Empirical Process Theory for Statistics

Jon A. Wellner

University of Washington, Seattle

Talk to be given at
School of Statistics and Management Science
Shanghai University of Finance and Economics
Shanghai, China

23 June 2015

Talk, Shanghai;
School of Statistics and Management Science

- Lecture Outline:
  - ▷ 1. Introduction, history, selected examples.
  - ▷ 2. Some basic inequalities and Glivenko-Cantelli theorems.
  - ▷ 3. Using the Glivenko-Cantelli theorems: first applications.
  - ▷ 4. Donsker theorems and some inequalities.
  - ▷ 5. Peeling methods and rates of convergence.
  - ▷ 6. Some useful preservation theorems.

Based on Courses given at Torgnon, Cortona, and Delft (2003-2005). Notes available at:

http://www.stat.washington.edu/jaw/
RESEARCH/TALKS/talks.html

Part I: Introduction, history, selected examples

- 1. Classical empirical processes
- 2. Modern empirical processes
- 3. Some examples

1. Classical empirical processes. Suppose that:

- $X_1, \ldots, X_n$ are i.i.d. with d.f. F on $\mathbb{R}$.
- $\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^n 1_{[X_i \le x]}$, the empirical distribution function.
- $\{\mathbb{Z}_n(x) \equiv \sqrt{n}(\mathbb{F}_n(x) - F(x)) : \ x \in \mathbb{R}\}$, the empirical process.

Two classical theorems:

Theorem 1. (Glivenko-Cantelli, 1933).

$$\|\mathbb{F}_n - F\|_\infty \equiv \sup_{-\infty < x < \infty} |\mathbb{F}_n(x) - F(x)| \to_{a.s.} 0.$$

Theorem 2. (Donsker, 1952).

$$\mathbb{Z}_n \Rightarrow \mathbb{Z} \equiv \mathbb{U}(F) \quad \text{in} \quad D(\mathbb{R}, \|\cdot\|_\infty)$$

where $\mathbb{U}$ is a standard Brownian bridge process on $[0,1]$; i.e. $\mathbb{U}$ is a zero-mean Gaussian process with covariance

$$E(\mathbb{U}(s)\mathbb{U}(t)) = s \wedge t - st, \quad s,t \in [0,1].$$

This means that we have

$$Eg(\mathbb{Z}_n) \to Eg(\mathbb{Z})$$

for any bounded, continuous function $g : D(\mathbb{R}, \| \cdot \|_\infty) \to \mathbb{R}$ and

$$g(\mathbb{Z}_n) \to_d g(\mathbb{Z})$$

for any continuous function $g : D(\mathbb{R}, \| \cdot \|_\infty) \to \mathbb{R}$ (ignoring measurability issues).

2. General empirical processes (indexed by functions)

Suppose that:

- $X_1, \ldots, X_n$ are i.i.d. with probability measure P on $(\mathcal{X}, \mathcal{A})$.
- $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, the empirical measure; here

$$\delta_x(A) = 1_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \in A^c \end{cases} \quad \text{for} \quad A \in \mathcal{A}.$$

  Hence we have

$$\mathbb{P}_n(A) = n^{-1} \sum_{i=1}^n 1_A(X_i), \quad \text{and} \quad \mathbb{P}_n(f) = n^{-1} \sum_{i=1}^n f(X_i).$$

- $\{\mathbb{G}_n(f) \equiv \sqrt{n}(\mathbb{P}_n(f) - P(f)) : \ f \in \mathcal{F} \subset L_2(P)\}$, the empirical process indexed by $\mathcal{F}$

Note that the classical case corresponds to:

- $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}, \mathcal{B})$.
- $\mathcal{F} = \{1_{(-\infty,t]}(\cdot) : \ t \in \mathbb{R}\}$.

Then

$$
\begin{aligned}
&\mathbb{P}_n(1_{(-\infty,t]}) = n^{-1} \sum_{i=1}^{n} 1_{(-\infty,t]}(X_i) = \mathbb{F}_n(t), \\
&P(1_{(-\infty,t]}) = F(t), \\
&\mathbb{G}_n(1_{(-\infty,t]}) = \sqrt{n}(\mathbb{P}_n - P)(1_{(-\infty,t]} = \sqrt{n}(\mathbb{F}_n(t) - F(t)) \\
&\mathbb{G}(1_{(-\infty,t]}) = \mathbb{U}(F(t))\,.
\end{aligned}
$$

Two central questions for the general theory:

A. For what classes of functions $\mathcal{F}$ does a natural generalization of the Glivenko-Cantelli theorem hold? That is, for what classes $\mathcal{F}$ do we have

$$\|\mathbb{P}_n - P\|_{\mathcal{F}}^* \rightarrow_{a.s.} 0$$

If this convergence holds, then we say that $\mathcal{F}$ is a $P-$Glivenko-Cantelli class of functions.

B. For what classes of functions $\mathcal{F}$ does a natural generalization of Donsker's theorem hold? That is, for what classes $\mathcal{F}$ do we have

$$\mathbb{G}_n \Rightarrow \mathbb{G}_P \quad \text{in} \quad \ell^\infty(\mathcal{F})?$$

If this convergence holds, then we say that $\mathcal{F}$ is a $P-$Donsker class of functions.

Here $\mathbb{G}_P$ is a 0−mean $P-$Brownian bridge process with uniformly-continuous sample paths with respect to the semi-metric $\rho_P(f,g)$ defined by

$$\rho_P^2(f,g) = Var_P(f(X) - g(X)),$$

$\ell^\infty(\mathcal{F})$ is the space of all bounded, real-valued functions z from $\mathcal{F}$ to $\mathbb{R}$:

$$\ell^\infty(\mathcal{F}) = \left\{ z : \mathcal{F} \mapsto \mathbb{R} \middle| \, \|z\|_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} |z(f)| < \infty \right\},$$

and

$$E\{\mathbb{G}_P(f)\mathbb{G}_P(g)\} = P(fg) - P(f)P(g).$$

3. Some Examples

A commonly occurring problem in statistics: we want to prove consistency or asymptotic normality of some statistic which is *not* a sum of independent random variables, but which can be related to a natural sum of random functions indexed by a parameter in a suitable (metric) space.

Example 1. Suppose that $X_1, \ldots, X_n$ are i.i.d. real-valued with $E|X_1| < \infty$, and let $\mu = E(X_1)$. Consider the absolute deviations about the sample mean,

$$D_n = \mathbb{P}_n|X - \overline{X}_n| = n^{-1} \sum_{i=1}^{n} |X_i - \overline{X}_n|.$$

Since $\overline{X}_n \rightarrow_{a.s.} \mu$, we know that for any $\delta > 0$ we have $\overline{X} \in [\mu - \delta, \mu + \delta]$ for all sufficiently large n almost surely. Thus we see that if we define

$$D_n(t) \equiv \mathbb{P}_n|x - t| = n^{-1} \sum_{i=1}^{n} |X_i - t|,$$

then $D_n = D_n(\overline{X}_n)$ and study of $D_n(t)$ for $t \in [\mu - \delta, \mu + \delta]$ is equivalent to study of the empirical measure $\mathbb{P}_n$ indexed by the class of functions

$$\mathcal{F}_\delta = \{x \mapsto |x - t| \equiv f_t(x) : \quad t \in [\mu - \delta, \mu + \delta]\}.$$

To show that $D_n \to_{a.s.} d \equiv E|X - \mu|$, we write

$$\begin{aligned} D_n - d &= \mathbb{P}_n|X - \overline{X}_n| - P|X - \mu| && (1) \\ &= (\mathbb{P}_n - P)(|X - \overline{X}_n|) + P|X - \overline{X}_n| - P|X - \mu| \\ &\equiv I_n + II_n. && (2) \end{aligned}$$

Now

$$\begin{aligned} |I_n| &= |(\mathbb{P}_n - P)(|X - \overline{X}_n|)| \\ &\le \sup_{t:|t-\mu|\le\delta} |(\mathbb{P}_n - P)|X - t|| = \sup_{f \in \mathcal{F}_\delta} |(\mathbb{P}_n - P)(f)| \\ &\to_{a.s.} 0 && (3) \end{aligned}$$

if $\mathcal{F}_\delta$ is $P-$Glivenko-Cantelli.

