Measuring Association Between Random Vectors

Oliver Grothe[1] Julius Schnieders[1] Johan Segers[2]

[1]University of Cologne
Department of Economic and Social Statistics

[2]Université catholique de Louvain
Institut de statistique, biostatistique et sciences actuarielles

ESSEC La Défense, May 19, 2014

Association as a particular form of dependence

Positive/Negative association between random variables:

- Large values of one variable tend to go together with large/small values of the other one
- Particular form of *dependence*
  - If X is symmetrically distributed around 0 and if $Y = X^2$, then X and Y are perfectly dependent but not associated.

Of obvious practical interest:

- risk concentration/diversification

Quantifying association via linear correlation: the marginal distributions come into play

Example

If

$$X \sim N(0,1), \qquad Y = X^2$$

then

$$\begin{aligned} \operatorname{cov}(X, Y) &= 0 \\ \operatorname{cov}(e^X, Y) &= \sqrt{e} > 0 \\ \operatorname{cov}(-e^{-X}, Y) &= -\sqrt{e} < 0 \end{aligned}$$

Still, both functions $f_1(x) = e^x$ and $f_2(x) = -e^{-x}$ are increasing: what is "large" and what is "small"?

Reasons to measure association in a margin-free way

We might want to measure association in a margin-free way if:

- The state spaces of X and Y are not directly comparable
  - Decathlon: high jump versus 100 metres
- Outliers have a disproportionately large influence
  - Very large positive or negative returns on asset prices

Margin-free: invariant w.r.t. increasing transformations

Examples:

- SPEARMAN's rho (1904)
- GINI's gamma (1914)
- KENDALL's tau (1938)
- BLOMQVIST's beta (1950)
- ...

[surveys in Joe (1990, 1997), Schmid and Schmidt (2007), Schmid, Schmidt, Blumentritt, Gaisser and Ruppert (2010)]

Association within and between random vectors

Association between scalar outcomes:

- *within* the bivariate random vector $\boldsymbol{Z} = (X, Y)$
- *between* random variables X and Y

Possible generalizations to vector-valued outcomes:

- *within* a d-dimensional random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$
- *between* random vectors $\boldsymbol{X} = (X_1, \ldots, X_p)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ $\rightarrow$ *This talk*
- *between* random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d$

Quantifying association between random vectors

- Classical measures of association:
  - Canonical correlation [Hotelling 1936]
  - RV coefficient [Escoufier 1973, Robert and Escoufier 1976]

  However:
  - positive values only: no distinction between positive/negative association
  - linear multivariate analysis
- Testing for independence between two random vectors
  - Matrices of Spearman and Kendall coefficients [El Maache and Lepage 2003]
  - Distances between densities [Székely, Rizzo and Bikorov 2007; Székely and Rizzo 2009]
  - Distances between copulas [Quessy 2010]

  Again: no 'direction' for the association

Inference is to be rank-based and nonparametric

Invariance w.r.t. increasing transformations of the components

- No assumptions on existence of moments required
- Inference is based on vectors of ranks

Not even a parametric model for the ‘dependence’ (copula)

- Ideal for data exploration
- Nonparametric inference

Contributions

1. Measures of association between random vectors
   - distinction between positive and negative associations
   - margin free: invariance with respect to increasing transformations
   - model free
2. Nonparametric, rank-based inference procedures
   - point estimators taking the form of U-statistics
   - consistent and asymptotically normal

Measuring Association Between Random Vectors

Measuring Association Between Random Vectors

Measuring Association Between Random Vectors

Covariance and positive quadrant dependence

Random variables X and Y are *positive quadrant dependent* if

$$P(X \leqslant x,\ Y \leqslant y) \geqslant P(X \leqslant x)\, P(Y \leqslant y), \qquad \forall x, y \in \mathbb{R}$$

[Lehmann 1966]

If X and Y have finite second moments, then

$$\operatorname{cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{P(X \leqslant x,\ Y \leqslant y) - P(X \leqslant x)\, P(Y \leqslant y)\}\, \mathrm{d}x\, \mathrm{d}y$$

[Lehmann 1966, acknowledging Hoeffding 1940]

Lemma

X and Y are positive quadrant dependent iff

$$\operatorname{cov}\{f(X), g(Y)\} \geqslant 0 \qquad \forall \textit{non-decreasing } f,\ g$$

[Esary, Proschan and Walkup 1967]

