A copula-based measure for quantifying asymmetry in dependence and associations

Robert R. Junker1*, Florian Griessenberger2, Wolfgang Trutschnig2

1University of Salzburg, Department of Biosciences, Hellbrunnerstrasse 34, 5020 Salzburg
2University of Salzburg, Department of Mathematics, Hellbrunnerstrasse 34, 5020 Salzburg
*Corresponding author: robert.junker@sbg.ac.at, telephone number: +43/662/8044-5512

Keywords: asymmetry, climate, dependence, key species, network, Pearson's r, Spearman's rho

Abstract

Asymmetry is an inherent property of bivariate associations and therefore must not be ignored. The currently applicable dependence measures mask the potential asymmetry of the underlying dependence structure by implicitly assuming that quantity Y is equally dependent on quantity X, and vice versa, which is generally not true. We introduce the copula-based dependence measure \( qad \) that quantifies asymmetry. Specifically, \( qad \) is applicable in general situations, is sensitive to noise in data, detects asymmetry in dependence and reliably quantifies the information gain/predictability of quantity Y given knowledge of quantity X, and vice versa. Using real-world data sets, we demonstrate the relevance of asymmetry in associations. Asymmetry in dependence is a novel category of information that provides substantial information gain in analyses of bivariate associations.

Introduction

Despite the prevalence of asymmetry in many systems, it is ignored by measures quantifying the statistical dependence between variables. In the two-dimensional setting, both the classical Pearson's correlation and Spearman rank correlation are symmetric, i.e., \( \rho(X, Y) = \rho(Y, X) \) holds; therefore, neither measure is capable of detecting asymmetries. The same is true for approaches that do not rely on specific forms of dependence and thus identify nonlinear and non-monotonic relationships (1). According to Renyi's (2) axioms, every dependence measure \( d \) should be symmetric or mutual in the sense that \( d(X, Y) = d(Y, X) \). However, considering, for instance, a two-dimensional sample in the form of a parabola (Fig. 1) or a sinusoidal curve (Fig. 2g), the dependence structure is clearly asymmetric. In such cases, knowing X strongly improves the predictability of Y, whereas the opposite is not true to the same extent. The
dependence of random variables is asymmetric by nature, and ignoring asymmetries means ignoring valuable information. To provide a statistical tool to quantify asymmetry in dependence, we introduce the quantification of asymmetric dependence \((qad)\) that induces the following related measures: \(q(X, Y)\) denoting the dependence of \(Y\) on \(X\) (or, equivalently, the influence of \(X\) on \(Y\)), \(q(Y, X)\) the dependence of \(X\) on \(Y\), \(\frac{1}{2}(q(X, Y) + q(Y, X))\) the average dependence, and \(a = q(X, Y) - q(Y, X)\) the asymmetry of the dependence. The four measures are associated with \(p\)-values related to the significance of the dependence and asymmetry.

Big data and small data sets require robust statistical analyses to extract information regarding the relationships among variables. In this context, asymmetric or directional dependence adds additional and important information to the dependence structure of two variables \(X\) and \(Y\). Both inductive and deductive approaches strongly benefit from the quantification of asymmetric dependence; notably, either more precise hypotheses and predictions will emerge from large data sets, or more precise hypotheses can be tested, i.e., quantifying the dependence of \(Y\) on \(X\) (or vice versa), instead of the overall dependence between \(X\) and \(Y\). Treating dependence structures as symmetric by default goes hand in hand with ignoring key factors and relevant information. For example, if the causal relationship between two variables is known, the dependence of the affected variable on the causing variable is desired information that should be unaffected by the statistical dependence of the causing variable on the affected variable. In cases where no a priori knowledge of the causal relationship is available, directional dependence is a useful measure for exploring and estimating the association between variables in a more detailed and realistic way than classical (symmetric) dependence measures. Furthermore, the asymmetry \(a = q(X, Y) - q(Y, X)\) provides additional information on bivariate associations. Although directional dependence is (by construction) not the same as causality, asymmetry in dependence may help to identify causal relationships among variables. In the era of big data, asymmetric directional dependence may be of particular importance for exploring pairwise relationships, identifying pairs of variables that feature either symmetric or asymmetric relationships, or pinpointing variables that affect a number of variables but are (mostly) unaffected by others and, therefore, may be key factor in their given system.

As a copula-based measure, \(qad\) captures all scale-invariant dependence in general situations that extend beyond standard parametric families or linear relationships and thus requires no assumptions regarding the underlying distribution of the data. These properties enable an almost universal applicability to data sets of various disciplines. We chose real-world data sets
to illustrate the usefulness of qad and its potential in identifying important variables in large data sets.

The copula-based asymmetric dependence measure qad

The asymmetric, scale-invariant dependence of pairs of variables can be modeled via copulas (3-5). Copulas link multivariate distribution functions and their one-dimensional marginals and thereby capture all scale-invariant dependence (6). Moreover, qad can be used to detect a) the dependencies of any functional type, which can be seen as generalized asymmetric versions of the coefficient of determination $R^2$, and b) the asymmetry of the dependence between two random variables. For a concise mathematical description of the applied methodology, details regarding the motivation for this approach, and explicitly calculated examples, please refer to Supporting information 1. In short, given a two-dimensional sample $(x_1, y_1), \ldots, (x_n, y_n)$ from random vector $(X, Y)$, the qad method involves calculating the so-called empirical copula $\hat{C}_n$, aggregating/smoothing the copula to obtain an empirical checkerboard $\hat{C}_n$ (the unit square $[0, 1]^2$ is partitioned into $N^2$ squares of edge length $\frac{1}{N}$, the mass per square is aggregated, and $N$ is referred to as the resolution) and then calculating the normalized $D_1$-distance of $\hat{C}_n$ and the product copula $\Pi$ (modeling independence). The normalization constant 3 assures that the range of the dependence measure is within the interval of $[0,1]$ (3). The resulting value is denoted by $q_n(X, Y)$. After computing the quantities $q_n(X, Y)$ and $q_n(Y, X)$, the asymmetry $\alpha$ of the sample $(x_1, y_1), \ldots, (x_n, y_n)$ is calculated as $\alpha = q(X, Y) - q(Y, X)$. A permutation test (based on permutations of the conditional distributions of the empirical checkerboard) is then used to establish a $p$-value for testing the hypothesis $H_0 : q(X, Y) = q(Y, X)$ of symmetric dependence. The empirical checkerboard copula underlying qad together with the empirical marginals can also be exploited to predict Y given X, and vice versa. This process is facilitated by plots of the probability of Y for a given X, and vice versa (Fig. 1). Note that prediction is possible only within the range of measured X- and Y-values. Moreover, qad is calculated independent of a parametric regression function; therefore, no extrapolation is possible. To allow for the cross-platform application of qad, we wrote the R package qad, which is freely available from CRAN (7). The qad method is a non-parametric approach and thus involves no assumptions regarding the underlying distribution of the data. In addition, the approach is robust to outliers and returns reliable results. However, because the qad method builds upon copulas and copulas are (according to Sklar’s Theorem) only unique for continuous random variables, discrete variables (i.e., ties in samples) must be handled with care. Although our simulations
indicate adequate and reliable performance for data sets with ties (indicating point masses of the random variables \( X, Y \)), the number of unique values should be sufficiently large in comparison with the sample size. The resolution of the empirical checkerboard copula is proportional to the square root of the sample size; thus, as in any statistical method, the qad results become more reliable as the sample size increases. We recommend a sample size of no smaller than \( n = 16 \) (i.e., resolution = 4 x 4).

Results and Discussion

Properties of qad. Dependence measures such as qad that capture the dependence in general situations should assign similar scores of dependence to equally noisy data in a manner independent of the concrete functional relationship (1). Accordingly, the dependence \( q \) between two random variables decreases with increasing noise irrespective of the functional relationship between \( X \) and \( Y \) (Fig. 2). In a linear setting, qad returned dependence values \( q \) closer to Pearson’s \( r \) and Spearman’s \( \rho \) than the maximum information coefficient MIC (1), a symmetric dependence measure defined for general situations (Fig. 2a). Unlike Pearson’s \( r \), Spearman’s \( \rho \) and MIC, which are symmetric measures by construction, qad indicated asymmetry in settings in which (on average) more information on \( Y \) could be obtained by knowing the value of \( X \) than vice versa, i.e., \( q(X, Y) > q(Y, X) \) (Fig. 2c, d, f, g, i). Recently, the use of Pearson’s product moment correlation as a standard method for quantifying dependence was criticized for underestimating dependence (8) and for being a predictor of limited applicability in complex systems (9). In order to specifically test the relation between Pearson’s \( r \) and qad we utilized the R-function plot_r() provided by Vanhove (10), which generates 16 datasets with different distributions (functional relations with noise component) all having the same Pearson correlation coefficient \( r \). For \( n = 200 \) predefined \( r \)-values in the interval [-0.999, 0.999], we generated these 16 datasets and then quantified dependence using qad, Spearman’s \( \rho \), and the maximum information coefficient MIC (Supporting information 4). In datasets with linear relationships and errors following a normal or a uniform distribution, qad produced values similar to \( r \), just as Spearman’s \( \rho \) and MIC (Fig. 3 a-h) did. In datasets with non-linear underlying dependence structures, however, Pearson’s \( r \) and Spearman’s \( \rho \) underestimated the dependence of the variables (Fig. 3 i, j, m, n) whereas qad and MIC detected a higher dependency in these cases. Furthermore, the graphics illustrate the well-known fact that that Pearson’s \( r \) is particularly sensitive to (single) outliers (Fig. 3 k, l), whereas qad and the other rank-based quantities are not strongly affected. All dependence measures returned similar values for bimodal distributions (Fig. 3 o). For coarse data, qad, Pearson’s \( r \) and Spearman’s returned similar values, MIC failed
to detect the underlying dependence structure (Fig. 3 p). Finally, *qad* was the only dependence measure capable of detecting asymmetries (Fig. 3 i, j, m, n) and thus provides additional valuable information on the underlying dependence structure.

To demonstrate the applicability and added value of *qad*, we used data sets from various disciplines to explore the asymmetry in the corresponding dependence structures.

*Cause–Effect Pairs*

We quantified the asymmetric dependence of 74 *Cause–Effect Pairs* (11, 12) to evaluate the causal discovery rate of *qad*. The *Cause–Effect Pair* benchmark data constitute a collection of data sets with known causalities for the evaluation of bivariate causal discovery methods (11). In 63.5% of all the tested pairs, the affected variable was more dependent on the causing variable than *vice versa*; therefore, the causal relationships were correctly discovered in these cases (see Supporting information SI2). The results suggest that the causal discovery rate of *qad* is as good as the best performing method specifically designed for causal discovery tested based on the same data set (11, 13). Moreover, the findings confirm that statistical dependence is not the same concept as causality; nevertheless, *qad* may help to infer causality.

