Quantifying asymmetric dependence with the R-package qad

Wolfgang Trutschnig[1]

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

Symposium: Ecology — Geomorphology — Statistics

Multidiversity in environmental successions - interdisciplinary views on the emergence of ecological complexity

Salzburg, March 28-29, 2019

[1]Department for Mathematics, University of Salzburg

[2]Department of Ecology and Evolution, University of Salzburg

How it started:

- 2015/16: First collaboration of the statistics group (Bathke & Trutschnig) with Robert's group on dynamic range boxes.
- R.R. Junker, J. Kuppler, A.C. Bathke, M.L. Schreyer, W. Trutschnig: Dynamic range boxes - A robust non-parametric approach to quantify size and overlap of n-dimensional hypervolumes, *Methods in Ecology and Evolution* 7(12), 1503-1513 (2016)
- After the paper was published Robert asked me: Can you sketch a problem you are working on in dependence modeling in a way comprehensible for non-mathematicians?
- Answer:
  - I try to quantify how much influence one variable/feature X has on another variable/feature Y and vice versa.
  - Main objective is to find a nonparametric, model-independent and scale-invariant version of the famous coefficient of determination R^2
  - A picture helps...

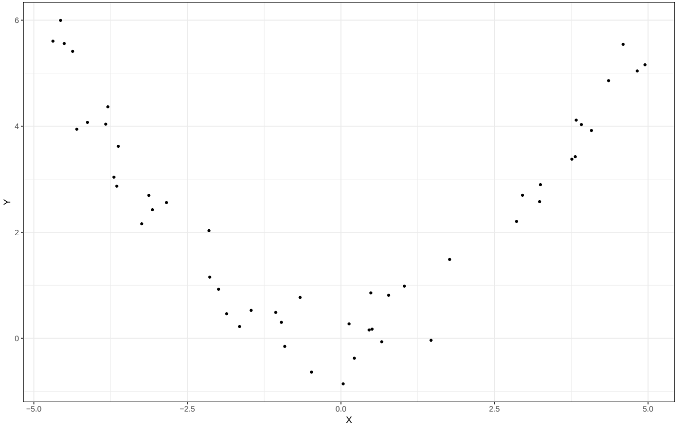


Figure: Bivariate sample $(x_1, y_1), \ldots, (x_n, y_n)$ of size $n = 50$ from the model $Y = \frac{X^2}{4} + \varepsilon$

- Which variable is easier to predict given the value of the other one?
- What would you say, and why?

- But do strongly asymmetric dependence structures really exist in nature?
- Examples:
  - Average speed vs. fuel consumption (measurements)
  - Wave length vs. reflection of light, etc.

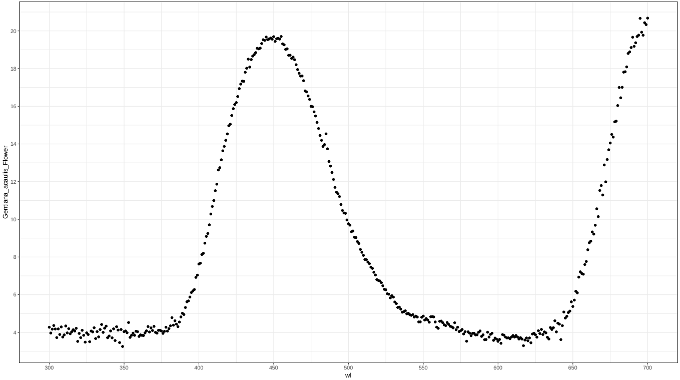


Figure: Wave length vs. reflection of light (measurements) for a purple flower

- Taking asymmetry in dependence for granted:
- How can dependence be quantified?
- How can asymmetry in dependence be quantified?

- All statistics courses mention 'independence': Two random variables X and Y are called independent, if X has no influence on Y AND vice versa.
- Toy example: X...result of rolling a dice, Y...result of rolling the dice a second time.
- If we know the outcome of X, does it help to predict Y?
- The probabilities of Y remain unchanged - **we do not gain any knowledge about Y if we know X and vice versa**.
- In other words: **Knowing X does not reduce the uncertainty of Y and vice versa**.

