty$. Here we target $\mu : [a, b] \to \mathbb{R}$, the mean function μ is assumed to be smooth in a certain sense. The design points X_i are assumed to lie within $[a, b]$.

Nadaraya Watson (NW) estimator:

$$\widehat{\mu}_{NW}(X_0) = \frac{\sum_{i=1}^n K_h(X_i - X_0)Y_i}{\sum_{j=1}^n K_h(X_j - X_0)}, X_0 \in [a, b]$$

where K_h is a kernel with a bandwidth h. The denominator is a weighting term with sum 1. $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$. The $K(\cdot)$ is ususally a symmetric pdf.

Kernel examples::

- Uniform kernel: $K(x) = \frac{1}{2}I_{[-1,1]}(x)$
- Epanechikov Kernel: $K(x) = \frac{3}{4(1-x^2)}I_{[-1,1]}(x)$
- Gaussian kernel: $K(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$

$\mu(x) \approx \mu(x_0)$ for $x \in [x_0 - h, x_0 + h]$ for h small. small bias, large deviance. Nonparametric regression (NR) is more sensitive to h than to K. When defining an NR estimation, the bandwidth h and K need to be selected. The bandwidth h is by far the more imporatnt one.

$$\begin{aligned} MSE(\hat{\mu}(X_n)) &= E(\hat{\mu}(X_0) - \mu(X_0))^2 \\ &= (E\hat{\mu}(X_0) - \mu(X_0))^2 + Var(\hat{\mu}(X_0) \\ &= bias^2 + Var^2 \end{aligned}$$

Boundary Issue: the regression is always biased. A window centerred at or near a boundary point would contain less data, so $Var(\hat{\mu}(a))$ would be larger than $Var(\hat{\mu}(X_0)$ for $X_0 \in [a, b]$ in practice. A more severe issue is the bias, that $bias(\hat{\mu}(a))$ is a magnitude larger than that for an interior point. $\implies$ Solution (i): truncate the support, report the results for a more limited domain $[a + \epsilon, b - \epsilon]$.

TBC next time.

4.1.1 Local Polynomial Estimation

Local polynomial (LP) estimation extends the NW estimator. The idea is to fit a local polynomial up to degree p to improve approximation.

e.g. Local constant ($p = 0$) estimator : is the NW estimator. Local linear ($p = 1$) estimator : .

The derivatives of $\mu(t)$ can be estimated by the appropriate orders of derivative in the local polynomial. Ingeneral, we approximate

$$\mu(t) \approx \beta_0 + \sum_{j=1}^{p} (X - X_0)^j \beta_j$$

in each local window $X_0 \pm h$

The LP estimation at $x_0 \in [a, b]$ minimizes the locally weighted sum of squares

$$\sum_{i=1}^{n} (Y_i - (\beta_0 + \sum_{j=1}^{p} (X_i - X_0)^j \beta_j))^2 K_h(X_i - X_0)$$

over $\beta_j, j = 0, \ldots, p$

Let

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$X_{n\times(p+1)} = \begin{pmatrix} 1 & x_1 - x_0 & \cdots & x_1 - x_0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_0 & \cdots & x_n - x_0 \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\hat{\beta} = (X' K_h X)^{-1} X' K_h Y$$

The estimate for the local polynomial function around x_0 is

$$\hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j (x - x_0)^j \tag{13}$$

Then $\mu^{(j)}$ is esimated by $\hat{\mu}^{(j)} := \hat{\beta}_j \cdot j!, j = 0, \ldots, p$.

Let $e_j = (0 \ldots 1 \cdots 0)'$, then

$$\hat{\mu}^{(j)}(x_0) = j! e'_{j+1} (X' K_h X)^{-1} X' K_h Y, j = 0, \ldots, p \tag{14}$$

The term before Y does not depend on itself, so it is a linear smoother.

Demonstration in R.

- When the band width is too large, over smoothed, higher bias
- When the band width is too small, outliers will affect some local esimators, too few observations for some less dense intervals, will get something completely random. The solution is to use Gaussian kernel (which has the whole real line as its support, therefore any point can buy information from every other points.)
- Check the design plot (kernel density estimate)

4.2 Basis expansion

Another way to approximate a nonlinear regression function $\mu : [0,1] \to \mathbb{R}$ is through an increasingly complex vector space of functions. Previously a global polynomial regression is

$$E(Y|X) \approx \beta_0 + \beta_1 x + \cdots + \beta_p x^p \tag{15}$$

If we let p to be rather large, then the underlying μ is rather well approximated. Precisely:

Theorem: (Weiersttross Approximation Theorem) Suppose f is continuous function from $[0,1]$ to $\mathbb{R}$. For any $\epsilon > 0$, there exists a polynomial f_p with degree $p = p(\epsilon)$ such that

$$\|f - f_p\|_\infty < \epsilon \tag{16}$$

where for $g : [0,1] \to \mathbb{R}$, $\|g\|_\infty := \sup_{x\in[0,1]} |g(x)|$.

In general, assume we are giecn a set of p basis functions $\{\phi_1, \ldots, \phi_p\}$ where each ϕ_j maps from $[0,1]$ into $\mathbb{R}, j = 1, \ldots, p$. We hope

$$E(Y|X = x) = \mu(x) \approx \sum_{j=1}^{p} \gamma_j \phi_j(x) =: \gamma' \Phi(x)$$

for some $\gamma \in \mathbb{R}^p, \phi(x) = (\phi_1(x) \cdots \phi_p(x))', x \in [0,1]$. By convention, all vector notations are cd vectors. Since $E(Y_i|X_i) = \mu(X_i) \approx \gamma'\phi(X_i)$, we can estimate μ by solving

$$\min \|\boldsymbol{Y} - \Phi\gamma\|$$

where $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\Phi}_{n\times p} = [\phi_j(x_i)]_{i=1,j=1}^{n\ p}$.

The solution is $\hat{\gamma} = (\Phi'\Phi)^{-1}\Phi' Y$. The regression function estimate is $\hat{\mu}(x) = \hat{\gamma}'\phi(x), x \in [0,1]$, and $\hat{Y} = \Phi\hat{\gamma} = \Phi(\Phi'\Phi)^{-1}\Phi' Y$.

Polynomial: . $\phi_j(x) = x^{j-1}, j = 1, \ldots, p, x \in [0,1]$. A severe disadvantage is that when p is large (≥ 20), the estimation becomes numerically unstable. This can be potentially be alleviated by working with the Legendre basis.

Trigonometrix basis: . If $\mu(x)$ is thought to be periodic, we can choose

$$\phi_j = \begin{cases} 1 & j = 1 \\ \sqrt{2}\sin(\pi j x) & j \geq 2 \text{ even} \\ \sqrt{2}\cos(\pi(j-1)x) & j \geq 2 \text{ odd} \end{cases} \quad x \in [0,1]$$

This is ideal if data contains periodic features, e.g. when modeling daily or yearly temperature/ preciitation.

4.3 Regression Splines

More general than polynomial basis is the spline basis defined on $[0,1]$ with interior knots $0 < \mathcal{K}_1 < \cdots < \mathcal{K}_J < 1$. The $\mathcal{K}_j$ are commonly equally spaced. On knot set $\mathcal{K} = [\mathcal{K}_1 \cdots \mathcal{K}_J]$, a spline function $S(x)$ of order d is a polynomial of degree $(d-1)$ between two adjacent knots, and is $(d-2)$ times continuously differentiable on the knots. Here $d = 2$ means the splines are merely continous at the knots.

The collection of all order d splines on knots $\mathcal{K}$ is a (functional) vector space denotes $S(d, \mathcal{K})$. TO calculate the dimension: on each $[\mathcal{K}_i, \mathcal{K}_{i+1}]$, the polynomial function has d degrees of freesom, and the 0-th, ..., (d-2)- th derivative at each of the J knots must match up. So

$$\dim(S(d,\mathcal{K})) = d \times (J+1) - (d-1) \times J = d + J \tag{17}$$

There are many bass that spans $S(d, \mathcal{K})$ such as the truncated polubomial basis

$$B_{trunc} = \{1, x, \cdots, x^{d-1}, (x - \mathcal{K}_1)_+^{d-1}, \cdots, (x - \mathcal{K}_J)_+^{d-1}\} \tag{18}$$

$S(4, \mathcal{K})$: : we have a class of cubic spline functions. This is the most commonly used spline class. The truncated poly basis looks like :

The most commonly used spline basis is the B-spline basis, which has nice numerical properties. To generate a B-spline basis, augment the knot set $\{\mathcal{K}_1, \cdots, \mathcal{K}_J\}$ to

$$\mathcal{K}_{-(d-1)} = \cdots = \mathcal{K}_0 = 0 < \mathcal{K}_1 < \cdots < \mathcal{K}_J < 1 = \mathcal{K}_{J+1} = \cdots = \mathcal{K}_{J+d}$$

(dJd, d knot at each boundary point). We have in total $(J + 2d)$ augmented knots. The B-spline basis functions are defined in the following recursive manner:

$$\{B_{i.j}(x), x \in [0,1] : -(d-1) \leq i \leq J + d, j \in 1, \cdots, d\} \tag{19}$$

$$B_{i,1}(x) = \begin{cases} 1 & \mathcal{K}_i \leq x \leq \mathcal{K}_{i+1} \\ 0 & otherwise \end{cases} \tag{20}$$

$$B_{i,j+1}(x) = \frac{x - \mathcal{K}_i}{\mathcal{K}_{i+j} - \mathcal{K}_i} B_{ij}(x) + \frac{\mathcal{K}_{i+j+1} - x}{\mathcal{K}_{i+j+1} - \mathcal{K}_{i+1}} B_{i+1,j}(x) \tag{21}$$

where $i = -(d-1), \cdots, d + J - 1$, $j = 1, \cdots, d-1$. If we define $0/0 = 0$ here, then $\{B_{i,d}\}_{i=-(d-1)}^{J}$ isthe B-spline basis of $S(d, \mathcal{K})$.

B-spline basis function properties

- $B_{i,d}(x) \geq 0, x \in [0,1]$, and $\sum_{i=-(d-1)}^{J} B_{i,d}(x) = 1, x \in [0,1]$
- B-spline basis functions have compact support: $B_{i,d}(x) > 0$ only on $[\mathcal{K}_i, \mathcal{K}_{i+d}]$
- Vice versa, on $[\mathcal{K}_i, \mathcal{K}_{i+1}]$ at most d basis functions are non-zero. leading to a banded design matrix, to handle inverse difficulties!

Tuning parameter: # of interior knots J. J increases $\implies$ bias decreases, variance increases.

4.4 Smoothing Spline

Smoothing spline avoids the issue of selecting the number of interior knots by placing a knot at each unique data location $x = (x_1, \cdots, x_n)$. To avoid overfitting, one imposes that the estimated function should be smooth by placing a roughness penalty. Given a regression function f, the roughness penalty is

$$PEN_m(f) = \int_0^1 [f^{(m)}(x)]^2 dx \tag{22}$$

it is zero only when f is polynomials up to degree $m-1$. The most common choice of m is $m = 2$, which would "shrink" the regression function towards a linear function.

Assuming the design points $x = \{x_i\}_{i=1}^n$ are all distinct and in increasing order, the smoothing spline estimate solves

$$\arg \min_{f \in S(d,x)} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_0^1 [f^{(m)}(x)]^2 dx \tag{*}$$

with a tuning parameter $\lambda > 0$. It is a finite dimensional problem.

If we choose B-spline basis $B(x) = [B_{-(d-1)}(x) \cdots B_n(x)]'$, then $f \in S(d, X)$ can be written as $f(x) = \gamma' B(x), x \in [0,1], \gamma \in \mathbb{R}^{d+n}$. (*) is equivalent to

$$\arg \min_{\gamma \in \mathbb{R}^{d+n}} \sum_{i=1}^{n} (Y_i - B'(x_i)\gamma)^2 + \lambda \int_0^1 \gamma' B^{(m)}(x) B^{(m)\prime}(x) \gamma dx \tag{**}$$

In the matrix form, let $\Phi_{n \times (d+n)} = [B(x_1) \cdots B(x_n)]'$ and $R_{(d+n)\times(d+n)} = \int_0^1 B^{(m)}(x) B^{(m)\prime}(x) dx$ where $[R]_{ij} = \int_0^1 B^{(m)}_{-d+i}(x) B^{(m)\prime}_{-d+j}(x) dx$.

We solve

$$\arg \min_{\gamma \in \mathbb{R}^{d+n}} \|Y - \Phi\gamma\|^2 + \lambda \gamma' R \gamma$$

obtaining

$$\hat{\gamma} = (\Phi'\Phi + \lambda R)^{-1}\Phi' Y =: S_\lambda Y \tag{Λ}$$

The fitted values are

$$\hat{Y} = \Phi S_\lambda Y$$

So the smoothing spline is a linear smoother. Tuning parameter λ increases, then bias increases, variance decreases.

4.5 Penalized Splines

A hybrid of Regression spline and Smoothing spline. To improve speed, penalized splines use a reduced set of J knots, $J << n$, and a roughness penalty. Denote the knot set as $\mathcal{K} = \{\mathcal{K}_j\}_{j=1}^J$. A version of the penalized spline solves

$$\arg \min_{f \in S(d,\mathcal{K})} \sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \int_0^1 [f^{(m)}(x)]^2 dx \tag{**'}$$

The solution has the same form as (Λ) expect that the dimensions of Φ and R are reduced from $n+d$ to $J+d$. Two tuning pairs, λ decreases, J increasing $\implies$ bias decreases, variance increases.

4.6 Selecting Tuning Parameters

A common goal of tuning parameter selection in nonparametric smoothing is to minimize the expected (conditional) prediction error.

$$Err_\lambda = E((Y - \hat{f}_\lambda(X))^2 | D)$$

where $\hat{f}_\lambda$ is the fitted model with data $D := \{(X_i, Y_i)\}_{i=1}^n$, and (X, Y) is a pair of new observations independent of D. This is a taget we want to minimize over λ. Here $\hat{f}_\lambda$ is indexed by the tuning parameter λ, which can vary if we use different methods.

To estimate $Err(\lambda)$, the in-dample training error

$$\overline{ERR}(\lambda) = \frac{1}{n} RSS(\lambda) := \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}_\lambda(X_i))^2$$

will under estimate $Err(\lambda)$ bevause $\hat{f}_\lambda$ is adapted to the training set. More complex model would lead to a smaller training error. we dont want to end up with the most complex model every time!

Algorithm: Cross-validation (CV) estimates $Err(\lambda)$ by partitioning the data.

1. Randomly divide the data into k folds.
2. Each time, hold out one fold, train using $(k-1)$ folds. The hold-out set is used to evaluate the prediction error.

In formula, let $\mathcal{K} : \{1, \cdots, n\} \to \{1, \cdots, k\}$ be a function that assigns a hold-out set to each observation, which is generated by randomization. The CV estimate of $Err(\lambda)$ is

$$CV(\hat{f}_\lambda) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}_\lambda^{-\mathcal{K}(i)}(X_i))^2$$

where $\hat{f}_\lambda^{-\mathcal{K}(i)}$ denotes the smoother fitted with all but the $\mathcal{K}(i)$-th fold. For each model $\hat{f}_\lambda$, to evaluate $CV(\hat{f}_\lambda)$, a total of k evaluations are needed. k is often 5 or 10.

Leave-one-out CV: If we choose $k = n$, then we have the leave one our cross validation (LOOCV). LOOCV is very computationally heavy. Fortunately for many linear smoothers, e.g. local polynomials and smoothing splines, the LOOCV can be evaluated efficienly. For a linear smoother, we have $\hat{Y} = H_\lambda Y$. Then

$$LOOCV(\lambda) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}_\lambda^{-i}(X_i))^2 = \frac{1}{n} \sum_{i=1}^n (\frac{(Y_i - \hat{f}_\lambda(X_i))^2}{1 - h_{ii}})^2$$

where h_{ij} denotes the (i,j)-th element in H_λ.

Idea: Omit the subscript λ. Consider linear smoothers s.t. $H1 = 1$ where $1_{n\times 1} = [1 \cdots 1]'$. WLOG consider the prediction of the $1H????$ observation. Since

$$\hat{Y}_1 = \sum_{i=1}^{n} h_{1i} Y_i$$

is a weighted average of Y_i, and leave-one-out prediction

$$\hat{Y}_1^{-1} = \sum_{i=1}^{n} h_{1i,-1} Y_i$$

We would like the weights to satisfy

$$h_{1i,-1} = \frac{h_{1i}}{1 - h_{11}} \quad i = 2, \cdots, n \tag{*}$$

Nomalizing precedure.
If the (*) holds, then

$$\begin{aligned} \hat{Y}_1^{-1} - Y_1 &= \sum_{i=2}^{n} \frac{h_{1i} Y_i}{1 - h_{11}} - \frac{1 - h_{1i} Y_i}{1 - h_{11}} \\ &= \sum_{i=1}^{n} \frac{h_{1i} Y_i}{1 - h_{11}} - \frac{Y_i}{1 - h_{11}} \\ &= \frac{\hat{Y}_1}{1 - h_{11}} - \frac{Y_1}{1 - h_{11}} = \frac{\hat{Y}_1 - Y_1}{1 - h_{11}} \end{aligned} \tag{**}$$

So $LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} (\hat{Y}_1^{-1} - Y_1)^2 \overset{(**)}{=} \frac{1}{n}\sum_{i=1}^{n} (\frac{\hat{Y}_i^{-1} - Y_i)^2}{1-h_{ii}})^2$

To select the optimal $\hat{\lambda}$, one can apply grid search and obtain $LOOCV(\lambda)$ for each candidate λ. For tuning the bandwidth or the roughness penalty, one often search on an equally spaced grid on the log scale, e.g. $\log(h) \in \{-1, -2.5, \cdots, 0\}$.