But convergence of the second term in (2) is easy: by the triangle inequality

$$II_n = |P|X - \overline{X}_n| - P|X - \mu|| \quad \le \quad P|\overline{X}_n - \mu| = |\overline{X}_n - \mu|$$
$$\to_{a.s.} \quad 0.$$

How to prove (3)? Consider the functions $f_1, \ldots, f_m \in \mathcal{F}_\delta$ given by

$$f_j(x) = |x - (\mu - \delta(1 - j/m)|, \quad j = 0, \ldots, 2m.$$

For this finite set of functions we have

$$\max_{0 \le j \le 2m} |(\mathbb{P}_n - P)(f_j)| \to_{a.s.} 0$$

by the strong law of large numbers applied $2m + 1$ times. Furthermore ...

it follows that for $t \in [\mu - \delta(1 - j/m), \mu - \delta(1 - (j+1)/m)]$ the functions $f_t(x) = |x - t|$ satisfy (picture!)

$$L_j(x) \equiv f_{j/m}(x) \wedge f_{(j+1)/m}(x) \le f_t(x) \le f_{j/m}(x) \vee f_{(j+1)/m}(x) \equiv U_j(x)$$

where

$$U_j(x) - f_t(x) \le \frac{1}{m}, \quad f_t(x) - L_j(x) \le \frac{1}{m}, \quad U_j(x) - L_j(x) \le \frac{1}{m}.$$

Thus for each m

$$\begin{aligned} &\|\mathbb{P}_n - P\|_{\mathcal{F}_\delta} \\ &\quad \equiv \quad \sup_{f \in \mathcal{F}_\delta} |(\mathbb{P}_n - P)(f)| \\ &\quad \le \quad \max\left\{ \max_{0 \le j \le 2m} |(\mathbb{P}_n - P)(U_j)|, \ \max_{0 \le j \le 2m} |(\mathbb{P}_n - P)(L_j)| \right\} + 1/m \\ &\quad \to_{a.s.} \quad 0 + 1/m \end{aligned}$$

Taking m large shows that (3) holds.

This is a bracketing argument, and generalizes easily to yield a quite general bracketing Glivenko-Cantelli theorem.

How to prove $\sqrt{n}(D_n - d) \to_d$? We write

$$
\begin{aligned}
\sqrt{n}(D_n - d) &= \sqrt{n}(\mathbb{P}_n|X-\overline{X}_n| - P|X-\mu|) \\
&= \sqrt{n}(\mathbb{P}_n|X-\mu| - P|X-\mu|) \\
&\quad + \sqrt{n}(P|X-\overline{X}_n| - P|X-\mu|) \\
&\quad + \sqrt{n}(\mathbb{P}_n - P)(|X-\overline{X}_n|) - \sqrt{n}(\mathbb{P}_n - P)(|X-\mu|) \\
&= \mathbb{G}_n(|X-\mu|) + \sqrt{n}(H(\overline{X}_n) - H(\mu)) \\
&\quad + \mathbb{G}_n(|X-\overline{X}_n| - |X-\mu|) \\
&= \mathbb{G}_n(|X-\mu|) + H'(\mu)(\overline{X}_n - \mu) \\
&\quad + \sqrt{n}(H(\overline{X}_n) - H(\mu) - H'(\mu)(\overline{X}_n - \mu)) \\
&\quad + \mathbb{G}_n(|X-\overline{X}_n| - |X-\mu|) \\
&\equiv \mathbb{G}_n(|X-\mu| + H'(\mu)(X-\mu)) + I_n + II_n
\end{aligned}
$$

where ...

$$
\begin{aligned}
H(t) &\equiv P|X-t|,\\
I_n &\equiv \sqrt{n}(H(\overline{X}_n) - H(\mu) - H'(\mu)(\overline{X}_n - \mu)),\\
II_n &\equiv \mathbb{G}_n(|X-\overline{X}_n|) - \mathbb{G}_n(|X-\mu|)\\
&= \mathbb{G}_n(|X-\overline{X}_n| - |X-\mu|)\\
&= \mathbb{G}_n(f_{\overline{X}_n} - f_\mu)\,.
\end{aligned}
$$

Here $I_n \rightarrow_p 0$ if $H(t) \equiv P|X-t|$ is differentiable at μ. The second term

$$
II_n \equiv \mathbb{G}_n(f_{\overline{X}_n} - f_\mu) \rightarrow_p 0
$$

if $\mathcal{F}_\delta$ is a Donsker class of functions! This is a consequence of asymptotic equicontinuity of $\mathbb{G}_n$ over the class $\mathcal{F}$: for every $\epsilon > 0$

$$
\lim_{\delta\searrow 0} \limsup_{n\rightarrow\infty} Pr^*\Big(\sup_{f,g:\ \rho_P(f,g)\le\delta} |\mathbb{G}_n(f) - \mathbb{G}_n(g)| > \epsilon\Big) = 0.
$$

Example 2. Copula models: the pseudo-MLE.
Let $c_\theta(u_1, \ldots, u_p)$ be a copula density with $\theta \subset \Theta \subset R^q$. Suppose that $X_1, \ldots, X_n$ are i.i.d. with density

$$f(x_1, \ldots, x_p) = c_\theta(F_1(x_1), \ldots, F_p(x_p)) \cdot f_1(x_1) \cdots f_p(x_p)$$

where $F_1, \ldots, F_p$ are absolutely continuous d.f.'s with densities $f_1, \ldots, f_p$.

Let

$$\mathbb{F}_{n,j}(x_j) \equiv n^{-1} \sum_{i=1}^{n} 1\{X_{i,j} \le x_j\}, \qquad j = 1, \ldots, p$$

be the marginal empirical d.f.'s of the data. Then a natural pseudo-likelihood function is given by

$$l_n(\theta) \equiv \mathbb{P}_n \log c_\theta(\mathbb{F}_{n,1}(x_1), \ldots, \mathbb{F}_{n,p}(x_p)).$$

Thus it seems reasonable to define the pseudo-likelihood estimator $\widehat{\theta}_n$ of θ by the $q-$dimensional system of equations

$$\Psi_n(\widehat{\theta}_n) = 0$$

where

$$\Psi_n(\theta) \equiv \mathbb{P}_n(\dot{\ell}_\theta(\theta; \mathbb{F}_{n,1}(x_1), \ldots, \mathbb{F}_{n,p}(x_p))$$

and where

$$\dot{\ell}_\theta(\theta; u_1, \ldots, u_p) \equiv \nabla_\theta \text{log} c_\theta(u_1, \ldots, u_p).$$

We also define $\Psi(\theta)$ by

$$\Psi(\theta) \equiv P_0(\dot{\ell}_\theta(\theta, F_1(x_1), \ldots, F_p(x_p)).$$

