# Stats 319 Review: Distance correlation

(AKA Brownian distance covariance)
Székely, Rizzo, and Bakirov (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, **35**, 2769-94.

Haben Michael

April 15, 2015

# Outline

Let R(X,Y) be a measure of the independence of two random variables X and Y. Desirable properties could include:

1. R(X,Y)=0 iff X and Y are independent
2. R invariant under transformations $(X, Y) \mapsto (\epsilon X, \epsilon Y), \epsilon > 0$
3. X and Y may have arbitrary (finite) dimension

# Outline

- Pearson product-moment correlation

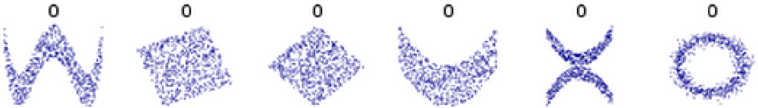
$$\rho = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}}, \quad X, Y \in \mathbb{R}$$

$$\hat{\rho} = \sum_i^n (X_i - \bar{X})(Y_i - \bar{Y}) / \sqrt{\sum_i^n (X_i - \bar{X})^2 \sum_i^n (Y_i - \bar{Y})^2}$$

- Pearson product-moment correlation

$$\rho = \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}}, \quad X, Y \in \mathbb{R}$$

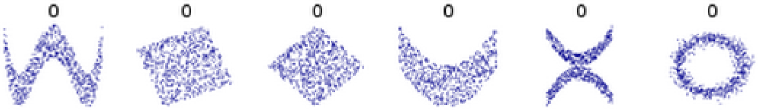
$$\hat{\rho} = \sum_i^n (X_i - \bar{X})(Y_i - \bar{Y}) / \sqrt{\sum_i^n (X_i - \bar{X})^2 \sum_i^n (Y_i - \bar{Y})^2}$$



- Pearson product-moment correlation

$$\rho = \frac{Cov[X, Y]}{\sqrt{Var[X]Var[Y]}}, \quad X, Y \in \mathbb{R}$$

$$\hat{\rho} = \sum_i^n (X_i - \bar{X})(Y_i - \bar{Y}) / \sqrt{\sum_i^n (X_i - \bar{X})^2 \sum_i^n (Y_i - \bar{Y})^2}$$




- doesn't meet 1st property for X,Y not jointly normal

- rank-based methods, e.g., spearman correlation—product-moment on ranks $X_i \mapsto \hat{X}_i := rank(X_i)$, $Y_i \mapsto \hat{Y}_i := rank(Y_i)$

$$\hat{\rho}_{spearman} = \sum_i^n (\hat{X}_i - \bar{\hat{X}})(\hat{Y}_i - \bar{\hat{Y}}) / \sqrt{\sum_i^n (\hat{X}_i - \bar{\hat{X}})^2 \sum_i^n (\hat{Y}_i - \bar{\hat{Y}})^2}$$

- rank-based methods, e.g., spearman correlation—product-moment on ranks $X_i \mapsto \hat{X}_i := rank(X_i)$, $Y_i \mapsto \hat{Y}_i := rank(Y_i)$

$$\hat{\rho}_{spearman} = \sum_i^n (\hat{X}_i - \bar{\hat{X}})(\hat{Y}_i - \bar{\hat{Y}}) / \sqrt{\sum_i^n (\hat{X}_i - \bar{\hat{X}})^2 \sum_i^n (\hat{Y}_i - \bar{\hat{Y}})^2}$$

- doesn't meet 1st property for non-monotone X,Y relationships

Renyi correlation (Rényi 1959)

$$sup\left\{corr(f(X), g(Y)) : f \in L_2(X), g \in L_2(Y)\right\}$$

- $= 0$ iff independent
- $= 1$ implies $\mathbb{P}[f(X) = g(Y)] = 1$ for some "non-trivial" functions $f, g$
- $= |\rho|$ for bivariate normal $X, Y$ with correlation $\rho$
- but, much harder to approximate

