On function-parametric empirical processes and unitary operators

Estáte V. Khmaladze,
Victoria University,
Wellington Workshop in Probability Theory
and Mathematical Statistics

Statistical roots

For an i.i.d.(F) sample $X_1, \ldots, X_n$ of in $[0,1]^d$, and its binomial process $N_n(x) = \sum_{i=1}^{n} I_{\{X_i \leq x\}}$, consider empirical process:

$$v_{nF}(x) = \frac{1}{\sqrt{n}}[N_n(x) - nF(x)], \quad x \in [0,1]^d.$$

Let $v_F(x)$ be its weak limit, i.e., F- Brownian bridge (BB).

Centered differently, $N_n(x)$ leads to another process:

$$b_{nF}(x) = \frac{1}{\sqrt{n}}[N_n(x) - \int_0^x (n - N_n(y)) \frac{F(dy)}{1 - F(y)}].$$

Let b_F denote another BM, weak limit of b_n.

Connection between Brownian bridges and Brownian Motions

Consider F-Brownian motion in unit cube $[0,1]^d$:

$$w_F(x), \text{ and } Ew_F^2(x) = F(x), \; x \in [0,1]^d,$$

along with its function-parametric version:

$$w_F(\phi) = \int_{y \in [0,1]^d} \phi(y) w_F(dy), \; \phi \in L_2(F).$$

We have two classical relationships between Brownian motion and the "corresponding" Brownian bridge.

One,

$$v_F(x) = w_F(x) - F(x)w_F(1), \quad 1 = (1, \dots, 1), \qquad (a)$$

represents v_F as the projection of BM w_F. In function-parametric form it is

$$v_F(\phi) = w_F(\phi) - \langle \phi, q \rangle_F w_F(q), \text{ and } q(x) \equiv 1$$

Another, for $d = 1$, i.e. on $[0, 1]$,

$$v_F(dx) = \frac{-v_F(x)}{1 - F(x)} F(dx) + b_F(dx), \qquad (b)$$

with b_F also F-BM. Unlike (a), this (b) is one-to-one.
b_F is the innovation martingale of v_F.

Side remark

A d-dimensional version of

$$v_F(dx) = \frac{-v_F(x)}{1 - F(x)} F(dx) + b_F(dx),\ x \in \mathbb{R}^d,$$

occurred late, Khm(1989,1993).

Cramér and Kolmogorov, 1963, Sukhumi, Georgia.

no q

For a little wile we stay with

$$v_F(\phi) = w_F(\phi) - \langle \phi, q \rangle_F w_F(q), \text{ and } q(x) \equiv 1$$

where we may add to v_F the q as an index, to stress that this our bridge is orthogonal to the function q:

$$v_F(q) = 0.$$

with q

For a little wile we stay with

$$v_F^q(\phi) = w_F(\phi) - \langle\phi, q\rangle_F w_F(q), \text{ and } q(x) \equiv 1$$

where we may add to v_F the q as an index, to stress that this our bridge is orthogonal to the function q:

$$v_F^q(q) = 0.$$

with q

For a little wile we stay with

$$v_F^q(\phi) = w_F(\phi) - \langle \phi, q \rangle_F w_F(q), \text{ and } q(x) \equiv 1$$

where we may add to v_F the q as an index, to stress that this our bridge is orthogonal to the function q:

$$v_F^q(q) = 0.$$

Besides,

$$E[v_F^q(\phi)]^2 = \|\phi\|_F^2 - \langle \phi, q \rangle^2$$

with q

What happens if we replace q by another h with $\|h\|_F = 1$?

$$v_F^h(\phi) = w_F(\phi) - \langle \phi, h\rangle w_F(h)$$

is again a projection of w_F, now parallel to h, and

$$v_F^h(h) = 0.$$

Besides,

$$E[v_F^h(\phi)]^2 = \|\phi\|_F^2 - \langle \phi, h\rangle^2$$

– almost no change.