Then we expect that

$$0 = \Psi_n(\widehat{\theta}_n) = \Psi_n(\theta_0) - \left\{ -\dot{\Psi}_n(\theta_n^*) \right\} (\widehat{\theta}_n - \theta_0) \tag{4}$$

where

$$\Psi_n(\theta_0) \;=\; \mathbb{P}_n \dot{\ell}_\theta(\theta_0, \mathbb{F}_{n,1}(x_1), \ldots, \mathbb{F}_{n,p}(x_p)),$$

and

$$\begin{aligned} -\dot{\Psi}_n(\theta_n^*) &= -\mathbb{P}_n \ddot{\ell}_{\theta,\theta}(\theta_n^*, \mathbb{F}_{n,1}(x_1), \ldots, \mathbb{F}_{n,p}(x_p)) \\ &\to_p -P_0(\ddot{\ell}_{\theta,\theta}(\theta_0, F_1(x_1), \ldots, F_p(x_p)) \qquad (5) \\ &\equiv B \equiv I_{\theta\theta}, \qquad (6) \end{aligned}$$

a $q \times q$ matrix. On the other hand ...

$$\sqrt{n}\Psi_n(\theta_0) = \sqrt{n}\mathbb{P}_n\dot{\ell}_\theta(\theta_0, \mathbb{F}_{n,1}(x_1), \ldots, \mathbb{F}_{n,p}(x_p))$$

where

$$\begin{aligned}
&\dot{\ell}_\theta(\theta_0, \mathbb{F}_{n,1}(x_1), \ldots, \mathbb{F}_{n,p}(x_p)) \\
&\quad = \dot{\ell}_\theta(\theta_0, F_1(x_1), \ldots, F_p(x_p)) \\
&\qquad\qquad + \sum_{j=1}^{p} \ddot{\ell}_{\theta,j}(\theta_0, u_1^*, \ldots, u_p^*) \cdot (\mathbb{F}_{n,j}(x_j) - F_j(x_j)),
\end{aligned}$$

$$\ddot{\ell}_{\theta,j}(\theta_0, u_1, \ldots, u_p) \equiv \frac{\partial}{\partial u_j}\dot{\ell}_\theta(\theta_0, u_1, \ldots, u_p),$$

and where $|u_j^*(x_j) - F_j(x_j)| \le |\mathbb{F}_{n,j}(x_j) - F_j(x_j)|$ for $j = 1, \ldots, p$. Thus we expect that

$$
\begin{aligned}
&\sqrt{n}\Psi_n(\theta_0) \\
&\quad = \sqrt{n}\mathbb{P}_n(\dot\ell_\theta(\theta_0, \mathbb{F}_{n,1}(x_1), \ldots, \mathbb{F}_{n,p}(x_p)) \\
&\quad \doteq \mathbb{G}_n\left(\dot\ell_\theta(\theta_0, F_1(x_1), \ldots, F_p(x_p))\right) \\
&\qquad\qquad + \mathbb{P}_n\left(\sum_{j=1}^p \ddot\ell_{\theta,j}(\theta_0, u_1^*, \ldots, u_p^*)\cdot\sqrt{n}(\mathbb{F}_{n,j}(x_j) - F_j(x_j))\right) \\
&\quad = \mathbb{G}_n\left(\dot\ell_\theta(\theta_0, F_1(x_1), \ldots, F_p(x_p))\right) \\
&\qquad\qquad + P_0\left(\sum_{j=1}^p \ddot\ell_{\theta,j}(\theta_0, u_1^*, \ldots, u_p^*)\cdot\sqrt{n}(\mathbb{F}_{n,j}(x_j) - F_j(x_j))\right) \\
&\qquad\qquad + (\mathbb{P}_n - P_0)\left(\sum_{j=1}^p \ddot\ell_{\theta,j}(\theta_0, u_1^*, \ldots, u_p^*)\cdot\sqrt{n}(\mathbb{F}_{n,j}(x_j) - F_j(x_j))\right).
\end{aligned}
$$

In this last display the third term will be negligible (via asymptotic equicontinuity!) and the second term can be rewritten as

$$
\begin{aligned}
& P_0\left(\sum_{j=1}^{p} \ddot{\ell}_{\theta,j}(\theta_0, u_1^*, \ldots, u_p^*) \cdot \sqrt{n}(\mathbb{F}_{n,j}(x_j) - F_j(x_j))\right) \\
&= \sum_{j=1}^{p} P_0 \ddot{\ell}_{\theta,j}(\theta_0, u_1^*(x_1), \ldots, u_p^*(x_p)) \cdot \sqrt{n}(\mathbb{F}_{n,j}(x_j) - F_j(x_j)) \\
&\doteq \mathbb{G}_n\left(\sum_{j=1}^{p} \int_{R^p} \ddot{\ell}_{\theta,j}(\theta_0, F_1(x_1), \ldots, F_p(x_p)) \right. \\
&\qquad \left. \cdot \Big(1\{X_j \le x_j\} - F_j(x_j)\Big)\, dC_\theta(F_1(x_1), \ldots, F_p(x_p))\right) \\
&= \mathbb{G}_n\left(\sum_{j=1}^{p} \int_{[0,1]^p} \ddot{\ell}_{\theta,j}(\theta_0, u_1, \ldots, u_p) \right. \\
&\qquad \left. \cdot \Big(1\{F_j(X_j) \le u_j\} - u_j\Big)\, dC_\theta(u_1, \ldots, u_p)\right) \\
&= \mathbb{G}_n\left(\sum_{j=1}^{p} W_j(X_j)\right)
\end{aligned}
$$

Example 3. Kendall's function.
Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ are i.i.d. F_0 on $\mathbb{R}^2$, and let $\mathbb{F}_n$ denote their (classical) empirical distribution function

$$\mathbb{F}_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]\times(-\infty,y]}(X_i, Y_i).$$

Consider the empirical distribution function of the random variables $\mathbb{F}_n(X_i, Y_i)$, $i = 1, \ldots, n$:

$$\mathbb{K}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{[\mathbb{F}_n(X_i,Y_i)\le t]}, \quad t \in [0, 1].$$

As in example 1, the random variables $\{\mathbb{F}_n(X_i, Y_i)\}_{i=1}^{n}$ are dependent, and we are already studying a stochastic process indexed by $t \in [0, 1]$. The empirical process method leads to study of the process $\mathbb{K}_n$ indexed by *both* $t \in [0, 1]$ *and* $F \in \mathcal{F}_2$, the class of all distribution functions F on $\mathbb{R}^2$:

$$\mathbb{K}_n(t, F) \equiv \frac{1}{n}\sum_{i=1}^{n} 1_{[F(X_i,Y_i)\le t]} = \mathbb{P}_n 1_{[F(X,Y)\le t]}$$

with $t \in [0, 1]$ and $F \in \mathcal{F}_2$ ... or the smaller set

$$\mathcal{F}_{2,\delta} = \{F \in \mathcal{F}_2 : \ \|F - F_0\|_\infty \leq \delta\}.$$

Example 4. Completely monotone densities.
Consider the class $\mathcal{P}$ of completely monotone densities p_G given by

$$p_G(x) = \int_0^\infty z \mathrm{exp}(-zx) dG(z)$$

where G is an arbitrary distribution function on $\mathbb{R}^+$. Consider the maximum likelihood estimator $\widehat{p}$ of $p \in \mathcal{P}$: i.e.

$$\widehat{p} \equiv \mathrm{argmax}_{p \in \mathcal{P}} \mathbb{P}_n \mathrm{log}(p).$$

Question: Is $\widehat{p}$ Hellinger consistent? That is, do we have

$$h(\widehat{p}_n, p_0) \rightarrow_{a.s.} 0?$$

Part II: Some basic inequalities and Glivenko-Cantelli theorems

- 1. Tools for consistency: two basic inequalities.
- 2. Tools for consistency:
  a further basic inequality for convex $\mathcal{P}$.
- 3. More basic inequalities:
  least squares estimators; penalized ML.
- 4. Glivenko-Cantelli theorems.

