# Quantifying And Estimating Asymmetric Dependence

**Wolfgang Trutschnig**[1]
(joint work with **Florian Griessenberger**[1] and **Robert R. Junker**[2])

10th International Workshop on Simulation and Statistics

Salzburg, 2019-09-02

[1] Department for Mathematics, University of Salzburg
[2] Department of Ecology and Evolution, University of Salzburg

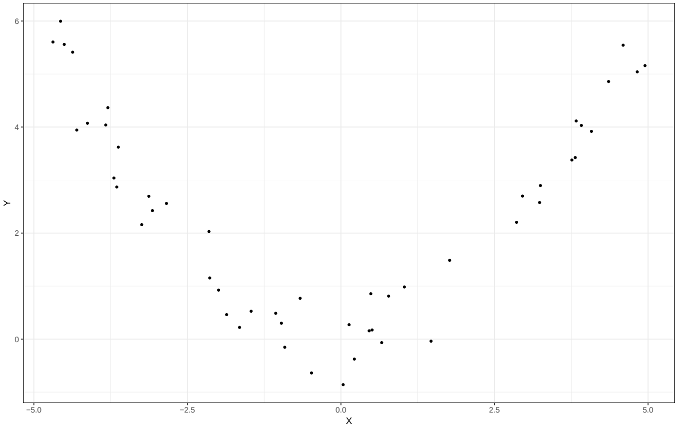



Figure: Bivariate sample of $(X, Y)$ of size $n = 50$

- Which variable is easier to predict given the value of the other one, and why?
- This talk is about one approach to estimate asymmetry for 2d samples.

- The following question arose in the context of an applied project (offer optimization in supermarkets and cannibalism effects) in 2010:
- Is there a non-parametric, scale-free version $\zeta$ of $R^2$ that quantifies the dependence of a r.v. $Y$ on a r.v. $X$ and vice versa?
- Desired natural properties:
  - $\zeta(X,Y) \in [0,1]$.
  - $\zeta(X,Y)$ is scale-free.
  - $\zeta(X,Y) = 0$ iff $X \perp Y$.
  - $\zeta(X,Y) = 1$ if $Y = \varphi(X)$ for some measurable $\varphi$ [a.k.a. $Y$ is completely dependent on $X$].
  - $\zeta(Y,X) \neq \zeta(X,Y)$ is possible.
- None of the standard 'dependence measures' I found in the literature 2010 fulfilled these properties.
- Schweitzer and Wolff's $\sigma(X,Y)$ can be arbitrarily small although $Y$ is completely dependent on $X$, the same is true for Spearman's $\rho$ and Kendall's $\tau$.
- What to do?

- Let's concentrate on continuous random variables $X, Y$.
- Focus on the copula $A$ underlying $(X, Y)$ and work with conditional distributions of $Y$ given $X$ and vice versa.
- In other words: Work with the Markov kernel $K_A(x, E)$ of the copula $A$.
- If $\mu_A$ denotes the doubly stochastic measure corresponding to $A$ then we have

$$\mu_A(E \times F) = \int_E K_A(x, F) d\lambda(x)$$

  for all $E, F \in \mathcal{B}([0, 1])$.
- A copula is called completely dependent, if there exists a $\lambda$-preserving transformation $h : [0, 1] \to [0, 1]$ such that $\mu_A(\Gamma(h)) = 1$ (or, equivalently, if all conditional distributions are degenerated).
- $\mathcal{C}$...family of all copulas; $\mathcal{C}_d$ family of all completely dependent copulas.
- Markov kernels can be used to construct metrics stronger than the uniform one $d_\infty$.

$$D_\infty(A,B) := \sup_{y\in[0,1]} \int_{[0,1]} \Big| K_A(x,[0,y]) - K_B(x,[0,y]) \Big| \, d\lambda(x)$$

$$D_1(A,B) := \int_{[0,1]} \int_{[0,1]} \Big| K_A(x,[0,y]) - K_B(x,[0,y]) \Big| \, d\lambda(x)\, d\lambda(y)$$

- $D_1(A,B)$ is the expected $L^1$-distance of the conditional distribution functions.

## Theorem (T., JMAA, 2011)

*Suppose that $A, A_1, A_2, \ldots$ are copulas. Then the following three conditions are equivalent:*

(a) $\lim_{n\to\infty} D_1(A_n, A) = 0$.

(b) $\lim_{n\to\infty} D_\infty(A_n, A) = 0$.