*Global climate*

In the context of global warming and changes in precipitation regimes, information regarding past, present, and future climate at a local scale is required to assess the effects of climate on the environment. We retrieved bioclimatic variables for *n* = 1862 locations homogenously distributed over the global landmass from CHELSA (14, 15) to test for asymmetric dependence between variables. As expected, the knowledge of one of the bioclimatic variables reduced the variability of the other variables, as reflected by the high dependence between pairs of variables (mean *q* ± sd = 0.54 ± 0.18, see Supporting information SI2). Many of the associations between the bioclimatic variables were asymmetric (mean asymmetry *α* ± sd = 0.06 ± 0.05, range = 0.00 – 0.23, see Supporting information SI2). For instance, annual precipitation can be better predicted by mean temperature (*q* = 0.61, *p* < 0.001) than *vice versa* (*q* = 0.54, *p* < 0.001, asymmetry *a* = 0.08, *p* < 0.001, Fig. 4). Thus, *qad* helps to identify variables that are more predictive or predictable (on average) than others (in the context of climate, and in any other context), which is important knowledge for study design, particularly if assessments of many variables are cost and labor intensive.

*Microbiomes*
Diverse and complex microbial communities have become accessible due to high-throughput sequencing. Ecological relationships such as mutualism, competition, and commensalism between organisms shape the abundance distribution of taxa in communities. Often, correlation analysis testing for relationships between the abundance of pairs of taxa is used as basis for network inference, which facilitates the interpretation of microbiome structure. Ecological relationships between organisms may be reciprocal in the sense that taxa mutually affect each other, either positively (mutualism) or negatively (competition). They may, however, also be directed such that a given taxon is facilitating or inhibiting the growth of another taxon without being affected by itself by the other taxon (e.g. commensalism, amensalism). Conventional correlation analysis neither detects directed relationships nor discriminates between directed and mutual relationships, and is therefore of limited value for community dynamics (9). Here we used a dataset of bacteria associated with surfaces of the plant *Metrosideros polymorpha* (16) and tested for asymmetric relationships between pairs of \(n = 93\) operational taxonomic units (OTUs) that were observed in at least 75% of all samples (\(n = 125\)). \(q\)-values (mean ± SD: 0.32 ± 0.094) were on average higher than Pearson’s \(r^2\) (0.07 ± 0.11) indicating that the majority of relationships between OTU pairs are not well described as a linear or isotonic interaction. Mostly, asymmetry in the dependence between OTUs was weak indicated by a relatively low mean value of asymmetry \(|a|\) (0.034 ± 0.028). However, some pairwise interactions between OTUs were strongly asymmetric reaching \(|a|-values of >0.20\). The quantification of asymmetric dependencies allows a novel definition of key species \(S_f\), which influence the abundance of other species \(S_j\) but are less influences by others. Thus, key species \(S_f\) influence other species \(S_j\) on average to a larger extend than they are dependent on other species \(S_j\). Median influence of species \(S_f\) is calculated as the median values of \(I_f\) calculated as

\[
I_f^j = q(S_f, S_j) - q(S_j, S_f), 1 \leq j \leq n, j \neq f
\]

If median \((l_f) > 0\) species \(S_f\) influences most other species \(S_j\) stronger than *vice versa*. As mentioned before, \(q(S_f, S_j)\) denotes the dependence of \(S_j\) on \(S_f\) and \(q(S_j, S_f)\) denotes the dependence in the other direction (Fig. 5a). In seven OTUs the corresponding medians of \(I_f\) were significantly larger than zero indicating a stronger influence on most of the other OTUs than they are dependent on the other OTUs, i.e. these OTUs are defined as key species in this dataset (Fig. 5a; OTU 60, 8, 98, 56, 3433, 3977, 741; Supporting information 2). The four OTUs with the highest median influence-value \((l_f)\) are members of class Gammaproteobacteria, three of them are within the order Pseudomonadales (genera Acinetobacter (OTU 8 and 60) and Pseudomonas (OTU 56)), one in the order Enterobacteriales (OUT 98, Supporting information 2). Interestingly, in an experimental study
Acinetobacter and Pseudomonas had been identified to play key roles in shaping abundances of bacteria associated with plant surfaces (17) suggesting that our results are not random findings but represent ecologically meaningful outcomes of qad. Using q-values in an adjacency matrix to create a weighted and directed network visualizing asymmetric dependencies q between OTUs confirmed the prominent role of the key-OTUs in shaping the abundances of a number of other OTUs. Most of the seven key-OTUs occupy central positions (betweenness centrality: 0 – 1882, mean ± sd = 431.7 ± 639.1; degree: 3 – 48, 16.0 ± 14.8) indicating their role as network hubs (hub-score: 0.04 – 1; 0.38 ± 0.33; Fig. 5b). In contrast, in an undirected and weighted network based on Pearson’s $r^2$, key-OTUs defined by qad occupied peripheral positions (betweenness centrality: 0 – 17, mean ± sd = 5.1 ± 6.4; degree: 0 – 11, 6.3 ± 3.7). Accordingly, these OTUs were not identified as hubs (hub-score: 0.00 – 0.05; 0.019 ± 0.016; Fig. 5c). Furthermore, network roles of all OTUs in networks based on qad and r were independent in betweenness centrality and degree (qad: q ≤ 0.203, p ≥ 0.08; Pearson’s product-moment correlation: |t| ≤ 0.103, df = 91, p ≥ 0.61). The hub-scores obtained by the OTUs in both networks showed a weak and even negative association (qad: q = 0.34, p ≤ 0.001; Pearson’s product-moment correlation: |t| ≤ -2.39, df = 91, p = 0.019). Betweenness centrality, degree, and hub-score have previously been used to identify key-species in diverse microbial communities (18). These values can, however, vary depending on the statistical analysis performed to obtain the dependence of OTUs such as Pearson’s r, Spearman’s rho, MIC (18), or qad. qad provides two novel indices – the asymmetric dependence q and the asymmetry a – that may help to detect key-species in diverse and complex microbiome data. We do not recommend to exclusively relying on qad for network inference and the detection of key-species, but without doubt qad has the potential to discover ecologically important species that would remain undiscovered using symmetric approaches.

World development indicators

The World Bank provides data sets on national development indicators related to, for example, the economy, education, health, and infrastructure of states (World Development Indicators WDI, The World Bank, last accessed July 2017). Apart from hypothesis-driven studies, data exploration to develop hypotheses and find interesting associations in large data sets is a major challenge in the era of big data. We explored the WDI data set published for 2015 to identify indicator pairs that feature strong asymmetry in dependence and thus stand out from other indicator pairs (see Supporting information SI2). This approach differs from previously applied strategies to identify interesting pairs of variables in large data sets, where pairs with a strong
dependence (linear or nonlinear) were selected (1). The birth rate and death rate indicators ($qad$: mean $q = 0.42$, $p < 0.001$) displayed strong asymmetry in dependence ($qad$: $a = 0.2016$, $p < 0.001$; Fig. 6) and thus stood out from the majority of pairs tested (higher asymmetry than in 99.8% of all indicator pairs). Further data exploration revealed that the gross domestic product per capita (GDP) was a potential underlying factor for this nonlinear relationship. The GDP was strongly and symmetrically associated with the birth rate ($qad$: mean $q = 0.67$, $p < 0.001$; $a = 0.02$, $p = 0.59$) and weakly and symmetrically related to the death rate ($qad$: mean $q = 0.31$, $p < 0.001$; $a = 0.03$, $p = 0.39$, see Supporting information SI2). Generally, in countries with a GDP below the mean GDP of all countries, the birth rate and death rate are positively associated; in countries with a GDP above the mean GDP, we found a negative association (Fig. 6, see Supporting information SI2). The relationship between the economy and population growth, i.e., the relationship between the birth and death rates, has been previously discussed (19, 20) and has important implications for societies. We conclude that $qad$ strongly facilitates the detection of interesting associations in large data sets and thus contributes to more thorough exploration of data sets and the improved detection of meaningful patterns compared to traditional methods.

Our theoretical and real-world examples clearly demonstrate the necessity and universal applicability of the quantification of (a)symmetric dependence $qad$ for extracting important information from bivariate associations. Asymmetry in dependence will facilitate the detection and extraction of patterns from big data and the testing of hypotheses. The $qad$ provides information pertaining to a novel property of bivariate associations – asymmetry in dependence – which will enhance the understanding of large and small data sets.

References and Notes:

1. Reshef DN, et al. (2011) Detecting novel associations in large data sets. *Science* 334(6062):1518-1524.
2. Renyi A (1959) On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae* 10:441-451.
3. Trutschnig W (2011) On a strong metric on the space of copulas and its induced dependence measure. *J Math Anal Appl* 384(2):690-705.
4. Dette H, Siburg KF, & Stoimenov PA (2013) A copula-based non-parametric measure of regression dependence. *Scand J Stat* 40(1):21-41.
5. Boonmee T & Tasena S (2016) Measure of complete dependence of random vectors. *J Math Anal Appl* 443(1):585-595.
6. Durante F & Sempi C (2015) *Principles of Copula Theory* (CRC/Chapman & Hall, Boca Raton).
7. R Core Team (2016) R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna).
8. Humphreys RK, Puth M-T, Neuhäuser M, & Ruxton GD (2018) Underestimation of Pearson’s product moment correlation statistic. *Oecologia*:1-7.
9. Coenen AR & Weitz JS (2018) Limitations of correlation-based inference in complex virus-microbe communities. *mSystems* 3(4):e00084-00018.
10. Vanhove J (2016) What data patterns can lie behind a correlation coefficient? (http://janhove.github.io/teaching/2016/11/21/what-correlations-look-like).
11. Mooij JM, Peters J, Janzing D, Zscheischler J, & Scholkopf B (2016) Distinguishing cause from effect using observational data: methods and benchmarks. *J Mach Learn Res* 17.
12. Dua D & Karra Taniskidou E (2017) UCI Machine Learning Repository. (Irvine, CA: University of California, School of Information and Computer Science).
13. Hoyer PO, Janzing D, Mooij JM, Peters J, & Schölkopf B (2009) Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, pp 689-696.
14. Karger DN, *et al.* (2017) Data Descriptor: Climatologies at high resolution for the earth's land surface areas. *Sci Data* 4.
15. Karger DN, *et al.* (2017) Data from: Climatologies at high resolution for the earth's land surface areas. *Dryad Digital Repository* https://doi.org/10.5061/dryad.kd1d4.
16. Junker RR & Keller A (2015) Microhabitat heterogeneity across leaves and flower organs promotes bacterial diversity. *FEMS Microbiol Ecol* 91: fiv097.
17. Maignien L, DeForce EA, Chafee ME, Eren AM, & Simmons SL (2014) Ecological Succession and Stochastic Variation in the Assembly of Arabidopsis thaliana Phyllosphere Communities. *mBio* 5(1).
18. Banerjee S, Schlaeppi K, & Heijden MG (2018) Keystone taxa as drivers of microbiome structure and functioning. *Nat Rev Microbiol*:1.
19. Harper S (2014) Economic and social implications of aging societies. *Science* 346(6209):587-591.
20. Lindo JM (2015) Aggregation and the estimated effects of economic conditions on health. *Journal of health economics* 40:83-96.
Acknowledgments: The study was funded by the Austrian Science Fund (FWF, Y 1102-B29).