Algorithm: Generalized Cross-validation (GCV) criterion function approximates $LOOCV(\lambda)$ by

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \hat{f}_\lambda(X_i))^2}{1 - \frac{1}{n} tr(H_\lambda))}$$

where $tr(A) := \sum_{i=1}^{n} A_{ii}$. Here $tr(H_\lambda)$ is interpreted as the degree of freedom of the smoother, and is denoted as $df(\lambda)$. We also have

$$GCV(\lambda) = \frac{n \sum_{i=1}^{n} (Y_i - \hat{f}_\lambda(X_i))^2}{(n - df(\lambda))^2} = \frac{n \cdot RSS(\lambda)}{(n - df(\lambda))^2}$$

GCV could be faster than the CV criterion, since $tr(H_\lambda)$ is easier to keep track of. GCV may alleviate the issue of undersmoothing, which CV is prone to. In practice, GCV may still oversmooth. A solution is by applying an adjustment, i.e. instead of using $\lambda = \lambda^*$ whereλ^* by GCV. Use $\lambda = 1.4\lambda^*$ (Empirical result, Hu Chong). k-fold CV in general has larger variance than GCV.

Local polynomial has better boundary control. For higher-dimensional data, kernel smoothing may be more accurate due to data sparsity.

4.7 Multivaritate Smoothing

We now consider a d-dimensional multivariate predictor $X = [X_{1d}, \ldots, X_{id}]'$ and a univariate response Y_i:

$$Y_i = \mu(X_{1d}, \ldots, X_{id}) + \varepsilon_i, i = 1, \ldots, n$$

where $E(\varepsilon_i|X_i) = 0, Var(\varepsilon_i|X_i) = \sigma^2(X_i)$ for some variance function $\sigma^2 : \mathbb{R}^d \to \mathbb{R}$.

If μ_i is smooth at X_0, the first order Taylor expansion of μ holds at $X_0 = [X_{01}, \ldots, X_{0d}]'$ and

$$\begin{aligned}\mu(X) &\approx \mu(X_0) + \frac{\partial}{\partial X}\mu(X_0)'(X - X_0) \\ &= \mu(X_0) + \sum_{j=1}^{d} \frac{\partial}{\partial x_j}\mu(X_0)(x_j - x_{0j})\end{aligned}$$

for X lying in a small neighborhood of X_0. This justifies to use a local linear approximation when estimating $\mu(X_0)$. The multivariate local linear estimator of $\mu(X_0)$ is $\hat{\mu}(X_0) = \hat{\alpha}_0$, where

$$\hat{\alpha} = [\hat{\alpha}_0, \hat{\alpha}_0, \ldots, \hat{\alpha}_d] = \arg\min_{\alpha \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \{Y_i - \alpha_0 - \sum_{j=1}^{d} \alpha_j (X_{ij} - X_{0j})\}^2 K_h(X_i - X_0)$$

where $h = [h_1, \ldots, d_d]'$ is a bandwidth vector $K_h* = \Pi_{j=1}^{d} K(\frac{x_j}{h_j})$ and $K(\cdot)$ is a univariate kernel. Let

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & X_{11} - X_{01} & \cdots & X_{1d} - X_{0d} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} - X_{01} & \cdots & X_{1d} - X_{0d} \end{pmatrix}$$

$K_h = diag(K_h(X_1 - x_0) \cdots K_h(X_n - x_0)$ Then the objective function is

$$(Y - X_0\alpha)' K_h (Y - X_0\alpha)$$

The minimizer is given by

$$\hat{\alpha} = (X_0' K_h X_0)^{-1} X_o' K_h Y$$

So

$$\hat{\mu}(X_0) = \hat{\alpha}_0, \frac{\hat{\partial \mu}}{\partial x_j}(X_0) = \hat{\alpha}_j, j = 1, \ldots, d.$$

Higher order local polynomials are useful when estimating higher order derivatives. To include local polynomials up to order p. The i-th row of the design matrix i, the concaternation of

$$1, [X_{ij} - X_{0j}]_{j=1}^{d}, \ldots, [\Pi_{j=1}^{d} (X_{ij} - X_{0j})^{v_j}]_{\sum_1^d v_j = p, v_j \in \mathbb{N}}$$

$\hat{\mu}(X_0)$ is a linear smoother.
R demonstration:

- The design matrix may have pattern, which exert errors for some estimates. solution: chop off some areas.
- Therefore it is important to check the design matrix. 'locfit' by default choose a large window to prevent warnings. Be aware to check that.
- There may be influencing points if we take log transformation, we can take the started log of the response.

4.7.1 Curse of dimensionality

Suppose the deisgn points lie in $[0,1]^d$. Let $W(X_0)$ be a local cubical window around X_0 with side lengths h, and $I(X_0) = \{i : X_i \in W(X_0)\}$ be the subjuect indices falling within $W(X_0)$. Since $\hat{\mu}(X_0)$ is a linear smoother, we have $\hat{\mu}(X_0) = \sum_{i \in I(X_0)} w_i Y_i$, where $\sum_{i \in I(X_0)} w_i = 1$. Assume the design points X_i are non-random, and they form an equally-spaced grid in $[0,1]^d$,

$$E(\hat{\mu}(X_0)) = \sum_{i \in I(X_0)} w_i \mu(X_i) \approx \sum_{i \in I(X_0)} w_i [\mu(X_0) + (X_i - X_0)' \frac{\partial}{\partial X}(X_0)] \approx \mu(X_0)_{i \in I(X_0)} w_i + \sum_{i \in I(X_0)} w_i \mathcal{O}(h) = \mu(X_0) + \mathcal{O}(h)$$

Actually, the bias term $E(\hat{\mu}(X_0)-\mu(X_0))$ can be improved to $\mathcal{O}(h^2)$ for an interior pt X_0. The order of the bias does not depend on the dimension d. The volumn of the interior region is $(1-2h)^d$, which decreases exponentially as d increases, so in a high dimensional nonparametric regression, almost all points are boundary points, which poses practical issues. However for $Var(\hat{\mu}(X_0))$ we have on average $W(X_0)$ contains $|W(X_0)| = nh^d$ samples. If weights $w_i = 1/nh^d$, then

$$Var(\hat{\mu}(X_0)) \approx \sum_{i \in I(X_0)} w_i^2 = 1/nh^d$$

To have $Var(\hat{\mu}(X_0)) = c$ where c is a constant, $n = 1/ch^d$ which increases exponentially as d increases.

Question: why weights are in this form?

Big-O and Little-O

Let $\{X_n\}$ be a sequence of r.v.s and a_n be a sequence of positive real numbers.

$X_n = \mathcal{O}_p(1)$ if $\forall \varepsilon > 0, \exists n_0, M < \infty \, s.t. \, P(|X_n| > M) < \varepsilon, \forall n \geq n_0$

$X_n = o_p(1)$ if $X_n \xrightarrow{p} 0$

$X_n = \mathcal{O}_p(a_n)$ if $X_n/a_n = \mathcal{O}_p(1)$

$X_n = o_p(a_n)$ if $X_n/a_n = o_p(1)$

To remember: big-O represents the same order, little-O represents lower order.

Theorem: (Ruppert and Wand 1994) Assume $h_1 = \ldots h_d = j$. Under RCs,

$$E[\hat{\mu}(X_0) - \mu(X)|\mathbb{X}] = 1/2\mu_2^d(K)tr(D(X_0))h^2 + o_p(h^2)$$

$$Var[\hat{\mu}(X_0)|\mathbb{X}] = (nh^d)^{-1}v_0(k)^d\sigma^2(X_0)f(X_0)^{-1} + o_p(nh^{d-1})$$

as $h \to 0, nh^d \to \infty$ where $D(X_0)$ is the Hessian of μ at X_0, $\mu_2(K) = \int_0^1 x^2K(x)dx, v_0(K) = \int_0^1 K^2(x)dx$, and $f(\cdot)$ is the pdf of X_1.

$$MSE = bias^2 + var = c_1h^4 + c_2\frac{1}{nh^d} + \text{smaller terms}$$

Since the slower rate dominates the MSE, set $h^4 = \frac{1}{nh^d}$, we have $h \asymp n^{-d+4}$. Plug in h^* the best rate for MSE is $n^{-\frac{2}{d+4}}$. $E(Y|X) = \mu(x_1, \ldots, x_d)$. Additional models: $\mu(x_1, \ldots, x_d) = f_1(x_1) + \cdots + f_d(x_d)$

5 Asymptotic theory

(Fan& Fijbel 1996) Here we consider a univariate presictor $d = 1$. For local polynomial with degree p, we need the following conditions:

- $(\mathcal{K})$ The kernel $\mathcal{K}(\cdot)$ is a symmetric continuous density function supported on $[-1, 1]$, such that $\mu_j := \int_{-1}^1 x^J\mathcal{K}(x)dx, j = 0, \ldots, 2p+2$ and $v_j := \int_{-1}^1 \mathcal{K}^2(x)x^jdx, j = 0, \ldots, 2p$ exist and are finite.
- (H) the bandwidth $h \to 0$ and $nh \to \infty$
- (D) $X_i \overset{iid}{\sim} f(x)$ where f is a continuous density supprted on a compact inverval $[a, b]$, and $f(x) > 0, \forall x$.
- (N) $var(\varepsilon_i|X_i) = \sigma^2(X_i) < \infty$ where $\sigma^2(\cdot)$ is a continuous function
- (S) $\mu^{(p+1)}$ exists and is bounded.

Use $\mathbb{X} = \{X_i\}_{i=1}^n$ to denote all the design points.

$(p = 1)$: The local linear smoothr $\hat{\mu}(X_0)$ for estimating $\mu(X_0)$ is expcted as follows:

$$S_{jn} = \sum_{i=1}^n \mathcal{K}_h(X_i - X_0)(X_i - X_0)^j, j = 0, 1, 2$$

$$R_{jn} = \sum_{i=1}^n \mathcal{K}_h(X_i - X_0)(X_i - X_0)^jY_i, j = 0, 1$$

$$\implies X'\mathcal{K}_h X = \begin{pmatrix} S_{0n} & S_{1n} \\ S_{1n} & S_{2n} \end{pmatrix}$$

$$X'\mathcal{K}_h Y = \begin{pmatrix} R_{0n} \\ R_{1n} \end{pmatrix}$$

$$\hat{\mu}(X_0) = (1 \quad 0) \begin{pmatrix} S_{0n} & S_{1n} \\ S_{1n} & S_{2n} \end{pmatrix}^{-1} \begin{pmatrix} R_{0n} \\ R_{1n} \end{pmatrix} = \frac{S_{2n}R_{0n} - S_{1n}R_{1n}}{S_{0n}S_{2n} - S_{1n}^2}.$$

Theorem: For the local linear smoother w/p $p = 1$ for $\mu(X_0)$ where $X_0 \in (a, b)$. Under conditions (K, H, D, N, S), as $h \to 0, nh \to \infty$, we have the asymptotic conditional variance is

$$Var(\hat{\mu}(X_0)|\mathbb{X}) = \frac{\sigma^2(X_0)v_0}{f(X_0)} \cdot \frac{1}{nh} + o_p(\frac{1}{nh})$$

And the A. conditional bias is

$$bias(\hat{\mu}(X_0)|\mathbb{X}) = E(\hat{\mu}(X_0)|\mathbb{X}) - \mu(X_0) = \mu''(X_0)\mu_2 h^2 + o_p(h^2)$$

In summary the conditional MSE is (*) $(\frac{\sigma^2(X_0)v_0}{f(X_0)} \cdot \frac{1}{nh} + (\mu''(X_0))^2\mu_4 h^4)(1 + o_p(1))$

Since we only have discrete observations, we can only derive conditional exprectation and variance given observed points.

Proof. Note $E(Y_i|\mathbb{X}) = \mu(X_i), var(Y_i|\mathbb{X}) = \sigma^2(X_i)$. The conditional variance is

$$\frac{S_{2n}^2 T_{0n} + S_{1n}^2 T_{2n} - 2S_{2n}S_{1n}T_{1n}}{(S_{2n}S_{0n} - S_{1n})^2}$$

where $T_{jn} = \sum_{i=1}^n K_h(X_i - x_0)^2(X_i - x_0)^j\sigma^2(X_i), j = 0, 1, 2.$

We obtain the rates in (*) term by term using the following lemma.

Lemma:

$$\left|\frac{S_{jn}}{nh^j} - f(x_0)\mu_j\right| = o_p(1),$$

$$\left|\frac{T_{jn}}{nh^{j-1}} - f(x_0)\sigma^2(x_0)v_j\right| = o_p(1), j = 0, 1, 2.$$

Proof.

$$\begin{aligned}
E(S_{jn}) &= nE(\mathcal{K}_h(X_1 - x_0)(X_1 - x_0)^j) \\
&= n\int_a^b \frac{1}{h}\mathcal{K}((X_1 - x_0)/h)(X_1 - x_0)^j f(x)dx \\
&= n\int_{x_0-h}^{x_0+h} \frac{1}{h}\mathcal{K}((X_1 - x_0)/h)(X_1 - x_0)^j f(x)dx \\
&= n\int_{-1}^{1} \mathcal{K}(t)t^j h^j f(x_0 + th)dt \text{change of variable: } t = \frac{X - x_0}{h} \\
&= nh^j\int_{-1}^{1} \mathcal{K}(t)t^j(f(x_0) + f(x_0 + th) - f(x_0))dt \\
&= nh^j(\int_{-1}^{1} \mathcal{K}(t)t^j f(x_0)dt + \int_{-1}^{1} \mathcal{K}_h(t)t^j(f(x0 + th) - f(x_0))dt) \\
&= nh^j(\mu_j f(x_0) + o(1))(DCT)
\end{aligned}$$

$$
\begin{aligned}
Var(S_{jn}) &= nVar(\mathcal{K}_h(X_1-x_0)(X_1-x_0)^j) \\
&\le nE(\mathcal{K}_h(X_1-x_0)^1(X_1-x_0)^{2j}) \\
&= n\int_{x_0-h}^{x_0+h}\frac{1}{h^2}\mathcal{K}((X_1-x_0)/h)^2(X_1-x_0)^{2j}f(x)dx \\
&= n\int_{-1}^{1}\mathcal{K}^2(t)t^{2j}h^{2j-1}f(x_0+th)dt \\
&= nh^{2j-1}\int_{-1}^{1}\mathcal{K}_h(t)t^{2j}(f(x_0)+f(x_0+th)-f(x_0))dt \\
&= nh^{2j-1}(\int_{-1}^{1}\mathcal{K}^2(t)t^{2j}f(x_0)dt+\int_{-1}^{1}\mathcal{K}_h(t)t^{j}(f(x_0+th)-f(x_0))dt) \\
&= nh^{2j-1}(v_{2j}f(x_0)+o(1))(DCT)
\end{aligned}
$$

$SD(\frac{S_{jn}}{nh^j} = O\left(\sqrt{\frac{1}{nh}}\right)$
By the Chebyshev's Inequality, we have

$$|\frac{S_{jn}}{nh^j}-E(\frac{S_{jn}}{nh^j})| = |\frac{S_{jn}}{nh^j}-f(X_0)\mu_j| = O_p(\sqrt{1/nh}) = o_p(1), nh\to\infty$$

Similarly $|\frac{T_{jn}}{nh^j}-f(X_0)\sigma^2(X_0)v_j| = o_p(1)$ □
Now $S_{jn} = nh^j f(x_0)\mu_j(1+o_p(1)), T_{jn} = nh^{j-1}f(x_0)v_j\sigma^2(x_0)(1+o_p(1))$, plug these into (*) we have

$$* = \frac{\sigma^2(X_0)}{f(X_0)}\frac{\mu_2^2v_1+\mu_1v_2-2\mu_1\mu_2v_1}{(\mu_2\mu_0-\mu_1^2)^2}\cdot\frac{1}{nh}+\text{smaller order terms}$$

where $\mu_1 = \int_{-1}^{1}x\mathcal{K}(x)dx = 0$ by symmetry of $\mathcal{K}$. $\mu_0 = \int_{-1}^{1}\mathcal{K}(x)dx = 1$. So $* = \frac{\sigma^2(X_0)v_0}{f(x_0)}\frac{1}{nh}+O_p(\frac{1}{nh})$.
For conditional bias:

$$
\begin{aligned}
E(\hat{\mu}(X_0)-\mu(X_0)|\mathbb{X}) &= \frac{S_{2n}(E(R_{0n|\mathbb{X}})-\mu_2(X_0)S_{0n})-S_{1n}(E(R_{1n|\mathbb{X}})-\mu_2(X_0)S_{1n})}{S_{0n}S_{2n}-S_{1n}^2} \\
E(R_{jn}|\mathbb{X}) &= \sum_{j=1}^{n}K_h(X_i-x)(X_i-x)^j\mu(X_i)
\end{aligned}
$$

By the Taylor's theorem

$$\mu(X) = \mu(X_0)+(X-X_0)\mu'(X_0)+(X-X_0)^2\mu''(\xi), \xi \text{ between } X_1 \text{ and } X$$

By applying a similar derivation to the previous lemma (**)

$$** = f(X_0)\mu_{j+2}\mu''(X_0)nh^j+o_p(nh^j)$$

By lemma,

$$E(\hat{\mu}(X_0)-\mu(X_0)|\mathbb{X}) = \frac{(f^2(X_0)\mu_2^2\mu''(X_0)h^4+f^2(X_0)\mu_1\mu_2\mu''(X_0)h^4)n^2}{f(X_0)^2(\mu_2\mu_0-\mu_1^2)n^2h^2}+\text{smaller order terms} = \mu''(X_0)h^2\mu_2$$

Since $\mu_3 = \int_{-1}^{1}x^3\mathcal{K}(x)dx = 0, \mu_1 = 0$.
Note for the theorem, LP gives best fit if the original function is up to second order differentiable.
The h^4 and $\frac{1}{nh}$ corresponds to the bias-variance tradeoff. The optimal rate for the conditional MSE is obrained when $h^4 = \frac{1}{nh} \iff h = n^{-1/5}$. The corresponding MSE is of order $n^{-4/5}$. This implies by Chebyshev's inequality.

$$\hat{\mu}(X_0)-\mu(X_0) = O_p(n^{-2/5})$$

The rate $n^{-2/5}$ is often called the 1 dimensional nonparametric rate. This is the results of the smoothness assumption that μ has $(p+1) = 2$ nd order derivative, which is bounded.