(c) *The corresponding Markov operators $T_{A_n}$ converge to $T_A$ in the strong operator topology $L^1([0,1], \mathcal{B}([0,1]), \lambda)$.*

## Theorem (T., JMAA, 2011)

*The metric space $(\mathcal{C}, D_1)$ is complete and separable. No closed ball $\overline{B}_{D_1}(A, r)$ with $A \in \mathcal{C}$ and $r > 0$ is compact. The family $\mathcal{C}_d$ is closed (but not compact). Convergence w.r.t. $D_1$ implies pointwise/uniform convergence but no vice versa.*

## Theorem (T., JMAA, 2011)

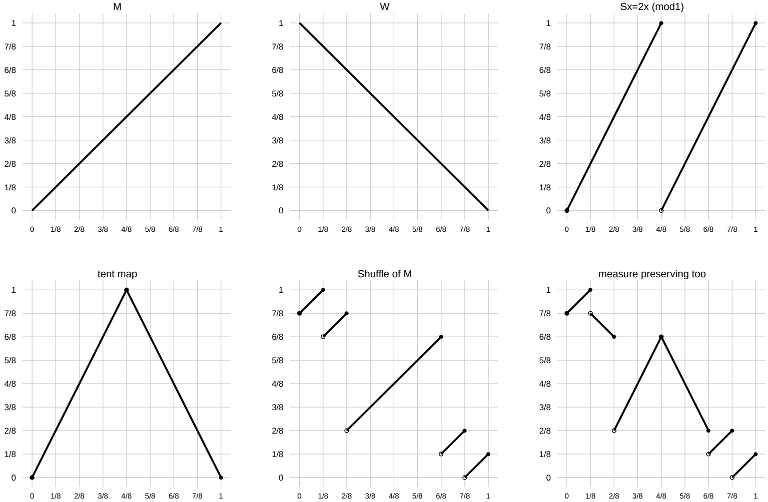
*The following assertions hold for every $A \in \mathcal{C}$:*

1. $D_1(A, \Pi) \leq 1/3$.
2. $D_1(A, \Pi) = 1/3$ *if and only if* $A \in \mathcal{C}_d$.

- Define the dependence measure $\zeta_1 : \mathcal{C} \to [0, 1]$ by

$$\zeta_1(A) := 3\, D_1(A, \Pi).$$

- $\zeta_1(A) = 0$ if and only if $A = \Pi$ (independence)
- $\zeta_1(A) = 1$ if and only if $A \in \mathcal{C}_d$ (complete dependence).


M
W
Sx=2x (mod1)
tent map
Shuffle of M
measure preserving too
0
1/8
2/8
3/8
4/8
5/8
6/8
7/8
1

## Example (Farlie-Gumbel-Morgenstern Familie)

- The FGM family $(G_\theta)_{\theta \in [-1,1]}$ is defined by
$$G_\theta(x,y) = xy + \theta xy(1-x)(1-y).$$
- $G_\theta$ is absolutely continuous and $K_{G_\theta}(\cdot,\cdot)$, given by
$$K_{G_\theta}(x,[0,y]) := y + \theta y(1-2x)(1-y) \quad \forall (x,y) \in [0,1]^2,$$
is the corresponding Markov kernel.
- $(G_\theta)_{\theta \in [-1,1]}$ is continuous in $\theta$ w.r.t. $D_1$ and we have
$$\zeta_1(G_\theta) = \frac{|\theta|}{4}$$
for every $\theta \in [-1,1]$.

- The metric $D_1$ has several other nice properties and has been extended to the multivariate setting in 2014 (Fernández Sánchez & T., JTP, 2015).
- The dependence measure $\zeta_1$ is not straightforward to extend $\rightarrow$ open work.
- 2017: Discussion with Robert Junker (professor for ecology in Salzburg) on ways to quantify the influence of one species on other ones.
- Check if a species is an influencer or is being influenced more by others.
- Natural idea: Try to estimate $\zeta_1(X, Y) = \zeta_1(A)$ based on samples of $(X, Y)$.
- Plug-in the empirical copula $\hat{E}_n$ and use $\zeta_1(\hat{E}_n)$ as estimator, done?!