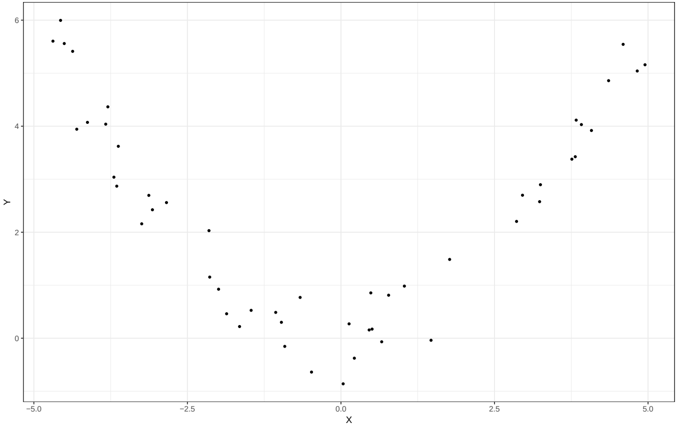


Figure: Bivariate sample $(x_1, y_1), \dots, (x_n, y_n)$ of size $n = 50$ from the model $Y = \frac{X^2}{4} + \varepsilon$

- Doesn't correlation quantify dependence?
- Why not work with Pearson, Spearman, or Kendall correlation?

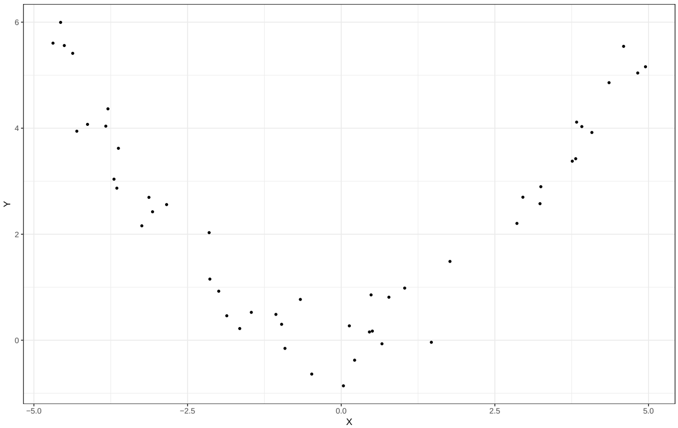


Figure: Bivariate sample $(x_1, y_1), \ldots, (x_n, y_n)$ of size $n = 50$ from the model $Y = \frac{X^2}{4} + \varepsilon$

- For the sample we get the following: $r = -0.011, \rho = -0.098, \tau = -0.081$
- Even worse: We get the same values if we interchange X and Y...

Consequences:

- **Correlation does not quantify dependence** (neither Pearson, nor Spearman, nor Kendall correlation quantifies dependence).
- Another approach is needed.

Wish list for such a quantification q:

- $q(X,Y)$ can be calculated for all continuous random variables X and Y (without having at hand an underlying model)
- $q(X,Y) \in [0,1]$ (normalization)
- $q(X,Y) = 0$ if and only if X and Y are independent (**independence**)
- $q(X,Y) = 1$ if and only if Y is a function of X (**complete dependence, full predictability**)
- It may happen that $q(X,Y) \neq q(Y,X)$ (**asymmetry**)
- Additionally: Scale changes should not affect q (**scale-invariance**)

- Robert had a big smile on his face when I told him that such a measure q existed and that I had developed and published it in 2011.
- He saw the potential of q not only for ecology (key species, invasive species, networks, etc.) but for data analytics in general.
- The smile disappeared when I told him that it was still unknown how to estimate q based on samples and that I had not found a general, consistent estimator yet...
- ...and that a superstar in my field of research (=dependence modeling) had conjectured that no such estimator existed...

- ...sometimes even superstars are mistaken.
- We found such an estimator but it took a while.
- The estimator was developed and studied in Florian Griessenberger's master thesis (2018).
- Afterwards the estimator (a so-called empirical checkerboard copula) and additional tools were implemented in the R-package qad (available CRAN) $\rightarrow$ see Florian's presentation of qad tomorrow at 14:40.