# Outline

- Distance correlation exploits that the characteristic functions of X and Y factor,

$$f_{X,Y}(t,s) = f_X(t)f_Y(s) \qquad X, t \in \mathbb{R}^p, Y, s \in \mathbb{R}^q,$$

  iff X and Y are independent, whatever the dimension of X and Y

  $(f_{X,Y}(t,s) := \mathbb{E}exp(itX + isY), f_X(t) := \mathbb{E}exp(itX), f_Y(s) := \mathbb{E}exp(isY))$

- Distance correlation exploits that the characteristic functions of X and Y factor,

$$f_{X,Y}(t,s) = f_X(t)f_Y(s) \qquad X,t \in \mathbb{R}^p, Y,s \in \mathbb{R}^q,$$

  iff X and Y are independent, whatever the dimension of X and Y

  $(f_{X,Y}(t,s) := \mathbb{E}exp(itX + isY), f_X(t) := \mathbb{E}exp(itX), f_Y(s) := \mathbb{E}exp(isY))$
- Use a weighted $L2$ distance on $\mathbb{R}^{p+q}$ between the LHS and RHS

## Definition

$$\begin{aligned}\mathcal{V}^2(X,Y;w) &= ||f_{X,Y}(t,s) - f_X(t)f_Y(s)||_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)dtds\end{aligned}$$

## Definition

$$
\begin{aligned}
\mathcal{V}^2(X, Y; w) &= ||f_{X,Y}(t,s) - f_X(t)f_Y(s)||_w^2 \\
&= \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)dtds \\
\mathcal{V}^2(X; w) = \mathcal{V}^2(X, X; w) &= ||f_{X,X}(t,s) - f_X(t)f_X(s)||_w^2 \\
&= \int_{\mathbb{R}^{p+q}} |f_{X,X}(t,s) - f_X(t)f_X(s)|^2 w(t,s)dtds
\end{aligned}
$$

## Definition

$$
\begin{aligned}
\mathcal{V}^2(X, Y; w) &= ||f_{X,Y}(t,s) - f_X(t)f_Y(s)||_w^2 \\
&= \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)dtds \\
\mathcal{V}^2(X; w) = \mathcal{V}^2(X, X; w) &= ||f_{X,X}(t,s) - f_X(t)f_X(s)||_w^2 \\
&= \int_{\mathbb{R}^{p+q}} |f_{X,X}(t,s) - f_X(t)f_X(s)|^2 w(t,s)dtds \\
\mathcal{R}(X, Y; w) &= \frac{\mathcal{V}(X,Y;w)}{\sqrt{V(X;w)V(Y;w)}}
\end{aligned}
$$

- taking weights $w(t,s) = c_p c_q |t|_p^{1+p} |s|_q^{1+q}$, $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$
- non-negative, $\mathcal{R}^2(X, Y; w) = 0$ iff $X$ and $Y$ are independent

- Where does this choice of weights come from?
  - $w$ should be non-integrable. Otherwise it can be shown that $\lim_{\epsilon \to \infty} \mathcal{R}^2(\epsilon X, \epsilon Y; w) \to \rho^2(X, Y)$
  - 2005 result, $\int_{\mathbb{R}^d} (1 - cos\langle t, x\rangle)/|t|^{d+1} dt = c_d|x|$
  - good properties (below) "only the weight functions [$w$] lead to distance covariance type statistics" meaning/proof?

- Where does this choice of weights come from?
  - $w$ should be non-integrable. Otherwise it can be shown that $\lim_{\epsilon \to \infty} \mathcal{R}^2(\epsilon X, \epsilon Y; w) \to \rho^2(X, Y)$
  - 2005 result, $\int_{\mathbb{R}^d} (1 - cos\langle t, x\rangle)/|t|^{d+1} dt = c_d|x|$
  - good properties (below) "only the weight functions [$w$] lead to distance covariance type statistics" meaning/proof?
- The choice is interesting because it is what appears to distinguish Székeley & Rizzo from prior work using characteristic functions to measure independence. Bickel mentions:
  - Feuerverger, A. and Mureika, R.A. (1977). The empirical characterisic function and its applications. *Ann. Statistic.* **5** 88-97.
  - Chen, A. and Bickel, P.J. (2005).