1. Tools for consistency: two basic inequalities

Density estimation Suppose that:

- $\mathcal{P}$ is a class of densities with respect to a fixed $\sigma-$finite measure μ on a measurable space $(\mathcal{X}, \mathcal{A})$.
- Suppose that $X_1, \ldots, X_n$ are i.i.d. P_0 with density $p_0 \in \mathcal{P}$.
- Then the Maximum Likelihood Estimator (MLE) for the class $\mathcal{P}$ is

$$\widehat{p}_n \equiv \text{argmax}_{p \in \mathcal{P}} \mathbb{P}_n \log(p) .$$

Here are two "basic inequalities" for density estimation.

Proposition 1.1. (Van de Geer). Suppose that $\widehat{p}_n$ maximizes $\mathbb{P}_n \log(p)$ over $\mathcal{P}$. then

$$h^2(\widehat{p}_n, p_0) \le (\mathbb{P}_n - P_0)\left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) 1\{p_0 > 0\}\,.$$

Proposition 1.2. (Birgé and Massart). If $\widehat{p}_n$ maximizes $\mathbb{P}_n \log(p)$ over $\mathcal{P}$, then

$$\begin{aligned} & h^2((\widehat{p}_n + p_0)/2, p_0) \\ \le \;\; & (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right) 1_{[p_0>0]}\right)\,, \end{aligned}$$

and

$$h^2(\widehat{p}_n, p_0) \le 24 h^2\left(\frac{\widehat{p}_n + p_0}{2}, p_0\right)\,.$$

- Proposition 1.1 leads to the class of functions

$$\mathcal{F} = \left\{ \left(\sqrt{\frac{p}{p_0}} - 1 \right) : \quad p \in \mathcal{P} \right\}.$$

and the question: Is $\mathcal{F}$ a P_0−Glivenko class?

- Proposition 1.2 leads to the class of functions

$$\mathcal{F} = \left\{ \left(\frac{1}{2} \log \left(\frac{p + p_0}{2p_0} \right) 1_{[p_0 > 0]} \right) : \quad p \in \mathcal{P} \right\}.$$

and the question: Is $\mathcal{F}$ a P_0−Glivenko class?

Proof, proposition 1.1: Since $\widehat{p}_n$ maximizes $\mathbb{P}_n \log p$,

$$
\begin{aligned}
0 &\le \frac{1}{2}\int_{[p_0>0]} \log\left(\frac{\widehat{p}_n}{p_0}\right) d\mathbb{P}_n \\
&\le \int_{[p_0>0]} \left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) d\mathbb{P}_n \\
&\qquad \text{since } \log(1+x) \le x \\
&= \int_{[p_0>0]} \left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) d(\mathbb{P}_n - P_0) \\
&\qquad + P_0\left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) 1\{p_0 > 0\} \\
&= \int_{[p_0>0]} \left(\sqrt{\frac{\widehat{p}_n}{p_0}} - 1\right) d(\mathbb{P}_n - P_0) - h^2(\widehat{p}_n, p_0)
\end{aligned}
$$

where the last equality follows by direct calculation and the definition of the Hellinger metric h. □

Proof, Proposition 1.2: By concavity of log,

$$\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right) 1_{[p_0>0]} \geq \frac{1}{2}\log\left(\frac{\widehat{p}_n}{p_0}\right) 1_{[p_0>0]}\,.$$

Thus

$$\begin{aligned}
0 &\leq \mathbb{P}_n\left(\frac{1}{4}\log\left(\frac{\widehat{p}_n}{p_0}\right)1_{[p_0>0]}\right) \leq \mathbb{P}_n\left(\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right)\\
&= (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right)\\
&\quad + P_0\left(\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right)\\
&= (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right) - \frac{1}{2}K(P_0, (\widehat{P}_n + P_0)/2)\\
&\leq (\mathbb{P}_n - P_0)\left(\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right)1_{[p_0>0]}\right) - h^2(P_0, (\widehat{P}_n + P_0)/2)\,.
\end{aligned}$$

where we used Exercise 1.2 at the last step. The second claim follows from Exercise 1.4. $\square$

Exercise 1.2: (Pinsker inequalities)
(a) $K(P,Q) \ge 2h^2(P,Q) = \int [\sqrt{p} - \sqrt{q}]^2 d\mu$.
(b) $K(P,Q) \ge (1/2) \left(\int |p-q| d\mu \right)^2 = 2d_{TV}^2(P,Q)$.

Exercise 1.4:

$$2h^2(P,(P+Q)/2) \le h^2(P,Q) \le 12h^2(P,(P+Q)/2).$$

Corollary 1.1. (Hellinger consistency of MLE). Suppose that either

$$\{(\sqrt{p/p_0} - 1)1\{p_0 > 0\} : \; p \in \mathcal{P}\}, \;\text{ or }\; \left\{ \frac{1}{2} \log\left(\frac{p + p_0}{2p_0} \right) 1_{[p_0 > 0]} : \; p \in \mathcal{P} \right\}$$

is a P_0-Glivenko-Cantelli class. Then $h(\widehat{p}_n, p_0) \to_{a.s.} 0$.

2. Tools for consistency: a further basic inequality.

- For $0 < \alpha \leq 1$, let $\varphi_\alpha(t) = (t^\alpha - 1)/(t^\alpha + 1)$ for $t \geq 0$, $\varphi(t) = -1$ for $t < 0$. Thus φ_α is bounded and continuous for each $\alpha \in (0,1]$.

- For $0 < \beta < 1$ define

$$h_\beta^2(p,q) \equiv 1 - \int p^\beta q^{1-\beta} d\mu \, .$$

- Note that

$$h_{1/2}^2(p,q) \equiv h^2(p,q) = \frac{1}{2}\int \{\sqrt{p} - \sqrt{q}\}^2 d\mu$$

yields the Hellinger distance between p and q. By Hölder's inequality, $h_\beta(p,q) \geq 0$ with equality if and only if $p = q$ a.e. μ.

Proposition 1.3. Suppose that $\mathcal{P}$ is convex. Then

$$h^2_{1-\alpha/2}(\widehat{p}_n, p_0) \le (\mathbb{P}_n - P_0)\left(\varphi_\alpha\left(\frac{\widehat{p}_n}{p_0}\right)\right).$$

In particular, when $\alpha = 1$ we have, with $\varphi \equiv \varphi_1$,

$$\begin{aligned} h^2(\widehat{p}_n, p_0) = h^2_{1/2}(\widehat{p}_n, p_0) &\le (\mathbb{P}_n - P_0)\left(\varphi\left(\frac{\widehat{p}_n}{p_0}\right)\right) \\ &= (\mathbb{P}_n - P_0)\left(\frac{2\widehat{p}_n}{\widehat{p}_n + p_0}\right). \end{aligned}$$

Corollary 1.2. Suppose that $\{\varphi(p/p_0) : p \in \mathcal{P}\}$ is a P_0-Glivenko-Cantelli class. Then for each $0 < \alpha \le 1$, $h_{1-\alpha/2}(\widehat{p}_n, p_0) \rightarrow_{a.s.} 0$.