Measuring association in a margin-free way: Weighing the margins

Choose measures μ and ν on $\mathbb{R}^2$ and consider

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(X \leqslant x,\ Y \leqslant y)\,\mathrm{d}\mu(x,y)$$
$$-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(X \leqslant x)\,P(Y \leqslant y)\,\mathrm{d}\nu(x,y)$$

μ	ν	assocation measure
Lebesgue	Lebesgue	covariance
$P_{X,Y}$	$P_X \otimes P_Y$	proportional to Kendall's tau
$P_X \otimes P_Y$	$P_X \otimes P_Y$	proportional to Spearman's rho

Special case: Kendall's tau

If (X_1, Y_1) and (X_2, Y_2) are iid (X, Y), then

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(X \leqslant x,\ Y \leqslant y)\, \mathrm{d}P(X \leqslant x, Y \leqslant y) \\
&- \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(X \leqslant x)\, P(Y \leqslant y)\, \mathrm{d}P(X \leqslant x)\, \mathrm{d}P(Y \leqslant y) \\
&= \mathrm{cov}(\mathbf{1}\{X_1 \leqslant X_2\},\ \mathbf{1}\{Y_1 \leqslant Y_2\})
\end{aligned}$$

Standardizing to obtain a result in $[-1, 1]$, we find Kendall's tau:

$$\begin{aligned}
\tau(X, Y) &= \mathrm{cor}(\mathbf{1}\{X_1 \leqslant X_2\},\ \mathbf{1}\{Y_1 \leqslant Y_2\}) \\
&= 4\,\mathrm{cov}(\mathbf{1}\{X_1 \leqslant X_2\},\ \mathbf{1}\{Y_1 \leqslant Y_2\})
\end{aligned}$$

provided the distributions of X and Y are continuous.

Properties of Kendall's tau

Special values:

$\tau = 1$ iff $Y = f(X)$ with f increasing

$\tau = 0$ if (but not only if) X and Y are independent

$\tau = -1$ iff $Y = f(X)$ with f decreasing

Compare: $\text{cor}(X, Y) = \pm 1$ iff $Y = aX + b$ with $\text{sign}(a) = \pm 1$

Margin free: If f and g are increasing, then

$$\tau(X, Y) = \tau\big(f(X), g(Y)\big)$$

Special case: Spearman's rho

By the Hoeffding/Lehmann covariance formula,
provided $F(x) = P(X \leqslant x)$ and $G(y) = P(Y \leqslant y)$ are continuous:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{P(X \leqslant x,\ Y \leqslant y) - P(X \leqslant x)\, P(Y \leqslant y)\}\, \mathrm{d}F(x)\, \mathrm{d}G(y)$$

$$= \mathrm{cov}\{F(X), G(Y)\}$$

Standardizing to obtain a result in $[-1, 1]$, we find Spearman's rho:

$$\rho_S(X, Y) = \mathrm{cor}\{F(X), G(Y)\}$$

$$= 12\ \mathrm{cov}\{F(X), G(Y)\}$$

Enjoys similar properties as Kendall's tau

Measuring Association Between Random Vectors

For random vectors, quantify association in a similar way

Random vectors $\boldsymbol{X}$ in $\mathbb{R}^p$ and $\boldsymbol{Y}$ in $\mathbb{R}^q$. Distribution functions

$$\begin{aligned} H(\boldsymbol{x},\boldsymbol{y}) &= P(\boldsymbol{X} \leqslant \boldsymbol{x},\ \boldsymbol{Y} \leqslant \boldsymbol{y}), \\ F(\boldsymbol{x}) &= P(\boldsymbol{X} \leqslant \boldsymbol{x}), \\ G(\boldsymbol{y}) &= P(\boldsymbol{Y} \leqslant \boldsymbol{y}) \end{aligned}$$

Measuring association: compare $H(\boldsymbol{x},\boldsymbol{y})$ and $F(\boldsymbol{x})\,G(\boldsymbol{y})$:

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^q} H(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\mu(\boldsymbol{x},\boldsymbol{y}) - \int_{\mathbb{R}^p}\int_{\mathbb{R}^q} F(\boldsymbol{x})\,G(\boldsymbol{y})\,\mathrm{d}\nu(\boldsymbol{x},\boldsymbol{y})$$

for appropriate choices of μ and ν.