Author contributions: RRJ and WT designed the study; RRJ, FG, and WT analyzed the data; RRJ and WT drafted the manuscript; and all authors contributed to the final version.

Supplementary Information:

Supporting information 1 – Concise mathematical description of \( qad \), the motivation behind this approach, and explicitly calculated examples

Supporting information 2 – Methods, data, and additional results regarding the theoretical and real-world data sets

Supporting information 3 – Data used in the study.

Supporting information 4 – R codes
Fig. 1 Quantification of asymmetric dependence \( qad \). The quadratic function \( (a, d, g) \) illustrates asymmetric dependence. \( a \) For each value of \( X \), the value of \( Y \) is well predicted, whereas for each \( Y \), two values (or regions) of \( X \) are generally possible. Accordingly, \( q(X, Y)\approx 0.96, p < 0.001 \), and \( q(Y, X)= 0.48, p < 0.001 \), resulting in strong and significant asymmetry \( a = 0.48 \) at \( p < 0.001 \). Based on the pseudo-observations (normalized ranks, uniform distribution of data, \( b, e, h \)) the empirical checkerboard copula is calculated (\( c, f, i \)). The resolution of the checkerboard copula (number of stripes per dimension) depends on the sample size. For each vertical or horizontal stripe, the estimated conditional distribution is given, and the values sum to 1. Therefore, the checkerboard copula can be used to predict \( Y \) given the knowledge of \( X \), and \textit{vice versa}. The precision of the prediction decreases with the sample size for \( d-f \) \( q(X, Y) = \).
0.90, $p < 0.001$; $q(Y, X) = 0.45, p < 0.001$; $a = 0.45, p < 0.001$). For noisy and less depend data, we obtain (g-i: $q(X, Y) = 0.36, p < 0.001$; $q(Y, X) = 0.24, p < 0.001$; $a = 0.12, p < 0.001$).
Fig. 2 Comparison of qad to existing methods quantifying the dependence of two random variables. $q(X, Y), q(Y, X)$, the maximum information coefficient MIC, Pearson’s $r$ and Spearman’s rho are calculated for nine different dependence structures as a function of increasing vertical noise (a-i). The gray points in the background depict samples from the corresponding dependence structures. Absolute values of the Pearson correlation $r$ and Spearman correlation $rho$ are plotted, and negative values were obtained in (d) and (e). In the linear case, the qad values range between Pearson’s $r$, Spearman’s $rho$ and MIC (a). For the nonlinear function types, qad and MIC yields higher dependence values than Pearson’s $r$ and Spearman’s $rho$. Note that the dependence structures depicted in (c), (d), (f), (g), and (i) are asymmetric, as reflected by the two $q$-values because we have $q(X, Y) > q(Y, X)$, particularly in situations with little noise. The other measures of dependence are not able to provide information about asymmetry in dependence.
Comparison of \( qad \) to Pearson’s correlation coefficient \( r \) and further existing methods quantifying the dependence of two random variables. The R-function plot_r() provided by Vanhove (10) generates 16 datasets with different distributions (functional relations with noise component) all having the same Pearson correlation coefficient \( r \). For each pattern, we simulated data sets with a given \( r \) and quantified dependence using \( qad \), Spearman’s \( \rho \), and MIC. Absolute values for Spearman’s \( \rho \) are given. Example data with \( r = 0.5 \) are depicted in the background in gray in each figure. The patterns are (a) normal x, normal residuals; (b) uniform x, normal residuals; (c) skewed x, normal residuals; (d) skewed x, normal residuals; (e) normal x, skewed residuals; (f) normal x, skewed residuals; (g) increasing spread; (h) decreasing spread; (i) quadratic trend; (j) sinusoid relationship; (k) a single positive outlier; (l) a single negative outlier; (m) bimodal residuals; (n) two groups; (o) sampling at the extremes; (p)
coarse data. For further information see Vanhove (10). Note that the (dependence structures underlying the) samples depicted in (i), (j), (m), and (n) are asymmetric, which is also detected by $qad$ ($q(X, Y) > q(Y, X)$), the other measures of dependence are not able to inform about asymmetry in dependence.
**Fig. 4** Association between the annual mean temperature AMT [°C] and annual precipitation AP [mm]. (a) Grid of \( n = 1862 \) locations homogenously distributed over the global landmass for which bioclimatic variables were retrieved. AMT is color coded, and AP is proportional to the point size. Low precipitation (small points) occurs in hot areas (red points, e.g., Sahara) and cold areas (blue, e.g., Arctic and Antarctica). High precipitation (large points) is mainly restricted to hot, tropical areas. (b) Accordingly, the association between AMT and AP is asymmetric. The knowledge of AMT strongly improved the predictability of AP \( (q = 0.61, p < 0.001) \), whereas the predictability was weaker in the opposite case \( (q = 0.54, p < 0.001; \text{Asymmetry } a = 0.08, p < 0.001) \). (c) The association between AMT and AP visualized as a two-dimensional empirical checkerboard distribution (retransformed empirical checkerboard copula), which can be used to predict Y-values given X-values, and **vice versa**. For instance, in an area with an AMT of -50.8...
°C (first column), the estimated probability of AP totaling between 1 mm and 6 mm is 51.6%, between 6 mm and 14 mm is 35.8%, and between 14 mm and 26 mm is 12.6% (red frame).
Fig. 5 Bacterial community associated with above ground surfaces of *Metrosideros polymorpha*, a tree species endemic to the Hawaiian Islands. (a) To identify key species following the definition that key species strongly influence the abundance of (many) other species but are influenced to a much smaller extend by other species, we calculated the median of the influences $I^f_j = q(S_f, S_j) - q(S_f, S_f)$ of each species $S_f$ and every other species $S_j$. For each OTU $S_f$, the median, and quartiles of $I^f_1, \ldots, I^f_n$ is calculated. OTUs are ordered by increasing median influence $I$. In seven OTUs the corresponding medians were significantly greater than zero indicating that these OTUs have a stronger influence on most of the other OTUs than they are dependent on the other OTUs, i.e. these OTUs are defined as key species in this dataset (Fig. 5a; OTU 60, 8, 98, 56, 3433, 3977, 741; Supporting information 2). The dashed pink line above the boxplots depicts the average influence $\frac{1}{n-1} \sum_{1 \leq j \leq n, j \neq f} I^f_j$ of each focal species $S_f$ on the other species $S_1, \ldots, S_n$. The green line denotes the mean influence $\frac{1}{n-1} \sum_{1 \leq j \leq n, j \neq f} I^f_j$ of the other species on species $S_f$ (b) Weighted and directed network visualizing asymmetric dependencies between OTUs. Arrows of edges point towards the dependent species and away
from the influencing species. Width of edges is proportional to dependence $q$. Median influence $I$ is lowest in nodes shown in green, and highest in nodes shown in pink; OTUs with medium influence values are plotted in blue, see scale in (a). (c) Undirected and weighted network based on Pearson’s $r$. Edge width is proportional to dependence $r$. Colors of nodes correspond to median influence based on $q$ to visualize different network roles in both networks.
Fig. 6 Association between the birth rate and death rate. The gross domestic product (GDP per capita) of each country is color coded. The death rate is well predicted by the birth rate \((q(\text{birth rate, death rate}) = 0.53, p < 0.001)\). The birth rate is less well predicted by the death rate \((q(\text{death rate, birth rate}) = 0.33, p < 0.001; \text{asymmetry: } a = 0.20, p < 0.001)\). Countries with low GDPs exhibit higher birth rates, and the corresponding death rates seem to be positively correlated with the birth rates; countries with higher GDPs display lower birth rates, and the birth and death rates seem to be negatively correlated. The detection of the relation between the birth and death rates was possible due to the strong and significant asymmetry in the relation, which is higher than that for 99.8% of all variable pairs tested. Detection exclusively based on the strength of the dependence \((qad: \text{mean } q = 0.42; \text{Pearson’s } r = 0.063)\) is unlikely, as the \text{mean } q and r are within the interquartile range of all variable pairs \((\text{mean } q: 0.27 – 0.43; r: 0.05)\).
– 0.28). Thus, the detection of interesting relationships and meaningful explanations is strongly facilitated by qad.
A copula-based measure for quantifying asymmetry in dependence and associations

Robert R. Junker\textsuperscript{1*}, Florian Griessenberger\textsuperscript{2}, Wolfgang Trutschnig\textsuperscript{2}

Supporting information \textsuperscript{1}

1 Mathematical description of qad

The dependence measure $\zeta_1$, introduced in (1), is a copula-based dependence measure and hence scale-invariant. Considering that $\zeta_1$ is based on conditional distributions (Markov kernels) and that estimating conditional distribution is a difficult endeavor, it is a-priori unclear if good estimators can be derived in full generality, i.e. without any regularity assumptions on the underlying two-dimensional distribution function or the underlying copula, at all. Nevertheless we proved that the so-called empirical checkerboard estimator introduced in this paper is strongly consistent for all continuous random variables $X, Y$.

The rest of S1 is organized as follows. Section 2 gathers some preliminaries and notations that will be used throughout S1. Section 3 recalls the definition of empirical copulas which are then aggregated in Section 4 to so-called empirical checkerboard copulas ECC. Our main result saying that calculating $\zeta_1$ of the ECC and choosing an adequate resolution for the aggregation yields a strongly consistent estimator for $\zeta_1(X, Y)$ is derived and proved in Section 4 too. A simulation study for different dependence structures including some extreme cases concludes S1.