Proof. Since $\mathcal{P}$ is convex and $\widehat{p}_n$ maximizes $\mathbb{P}_n \log p$ over $\mathcal{P}$, it follows that

$$\mathbb{P}_n \log \frac{\widehat{p}_n}{(1-t)\widehat{p}_n + tp_1} \ge 0$$

for all $0 \le t \le 1$ and every $p_1 \in \mathcal{P}$; this holds in particular for $p_1 = p_0$. Note that equality holds if $t = 0$. Differentiation of the left side with respect to t at $t = 0$ yields

$$\mathbb{P}_n \frac{p_1}{\widehat{p}_n} \le 1 \qquad \text{for every} \quad p_1 \in \mathcal{P}\,.$$

If $L : (0, \infty) \mapsto R$ is increasing and $t \mapsto L(1/t)$ is convex, then Jensen's inequality yields

$$\mathbb{P}_n L\left(\frac{\widehat{p}_n}{p_1}\right) \ge L\left(\frac{1}{\mathbb{P}_n(p_1/\widehat{p}_n)}\right) \ge L(1) = \mathbb{P}_n L\left(\frac{p_1}{p_1}\right)\,.$$

Choosing $L = \varphi_\alpha$ and $p_1 = p_0$ in this last inequality and noting that $L(1) = 0$, it follows that

$$\begin{aligned} 0 &\le \mathbb{P}_n \varphi_\alpha(\widehat{p}_n/p_0) \\ &= (\mathbb{P}_n - P_0)\varphi_\alpha(\widehat{p}_n/p_0) \quad + \quad P_0\varphi_\alpha(\widehat{p}_n/p_0)\,; \end{aligned} \tag{7}$$

see van der Vaart and Wellner (1996) page 330, and Pfanzagl

(1988), pages 141 - 143. Now we show that

$$P_0\varphi_\alpha(p/p_0) = \int \frac{p^\alpha - p_0^\alpha}{p^\alpha + p_0^\alpha} dP_0 \le -\left(1 - \int p_0^\beta p^{1-\beta} d\mu\right) \qquad (8)$$

Note that this holds if and only if

$$-1 + 2\int \frac{p^\alpha}{p_0^\alpha + p^\alpha} p_0 d\mu \le -1 + \int p_0^\beta p^{1-\beta} d\mu\,,$$

or

$$\int p_0^\beta p^{1-\beta} d\mu \ge 2\int \frac{p^\alpha}{p_0^\alpha + p^\alpha} p_0 d\mu\,.$$

But this holds if

$$p_0^\beta p^{1-\beta} \ge 2\frac{p^\alpha p_0}{p_0^\alpha + p^\alpha}\,.$$

With $\beta = 1 - \alpha/2$, this becomes

$$\frac{1}{2}(p_0^\alpha + p^\alpha) \ge p_0^{\alpha/2} p^{\alpha/2} = \sqrt{p_0^\alpha p^\alpha}\,,$$

and this holds by the arithmetic mean - geometric mean inequality, $\sqrt{ab} \leq (a+b)/2$. Thus (8) holds. Combining (8) with (7) yields the claim of the proposition.

The corollary follows by noting that $\varphi(t) = (t-1)/(t+1) = 2t/(t+1) - 1$. $\square$

3. More basic inequalities: penalized ML & LS

Penalized ML:

- Suppose that $\mathcal{P}$ is a collection of densities described by a "penalty functional" $I(p)$:

$$\mathcal{P} = \{p : \mathbb{R} \rightarrow [0, \infty) : \int p(x)dx = 1, \; I^2(p) < \infty\}$$

For example, $I^2(p) = \int (p''(x))^2 dx$.

- Suppose that

$$\widehat{p}_n = \text{argmax}_{p \in \mathcal{P}} \left(\mathbb{P}_n \log(p) - \lambda_n^2 I^2(p) \right);$$

here λ_n is a smoothing parameter.

Basic inequality: (van de Geer, 2000, page 175): For $p_0 \in \mathcal{P}$

$$h^2(\widehat{p}_n, p_0) + 4\lambda_n^2 I^2(\widehat{p}_n) \le 16(\mathbb{P}_n - P_0)\frac{1}{2}\log\left(\frac{\widehat{p}_n + p_0}{2p_0}\right) + 4\lambda_n^2 I^2(p_0).$$

Least squares regression:

- Suppose that $Y_i = g_0(z_i) + W_i$, where $EW_i = 0$, $Var(W_i) \leq \sigma_0^2$.
- $Q_n = n^{-1} \sum_{i=1}^n \delta_{z_i}$, $\|g\|_n^2 \equiv n^{-1} \sum_{i=1}^n g(z_i)^2$.
- $\|y - g\|_n^2 = n^{-1} \sum_1^n (Y_i - g(z_i))^2$.
- $\langle w, g \rangle_n = n^{-1} \sum_1^n W_i g(z_i)$.
- $\widehat{g}_n \equiv \text{argmin}_{g \in \mathcal{G}} \|y - g\|_n^2$.

Basic inequality: (van de Geer, 2000, page 55).

$$\begin{aligned} \|\widehat{g}_n - g_0\|_n^2 &\leq 2\langle w, \ \widehat{g}_n - g_0 \rangle_n \\ &= 2n^{-1} \sum_{i=1}^n W_i \left(\widehat{g}_n(z_i) - g_0(z_i) \right). \end{aligned}$$

4. Glivenko-Cantelli Theorems:

Bracketing:

Given two functions l and u on $\mathcal{X}$, the *bracket* $[l,u]$ is the set of all functions $f \in \mathcal{F}$ with $l \le f \le u$. The functions l and u need not belong to $\mathcal{F}$, but are assumed to have finite norms. An $\epsilon-$*bracket* is a bracket $[l,u]$ with $\|u-l\| \le \epsilon$. The *bracketing number* $N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of $\epsilon-$brackets needed to cover $\mathcal{F}$. The *entropy with bracketing* is the logarithm of the bracketing number.

Theorem 1. Let $\mathcal{F}$ be a class of measurable functions such that $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Then $\mathcal{F}$ is $P-$Glivenko-Cantelli; that is

$$\|\mathbb{P}_n - P\|_{\mathcal{F}}^* = \left(\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \right)^* \to_{a.s.} 0\,.$$

Proof. Fix $\epsilon > 0$. Choose finitely many $\epsilon-$brackets $[l_i, u_i]$, $i = 1, \ldots, m = N(\epsilon, \mathcal{F}, L_1(P))$, whose union contains $\mathcal{F}$ and such that $P(u_i - l_i) < \epsilon$ for all $1 \leq i \leq m$. Thus, for every $f \in \mathcal{F}$ there is a bracket $[l_i, u_i]$ such that

$$(\mathbb{P}_n - P)f \leq (\mathbb{P}_n - P)u_i + P(u_i - f) \leq (\mathbb{P}_n - P)u_i + \epsilon\,.$$

Similarly,

$$(P - \mathbb{P}_n)f \leq (P - \mathbb{P}_n)l_i + P(f - l_i) \leq (P - \mathbb{P}_n)l_i + \epsilon\,.$$

□

It is not hard to see that bracketing condition of Theorem 1 is sufficient but not necessary.

In contrast, our second Glivenko-Cantelli theorem gives conditions which are both necessary and sufficient.

A simple setting in which this theorem applies involves a collection of functions $f = f(\cdot, t)$ indexed or parametrized by $t \in T$, a compact subset of a metric space $(\mathbb{D}, d)$. Here is the basic lemma; it goes back to Wald (1949) and Le Cam (1953).

Lemma 1. Suppose that $\mathcal{F} = \{f(\cdot, t) : \ t \in T\}$ where the functions $f : \mathcal{X} \times T \mapsto R$, are continuous in t for $P-$ almost all $x \in \mathcal{X}$. Suppose that T is compact and that the envelope function F defined by $F(x) = \sup_{t \in T} |f(x, t)|$ satisfies $P^* F < \infty$. Then

$$N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$$

for every $\epsilon > 0$, and hence $\mathcal{F}$ is $P-$Glivenko-Cantelli.