# Outline

On the other hand, simple to compute statistics are part of the appeal to applied statisticians

- compute Euclidean distances within samples

$$\begin{aligned} a_{kl} &:= |X_k - X_l| \\ b_{kl} &:= |Y_k - Y_l| \end{aligned}$$

- and arrange in a matrix

$$\left[\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & a_{1\cdot} \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & a_{n\cdot} \\ \hline a_{\cdot 1} & \dots & a_{\cdot n} & a_{\cdot\cdot} \end{array}\right]$$

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} - a_{1.} & \dots & a_{1n} - a_{1.} \\ \vdots & \ddots & \vdots \\ a_{n1} - a_{n.} & \dots & a_{nn} - a_{n.} \end{bmatrix}$$

$$\begin{bmatrix} a_{11} - a_{1.} - a_{.1} + a_{..} & \dots & a_{1n} - a_{1.} - a_{.n} + a_{..} \\ \vdots & \ddots & \vdots \\ a_{n1} - a_{n.} - a_{.n} + a_{..} & \dots & a_{nn} - a_{n.} - a_{.n} + a_{..} \end{bmatrix} =: \underset{n\times n}{(A_{kl})}$$

and analogously for $B_{kl}$, giving 2 centered interpoint distance matrices

## Definition

$$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n^2} \sum_{k,l} A_{kl} B_{kl}$$

$$\mathcal{V}_n^2(\boldsymbol{X}) = \mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{X}) = \frac{1}{n^2} \sum_{k,l} A_{kl}^2$$

$$\mathcal{R}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y})}{\sqrt{\mathcal{V}_n^2(\boldsymbol{X})\mathcal{V}_n^2(\boldsymbol{Y})}}$$

where $\underset{n \times p}{\boldsymbol{X}}, \underset{n \times q}{\boldsymbol{Y}}$ contain $n$ samples iid $\overset{d}{=} X, Y$

It turns out that this definition of $\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y})$ is equivalent to the empirical formulation of $\mathcal{V}(X, Y)$:

**Theorem**

$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n^2}\sum_{k,l} A_{kl}B_{kl} = ||f_{X,Y}^n(t, s) - f_X^n(t)f_Y^n(s)||_w^2$

Here $f_{X,Y}^n(t, s) := \frac{1}{n}\sum_k^n exp(i\langle t, X_k\rangle + i\langle s, Y_k\rangle)$ is the empirical characteristic function of the sample $(\boldsymbol{X}, \boldsymbol{Y})$, and analogously for $f_X^n(t), f_Y^n(s)$

**Theorem**

*If $X$ and $Y$ are integrable, then almost surely*
$\lim_{n\to\infty} \mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{V}(X, Y)$ *and* $\lim_{n\to\infty} \mathcal{R}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{R}(X, Y)$

### Theorem

*If $X$ and $Y$ are integrable, then almost surely*
$$\lim_{n\to\infty} \mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{V}(X, Y) \text{ and } \lim_{n\to\infty} \mathcal{R}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{R}(X, Y)$$

$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y})$ is biased for $\mathcal{V}^2(X, Y)$ (later work)

- Worked example: $X \sim bernoulli(p), Y \sim bernoulli(q)$ iid

$$A_{kl} = \begin{cases} -2\bar{X}^2, & X_k = X_l = 0 \\ -2(1-\bar{X})^2, & X_k = X_l = 1 \\ -2(\bar{X} - \bar{X}^2), & \text{otherwise} \end{cases}$$

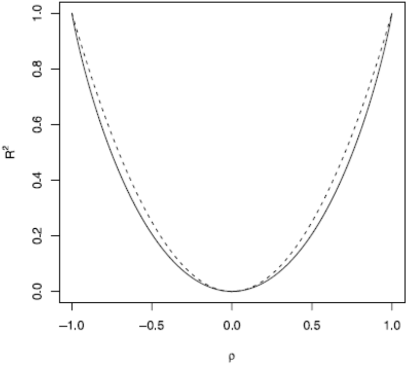
$$\underset{n \times n}{(A_{kl})} = \frac{-2}{n}(\bar{X}\mathbb{1} - X)(\bar{X}\mathbb{1}^T - X^T) =: \frac{-2}{n} u_x u_x^T$$