2 Notation and preliminaries

Throughout S1 $\mathbb{R}$ will denote the real numbers, $\mathbb{N}$ the natural numbers and $\mathcal{C}$ will denote the family of all two-dimensional copulas (for background on copulas we refer to (2, 3)). For every copula $A \in \mathcal{C}$ the corresponding doubly stochastic measure will be denoted by $\mu_A$. As usual, $d_\infty(A, B)$ will denote the uniform metric on $\mathcal{C}$, i.e.

$$d_\infty(A, B) := \max_{(x,y) \in [0,1]^2} |A(x,y) - B(x,y)|.$$
It is well known that \((C,d_\infty)\) is a compact metric space. For every metric space \((\Omega,d)\) the Borel \(\sigma\)-field will be denoted by \(\mathcal{B}(\Omega)\) and \(\lambda\) will denote the Lebesgue measure on \(\mathcal{B}([0,1])\).

A mapping \(K : \mathbb{R} \times \mathcal{B}(\mathbb{R}) \to [0,1]\) is called a Markov kernel from \(\mathbb{R}\) to \(\mathcal{B}(\mathbb{R})\) if \(x \mapsto K(x,B)\) is measurable for every fixed \(B \in \mathcal{B}(\mathbb{R})\) and \(B \mapsto K(x,B)\) is a probability measure for every fixed \(x \in \mathbb{R}\). A Markov kernel \(K : \mathbb{R} \times \mathcal{B}(\mathbb{R}) \to [0,1]\) is called regular conditional distribution of a (real-valued) random variable \(Y\) given (another random variable) \(X\) if for every \(B \in \mathcal{B}(\mathbb{R})\)

\[ K(X(\omega),B) = E(1_B \circ Y|X)(\omega) \]

holds \(P\)-almost sure, whereby \(1_B(x)\) denotes the indicator function. It is well known that a regular conditional distribution of \(Y\) given \(X\) exists and is unique \(P_X\)-almost sure. For every \(A \in C\) the corresponding regular conditional distribution (i.e. the regular conditional distribution of \(Y\) given \(X\) in the case that \((X,Y) \sim A\)) will be denoted by \(K_A(\cdot,\cdot)\). Note that for every \(A \in C\) and Borel sets \(E,F \in \mathcal{B}([0,1])\) we have

\[
\int_E K_A(x,F)d\lambda(x) = \mu_A(E \times F).
\]

For more details and properties of conditional expectations and regular conditional distributions see (4, 5).

We will mainly work with the metrics \(D_1\) and \(D_\infty\) introduced in (1). These metrics are defined by

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

and

\[
D_\infty(A,B) := \sup_{y \in [0,1]} \Phi_{A,B}(y)
\]

respectively. It can be shown that \(\phi_{A,B}\) is Lipschitz-continuous with Lipschitz constant 2 and that both metrics generate the same topology (without being equivalent). The resulting metric space \((C,D_1)\) is complete and separable and it can be shown that, firstly, \(D_1(A,\Pi)\) attains only values in \([0,\frac{1}{3}]\) and that, secondly, \(D_1(A,\Pi)\) is maximal if and only if \(A\) is completely dependent, i.e. if a \(\lambda\)-preserving transformation \(h : [0,1] \to [0,1]\) exists such that \(K_A(x,\{h(x)\}) = 1\) for \(\lambda\)-a.e.
$x \in [0,1]$. In the sequel we will let $\mathcal{C}_d$ denote the family of all completely dependent copulas, and write $A_h$ and $K_h(\cdot,\cdot)$ for the completely dependent copula and the Markov kernel of the completely dependent copula induced by the $\lambda$-preserving transformation $h$ respectively. For equivalent definitions and properties of completely dependent copulas we refer to (1) and the references therein.

The asymmetric dependence measure $q_{ad}$ calculated in the R-package coincides with $\zeta_1$ defined in (1).

**Definition 2.1** Let $X,Y$ be two continuous random variables with joint distribution function $H$ and copula $A$. Then the dependence measure $\zeta_1$ is defined by

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

As a direct consequence of the properties of $D_1$ it follows immediately that for all continuous random variables $X,Y$ we have $\zeta_1(X,Y) \in [0,1]$, that $\zeta_1(X,Y) = 0$ if and only if $X$ and $Y$ are independent, and that $\zeta_1(X,Y) = 1$ if and only if the copula $A$ of $(X,Y)$ is completely dependent (or, equivalently, if there exists some Borel measurable $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f \circ X$ a.s.).

### 3 Empirical copulas for general samples (possibly having ties)

Let $(X,Y)$ be a random vector with joint distribution $H$, margin distributions $F,G$ and copula $A \in \mathcal{C}$. Furthermore let $(x_1,y_1), \ldots, (x_n,y_n)$ denote a sample of $(X,Y)$, $H_n$ the bivariate empirical distribution function and $F_n,G_n$ the one-dimensional empirical distribution functions. As a consequence of Sklar’s Theorem there exists a unique subcopula $A'_n : \text{Range}(F_n) \times \text{Range}(G_n) \to \text{Range}(H_n)$, which fulfills

$$H_n(x,y) = A'_n(F_n(x), G_n(y)).$$

For an exact definition of subcopulas and their properties we refer to (3). If $X,Y$ are continuous random variables then with probability one the domain of $A'_n$ is the equidistant raster $R_n = \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\}^2$, if not, then ties occur with probability greater than zero. There are uncountably many ways to extend a given (strict) subcopula to a copula - if we only consider the bilinear interpolation/extension $A_n \in \mathcal{C}$ then we can assign every sample $(x_1,y_1), \ldots, (x_n,y_n)$ a unique copula. To this (absolutely continuous checkerboard) copula we will refer to as the empirical copula of the sample $(x_1,y_1), \ldots, (x_n,y_n)$ in the sequel (6, 7).
More precisely we proceed as follows (we only include the description for the sake of completeness): Let \((u_i, v_i) := (F_n(x_i), G_n(y_i)), i = 1, \ldots, n\) denote the pseudo-observations and \((u'_1, v'_1), \ldots, (u'_m, v'_m)\) the distinct pairs of pseudo-observations with \(m \leq n\). Set \(S_1 := \{0, u_1, \ldots, u_m\}, S_2 := \{0, v_1, \ldots, v_m\}\) and define the quantities \(t_i, r_i, s_i\) by

\[
\begin{align*}
t_i & := \sum_{j=1}^{n} 1_{(u'_j, v'_j)}(u_j, v_j) & i = 1, \ldots, m \\
r_i & := \sum_{j=1}^{n} 1_{u_j}(u_j) & i = 1, \ldots, m \\
s_i & := \sum_{j=1}^{n} 1_{v_j}(v_j) & i = 1, \ldots, m.
\end{align*}
\]

Define the empirical subcopula \(A'_n : S_1 \times S_2 \to \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\}\) by

\[
A'_n(s_1, s_2) = \frac{1}{n} \sum_{i=1}^{m} t_i \cdot 1_{[0, s_1] \times [0, s_2]}(u'_i, v'_i) = \frac{1}{n} \sum_{i=1}^{n} 1_{[0, s_1] \times [0, s_2]}(u_i, v_i)
\]

and extend \(A'_n\) bilinearly to a copula \(A_n\) by considering the product copula \(\Pi \in \mathcal{C}\), defining transformations \(w_i : [0, 1]^2 \to [u'_i - \frac{r_i}{n}, u'_i] \times [v'_i - \frac{s_i}{n}, v'_i]\) by

\[
w_i(x, y) = \left( u'_i - \frac{r_i}{n} + \frac{r_i}{n} x, v'_i - \frac{s_i}{n} + \frac{s_i}{n} y \right)
\]

and setting \(\mu_{A_n} := \frac{1}{n} \sum_{i=1}^{m} t_i \mu_{\Pi}^{w_i}\). In the following simple example we calculate the empirical copula of a count-data sample of size 6.

**Example 3.1.** Consider the sample \((10, 10), (6, 3), (5, 1), (6, 4), (4, 1), (6, 3)\) for which we obviously have \(m = 5 < 6 = n\). The distinct pseudo-observations and the quantities \(t_i, r_i, s_i\) are gathered in the subsequent table (Tab. S1.1). As direct consequence we get \(S_1 = \{0, \frac{1}{6}, \frac{2}{6}, \frac{5}{6}, 1\}\) and \(S_2 = \{0, \frac{2}{6}, \frac{3}{6}, \frac{5}{6}, 1\}\).

| \(i\)  | \(1\) | \(2\) | \(3\) | \(4\) | \(5\) |
|-------|-----|-----|-----|-----|-----|
| \((u_i, v_i)\) | \(1, 1\) | \(\frac{5}{6}, \frac{4}{6}\) | \(\frac{5}{6}, \frac{2}{6}\) | \(\frac{5}{6}, \frac{5}{6}\) | \(\frac{1}{6}, \frac{2}{6}\) |
| \(t_i\) | 1 | 2 | 1 | 1 | 1 |
| \(r_i\) | 1 | 3 | 1 | 3 | 1 |
| \(s_i\) | 1 | 2 | 2 | 1 | 2 |

Tab. S1.1: Pseudo-observations and the quantities \(t_i, r_i, s_i\) of example 3.1.

\(\{0, \frac{2}{6}, \frac{3}{6}, \frac{5}{6}, 1\}\). The resulting empirical copula \(A_6\) (i.e. the bilinear extension of \(A'_6\)) is absolutely continuous with the density depicted in Fig. S1.1.
Remark 3.2. The R-package calculates the empirical copula for arbitrary samples - if there are no ties the package calls the function C.n from the copula package, otherwise it calculates $A_n$ as described above.

Remark 3.3. The following results only consider the case of pairs $(X, Y)$ whose two-dimensional distribution function $H$ is continuous and hence the underlying copula unique. Nevertheless qad also produced very good results in the context of random variables with non-trivial discrete components. Our plan is to study the extensibility of the consistency results presented on the next pages and in Section 4 to the general setting in the near future.

The following result is key for proving consistency of the empirical checkerboard copula w.r.t. metric $D_1$ in the next section.

Theorem 3.4. (8) Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be a random sample from $(X, Y)$ and assume that $(X, Y)$ has continuous joint distribution function $H$ and copula $A$. Then the following asymptotic
result holds for the empirical copula $A_n$ with probability 1:

$$d_m(A_n, A) = O\left(\sqrt{\frac{\log(\log(n))}{n}}\right) \quad (1)$$

The following first result shows that working with the empirical copula (without further ado) does not yield a useful estimator of $\zeta_1(A)$.