A generalization of Kendall's tau: just as in the bivariate case

If (X_1, Y_1) and (X_2, Y_2) are iid (X, Y), then

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)\,\mathrm{d}H(x,y) \\
&-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x)\,G(y)\,\mathrm{d}F(x)\,\mathrm{d}G(y) \\
&= \mathrm{cov}(\mathbf{1}\{X_1 \leqslant X_2\},\, \mathbf{1}\{Y_1 \leqslant Y_2\})
\end{aligned}$$

Standardizing, we obtain a multivariate analogue of Kendall's tau:

$$\begin{aligned}
\tau(X, Y) &= \mathrm{cor}(\mathbf{1}\{X_1 \leqslant X_2\},\, \mathbf{1}\{Y_1 \leqslant Y_2\}) \\
&= \frac{\mathrm{cov}(\mathbf{1}\{X_1 \leqslant X_2\},\, \mathbf{1}\{Y_1 \leqslant Y_2\})}{\sqrt{\mathrm{var}(\mathbf{1}\{X_1 \leqslant X_2\})\ \mathrm{var}(\mathbf{1}\{Y_1 \leqslant Y_2\})}}
\end{aligned}$$

If $p + q \geqslant 3$, the standardization depends on F and G.

Generalizing Spearman's rho: the joint survival function comes into play

Let $(\boldsymbol{X}_1, \boldsymbol{Y}_1), (\boldsymbol{X}_2, \boldsymbol{Y}_2), (\boldsymbol{X}_3, \boldsymbol{Y}_3)$ be iid $(\boldsymbol{X}, \boldsymbol{Y})$. Then

$$\begin{aligned}\iint \{H(\boldsymbol{x},\boldsymbol{y}) - F(\boldsymbol{x})\, G(\boldsymbol{y})\}\, \mathrm{d}F(\boldsymbol{x})\, \mathrm{d}G(\boldsymbol{y}) &= \operatorname{cov}(\mathbf{1}\{\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2\},\, \mathbf{1}\{\boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_3\}) \\ &= \operatorname{cov}\big(\bar{F}(\boldsymbol{X}),\, \bar{G}(\boldsymbol{Y})\big)\end{aligned}$$

$$\begin{aligned}\iint \{\bar{H}(\boldsymbol{x},\boldsymbol{y}) - \bar{F}(\boldsymbol{x})\, \bar{G}(\boldsymbol{y})\}\, \mathrm{d}F(\boldsymbol{x})\, \mathrm{d}G(\boldsymbol{y}) &= \operatorname{cov}(\mathbf{1}\{\boldsymbol{X}_1 \geqslant \boldsymbol{X}_2\},\, \mathbf{1}\{\boldsymbol{Y}_1 \geqslant \boldsymbol{Y}_3\}) \\ &= \operatorname{cov}\big(F(\boldsymbol{X}),\, G(\boldsymbol{Y})\big)\end{aligned}$$

in terms of the *joint survival functions*

$$\begin{aligned}\bar{H}(\boldsymbol{x},\boldsymbol{y}) &= P(\boldsymbol{X} \geqslant \boldsymbol{x},\, \boldsymbol{Y} \geqslant \boldsymbol{y}), \\ \bar{F}(\boldsymbol{x}) &= P(\boldsymbol{X} \geqslant \boldsymbol{x}), \\ \bar{G}(\boldsymbol{y}) &= P(\boldsymbol{Y} \geqslant \boldsymbol{y})\end{aligned}$$

If $p \geqslant 2$, it is no longer true that $\bar{F}(\boldsymbol{x})$ equals $1 - F(\boldsymbol{x})$ etc.

Generalizations of Spearman's rho: pick your choice

Several alternative generalizations of Spearman's rho:

$$\begin{aligned}\rho(\boldsymbol{X},\boldsymbol{Y}) &= \operatorname{cor}\big(F(\boldsymbol{X}),\, G(\boldsymbol{Y})\big)\\ &= \frac{\operatorname{cov}\big(F(\boldsymbol{X}),\, G(\boldsymbol{Y})\big)}{\sqrt{\operatorname{var} F(\boldsymbol{X})\ \operatorname{var} G(\boldsymbol{Y})}}\end{aligned}$$

$$\begin{aligned}\bar{\rho}(\boldsymbol{X},\boldsymbol{Y}) &= \operatorname{cor}\big(\bar{F}(\boldsymbol{X}),\, \bar{G}(\boldsymbol{Y})\big)\\ &= \frac{\operatorname{cov}\big(\bar{F}(\boldsymbol{X}),\, \bar{G}(\boldsymbol{Y})\big)}{\sqrt{\operatorname{var} \bar{F}(\boldsymbol{X})\ \operatorname{var} \bar{G}(\boldsymbol{Y})}}\end{aligned}$$

$$\rho^*(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{2}\big(\rho(\boldsymbol{X},\boldsymbol{Y}) + \bar{\rho}(\boldsymbol{X},\boldsymbol{Y})\big)$$

Again, if $p + q \geqslant 3$, the standardizations depend on F and G.