Structure for the rest of this talk:

- Sketch how the estimator works (no heavy mathematics, only the underlying ideas).
- Sketch how qad-based testing and forecasting works.
- Illustrate qad in terms of several examples and simulations.
- Please interrupt is something is unclear or if questions arise!

How the qad estimator is calculated

(S0) Suppose that $(x_1, y_1), \ldots, (x_n, y_n)$ is a sample from (X, Y).

(S1) Calculate the normalized ranks of the sample; we get values of the form $(\frac{i}{n}, \frac{j}{n})$ with $i, j \in \{1, \ldots, n\}$.

(S2) Calculate the so-called empirical copula $\hat{E}_n$ and aggregate it to the empirical checkerboard copula $\hat{C}_n$.

(S3) Calculate how different the checkerboard distribution and the uniform distribution on the unit square (modelling independence) are[1]; i.e. calculate $q_n(X, Y) = 3D_1(\hat{C}_n, \Pi))$.

- It can be proved mathematically that $q_n(X, Y) \approx q(X, Y)$ for sufficiently large n (mathematically speaking: The estimator is strongly consistent).
- Let's have a look at the construction for our specific U-shaped sample.

[1] More precisely: the conditional distribution functions are compared with the distribution function of the uniform distribution on $[0, 1]$.

The qad estimator

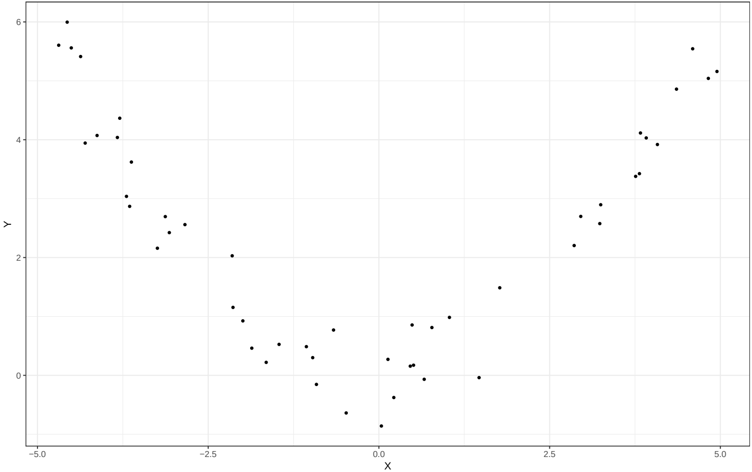


Figure: Bivariate sample $(x_1, y_1), \ldots, (x_n, y_n)$ of size $n = 50$ from the model $Y = \frac{X^2}{4} + \varepsilon$

The qad estimator

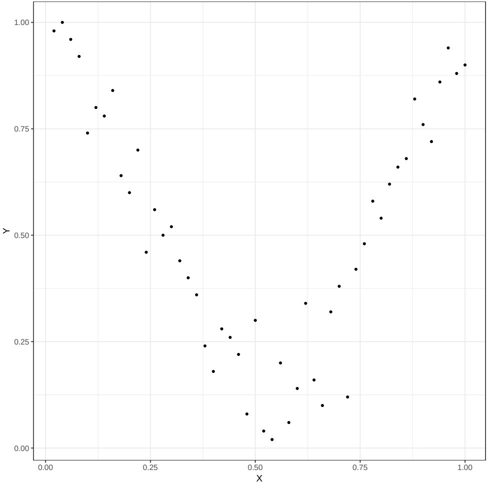


Figure: Normalized ranks of the sample; notice the scale change.