**Theorem 3.5.** Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be a random sample of $(X, Y)$, where $(X, Y)$ has continuous joint distribution $H$ and copula $A = \Pi$. Then with probability 1 we have

$$\lim_{n \to \infty} D_1(A_n, \Pi) = \frac{1}{3}.$$  

**Proof.** For every $x \in [0, 1]$ there exists exactly one $i(x) \in \{0, 1, \ldots, n\}$ such that

$$K_{A_n}\left(x, \left[\frac{i(x)}{n}, \frac{i(x)+1}{n}\right]\right) = 1.$$  

As direct consequence we get

\[
D_1(A_n, \Pi) = \int_{[0,1]} \int_{[0,1]} |K_{A_n}(x, [0,y]) - y| d\lambda(x) d\lambda(y) 
\geq \int_{[0,1]} \int_{\left[\frac{i(y)}{n}\right]} \left[\frac{i(y)+1}{n}\right] |K_{A_n}(x, [0,y]) - y| d\lambda(y) d\lambda(x) 
= \int_{[0,1]} \left( \int_{\left[\frac{i(y)}{n}\right]} \frac{i(y)}{n} d\lambda(y) + \int_{\left[\frac{i(y)+1}{n}\right]} \frac{1}{n} d\lambda(y) \right) d\lambda(x) 
= \int_{[0,1]} \frac{i(x)^2}{2n^2} + 1 - \frac{i(x)}{n} - \frac{(i(x)+1)^2}{2n^2} d\lambda(x) 
= \frac{1}{2} + \frac{1}{n} + \frac{1}{2n^2} + \int_{[0,1]} \frac{i(x)^2}{n^2} dx - \int_{[0,1]} \frac{i(x)}{n} dx + \int_{[0,1]} \frac{i(x)}{n^2} dx. 
\]

A straightforward calculation shows $\lim_{n \to \infty} P_1^1 = \frac{1}{3}, \lim_{n \to \infty} P_2^1 = \frac{1}{2}$ and $\lim_{n \to \infty} P_3^1 = 0$, from which the desired result follows immediately.

One possibility to overcome this problem is to smooth or aggregate the empirical copula. Aggregation leads to so-called checkerboard copulas.

### 4 Checkerboard copulas

We follow (9) and proceed as follows. Fix $N \in \mathbb{N}$, define the squares $R_{ij}^N$ for $i, j \in \{1, \ldots, N\}$ by

$$R_{ij}^N = \left[\frac{i-1}{N}, \frac{i}{N}\right] \times \left[\frac{j-1}{N}, \frac{j}{N}\right].$$


and let \( \text{int}(R^N_{ij}) \) denote the interior of \( R^N_{ij} \).

**Definition 4.1.** A copula \( A_N \in \mathcal{C} \) is called \( N \)-checkerboard copula, if \( A_N \) is absolutely continuous and (a version of) its density \( k_{A_N} \) is constant on the interior of each square \( R^N_{ij} \). We call \( N \) the resolution of \( A_N \), denote the set of all \( N \)-checkerboard copulas with \( \mathcal{CB}_N \), and set \( \mathcal{CB} = \bigcup_{N=1}^{\infty} \mathcal{CB}_N \).

**Definition 4.2.** For \( A \in \mathcal{C} \) and \( N \in \mathbb{N} \) the (absolute continuous) copula \( \mathcal{CB}_N(A) \in \mathcal{CB}_N \), defined by

\[
\mathcal{CB}_N(A)(x, y) := \int_0^x \int_0^y N^2 \sum_{i,j=1}^N \mu_A(R^N_{ij}) 1_{\text{int}(R^N_{ij})}(s,t) d\lambda(t) d\lambda(s)
\]

is called \( N \)-checkerboard approximation of \( A \) or simply \( N \)-checkerboard of \( A \).

According to (1) \( \mathcal{CB} \) is dense in \( (\mathcal{C}, D_1) \). Furthermore, the following result going back to (9) holds (thereby \( A' \) denotes the transpose of \( A \)).

**Theorem 4.3.** For every copula \( A \in \mathcal{C} \) the following equality holds:

\[
\lim_{N \to \infty} D_1(\mathcal{CB}_N(A), A) = 0 = \lim_{N \to \infty} D_1(\mathcal{CB}_N(A)' , A').
\]

Our simple idea for deriving a 'good' estimator for \( \zeta_1(A) \) is to plug-in the empirical copula \( A_n \) in Equation (2) and consider a resolution \( N \) fulfilling \( N < n \). In the sequel we will refer to \( \mathcal{CB}_N(A_n) \) as empirical \( N \)-checkerboard copula.

**Example 4.4.** Suppose the vector \((X,Y)\) has distribution function \( \Pi \) (in the sequel we will simply write \((X,Y) \sim \Pi \) and that \((x_1,y_1),\ldots,(x_n,y_n)\) is a sample of size \( n = 100 \) from \((X,Y)\). Fig. S1.2 depicts a scatterplot of the sample and Fig. S1.3 depicts the density of the corresponding empirical 5-checkerboard.

The qad packages carries out the following steps for estimating \( \zeta_1(A) \) given a sample \((x_1,y_1),\ldots,(x_n,y_n)\) from \((X,Y)\) with copula \( A \).

1. Calculation of the empirical copula \( A_n \).
2. Selection of an appropriate resolution \( N \) and calculation of the empirical \( N \)-checkerboard \( \mathcal{CB}_N(A_n) \) according to equation (2).
Fig. S1.2: Scatterplot of the sample \((x_1, y_1), \ldots, (x_n, y_n)\) of size \(n = 100\) from \((X, Y) \sim \Pi\).

Fig. S1.3: 3d plot (left panel) and image plot (right panel) of the density of the empirical 5-checkerboard of the sample from Fig. S1.2.

3. Estimation of \(\zeta_1(A)\) in terms of \(\zeta_1(C_B N(A_n)) = 3D_1(C_B N(A_n), \Pi)\).

Choosing an adequate resolution \(N = N(n)\) the estimator \(\zeta_1(C_B N(A_n))\) is strongly consistent - we proceed in several steps to proof this result. We start with the following lemma linking \(D_\infty\).
and $d_\infty$ for checkerboards.

**Lemma 4.5.** For all $A_N, B_N \in C_\mathcal{B}_N$ the following inequality holds:

$$D_\infty(A_N, B_N) \leq 2(N - 1)d_\infty(A_N, B_N).$$

Moreover the inequality is sharp, i.e. for every $N \in \mathbb{N}$ we can find copulas $A_N, B_N \in C_\mathcal{B}_N$ for which equality in (3) holds.

**Proof.** Fix an arbitrary $i \in \{1, \ldots, N\}$. Then for every $x \in \left[\frac{i-1}{N}, \frac{i}{N}\right]$ and $y \in [0, 1]$ continuity of $A$, disintegration and the fact that $K_{A_N}(x,\cdot)$ is constant on each interval of the form $(\frac{i-1}{N}, \frac{i}{N})$ yields

$$A_N\left(\frac{i}{N}, y\right) - A_N\left(\frac{i-1}{N}, y\right) = \mu_{A_N}\left(\left[\frac{i-1}{N}, \frac{i}{N}\right] \times [0, y]\right)$$

$$= \int_{\frac{i-1}{N}}^{\frac{i}{N}} K_{A_N}(s, [0, y]) \, d\lambda(s)$$

$$= \frac{1}{N} K_{A_N}(x, [0, y]).$$

Altogether we get

$$D_\infty(A_N, B_N) = \sup_{y \in [0, 1]} \int_{[0, 1]} |K_{A_N}(x, [0, y]) - K_{B_N}(x, [0, y])| \, d\lambda(x)$$

$$= \sup_{y \in [0, 1]} \sum_{i=1}^{N} \int_{\frac{i-1}{N}}^{\frac{i}{N}} |K_{A_N}(x, [0, y]) - K_{B_N}(x, [0, y])| \, d\lambda(x)$$

$$= \sup_{y \in [0, 1]} \sum_{i=1}^{N} \left| A_N\left(\frac{i}{N}, y\right) - A_N\left(\frac{i-1}{N}, y\right) - B_N\left(\frac{i}{N}, y\right) + B_N\left(\frac{i-1}{N}, y\right) \right|$$

$$= \sup_{y \in [0, 1]} \left\{ \sum_{i=2}^{N-1} \left| A_N\left(\frac{i}{N}, y\right) - A_N\left(\frac{i-1}{N}, y\right) - B_N\left(\frac{i}{N}, y\right) + B_N\left(\frac{i-1}{N}, y\right) \right| + \left| A_N\left(\frac{1}{N}, y\right) - B_N\left(\frac{1}{N}, y\right) \right| + \left| A_N\left(\frac{N-1}{N}, y\right) - B_N\left(\frac{N-1}{N}, y\right) \right| \right\}$$

$$\leq \left( \sum_{i=2}^{N-1} 2d_\infty(A_N, B_N) \right) + 2d_\infty(A_N, B_N)$$

$$= 2(N - 1)d_\infty(A_N, B_N),$$

which completes the proof of ineq. (3).

Fig. S1.4 and Fig. S1.6 depict copulas $A_N, B_N$ for which equality holds. Since the construction idea easily extends to arbitrary $N \in \mathbb{N}$ the proof is complete.
Fig. S1.4: Checkerboard copulas $A_N$ (left) and $B_N$ (right) with $N = 8$ for which equality in Lemma 4.5 holds since $D_{\infty}(A_N, B_N) = 7/8$ and $d_{\infty}(A_N, B_N) = 1/16$ (see Fig. S1.5).

Fig. S1.5: The maps $y \mapsto \Phi_{A_N, B_N}(y)$ (left) and $(x, y) \mapsto |A_N(x, y) - B_N(x, y)|$ (right) for $A_N$ and $B_N$ defined in Fig. S1.4.

As next step we derive a slightly improved inequality linking $D_1$ and $D_{\infty}$ (compare with Theorem 6 in (1) and Lemma 3 in (10)): 
Fig. S1.6: Checkerboard copulas $A_N$ (left) and $B_N$ (right) with $N = 16$ for which equality in Lemma 4.5 holds since $D_\infty(A_N, B_N) = 15/16$ and $d_\infty(A_N, B_N) = 1/32$.

**Lemma 4.6.** Let $N \in \mathbb{N}$ and $A_N, B_N \in C_B$. Then the following inequality holds:

$$D_1(A_N, B_N) \leq \left( \frac{N - 1}{N} \right) D_\infty(A_N, B_N). \quad (3)$$

Inequality (3) is best possible and we have equality if there exists some $c \in [0, \frac{2}{N}]$ such that $\Phi_{A_N, B_N}(y) = c$ on $\left[ \frac{1}{N} \cdot \frac{N-1}{N} \right]$.

**Proof.** Considering the facts that $\Phi_{A_N, B_N}(0) = \Phi_{A_N, B_N}(1) = 0$, that $y \mapsto \Phi_{A_N, B_N}(y)$ is piecewise linear and Lipschitz-continuous with Lipschitz-constant 2 we get

$$\max_{y \in [0,1]} \Phi_{A_N, B_N}(y) = \max_{y \in \left[ \frac{1}{N} \cdot \frac{N-1}{N} \right]} \Phi_{A_N, B_N}(y),$$

from which the assertions follow.