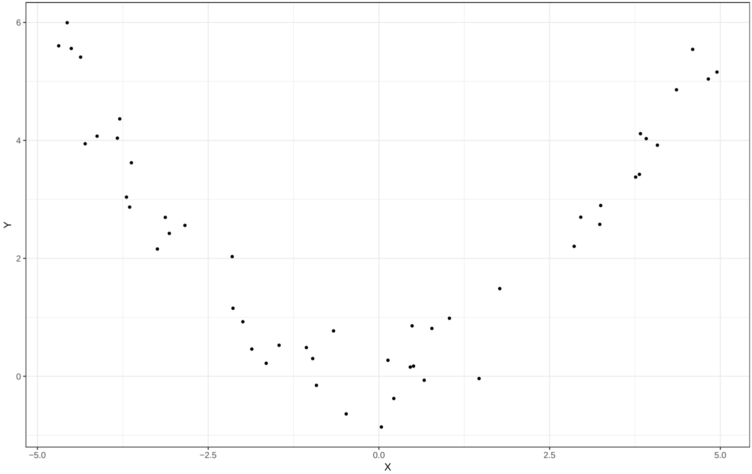



Figure: Bivariate sample of $(X, Y)$ of size $n = 50$.

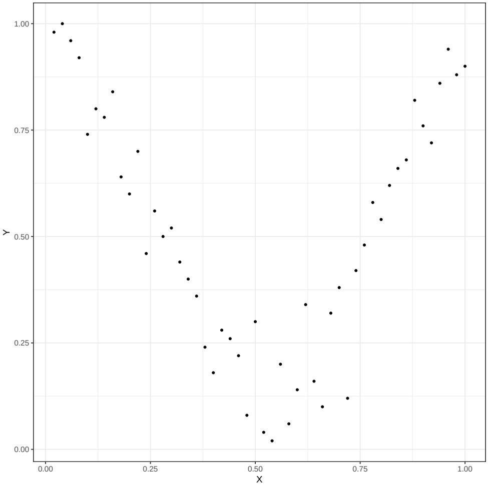



Figure: Normalized ranks of the sample.

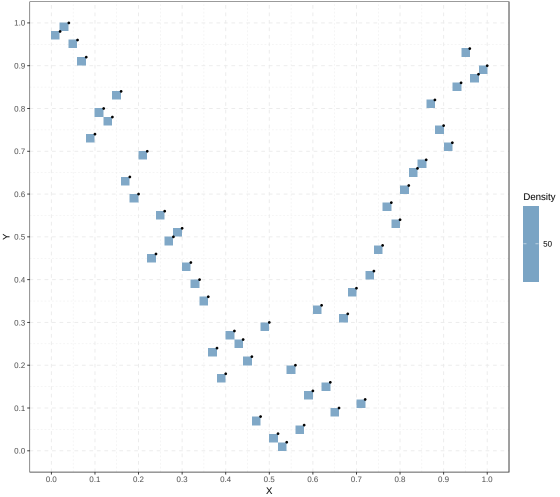



Figure: Empirical copula $\hat{E}_n$ (uniform density on each of the little squares).

- In our case we get $\zeta_1(\hat{E}_n) \sim 1$.
- $\hat{E}_n$ almost looks like a shuffle...
- Substituting the filled square with little copies of the minimum copula $M$ yields a completely dependent copula $\hat{E}_n^M$ (a.k.a. empirical checkmin copula), so $\zeta(\hat{E}_n^M) = 1$.
- The same is true for all empirical copulas:
- If $\hat{E}_n$ is the empirical copula of a sample of $(X, Y)$ and $X, Y$ are continuous then
$$\lim_{n \to \infty} \zeta_1(\hat{E}_n) = 0 \ [\mathbb{P}].$$
- Long story short: The plug-in estimator does not work.
- Estimating conditional distributions is a difficult endeavor.
- $D_1$ and $\zeta_1$ are based on conditional distributions...
- Possible way out: Aggregate/Smooth $\hat{E}_n$.

