

Reproducing Kernel Hilbert Spaces and
Nonparametric Estimation

**Belkacem Abdous, Alain Berlinet
and Nicolas Hengartner**

**SEMINAIRE EUROPEEN - EUROPEAN
SEMINAR**

Paris 6

6 March, 2006

Plan

- Introduction & Motivation
- Representation of the estimator
- Asymptotic behavior : a.s. convergence and asymptotic normality
- Applications

Introduction & Motivation

- Let X_1, \dots, X_n be independent observations from an unknown d.f. F and denote by $\hat{F}_n(x)$ the empirical distribution

$$\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^n \mathbf{I}_{(X_j \leq x)}.$$

- Let $\Phi(x, F)$ be a functional to be estimated from X_1, \dots, X_n . A typical example is the density

$$\Phi(x, F) = F'(x).$$

Assume that $\Phi(x, F)$ is $(r + 1)$ times continuously differentiable at x .

- Then $\Phi(y, F)$ can be locally approximated by a polynomial of order r . For some $\xi \in (x, y)$, one has

$$\begin{aligned}\Phi(y, F) &= \sum_{i=0}^r \frac{(y-x)^i}{i!} \Phi^{(i)}(x, F) \\ &\quad + \frac{(y-x)^{r+1}}{(r+1)!} \Phi^{(r+1)}(\xi, F).\end{aligned}$$

- The d.f. F being unknown, a pilot estimator of $\Phi(x, F)$ is provided by

$$\Phi_n(x) = \Phi(x, \hat{F}_n).$$

- Next, assume that $\Phi(z, F)$ is Hadamard differentiable in F , so that $\Phi(x, \widehat{F}_n)$ can be linearized as follows

$$\begin{aligned} \Phi(x, \widehat{F}_n) - \Phi(x, F) &= \Psi(x, F)(\widehat{F}_n - F) \\ &\quad + o\left(\left\|\widehat{F}_n - F\right\|_{\infty}\right) \end{aligned}$$

- Hence, the process $\sqrt{n}(\Phi(x, \widehat{F}_n) - \Phi(x, F))$ is approximated by $\sqrt{n}(\widehat{F}_n(x) - F(x)) \cdot \Psi(x, F)$ which is known to converge to a Brownian Bridge.

- However $\Phi(\cdot, \hat{F}_n)$ is a rough estimator and some smoothing is needed. To estimate $\Phi(x, F)$ and its r first derivatives we will minimize

$$J(a_0, \dots, a_r) =$$

$$\int K\left(\frac{x-z}{h}\right) \left\{ \Phi_n(z) - \sum_{k=0}^r a_k (x-z)^k \right\}^2 dz,$$

- h (the smoothing parameter) controls the size of the neighborhood of x , K denotes a weight function (or kernel).

This general setting produces good estimators in a broad variety of statistical problems.

A nice representation of these estimators will easily provide asymptotic results.

A representation of the estimator

- Assume that K is a probability density supported on the interval $[-1, 1]$
- Let $(\hat{a}_0, \dots, \hat{a}_r)$ be the minimizer of $J(a_0, \dots, a_r)$
- Let \mathcal{P}_r be the space of polynomials of degree less than or equal to r and let $Q_0(z), Q_1(z), \dots, Q_r(z)$ be an orthonormal basis of \mathcal{P}_r considered as a subspace of $L_2(K \lambda)$.

Let

$$K^{[m,r]}(u) = \sum_{k=0}^r Q_k(u) \left. \frac{d^m}{dw^m} Q_k(w) \right|_{w=0} K(u).$$

- Then one can show that

$$\hat{a}_m = \frac{1}{m!h^m} \int_{-\infty}^{\infty} \Phi_n(z) \frac{1}{h} K^{[m,r]} \left(\frac{z-x}{h} \right) dz.$$

- Indeed, rewrite

$$J = \int_{-1}^1 \left\{ \Phi_n(x + uh) - \sum_{k=0}^r a_k h^k u^k \right\}^2 K(u) du.$$

- The polynomial $\hat{p}_r(u) = \sum_{k=0}^r \hat{a}_k h^k u^k$ that minimizes J is the $L_2(K\lambda)$ projection of $\Phi_n(x + hu)$ onto \mathcal{P}_r .