Considering $N$-checkerboard approximations of arbitrary copulas can not increase their $d_\infty$-distance - the following lemma holds:

**Lemma 4.7.** For arbitrary $A, B \in \mathbb{C}$ and their $N$-checkerboard approximations $C_B N(A)$, $C_B N(B) \in \mathbb{C}_B$ the following inequality holds:

$$d_\infty(C_B N(A), C_B N(B)) \leq d_\infty(A, B). \quad (4)$$
The inequality is sharp for every $N \in \mathbb{N}$.

Proof. The inequality is easily derived via

$$
\begin{align*}
d_{\infty}(C_{N}(A), C_{N}(B)) &= \sup_{(x,y) \in \{0,1\}^2} |C_{N}(A)(x,y) - C_{N}(B)(x,y)| \\
&= \sup_{(x,y) \in \{\frac{k}{N}, \ldots, \frac{k+1}{N}\}^2} |C_{N}(A)(x,y) - C_{N}(B)(x,y)| \\
&= \sup_{(x,y) \in \{\frac{k}{N}, \ldots, \frac{k+1}{N}\}^2} |A(x,y) - B(x,y)| \\
&\leq \sup_{(x,y) \in \{0,1\}^2} |A(x,y) - B(x,y)| \\
&= d_{\infty}(A,B).
\end{align*}
$$

Considering the fact that for $A,B \in C_{N}$ we obviously have $C_{N}(A) = A$ and $C_{N}(B) = B$, obviously the inequality is best possible.

**Corollary 4.8.** For arbitrary $A,B \in C$ and their $N$-checkerboard approximations $C_{N}(A), C_{N}(B) \in C\mathcal{B}$ the following inequality holds:

$$
D_1(C_{N}(A), C_{N}(B)) \leq 2 \frac{(N-1)^2}{N} d_{\infty}(A,B). \quad (5)
$$

We finally arrive at the main result saying that the empirical checkerboard is a strongly consistent estimator of the underlying copula w.r.t. metric $D_1$ - this result is key for finally proving consistency of $\zeta_1(C_{N}(A_n))$.

**Theorem 4.9.** Let $(X_1,Y_1), (X_2,Y_2), \ldots$ be a random sample from $(X,Y)$ and assume that $(X,Y)$ has continuous joint distribution function $H$ and copula $A$. Setting $N(n) := [n^s]$ for some $s \in (0, \frac{1}{2})$ the following identity holds with probability 1:

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

Proof. Applying Corollary 4.8 and the triangle inequality yields

$$
\begin{align*}
D_1(C_{N(n)}(A_n), A) &\leq D_1(C_{N(n)}(A_n), C_{N(n)}(A)) + D_1(C_{N(n)}(A), A) \\
&\leq 2 \frac{(N(n)-1)^2}{N(n)} d_{\infty}(A_n, A) + D_1(C_{N(n)}(A), A) \\
&\leq 2N(n) d_{\infty}(A_n, A) + D_1(C_{N(n)}(A), A)
\end{align*}
$$
According to Theorem 3.4 there exists a $\Lambda \in \mathcal{A}$ with $\mathbf{P}(\Lambda) = 1$ such that for every $\omega \in \Lambda$ we can find a constant $C(\omega) > 0$ and an index $n_0 = n_0(\omega) \in \mathbb{N}$ such that for all $n \geq n_0$

$$d_\omega(A_n(\omega), A) \leq C(\omega) \sqrt{\frac{\log(\log(n))}{n}}$$

holds. Let $\varepsilon > 0$ be fixed. Theorem 4.3 implies the existence of an index $n_1 \in \mathbb{N}$ fulfilling

$$D_1(C B_{N(n)}(A_n), A) < \frac{\varepsilon}{2}$$

for all $n \geq n_1$. For every $\omega \in \Lambda$ and $n \geq \max\{n_0, n_1\}$ we finally obtain

$$D_1(C B_{N(n)}(A_n(\omega)), A) \leq 2 N(n) C(\omega) \sqrt{\frac{\log(\log(n))}{n}} + D_1(C B_{N(n)}(A_n), A)$$

$$< 2 C(\omega) |n^s| n^{-\frac{1}{2}} \sqrt{\log(\log(n))} + \frac{\varepsilon}{2}$$

from which the result follows since $0 < s < \frac{1}{2}$ by assumption.

The simulations in the next section insinuate that $C B_{N(n)}(A_n)$ might also be a strongly consistent estimator for more flexible choices of $N(n)$, particularly for the case $N(n) := \lfloor n^r \rfloor$ and some $s \geq \frac{1}{2}$ - a clarification of this question is future work, in the R-package we considered $s = \frac{1}{2}$. We conclude this section with the main result on the estimator $\zeta_1(C B_{N(n)}(A_n))$ of $\zeta_1(A)$ - this very result is the main reason for using the qad-package.

**Theorem 4.10** Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be a random sample from $(X, Y)$ and assume that $(X, Y)$ has continuous joint distribution function $H$ and copula $A$. Fix $s \in (0, \frac{1}{2})$ and set $N(n) := \lfloor n^s \rfloor$ for every $n \in \mathbb{N}$. Then with probability 1 we have

$$\lim_{n \to \infty} \zeta_1(C B_{N(n)}(A_n)) = \zeta_1(A).$$

**Proof.** Direct consequence of Theorem 4.9 and the fact that

$$|D_1(C B_{N(n)}(A_n), \Pi) - D_1(A, \Pi)| \leq D_1(C B_{N(n)}(A_n), A).$$

**5 Simulations**

If not specified differently throughout this section we consider $s_0 = \frac{1}{2}$ and set $N(n) = \lfloor n^{s_0} \rfloor$. In order to illustrate the performance of the estimator $\zeta_1(C B_{N(n)}(A_n))$ we consider Marshall-Olkin
and FGM copulas as well as a completely dependent copula $A_h$ fulfilling that $\zeta_1(A_h)$ is much bigger than $\zeta_1(A_h')$.

**Example 5.1.** [Marshall Olkin family] The Marshall Olkin (MO) family of copulas $(M_{\alpha,\beta})_{\alpha,\beta\in[0,1]}$ is defined by

$M_{\alpha,\beta}(x,y) := \begin{cases} x^{1-\alpha} y & x^\alpha \geq y^\beta \\ x^{1-\beta} & x^\alpha < y^\beta. \end{cases}$

It contains $\Pi$ ($\alpha = 0$ or $\beta = 0$) as well as $M$ ($\alpha = \beta = 1$). It was shown in (1), that in case of $\alpha, \beta > 0$

$$\zeta_1(M_{\alpha,\beta}) = 3\alpha(1 - \alpha)^2 + \frac{6}{\beta} \frac{1 - (1 - \alpha)^2}{z} - \frac{6}{\beta} \frac{1 - (1 - \alpha)^{2+1}}{z+1},$$

whereby $z = \frac{1}{\alpha} + \frac{2}{\beta} - 1$.

To test the performance of qad for different dependence structures we consider MO copulas with parameter $(\alpha, \beta) \in \{(1, 0), (1, 1), (0.3, 1), (1, 0.7), (0.5, 0.5)\}$, see Fig. S1.7, Fig. S1.9, Fig. S1.11, Fig. S1.13 and Fig. S1.15.

We generated samples of size $n \in \{10, 50, 100, 500, 1000, 5000, 10000\}$, calculated the empirical checkerboard copula $A_n$ as well as $\zeta_1(CB_{N(n)}(A_n))$. These steps were repeated $R = 500$ times and the obtained results are depicted as boxplots in Fig. S1.8, Fig. S1.10, Fig. S1.12, Fig. S1.14 and Fig. S1.16.

**Example 5.2.** [FGM family] The Farlie-Gumbel-Morgenstern family $(G_\theta)_{\theta\in[-1,1]}$ is defined by

$G_\theta(x,y) := xy + \theta xy(1-x)(1-y)$.

According to (1) $\zeta_1(G_\theta)$ is given by

$$\zeta_1(G_\theta) = \frac{|\theta|}{4}.$$

Replacing the Marshall-Olkin copula by a Farlie-Gumbel-Morgenstern copula and proceeding analogously as before yields in the special case with $\theta \in \{-1, -0.5\}$ the results, depicted in Fig. S1.17, Fig. S1.18, Fig. S1.19 and Fig. S1.20.

**Example 5.3.** [Highly asymmetric completely dependent copulas] To test the performance of qad in a highly asymmetric situation we consider the completely dependent copula $A_{h_a}$ for $h = ax(mod 1)$ and $a = 5, 10, 50$, see Fig. S1.21, Fig. S1.23 and Fig. S1.25. Notice that in these
cases we have $\zeta_1(A_{h_a}) = 1$ whereas $\zeta_1(A'_{h_a})$ is (particularly for large $a$) close to 0. For fixed $a \in \{5, 10, 50\}$ we generated samples of size $n \in \{10, 50, 100, 500, 1.000, 5.000, 10.000\}$, calculated the empirical checkerboard copula $A_n$ as well as $\zeta_1(\mathcal{CB}(A_n))$. These steps were repeated $R = 1.000$ times. The obtained results are depicted as boxplots in Fig. S1.22, Fig. S1.24 and Fig. S1.26 respectively. As expected, the parameter $a$ has a big influence on the precision of the estimator, the bigger $a$ the longer it takes qad to detect the asymmetry. Notice that, contrary to qad, classical dependence measures like Schweizer and Wolff’s $\sigma$ (11) are not capable of detecting the asymmetry since both $d_\infty(A_{h_a}, \Pi)$ and $d_\infty(A_{h_a'}, \Pi)$ are very small.
Fig. S1.7: Sample of size 10,000 from a MO copula with parameter $\alpha = 1$ and $\beta = 0$.

Fig. S1.8: Boxplots summarizing the 500 obtained estimates for $\zeta_1(M_{\alpha, \beta})$ (magenta) and $\zeta_1(M'_{\alpha, \beta})$ (gray). The dashed lines depict the true dependence measure $\zeta_1(M_{\alpha, \beta})$ and $\zeta_1(M'_{\alpha, \beta})$ for $\alpha = 1$ and $\beta = 0$. 
Fig. S1.9: Sample of size 10,000 from a MO copula with parameter $\alpha = 1$ and $\beta = 1$.

Fig. S1.10: Boxplots summarizing the 500 obtained estimates for $\zeta_1(M_{\alpha,\beta})$ (magenta) and $\zeta_1(M'_{\alpha,\beta})$ (gray). The dashed lines depict the true dependence measure $\zeta_1(M_{\alpha,\beta})$ and $\zeta_1(M'_{\alpha,\beta})$ for $\alpha = 1$ and $\beta = 1$. 
Fig. S1.11: Sample of size 10,000 from a MO copula with parameter $\alpha = 0.3$ and $\beta = 1$.