Correlation does not affect conditional mean, since mean is linear operator, but it will affect the conditional variance.

Theorem: (Uniform convergence, Mack and Silverman 1982) Under the condition of the last theorem, if h has the same order as $(\log(n)/n)^{1/5}$, the local linear estimator $\hat{\mu}(X_0)$ satisfies

$$\sup_{x\in[a,b]} |\hat{\mu}(X) - \mu(X)| = O(\frac{n}{\log(n)})^{-2/5} a.s.$$

Conditional bias & variance for a general local polynomial

Consider

$$S = \left[\mu_{j+j'}\right]_{j,j'=0}^{p} = \begin{bmatrix} \mu_0 & \mu_1 & \cdot\mu_p \\ \mu_1 & \mu_2 & \cdot\mu_{p+1} \\ \vdots & \vdots & \vdots \\ \mu_p & \mu_{p+1} & \cdot\mu_{2p} \end{bmatrix}$$

$$S^* = [v_{j+j'}]_{j,j'=0}^{p}$$

$$C_p = \begin{bmatrix} \mu_{p+1} \\ \vdots \\ \mu_{2p+1} \end{bmatrix} \qquad \tilde{C}_p = \begin{bmatrix} \mu_{p+2} \\ \vdots \\ \mu_{2p+2} \end{bmatrix}$$

Theorem: (LP interior point, general p and k, Fan and Giljbel,1996) Under (K,H,D,N,S), $h \to 0, nh \to \infty$, the asymptotic conditional variance of $\hat{\mu}^{(k)}(X_0), k \le p, X_0 \in (0,1)$ is

$$var(\hat{\mu}^{(k)}(X_0)|\mathbb{X}) = e_{k+1}^T S^{-1} S^* S^{-1} e_{k+1} \frac{v!^2 \sigma^2(X_0)}{f(X_0)} \cdot \frac{1}{nh^{1=2k}} + O_p(\frac{1}{nh^{1+2k}})$$

The asymptotic conditional bias for $p = k$ odd is given by

$$bias(\hat{\mu}^{(k)}(X_0)|\mathbb{X}) = e_{k+1}^T S^{-1} C_p \frac{k!}{(p+1)!} \mu^{(p+1)}(X_0) h^{p+1-k} + O_p(h^{p+1-k})$$

Further for $p - v$ even the asymptotic conditional bias is

$$bias(\hat{\mu}^{(k)}(X_0)|\mathbb{X}) = e_{k+1}^T S^{-1} \tilde{C}_p \frac{k!}{(p+2)!} \mu^{(p+2)}(X_0) h^{p+2-k} + O_p(h^{p+2-k})$$

6 Functional Data

6.1 Spase FPCA

Principal component analysis through conditional expectation (PACE) by Yao Muller Wang (2005) https://projecteuclid.org/euclid.aos/1140191677.

Let $X_1(t), \ldots, X_n(t), t \in \mathcal{T}$ be a sample of iid unobserved 2nd order stochastic process. Available are discrete observations $\{T_{ij}, Y_{ij}\}_{i=1,j=1}^{n,m_i}$

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}$$

Here n is the number of subjects, and m_i is the number of obs per curve. The time domain $\mathcal{T}$ is a compact set in the Euclidean space. For concreteness, assume $\mathcal{T} = [0,1]$. We assume $T_{ij} \overset{iid}{\sim} T$ a random variable supported on $\mathcal{T}$. and that the T_{ij} are independent of the X_i so the observation time is uninformative.

The measurement errors $\{\epsilon_{ij}\}_{i,j}$ are iid, independent of the X_i and T_{ij} and satisfies $E(\varepsilon_{ij}) = 0, var(\varepsilon_{ij}) = \sigma^2$. In FPCA, we want to target the KL expantion

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t), t \in [0,1]$$

$i = 1, \ldots, n$. Recall $\mu(t) = E(X(t)), F(t,s) = Cov(X(t), X(s))$. Let $\mathcal{G} : L^2 \to L^2, \mathcal{G}(f)(\cdot) = \int_0^1 G(\cdot, s) f(s) ds$ be the covariance operator associated with G. Write $\mathcal{G} = \sum_{k=1}^{\infty} \lambda_k \phi_k \otimes \phi_k$ be the eigendecomposition. Here $\lambda_1 \geq \lambda_2 \geq \cdots 0$ are the eigen values, and $\phi_1, \phi_2, \ldots$ are orthonormal eigen functions. Here $\xi_{ik} = \int_0^1 (X_i(t) - \mu(t)) \xi_k(t) dt$ is the k-th FPC score for subject i.

6.1.1 Mean Estimate

Note the cross-sectional estimate does not work for $\mu(t)$. However,

$$\begin{aligned} E(Y_{ij}|T_{ij}) &= E(X_i(T_{ij}) + \varepsilon_{ij}|T_{ij}) \\ &= E(X_i(T_{ij})|T_{ij}) \quad (\varepsilon_{ij} \perp T_{ij}) \\ &= \mu(T_{ij}) \quad (X_i \perp T_{ij}) \\ \Longrightarrow & \\ Y_{ij} &= \mu(T_{ij}) + u_{ij} \end{aligned}$$

where $u_{ij} = X_i(T_{ij}) - \mu(T_{ij}) + \varepsilon_{ij}$ with $E(u_{ij}|T_{ij}) = 0, var(u_{ij}|T_{ij}) = G(T_{ij}, T_{ij}) \to \sigma^2$. This motivates to esimate $\mu(t)$ by applying a LL amoother on the pooled scatterplot $\{(T_{ij}, Y_{ij})\}$ for $t \in [0,1]$, set $\hat{\mu}(t) = \hat{\alpha}_0$ where

$$(\hat{\alpha}_0, \hat{\alpha}_1) = \arg \min_{\alpha_0, \alpha_1 \in \mathbb{R}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} [Y_{ij} - \alpha_0 - \alpha_1 (T_{ij} - t)]^2 K_{h_\mu}(T_{ij} - t)$$

for some bandwidth $h_\mu > 0$, and residual kernel $K_h(s) = 1/h \cdot K(x/h)$, K is a kernel.

6.1.2 Covariance surface estimate

The idea is to construct "raw covariances" which has conditional mean equal to the $G(\cdot_1, \cdot_2)$. Denote $G_{ijj'} = (Y_{ij} - \mu(T_{ij})) \cdot (Y_{ij'} - \mu(T_{ij'}))$ as the raw covariance, $i = 1, \ldots, n, j, j' = 1, \ldots, m_i$. We have

$$\begin{aligned} & E(G_{ijj'}|T_{ij}, T_{ij'}) \\ =& E[Y_{ij} - \mu(T_{ij}) \cdot (Y_{ij'} - \mu(T_{ij'})|T_{ij}, T_{ij'}] \\ =& E[(X_i(T_{ij}) - \mu(T_{ij}) \cdot (X_i(T_{ij'}) - \mu(T_{ij'}) + (X_i(T_{ij}) - \mu(T_{ij}) \varepsilon_{ij'} + (X_i(T_{ij'}) - \mu(T_{ij'}) \varepsilon_{ij} + \varepsilon_{ij} \varepsilon_{ij'}|T_{ij}, T_{ij'}] \\ =& \begin{cases} G(T_{ij}, T_{ij'}) & j \neq j' \\ G(T_{ij}, T_{ij'}) + \sigma^2 & j = j' \end{cases} \end{aligned}$$

In order to estimate $G(s,t)$, apply a 2-dimensional LL smoother to the "off-diagonal" raw covariances

$$\{(T_{ij}, T_{ij'}, G_{ijj'})|i = 1, \ldots, n, j \neq j', j, j' = 1, \ldots, m + i\}$$

Set $\hat{G}(t,s) = \hat{\beta}_0$,

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \arg \min_{\beta_0, \beta_1, \beta_2 \in \mathbb{R}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{j \neq j', j'=1}^{m_i} [G_{ijj'} - \beta_0 - \beta_1 (T_{ij} - t) - \beta_2 (T_{ij'} - s)]^2 K_{h_G}(T_{ij} - t) K_{h_G}(T_{ij} - s)$$

where $h_G > 0$ is the bandwidth for smoothing $G(t,s)$. The eigenvalues and eigenfunctions are estimated by those of $\hat{G}(t,s)$ demoted as $(\hat{\lambda}_k, \hat{\phi}_k), k = 1, 2, \cdots$

Since the covariance matrix is symmetric, we use the same bandwidth for t and s.

Note in practice, need to work with $\widehat{G}_{ijj'} = [Y_{ij} - \hat{\mu}(T_{ij})][Y_{ij'} - \hat{\mu}(T_{ij'})]$ instead of $G_{ijj'}$

Caution: check the design plot, i.e. the scatter plot of $\{(T_{ij}, T_{ij'})\}_{i, j \neq j'}$. Some longitutional studies follow patients for a shorter period of time than the age window of interest, generating "snippets".

6.1.3 Variance estimation σ^2

In view of (*), the diagonal raw covariances $\{T_{ij}, G_{ijj'}\}$ are used to estimate $\sigma^2 = var(\varepsilon_{ij})$ since $E(G_{ijj'}|T_{ij}, T_{ij'}) = G(T_{ij}, T_{ij'}) + \sigma^2$. Write $V(t) = G(t,t) + \sigma^2, t \in [0,1]$. So $\sigma^2 = \int_0^1 [V(t) - G(t,t)]dt$, set $\hat{V}(t) = \hat{\gamma}_0$ where

$$\hat{\gamma}_0, \hat{\gamma}_1 = \arg\min_{\gamma_0,\gamma_1 \in \mathbb{R}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} [Y_{ij} - \alpha_0 - \alpha_1(T_{ij} - t)]^2 K_{h_{\sigma^2}}(T_{ij} - t)$$

where $h_{\sigma^2} > 0$ is the bandwidth, then set

$$\hat{\sigma}^2 = \int_0^1 \hat{V}(t) - \hat{G}(t,t)dt$$

6.1.4 ξ_{ik} estimation

Due to the sparse obs, any Riemann sum approximation to $\int_0^1 (X_i(t) - \mu(t))\phi_k(t)dt$ fails. Instead consider the best linear unbiased prediction (BLUP) of ξ_{ik} given the available observations $Y_i = [Y_{i1}, \ldots, Y_{im_i}]^T$.

Theorem: For a pair of random variables $(Z, W) \in \mathbb{R}^{p+q}$, the BLUP of Z given W is

$$B(Z|W) = Cov(Z,W)Cov^{-1}(W)(W - E(W)) + E(Z)$$

where $Cov(W) = Cov(W,W)$. Here $B(Z|W)$ is the linear predicter $A^*W + b^*$ that minimizes

$$E\|Z - AW - b\|^2$$

For our functional data, let the follwing $E(\cdot)$ and $Var(\cdot)$ to be conditional on the T_{ij}. Let $\xi_{ik} = [\xi_{i1}, \ldots, \xi_{ik}]^T, k = 1, 2, \ldots$.

$$\tilde{\xi}_{ik} := B(\xi_{ik}|Y_i) = Cov(\xi_{ik}, Y_i)Cov(Y_i)^{-1}(Y_i - \mu_i)$$

where

$$\begin{aligned}
[\mu_i]_j &= \mu(Y_{ij}) \\
[cov(Y_i)]_{jj'} &= cov(Y_{ij}, Y_{ij'}) = G(T_{ij}, T_{ij'}) + \sigma^2 1[j = j'], j, j' = 1, \ldots, m_i \\
[cov(\xi_{ik}, Y_i)]_{kj} &= cov(\xi_{ik}, X_i(T_{ij}) + \varepsilon_{ij}) = cov(\xi_{ik}, X_i(T_{ij})) = \lambda_k \phi_k(T_{ij}), \text{ by the last theorem in intro chap 4.1}
\end{aligned}$$

Question: where is the expectation of ξ_{ik}????

Denote $\Sigma_{Y_i} = cov(Y_i), \phi_{ik} = [\phi_k(T_{i1}) \cdots \phi_k(T_{im_i})]^T, \Phi_i = [\phi_i 1 \cdots \phi_{ik}]^T, \Omega_k = diag(\lambda_1, \ldots, \lambda_k)$. Now,

$$\tilde{\xi}_{ik} = [\tilde{\xi}_{i1} \cdots \tilde{\xi}_{ik}]^T = \Omega_k \Phi_i \Sigma_{Y_i}^{-1}(Y_i - \mu_i)$$

which is estimated by the plug-in estimate

$$\hat{\xi}_{ik} = [\hat{\xi}_{i1} \cdots \hat{\xi}_{ik}]^T = \hat{\Omega}_k \hat{\Phi}_i \hat{\Sigma}_{Y_i}^{-1}(Y_i - \hat{\mu}_i)$$

Note $\hat{\xi}_{ik}$ is consistent for $\tilde{\xi}_{ik}$ but not ξ_{ik} as $n \to \infty$.

To estimate the original trajectories for $k = 1, 2, \ldots$, approach

$$X_{ik}(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t), t \in [0,1]$$

through

$$\tilde{X}_{ik}(t) = \mu(t) + \sum_{k=1}^{K} \tilde{\xi}_{ik}\phi_k(t), t \in [0,1]$$

The plugin estimate is

$$\hat{X}_{ik}(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_k(t), t \in [0,1]$$

Note if you only have very few observations for each sample, the best you can get is still bad even whe you got many samples. Therefore we are targeting BLUP, not the true principal component.

Determine whether the truncated version well approximates the continuous one, look at the first eigen values and check the proportion of variance explained.

Smaller sample size, we always choose a larger bandwidth.

Example: Gaussian Process

From here on, assume the $X_i(t)$ are Gaussian processes, and the ε_{ij} are also Gaussian. Then (Y_i, ξ_{ik}) are jointly Gaussian. (c.f. The definition of stochastic integral in Intro Chap 4.2 and Ash& Gardner (1972) Thm 1.4.2) Question: cannot find this book.