The qad estimator

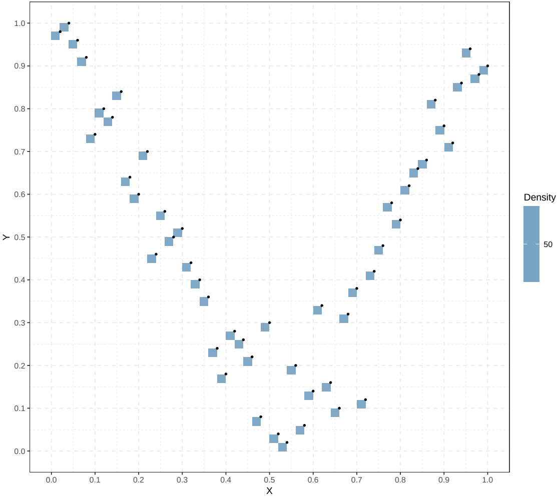


Figure: Empirical copula $\hat{E}_n$; the density is uniform on each of the little squares

The qad estimator

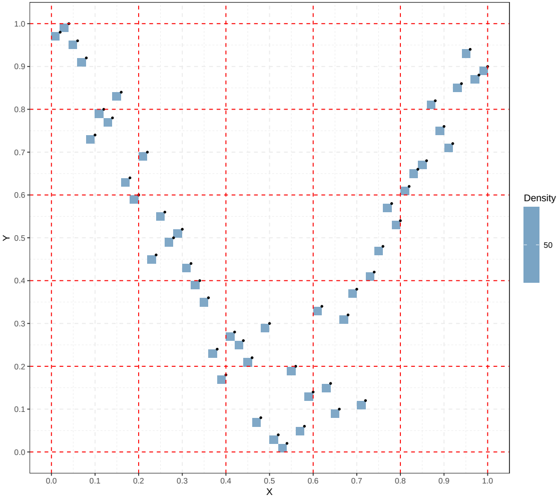


Figure: Empirical copula $\hat{E}_n$ and the partition to which we aggregate

The qad estimator

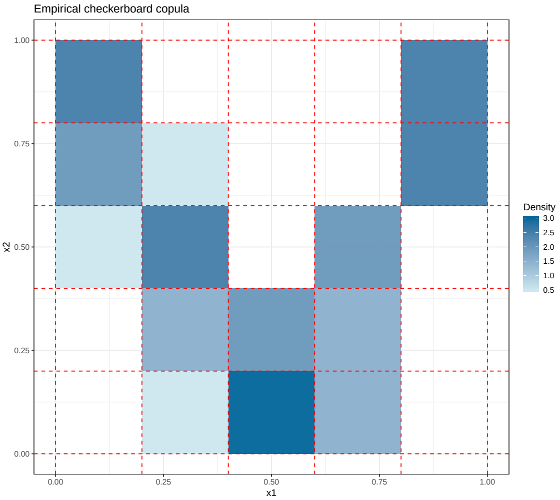


Figure: Empirical checkerboard copula $\hat{C}_n$ and its density on each of the big squares. The higher the concentration of the mass in y-direction the higher the dependence of Y on X.

The qad estimator

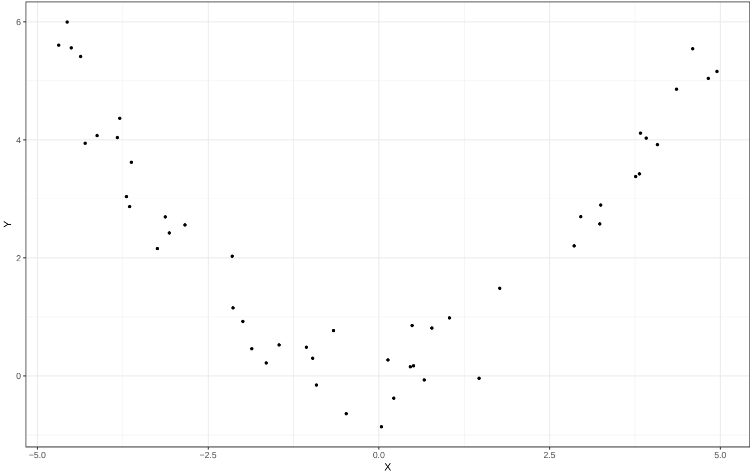
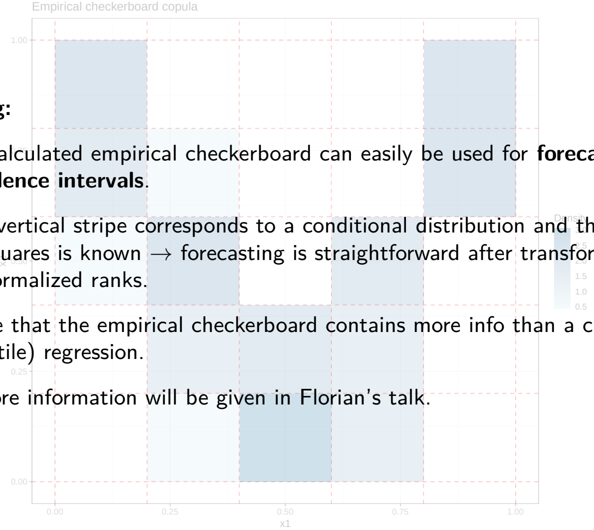


