♠ On Consistent and Nonparametric ♠ ♢♡ Tests for Dependence ♡♢

Andrey Feuerverger

Department of Statistics, University of Toronto

andrey@utstat.toronto.edu

Columbia University, 2 May 2014

Some basic problems

There are <u>many</u> needs for <u>general</u> tests of dependence:

- If Y has non-monotone regression on X, and X is sampled randomly
- Testing random number generators
- Signal processing: (Are the uncorrelated coefficients of a Karhunen-Loeve expansion independent?)
- Nonlinear time series modeling / identification / estimation:

  $$X_t = h_\beta(X_{t-1}, \cdots) + \xi_t \,,$$

  (a very general model, with quite arbitrary h_β...)

  Can a test of "H_0: The ξ_t are iid" be used to estimate β?
- Testing for tail dependence (copulas...)

The classical consistent tetsts

- Hoeffding 1948, Blum, Kiefer & Rosenblatt 1961:

$$\int\int \left(F_n(x,y) - F_n^X(x)F_n^Y(y)\right)^2 dF_n(x,y)$$

  Rank test; H_0-distribution free

- Rosenblatt 1975: (based on kernel density estimates)

$$\int\int \left(f_n(x,y) - f_n^X(x)f_n^Y(y)\right)^2 w(x,y)\,dx\,dy$$

  Not distribution-free [that doesn't matter much]
  But: Such tests lack power! Why ??

- The asymptotics are incompatible: Under H_1 both $\Rightarrow N(\cdot,\cdot)$, but under H_0, they are respectively $\sum \lambda_j \xi_1^2$ and $N(\cdot,\cdot)$. Why??

Some WILD abandon ! ...

Start with $\frac{1}{n}\sum(X_i-\bar{X})(Y_i-\bar{Y})$.

Replace X_i, Y_i with 'scores' X_i' and Y_i', and write in 'U-statistic form':

$$= \frac{1}{2n^2}\sum_j\sum_k \left(X_j'-X_k'\right)\left(Y_j'-Y_k'\right)$$

Interject absolute value signs: (ignore the '2')

$$\text{So let} \quad \xi \equiv \frac{1}{n^2}\sum_j\sum_k \left|X_j'-X_k'\right|\left|Y_j'-Y_k'\right|.$$

From this, remove its 'first order Hajek projection'

$$\sum_{j=1}^{n} E_0\left[\,\xi - E_0\xi \;\middle|\; \left(X_j',Y_j'\right)\right]$$

(under the H_0-based random assignment of Y' scores to X' scores)

[It's a CLT component of form $\sum_j g(X_j',Y_j')$ which contains no information]

Wild abandon (continued)...

Computing the Hajek projection just involves a bunch of algebra.
When this projection is removed from ξ we arrive at:

$$\begin{aligned}
&\frac{1}{n^2}\sum_j\sum_k |X'_j - X'_k|\,|Y'_j - Y'_k| \\
&- \frac{2}{n(n-1)(n-2)}\sum_j\sum_\ell\sum_m |X'_j - X'_\ell|\,|Y'_j - Y'_m| \\
&+ \frac{2}{n^2(n-1)(n-2)}\sum_j\sum_k\sum_\ell\sum_m |X'_j - X'_k|\,|Y'_\ell - Y'_m|
\end{aligned}$$

- Requires only $O(n^2)$ computations. (Also: Third term is constant)
- Converges in distribution $\Rightarrow \sum \lambda_j \chi_1^2$ under H_0
- Yields a consistent, and powerful test for H_0! But why?

Approaching dependence from the Fourier domain

The cf and the ecf : $\quad \varphi(t) \equiv E\, e^{itX}\,, \quad \varphi_n(t) \equiv \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}$

$$\sup_{|t|\le T} |\varphi_n(t) - \varphi(t)| \to 0\,, \quad \text{even if } \ T \equiv T_n = \exp(o(n))$$

Since $\varphi_n(t)$ just averages iid versions of e^{itX}, we get

$$n \cdot \mathrm{Cov}\,\{\varphi_n(s), \varphi_n(t)\} \;=\; E\, e^{isX}\, \overline{e^{itX}} - E\, e^{isX} E\, \overline{e^{itX}} \;=\; \varphi(s-t) - \varphi(s)\overline{\varphi(t)}$$

• Boundedness implies asymptotic joint normality at points $(t_1, \ldots, t_k)$.
• Weak convergence of $\sqrt{n}(\varphi_n(t) - \varphi(t))$ to a Gaussian process requires only very mild moment conditions.