Fig. S1.12: Boxplots summarizing the 500 obtained estimates for $\zeta_1(M_{\alpha,\beta})$ (magenta) and $\zeta_1(M'_{\alpha,\beta})$ (gray). The dashed lines depict the true dependence measure $\zeta_1(M_{\alpha,\beta})$ and $\zeta_1(M'_{\alpha,\beta})$ for $\alpha = 0.3$ and $\beta = 1$. 
Fig. S1.13: Sample of size 10,000 from a MO copula with parameter $\alpha = 1$ and $\beta = 0.7$.

Fig. S1.14: Boxplots summarizing the 500 obtained estimates for $\zeta_1(M_{\alpha,\beta})$ (magenta) and $\zeta_1(M'_{\alpha,\beta})$ (gray). The dashed lines depict the true dependence measure $\zeta_1(M_{\alpha,\beta})$ and $\zeta_1(M'_{\alpha,\beta})$ for $\alpha = 1$ and $\beta = 0.7$. 
Fig. S1.15: Sample of size 10,000 from a MO copula with parameter $\alpha = 0.5$ and $\beta = 0.5$.

Fig. S1.16: Boxplots summarizing the 500 obtained estimates for $\zeta_1(M_{\alpha,\beta})$ (magenta) and $\zeta_1(M'_{\alpha,\beta})$ (gray). The dashed lines depict the true dependence measure $\zeta_1(M_{\alpha,\beta})$ and $\zeta_1(M'_{\alpha,\beta})$ for $\alpha = 0.5$ and $\beta = 0.5$. 

Fig. S1.17: Sample of size 10,000 from a FGM copula with parameter $\theta = -1$.

Fig. S1.18: Boxplots summarizing the 500 obtained estimates for $\zeta_1(G_\theta)$ (magenta) and $\zeta_1(G'_\theta)$ (gray). The dashed lines depict the true dependence measure $\zeta_1(G_\theta)$ and $\zeta_1(G'_\theta)$ for $\theta = -1$. 
Fig. S1.19: Sample of size 10,000 from a FGM copula with parameter $\theta = -0.5$.

Fig. S1.20: Boxplots summarizing the 500 obtained estimates for $\zeta_1(G_\theta)$ (magenta) and $\zeta_1(G'_\theta)$ (gray). The dashed lines depict the true dependence measure $\zeta_1(G_\theta)$ and $\zeta_1(G'_\theta)$ for $\theta = -0.5$. 
Fig. S1.21: Support of $A_{h_o}$ for $a = 5$.

Fig. S1.22: Boxplots summarizing the 1000 obtained estimates for $\zeta_1(A_{h_o})$ (magenta) and $\zeta_1(A_{h_o}^t)$ (gray) for the case $a = 5$. 

Sample size
Measure
$\zeta_1(A_{h_o})$
$\zeta_1(A_{h_o}^t)$
Fig. S1.23: Support of $A_{h_a}$ for $a = 10$.

Fig. S1.24: Boxplots summarizing the 1000 obtained estimates for $\zeta_1(A_{h_a})$ (magenta) and $\zeta_1(A'_{h_a})$ (gray) for the case $a = 10$. 

24
Fig. S1.25: Support of $A_{h_a}$ for $a = 50$.

Fig. S1.26: Boxplots summarizing the 1000 obtained estimates for $\zeta_1(A_{h_a})$ (magenta) and $\zeta_1(A_{h_a}^I)$ (gray) for the case $a = 50$. 
References

[1] Trutschnig W (2011) On a strong metric on the space of copulas and its induced dependence measure. *Journal of Mathematical Analysis and Applications* 384(2):690 – 705.

[2] Durante F, Sempi C (2015) *Principles of copula theory*. (CRC Press, Hoboken, NJ).

[3] Nelsen RB (2006) *An Introduction to Copulas (Springer Series in Statistics)*. (Springer-Verlag, Berlin, Heidelberg).

[4] Kallenberg O (2002) *Foundations of modern probability*, Probability and its Applications (New York). (Springer-Verlag, New York).

[5] Klenke A (2008) *Wahrscheinlichkeitstheorie*, Springer-Lehrbuch Masterclass Series. (Springer Berlin Heidelberg).

[6] Deheuvels P (1979) La fonction de dépendance empirique et ses propriétés, un test non paramétrique d’indépendance. *Académie Royale de Belgique, Bulletin de la Classe des Sciences, 5e série* 65(6):274–292.

[7] Genest C, Nešlehová JG, Rémillard B (2014) On the empirical multilinear copula process for count data. *Bernoulli* 20(3):1344–1371.

[8] Janssen P, Swanepoel J, Veraverbeke N (2012) Large sample behavior of the bernstein copula estimator. *Journal of Statistical Planning and Inference* 142:1189 – 1197.

[9] Li X, Mikusiński P, Taylor M (1998) Strong approximation of copulas. *Journal of Mathematical Analysis and Applications* 225(2):608 – 623.

[10] Fernández Sánchez J, Trutschnig W (2015) Conditioning-based metrics on the space of multivariate copulas and their interrelation with uniform and levelwise convergence and iterated function systems. *Journal of Theoretical Probability* 28(4):1311–1336.

[11] Schweizer B, Wolff EF (1981) On nonparametric measures of dependence for random variables. *Ann. Statist.* 9(4):879–885.
A copula-based measure for quantifying asymmetry in dependence and associations

Robert R. Junker1*, Florian Griessenberger2, Wolfgang Trutschnig2

Supporting information 2

Parabola [Fig. 1]
We generated $n$ data points from a quadratic function with predefined noise ($y = x^2 + \text{noise}$; Supplementary material 4 – Fig. 1). Noise is generated by drawing random numbers from a uniform distribution on the interval $[-a, a]$. Original data and pseudo-observations (normalized ranks) were plotted. Based on pseudo-observations, the empirical checkerboard copula was calculated using the function $qad()$ provided in the R-package $qad$, which also serves as a tool to predict $Y$ given $X$ and vice versa.

Dependence between two random variables [Fig. 2]
In order to demonstrate the properties of $qad$ and to compare it to other methods quantifying dependence, we generated $n$ data points from nine different dependence structures with increasing vertical noises (Supplementary material 4 – Fig. 2) and quantified dependence using $qad$, Pearson’s $r$, Spearman’s $\rho$, and the maximum information coefficient MIC ($1$). Noise was generated by drawing random numbers from a uniform distribution on the interval $[-a, a]$. For each dependence structure, we increased the noise between 0 and 1 in $1/99$ increments, and for each noise value we repeated each simulation $n = 10$ times, resulting in $n = 1,000$ simulations per dependence structure.

CauseEffectPairs
We quantified the asymmetric dependence of $n = 74$ CauseEffectPairs (2, 3) to evaluate the causal discovery rate of $qad$. A total of $n = 107$ CauseEffectPairs can be downloaded (http://webdav.tuebingen.mpg.de/cause-effect/) from which we excluded $n = 33$ with too many ties (double entries) preventing a precise estimation of the empirical checkerboard copula (see Supplementary material 1). For each of the datasets the ground truth, i.e. whether $y$ depends on $x$ or vice versa, is given. We quantified the asymmetric dependence $q(X, Y)$ and $q(Y, X)$ as well as the asymmetry $a = q(X, Y) - q(Y, X)$ for each CauseEffectPair (Supplementary material 4). In the case that $q(X, Y) > q(Y, X)$ and the ground truth is a dependence of $Y$ on $X$, we scored a correct causal discovery (Tab. S2_1). The same is true for $q(X, Y) < q(Y, X)$ if the ground truth is a dependence of $X$ on $Y$. Directional dependency of CauseEffectPairs was mostly not significantly asymmetric (only in 5 out of 74 cases). In these cases, however, causal discovery rate was 80%. These results confirm that statistical dependence is not the same concept as causality, but $qad$ may help to infer causality. Nevertheless, expert knowledge is required to correctly interpret the data and to discuss the cause of (a)symmetric dependencies.

Tab. S2_1 Causal discovery by $qad$. For $n = 74$ CauseEffectPairs we quantified the asymmetric dependence $q(X, Y)$ and $q(Y, X)$ and asymmetry in dependence $a$ between $X$ and $Y$ using $qad$. 

1
Significant dependencies and asymmetries are listed in bold. Ground truth is provided for each dataset. In the case that \( q(X, Y) > q(Y, X) \) and the ground truth is a dependence of \( Y \) on \( X \), we scored a correct causal discovery (TRUE), otherwise \( qad \) failed to discover causality (FALSE). The same is true if \( q(X, Y) < q(Y, X) \) and the ground truth is a dependence of \( X \) on \( Y \).