By Gaussiality

$$E(\xi_{ik}|Y_i) = B(\xi_{ik}|Y_i)$$

and $\tilde{\xi}_{ik}$ is the best prediction of ξ_{ik} given Y_i. Also

$$\Omega_k = cov(\xi_{ik}|Y_i) = E(cov(\xi_{ik}|Y_i)) = cov(\xi_{ik}) - cov(E(\xi_{ik}|Y_i)) = \Omega_k - \Omega_k \Phi \Sigma_{Y_i}^{-1} \Phi_i^T \Omega_k$$

therefore

$$\xi_{ik}|Y_i \sim N(\tilde{\xi}_i k, \Omega_k) \quad (**)$$

Then $X_{ik}(t) - \tilde{X}_{ik}(t) = \sum_{k=1}^{K}(\xi_{ik} - \tilde{\xi}_{ik})\phi_k(t) = (\xi_{ik} - \tilde{\xi}_{ik})^T \phi_k(t)$ where $\phi_K(t) = [\phi_1(t) \cdots \phi_k(t)]^T$. By (**),

$$X_{ik}(t) - \tilde{X}_{ik}(t) = \sum_{k=1}^{K}(\xi_{ik} - \tilde{\xi}_{ik})\phi_k(t) = (\xi_{ik} - \tilde{\xi}_{ik})^T \phi_k(t)$$

So a $100(1-\alpha)\%$ pointwise confidence interval for $X_{ik}(t), t \in [0,1]$ is

$$\xi_{ik} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\tilde{\xi}_{ik}^T \phi_k(t)}$$

Also,

$$\begin{aligned}
&P(|(\xi_{ik} - \tilde{\xi}_{ik})^T \phi_k(t)|^2 \le \chi^2_{k,1-\alpha} \phi_k(t)^T \Omega_k \phi_k(t), \forall t | Y_i) \\
\ge_{CS}\ &P(\|\Omega_k^{-1/2}(\xi_{ik} - \tilde{\xi}_{ik})\|^2 \|\Omega_k^{1/2}\phi_k(t)\|^2 \le \chi^2_{k,1-\alpha} \phi_k^T(t) \Omega_k \phi_k(t), t \in [0,1] | Y_i) \\
=\ &1 - \alpha
\end{aligned}$$

Since $\Omega_k^{-1/2}(\xi_{ik} - \tilde{\xi}_{ik})|Y_i \sim N(0, I_k)$. So a $100(1-\alpha)\%$ simultaneous CI for $X_{ik}(t), t \in [0,1]$ is

$$\tilde{X}_{ik}(t) \pm \sqrt{\chi^2_{k,1-\alpha} \phi_k^T(t) \Omega_k \phi_k(t)}$$

Select K

1. FVE
2. CV: use a task specific loss (criterion) function and select K that minimizes the loss. Pseudo AIC (Tao et a (2005)). If the true KL expansion has K terms, i.e. $X_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t)$ then

   $$Y_i|\xi_{ik} \sim N(\mu_i + \phi_i \xi_{ik}, \sigma^2 I_{m_i})$$

   The estimated log-like function is

   $$\hat{L}(K) = \sum_{i=1}^{n} -\frac{m_i}{2}\log(2\pi)\frac{m_i}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\|Y_i - \hat{\mu}_i \hat{\Phi}_i \hat{\xi}_{ik}\|^2$$

   Let $N = \sum_{i=1}^{n} m_i, \hat{\sigma}_K^2 = \frac{1}{N}\sum_{i=1}^{n}\|Y_i - \hat{\mu}_i - \hat{\Phi}_i\hat{\xi}_{ik}\|^2$,

   (a) The Pseudo-AIC criterion is

   $$\widehat{pAIC}_K = -2\hat{L}(K) + 2K = -\log(2\pi)C + N\log(\hat{\sigma}^2) + \frac{\hat{\sigma}_K^2}{\hat{\sigma}^2}N + 2K$$

   (b) The Pseudo-BIC criterion is

   $$\widehat{pBIC}_K = -2\hat{L}(K) + \log(n)K$$

   (c) Li Wang Carrol (2013): AIC for dense data, i.e. $m_i \asymp m \to \infty$

   $$\widehat{AIC(K)} = N\ log(\hat{\sigma}_K^2) + N + 2nK$$

Here $Y_{ij} - \mu_{ij} = \sum_{k=1}^{K} \xi_{ik}\phi_k(T_{ij}) + \varepsilon_{ij}$ is viewed as a regression. Each ξ_{ik} is counted as a parameter.

Bandwidth selection

h_μ, h_G could be selected by CV or GCV, but an upward adjustment may be needed due to the dependency of repeated measurements.

6.1.5 Asymptotic Results

For simplicity assume $m_i = m, i = 1, \ldots, n$. Full general result can be found in Zhang and Wang (2016)'s Annals paper, who derived the results presebted in below. See also Li and Hsing (2010).

We are under the PACE observation models. Other than the requirements for the data model, a set of conditions are required for the LL smoother $\hat{\mu}, \hat{G}$ as well as X, ε.

Theorem: (Theorem 1) Under conditions $(K, \gamma, S_\mu, H_\mu, M_\mu)$ defined below,

$$\sup_{t\in[0,1]} |\hat{\mu}(t) - \mu(t)| = O(h_\mu^2 + \sqrt{\log(n)(\frac{1}{mnh_\mu} + \frac{1}{n})}) a.s.$$

Theorem: (Col 1)

1. (Sparse) If $m \le C < \infty$, then $\sup_{t\in[0,1]} |\hat{\mu}(t) - \mu(t)| = O(h_\mu^2 + \sqrt{\log(n)(\frac{\log(n)}{nh_\mu} + \frac{1}{n})})$ a.s.
2. (Dense) If $m/[n/\log(n)]^{1/4} \to d, 0 < d \le \infty$, $h_\mu = O(n/\log(n)]^{1/4})$, and $h_\mu m$ is bounded away from 0, then $\sup_{t\in[0,1]} |\hat{\mu}(t) - \mu(t)| = O(\sqrt{\frac{\log(n)}{n}}$ a.s.

Note: under a similar set of conditions

$$\|\hat{\mu} - \mu\| = [\int_{\mathcal{T}} (\hat{\mu}(t) - \mu(t))^2 dt]^{1/2} = O_p(h_\mu^2 + \sqrt{\frac{1}{mnh_\mu} + \frac{1}{n}})$$

- $(\mathcal{K})$ The kernel $\mathcal{K}(\cdot)$ is a symmetric density function supported on $[-1, 1]$, and $\mathcal{K}$ is Lipschitz, i.e. $\exists 0 < K < \infty$ such that
$$|\mathcal{K}(u) - \mathcal{K}(v)| \le L|u - v|, \quad u, v \in [-1, 1]$$
- (D) $T_{ij} \overset{iid}{[} 0, 1]$ where f is a continuous density is bounded away from $0, \infty$, and $0 < \min_{t\in\mathcal{T}} f(x) \le \max_{t\in T} f(t0 < \infty$.
- (S_μ) μ'' exists and is bounded on $\mathcal{T}$.
- (S_G) $\frac{\partial^2}{\partial t^2}G(t, s)$, $\frac{\partial^2}{\partial s^2}G(t, s)$, $\frac{\partial^2}{\partial t \partial s}G(t, s)$ exists and is bounded on $\mathcal{T}^2$.
- (H_μ) $h_\mu \to 0, \log(n)\frac{1}{mnh_\mu} \to 0$
- (H_G) $m \ge 4, h_G \to 0, \log(n)\frac{1}{m^2nh_G^2} \to 0$
- (M_μ) For some $\alpha > 2, E(\sup_{t\in[0,1]} |X(t) - \mu(t)|^\alpha) < \infty, E|\xi_{ij}|^\alpha < \infty$ and
$$(\frac{1}{m}h_\mu + h_\mu^2)\left(\frac{\log(n)}{n}\right)^{2/\alpha-1} \to \infty$$
- (M_G) For some $\beta > 2, E(\sup_{t\in[0,1]} |X(t) - \mu(t)|^{2\beta}) < \infty, E|\xi_{ij}|^{2\beta} < \infty$ and
$$\left(\frac{1}{m^2}h_G^2 + \frac{1}{m}h_G^2 + h_G^4\right)\left(\frac{\log(n)}{n}\right)^{2/\beta-1} \to \infty$$

6.1.6 Covariance Estimation

Denote

$$b_n = h_\mu^2 + \left[\log(n)(\frac{1}{mnh_\mu} + \frac{1}{n})\right]^{1/2} + h_G^2 + \left[\log(n)(\frac{1}{m^2nh_G^2} + \frac{1}{n})\right]^{1/2}$$

Theorem: (Thm 2) Under conditions (K,S, H_g, S_G, M_G)

$$\sup_{t,s\in\mathcal{T}} |\hat{G}(t,s) - G(t,s) = O(b_n)$$

almost surely. Note in the sparse case when $m \le c < \infty$, if $h_\mu = (\frac{\log(n)}{n})^{1/5}, h_G(\frac{\log(n)}{n})^{1/6}$, then

$$\sup_{t,s\in\mathcal{T}} |\hat{G}(t,s) - G(t,s) = (\frac{\log(n)}{n})^{1/3}$$

6.1.7 Eigenanalysis

The rate of convergence of $(\hat{\lambda}_n, \hat{\phi}_k)$ follows from those of $\hat{\mathcal{G}}$. We need a useful lemma. Let $\mathcal{G}_1, \tilde{\mathcal{G}}_1 : L^2 \to L^2$ are two compact self-adjoint operators. Their spectral decompositions are

$$\mathcal{G}_1 = \sum_{j=1}^{\infty} \lambda_{1j}\phi_{1j} \otimes \phi_{1j}, \quad \tilde{\mathcal{G}}_1 = \sum_{j=1}^{\infty} \lambda_{1j}\tilde{\phi}_{1j} \otimes \tilde{\phi}_{1j}$$

To compare eigenfunctions, take $\tilde{\phi}_{1j}$ ne s.t $\langle \phi_{1j}, \tilde{\phi}_{1j} \rangle \ge 0$. By potentially flipping the sign of $\tilde{\phi}_{1j}$,

Lemma: (Bosq 2000, lemma 4.2-4,3))

1. $\sup_{j=1,2,\dots} |\lambda_{1j} - \tilde{\lambda}_{1j}| \le \|\mathcal{G}_1 - \tilde{\mathcal{G}}_1\|$ where $\|\mathcal{G}\|$ is the operator norm of $\mathcal{G}$.
2. If the multiplicity of λ_{1j} is 1, then for $j = 1, 2, \dots$, $\|\phi_{1j} - \tilde{\phi}_{1j}\| \le 2\sqrt{2}\psi_j\|\mathcal{G}_1 - \tilde{\mathcal{G}}_1\|$ where $\psi_j = \begin{cases} (\lambda_1 - \lambda_2)^{-1} & j = 1 \\ \max\{(\lambda_{j-1} - \lambda_j)^{-1}, (\lambda_j - \lambda_{j+1})^{-1}\} & j = 2 \end{cases}$ is the inverse of the i-th eigengap.

For more precise results, see the parturbation theorem (Chap 5 HE)

Corollary: (For FPCA)
Under the conditions of thm2,

$$\sup_{k=1,2,\dots} |\hat{\lambda}_k - \lambda_k| = O(b_n) \quad a.s.(1)$$

Assume $\lambda_1 > \lambda_2 > \dots > 0$, For any $j = 1, 2, \dots$

$$|\hat{\phi}_j - \phi_j| = O(b_n) \quad a.s.(2)$$

and

$$\sup_{t\in\mathcal{T}} |\hat{\phi}_j(t) - \phi_j(t)| = O(b_n) \quad a.s.(3)$$

Proof. Let $\mathcal{G}, \hat{\mathcal{G}}$ be the integral operator associated with $G, \hat{G}$ resp. Then

$$\|\hat{\mathcal{G}} - \mathcal{G}\| \le \|\hat{\mathcal{G}} - \mathcal{G}\|_{HS} = \left[\int_{\mathcal{T}}\int_{\mathcal{T}} [\hat{G}(t,s) - G(t,s)]^2 dtds\right]^{1/2} = O(b_n) \quad a.s.$$

Now (1) and (2) are implied be the previous lemma. For (3), note

$$\begin{aligned}
|\hat{\lambda}_j\hat{\phi}_j(t) - \lambda_j\phi_j(t)| &= \left|\int_{\mathcal{T}} \hat{G}(t,s)\hat{\phi}_j(s) - G(t,s)\phi_j(s)ds\right| \\
&= \left|\int_{\mathcal{T}} (\hat{G}(t,s) - G(t,s))\hat{\phi}_j(s) + G(t,s)(\hat{\phi}_j(s) - \phi_j(s))ds\right| \\
&\leq_{abs} \int_{\mathcal{T}} \left|\hat{G}(t,s) - G(t,s)\right| \left|\hat{\phi}_j(s)\right| + |G(t,s)| \left|\hat{\phi}_j(s) - \phi_j(s)\right| ds \\
&\leq_{CS} \sup_{t,s} \left|\hat{G}(t,s) - G(t,s)\right| \left\|\hat{\phi}_j(s)\right\| + \left|\int G(t,s)^2 ds\right|^{-1/2} \left\|\hat{\phi}_j(s) - \phi_j(s)\right\|
\end{aligned}$$

Now take sup ober t on both sides,

$$\begin{aligned}
\sup_t |\hat{\lambda}_j\hat{\phi}_j(t) - \lambda_j\phi_j(t)| &= \sup_t \left|(1 - \frac{\hat{\lambda}_j}{\lambda_j})\hat{\phi}_j(t) + \frac{\hat{\lambda}_j(t) - \lambda_j\phi_j(t)}{\lambda_j}\right| \\
&\leq \left|(1 - \frac{\hat{\lambda}_j}{\lambda_j}\right| \sup_t \left|\phi_j(t) + \hat{\phi}_j(t) - \phi_j(t)\right| + O(b_n) \quad a.s. \\
&= O(b_n) + O(b_n) = O(b_n) \quad a.s.
\end{aligned}$$

6.2 Estimating derivative of functional data

The deruvative functions $X^{(v)}(t)$ are of special interests as they reflect a different physical process from the original $X(t)$.

Data: datai same as PACE. Observe $\{(T_{ij}, Y_{ij})\}_{i=1,j=1}^{n,m_i}, Y_{ij} = X_i(\Gamma_{ij}) + \varepsilon_{ij}$.

KL expansion of X_i is

$$X_i(t) = \mu(t) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t), t \in \mathcal{T}$$

(Lin and Muller (2009) JASA, FPCA)

Take derivative of the KL expansion

$$X_i^{(v)}(t) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k^{(v)}(t)$$

FPCA method targets the truncated representation

$$X_{i,FPC}^{(v)}(t) = \sum_{k=1}^{\mathcal{K}} \xi_{ik}\phi_k^{(v)}(t)$$

To estimate $\mu^{(v)}$, apply $(v+1)$ - degree local polynomial and obtain

$$(\hat{\alpha}_0, \ldots, \hat{\alpha}_{v+1}) = \arg\min_{\alpha_0,\ldots,\alpha_{v+1}} \sum_{i=1}^{n}\sum_{j=1}^{m_i} K(\frac{T_{ij} - t}{h_{\mu,v}})\{Y_{ij} - \sum_{m=1}^{v+1} \alpha_m(T_{ij} - t)^m\}^2$$

where $h_{\mu,v} > 0$ is th BW for estimating the v-th derivative by matching up terms in Taylor series, we have $\hat{\mu}^{(v)} = \hat{\alpha}_v \cdot v!$

Here we use $(v+1)$ to get a better perfomance when estimating v-th derivative. Yet it may not be the case when the dataset is small. Local linear maybe better.

Local polynomial: $\sum \alpha_m(T_{ij} - t)^m$; Taylor: $\mu(T_{ij}) - \mu(t) = \sum \frac{\mu^{(m)}(t)}{m!} \cdot (T_{ij} - t)^m$ To estimate ξ_{ik}, invoke the PACE procedure and target

$$\tilde{\xi}_{ik} = \lambda_\lambda \phi_{ik}' \Sigma_{Y_i}^{-1}(Y_i - \mu_i)$$

To estimate $\phi_k^{(v)}(t)$, note that if $\hat{\phi}_k(t)$ is available and smooth, one can use its (v)-th order derivative as an estimate of $\phi_k^{(v)}(t)$.

The numerical derivative is wiggly since the best bandwidth for estimate $\phi(t)$ is always too small for estimating the derivatives.

Therefore, a more mothodologically justified method is to derive the expression

$$\phi_k(t) = \frac{1}{\lambda_k}\int G(t,s)\phi_k(s)ds$$

wrt t, obtaining

$$\phi_k^{(v)}(t) = \frac{1}{\lambda_k}\frac{\partial^v}{\partial t^v}\int G(t,s)\phi_k(s)ds = \frac{1}{\lambda_k}\int \frac{\partial^v}{\partial t^v}G(t,s)\phi_k(s)ds$$

assuming $\frac{\partial^v}{\partial t^v}G(t,s)$ exists and is dominated by a square integrable function. Transform the problem from estimating a derivative function to estimating a covariance matrix.

Denote $G^{v,u}(t,s)$ as $\frac{\partial^{v+u}}{\partial t^v \partial s^u}G(t,s)$.

To estimate $G^{v,0}(t,s), t,s \in \mathcal{T}$,

$$\begin{aligned}&(\hat{\alpha}_{1,0},\ldots,\hat{\alpha}_{1,v+1},\hat{\alpha}_{21})\\ =&\arg\min_{\alpha_{2,1},\alpha_{1,m},m=0,\ldots,v+1}\sum_{i=1}^{n}\sum_{1\le j\ne j'\le m_i}K\left(\frac{T_{ij}-t}{h_{G_v}}\right)K\left(\frac{T_{ij'}-s}{h_{G_v}}\right)\left[G_{ijj'}-\sum_{m=0}^{v+1}\alpha_{1m}(T_ij-t)^m-\alpha_{21}(T_{ij'}-s)\right]^2\end{aligned}$$

$$\hat{G}_k^{(0,v)}(t) = \frac{1}{\hat{\lambda}_k}\int \hat{G}^{(v,0)}(t,s)\hat{\phi}_k(s)ds.$$

(Dai, Tao and Muller (2018), Statistica Sinica, DPCA)

An issue of the FPCA method is that $\phi_k^{(v)}$ is no longer the eigen function of $X_i^{(v)}$, therefore $X_{ik}^{(v)}$ is not the most parsimonious expansion. The DPCA method instead targets the KL expansion of $x^{(v)}$ directly.