The qualitative statement of the preceding lemma can be quantified as follows:

Lemma 2. Suppose that $\{f(\cdot, t) : t \in T\}$ is a class of functions satisfying

$$|f(x,t) - f(x,s)| \leq d(s,t)F(x)$$

for all $s, t \in T$, $x \in \mathcal{X}$ for some metric d on the index set, and a function F on the sample space $\mathcal{X}$. Then, for any norm $\|\cdot\|$,

$$N_{[\,]}(2\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\epsilon, T, d)\,.$$

For our second Glivenko-Cantelli theorem, we need:

- An envelope function F for a class of functions $\mathcal{F}$ is any function satisfying

$$|f(x)| \leq F(x) \quad \text{for all} \quad x \in \mathcal{X} \quad \text{and for all} \quad f \in \mathcal{F}.$$

- A class of functions $\mathcal{F}$ is $L_1(P)$ bounded if $\sup_{f\in\mathcal{F}} P|f| < \infty$.

Theorem 2.. (Vapnik and Chervonenkis (1981), Pollard (1981), Giné and Zinn (1984)). Let $\mathcal{F}$ be a $P-$measurable class of measurable functions that is $L_1(P)-$bounded. Then $\mathcal{F}$ is $P-$Glivenko-Cantelli if and only if both

(i) $P^*F < \infty$.

(ii)

$$\lim_{n\to\infty} \frac{E^*\log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))}{n} = 0$$

for all $M < \infty$ and $\epsilon > 0$ where $\mathcal{F}_M$ is the class of functions $\{f1\{F \le M\} : f \in \mathcal{F}\}$.

For n points $x_1, \ldots, x_n$ in $\mathcal{X}$ and a class of $\mathcal{C}$ of subsets of $\mathcal{X}$, set

$$\Delta_n^{\mathcal{C}}(x_1, \ldots, x_n) \equiv \#\left\{C \cap \{x_1, \ldots, x_n\} : \ C \in \mathcal{C}\right\}.$$

Corollary. (Vapnik-Chervonenkis-Steele GC theorem) If $\mathcal{C}$ is a $P-$measurable class of sets, then the following are equivalent:
(i) $\|\mathbb{P}_n - P\|_{\mathcal{C}}^* \to_{a.s.} 0$
(ii) $n^{-1} E \log \Delta^{\mathcal{C}}(X_1, \ldots, X_n) \to 0$; where,

The second hypothesis is often verified by applying the theory of VC (or Vapnik-Chervonenkis) classes of sets and functions. Let

$$m^{\mathcal{C}}(n) \equiv \max_{x_1, \ldots, x_n} \Delta_n^{\mathcal{C}}(x_1, \ldots, x_n),$$

and let

$$V(\mathcal{C}) \equiv \inf\{n : \ m^{\mathcal{C}}(n) < 2^n\},$$
$$S(\mathcal{C}) \equiv \sup\{n : \ m^{\mathcal{C}}(n) = 2^n\}.$$

Examples:

(1) $\mathcal{X} = \mathbb{R}$, $\mathcal{C} = \{(-\infty, t] : \ t \in \mathbb{R}\}$: $S(\mathcal{C}) = 1$.

(2) $\mathcal{X} = \mathbb{R}$, $\mathcal{C} = \{(s, t] : \ s < t, s, t \in \mathbb{R}\}$: $S(\mathcal{C}) = 2$.

(3) $\mathcal{X} = \mathbb{R}^d$, $\mathcal{C} = \{(s, t] : s < t, s, t \in \mathbb{R}^d\}$: $S(\mathcal{C}) = 2d$.

(4) $\mathcal{X} = \mathbb{R}^d$, $H_{u,c} \equiv \{x \in \mathbb{R}^d : \ \langle x, u \rangle \le c\}$,
$\mathcal{C} = \{H_{u,c} : \ u \in \mathbb{R}^d, \ c \in \mathbb{R}\}$: $S(\mathcal{C}) = d + 1$.

(5) $\mathcal{X} = \mathbb{R}^d$, $B_{u,r} \equiv \{x \in R^d : \|x - u\| \le r\}$;
$\mathcal{C} = \{B_{u,r} : \ u \in \mathbb{R}^d, \ r \in \mathbb{R}^+\}$: $S(\mathcal{C}) = d + 1$.

Definition. The *subgraph* of $f : \mathcal{X} \to \mathbb{R}$ is the subset of $\mathcal{X} \times \mathbb{R}$ given by $\{(x, t) \in \mathcal{X} \times \mathbb{R} : \ t < f(x)\}$. A collection of functions $\mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$ is called a VC-subgraph class if the collection of subgraphs in $\mathcal{X} \times \mathbb{R}$ is a VC - class of sets. For a VC-subgraph class $\mathcal{F}$, let $V(\mathcal{F}) \equiv V(\text{subgraph}(\mathcal{F}))$.

Theorem. For a VC-subgraph class with envelope function F and $r \geq 1$, and for any probability measure Q with $\|F\|_{L_r(Q)} > 0$,

$$N(2\epsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq KV(\mathcal{F}) \left(\frac{16e}{\epsilon^r}\right)^{S(\mathcal{F})}.$$

Here is a specific result for monotone functions on $\mathbb{R}$:

Theorem. Let $\mathcal{F}$ be the class of all monotone functions $f : \mathbb{R} \to [0, 1]$. Then:
(i) (Birman and Solomojak (1967), van de Geer (1991)):

$$\log N_{[\,]}(\epsilon, \mathcal{F}, L_r(Q)) \leq \frac{K}{\epsilon}$$

for every probability measure Q, every $r \geq 1$, and a constant K depending on r only.
(ii) (via convex hull theory):

$$\sup_Q \log N(\epsilon, \mathcal{F}, L_2(Q)) \leq \frac{K}{\epsilon}$$

Part III: Using the Glivenko-Cantelli theorems: first applications

- 1. Preservation of Glivenko-Cantelli theorems.
  - ▷ Preservation under continuous functions.
  - ▷ Preservation under partitions of the sample space.
- 2. First applications
  - ▷ Example 1: current status data
  - ▷ Example 2: Mixed case interval censoring
  - ▷ Example 3: Completely monotone densities.

1. Preservation of Glivenko-Cantelli theorems.

Theorem 1. (van der Vaart & W, 2001). Suppose that $\mathcal{F}_1, \ldots, \mathcal{F}_k$ are $P-$ Glivenko-Cantelli classes of functions, and that $\varphi : \mathbb{R}^k \to \mathbb{R}$ is continuous. Then $\mathcal{H} \equiv \varphi(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ is $P-$ Glivenko-Cantelli provided that it has an integrable envelope function.

Corollary 1. (Dudley, 1998). Suppose that $\mathcal{F}$ is a Glivenko-Cantelli class for P with $PF < \infty$, and g is a fixed bounded function ($\|g\|_\infty < \infty$). Then the class of functions
$g \cdot \mathcal{F} \equiv \{g \cdot f : f \in \mathcal{F}\}$ is a $P-$Glivenko-Cantelli class.

Corollary 2. (Giné and Zinn, 1984). Suppose that $\mathcal{F}$ is a uniformly bounded strong Glivenko-Cantelli class for P, and $g \in \mathcal{L}_1(P)$ is a fixed function. Then the class of functions
$g \cdot \mathcal{F} \equiv \{g \cdot f : f \in \mathcal{F}\}$ is a $P-$Glivenko-Cantelli class.