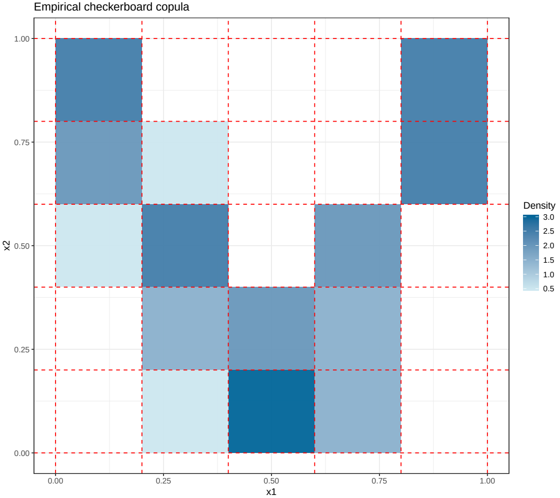


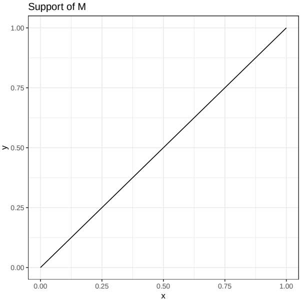

Figure: Density of the empirical checkerboard approximation $\mathfrak{CB}_5(\hat{E}_n)$ of $\hat{E}_n$. Plugging in $\mathfrak{CB}_5(\hat{E}_n)$ yields $\zeta_1(\mathfrak{CB}_5(\hat{E}_n)) = q_n(X,Y) = 0.8$; Flipping $X$ and $Y$ yields $q_n(Y,X) = 0.43$.

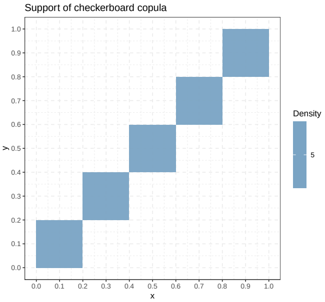
## Definition

Suppose that $A \in \mathcal{C}, N \in \mathbb{N}$. The absolute continuous copula $\mathfrak{CB}_N(A) \in \mathcal{CB}_N$ defined by

$$\mathfrak{CB}_N(A)(x,y) := \int_0^x \int_0^y N^2 \sum_{i,j=1}^{N} \mu_A(R_{ij}^N) \mathbf{1}_{R_{ij}^N}(s,t)\, d\lambda(t) d\lambda(s)$$

is called $N$-checkerboard approximation of $A$. $N$ is called the resolution of $\mathfrak{CB}_N(A)$.



### Theorem (Griessenberger & Junker & T., submitted, 2019; arXiv)

*Let $(x_1, y_1), \dots, (x_n, y_n)$ be a sample of $(X, Y)$ with copula $A$. Furthermore consider $N(n) := \lfloor n^s \rfloor$ where $s$ fulfills $0 < s < \frac{1}{2}$. Then*

$$\lim_{n\to\infty} D_1 \left(\mathfrak{CB}_{N(n)}(\hat{E}_n), A\right) = 0 \ [\mathbb{P}].$$

### Theorem (Griessenberger & Junker & T., submitted, 2019; arXiv)

*Same setting as above. Then $\zeta_1(\mathfrak{CB}_{N(n)}(\hat{E}_n))$ is a strongly consistent estimator of $\zeta_1(A)$.*

- R-package **qad**[1] (available on CRAN) calculates the empirical checkerboard copula and the estimator $\zeta_1(\mathfrak{CB}_{N(n)}(\hat{E}_n))$.
- Next talk: Florian Griessenberger will show what the package can be used for and how our dependence estimator performs in comparison to various other ones.

[1] short for 'quantification of asymmetric dependence'

## Simulations - extreme cases

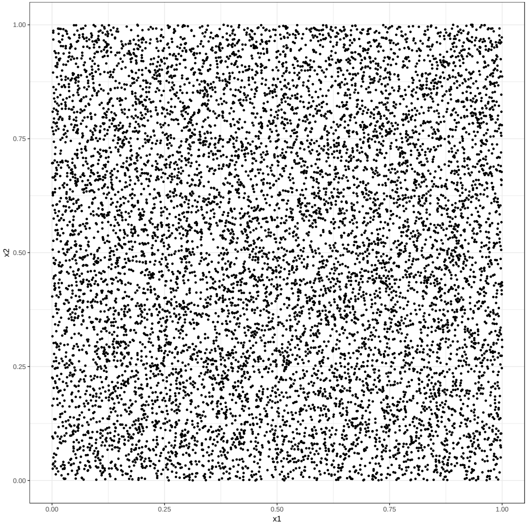



Figure: Sample of size 10.000 from the product copula Π describing independence.