Firstly by differentiation under the integral sign

$$E(X^{(v)}(t)) = \frac{d^v}{dt^v}E(X(t)) = \mu^{(v)}(t)$$

So $\mu^{(v)}(t)$ is the mean function of $X^{(v)}, t\in\mathcal{T}$. The covairance of $X^{(v)}$ is

$$\begin{aligned}G_v(t,s) &:= cov(X^{(v)}(t),X^{(v)}(s))\\ &= E\{[X^{(v)}(t)-\mu^{(v)}(t)][X^{(v)}(s)-\mu^{(v)}(s)]\}\\ &= E\{\frac{\partial^v}{\partial t^v}\frac{\partial^v}{\partial s^v}[X(t)-\mu(t)][X(s)-\mu(s)]\}\\ &= \frac{\partial^{2v}}{\partial t^v\partial s^v}E\{[X(t)-\mu(t)][X(s)-\mu(s)]\}\\ &= G^{(v,v)}(t,s)\end{aligned}$$

To estimate $G^{(v,v)}(t,s)$ apply a bivariate LP with degree up to $2v+1$, minimizing

$$\sum_{i=1}^{n}\sum_{1\le j\ne j'\le m_i}K(\frac{T_{ij}-t}{h_{G_{vv}}})K(\frac{T_{ij}-s}{h_{G_{vv}}})[G_{ijj'}-\sum_{0\le p+q\le 2v+1}\alpha_{pq}(T_ij-t)^p(T_ij-s)^q]^2$$

Set $\hat{G}^{v,v}(t,s) = \hat{\alpha}_{vv}\cdot(v!)^2$

Let $\mathcal{G}_v : L^1\to L^2$ be the integral operator associated with $G_v = G^{v,v}$ and write its eigendecomposition

$$\mathcal{G}_v = \sum_{k=1}^{\infty}\lambda_{kv}\phi_{kv}\otimes\phi_{kv}$$

The eigenpairs (λ_{kv},ϕ_{kv}) are estimated by those $(\hat{\lambda}_{kv},\hat{\phi}_{kv})$ of $\hat{\mathcal{G}}_v$, associated with $\hat{G}_v$. ϕ_{kv} is called derivative eigenfunction. The KL expansion of $X_i^{(v)}$ is

$$X_i^{(v)}(t) = \mu^{(v)}(t) + \sum_{k=1}^{\infty}\xi_{ikv}\phi_{kv}(t), t\in\mathcal{T} \quad (3)$$

where $\xi_{ikv} = \int_{\mathcal{T}} (X_i^{(v)}(t) - \mu^{(v)})\phi_{kv}(t)dt$ is called the derivative principal component (DPC) score.

We use the BLUP to estimate ξ_{ikv} which gives

$$\tilde{\xi}_{ikv} = BLUP(\xi_{ikv}|Y_i) = g_{ikv}^T \Sigma_{Y_i}^{-1}(Y_i - \mu_i) \quad (4)$$

where

$$\begin{aligned}
[g_{ikv}]_j &= cov(\xi_{ikv}, Y_{ij}|T_{ij}) \\
&= cov(\xi_{ikv}, X_i(T_{ij})|T_{ij}) \\
&= E\{\int [X_i^{(v)}(t) - \mu^{(v)}]\phi_{kv}(t)dt[X_i^{(v)}(T_{ij}) - \mu^{(v)}(T_{ij})]\} \\
&= \int E\{[X_i^{(v)}(t) - \mu^{(v)}][X_i^{(v)}(T_{ij}) - \mu^{(v)}(T_{ij})]\}\phi_{kv}(t)dt \\
&= \int G^{(v,0)}(t, T_{ij})\phi_{kv}(t)dt
\end{aligned}$$

One applies plugin estimate for all quantities in (3) and (4).

6.3 Functional Concurrent Regression (FCR)

c.f. Senturk Muller (2010), Senturk Nguyen (2011) https://www.tandfonline.com/doi/pdf/10.1198/jasa.2010.tm09228?needAccess=true

Model: Consider a sample of paired functions $(X_i(t), Y_i(t)), t = 1, \ldots, n$, where

$$Y_i(t) = \alpha(t) + X_i(t)^T\beta(t) + \varepsilon_i(t), t \in \mathcal{T}(0) = \alpha(t) + \sum_{k=1}^{d} X_{ik}(t)\beta_k(t) + \varepsilon_i(t)(1)$$

. Here $Y_i(t)$ is a response function, $X_i(t) = [X_{i1}(t), \ldots, X_{in}(t)] \in \mathbb{R}^d$ is a vector of prefictor functions modeled as random. $\beta(t) = [\beta_1(t), \ldots, \beta_n(t)]$ is a vector of coefficient function. and f$\varepsilon_i(t)$ is the error process with mean 0 and $E(\varepsilon_i(t)X_i(t)) = 0, t \in \mathcal{T}, \sup_{t\in\mathcal{T}} var(\varepsilon_i(t)) < \infty$.

Concurrent means the value of response at time t only depends on the observations at t.

$X_{ik}(t) := z_{ik}, t \in \mathcal{T}$ is a special case. Model (1) can handle scalar predictors. the k-th term in (1) becomes $z_{ik}\beta_k(t)$

The function ANOVA is a special case. If there are d groups, and observe $(g_i, Y_i(t))$, the model becomes

$$Y_i(t) = \sum_{k=1}^{d} z_k\mu_k(t) + \varepsilon_i(t)$$

where $z_{ik} = \begin{cases} 1, & g_i = k, k = 1, \ldots, d \\ 0, & g_i \neq k \end{cases}$ and $\mu_k(t)$ is the mean function of group k.

Take covariance with $X_i(t)$ on both sides of (0), obrain the normal equation

$$cov(X_i(t), Y_i(t)) = cov(X_i(t), X_i(t))\beta(t), t \in \mathcal{T}$$

dimension: $d \times 1, d \times d, d \times 1$. So

$$\beta(t) = cov(X_i(t))^{-1}cov(X_i(t), Y_i(t))$$

Denote $G_{kk'}(t, s) := cov(X_{ik}(t), X_{ik'}(s))$ and $G_{kY} := cov(X_{ik}(t), Y_i(s))$.

$$cov(X_i(t)) = \begin{bmatrix} G_{11}(t,t) & G_{12}(t,t) & \cdots & G_{1d}(t,t) \\ \vdots & & & \vdots \\ G_{d1}(t,t) & G_{d2}(t,t) & \cdots & G_{dd}(t,t) \end{bmatrix}$$

$$cov(X_i(t), Y_i(t)) = \begin{bmatrix} G_{1Y}(t,t) \\ \vdots \\ G_{dY}(t,t) \end{bmatrix}$$

Take expected value on bothe sides of (0), with $\mu_Y(t) := E(Y_i(t))$ and $\mu_X := E(X_i(t))$

$$\alpha(t) = \mu_Y(t) = \mu_X(t)^T\beta(t)$$

One can also consider a functional concurrent R^2

$$R^2(t) = cov(X_{ik}(t), Y_i(s))^T cov(X_{ik}(t))^{-1} cov(X_{ik}(t), Y_i(s))$$

which quantifies the percent of variance explained by the concurrent regression relationship at t.

Dense case

Suppose $(X_i(t), Y_i(t))$ are fully observed over $t \in \mathcal{T}$ without noise. Then

$$\hat{\mu}_X(t) = \hat{X}(t) = \frac{1}{n}\sum_i X_i(t), \quad \hat{\mu}_Y(t) = \bar{Y}(t) = \frac{1}{n}\sum_i Y_i(t)$$

$$\hat{cov}(X_i(t)) = \frac{1}{n}\sum_i [X_i(t) - \bar{X}(t)][X_i(t) - \bar{X}(t)]^T$$

$$\hat{cov}(X_i(t), Y_i(t)) = \frac{1}{n}\sum_i [X_i(t) - \bar{X}(t)][Y_i(t) - \bar{Y}(t)]^T$$

$$\hat{\beta}(t) = \hat{cov}(X_i(t))^{-1}\hat{cov}(X_i(t), Y_i(t))$$
$$\hat{\alpha}(t) = \hat{\mu}_Y(t) - \hat{\mu}_X(t)^T\hat{\beta}(t)$$

If X_i and Y_i are smooth over t, then $\hat{\beta}$ will also be smooth If X_i and Y_i are tough but prior knowledge indicates that β should be smooth, the smoothing techniques can be applied when constructing $\hat{cov}(X_i(t)), \hat{cov}(X_i(t), Y_i(t)i(t))$

Note the observation at t is small we need to borrow information from neighborhood.

Sparse case

$(X_i(t), Y_i(t))$ are unobserved . Instead observe $(T_{ij}, \tilde{X}_{ij}, Y_{ij})$ where

$$\tilde{X}_{ij} = X_i(T_{ij}) + \varepsilon_{ij}, \quad i.e\tilde{X}_{ijk} = X_{ik}(T_{ij}) + \varepsilon_{ijk}, k = 1, \ldots, d$$

$$Y_{ij} = Y_i(T_{ij}) + \varepsilon_{ij}$$

for $i = 1, \ldots, n, j = 1, \ldots, m_i$. Here $\varepsilon_{ij} = [\varepsilon_{i1j}, \ldots, \varepsilon_{idj}, \varepsilon_{ij}]^T$ and bare iid with mean zero, and have finite variance $\sigma_1^2, \ldots, \sigma_d^2, \sigma_Y^2$ resp.

The mean functions $\mu_{X_1}, \ldots, \mu_{X_d}, \mu_Y$ are estimated by the LL smoother, we habe seen how to estimate $G_{kk}(t, s) = cov(X_{ik}(t), X_{ik}(s))$ by performing 2D local polyomial smoothing on the row covariances obrained from the observations of the k-th prediction process. For $k \neq k' \in \{1, \ldots, d\}$, to estimate $G_{kk'}(t, s) = cov(X_{ik}(t), X_{ik'}(s))$. Note that the row cross-covariance

$$G_{ijj'kk'} := [X_{ikj} - \mu_{X_k}(T_{ij})][X_{ik'j'} - \mu_{X_{k'}}(T_{ij'})]^T$$

has conditional mean

$$E(G_{ijj'kk'}|T_{ij}, T_{ij'}) = G_{kk'}(T_{ij}, T_{ij'})$$

Set $\hat{G}_{kk'}(t, s) = \hat{\alpha}_0$,

$$(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2) = \arg\min_{\alpha_0, \alpha_1, \alpha_2} \sum_{i=1}^{n}\sum_{j,j'=1}^{m_i} K(\frac{T_{ij} - t}{h_1})K(\frac{T_{ij'} - t}{h_2})\{G_{ijj'kk'} - (\alpha_0 + \alpha_1(T_{ij} - t) + \alpha_2(T_{ij'} - s))\}$$

Finally, plug in the results we get $\hat{\beta}, \hat{\alpha}$.

6.4 Scalar response functional linear models

Model $(X_i, Y_i), i = 1, \ldots, n$ are iid pairs of observations, where $Y_i \in \mathbb{R}$ is a scalar response and $X_i(t), t \in \mathcal{T}$ is a densely observed functional predictor, which is a L^2- stochastic process. The relationship between X_i, Y_i is linear satisfying

$$Y_i = \alpha + \int_{\mathcal{T}} X_i(s)\beta(s)ds + \varepsilon_i \quad (1)$$

where ε_i is a zero mean random error independent of X_i with $var(\varepsilon_i) = \sigma^2, \beta \in L^2, \alpha \in \mathbb{R}$.

The objective to estimate α, β.

Normal equation

Take cov with $X_i(t)$ on both sides of (1).

$$\begin{aligned} cov(X_i(t), Y_i) &= cov(X_i(t), \int_{\mathcal{T}} X_i(s)\beta(s)ds) \\ &= E(\int (X_i(t) - \mu_X(t))(X_i(s) - \mu_X(s))\beta(s)ds) \\ &= \int G(t,s)\beta(s)ds \\ &=: \mathcal{G}\beta \end{aligned}$$

where μ_X is the mean funtion of X_i, G is the covariance ufnction and $\mathcal{G}$ is the ingral operator associated with G. Write $g(t) = cov(X_i(t), Y_i)$, then the normal euqation is

$$\mathcal{G}\beta = g$$

There is difficulty in inverting $\mathcal{G}$.

For simplicity and identifiability, assume $ker(\mathcal{G}) = \{0\}$, i.e. the null space. If $ker(\mathcal{G}) = \{f : \mathcal{G}f = 0\} \neq \{0\}$ then if β_0 is a solution of the normal equation then $\beta_0 + \delta$ is also a solution and $\delta \in ker(\mathcal{G})$.

Recall that the covariance operator $\mathcal{G} : L^2 \to L^2$ has the eigen decomposition

$$\mathcal{G} = \sum_{k=1}^{\infty} \lambda_k \phi_k \otimes \phi_k$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ that goes down to zero and the ϕ_k is a CONS for $Im(\mathcal{G})$. If all λ_k are nonzero, the inverse of $\mathcal{G}$ is fomally written as

$$\mathcal{G}^{-1} = \sum_{k=1}^{\infty} \lambda_k^{-1} \phi_k \otimes \phi_k \quad \mathcal{G}^{-1} : \overline{Im(\mathcal{G})} \to L^2$$

is not necessarily converging, just an algebraic representation.

The inverse is an unbounded operator since

$$\|\mathcal{G}^{-1}\phi_l\| = \|\sum_{k=1}^{\infty} \lambda_k^{-1} \langle \phi_k, \phi_l \rangle\| = \|\lambda_l^{-1}\phi_l\| = \lambda_l^{-1} \to \infty, as\, l \to \infty$$

In summary, the problem of inverting $\mathcal{G}$ is ill-posed, i.e. a small perturbation in $\mathcal{G}$ mau lead to a larger pertubation in the inverse.

Also the sample covariance operator

$$\mathcal{G}_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) \otimes (X_i - \bar{X})$$

associated with the sample covariance function

$$\mathcal{G}(t,s) = \frac{1}{n} \sum_{i=1}^{n} (X_i(t) - \bar{X}(t))(X_i(s) - \bar{X}(s))$$

has finite rank $(n-1)$ so the inverse of $\mathcal{G}_n$ does not exists.

To estimate

$$\beta = \mathcal{G}^{-1}g = \sum_{k=1}^{\infty} \lambda_k \langle g, \phi_k \rangle \phi_k \quad (2)$$

Two stategies are common: regulatization and truncation.

Regularization Approach

(Ramsay Silverman 2005, KR 2019 4.5) Target to minimize

$$P_\lambda(\alpha, \beta) = \sum_{i=1}^{N} (Y_i - \alpha - \int \beta(t) X_i(t) dt + \lambda \int (\beta^{(m)}(t))^2 dt$$

where $m \geq 0$ is the order of the derivative one wants to penalize. Usually $m = 0, 2$

$m = 0$ corresponds to a ridge penalty. Solution to the population version of (3) is

$$\beta\lambda = (\mathcal{G} + \lambda I)^{-1} g = \sum_{k=1}^{K} (\lambda_k + \lambda)^{-1} \langle g, \phi_k \rangle \phi_k$$

For dense FD, this can be implemented with basis-expansion approximating $X_i(t)$ by

$$\tilde{X}_i(t) = c_i^T B(t)$$

where $B(t) = [B_1(t), \ldots, B_K(t)]^T$ is a set of basis function (B-Spline, Fourier bases, etc). And $C_i = [C_{i1}, \ldots, C_{iK}]^T \in \mathbb{R}^K$ are the basis coefficient. The C_i are estimated using least squares with the dense observations for X_i. Choose K to be rather large so that the approximation error $\tilde{X}_i(t) - X_i(t)$ is negligible. Approximate $\beta(t)$ by $b^T B(t), b \in \mathbb{R}^K$. Then the linear model (1) is written as

$$Y = 1\alpha + CJb + \varepsilon$$

where dim: $n \times 1, n \times 1, n \times K$ $K \times K$ $K \times 1, n \times 1$. And $1 = [1, \ldots, 1]^T, C = [C_1, \ldots, C_K]^T, J = \int B(t) B(t)^T dt$ and ε contains noise. The penalized least square problem (3) becomes

$$\|Y - 1a - CJb\| + \lambda b^T \Omega b$$

where $\Omega = \int B^{(m)}(t)[B^{(m)}(t)]^T dt$. The tuning parameter λ can be chosen by CV. The solution is

$$\hat{\beta}(t) = \hat{b}^T B(t), \quad \hat{b} = [(CJ)^T (CJ) + \lambda\Omega]^{-1} (CJ)^T Y$$

Truncation Approach

Target

$$\beta_k = \sum_{k=1}^{K} b_k \phi_k$$

where $b_k = \langle \beta, \phi_k \rangle$. Q: advantage of orthonormal basis: we can use itemwise regression, and no need to involve matrix calculation.

Implemenation: express $X_i - \mu_X, \beta$ onto the eigen basis, obtaining

$$X_i - \mu_X = \sum_{k=1}^{\infty} \xi_{ik} \phi_k \quad \beta = \sum_{k=1}^{\infty} b_k \phi_k$$

then

$$\langle X_i - \mu, \beta \rangle = \langle \sum_{k=1}^{\infty} \xi_{ik} \phi_k, \sum_{k=1}^{\infty} b_k \phi_k \rangle = \sum_{k=1}^{\infty} \xi_{ik} b_k \approx \sum_{k=1}^{K} \xi_{ik} b_k \quad (4)$$

The regression model is equivalently written as

$$Y_i = \mu_Y + \int (X_i(s) - \mu_X(s)) \beta(s) ds + \varepsilon_i = \mu_Y + \langle X_i - \mu_X, \beta \rangle + \varepsilon_i$$

where $\mu_Y = E(Y)$, so

$$Y_i \approx \mu_Y + \sum_{k=1}^{K} \xi_{ik} b_k + \varepsilon_i \quad (5)$$

Estimate of the b_k are obrained by regressing Y_i on $\{\xi_{i1}, \ldots, \xi_{iK}\}$.
In practice, the FPC regression algorithm is performed as follows

- 0. Decide K
- 1. Obtain the FPC scores $\hat{\xi}_{ik}, k-1, \ldots, K, i = 1, \ldots, n$ from the functional observations. Also obtain the $\hat{\phi}_k$
- 2. Regress Y_i on $\hat{\xi}_{ik}, k = 1, \ldots, K$, obraining estimates

$$\hat{b}_k = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i \hat{\xi}_{ik}}{\frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_{ik}}$$

  which estimates $b_k = \frac{\langle g, \phi_k \rangle}{\lambda_k}$

- 3. Reconstruct

$$\hat{\beta}_K(t) = \sum_{k=1}^{K} \hat{b}_k \hat{\phi}_k(t)$$

  as an estimate of $\beta(t)$

K is rather small compared to the sample size. To select K, one can apply cross validation.
If the data is sparse, ξ_{ik} has to be estimated by BLUP, which is not consistent.