Measuring Association Between Random Vectors

Measuring Association Between Random Vectors

Kendall's tau can be represented in terms of estimable parameters of degree two

Let $(\boldsymbol{X}_1, \boldsymbol{Y}_1), (\boldsymbol{X}_2, \boldsymbol{Y}_2)$ be iid $(\boldsymbol{X}, \boldsymbol{Y})$. We have

$$\begin{aligned}\tau(\boldsymbol{X}, \boldsymbol{Y}) &= \operatorname{cor}(\mathbf{1}\{\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2\}, \mathbf{1}\{\boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_2\}) \\ &= \frac{p_{\boldsymbol{X},\boldsymbol{Y}} - p_{\boldsymbol{X}}\, p_{\boldsymbol{Y}}}{\sqrt{p_{\boldsymbol{X}}\,(1 - p_{\boldsymbol{X}})\, p_{\boldsymbol{Y}}\,(1 - p_{\boldsymbol{Y}})}}\end{aligned}$$

where

$$\begin{aligned}p_{\boldsymbol{X},\boldsymbol{Y}} &= P(\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2,\ \boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_2), \\ p_{\boldsymbol{X}} &= P(\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2), \\ p_{\boldsymbol{Y}} &= P(\boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_2)\end{aligned}$$

The latter three probabilities can be estimated using U-statistics of degree $m = 2$.

[Hoeffding 1948; van der Vaart 1998; Quessy, Saïd and Favre 2012]

U-statistics: generalizations of sample means

Let $X_1, \ldots, X_m$ be iid X, in some space $\mathscr{X}$.
Aim: estimate a 'parameter' θ of the form

$$\theta = \mathrm{E}[g(X_1, \ldots, X_m)]$$

for some known function $g : \mathscr{X}^m \to \mathbb{R}$.

- Example: $p_X = P(X_1 \leqslant X_2) = \mathrm{E}[\mathbf{1}(X_1 \leqslant X_2)]$

The U-statistic estimator for θ based on a sample $X_1, \ldots, X_n$ is

$$\hat{\theta}_m = \frac{1}{n(n-1)\cdots(n-m+1)} \sum g(X_{i_1}, \ldots, X_{i_m})$$

Sum over all permutations $(i_1, \ldots, i_m)$ of $\{1, \ldots, n\}$ of length m

Hoeffding's decomposition theorem: Linear expansion of a U-statistic

If $\mathrm{E}[g^2(X_1, \ldots, X_m)] < \infty$, then

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{m}{\sqrt{n}} \sum_{i=1}^{n} h_1(X_i) + o_p(1)$$

where

$$h_1(x_1) = \mathrm{E}[g_{\mathrm{sym}}(x_1, X_2, \ldots, X_m)] - \theta$$

$$g_{\mathrm{sym}}(x_1, \ldots, x_m) = \frac{1}{m!} \sum g(x_{i_1}, \ldots, x_{i_m})$$

[Hoeffding 1948]

Asymptotic normality

Hoeffding's decomposition yields joint asymptotic normality of a vector of U-statistics, e.g.

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, m^2\sigma_1^2) \qquad (n \to \infty)$$

$$\sigma_1^2 = \operatorname{var} h_1(X_1)$$

The asymptotic (co)variance(s) can be estimated consistently by

- U-statistics
- jackknife [Lee 1990]
- the sample (co)variance of

$$\hat{h}_{1,n}(X_i), \qquad i = 1, \ldots, n$$

$$\hat{h}_{1,n}(x) = U\text{-statistic}$$

Measuring Association Between Random Vectors

Kendall's tau can be represented in terms of estimable parameters of degree two

Let $(\boldsymbol{X}_1, \boldsymbol{Y}_1), (\boldsymbol{X}_2, \boldsymbol{Y}_2)$ be iid $(\boldsymbol{X}, \boldsymbol{Y})$. We have