Figure: For the sample $(x_1, y_1), \ldots, (x_n, y_n)$ qad yields $q_n(X, Y) = 0.8$ and $q_n(Y, X) = 0.43$.

Testing:

- For the considered sample qad yields $q_n(X,Y) = 0.8$ and $q_n(Y,X) = 0.43$.
- So the **asymmetry in dependence** a is $a = q_n(X,Y) - q_n(Y,X) = 0.37$.
- When applying the qad-function in the qad R-package a **permutation test** for equal dependence in both directions can be executed (for syntax and function calls see Florian's talk).
- Basic idea of the implemented permutation test: Consider the doubled sample $(x_1,y_1),\ldots,(x_n,y_n),(y_1,x_1),\ldots,(y_n,x_n)$, randomly draw n observations from it, calculate the corresponding qad value and the corresponding asymmetry in dependence.
- Repeat for $R = 1.000$ times and check how often the asymmetry is at least as big as the one of the original sample $(x_1,y_1),\ldots,(x_n,y_n)$.
- The resulting p.value (based on 1.000 runs) for our sample fulfills $p < 0.001$, i.e. the null for symmetric dependence is rejected.



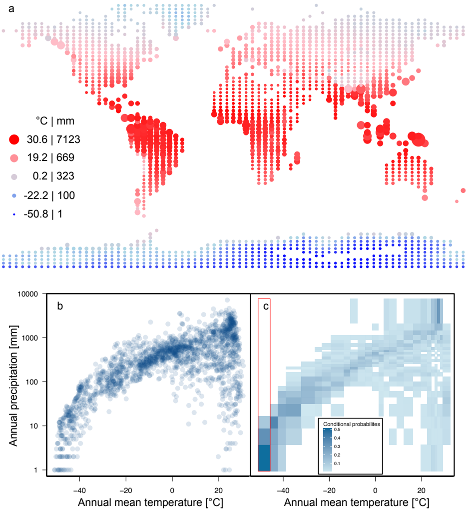
Forecasting:

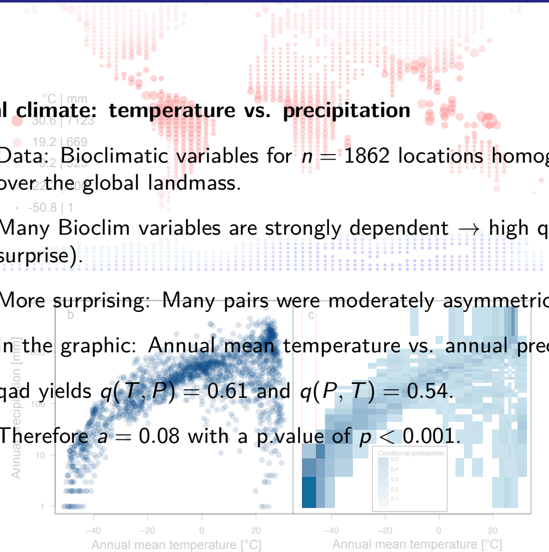
- The calculated empirical checkerboard can easily be used for **forecasting and confidence intervals**.
- Each vertical stripe corresponds to a conditional distribution and the mass of the squares is known $\rightarrow$ forecasting is straightforward after transforming back the normalized ranks.
- Notice that the empirical checkerboard contains more info than a classical (quantile) regression.
- $\rightarrow$ more information will be given in Florian's talk.