So why not base tests on empirical version of $\Gamma(s,t) \equiv \varphi(s,t) - \varphi^X(s)\varphi^Y(t)$:

$$\Gamma_n(s,t) \equiv \varphi_n(s,t) - \varphi_n^X(s)\varphi_n^Y(t) \quad ??$$

Inference in the Fourier domain remains efficient

By Fourier transforming, statistical efficiency is not lost:

For instance, if $X_1, \ldots, X_n$ are an iid sample from $\left\{f_\theta(x) = F'_\theta(x), \theta \in \Theta\right\}$, the MLE equations for θ can be written as

$$\int_{-\infty}^{\infty} \frac{\partial \log f_\theta(x)}{\partial \theta}\, d\left(F_n(x) - F_\theta(x)\right) \;=\; 0\,.$$

But parseval's identity $\left(\int \hat{g}h = \int g\hat{h}\right)$ allows us to view this as

$$\int_{-\infty}^{\infty} (\varphi_n(t) - \varphi_\theta(t))\, W_\theta(t)\, dt \;=\; 0\,,$$

where $W_\theta(t)$ is the inverse Fourier transform of $\partial \log f_\theta(x)/\partial\theta$.

A lesson from Goodness of fit...

Recall the Cramer-von Mises statistic

$$W_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \, dF(x) \,,$$

and the significant improvement to it due to Anderson and Darling:

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{F(x)[1 - F(x)]} \, dF(x) \,;$$

and while on the subject, the versions

$$n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{1 - F(x)} \, dF(x) \,, \quad \text{and} \quad n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{[1 - F(x)]^2} \, dF(x) \,,$$

due to Sinclair, Spurr & Ahmad (1990), and Chernobai et al (2005).

A key point is that such weights increase emphasis on tails.

Back to dependence & the Fourier domain...

Returning to the "dependence function" and the "dependence process":

$$\Gamma(s,t) \equiv \varphi(s,t) - \varphi^X(s)\varphi^Y(t) \quad \text{and} \quad \Gamma_n(s,t) \equiv \varphi_n(s,t) - \varphi_n^X(s)\varphi_n^Y(t)$$

These behave well: $E\,\Gamma_n(s,t) = \frac{n-1}{n}\Gamma(s,t)$, $\Gamma_n(s,t) \to \Gamma(s,t)$ as $n \to \infty$, etc.
And: Under the hypothesis H_0 of independence we get a simple result

$$\mathrm{Cov}_0\left\{\Gamma_n(s_1,t_2), \Gamma_n(s_2,t_2)\right\}$$

$$= \frac{n-1}{n^2}\left[\varphi^X(s_1-s_2) - \varphi^X(s_1)\overline{\varphi^X(s_2)}\right]\left[\varphi^Y(t_1-t_2) - \varphi^Y(t_1)\overline{\varphi^Y(t_2)}\right].$$

Under standard normality this implies

$$n \cdot \mathrm{Var}\,\Gamma_n(s,t) \;\to\; \left(1-e^{-s^2}\right)\left(1-e^{-t^2}\right).$$

(Remark: This simple form does not happen in higher dimensions!)

The shape of "dependence functions"

What is the shape of the functions

$$\Gamma(s,t) \equiv \varphi(s,t) - \varphi^X(s)\varphi^Y(t) \quad \text{and} \quad \Gamma_n(s,t) \equiv \varphi_n(s,t) - \varphi_n^X(s)\varphi_n^Y(t) \ \ ??$$

- $\Gamma(s,t)$ and $\Gamma_n(s,t)$ are bounded
- $\Gamma(s,t) = \overline{\Gamma(-s,-t)}$; likewise for Γ_n.
  (So any half-plane through origin is a natural domain...)
- On the $s = 0$ and/or $t = 0$ axes, $\Gamma(s,t) = 0$ and $\Gamma_n(s,t) = 0$.
- If (X,Y) has a density, then $\Gamma(s,t) \to 0$ as $||(s,t)|| \to \infty$.
- Typically $\sqrt{n}(\Gamma_n(s,t) - \Gamma(s,t)) \Rightarrow$ Gaussian process (on compacts)
  (Holds for normal scores under independence.)

Proposed test statistic

Proposed test statistics for dependence:

$$T_n \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{|\Gamma_n'(s,t)|^2}{\left(1-e^{-s^2}\right)\left(1-e^{-t^2}\right)}\, W(s,t)\, ds\, dt$$

Primes mean we are using normal scores.
Denominator $\to 0$ as $s \to 0$ or $t \to 0$, but can define ratio by continuity...
The ratio emphasizes moments/tails of the joint distribution.
Without such denominator, substantial power gets lost.
The ratio has constant variance, so need an integrable weight $W(s,t)$.

$$\text{Try}: \quad W(s,t) \;=\; \left(\frac{1-e^{-s^2}}{s^2}\right)\left(\frac{1-e^{-t^2}}{t^2}\right).$$

This $W(s,t)$ is shaped like a bell.