$$\begin{aligned}\tau(\boldsymbol{X}, \boldsymbol{Y}) &= \operatorname{cor}(\mathbf{1}\{\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2\}, \mathbf{1}\{\boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_2\}) \\ &= \frac{p_{\boldsymbol{X},\boldsymbol{Y}} - p_{\boldsymbol{X}}\, p_{\boldsymbol{Y}}}{\sqrt{p_{\boldsymbol{X}}\,(1 - p_{\boldsymbol{X}})\, p_{\boldsymbol{Y}}\,(1 - p_{\boldsymbol{Y}})}}\end{aligned}$$

where

$$\begin{aligned}p_{\boldsymbol{X},\boldsymbol{Y}} &= P(\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2,\ \boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_2), \\ p_{\boldsymbol{X}} &= P(\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2), \\ p_{\boldsymbol{Y}} &= P(\boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_2)\end{aligned}$$

The latter three probabilities can be estimated using U-statistics of degree $m = 2$.

[Hoeffding 1948; van der Vaart 1998; Quessy, Saïd and Favre 2012]

U-statistic estimator for Kendall's tau

Represent $p_{\boldsymbol{X},\boldsymbol{Y}}$ as an estimable parameter of degree $m = 2$:

$$\begin{aligned} p_{\boldsymbol{X},\boldsymbol{Y}} &= P(\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2,\ \boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_2) \\ &= \mathrm{E}[\underbrace{\mathbf{1}(\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2)\,\mathbf{1}(\boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_2)}_{=g((\boldsymbol{X}_1,\boldsymbol{Y}_1),(\boldsymbol{X}_2,\boldsymbol{Y}_2))}] \end{aligned}$$

The corresponding U-statistic estimator is

$$\hat{p}_{\boldsymbol{X},\boldsymbol{Y};n} = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{1}(\boldsymbol{X}_i \leqslant \boldsymbol{X}_j)\,\mathbf{1}(\boldsymbol{Y}_i \leqslant \boldsymbol{Y}_j)$$

Similarly for $p_{\boldsymbol{X}} = p_{\boldsymbol{X},\boldsymbol{X}}$ and $p_{\boldsymbol{Y}} = p_{\boldsymbol{Y},\boldsymbol{Y}}$. Finally, put

$$\hat{\tau}_n(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\hat{p}_{\boldsymbol{X},\boldsymbol{Y},n} - \hat{p}_{\boldsymbol{X},n}\,\hat{p}_{\boldsymbol{Y},n}}{\sqrt{\hat{p}_{\boldsymbol{X},n}\,(1 - \hat{p}_{\boldsymbol{X},n})\,\hat{p}_{\boldsymbol{Y},n}\,(1 - \hat{p}_{\boldsymbol{Y},n})}}$$

The estimator is asymptotically normal

Hoeffding's decomposition yields joint asymptotic normality of

$$\sqrt{n}\begin{pmatrix} \hat{p}_{\boldsymbol{X},\boldsymbol{Y},n} - p_{\boldsymbol{X},\boldsymbol{Y}} \\ \hat{p}_{\boldsymbol{X},n} - p_{\boldsymbol{X}} \\ \hat{p}_{\boldsymbol{Y},n} - p_{\boldsymbol{Y}} \end{pmatrix}$$

with explicit 3×3 covariance matrix Σ.
From the delta method, we get asymptotic normality of

$$\sqrt{n}\big(\hat{\tau}_n(\boldsymbol{X},\boldsymbol{Y}) - \tau(\boldsymbol{X},\boldsymbol{Y})\big)$$

The expression for the asymptotic variance is longish, but explicit, and can be estimated consistently.

Spearman's rho can be estimated via a function of a vector of U-statistics of degree three

Recall

$$\begin{aligned}\rho(\boldsymbol{X}, \boldsymbol{Y}) &= \operatorname{cor}\bigl(\bar{F}(\boldsymbol{X}),\, \bar{G}(\boldsymbol{Y})\bigr) \\ &= \frac{\operatorname{cov}\bigl(\bar{F}(\boldsymbol{X}),\, \bar{G}(\boldsymbol{Y})\bigr)}{\sqrt{\operatorname{var} \bar{F}(\boldsymbol{X})\ \operatorname{var} \bar{G}(\boldsymbol{Y})}}\end{aligned}$$