Examples

- All examples and simulations mentioned in the sequel are part of the following preprint:
- Robert R. Junker, Florian Griessenberger, Wolfgang Trutschnig: A scale-invariant measure for quantifying asymmetry in dependence and associations, submitted for publication
- The preprint is available on arXiv and can be downloaded from https://arxiv.org/abs/1902.00203
- The preprint contains a general, non-mathematical description of qad, a separate section with all the mathematics behind it, and R-Codes for the examples.

Examples

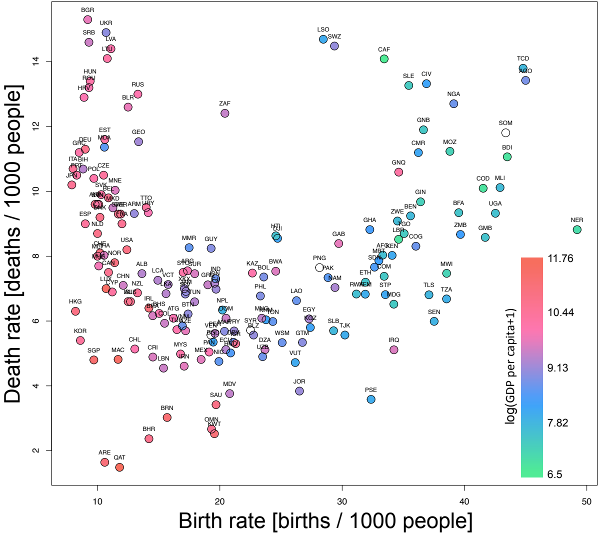

a
°C | mm
30.6 | 7123
19.2 | 669
0.2 | 323
-22.2 | 100
-50.8 | 1
b
c
10000
1000
100
10
1
Annual precipitation [mm]
−40
−20
0
20
Annual mean temperature [°C]
Conditional probabilites
0.5
0.4
0.3
0.2
0.1



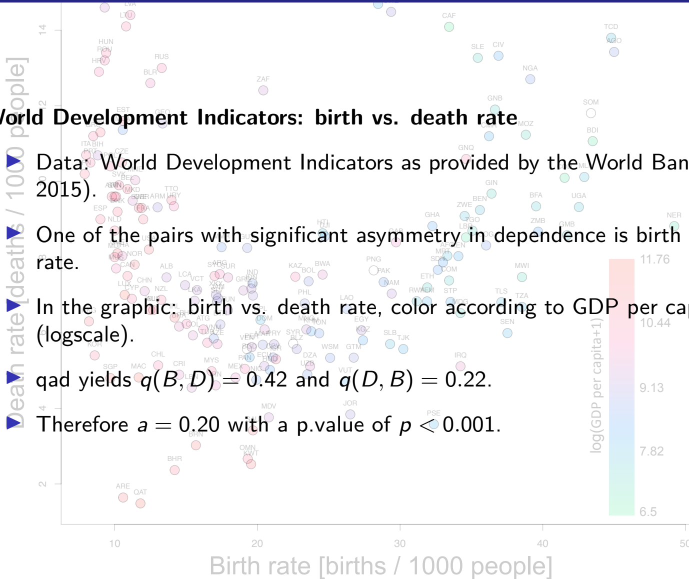
Global climate: temperature vs. precipitation

- Data: Bioclimatic variables for $n = 1862$ locations homogeneously distributed over the global landmass.
- Many Bioclim variables are strongly dependent $\rightarrow$ high qad values (no real surprise).
- More surprising: Many pairs were moderately asymmetric in dependence.
- In the graphic: Annual mean temperature vs. annual precipitation (logscale).
- qad yields $q(T, P) = 0.61$ and $q(P, T) = 0.54$.
- Therefore $a = 0.08$ with a p.value of $p < 0.001$.