Proposed test statistic

With this choice for $W(s,t)$, the test statistic becomes

$$T_n \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{|\Gamma_n'(s,t)|^2}{s^2\,t^2}\,ds\,dt\,.$$

- Define by continuity at $s=0, t\neq 0$; $s\neq 0, t=0$; and at $s=0, t=0$.
- Apply 'U-statistic' identity to numerator.
- After some algebra & calculus, get:

$$T_n/\pi^2 \;=\; \frac{1}{n^2}\sum_j\sum_k \left|X_j'-X_k'\right|\left|Y_j'-Y_k'\right|$$

$$-\;\frac{2}{n^3}\sum_j\sum_\ell\sum_m \left|X_j'-X_\ell'\right|\left|Y_j'-Y_m'\right| \;+\; \frac{1}{n^4}\sum_j\sum_k\sum_\ell\sum_m \left|X_j'-X_k'\right|\left|Y_\ell'-Y_m'\right|.$$

All three terms $O(n^2)$ computable; third one is constant (depends only on n).
Szekeley gives nice/deep interpretations for expressions of this type.

Proposed test statistic

We can simplify the test statistic

$$\frac{1}{n^2}\sum_j\sum_k \left|X'_j - X'_k\right| \left|Y'_j - Y'_k\right|$$

$$-\frac{2}{n^3}\sum_j\sum_\ell\sum_m \left|X'_j - X'_\ell\right| \left|Y'_j - Y'_m\right| + \frac{1}{n^4}\sum_j\sum_k\sum_\ell\sum_m \left|X'_j - X'_k\right| \left|Y'_\ell - Y'_m\right|.$$

The second term involves repeated approximation to

$$q(x) \equiv E|x - Z| = \sqrt{2/\pi}\, e^{-x^2/2} + 2x\Phi(x) - x, \quad Z \sim N(0,1),$$

while the third term approximates the square of $E\,q(Z) = 2/\sqrt{\pi}$.

A modified/simplified test statistic is then

$$\frac{1}{n^2}\sum_j\sum_k \left|X'_j - X'_k\right| \left|Y'_j - Y'_k\right| - \frac{2}{n}\sum_j q(X'_j)q(Y'_j) + \frac{4}{\pi}$$

Proposed test statistic

Other weight functions are possible; particularly tractable are

$$W(s,t) \;=\; \left(\frac{1-e^{-s^2}}{s^2}\right)\cdot\left(\frac{1-e^{-t^2}}{t^2}\right)\cdot V(s)\cdot V(t)$$

which lead to test statistics that look like

$$\frac{1}{n^2}\sum_j\sum_k g(X'_j - X'_k)g(Y'_j - Y'_k)$$

$$-\;\frac{2}{n^3}\sum_j\sum_\ell\sum_m g(X'_j - X'_\ell)g(Y'_j - Y'_m) \;+\; \frac{1}{n^4}\sum_j\sum_k\sum_\ell\sum_m g(X'_j - X'_k)g(Y'_\ell - Y'_m).$$

Graphical tools

X and Y are independent iff $\operatorname{Cov}[f(X), g(Y)] \equiv 0$ over a separating class...

So consider plotting

$$\operatorname{Corr}\{\cos(sX),\ \cos(tY)\}$$

$$\operatorname{Corr}\{\cos(sX),\ \sin(tY)\}$$

$$\operatorname{Corr}\{\sin(sX),\ \cos(tY)\}$$

$$\operatorname{Corr}\{\sin(sX),\ \sin(tY)\}$$

Can call such plots and their variants Correlographs...

The mystery of Rosenblatt's test

Recall Rosenblatt (1975): $\int\int \left(f_n(x,y) - f_n^X(x) f_n^Y(y)\right)^2 w(x,y)\,dx\,dy.$

Why does it go wrong?

Disregard variance scaling, use unscored data, and consider

$$\int\int \left|\varphi_n(s,t) - \varphi_n^X(s)\varphi_n^Y(t)\right|^2 W(s,t)\,ds\,dt.$$

By Parseval's Theorem $\left(\int |g|^2 = c\int |\hat{g}|^2\right)$, this is same as

$$\int\int \left|\int\int V(x-u, y-v)\,d^2\left[F_n(u,v) - F_n^X(u)F_n^Y(v)\right]\right|^2 dx\,dy$$

where $V(x,y)$ is the Fourier transform of $\sqrt{W(s,t)}$.
The inner $\int\int$ is a convolution, but here bandwidth stays constant.