- $\mathcal{K}(u, v) = \sum_{k=0}^r Q_k(u)Q_k(v)$ being the reproducing kernel of \mathcal{P}_r we can write

$$\hat{p}_r(u) = \int_{-1}^1 \mathcal{K}(u, v) \Phi_n(x + hu) K(u) du,$$

- The coefficients \hat{a}_m are obtained by differentiating

$$\begin{aligned} m!h^m \hat{a}_m &= \left. \frac{d^m}{du^m} \hat{p}_r(u) \right|_{u=0} \\ &= \int_{-1}^1 \sum_{k=0}^r \left[\left. \frac{d^m}{du^m} Q_k(u) \right|_{u=0} \right] Q_k(v) \hat{\Phi}_n(x + hv) K(v) dv. \end{aligned}$$

- The conclusion follows by setting

$$K^{[m,r]}(u) = \sum_{k=0}^r \left[\frac{d^m}{dv^m} Q_k(v) \Big|_{v=0} \right] Q_k(u) K(u),$$

and operating another change of variables.

Asymptotic behavior: a.s. convergence

- Suppose that the first m derivatives of $\Phi(x, F)$ exist and they belong to $L^p(\mathbb{R})$ for some $1 \leq p \leq \infty$.
- Assume that the point x is a Lebesgue point of $\Phi^{(m)}(x, F)$
- Assume that there exists an open neighborhood \mathcal{N}_x of x such that for any sequence of distribution functions $\{F_n\}_{n=1}^{\infty}$ converging to F in the sup-norm,

$$\lim_{n \rightarrow \infty} \sup_{y \in \mathcal{N}_x} |\Phi(y, F_n) - \Phi(y, F)| = 0$$

almost surely.

- Then $\widehat{\theta}_{n,h}^{(m)}(x)$ converges almost surely to $\Phi^{(m)}(x, F)$ whenever the sequence of estimated distribution functions $\{F_n\}_{n=1}^{\infty}$ converges almost surely to F and the bandwidth converges to zero slowly enough.
- This convergence result applies to independent as well as dependent data.
- The assumption that $\Phi^{(m)}(x, F)$ belongs to $L^p(\mathbb{R})$ for some $1 \leq p \leq \infty$ can be replaced with a continuity assumption on $\Phi^{(m)}(x, F)$.

Asymptotic normality

- In general, the statistical functionals $\Phi^{(m)}(x, F)$ are typically not Hadamard differentiable whereas the base functional $\Phi(x, F)$ is.
- It follows that the projection of $\Phi(\cdot, F)$ onto the space of polynomials of degree r in $L_2(K\lambda)$ for fixed bandwidth $h > 0$, is also Hadamard differentiable.

- This implies that the coefficients of the resulting best fitting polynomial,

$$\bar{a}_m = \frac{1}{m!h^m} \int \Phi(x + hv, F) K^{[m,r]}(v) dv,$$

$$m = 0, \dots, r$$

viewed as a functional of F , are Hadamard differentiable as well.

- Hence there exists a function $\Delta_h(t)$ with

$$\int \Delta_h(t) F(dt) = 0$$
 for which

$$\hat{a}_m = \bar{a}_m + \int \Delta_h(t) (\hat{F}_n - F)(dt) + r_h(\hat{F}_n, F),$$

in which the remainder $r_h(\hat{F}_n, F) = o(\|\hat{F}_n - F\|)$.

- For fixed bandwidth h , under standard regularity conditions on $K^{[m,r]}$ together with the implicit assumption that

$$\sqrt{n} \|\hat{F}_n - F\|_\infty = O_p(1)$$

one concludes that the remainder term $r_h(\hat{F}_n, F)$ converges in probability to 0.

- Thus by virtue of Slutski's Lemma,

$$\left[\hat{a}_m - \bar{\hat{a}}_m \right] / \sqrt{V ar(\hat{a}_m)}$$

has the same limit as

$$\frac{\int \Delta_h(t)(\hat{F}_n - F)(dt)}{\sqrt{V ar(\hat{a}_m)}} = \frac{\sum_{i=1}^n \Delta_h(X_i)}{\sqrt{V ar(\hat{a}_m)}}.$$

- Asymptotic normality follows from standard arguments provided that $E[\Delta_h^2(X)] < \infty$.