Extension for FLM

Let Y be a scalar response, $X(t), t \in \mathcal{T}$ a functional predictor functional quadratic regression (Yao, Muller 2010)

$$E(Y|X) = \alpha + \int_{\mathcal{T}} \beta(t) X^c(t) dt + \int_{\mathcal{T}} \int_{\mathcal{T}} r(t,s) X^c(t) X^c(s) ds dt$$

where $X^c(t) = X(t) - \mu(t), t \in \mathcal{T}$. Now $\{\phi_k(t)\phi_l(t), t, s \in \mathcal{T}\}_{k,l=1}^{\infty}$ is an orthonormal basis for $Ł^2(\mathcal{T} \times \mathcal{T})$. Expand on the eigen bases,

$$\beta(t) = \sum_{k=1}^{\infty} b_k \phi_k(t), \quad r(t,s) = \sum_{k,l=1}^{\infty} \tilde{r}_{kl} \phi_k(t) \phi_l(t), t, s \in \mathcal{T}$$

So

$$E(Y|X) = \alpha + \sum_{k=1}^{\infty} \beta_l \xi_k + \sum_{k,l=1}^{\infty} \tilde{r}_{kl} \xi_k \xi_l$$

Here $\tilde{r}_{kl} = \int \int r(t,s) \phi_k(t) \phi_l(s) dt ds$
Diagnostics: Plot the scores versus residuals/ response. See also Chiou Muller (2007, CSDA)
Functional Generalized LInear models (Moller and Stadtmuller, 2005, Annals)

$$Y = g\{\alpha + \int \beta(t) X(t) dt\} + \varepsilon$$

where g is a known (inverse) link function, and the response Y can be non-Gaussian.
Generalized multi-level functional regression (Crainireanu Staicu Di 2013 JASA)
Respect to Y_i: hypertension (1 or 0), ovserve

$$W_{ij}(t) = \mu(t) + \eta_j(t) + X_i(T) + U_{ij}(t) + \varepsilon_{ij}(t), t \in [0,1], i = 1, \ldots, I, j = 1, \ldots, J$$

.

$W_{ij}(t)$ is the EEG signal of subject i in the j-th visit.
For each subject, baseline covariates Z_i are available, which contains age, fender, BMI etc. The model associating X_i and Y_i is

$$Y_i \sim Exp(\theta_i, \alpha)$$

where

$$\theta_i = \int X_i \beta(t) dt + Z_i^T r$$

Truncated linear models functional data (Hall Hooker 2016, JRSSB, Guan Lin Cao 2018)

$$Y = \alpha + \int_0^\theta \beta(t) X(t) dt + \varepsilon$$

where $X(t), t \in [0,1], 0 < \theta < 1$. Y only depends on a part of the time. The objective is to estimate θ.

Minimax properties for scalar response FLM

Know that $\hat{\beta}_K \to \beta_K$ for each K, as $n \to \infty$ by LLM. Also we have $\beta_K \to \beta$, as $K \to \infty$. The Goal is to find K and n such that $\hat{\beta} \to \beta$. Hall& Horowitz (2007) considers the rt of convergence for $\hat{\beta}_K$ to the true β. A set of assumptions are needed. Throughour , C is a positive constant.

- (M) X has finite 4-th moment in that $\int EX^4(t)dt < \infty, E(\xi_n^4) \leq C\lambda_n^2, \forall k = 1, 2, \ldots$
- (EG) For some $\alpha > 1$,
$$\lambda_k - \lambda_{k+1} \geq C^{-1} k^{-(\alpha+1)}, k = 1, 2 \ldots$$
- (RF) $|b_k| \leq Ck^{-r}$, $r > \frac{1}{2}\alpha + 1$
- (K) $K = K_n \asymp x^{\frac{1}{\alpha+2r}}$

Let $\mathcal{F}(C, \alpha, r)$ denote the set of distributions F of (X, Y) that satisfy (M, EG, RF) for a given values of C, α, r. Let $\mathcal{B}$ denote the class of measureable functions $\bar{\beta}$ of the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ generated by the FLM $Y = \alpha + \int \beta X + \varepsilon$

Theorem: If (M, EG, RF, K) hold, then

$$\lim_{D \to \infty} \lim \sup_{n \to \infty} \sup_{F \in \mathcal{F}} P_F \{ \int (\hat{\beta}_k - \beta)^2 > D m^{\frac{-(2r-1)}{\alpha+2r}} \} = 0$$

as $n \to \infty$. Furthermore,

$$\lim \inf_{n \to \infty} n^{\frac{2r-1}{\alpha+2r}} \inf_{\bar{\beta} \in \mathcal{B}} \sup_{F \in \mathcal{F}} \int E_F (\bar{\beta} - \beta)^2 = 0$$

6.5 Functional-response Functional Linear Regression

Yao Muller Wang (2005)

Complete observation model $\{X_i(s), Y_i(t), s \in \mathcal{T}_1, t \in \mathcal{T}_2\}$ iid relization of a parit of stochastic process $\{X(s), Y(t), s \in \mathcal{T}_1, t \in \mathcal{T}_2\}$. The relationship between X and Y is described by a linear model

$$Y_i(t) = \alpha(t) + \int_{\mathcal{T}_1} \beta(T, s) X_i(s) ds + \varepsilon_s(t), t \in \mathcal{T}_2$$

where $i = 1, \ldots, n, \varepsilon_i(t)$ is a zero-mean L^2 stochastic process independent of X_i. The normal equation by taking covariance of both sides with $X_i(s')$ is

$$G_{YX}(t, s') = \int_{\mathcal{T}_1} \beta(t, s) G_{XX}(s, s') ds, t \in \mathcal{T}_2, s \in \mathcal{T}_1$$

where $G_{XX}(s, s') = cov(X(s), X(s')), G_{YX(t,s')=cov(Y(t),X(s')),s,s' \in \mathcal{T}_1, t \in \mathcal{T}_2}$. For $t \in \mathcal{T}_2$,

$$G_{YX}(t, \cdot) = \mathcal{G}_{XX} \beta(t, \cdot) \text{ as alements in } L^2(\mathcal{T}_1)$$

where $\mathcal{G}_{XX}$ is the covariance operator assoviated with G_{XX}.

1. We have an inverse problem similar to the scalar-response case 2. This equation is useful if we want to perform hypothesis test on $\beta(t, s) = 0$, it is equivalent to test $G_{YX}(t, s) = 0, s \in \mathcal{T}_1, t \in \mathcal{T}_2$, which is easier to calculate. The equivalence holds if $Ker(\mathcal{G}_{XX}) = 0$.

FPC approach. Let μ_X, μ_Y denote the mean function of X, Y resp., and G_{XX}, G_{YY} be the corresponding covariance function. WRite

$$X^c(s) = X(s) - \mu_X(s), \quad s \in \mathcal{T}_1$$
$$Y^c(t) = Y(t) - \mu_Y(t), \quad t \in \mathcal{T}_2$$

the centered process.

Since

$$E(Y(t)|X) = \alpha(t) + \int_{\mathcal{T}_1} \beta(t,s)X(s)ds$$

and

$$\mu_Y(t) = \alpha(t) + \int_{\mathcal{T}_1} \beta(t,s)\mu_X(s)ds$$

we have

$$\alpha(t) = \mu_Y(t) - \int_{\mathcal{T}_1} \beta(t,s)\mu_X(s)ds$$

Now

$$\begin{aligned} E(Y^c(t)|X^c) &= E(Y^c(t)|X) \\ &= E(Y(t)|X) - \mu_Y(t) \\ &= \alpha(t) + \int_{\mathcal{T}_1} \beta(t,s)X^c(s)ds + \int_{\mathcal{T}_1} \beta(t,s)\mu_X(s)ds - \mu_Y(t) \\ &= \int_{\mathcal{T}_1} \beta(t,s)X^c(s)ds \end{aligned}$$

Write

$$X_i^c(t) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(s)$$
$$Y_i^c(t) = \sum_{m=1}^{\infty} \zeta_{im}\psi_m(s)$$

and here $\lambda_k = E(\xi_k^2)$ is the k-th eigenvalue for X, $\rho_m = E\xi_m^2$ is the m-th eigen value for Y. If the ϕ_k, ψ_m are both CONS for $L^2(\mathcal{T}_1), L^2(\mathcal{T}_2)$ resp., then the tensor product basis $\{\phi_k(s), \psi_,(t), s \in \mathcal{T}_1, t \in \mathcal{T}_2\}_{k,m=1}^{\infty}$ from a CONS of $L^2(\mathcal{T}_1 \times \mathcal{T}_2)$ of square integrable functions supported on $\mathcal{T}_1 \times \mathcal{T}_2$, thus

$$\beta(t,s) = \sum_{k=1}^{\infty}\sum_{m=1}^{\infty} \beta_{km}\phi_k(s)\psi_m(t)$$

where $\beta_{km} = \int_{\mathcal{T}_1}\int_{\mathcal{T}_2} \beta(t,s)\phi_k(s)\psi_m(t)dsdt, k, m = 1, 2, \ldots$ Then

$$\begin{aligned} Y^c(t) &= \int [\sum_K \sum_m \beta_{km}\phi_k(s)\psi_m(t)][\sum_l \xi_l\phi_l(s)]ds + \varepsilon(t) \\ &= \sum_m \psi_m(t) + \int [\sum_k \beta_{km}\phi_k(s)][\sum_l \xi_l\phi_l(s)]ds + \varepsilon(t) \\ &= \sum_m \phi_m(t) \sum_k \beta_{km}\xi_k + \varepsilon(t) \\ &= \sum_{k,m} \xi_k\beta_{km}\psi_m(t) + \varepsilon(t) \end{aligned}$$

Multiply both sides by ξ_l and take expected value , we have

$$E[\xi_l \sum_m \zeta_m\psi_m(t)] = E\sum_{k,m} \xi_k\xi_l\beta_{km}\psi_m(t)$$

So

$$\sum_m E(\xi_l, \zeta_m)\psi_m(t) = \sum_m \lambda_l\beta_{lm}\psi_m(t)$$

Since $E(\xi_k\xi_l) = 0$ if $k \neq l$.

The inner product with ψ_j on both sides,

$$E(\xi_k\zeta_j) = \lambda_l\beta_{lj}, \quad j, l = 1, 2, \ldots$$

Rename the indices and reorganizing terms, we have

$$\beta_{km} = \frac{\sigma_{km}}{\lambda_k}, \quad k, m = 1, 2, \ldots$$

where $\sigma_{km} = E(\xi_k\zeta_m)$, so

$$\beta(t, s) = \sum_{k,m} \frac{\sigma_{km}}{\lambda_k}\psi_m(t)\phi_k(s), s \in \mathcal{T}_1, t \in \mathcal{T}_2$$

The coefficient of determination is the ratio of explained variance and the total varaince.

$$k^2 = \frac{\int var(E(Y(t)|X))}{\int var(Y(t))dt} = \frac{E\int E(Y^c(t)|X^c)^2dt}{E\int (Y^c(t))^2dt}$$

$$= \frac{E(\sum_{k,m}(\xi_k\beta_{km})^2)}{E(\sum_m \zeta_m^2)}$$

$$= \frac{\sum_{k,m}\beta_{k,m}^2\lambda_k}{\sum_m \rho_m}$$

Estimation for the dense case

Obtain empirical eigenfunctions $\hat{\phi}_k, \hat{\phi}_m, \hat{\beta}_{km}$ by regressing $\{\hat{\zeta}_{im}\}_{i=1}^n$ on $\{\hat{\xi}_{ik}\}_{i=1}^n$. Then

$$\hat{\beta}(t, s) = \sum_{m=1}^{M}\sum_{k=1}^{K}\hat{\beta}_{km}\hat{\phi}_k(s)\hat{\psi}_m(t), s \in \mathcal{T}_1, t \in \mathcal{T}_2$$

Estimation for the sparse case

Rather than FPC regression, apply term-by-term estimate for $\beta_{km} = \sigma_{km}/\lambda_k$. Assume we observe

$$U_{iL} = X_i(S_{iL}) + \varepsilon_{X,iL}, s_{iL} \in \mathcal{T}_1, i = 1, \ldots, n, l = 1, \ldots, L_i$$

$$V_{ij} = Y_i(S_{ij}) + \varepsilon_{Y,ij}, s_{ij} \in \mathcal{T}_2, i = 1, \ldots, n, j = 1, \ldots, N_i$$

The mean function μ_X, μ_Y and covariance function G_{XX}, G_{YY} are estimated by the PACE procedure. The cross-covariance G_{YX} is obtained through smoothing the raw cross0covariances. For $\sigma_{k,m}$ note

$$\sigma_{km} = E(\xi_k\zeta_m) = E(\int_{\mathcal{T}_1} X^c(s)\phi_k(s)ds \int_{\mathcal{T}_2} Y^c(t)\psi_k(t)dt) = \int_{T_1}\int_{\mathcal{T}_2} G_{YX}(t, s)\phi_k(s)\psi_m(t)dtds$$

So

$$\hat{\sigma}_{k,m} = \int_{T_1}\int_{\mathcal{T}_2} \hat{G}_{YX}(t, s)\hat{\phi}_k(s)\hat{\psi}_m(t)dtds$$

Finally

$$\hat{\beta}(t, s) = \sum_{m=1}^{M}\sum_{k=1}^{K} \frac{\hat{\sigma}_{km}}{\hat{\lambda}_k}\hat{\phi}_k(s)\hat{\psi}_m(t)$$

The smaller M, K are, the smoother of the estimator functions. You cannot increase M, K too quickly due to inverse problem. The estimator is consistent.

Functional Additive Model (FAM)

$$E(Y(t)|X) = \mu_Y(t) + \sum_m\sum_k f_{km}(\xi_k)\psi_m(t)$$

extends the linear to nonlinear.

Funcational Historical Regression Model assumes the later observations are not responsible for the earlier observations.

7 Inference

Objective: try to derive functional CLT. Ref: HE Chap 2.6, 7.2, 7.4, 8.1.

7.1 Bochner integral

To define LLN and CLT for functional data, one needs the mean and covariance elements defined through the Bochner integral.

Suppose we have a function $f : E \to \mathbb{X}$ defined on a measurable space $(E, \mathcal{B}, \mu)$ that takes values in a Banach space $\mathbb{X}$.

Simple functions

A function f is simple if it can be written as

$$f(\omega) = \sum_{i=1}^{k} 1_{E_i}(\omega) g_i$$

for some finite k, $E_i \in \mathcal{B}$ and $g_i \in \mathbb{X}$. Any simple function $f(\omega) = \sum_{i=1}^{k} 1_{E_i}(\omega) g_i$ with $\mu(E_i) < \infty$ for all i is said to be Bochner integrable and its integral is defined to be

$$\int_E f d\mu = \sum_{i=1}^{k} \mu(E_i) g_i$$

The integral is also in $\mathbb{X}$.

$$f(\omega) = w.t. \quad f_n(\omega) \ \frac{1}{2}[1_{[0,1/2]}(\omega) \cdot (t/2) + 1_{[1/2,1]}(\omega) \cdot |t|]$$

:

Measurable functions

A measurable function f is said to be Bochner integrable if there exists a sequence $\{f_n\}$ of simple Bochner integrable functions s.t.

$$\lim_{n\to\infty} \int_E \|f_n - f\| d\mu = 0$$

In this case the Bochner integral of f is defined as

$$\int_E d\mu = \lim_{n\to\infty} \int f_n d\mu$$

The RHS of the last equation makes sense, because

$$\| \int f_n d\mu - \int f_m d\mu \| \overset{Triangular}{\leq} \int \|f_n - f_m\| d\mu \leq \int \|f_n - f\| + \|f - f_m\| d\mu \to 0, m, n \to \infty$$

by assumption. Therefore $\{\int f_m d\mu\}$ is Cauchy and converges to a limit in $\mathbb{X}$.

Theorem: Suppose $\mathbb{X}$ is a seperable Hilbert Space and f is measurable from E to $\mathbb{X}$ with $\int \|f\| d\mu < \infty$. Then f is Bochner integrable.