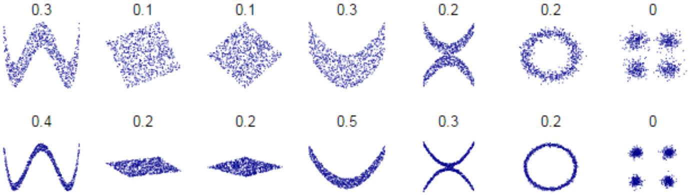
$$n^2 \mathcal{V}_n^2(X, Y) = \sum_{k,l} A_{kl} B_{kl} = \frac{4}{n^2} tr(u_x u_x^T u_y u_y^T) = \frac{4}{n^2} tr(u_y^T u_x u_x^T u_y)$$

$$= 4\left(\frac{1}{n}(Y - \mathbb{1}\bar{Y})^T (X - \mathbb{1}\bar{X})\right)^2 \propto Cov_{MLE}(X, Y)^2$$

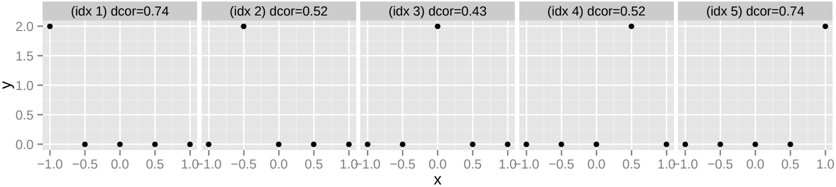
- More generally, $|X_k - X_l| \mapsto (X_k - X_l)^2$ reduces the statistics $\mathcal{V}_n, \mathcal{R}_n$ to MLEs for usual covariance and correlation (but then we lose empirical ch. fn. formulation, scale-free property, etc.)

- For bivariate normal $X, Y$ with correlation $\rho$, $\mathcal{R}^2(X, Y)$ (the population parameter) is a function of $\rho$
- Theorem: $inf_{\rho \neq 0} \frac{\mathcal{R}(X,Y)}{|\rho|} = (4(1 + \pi/3 - \sqrt{3}))^{-1/2} \approx .89$

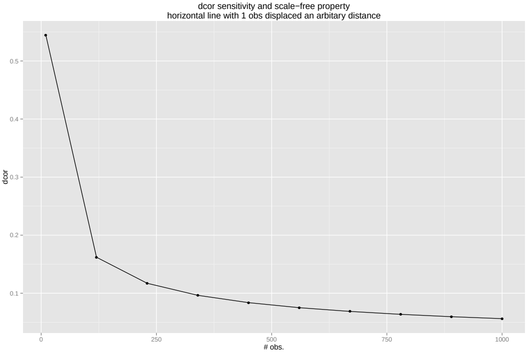





(Source: Wikipedia/Michael Newton paper)




- scale-free property: $\mathcal{R}^n(\epsilon \boldsymbol{X}, \epsilon \boldsymbol{Y}) = \mathcal{R}^n(\boldsymbol{X}, \boldsymbol{Y})$
- $\mathcal{R}^n(\boldsymbol{X}, \epsilon \boldsymbol{Y}) = \mathcal{R}^n(\boldsymbol{X}, \boldsymbol{Y})$, might want something continuous at $\mathcal{R}^n = 0$




$$X = (1, \ldots, n), Y = (0, \ldots, 0, d)$$
$$\sum A_{ij} B_{ij} = dn(n-1)/2, \ \sum B_{ij}^2 = nd^2, \ \sum A_{ij}^2 = \sum (i-j)^2 = n^2(n-1)^2/6, \ \mathcal{R}_n^2(X, Y) = \frac{dn(n-1)}{2} \sqrt{\frac{1}{nd^2} \frac{6}{n^2(n-1)^2}} = \Theta(1/\sqrt{n})$$