If $(\boldsymbol{X}_1, \boldsymbol{Y}_1), (\boldsymbol{X}_2, \boldsymbol{Y}_2), (\boldsymbol{X}_3, \boldsymbol{Y}_3)$ are iid $(\boldsymbol{X}, \boldsymbol{Y})$, then

$$\begin{aligned}\operatorname{cov}\bigl(\bar{F}(\boldsymbol{X}),\, \bar{G}(\boldsymbol{Y})\bigr) &= \operatorname{cov}\bigl(\mathbf{1}(\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2),\, \mathbf{1}(\boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_3)\bigr) \\ &= P(\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2,\, \boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_3) - P(\boldsymbol{X}_1 \leqslant \boldsymbol{X}_2)\, P(\boldsymbol{Y}_1 \leqslant \boldsymbol{Y}_3)\end{aligned}$$

Similarly for $\operatorname{var} \bar{F}(\boldsymbol{X})$ and $\operatorname{var} \bar{G}(\boldsymbol{Y})$.

All probabilities occurring in these expressions can be estimated by U-statistics of degree $m \in \{2, 3\}$.

Measuring Association Between Random Vectors

Measuring Association Between Random Vectors

Write the formulas in such a way as to reduce the number of nested loops

- A naive implementation of the U-statistics formulas requires multiple nested loops over the sample:
  - all permutations $(i_1, \ldots, i_m)$ of $\{1, \ldots, n\}$ of length m

  Patience needed if $n = 1\,000$ and $m = 3\ldots$
- In case of τ and ρ, a clever rewriting of the formulas reduces the calculations to at most double loops
  - Compare with the formulas for the sample variance:

$$\hat{\sigma}_n^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2}(X_i - X_j)^2 \qquad \text{double loop}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \qquad \text{single loop}$$

Perform the comparisons in a smart order

Estimators involve statistics of the form

$$\sum_{i=1}^{n} \sum_{j:j\neq i} \mathbf{1}\{\underbrace{X_{i,1} \leqslant X_{j,1}, \ldots, X_{i,p} \leqslant X_{j,p}}_{\boldsymbol{X}_i \leqslant \boldsymbol{X}_j}\}$$

For a given $i = 1, \ldots, n$:

1. Let $R_{i,k}$ be the rank of $X_{i,k}$ within $X_{1,k}, \ldots, X_{n,k}$
2. Let $k(i)$ be the component of $\boldsymbol{X}_i$ with the highest rank
3. Restrict the inner sum to those j such that $R_{j,k(i)} \geqslant R_{i,k(i)}$

Requires preliminary sorting of the data: $O(n \log n)$ algorithm

Implement the core computations in a low-level programming language

Even after all these tricks, nested loops are required: $O(n^2)$ steps

- Painful in a simulation study when n is large

Solution: do the computations in a low-level programming language

- for instance C or Fortran
- to be called from Matlab or R for convenience
- faster than computing directly in Matlab or R, even if vectorized

Measuring Association Between Random Vectors

Aims of the simulation study

To answer the following questions:

- Finite-sample bias and variance of the estimators of the association measures τ and ρ^*?
- Accuracy of the estimators of the standard errors?
- Influence of a number of factors:
  - Model
  - Dimension
  - Sample size
- Computational feasibility

Set-up

Models:
- multivariate Gaussian
- D-vine copula with bivariate Clayton margins

[Clayton 1978; Joe 1997; Bedford and Cooke 2001, 2002]

Sample sizes: $n = 50, 100, 1\,000$

Dimensions: $p = q = 3$ and $p = q = 4$

Number of repetitions: $10\,000$

'True' values: based on 30 repetitions of samples of size $50\,000$

Results: Everything works as it should

```
Imagine here some
huge multi-way tables
with tons of digits
```

- $\hat{\tau}_n$ has lower standard error and bias than $\hat{\rho}_n^*$
- Finite-sample bias, decreasing with sample size
  - Bias is highest for $\hat{\rho}_n^*$ at weak dependence and low sample size
- Standard deviations are estimated accurately
  for sample sizes above 100

Results: Everything works as it should

```
Imagine here some
huge multi-way tables
with tons of digits
```

- $\hat{\tau}_n$ has lower standard error and bias than $\hat{\rho}^*_n$
- Finite-sample bias, decreasing with sample size
  - Bias is highest for $\hat{\rho}^*_n$ at weak dependence and low sample size
- Standard deviations are estimated accurately
  for sample sizes above 100