7.2 Mean and Covariance of a Random Element in Hilbert space

Let $X : (\Omega, \mathcal{F}, P) \to \mathbb{H}$ be a RE takes values in a separable Hilbert Space $\mathbb{H}$. If $E\|X\| < \infty$, the mean (element) of X is defined as the Bochner integrabl

$$EX := \int_\Omega X(\omega) dP(\omega) \in \mathbb{H}$$

Theorem: If m is the mean of X, then

$$\langle m, f\rangle = E(\langle X, f\rangle)$$

for any $f \in \mathbb{H}$

Proof: Let X_n be a sequence of simple random elements s.t $E\|X_n - X\| \to 0$ as $n \to \infty$. Then

$$\langle \lim_{n\to\infty} EX_n, f\rangle = \lim_{n\to\infty} \langle EX_n, f\rangle =_{DCT} E\langle X, f\rangle$$

Since $|\langle X_n, f\rangle| \le \|X_n\|\|f\| \le (\|X\| + \varepsilon)\|f\|$m where the last is Lebesgue integral. In fact, (1) can be taken as an equivalent definition of the mean element, due to Riesz Representation theorem.

If define a covariance, consider an analogy to a Eudlidean case. If X is a p-dim random vector. Then

$$Cov(X) = E\left((X - EX)(X - EX)^T\right) = E\left((X - EX) \otimes (X - EX)\right)$$

Recall Hilbert Schmit operator is a bounded operator A on a Hilbert space H with finite Hilbert–Schmidt norm.

Assume that $E\|X\|^2 < \infty$, then the covariance operator for X is the element $\mathcal{K} : \mathbb{H} \to \mathbb{H}$ of $\mathcal{B}_{HS}(\mathbb{H})$ given by the Bochner integral

$$\mathcal{K} = E\left((X - EX) \otimes (X - EX)\right) := \int_\Omega (X - EX) \otimes (X - EX) dP \quad (3)$$

Recall that for $f, g, h \in \mathbb{H}$, $f \otimes g \in \mathcal{B}(\mathbb{H}, \mathbb{H})$, $\quad (f \otimes g)(h) = \langle f, g\rangle g$.

Theorem: Suppose that $E\|X\| < \infty$, then for $f, g \in \mathbb{H}$,

$$\mathcal{K}f = E(\langle X - m, f\rangle(X - m))$$

$$\langle \mathcal{K}f, g\rangle = E\langle X - m, f\rangle\langle X - m, g\rangle$$

The mean and covariance elements are closely related to the mean and covariance function under slightly more assumptions.

Theorem: Let $X = \{X(t), t \in \mathcal{T}\}$ be a mean-square continuous process that is jointly measurable. Consider $\mathbb{H} = L^2(\mathcal{T})$.

1. The mean function $\mu(t), t \in T$ regarded as an element in $\mathbb{H}$ coincide with the mean element m of $X \in \mathbb{H}$.
2. The integral operator $\mathcal{G}$ associated with the covariance function $\mathcal{G}$ coincide with the covariance operator $\mathcal{K}$ in (3).
3. For any $f \in \mathbb{H}$
$$I_X(f) = \int_E X(t) f(t) dt = \langle X, f\rangle$$

When $\mathbb{H} = L^2$. Consider the scalar-response FLM.

Stochastic Process	Random element
$Y = \alpha + \int X(t)\beta(t)dt + \varepsilon$	$Y = \alpha + \langle X, \beta\rangle + \varepsilon$
$\mu_X(t) = EX(t), t \in \mathcal{T}$	$m_X = EX \in \mathbb{H}$
$g(t) := cov(X(t), Y)t \in \mathcal{T}$	$g = E(Y \cdot (X - m))$
$G(t,s) = Cov(X(t), X(s))$	$\mathcal{K} = E(X-m) \otimes (X-m)$
$\mathcal{G}(f)(t) = \int \mathcal{G}(t,s)f(s)ds, t \in \mathcal{T}$	$\mathcal{K}(f) = E(\langle X - m, f\rangle(X-m))$
$= E\{[\int [X(s) - \mu(s)]f(s)ds][X(t) - \mu(t)]\}$	
Normal equation:	
$g(t) = \int G(t,s)\beta(s)ds$	$g = \mathcal{K}(\beta), \beta = \mathcal{K}^{-1}g$
	$\hat{\mathcal{K}} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}) \otimes (X_i = \bar{X}) = \sum_{k=1}^n \hat{\lambda}_k \hat{\phi}_k \otimes \hat{\phi}_k$
	$\hat{g} = \frac{1}{n}\sum_{i=1}^n Y_i(X_i - \bar{X})$
	$\hat{\beta}_k = (\sum_{k=1}^n \hat{\lambda}_k^{-1}\hat{\phi}_k \otimes \hat{\phi}_k)\hat{g}$

7.3 LLN and CLT in Hilbert Space

Let $\mathbb{H}$ be a separable Hilbert space. Assume we have iid sample $X_1, \ldots, X_n \sim X$. We estimate the mean and covariance elements via

$$\hat{m} = \frac{1}{n}\sum_{i=1}^n X_i \in \mathbb{H}$$

$$\hat{\mathcal{K}} - \frac{1}{n}\sum_{i=1}^n (X_i - \hat{m}) \otimes (X_i - \hat{m}) \in \mathcal{B}_{HS}(\mathbb{H})$$

Theorem: (LLN)

1. If $E\|X\| < \infty$, then
$$\hat{m} \to m \text{ a.s. in } \mathbb{H}$$
as $n \to \infty$
2. If $E\|X\|^2 < \infty$, then
$$\hat{\mathcal{K}} \to \mathcal{K} \text{ in } \mathcal{B}_{HS}(\mathbb{H})$$
as $n \to \infty$

To definve a weak convergence in $\mathbb{H}$. Let $P, P_n, n \geq 1$ be probability measures on $\mathbb{H}$ with the Borel σ =algebra defined on $\mathbb{H}$. We say P_n converges weakly to P, defined by $|_n \xrightarrow{\omega} P$ if

$$\int f(x)dP_n(x) \to \int f(x)dP(x)$$

for any bounded continuous function $f : \mathbb{H} \to \mathbb{R}$, as $n \to \infty$. For REs Y and $Y_n, n \geq 1$, we say Y_n converges ini distribution to Y or $Y_n \xrightarrow{d} Y$ if $P \cdot Y_n^{-1} \xrightarrow{\omega} P \cdot Y^{-1}$.

Theorem: The following are equivalent

1. $P_n \xrightarrow{\omega} P$
2. $P_n(A) \to P(A)$ for any Borel set A with $P(\partial A) = 0$ where ∂A is the boundary of A.

Theorem: (CLT) If $E\|X\|^2 < \infty$, then
$$\sqrt{n}(\hat{m} - m) \xrightarrow{d} \xi$$
where ξ is a Gaussian RE with zero mean and covariance operator equal to $\mathcal{K} = E(X_i - m) \otimes (X_i - m)$

Theorem: (Continuous mapping) Let h be a real-valued continous function , and $Y, Y_n, n \geq 1$ are REs. If $Y_m \xrightarrow{d} Y$, then
$$h(Y_n) \xrightarrow{d} h(Y)$$

Let $h(x) = \langle f, x\rangle$ where f is any fixed function in $\mathbb{H}$. then h is a continuous function. The continuous mapping theorem implies

$$\langle \sqrt{n}(\hat{m} - m), f\rangle = h(\sqrt{(\hat{m} - m)}) \to h(\xi) - \langle \xi, f\rangle$$

$$\stackrel{d}{=} N(0, E\langle \xi, f\rangle^2) = N(0, \langle \mathcal{K}f, f\rangle)$$

by last theorem : Pointwise convergence is weaker than the uniform convergence.

7.3.1 Inference for the mean

Consider a one-sample scenatio. Suppose we observe $X_1, \ldots, X_n \sim_{IID} X$ which are $\mathbb{H} = L^2(\mathcal{T})$ valued random elemtnts with mean element m and covariance $\mathcal{K}$. For a pre-specified function $m_0 \in \mathbb{H}$, we want to test

$$H_0 : E(X) = m_0 \quad H_1 : EX \neq m_0$$

Under H_0 the CLT in $\mathbb{H}$ gives

$$\sqrt{n}(\hat{m} - m) \stackrel{d}{\to} Z \sim N(0, \mathcal{K})$$

The norm-based test is based on the test statistic

$$T_{1n} = \|\sqrt{n}(\hat{m} - m_0)\|^2$$

Y_{1n} tends to be large under the H_1. Under H_0, since $\|\cdot\|^2$ is a continuous function, by the continuous mapping theorem,

$$T_{1n} \stackrel{d}{\to} \|z\|^2$$

Expand z on the eigen basis of X,

$$z = \sum_{k=1}^{\infty} \xi_k \phi_k$$

where $\xi_k = \langle a, \phi_k\rangle \sim N(0, \lambda_k)$. Now

$$\|z\|^2 = \|\sum_{k=1}^{\infty} \xi_k \phi_k\|^2 = \sum_{k=1}^{\infty} \xi_k^2 - \sum_{k=1}^{\infty} \lambda_k (\xi_k/\sqrt{\lambda_k})^2 \stackrel{d}{=} \sum_{k=1}^{\infty} \lambda_k W_k$$

where $W_k \sim_{IID} \chi_1^2$. The null distribution involves unknown λ_k. which are estimated by the sample eigen values of $\hat{\lambda}_k$. To obtain α- level critical value, use the $(1-\alpha)$ quantile of $\sum_{k=1}^{\infty} \hat{\lambda}_k W_k$, a weighted sum of X_1^2 random variables. In practivce, one can apply a finite truncation and monte carlo to obtain the quantiles easily.

The projection-based test statisitc is , for a chosen $K < \infty$ (in practice often chosen using FVE).

$$T_{2n} = \sum_{k=1}^{K} n(\frac{\langle \hat{m} - m_0, \hat{\phi}_k\rangle}{\hat{\lambda}_k^2})^2$$

Under H_0, by the consistency of $(\hat{\lambda}_k, \hat{\phi}_k), k = 1, \ldots, K$, and the continuous mapping theorem,

$$T_{2n} \stackrel{d}{\to} \sum_{k=1}^{K} \frac{\langle z, \phi_k\rangle}{\lambda_k} \stackrel{d}{=} \chi_k^2$$

since $\langle z, \phi_k\rangle/\sqrt{\lambda_k} \sim_{IID} N(0,1)$.

Whether the norm-based or the projection based test is better depends on the structure of the deviation from the H_0, i.e. $EX - m_0$. If $\langle EZ - m + 0, \phi_k\rangle = 0, k = 1, \ldots, K$, then the projection based test statistic T_{2n} has the same distribution under H_1 and H_0, and thus cannot detect any signal. On the other hand, if $EX - m_0 = c\phi$ where c is a non-zero scalar constant, then the projection method with $K-1$ is more powerful than the norm-based test.

Application: Consider a scalar-based response FLM. Since

$$g = \mathcal{K}\beta$$

where $g = E[(Y - EY)(X - EX)]$. If $Ker(\mathcal{K}) = \{0\}$, then testing the regression effect exists

$$H_0 : \beta = 0 \iff H_0 : g = 0$$

Under H_0, by the LLN and CLT for $(Y - EY)(X - EX)$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\overline{Y})(X_i-\overline{Y})\xrightarrow{d}N(0,E[(Y_i-EY)(X_i-EX)\otimes(Y_i-EY)(X_i-EX)])$$

the covariance part can be estimated by

$$\frac{1}{n}\sum_{i}(Y_i-\overline{Y}_i)(X_i(t)-\overline{X}(t))(Y_i-\overline{Y}_i)(X_i(s)-\overline{X}(s))$$

this CLT is not joint convergence. We are not able to construct confidence band for X or Y.

7.3.2 Functional One-way ANOVA

c.f. Jin-Ting Zhang 2007. Chap 5

Suppose we observe functions $Y_{ij}(t), t\in\mathcal{T}$ in k independent examples. The data satisfy

$$Y_{ij}(t)=\mu_i(t)+V_{ij}(t),\quad t\in\mathcal{T},\quad i=1,\ldots,K,\quad j=1,\ldots,n_i$$

where μ_i are unknown group mean functions, and the V_{ij} are independent subkect-specific effect that has mean zero and covariance operator $\mathcal{G}$ (homosked. assumption). View the functions as elements in $\mathbb{H}=L^2(\mathcal{T})$, we want to test

$$H_0:\mu_1=\mu_2=\ldots,\quad H_1:notH_0$$

The estimates are

$$\hat{\mu}_i=\overline{Y}_{i}=\frac{1}{n_i}\sum_{j}Y_{ij}$$

$$\hat{\mathcal{G}}=\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\cdot})\otimes(Y_{ij}-\overline{Y}_{i\cdot})$$

$$\hat{\mathcal{G}}(t,s)=\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}(t)-\overline{Y}_{i}(t))\otimes(Y_{ij}(s)-\overline{Y}_{i}(s))$$

By the CLT, we have

$$\sqrt{n_i}(\overline{Y}_{i\cdot}-\mu_i)\xrightarrow{d}z_i$$

where z_i is a zero mean Gaussian Random element with covariance $\mathcal{G}$ and the z_i are independent bevause we have independent samples.

If the Y_{ij} are themselves Gaussian, then $\sqrt{n_i}(\overline{Y}_{i\cdot}-\mu_i)$ follows the same Gaussian distribution as z_i does, for any finite n_i, let

$$SSR=\sum_{i=1}^{k}n_i\|\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}\|^2$$

one degree of freedom is lost due to the grand average here

$$SSE=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\|\bar{Y}_{i\cdot}-\bar{Y}_{j\cdot}\|^2$$

where

$$\bar{Y}_{\cdot\cdot}=\frac{1}{n}\sum_{i}\sum_{j}Y_{ij}$$

Theorem: Under H_0,

$$SSR \xrightarrow{d} \sum_{r=1}^{\infty} \lambda_r A_r, \quad A_r \sim_{IID} \xi^2_{k-1}$$

as $n_i \to \infty, i = 1, \ldots, k.$ where λ_r are the eigen values of $\mathcal{G}$.
If further the Y_{ij} are Gaussian, then

$$SSR \stackrel{d}{=} \sum_{r=1}^{\infty} \lambda_r A_r$$

$$SSE \stackrel{d}{=} \sum_{r=1}^{\infty} \lambda_r E_r$$

One can test H_0 by comparing the observed SSR with quantiles of $\sum_{r=1}^{\infty} \hat{\lambda}_r A_r$, where the $\hat{\lambda}_r$ are the estimated eigenvalues of $\hat{\mathcal{G}}$.

If Gaussianity is assumes, one can apply an F-test, by comparing the F-type statistic

$$f_n = \frac{SSR/(k-1)}{SSE/(n-k)}$$

with quantiles of $\frac{\sum_{r=1}^{\infty} \lambda_r A_r/(k-1)}{\sum_{r=1}^{\infty} \lambda_r E_r/(n-k)}$. Again quantiles are obtained from Monte Carlo. Zhang (2007) proposes an F-dist approximation.

7.3.3 Confidence Bands

Let $X_1(t), \ldots, X_n(t), t \in \mathcal{T}$ be IID realizations of a stochastic process $X(t), t \in \mathcal{T}$. Let μ and G denote the mean and covariance function, $\hat{\mu}, \hat{G}$ be the empirical estimates. Let $\sigma(t) = \sqrt{G(t,t)}, \hat{\sigma}(t) = \sqrt{\hat{G}(t,t)}$ By the scalar LLN and CLT, for $t \in \mathcal{T}$

$$P(\sqrt{n}|\frac{\hat{\mu}(t) - \mu(t)}{\sqrt{\hat{\sigma}(t)}}| \leq z_{1-\alpha/2}) \approx 1 - \alpha$$

A $(1-\alpha)100\%$ pointwise CI for μ is

$$\hat{\mu}(t) \pm \frac{1}{\sqrt{n} z_{1-\alpha/2} \hat{\sigma}(t)}, t \in \mathcal{T}$$

This does not account for multiplicity.
To obtain a simultaneous confidence band (SCB), we would like to find threshold C_α such that

$$P(\sup_{t \in \mathcal{T}} \sqrt{n}|\frac{\hat{\mu}(t) - \mu(t)}{\sqrt{\hat{\sigma}(t)}}| \leq C_\alpha) \approx 1 - \alpha$$

Some issues need to be taken care of:

1. sup need to be replaced by esssup for $f \in L^2(\mathcal{T})$

$$esssup(f) = \inf_a \{m(f(t) > a) = 0\}$$

where m is Lebesgue measure on $\mathcal{T}$.

2. In Hilbert space, sup is not a continuous operation wrt L^2 norm. Let $g : L^2(\mathcal{T}) \to \mathbb{R}$

$$g(f) := esssup_{t \in \mathcal{T}} |f(t)|$$

then g is not a continuous function wrt the L^2- norm.

Choi & Reimberr (2017) constructed hyper-elliptical confidence region for $\mu \in L^2(\mathcal{T})$ of the form

$$E_{\hat{\mu}} = \{h \in L^2(\mathcal{T}) : \sum_{j=1}^{\infty} \langle h - \hat{\mu}, \phi_j \rangle^2 / r_j^2 \leq 1\}$$

where r_j is radius.

Target to find the r_j such that

$$P(\mu \in E_{\hat{\mu}}) = P(\sum_{j=1}^{\infty} \langle h - \hat{\mu}, \phi_j \rangle^2 / r_j^2 \leq 1) \to 1 - \alpha$$

They propose to consider $r_j^2 = n^{-1}\lambda_j^{1/2} C$ to achieve the narrowest confidence band width, where C is roughly the quantile.

