Citations From References: 0 From Reviews: 0

MR4085751 62G10 60E15 62G07 62G20

Amini-Seresht, Ebrahim (IR-BASU-S); Milošević, Bojana (SE-BELGM)

New non-parametric tests for independence. (English summary)

J. Stat. Comput. Simul. **90** (2020), *no. 7,* 1301–1314.

In this article, two nonparametric tests of independence against a particular form of dependence are proposed. The alternative considered here is a stochastically increasing property (also known as positive regression dependence), which is defined as follows: For a pair of positive random variables (X, Y), Y is said to be *stochastically increasing* in X if, for all $y > 0$, $\Pr(Y > y|X = x)$ is increasing in $x > 0$.

For a given $y > 0$, let $\delta(s, t; y) := \Pr(Y > y|X = s) - \Pr(Y > y|X = t)$ whenever $s \geq t > 0$. The test statistics are built on the following measure of deviation:

$$\Delta_{(F,G)}(y) := E\left\{\delta\left(X_2, X_1; y\right)\middle|\, X_2 \geq X_1\right\},$$

where F and G are marginal distribution functions of X and Y, respectively. Given the conditional survival function $\overline{G}(y|x)$, $\Delta_{(F,G)}(y)$ can be expressed as

$$\Delta_{(F,G)}(y) = \int \overline{G}\left(y|\, x\right)\left\{2F(x) - 1\right\} dF(x).$$

The sup measure $\Delta^*_{(F,G)} := \sup_{y>0} \Delta_{(F,G)}(y)$ and the density-average measure $\Delta^{**}_{(F,G)} := \int \Delta_{(F,G)}(y)\, dG(y)$ are employed to gauge the deviation from independence. Because these measures include a few unknown quantities, the corresponding test statistics can be constructed by replacing the quantities with their consistent estimates. Specifically, using n *i.i.d.* observations, $\{(X_i, Y_i)\}_{i=1}^n$, $\overline{G}(y|x)$ can be estimated nonparametrically by

$$\overline{G}_n\left(y|\, x\right) := \frac{\sum_{j=1}^n k\left\{\left(x - X_j\right)/a_n\right\} \mathbf{1}\left(Y_j > y\right)}{\sum_{j=1}^n k\left\{\left(x - X_j\right)/a_n\right\}},$$

where $k(\cdot)$ is a kernel function, $a_n(> 0)$ is the sequence of bandwidths, and $\mathbf{1}(\cdot)$ is an indicator function. By further replacing (F, G) with their empirical measures (F_n, G_n), a natural estimator of $\Delta_{(F,G)}(y)$ can be obtained as

$$\Delta_{(F_n,G_n)}(y) := \frac{1}{n^2} \sum_{i,j,\ell} \frac{k\left\{\left(X_i - X_j\right)/a_n\right\} \mathbf{1}\left(Y_j > y\right)\left\{2(\mathbf{1}(X_i \geq X_\ell)) - 1\right\}}{\sum_{j=1}^n k\left\{\left(X_i - X_j\right)/a_n\right\}}.$$

In the end, the test statistics become

$$\Delta^*_{(F_n,G_n)} := \sup_{y>0} \Delta_{(F_n,G_n)}(y) \text{ and } \Delta^{**}_{(F_n,G_n)} := \int \Delta_{(F_n,G_n)}(y)\, dG_n(y).$$

In particular, $\Delta^{**}_{(F_n,G_n)}$ can be simplified as

$$\Delta^{**}_{(F_n,G_n)} = \frac{1}{n^3} \sum_{i,j} \frac{k\left\{\left(X_i - X_j\right)/a_n\right\}\left(S_j - 1\right)\left(2R_i - n\right)}{\sum_{j=1}^n k\left\{\left(X_i - X_j\right)/a_n\right\}},$$

where $R_i := \mathrm{Rank}(X_i) = \sum_{\ell=1}^n \mathbf{1}(X_i \geq X_\ell)$ and $S_j := \mathrm{Rank}(Y_j) = \sum_{h=1}^n \mathbf{1}(Y_j \geq Y_h)$.

Convergence properties of the test statistics $\Delta^*_{(F_n,G_n)}$ and $\Delta^{**}_{(F_n,G_n)}$ can be established by the following weak convergence of the empirical process $\{\Delta_{(F_n,G_n)}(y),\, y>0\}$:

$$\sqrt{na_n}\left\{\Delta_{(F_n,G_n)}(y)-\Delta_{(F,G)}(y)\right\}\Rightarrow \mathcal{Q}(y):=k_0^{1/2}\int\{f(x)\}^{-1/2}\{2F(x)-1\}\mathcal{B}\{G(y|x)\}\,dF(x)$$

in $D(0,\infty)$, where $k_0=\int k^2(u)\,du$ and $\mathcal{B}$ is a standard Brownian bridge on the unit interval $[0,1]$. This yields the limiting null distributions of $\Delta^*_{(F_n,G_n)}$ and $\Delta^{**}_{(F_n,G_n)}$ as

$$\sqrt{na_n}\left\{\Delta^*_{(F_n,G_n)}-\Delta^*_{(F,G)}\right\}\Rightarrow k_0^{1/2}\Gamma\sup_{y>0}\mathcal{B}\{G(y)\},\ \text{and}$$

$$k_0^{-1/2}\Gamma^{-1}\sqrt{na_n}\left\{\Delta^{**}_{(F_n,G_n)}-\Delta^{**}_{(F,G)}\right\}\Rightarrow N\left(0,\frac{1}{12}\right),$$

where $\Gamma:=\int\{f(x)\}^{-1/2}\{2F(x)-1\}\,dF(x)$. The null of independence is rejected if test statistics take large values.

Observe that the above convergence results depend on unknown marginal distributions of X and Y. Then, bootstrapping is adopted in the Monte Carlo study. However, the bootstrap algorithm does not look fully operational in the sense that it is subject to prior knowledge of the joint distribution of (X,Y). When implementing each test, we may have to either (i) estimate the unknown quantities nonparametrically and rely on its limiting null distribution or (ii) explore a nonparametric bootstrap test.

Masayuki Hirukawa