Rosenblatt's bandwidth $\to 0$ (for consistent density estimation),
AND ignores variance scaling.

The curse of *small* dimensions

The case of dimension $k > 2$ is not straightforward.

- Factorization $n \cdot \mathrm{Var}\,\Gamma_n(s,t) \to (1 - e^{-s^2})(1 - e^{-t^2})$ doesn't extend.
- The 'U-statistic' factoring trick fails
- For some insights, see Deheuvels' 1981 paper *'An asymptotic decomposition for multivariate distribution-free tests of dependence'*.

Szekeley (et al) provide effective extensions to higher dimensions, etc.

The copula point of view

Dependence is a copula property (Hoeffding, Sklar, etc.)

$$F(x,y) = C(F(x), F(y))$$

Copula $C(u,v)$ is a cdf with uniform marginals; density is $\partial^2 C(u,v)/\partial u \partial v$.

Given data $\{(X_i, Y_i), i = 1, \ldots, n\}$, having ranks $\{(R_i, S_i), i = 1, \ldots n\}$, we can estimate $C(u,v)$ by the empirical cdf $C_n(u,v)$ of the $(R_i/n, S_i/n)$.

One can test for dependence by comparing $C_n(u,v)$ with $C(u,v) = uv$,

- presumably at the points $(R_i/n, S_i/n)$, and
- presumably by taking (co)variability into account.

Presumably, this is the same as examining $F_n(x,y) - F_n^X(x) F_n^Y(y)$ at the points $(x,y) = (X_i, Y_i)$...

Can one focus such tests on tail dependence?

A wavelet point of view (!)

Antoniadis, Feuerverger & Goncalves, 2006 (Wavelet-based estimation for univariate stable laws, *Ann. Inst. Statist. Math.*) considered fitting stable laws via <u>wavelet transforms on the cf and ecf</u>:

Say $\{\psi_{j,k}\}$ is a complete wavelet family, and $\{\hat{\psi}_{j,k}\}$ their Fourier transforms. Idea is based on a generalized (nonlinear) least squares:

$$Y_{j,k} \equiv \langle dF_n, \hat{\psi}_{j,k} \rangle \equiv \frac{1}{n}\sum_{\ell}^{n} \hat{\psi}_{j,k}(X_\ell) \equiv \langle c_n, \psi_{j,k} \rangle \quad = \quad \langle f_\theta, \hat{\psi}_{j,k} \rangle + \varepsilon_{j,k}$$

Parseval's identity $[\ \int f\hat{g} = \int \hat{f} g\ ]$ lets us rewrite this as

$$Y_{j,k} \ = \ \langle c_\theta, \psi_{j,k} \rangle + \varepsilon_{j,k}$$

Here $\{f_\theta\}$, $\{c_\theta\}$ are the (4-parameter) stable law densities, and cf's.

- Covariance structure of $\varepsilon_{j,k}$ (*Re* and *Im* parts) has closed form.
- Wavelet computation is numerically (very) efficient

A wavelet point of view continued

$$Y_{j,k} \;=\; \langle c_\theta, \psi_{j,k} \rangle + \varepsilon_{j,k}$$

- Indices j, k involve translating and scaling a mother wavelet
- Use many j and k
- Method has arbitrarily high asymptotic efficiency (due to completeness)
- Complements MLE (which is hard to implement when densities don't have closed forms – as in convolutions)
- Novelty is the use of wavelets for a parametric inference
- Wavelets disbalance, tend to diagonalize correlation: reduce ill-conditioning
- Rearranges Fisher information, concentrates it into few coefficients
- Wavelets can 'zero in' at the origin and get at tail structure.

So what about testing for dependence ??

A wavelet point of view continued

So what about testing for dependence ??

Need to use multivariate wavelet transforms.
With wavelets we can zero in on tail/moment structure. BUT:

- Which wavelets should we use? Which j and k ?
- Should we apply them to $\varphi_n(u,v) - \varphi_n^X(u)\varphi_n^Y(v)$?
- Or should we apply them to $(\varphi_n(u,v) - \varphi_n^X(u)\varphi_n^Y(v))/(t^2u^2)$??
- Should we use transformed data: i.e. the φ' ???
  (And if so, which marginals are best ?)
- Computational details, covariance structures, etc., etc. ...
- Durbin-Knott type 'components' of the test...

These are research questions !!

A reference

Feuerverger, A. (1993).
A consistent test for bivariate dependence.
International Statistical Review, 61, 3, 419-433.

THANK YOU !