However, in point parametric version,

$$E[v_F^q(x)]^2 = F(x) - F^2(x),$$

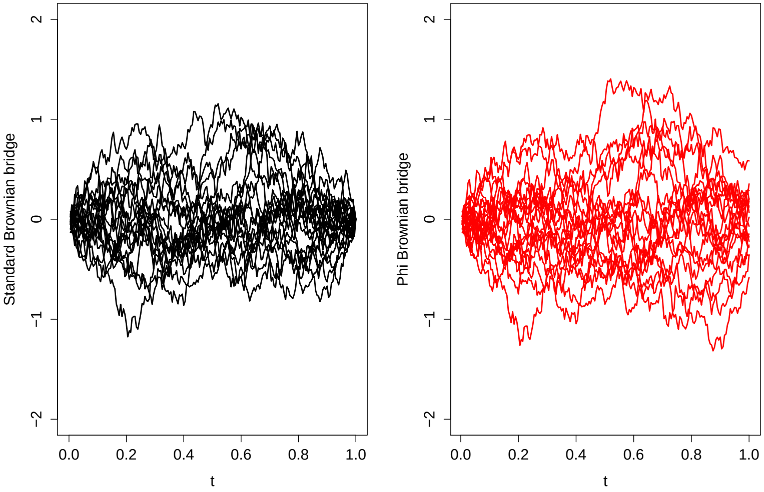
while

$$E[v_F^h(x)]^2 = F(x) - [\int^x h(y)F(dy)]^2$$

So, v_F^h is not a familiar Brownian bridge.

In many problems h can be a vector (of orthonormal function) – therefore, again, not a BB.

Trajectories of v_F^q and v_F^h



Black - usual bridges, but red..., but red – also bridges

Do we see v_F^h in reality?

. For example, when the failure rate

$$\frac{f_\theta(x)}{1 - F_\theta(x)}$$

depends on parameter, which needs to be estimated from the same sample, the process

$$\hat{b}_{nF}(x) = \frac{1}{\sqrt{n}}[N_n(x) - \int_0^x [n - N_n(y)] \frac{F_{\hat{\theta}}(dy)}{1 - F_{\hat{\theta}}(y)}]$$

converges in distribution to v_F^h with h quite different from q.

The estimated (or parametric) empirical process

$$\hat{v}_{nF}(x) = [N_n(x) - nF_{\hat{\theta}}(x)]/\sqrt{n},$$

converges to v_F^h, where h is now a vector-function.

Example with covariates

Consider pairs (X_i, Y_i) with are i.i.d.(F), $F(dx, dy) = F(dy|x)G(dx)$. However, now we do not care about the distribution G of X_i-s, and only formulate the hypothesis about $F(y|x) = F_\theta(y|x)$
Then

$$v_n(y, x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [I_{\{Y_i \leq y\}} - F_\theta(y|X_i)] I_{\{X_i \leq x\}}$$

Is not a BB at all – in x it behaves as a BM, but in y – a BB.

We call v_F^h an h-projected F-BM. If $h = q$ and F is uniform on $[0,1]^d$, then it is the just standard Brownian bridge.

New connections

Proposition 1. *Suppose F has a.e. positive density f on $[0,1]^d$. The process with the differential*

$$u(dx) = \frac{1}{\sqrt{f(x)}} v_F^q(dx) - \frac{\int_{[0,1]^d} \frac{1}{\sqrt{f(y)}} v_F^q(dy)}{1 - \int_{[0,1]^d} \sqrt{f(y)} dy} \left(1 - \sqrt{f(x)}\right) dx \qquad (1)$$

is the standard Brownian bridge.

From v_F^q to v_G^q

It is not necessary that F and G live on the unit cube. Suppose only that $G \ll F$ and denote the likelihood ratio

$$l^2(x) = \frac{dG}{dF}(x)$$

Proposition 2. *If v_F^q is F-BB, then the process, defined as the right-hand side of*

$$v_G^q(dx) = l(x)v_F^q(dx) - \int_{y\in\mathbb{R}^d} l(y)v_F(dy)\times \frac{1}{1-\int_{y\in\mathbb{R}^d} l(y)dF(y)}[l^2(x)-l(x)]f(x)dx$$

is G-BB. If $G \sim F$, then the relationship between v_G^q and v_F^q is one-to-one.

Even if G and F are discrete?

If G and F are m-dimensional discrete distributions, of course they are equivalent.

And v_G^q and v_F^q are again weak limits of the respective empirical processes,

Therefore Proposition 2 shows that the problems of testing F and testing G are equivalent.