Measuring Association Between Random Vectors

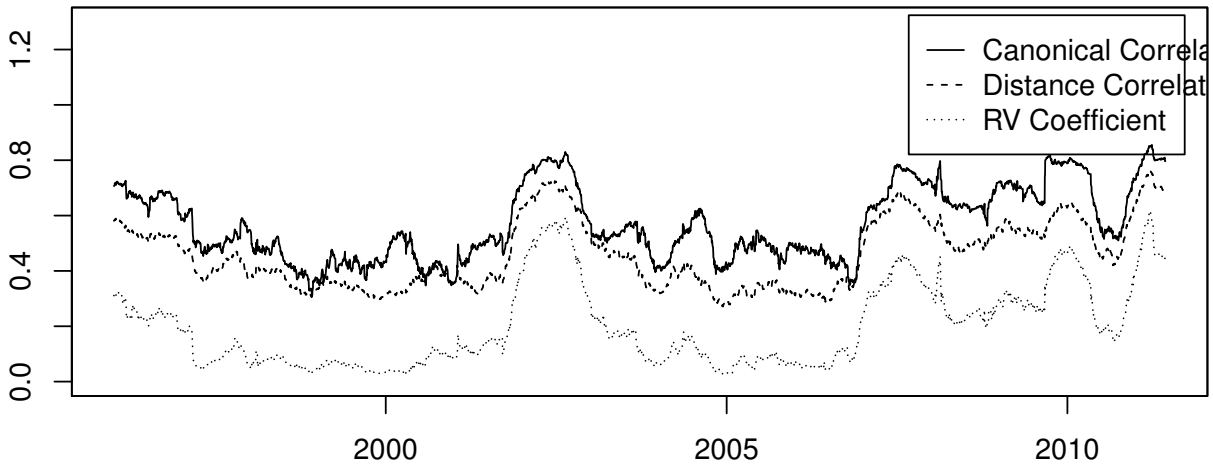
Stocks versus Bonds

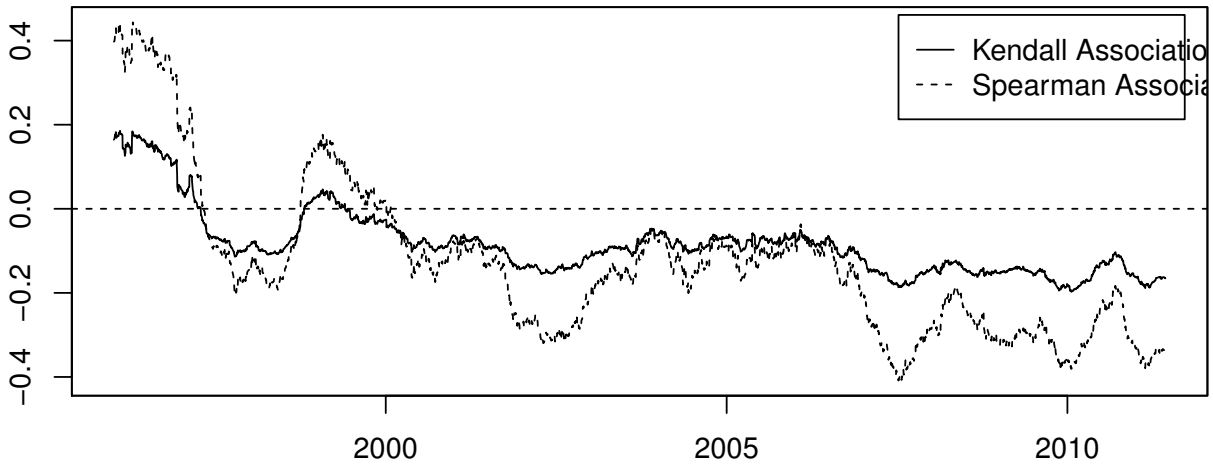
- In terms of crises, association between stocks and bonds can become negative
  - Investors searching for save havens like bonds of strong countries

[Gulko 2002, Ilmanen 2003]

- Data: daily returns
  - stock market indices of 5 major countries
    - All Ordinaries, CAC 40, DAX, Nikkei 225, S&P 500
  - government bond indices for the same countries
    - indices from The Bank of America Merrill Lynch
  - From January 3, 1996, to November 15, 2012
  - Forward looking moving window of 150 days