Theorem 2. (Partitioning of the sample space). Suppose that $\mathcal{F}$ is a class of functions on $(\mathcal{X}, \mathcal{A}, P)$, and $\{\mathcal{X}_i\}$ is a partition of $\mathcal{X}$: $\cup_{i=1}^{\infty} \mathcal{X}_i = \mathcal{X}$, $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ for $i \neq j$. Suppose that $\mathcal{F}_j \equiv \{f 1_{\mathcal{X}_j} : f \in \mathcal{F}\}$ is $P-$Glivenko-Cantelli for each j, and $\mathcal{F}$ has an integrable envelope function F. Then $\mathcal{F}$ is itself $P-$Glivenko-Cantelli.

First Applications:

Example 2.1. (Interval censoring, case I). Suppose that $Y \sim F$ on $\mathbb{R}^+$ and $T \sim G$. Here Y is the time of some event of interest, and T is an "observation time". Unfortunately, we do not observe (Y, T); instead what is observed is $X = (1\{Y \leq T\}, T) \equiv (\Delta, T)$. Our goal is to estimate F, the distribution of Y. Let P_0 be the distribution corresponding to F_0, and suppose that $(\Delta_1, T_1), \ldots, (\Delta_n, T_n)$ are i.i.d. as (Δ, T). Note that the conditional distribution of Δ given T is simply Bernoulli($F(T)$), and hence the density of (Δ, T) with respect to the dominating measure $\# \times G$ (here $\#$ denotes counting measure on $\{0, 1\}$) is given by

$$p_F(\delta, t) = F(t)^{\delta}(1 - F(t))^{1-\delta}.$$

Note that the sample space in this case is

$$\begin{aligned} \mathcal{X} &= \{(\delta, t) : \delta \in \{0, 1\}, t \in R^+\} = \{(1, t) : t \in R^+\} \cup \{(0, t) : t \in R^+\} \\ &:= \mathcal{X}_1 \cup \mathcal{X}_2. \end{aligned}$$

Now the class of functions $\{p_F : F \text{ a d.f. on } R^+\}$ is a universal Glivenko-Cantelli class by an application of GC-preservation Theorem 2, since on $\mathcal{X}_1$, $p_F(1,t) = F(t)$, while on $\mathcal{X}_2$, $p_F(0,t) = 1 - F(t)$ where F is a distribution F (and hence bounded and monotone nondecreasing). Furthermore the class of functions $\{p_F/p_{F_0} : F \text{ a d.f. on } R^+\}$ is P_0-Glivenko by an application of GC-preservation Theorem 1: Take

$$\mathcal{F}_1 = \{p_F : F \text{ a d.f. on } R^+\}, \qquad \mathcal{F}_2 = \{1/p_{F_0}\},$$

and $\varphi(u,v) = uv$. Then both $\mathcal{F}_1$ and $\mathcal{F}_2$ are P_0-Glivenko-Cantelli classes, φ is continuous, and $\mathcal{H} = \varphi(\mathcal{F}_1, \mathcal{F}_2)$ has P_0-integrable envelope $1/p_{F_0}$. Finally, by a further application of GC-preservation Theorem 2 with $\varphi(u) = (t-1)/(t+1)$ shows that the hypothesis of Corollary 2.1.1 holds: $\{\varphi(p_F/p_{F_0}) : F \text{ a d.f. on } R^+\}$ is P_0-Glivenko-Cantelli. Hence the conclusion of the corollary holds: we conclude that

$$h^2(p_{\widehat{F}_n}, p_{F_0}) \rightarrow_{a.s.} 0 \qquad \text{as} \qquad n \rightarrow \infty \, .$$

Now note that $h^2(p, p_0) \geq d_{TV}^2(p, p_0)/2$ and we compute

$$\begin{aligned} d_{TV}(p_{\widehat{F}_n}, p_{F_0}) &= \int |\widehat{F}_n(t) - F_0(t)| dG(t) \\ &\quad + \int |1 - \widehat{F}_n(t) - (1 - F_0(t))| dG(t) \\ &= 2 \int |\widehat{F}_n(t) - F_0(t)| dG(t)\,, \end{aligned}$$

so we conclude that

$$\int |\widehat{F}_n(t) - F_0(t)| dG(t) \to_{a.s.} 0$$

as $n \to \infty$. Since $\widehat{F}_n$ and F_0 are bounded (by one), we can also conclude that

$$\int |\widehat{F}_n(t) - F_0(t)|^r dG(t) \to_{a.s.} 0$$

for each $r \geq 1$, in particular for $r = 2$.

Example 2. (Mixed case interval censoring)

Suppose that:

- $Y \sim F$ on $R^+ = [0, \infty)$.
- Observe:
  - ▷ $T_K = (T_{K,1}, \ldots, T_{K,K})$ where K, the number of times is itself random.
  - ▷ The interval $(T_{K,j-1}, T_{K,j}]$ into which Y falls (with $T_{K,0} \equiv 0$, $T_{K,K+1} \equiv \infty$).
  - ▷ Here $K \in \{1, 2, \ldots\}$, and $\underline{T} = \left\{ T_{k,j}, j = 1, \ldots, k, k = 1, 2, \ldots \right\}$,
  - ▷ Y and $(K, \underline{T})$ are independent.
- $X \equiv (\Delta_K, T_K, K)$, with a possible value $x = (\delta_k, t_k, k)$, where $\Delta_k = (\Delta_{k,1}, \ldots, \Delta_{k,k})$ with $\Delta_{k,j} = 1_{(T_{k,j-1}, T_{k,j}]}(Y)$, $j = 1, 2, \ldots, k+1$.

- Suppose we observe n i.i.d. copies of X; $X_1, X_2, \ldots, X_n$, where $X_i = (\Delta^{(i)}_{K^{(i)}}, T^{(i)}_{K^{(i)}}, K^{(i)})$, $i = 1, 2, \ldots, n$. Here $(Y^{(i)}, \underline{T}^{(i)}, K^{(i)})$, $i = 1, 2, \ldots$ are the underlying i.i.d. copies of $(Y, \underline{T}, K)$.

note that conditionally on K and T_K, the vector Δ_K has a multinomial distribution:

$$(\Delta_K | K, T_K) \sim \text{Multinomial}_{K+1}(1, \Delta F_K)$$

where

$$\Delta F_K \equiv (F(T_{K,1}), F(T_{K,2}) - F(T_{K,1}), \ldots, 1 - F(T_{K,K}))\,.$$

Suppose for the moment that the distribution G_k of $(T_K|K=k)$ has density g_k and $p_k \equiv P(K=k)$. Then a density of X is given by

$$\begin{aligned} p_F(x) &\equiv p_F(\delta, t_k, k) \\ &= \prod_{j=1}^{k+1} (F(t_{k,j}) - F(t_{k,j-1}))^{\delta_{k,j}} g_k(t) p_k \end{aligned}$$

where $t_{k,0} \equiv 0$, $t_{k,k+1} \equiv \infty$. In general,

$$\begin{aligned} p_F(x) &\equiv p_F(\delta, t_k, k) \\ &= \prod_{j=1}^{k+1} (F(t_{k,j}) - F(t_{k,j-1}))^{\delta_{k,j}} \\ &= \sum_{j=1}^{k+1} \delta_{k,j}(F(t_{k,j}) - F(t_{k,j-1})) \qquad (9) \end{aligned}$$

is a density of X with respect to the dominating measure ν where ν is determined by the joint distribution of $(K, \underline{T})$, and it is this version of the density of X with which we will work throughout the rest of the example. Thus the log-likelihood function for F of $X_1, \ldots, X_n$ is given by

$$\begin{aligned} \frac{1}{n} l_n(F|\underline{X}) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K^{(i)}+1} \Delta_{K,j}^{(i)} \log \left(F(T_{K^{(i)},j}^{(i)}) - F(T_{K^{(i)},j-1}^{(i)}) \right) \\ &= \mathbb{P}_n m_F \end{aligned}$$

where

$$\begin{aligned} m_F(X) &= \sum_{j=1}^{K+1} \Delta_{K,j} \log \left(F(T_{K,j}) - F(T_{K,j-1}) \right) \\ &\equiv \sum_{j=1}^{K+1} \Delta_{K,j} \log \left(\Delta F_{K,j} \right) \end{aligned}$$

and where we have ignored the terms not involving F. We also note that

$$Pm_F(X) = P\left(\sum_{j=1}^{K+1} \Delta F_{0,K,j} \log\left(\Delta F_{K,j}\right)\right).$$