Examples

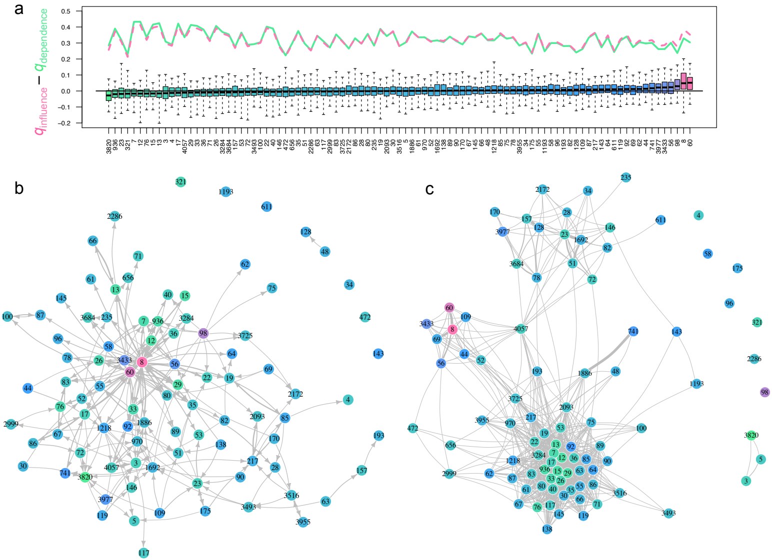

Death rate [deaths / 1000 people]
Birth rate [births / 1000 people]
log(GDP per capita+1)
11.76
10.44
9.13
7.82
6.5
2
4
6
8
10
12
14
10
20
30
40
50
BGR
UKR
SRB
LVA
LTU
HUN
HRV
RUS
BLR
ZAF
EST
GEO
DEU
CZE
MNE
TTO
ESP
NLD
USA
MMR
GUY
NOR
CAN
ALB
CHN
NZL
HKG
KOR
SGP
MAC
CHL
CRI
LBN
MYS
IRN
BRN
BHR
SAU
MDV
ARE
QAT
LSO
SWZ
CAF
SLE
CIV
TCD
NGA
GNB
SOM
CMR
MOZ
BDI
GNQ
COD
MLI
GIN
BFA
UGA
NER
BEN
ZWE
COG
GHA
ZMB
GMB
GAB
KEN
COM
ETH
STP
MDG
MWI
TLS
TZA
SEN
IRQ
PSE
JOR
VUT
LAO
EGY
KGZ
SLB
TJK
WSM
GTM
DZA
UZB
PNG
PAK
NAM
KAZ
BOL
BWA
PHL
SYR
BLZ
NPL
DOM
BTN
ATG
BHS

World Development Indicators: birth vs. death rate



- Data: World Development Indicators as provided by the World Bank (year 2015).
- One of the pairs with significant asymmetry in dependence is birth vs. death rate.
- In the graphic: birth vs. death rate, color according to GDP per capita (logscale).
- qad yields $q(B, D) = 0.42$ and $q(D, B) = 0.22$.
- Therefore $a = 0.20$ with a p.value of $p < 0.001$.

Examples



Microbiomes: abundance of OTUs

- Data: Abundances of bacteria associated with surfaces of the plant Metrosideros polymorpha.
- We calculated the qad values for all pairs of $n = 93$ operational taxonomic units (OTUs).
- We looked for key species, i.e. species which influence the abundance of other species but are less influenced by others.
- 7 OTUs were identified as key species.
- The 4 OTUs with the highest influence values were also identified in an experimental study as key players w.r.t. abundance of bacteria.
- qad goes beyond simple statistics and may produce ecologically meaningful outcomes.

Simulations

- Whenever new estimators are developed statisticians test their performance.
- Basic idea is (strong) consistency: For sufficiently large samples the estimator should be close to the true value.
- Toy example: $\mathcal{N}(1,2)$, for large n we expect $\overline{X}_n \approx 1$.
- We proved strong consistency mathematically.
- Simulations illustrate the speed of convergence as well as the small sample performance.
- Numerous dependence structures (no matter if they may appear in nature or not) were considered.
- The next slides only show two extreme cases.

Simulations

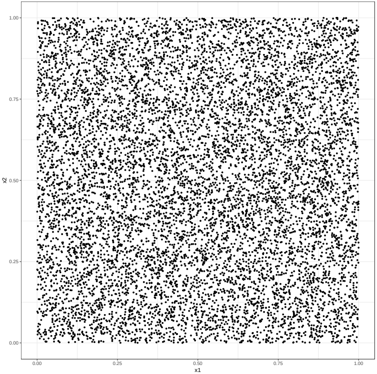


Figure: Sample of size 10.000 from the product copula Π describing independence.