From positive to negative assocation

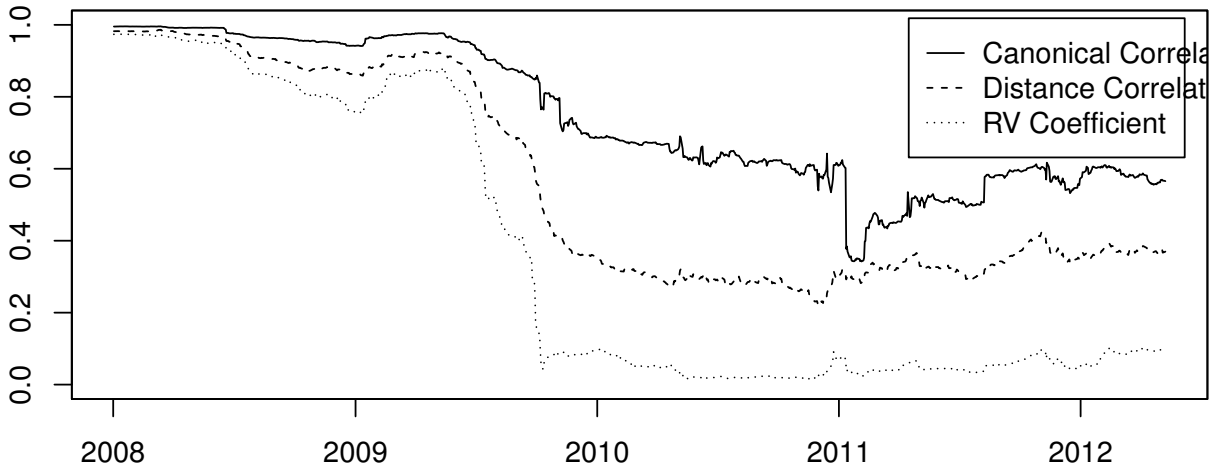


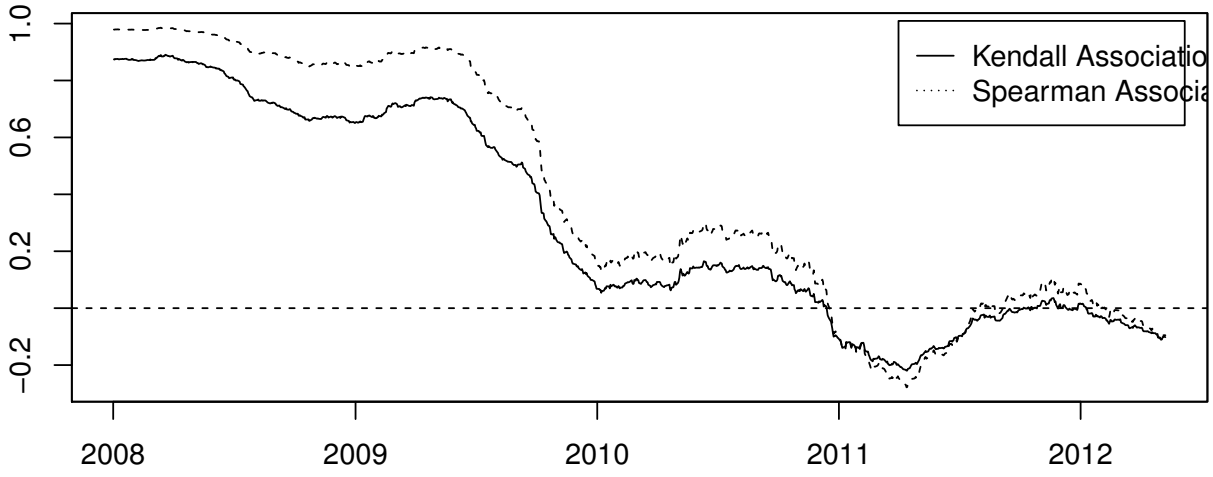


North versus South

- European sovereign debt crisis
- Association between north and south European bond markets changes as credit-worthiness of countries evolve
- Data: daily returns of Merrill Lynch government bond indices
  - North: France, Germany, the Netherlands
  - South: Italy, Portugal, Spain
  - From January 1, 2007 to November 15, 2012
  - Forward looking moving window of 150 days

With the crisis, association becomes negative





Measuring Association Between Random Vectors

Contributions

1. Measures of association between random vectors
   - positive/negative associations
   - margin free (copula-based)
   - model free
2. Nonparametric, rank-based inference procedures
   - U-statistics > consistent, asymptotically normal
3. Numerical examples
   - Finite-sample performance is satisfactory
   - Case studies demonstrate potential applicability

O. Grothe, J. Schnieders, J. Segers (2013)
Journal of Multivariate Analysis 123, 96–110