Applications

Hazard Functions

Selection biased models

Reliability and econometric functions

Spectral Density

and many others ...

Hazard functions

Right-censored data An item is removed from the study before the end of its *natural life*. (e.g. A lightbulb is accidentally broken before it burns).

Failure rate Approximate probability of failure in the time interval $[x, x + dx]$, given that the subject has survived to time x .

Let X_1, \dots, X_n denote the uncensored lifetimes ($\sim F_0$) and Y_1, \dots, Y_n denote the censoring variables ($\sim H$) which are independent of the X_i 's. Rather than observe the X_i 's, we observe

$$Z_j = (X_j, Y_j), \quad \text{and} \quad \Delta_j = \mathbb{I}_{\{X_j \leq Y_j\}}$$

The Z 's are an i.i.d. sample from a distribution F , where $1 - F(x) = (1 - F_0(x))(1 - H(x))$. An interesting class of target functions are

$$\eta(x) = \frac{(1 - H(x))f_0(x)}{Q(x)},$$

for some positive function $Q(x)$. In particular, with $Q(x) = 1 - F(x)$, $\eta(x)$ is the hazard function while

taking $Q(x) = (1 - H(x))$ leads to $\eta(x) = f_0(x)$, the density of the X 's. Patil, Wells and Marron (1994)^a suggest estimating the cumulative target function

$$\Phi(x, F^0) = \int_0^x \eta(u) du = \int_0^x \frac{dF^-(u)}{Q(u)},$$

where $F^-(x) = P[Z < x, \Delta = 1]$, by its empirical version

$$\hat{\Phi}_n(x) = \int_0^x \frac{d\hat{F}_n^-(u)}{Q_n(u)}.$$

Of interest is the estimation of $\eta(x) = \frac{d}{dx} \Phi(x, F^0)$.

^aPatil, Wells and Marron (1994). Some Heuristics of kernel based estimators of ratio of functions. *Nonparametric Statistics*, 4, 203-209.

Selection biased models

- *Weighted distributions or selection biased models*
data arise in many fields, e.g., missing data, survey sampling, damaged observations, sociological studies, reliability theory, economics, see Patil et al. (1988) ^a
- **Example:** (*Length biased data*) the probability of retention of a random draw X from a density f is proportional to the value of X .

^a PATIL, G.P., RAO, C.R. AND ZELEN, M. (1988). Weighted distribution. *In: Encyclopedia of Statistical Sciences*, Vol. 9 (S. Kotz and N.L. Johnson, Eds.), 565-571. Wiley, New York.

- Let Y be a nonnegative r.v. with d.f. F and probability density f . Suppose that we do not observe Y but rather an another random variable X with distribution function G and density function g related to f as follows

$$g(x) = w(x)f(x)/\mu_w, \quad x > 0,$$

where $w(x) \geq 0$ is known, and,

$$\mu_w \equiv \int_0^\infty w(x)f(x)dx < \infty.$$

- Given X_1, \dots, X_n a random sample from G , we want to estimate f . The empirical estimate of $F(x)$ is

given by

$$\hat{F}_n(x) = \frac{\hat{\mu}_w}{n} \sum_{j=1}^n \frac{1}{w(X_j)} \mathbb{I}\{X_j \leq x\},$$

where $\hat{\mu}_w = [n^{-1} \sum_{j=1}^n 1/w(X_j)]^{-1}$.

- To estimate the density $f(x)$, set $\Phi(x, F) = F(x)$, $\Phi_n(x) = \hat{F}_n(x)$, and focus on estimating the first derivative of $\Phi(x, F)$.