# Outline

| | |
|---|---|
| $\mathcal{V}^2(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ $(\Rightarrow \mathcal{V}^2(\boldsymbol{X}) \geq 0)$ | $Cov[X] \geq 0$ |
| $\mathcal{V}^2(X) = 0$ iff $X = \mathbb{E}[X]$ a.s. | $Cov[X] = 0$ iff $X = \mathbb{E}[X]$ a.s. |
| $\mathcal{V}^2(a + bQX) = b^2\mathcal{V}^2(X)$ for $Q^TQ = I$ | $Cov[a + bAX] = b^2 ACov[X]A^T$ |
| $\mathcal{V}(X + Y) \leq \mathcal{V}(X) + \mathcal{V}(Y)$ for $X, Y$ independent | $Cov[X + Y] = Cov[X] + Cov[Y]$ |
| $0 \leq \mathcal{R}(X, Y) \leq 1$ | $-1 \leq \rho \leq 1$ |

| | |
|---|---|
| $\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) \geq 0 \; (\Rightarrow \mathcal{V}_n^2(\boldsymbol{X}) \geq 0)$ | $Cov_n[X] \geq 0$ |
| $\mathcal{V}_n^2(\boldsymbol{X}) = 0 \Rightarrow X_i = X_j, \forall i, j$ | $Cov[X] = 0 \Rightarrow X_i = X_j$ |
| $\mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}) = 1 \Rightarrow span(\boldsymbol{X}) = span(\boldsymbol{Y})$ | |
| $\mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}) = 1 \Rightarrow \boldsymbol{Y} = a + b\boldsymbol{X}Q$ | $\hat{\rho}(X, Y) = 1 \Rightarrow Y = a + bX$ |
| some $a \in \mathbb{R}^q, b \in \mathbb{R}, Q$ orthogonal | |

# Outline

## Theorem

*If $X$ and $Y$ are integrable, then*

- *under independence,* $n\mathcal{V}_n^2/(a_{\cdot\cdot}b_{\cdot\cdot}) \rightsquigarrow Q = \sum_1^\infty \lambda_j Z_j^2$ *for* $Z_j$ *iid standard normal and* $\{\lambda_j\}$ *nonegative constants depending on the distributions of* $X$ *and* $Y$ *s.t.* $\mathbb{E}[Q] = 1$
  - *the test rejecting when* $n\mathcal{V}_n^2/(a_{\cdot\cdot}b_{\cdot\cdot}) > z_{1-\alpha/2}^2$ *is asymptotically level* $\alpha$
- *if* $X$ *and* $Y$ *are dependent,* $n\mathcal{V}_n^2/(a_{\cdot\cdot}b_{\cdot\cdot}) \to \infty$ *in probability*

the authors recommend permutation/randomization to get a reference distribution for small samples (used in R package)

# Outline

"Brownian distance covariance"

- $X$ is an $\mathbb{R}^n$-valued RV and $U$ a random process on $R^n$
- Denote $X_U = U(X) - \mathbb{E}[U(X)|U]$
- Consider $\mathbb{E}[X_U X'_U Y_V Y'_v]$
  - $= \mathcal{V}^2(X, Y)$ when $U, V$ are two independent copies of brownian motion (conditions...)
  - $= Cov^2(X, Y)$ when $U, V$ are the identity

1. Székely, Rizzo, and Bakirov (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, **35**, 2769-94.
2. Székely and Bakirov (2008). Brownian covariance and CLT for stationary sequences. TR 08-01, Dept Math. and Statistics, Bowling Green State Univ.
3. Discussion papers in *Annals of Applied Statistics* 2009, Vol. 3, No. 4.

The 2007 paper is mostly self-contained except for a few places:

- $\int_{\mathbb{R}^d} (1 - cos\langle t, x \rangle)/|t|^{d+1} dt = c_d |x|$ in Székely, Rizzo (2005)
- $\mathbb{P}[Q \geq z^2_{1-\alpha/2}] \leq \alpha$ in Székely, Rizzo (2003)
- $||\zeta||^2 \stackrel{d}{=} \sum^{\infty} \lambda_j Z_j^2$ for certain Gaussian processes, $\zeta$, results on V-statistics