From F-BB to b_F

Proposition 3. *Given $A \subseteq [0,1]^d$ choose $\eta_A^2(x)$ as a density on A. If the density of F is positive a.e. on $[0,1]^d$, the process*

$$b_A(dx) = \frac{v_F(dx)}{\sqrt{f(x)}} - \int_{y\in A} \eta_A(y) \frac{v_F(dy)}{\sqrt{f(y)}} \times \frac{1}{1 - \int_{y\in A} \eta_A(y)\sqrt{f(y)}dy} (\eta_A(x) - \sqrt{f(x)})dx$$

is a standard Brownian motion on $[0,1]^d \setminus A$, while

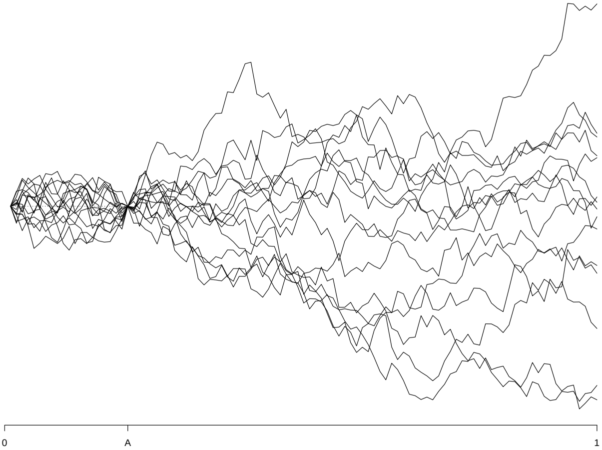
$$\int_{y\in A} \eta_A(y) b_A(dy) = 0.$$

On $[0,1]$, with $A = [0,\Delta]$ and uniform F, it means that

$$b_A(dx) = u(dx) + \frac{u(\Delta)}{\sqrt{\Delta} - \Delta}\, dx, \quad x > \Delta,$$

is standard BM.

Bundle of trajectories of b_F



A here is Δ of the previous slide and equals 0.2.

So, v_F^h is projection of w_F parallel to h, and v_G^r is projection of w_G parallel to r. Can we rotate one into the other?

Side remark

It is easy enough to connect w_F and w_G, although they live of different spaces. $L_2(F)$ and $L_2(G)$: if G and F are equivalent, $l(x) = \sqrt{dG/dF}(x)$, and if $\psi \in L_2(G)$, then $\phi = l\psi \in L_F$ and $\|\phi\|_F = \|\psi\|_G$. Hence

$$w_G(\psi) = w_F(l\psi)$$

However,

$$v_G(\psi) = v_F(l\psi)$$

is totally wrong.

To do this we use the following operator:

$$K\phi = \phi - \frac{2}{\|lr - h\|_F^2}(lr - h)\langle lr - h, \phi\rangle_F, \phi \in L_2(F).$$

This is self-adjoint unitary operator, which maps lr into h and vice versa.

Proposition 0. *If v_F^h is h-projected F-Brownian motion, then the process defined as*

$$v_G^r(\psi) = v_F^h(Kl\psi), \psi \in L_2(G),$$

is r-projected G-Brownian motion.

Components of χ^2 statistic

Consider "components" of K.Pearson's χ^2-statistic:

$$Y_{in} = \frac{\nu_{in} - np_i}{\sqrt{np_i}}, \quad Y_n = \{Y_{in}\}_{i=1}^m$$

As everybody knows,

$$\sum_{i=1}^m Y_{in}^2 = \sum_{i=1}^m \frac{(\nu_{in} - np_i)^2}{np_i}$$

is the statistic. It is 115 years old.

It is the only goodness statistic for discrete distributions.

The transformation of Y_n

Here is the unitary transformation:

$$Z_n = KY_n = Y_n - \langle Y_n, \sqrt{r}\rangle \frac{1}{1 - \langle \sqrt{p}, \sqrt{r}\rangle}(\sqrt{r} - \sqrt{p}). \qquad (2)$$

Proposition. *Let r be any m-dimensional discrete distribution: $\sum_{i=1}^{m} r_i = 1$; if*

$$Y_n \xrightarrow{d} Y = W - \langle W, \sqrt{p}\rangle \sqrt{p},$$

then

$$Z_n \xrightarrow{d} Z = W - \langle W, \sqrt{r}\rangle \sqrt{r}.$$

Therefore testing various discrete distributions of the same m are not different problems, but one single problem.

$$v_F^h(K\phi)$$

References

1. J. Doob (1949), Heuristic approach to the Kolmogorov - Smirnov theorems, *Ann. Math. Stat.*, 20, 393–403.
2. E.V. Khmaladze (2013), Note on distribution free testing for discrete distributions, *Ann. Stat,* 41, 2979– 2993.
3. T. T. M. Nguyen (2014), New approach to distribution free tests in contingency tables, Report R14-1, MSOR, Victoria University of Wellington.