## Simulations - extreme cases

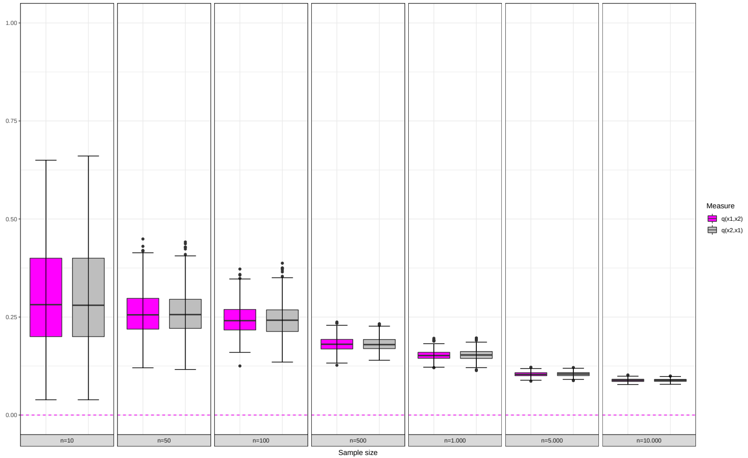



Figure: Boxplots summarizing the 1.000 obtained estimates for $\zeta_1(X, Y)$ (magenta) and $\hat{\zeta}_1(Y, X)$ (gray). The dashed lines depict the true quantities $\zeta_1(X, Y)$ and $\zeta_1(Y, X)$.

## Simulations - extreme cases

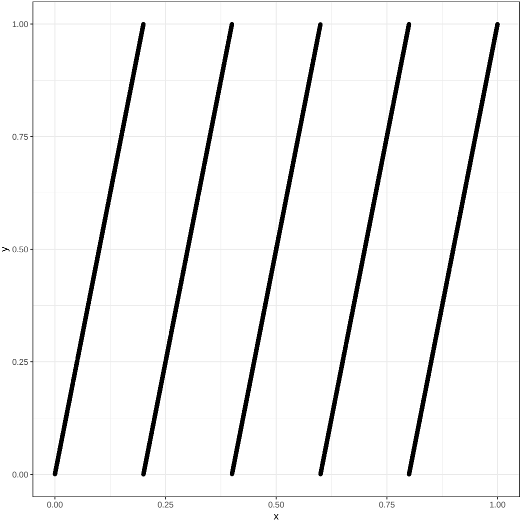



Figure: Sample of size 10.000 of a completely dependent copula $A_{h_a}$ for $h_a = ax(mod 1)$ and $a = 5$. Highly asymmetric setting!

## Simulations - extreme cases

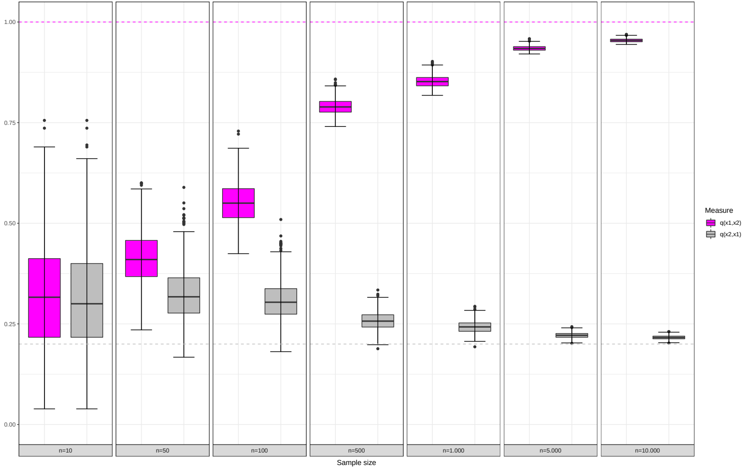



Figure: Boxplots summarizing the 1.000 obtained estimates for $\zeta_1(X, Y)$ (magenta) and $\hat{\zeta}_1(Y, X)$ (gray) for the case $a = 5$.

**Wrap-up:**

- Dependence and asymmetry in dependence is a key feature in bivariate associations.
- All standard 'dependence measures' ignore asymmetry.
- qad seems to be the first scale-invariant, model-free measure of dependence that overcomes this problem.
- $q(X, Y)$ can be interpreted as the information gained about $Y$ by knowing $X$.
- In general we have $q(X, Y) \neq q(Y, X)$.
- Many real datasets underline the usefulness of qad. Additionally, consistency has be proved mathematically.
- Nevertheless: There is a lot of work to be done: Extension to the discrete setting, extension to the multivariate setting, etc. ($\rightarrow$ part of Florian's PhD project).