Simulations

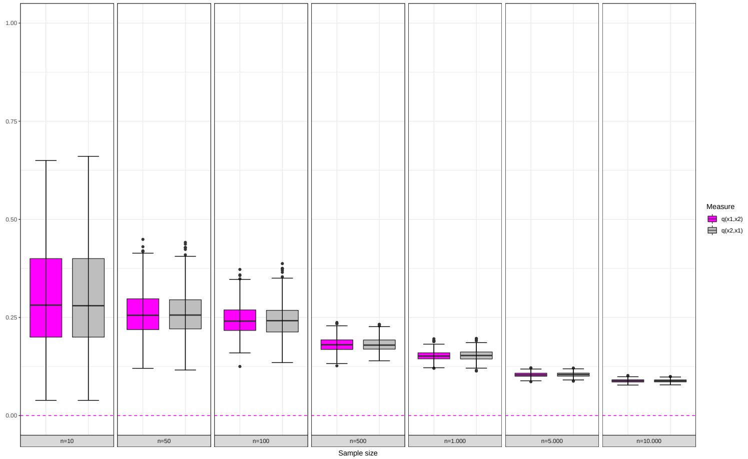


Figure: Boxplots summarizing the 1.000 obtained estimates $q_n(X,Y)$ (magenta) and $q_n(Y,X)$ (gray).
The dashed lines depict the true (=population) values $q(X,Y)$ and $q(Y,X)$.

Simulations

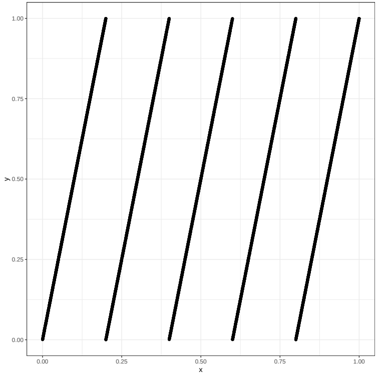


Figure: Sample of size 10.000 of a situation with $Y = f(X)$.

Simulations

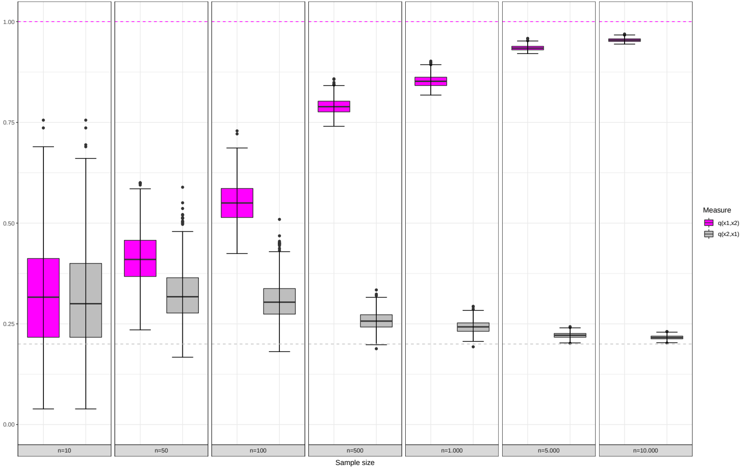


Figure: Boxplots summarizing the 1.000 obtained estimates $q_n(X,Y)$ (magenta) and $q_n(Y,X)$ (gray).
The dashed lines depict the true (=population) values $q(X,Y)$ and $q(Y,X)$.

Wrap-up:

- Asymmetric 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.
- $q(X, Y)$ describes 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 do for statisticians.

Future work:

- qad was developed for continuous data and not for count data (abundances, etc.).
- Nevertheless: It also produces good results for such data.
- **To do: Study the mathematical properties of the estimator in the count data setting** ($\rightarrow$ part of Florian's PhD project).
- So far we can only quantify dependence of pairs - the interplay between two variables might have an influence on a third variable but none of the variables individually.
- **To do: Extend qad to the general multivariate setting** ($\rightarrow$ part of Florian's PhD project).