By the CLT in $L^2(\mathcal{T})$ and continuous mapping

$$\sum_{j=1}^{\infty} \frac{\langle h - \hat{\mu}, \phi_j \rangle^2}{n^{-1}\lambda_j^{1/2}} \to \sum_{j=1}^{\infty} \lambda_j^{1/2} W_j =: W$$

where $W_j \sim^{IID} \chi_1^2$. Therefore, setting $C = C_\alpha$, the α-upper quantile of W, we have

$$P(\sum_{j=1}^{\infty} \frac{\langle h - \hat{\mu}, \phi_j \rangle^2}{n^{-1}\lambda_j^{1/2}} \leq C_\alpha) \to 1 - \alpha$$

The SCB based on a Scheffe's type projection is

$$\hat{\mu}(t) \pm \frac{1}{\sqrt{n}} \sqrt{C_\alpha \sum_{j=1}^{\infty} \lambda_j^{1/2} \phi_j^2(t)}, \quad t \in \mathcal{T}$$

A way out of the tedium working within a Hilbert space is to assume smoothness in the sample paths. Work instead in the Banach space of continuous functions

$$C(\mathcal{T}) = \{f : \mathcal{T} \to \mathbb{R}, f cont.\}$$

which is equipped with the supreme norm

$$\|f\|_\infty = \sup_{t \in \mathcal{T}} |f(t)|$$

If the X_i are Lipschitz with integrable Lipschitz constant (see Jain-Marcus 1982) then

$$\frac{\sqrt{n}(\hat{\mu}(t) - \mu(t))}{\hat{\sigma}(t)} \to^D Z =^D N(0, \rho) in C(\mathcal{T})$$

where ρ is the correlation function

$$P(\sup_{t \in \mathcal{T}} \frac{\sqrt{n}(\hat{\mu}(t) - \mu(t))}{\hat{\sigma}(t)} > C_\alpha) \to P(\|Z\|_\infty > C_\alpha)$$

as $n \to \infty$. Set C_α be such that $P(\|Z\|_\infty > C_\alpha) = \alpha$. To estimate C_α, simulate Gaussian processses $\tilde{Z}$ with mean 0 and covariance function

$$\hat{\rho}(t, s) = \hat{G}(t, s)/(\hat{\sigma}(t)\hat{\sigma}(s))$$

on a dense grid, and find the $\alpha-$ upper quantile of $\|\tilde{Z}\|_\infty$. A SCB for μ is

$$\hat{\mu}(t) \pm \frac{1}{\sqrt{n}} \hat{C}_\alpha \hat{\sigma}(t)$$

Degras (2011) and Cao Yang Todem (2011) considered this type of SCB with dense and noisy observations. For the sparse case, SCB is considered by Zheng Yang Hardle (2014).

8 Functional Classification

The classification setting: observe iid realizations $(X_i, Y_i) \sim (X, Y), i = 1, \ldots, n$ where X_i is the predictor, $Y_i \in \{0, 1\}$ is the response class.

Goal: predict Y from X, minimizing misclassification error.

1-versus-1 classifier: compair all the pairs to decide the overal winner.

1-versus-n classifier: get the credible scores

Let $X^{(k)} = X|Y = k$ be the conditional distribution of X given $Y = k$,

$$\pi_k = P(Y = k)$$

be the prior probability / mixture probability for observing group k, $k \in \{0, 1\}$.

8.1 Quadratic and linear discriminant analysis with multivariate predictor

Predictor X is a $d-$ dimentional random vector. Suppose

$$X^{(0)} \sim N(\mu_0, \Sigma_0), \quad X^{(1)} \sim N(\mu_1, \Sigma_1)$$

The classifier that minimizes misclassification error

$$P(\hat{Y} \neq Y) = E(\hat{Y} - Y)^2$$

Note $\hat{Y} \in \{0, 1\}$ being the predicted class label is given by the Bayes classifier which assigns a new observation with $X = x$ to group 1 iff

$$P(Y = 1|X = x) \geq P(Y = 0|X = x)$$

Move terms and apply Gassianity, assign X to group 1 iff

$$0 \leq \log \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = \log \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x|Y = 0X^{(0)}P(Y = 0)}$$

$$= -\frac{1}{2}\{(x - \mu_1)^T\Sigma_1^{-1}(x - \mu_1) - (x - \mu_0)^T\Sigma_1^{-1}(x - \mu_0)\} + \log \frac{|\Sigma_1|}{|\Sigma_0|} + \log \frac{\pi_1}{\pi_0} =: f_{QDA}(x)(1)$$

Here $f_{QDA}(x)$ is the criterion function for the quadratic discriminant analysis (QDA). The QDA classifier is

$$1\{f_{GDA}(x) \geq 0\}$$

If further assume $\Sigma_0 = \Sigma_1$ then (1) is simplified as

$$X^T\Sigma^{-1}(\mu_1 - \mu_0) - \frac{1}{2}\{\mu_1^T\Sigma^{-1}\mu_1 - (\mu_0)^T\Sigma^{-1}\mu_0\} + \log \frac{\pi_1}{\pi_0} =: f_{GDA}(x)(2)$$

We obtain the linear discriminant analysis (LDA) criterion function. Note we project X to a line. The definition of GDA and LDA applies to non-Gaussian data.

8.2 Functional Linear discrimination analysis

(Hall & Delaigle 2012, JRSSB)

Now the predictor X is a function to $L^2(\mathcal{T})$, the conditional distribution of X given Y satisfies

$$EX^{(0)}(t) = 0, \quad EX^{(1)}(t) = \mu(t), \quad cov(X^{(0)}(t), X^{(0)}(s)) = cov(X^{(1)}(t), X^{(1)}(s)) - G(t, s)$$

Here for simplicity assume that $X^{(0)}$ has mean 0. Let $\mathcal{G}$ be the covariance operator with eigen decomposition

$$\mathcal{G} = \sum_{j=1}^{\infty} \lambda_j \phi_j \otimes \phi_j$$

Note the inversion of Σ can raise problem, the functional LDA mimics (2) and uses a truncated version

$$\mathcal{G}_J = \sum_{j=1}^{J} \lambda_j \phi_j \otimes \phi_j$$

when performing the inverse. Let $\mu_j = \langle \mu, \phi_j \rangle$ and

$$\Phi_j = (\sum_{j=1}^{J} \lambda_j \phi_j \otimes \phi_j)^{-1}(\mu) = \sum_{j=1}^{J} \langle \phi_j, \mu \rangle \phi_j = \sum_{j=1}^{J} \lambda_j^{-1} \mu_j \phi_j$$

be the prejection direction.

Also let $\xi_j = \langle X, \phi_j \rangle$ and $x_j = \langle x, \phi_j \rangle$ be the prohection of a random and a fixed function respectively. The criterion function

$$f_{FLDA} = \langle x, \psi^J \rangle - \frac{1}{2}\langle \psi^J, \mu \rangle + \log(\frac{\pi_1}{\pi_0}) = \sum_{j=1}^{J} \lambda_j^{-1} \mu_j^2 \mu_j x_j - \frac{1}{2}\lambda_j^{-1} + \log(\frac{\pi_1}{\pi_0})$$

Given a new x, the classifier predicts

$$\hat{Y} = 1[f_{FLDA}(x) \geq 0]$$

Theorem: Assume the conditional distribution $X^{(0)}, X^{(1)}$ are both Gaussian, and $\pi_0 = \pi_1 = \frac{1}{2}$. Then the probability of misclassification using the FLDA classidier is

$$err = P(\hat{Y} \neq Y) = 1 - \Phi(\frac{1}{2}\sqrt{\sum_{j=1}^{J} \lambda_j^{-1}\mu_j^2})$$

Proof. Under $\pi_0 = \pi_1$,

$$f_{FLDA}(X) = \sum_{j=1}^{J} \lambda_j^{-1}\mu_j\xi_j - \frac{1}{2}\sum_{j=1}^{J} \lambda_j^{-1}\mu_j^2$$

Let z_j be IID N(0,1), random variables, $j = 1, \ldots J$. Then

$$\begin{aligned} f_{FLDA}(X^{(0)}) &= \sum_{j=1}^{J} \lambda_j^{-1/2}\mu_j z_j = \frac{1}{2}\sum_{j=1}^{J} \lambda_j^{-1}\mu_j^2 \\ &=_D N(-\frac{1}{2}\sum_{j=1}^{J} \lambda_j^{-1}\mu_j^2, \sum_{j=1}^{J} \lambda_j^{-1}\mu_j^2) \end{aligned}$$

and

$$f_{FLDA}(X^{(1)}) =_D N(\frac{1}{2}\sum_{j=1}^{J} \lambda_j^{-1}\mu_j^2, \sum_{j=1}^{J} \lambda_j^{-1}\mu_j^2)$$

Therefore

$$\begin{aligned} P(\hat{Y} \neq Y) &= \sum_{k=0}^{1} P(Y = k)P(\hat{Y} \neq k|Y = k) \\ &= \frac{1}{2}[P(f_{FLDA}(X) \geq 0|Y = 0) + P(f_{FLDA}(X) < 0|Y = 1)] \\ &= \frac{1}{2}[1 - \Phi(\frac{1}{2}\sqrt{\sum_{j=1}^{J} \lambda_j^{-1}\mu_j^2})] * 2 \\ &= 1 - \Phi(\frac{1}{2}\sqrt{\sum_{j=1}^{J} \lambda_j^{-1}\mu_j^2}) \quad \square \end{aligned}$$

1. If $\sum_{j=1}^{\infty} \lambda_j^{-1}\mu_j^2 < \infty$, then $\Psi^{(\infty)} = \lim_{J\to\infty} \Psi^J$ is well defined, and gives the smallest misclassification error if the $X^{(k)}$ are Gaussian.
2. If $\sum_{j=1}^{\infty} \lambda_j^{-1}\mu_j^2 = \infty$, then $err \to 0, J \to \infty$. BY using projection direction Ψ^J with a large J, it is possible to achieve 'near perfect classification'.
3. The second remark holds even if $X^{(k)}$ are non-Gaussian
4. In theory, a large J leads to a better performance, but in practice this is not necessaily true due to the errors in the estimation. J is usually selected by CV.

Estimation

Let $X^{(0)_i}$ be the i-th observation among $X_1, \ldots, X_n$ that have class label $Y = 0, i = 1, \ldots, n_0$ and similarly for $X^{(1)_i}$, $Y = 1, i = 1, \ldots, n_1$. We have $n = n_0 + n_1$.

Estimates are

$$\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n} \\
\widehat{EX_1^{(k)}} &= \frac{1}{n}\sum_{i=1}^{n_k} X_i^{(k)} \\
\hat{\mu} &= \widehat{EX_1^{(1)}} - \widehat{EX_1^{(0)}} \\
\hat{\mathcal{G}} &= \frac{1}{n}\sum_{k=0}^{1}\sum_{i=1}^{n_k}(X_i^{(k)} - \widehat{EX_1^{(k)}}) \otimes (X_i^{(k)} - \widehat{EX_1^{(k)}})
\end{aligned}$$

$(\hat{\lambda}_j, \hat{\phi}_j)$ are the eigen-value and eigen-fun pair of $\hat{\mathcal{G}}$.

Given a new function,

$$\hat{X}_j = \langle X, \hat{\phi}_j \rangle, \hat{\xi} = \langle X, \hat{\phi}_j \rangle$$

Estimate for $f_{FLDA}(x)$ is

$$\hat{f}_{FLDA}(x) = \sum_{j=1}^{J} \hat{\lambda}_j^{-1}\hat{\mu}_j - \frac{1}{2}\sum_{j=1}^{J}\hat{\lambda}_j^{-1}\hat{\mu}_j^2 + \log(\frac{\hat{\pi}_1}{\hat{\pi}_0})$$

8.3 Functional Quadratic Discriminant (Galeano Joseph Lillo, 2015)

Suppose

$$X^{(0)} \sim (\mu_0, \mathcal{G}_0), \quad X^{(1)} \sim (\mu_1, \mathcal{G}_1)$$

which are not necessarily Gaussian.

Write

$$\mathcal{G}_k = \sum_{j=1}^{\infty}\lambda_{jk}\phi_{jk} \otimes \phi_{jk}, \quad \mathcal{G}_{kJ} = \sum_{j=1}^{\infty}\lambda_{jk}\phi_k \otimes \phi_k$$

The squared Mahalanobis (semi) distance using the first J projections is w.r.t. $\mathcal{G}_{kJ}$,

$$d_J^k(x,y)^2 = \langle x - y, \mathcal{G}_{kJ}^{-1}(x-y)\rangle = \sum_{j=1}^{J}\lambda_{jk}^{-1}(x_{jk} - y_{jk})^2$$

where

$$X_{jk} = \langle X, \phi_{jk}\rangle, \quad X_{jk} = \langle X, \phi_{jk}\rangle$$

The functional GDA classifier is

$$1[f_{FQDA}(x) \geq 0]$$

$$f_{FQDA}(x)f = -\frac{1}{2}\left\{ d_J^1(x,\mu_1)^2 - d_J^0(x,\mu_0)^2 + \sum_{j=1}^{J}\log\frac{\lambda_{j1}}{\lambda_{j0}} + \log\frac{\pi_1}{\pi_0}\right\}$$

What about a general Bayes classifier through density ratio?

An issue for considering $\frac{f(X|Y=1)}{f(X|Y=0)}$ is that the densities of functional data X does not exist in general wrt a translation invariant measure. See Deleigle Hall (2010).

To see some ideas, consider $X \sim N(0, \mathcal{G})$. The projection $\xi_j = \langle X, \phi_j\rangle$ are independent,

$$f_J(x) := f(\xi_1, \ldots, \xi_J) =_{indep} \Pi_{i=1}^{J} f(\xi_j)$$

$$= \exp(-\frac{1}{2}\sum_{j=1}^{J}\log(2\pi\lambda_j) - \frac{1}{2}\sum_{j=1}^{J}\frac{\xi_j^2}{\lambda_j})$$

$$=_D \exp(-\frac{1}{2}\sum_{j=1}^{J}\log(2\pi\lambda_j) - \frac{1}{2}W)$$

where $W \sim \chi^2_J$. As $J \to \infty, f_J(x)$ diverges up with probability 1. So it is not possible to define

$$f(x) = \lim_{J\to\infty} f_J(x)$$

8.4 Functional Bayes Classifier (Dai Yao Muller, 2017)

Suppose

$$X^{(0)} \sim (\mu_0, \mathcal{G}_0), \quad X^{(1)} \sim (\mu_1, \mathcal{G}_1)$$

Consider the Bayes classifier through the density quotient of the projections

$$Q(x) = \frac{P(Y=1|X=x)}{P(Y=0|X=x)} \approx \frac{P(Y=1|\xi_1 = x_1, \ldots, x_J = x_J)}{P(Y=0|\xi_1 = x_1, \ldots, x_J = x_J)}$$

$$= \frac{\pi_1 f_1(x_1, \ldots, x_J)}{\pi_0 f_0(x_1, \ldots, x_J)} \quad (1)$$

where f_k is the conditional density of $\xi_1, \ldots, \xi_J$ under group k.

$$\xi_j = \langle X, \psi_j \rangle, x_j = \langle x, \psi_j \rangle, \{\psi_j\}_{j=1}^{\infty} \text{CONS}$$

EStimation of f_k involves J-dimensional density estimation and is subject to the curse of dimensionality. And thus the joint pdfs cannot be practically estimated. Apply two simplifying assumptions

- (A1) Common eigenfunctions for $X^{(k)}$

$$\mathcal{G}_k(t,s) = \sum_{j=1}^{\infty} \phi_j(t)\phi_j(s)$$

  where here the set of eigenfunctions are the same. Flurry (1988) Boente et al (2010)

- (A2) Independenve. The projections $\langle X^{(k)}, \phi_j \rangle, j = 1, 2, \ldots$ are independent under group $k = 0, 1$.

Under (A1), $\langle X^{(k)}, \phi_j \rangle$ are always uncorrelated. If $X^{(k)}$ is Gaussian, then (A2) will hold.

If both (A1) and (A2) hold, and we set $\psi_j = \phi_j$, then

$$(1) = \frac{\pi_1}{\pi_0}\Pi_{i=1}^{n}\frac{f_{j1}(x_j)}{f_{j0}}(x_j)$$

where f_{jk} is the univariate density of $\langle X^{(k)}, \phi_j \rangle$ which is easy to estimate. Upon taking log, obtain

$$Q_J(X) = \log(\frac{\pi_1}{\pi_0}) + \sum_{j=1}^{J}\log\frac{f_{i1}(x_j)}{f_{i0}(x_j)}$$

8.5 Functional Nonparametric Classifier

Let $p(x) = P(Y=1|X=x) = E(Y|X=x)$. The functional kNN classifier is

$$\hat{p}(x) = \frac{1}{k_n}\sum_{i=1}^{k_n} Y_{(i,n)}(x)$$

where $(X_{(1,n)}, Y_{(1,n)}(x), \ldots, X_{(n,n)}, Y_{(n,n)}(x))$ denotes a reordering of data according to increasing $\|X_i - x\|$. The rate of convergence of $\hat{p}(x)$ to $p(x)$ is of order $(\log(n))^{-C}, C > 0$ is a constant.

9 Presentations

- K-means clustering
- Outlier detection
- ZHVI data https://www.zillow.com/research/data/
- Functional Additive Mixed Model (FAMM)
- Multi-dimensional scaling
- SNE and t-SNE
- UMAP
- Alternating Lesat Squares
- grplasso
- Dynamic FPCA
  - spectral density operator