Reliability and econometric functions

- Let $X \geq 0$ be a r.v. ($\sim F$) with $\mu = E(X) < \infty$.
- There are various transforms of F which are of great importance in industrial reliability, biomedical science, life insurance, demography, econometric studies, etc. Some of these transforms are
- *Mean residual life function M , Lorenz curve L , Scaled total time on test function T (or total time of test transform).*

- *Mean residual life function M :*

$$M(x) = E(X - x | X > x)$$

- *Lorenz curve L :*

$$L(t) = \frac{1}{\mu} \int_0^t F^{-1}(s) ds$$

- *Scaled total time on test function T :*

$$T(t) = \frac{1}{\mu} \int_0^{F^{-1}(t)} (1 - F(x)) dx$$

- Motivations and more information about these functions can be found in Shorack and Wellner (1986, page 775).^a
- Assume that X_1, \dots, X_n is a sample drawn from F . Natural estimates $\Phi_n(x) = M_n(\cdot)$, $L_n(\cdot)$ and $T_n(\cdot)$ of $\Phi(x, F) = M(\cdot)$, $L(\cdot)$ and $T(\cdot)$ respectively are obtained by replacing, F and F^{-1} by their empirical estimates \hat{F}_n and \hat{F}_n^{-1}

^aSHORACK, G.R., AND WELLNER, J.A. (1986). *Empirical Processes with Applications to Statistics*. John Wiley, New York.

Spectral density

- Assume that $\{X_t\}$ is a zero mean univariate stationary time series with autocovariances $\gamma_k = E(X_t X_{t+k})$. Then the spectral density of $\{X_t\}$ is

$$\Phi(\omega) = \frac{1}{2\pi} \left\{ \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \cos(k\omega) \right\}, \quad \omega \in [-\pi, \pi].$$

- Given a sample X_1, \dots, X_n , an estimate of the spectral density at frequency ω is the so called

periodogram

$$\Phi_n(\omega) = \frac{1}{2\pi} \sum_{k=-(n-1)}^{n-1} \tilde{\gamma}_{nk} e^{-i\omega k},$$

where $\tilde{\gamma}_{nk}$ is a sample estimate of γ_k .

- For a smoothing technique of spectral densities by local polynomials, see Daniels (1962).^a

^a DANIELS, H.E. (1962). The estimation of spectral densities. *J. Roy. Statist. Soc. Ser. B*, 24, 185-198.

Item response Theory

- Let $\mathbf{X} = (X_1, \dots, X_J)$ the dichotomous item response variables for J items (scored 0 and 1 for negative/incorrect and positive/correct answers respectively)
- Assume that the basic and classical assumptions for IRT models hold, i.e. (i) unidimensionality, (ii) monotonicity, and (iii) local independence
- Thus, there exists a single scalar latent trait θ which explains responses of an individual to items.

- The Item Response Function (IRF) or Item Characteristic Curve (ICC) $p_i(\theta) = P[X_i = 1|\theta]$ is an increasing function of θ .
- We have

$$P(X_1 = x_1, \dots, X_J = x_J | \theta) = \prod_{i=1}^J p_i(\theta)^{x_i} [1 - p_i(\theta)]^{(1-x_i)}$$

- We have a multivariate regression problem

$$\begin{aligned}\mathbf{E}(\mathbf{X}|\theta) &= (p_1(\theta), \dots, p_J(\theta)) \\ &= \left(\frac{\exp(g_1(\theta))}{1 + \exp(g_1(\theta))}, \dots, \frac{\exp(g_J(\theta))}{1 + \exp(g_J(\theta))} \right)\end{aligned}$$

where $g_i(\theta) = \text{logit}(\theta) = \log \frac{p_i \theta}{1 - p_i(\theta)}$.

- Suppose that we have a sample of N persons with latent traits $(\theta_1, \dots, \theta_N)$ and responses to the J items

$$\mathbf{X}_1 = (X_{11}, \dots, X_{1J}), \dots, \mathbf{X}_N = (X_{N1}, \dots, X_{NJ})$$

- Approximation p_i 's scale: Fix θ_{0j} and write
(*regularity conditions*)

$$p_i(\theta) \approx p_i(\theta_0) + \sum_{k=1}^r \frac{p_i^{(k)}(\theta_0)}{k!} (\theta - \theta_0)^k = \sum_{k=0}^r b_{ik} (\theta - \theta_0)^k$$

- similarly, we can use $g_i(\cdot)$ scale