The (Nonparametric) Maximum Likelihood Estimator (MLE)

$$\widehat{F}_n = \text{argmax}_F \mathbb{P}_n \ell_n(F).$$

$\widehat{F}_n$ can be calculated via the iterative convex minorant algorithm proposed in Groeneboom and Wellner (1992) for case 2 interval censored data.

By Proposition 1 with $\alpha = 1$ and $\varphi \equiv \varphi_1$ as before, it follows that

$$h^2(p_{\widehat{F}_n}, p_{F_0}) \le (\mathbb{P}_n - P_0)\left(\varphi(p_{\widehat{F}_n}/p_{F_0})\right)$$

where φ is bounded and continuous from R to R. Now the collection of functions

$$\mathcal{G} \equiv \{p_F : \ F \in \mathcal{F}\}$$

is easily seen to be a Glivenko-Cantelli class of functions: this can be seen by first applying the GC-preservation theorem Theorem 1 to the collections $\mathcal{G}_k$, $k = 1, 2, \ldots$ obtained from $\mathcal{G}$ by restricting to the sets $K = k$. Then for fixed k, the collections $\mathcal{G}_k = \{p_F(\delta, t_k, k) : \ F \in \mathcal{F}\}$ are P_0-Glivenko-Cantelli classes since $\mathcal{F}$ is a uniform Glivenko-Cantelli class, and since the functions p_F are continuous transformations of the classes of functions $x \to \delta_{k,j}$ and $x \to F(t_{k,j})$ for $j = 1, \ldots, k+1$, and hence $\mathcal{G}$ is $P-$Glivenko-Cantelli by van de Geer's bracketing entropy bound for monotone functions. Note that single function p_{F_0} is trivially P_0- Glivenko-Cantelli since it is uniformly bounded, and the single function $(1/p_{F_0})$ is also P_0- GC since $P_0(1/p_{F_0}) < \infty$. Thus by the Glivenko-Cantelli preservation Theorem 1 with $g = (1/p_{F_0})$ and $\mathcal{F} = \mathcal{G} = \{p_F : \ F \in \mathcal{F}\}$, it follows that $\mathcal{G}' \equiv \{p_F/p_{F_0} : \ F \in \mathcal{F}\}$. Is P_0-Glivenko-Cantelli. Finally another application of preservation of the Glivenko-Cantelli property by continuous maps shows that the collection

$$\mathcal{H} \equiv \{\varphi(p_F/p_{F_0}) : \ F \in \mathcal{F}\}$$

is also P_0-Glivenko-Cantelli. When combined with Corollary 1.1, we find:

Theorem. The NPMLE $\widehat{F}_n$ satisfies

$$h(p_{\widehat{F}_n}, p_{F_0}) \rightarrow_{a.s.} 0\,.$$

To relate this result to a result of Schick and Yu (2000), it remains only to understand the relationship between their $L_1(\mu)$

and the Hellinger metric h between p_F and p_{F_0}. Let $\mathcal{B}$ denote the collection of Borel sets in R. On $\mathcal{B}$ we define measures μ and $\widetilde{\mu}$, as follows: For $B \in \mathcal{B}$,

$$\mu(B) = \sum_{k=1}^{\infty} P(K=k) \sum_{j=1}^{k} P(T_{k,j} \in B | K=k)\,, \tag{10}$$

and

$$\widetilde{\mu}(B) = \sum_{k=1}^{\infty} P(K=k) \frac{1}{k} \sum_{j=1}^{k} P(T_{k,j} \in B | K=k)\,. \tag{11}$$

Let d be the $L_1(\mu)$ metric on the class $\mathcal{F}$; thus for $F_1, F_2 \in \mathcal{F}$,

$$d(F_1, F_2) = \int |F_1(t) - F_2(t)| d\mu(t)\,.$$

The measure μ was introduced by Schick and Yu (2000); note that μ is a finite measure if $E(K) < \infty$. Note that $d(F_1, F_2)$ can also be written in terms of an expectation as:

$$d(F_1, F_2) = E_{(K,\underline{T})}\left[\sum_{j=1}^{K+1}\left|F_1(T_{K,j}) - F_2(T_{K,j})\right|\right]. \tag{12}$$

As Schick and Yu (2000) observed, consistency of the NPMLE $\widehat{F}_n$ in $L_1(\mu)$ holds under virtually no further hypotheses.

Theorem. (Schick and Yu). Suppose that $E(K) < \infty$. Then $d(\widehat{F}_n, F_0) \rightarrow_{a.s.} 0$.

Proof. We have shown that this follows from the Hellinger consistency proved above and the following lemma; see van der Vaart and Wellner (2000).

Lemma.

$$\frac{1}{2}\left\{\int |\widehat{F}_n - F_0| d\tilde{\mu}\right\}^2 \leq h^2(p_{\widehat{F}_n}, p_{F_0}) .$$

Example 3. (Completely monotone densities:)
Suppose that $\mathcal{P} = \{P_G : \ G$ a d.f. on $R\}$ where the measures P_G are scale mixtures of exponential distributions with mixing distribution G:

$$p_G(x) = \int_0^\infty ye^{-yx} dG(y)\,.$$

We first show that the map $G \mapsto p_G(x)$ is continuous with respect to the topology of vague convergence for distributions G. This follows easily since kernels for our mixing family are bounded, continuous, and satisfy $ye^{-xy} \to 0$ as $y \to \infty$ for every $x > 0$. Since vague convergence of distribution functions implies that integrals of bounded continuous functions vanishing at infinity converge, it follows that $p(x; G)$ is continuous with respect to the vague topology for every $x > 0$.

This implies, moreover, that the family $\mathcal{F} = \{p_G/(p_G + p_0) : G$ is a d.f. on $\mathbb{R}\}$ is pointwise, for a.e. x, continuous in G

with respect to the vague topology. Since the family of sub-distribution functions G on R is compact for (a metric for) the vague topology (see e.g. Bauer (1972), page 241), and the family of functions $\mathcal{F}$ is uniformly bounded by 1, we conclude from the basic bracketing lemma (Wald and LeCam) that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Thus it follows from Corollary 1.1 that the MLE $\widehat{G}_n$ of G_0 satisfies

$$h(p_{\widehat{G}_n}, p_{G_0}) \rightarrow_{a.s.} 0\,.$$

By uniqueness of Laplace transforms, this implies that $\widehat{G}_n$ converges weakly to G_0 with probability 1. This method of proof is due to Pfanzagl (1988); in this case we recover a result of Jewell (1982). See also Van de Geer (1999), Example 4.2.4, page 54.

Xièxiè!