| CauseEffectPair | \( n \) | \( q(X, Y) \) | \( q(Y, X) \) | \( a \) | ground truth | correct causal discovery |
|-----------------|-------|----------------|----------------|-----|---------------|---------------------------|
| 1               | 348   | 0.627          | 0.629          | -0.002 | X --> Y       | FALSE                     |
| 2               | 348   | 0.588          | 0.565          | 0.023 | X --> Y       | TRUE                      |
| 3               | 348   | 0.316          | 0.348          | -0.032 | X --> Y       | FALSE                     |
| 4               | 348   | 0.255          | 0.259          | -0.004 | X --> Y       | FALSE                     |
| 13              | 391   | 0.726          | 0.712          | 0.014 | X --> Y       | TRUE                      |
| 14              | 391   | 0.686          | 0.690          | -0.005 | X --> Y       | FALSE                     |
| 15              | 391   | 0.716          | 0.730          | -0.014 | X --> Y       | FALSE                     |
| 16              | 313   | 0.783          | 0.760          | 0.023 | X --> Y       | TRUE                      |
| 19              | 193   | 0.704          | 0.699          | 0.006 | X --> Y       | TRUE                      |
| 20              | 348   | 0.334          | 0.427          | -0.094 | X --> Y       | FALSE                     |
| 21              | 348   | 0.353          | 0.352          | 0.001 | X --> Y       | TRUE                      |
| 22              | 449   | 0.143          | 0.086          | 0.057 | X --> Y       | TRUE                      |
| 23              | 451   | 0.324          | 0.288          | 0.035 | X --> Y       | TRUE                      |
| 24              | 450   | 0.139          | 0.112          | 0.027 | X --> Y       | TRUE                      |
| 25              | 1029  | 0.373          | 0.358          | 0.015 | X --> Y       | TRUE                      |
| 26              | 1029  | 0.239          | 0.173          | 0.066 | X --> Y       | TRUE                      |
| 27              | 1029  | 0.111          | 0.111          | -0.001 | X --> Y       | FALSE                     |
| 28              | 1029  | 0.376          | 0.296          | 0.080 | X --> Y       | TRUE                      |
| 29              | 1029  | 0.276          | 0.264          | 0.013 | X --> Y       | TRUE                      |
| 30              | 1029  | 0.246          | 0.204          | 0.042 | X --> Y       | TRUE                      |
| 31              | 1029  | 0.220          | 0.193          | 0.027 | X --> Y       | TRUE                      |
| 39              | 393   | 0.217          | 0.224          | -0.007 | X --> Y       | FALSE                     |
| 43              | 10368 | 0.954          | 0.951          | 0.003 | X --> Y       | TRUE                      |
| 44              | 10368 | 0.904          | 0.898          | 0.006 | X --> Y       | TRUE                      |
| 45              | 10368 | 0.786          | 0.776          | 0.010 | X --> Y       | TRUE                      |
| 48              | 167   | 0.404          | 0.445          | -0.041 | Y --> X       | TRUE                      |
| 49              | 364   | 0.593          | 0.591          | 0.002 | Y --> X       | FALSE                     |
| 50              | 364   | 0.559          | 0.514          | 0.045 | Y --> X       | FALSE                     |
| 51              | 364   | 0.344          | 0.370          | -0.026 | Y --> X       | TRUE                      |
| 52              | 10225 | 0.942          | 0.944          | -0.003 | Y --> X       | TRUE                      |
| 53              | 988   | 0.873          | 0.870          | 0.003 | Y --> X       | FALSE                     |
| 54              | 391   | 0.860          | 0.854          | 0.006 | X --> Y       | TRUE                      |
| 55              | 71    | 0.222          | 0.294          | -0.073 | Y --> X       | TRUE                      |
| 56              | 191   | 0.400          | 0.413          | -0.013 | Y --> X       | TRUE                      |
| 57              | 191   | 0.354          | 0.390          | -0.036 | Y --> X       | TRUE                      |
| 58              | 191   | 0.368          | 0.397          | -0.029 | Y --> X       | TRUE                      |
| 59              | 191   | 0.385          | 0.407          | -0.022 | Y --> X       | TRUE                      |
|   |   |   |   |        |        |        |   |
|---|---|---|---|--------|--------|--------|---|
|60 |191 |0.325  |0.380  |-0.056  |Y --> X  |TRUE    |
|61 |191 |0.343  |0.373  |-0.030  |Y --> X  |TRUE    |
|62 |191 |0.361  |0.364  |-0.003  |Y --> X  |TRUE    |
|63 |191 |0.348  |0.375  |-0.026  |Y --> X  |TRUE    |
|64 |161 |0.687  |0.675  |0.012   |X --> Y  |TRUE    |
|65 |1330|0.429  |0.418  |0.011   |X --> Y  |TRUE    |
|66 |1330|0.654  |0.652  |0.002   |X --> Y  |TRUE    |
|67 |1330|0.561  |0.564  |-0.003  |X --> Y  |FALSE   |
|68 |497 |0.790  |0.775  |0.015   |Y --> X  |FALSE   |
|72 |1631|0.199  |0.189  |0.010   |X --> Y  |TRUE    |
|73 |5083|0.865  |0.864  |0.001   |Y --> X  |FALSE   |
|74 |193 |0.659  |0.678  |-0.019  |X --> Y  |FALSE   |
|75 |204 |0.737  |0.743  |-0.006  |X --> X  |TRUE    |
|76 |346 |0.727  |0.755  |-0.028  |X --> Y  |FALSE   |
|78 |720 |0.758  |0.737  |0.022   |X --> Y  |TRUE    |
|79 |720 |0.720  |0.706  |0.014   |Y --> X  |FALSE   |
|80 |720 |0.616  |0.626  |-0.010  |Y --> X  |TRUE    |
|81 |364 |0.638  |0.644  |-0.007  |X --> Y  |FALSE   |
|82 |364 |0.679  |0.662  |0.016   |X --> Y  |TRUE    |
|83 |364 |0.760  |0.758  |0.001   |X --> Y  |TRUE    |
|84 |3101|0.940  |0.938  |0.002   |Y --> X  |FALSE   |
|88 |260 |0.512  |0.465  |0.048   |X --> Y  |TRUE    |
|89 |130 |0.546  |0.548  |-0.002  |Y --> X  |TRUE    |
|90 |125 |0.379  |0.395  |-0.016  |Y --> X  |TRUE    |
|91 |148 |0.668  |0.682  |-0.014  |X --> Y  |FALSE   |
|92 |149 |0.789  |0.787  |0.002   |Y --> X  |FALSE   |
|93 |431 |0.708  |0.696  |0.012   |X --> Y  |TRUE    |
|97 |201 |0.278  |0.286  |-0.008  |X --> Y  |FALSE   |
|98 |93  |0.788  |0.772  |0.016   |X --> Y  |TRUE    |
|100|208 |0.501  |0.522  |-0.020  |X --> Y  |FALSE   |
|101|299 |0.891  |0.891  |0.000   |X --> Y  |FALSE   |
|102|108 |0.475  |0.382  |0.094   |X --> Y  |TRUE    |
|103|108 |0.155  |0.178  |-0.022  |X --> Y  |FALSE   |
|104|108 |0.278  |0.285  |-0.007  |X --> Y  |FALSE   |
|105|999 |0.556  |0.551  |0.005   |X --> Y  |TRUE    |
|106|113 |0.756  |0.752  |0.003   |Y --> X  |FALSE   |
Global climate [Fig. 4]

At times of global warming and changes in precipitation regimes, information on past, present, and future climate at a local scale is demanded to assess the effect of climate on the environment. Such data can be retrieved from databases that provide climate estimates in high geographic resolution. Commonly, estimates of temperature and precipitation are used to derive 19 bioclimatic variables that inform about climate features relevant to biological processes. Since these variables are based on a limited set of data, there is a strong underlying dependence structure between them, which, in particular, prevents their independent use in statistical models. We retrieved bioclimatic variables for \( n = 1862 \) locations homogenously distributed over the global landmass from CHELSA (4, 5) to test for asymmetric dependence between them (Fig. 4). Using these data we calculated \( q(X, Y) \), \( q(Y, X) \), (i.e. the mean dependence), and (i.e. the asymmetry of the dependence) and calculated the corresponding \( p \)-values informing about the significance of dependence and asymmetry (Supplementary material 4). Additionally, Pearson’s \( r^2 \) was calculated. Heatmaps visualize the dependencies between the 19 bioclimatic variables based on \( q_{ad} \) (Fig. S2_2, asymmetry \( |a| \) is shown in Fig. S2_3) and Pearson’s \( r^2 \) (Fig. S2_4).

As expected, the knowledge of one of the bioclimatic variables reduces the variability of the other variables, which is indicated by the high dependence between the pairs of variables (mean \( q \pm \text{sd} = 0.54 \pm 0.18 \)). Pearson’s product moment correlation indicated weaker dependencies (mean \( \pm \text{sd}: 0.32 \pm 0.29 \)). Many of the associations between the bioclimatic variables turned out to be asymmetric (mean asymmetry \( |a| \pm \text{sd} = 0.06 \pm 0.05 \), range = 0.00 – 0.23). Particularly, relationships between bioclimatic variables informing about temperature and those informing about precipitation are not related linearly (Fig. S2_2 and S2_4). For instance, the annual precipitation can be better predicted by mean temperature (\( q = 0.61, p < 0.001 \)) than vice versa (\( q = 0.54, p < 0.001 \), asymmetry \( a = 0.08, p < 0.001 \), Fig. 4). In general, precipitation related bioclimatic variables were better predictable than temperature related variables (Fig. S2_2).
Fig. S2.2 Asymmetric dependence $q$ between 19 bioclimatic variables measured at $n = 1862$ locations homogenously distributed over the global landmass. The heatmap depicts $q($variable on the y-axis, variable on the x-axis$)$, which is the dependence of the column-variable on the row-variable. Abbreviations: AMT: Annual Mean Temperature; MDR: Mean Diurnal Range; IT: Isothermality; TS: Temperature Seasonality; MTWM: Max Temperature of Warmest Month; MTCM: Min Temperature of Coldest Month; TAR: Temperature Annual Range; MTWeQ: Mean Temperature of Wettest Quarter; MTCQ: Mean Temperature of Coldest Quarter; MTWaQ: Mean Temperature of Warmest Quarter; MTDQ: Mean Temperature of Driest Quarter; MTCM: Min Temperature of Coldest Month; TAR: Temperature Annual Range; MTWeQ: Mean Temperature of Wettest Quarter; MTCQ: Mean Temperature of Coldest Quarter; MTWaQ: Mean Temperature of Warmest Quarter; PCQ: Precipitation of Coldest Quarter.
Asymmetry $a$ in dependence $q$ between 19 bioclimatic variables measured at $n = 1862$ locations homogenously distributed over the global landmass. Since $|a|$ is symmetric, the lower triangle of the heatmap differs only in the sign from the upper triangle of the heatmap. For abbreviations see Fig. S2_2.
Fig. S2_4 Dependence based on Pearson’s $r^2$ between 19 bioclimatic variables measured at $n = 1862$ locations homogenously distributed over the global landmass. By construction $r^2$ is symmetric, therefore the matrix is symmetric, too. For abbreviations see Fig. S2_2.

Using mean annual temperature and annual precipitation as examples, we demonstrate the properties of qad to quantify asymmetric dependence of bivariate associations. The observations (Fig. S2_5a) are transformed to pseudo-observations (normalized ranks, Fig. S2_5b), which are the basis for the empirical checkerboard copula (as estimator of the true underlying dependence structure, Fig. S2_5c). The resolution $N$ of the checkerboard copula (number of stripes per dimension) depends on the sample size and is defined as number of stripes $s = r^{1/2}$, with $r$ as the number of ranks per variable (note that the number of different ranks $r$ does not necessarily coincide with the number of samples $n$). In this example, the lower number of ranks of the two variables was $r = 696$, resulting in a resolution of $696^{1/2} = N = 26$ (i.e. Fig. S2_5c displays 26 horizontal and vertical stripes). For each vertical or horizontal stripe, the mass distribution of the empirical checkerboard copula is given and sums up to 1. As a consequence the empirical checkerboard copula can be used to predict $Y$ given the value of $X$ and vice versa. In our example the annual precipitation can be better predicted by mean temperature ($q(AMT, AP) = 0.61$, $p < 0.001$) than vice versa ($q(AP, AMT) = 0.54$, $p < 0.001$; asymmetry $a = 0.08$, $p < 0.001$, Fig. S2_5c).
**Fig. S2_5** Relationship between mean annual temperature and annual precipitation. The observations (a) are transformed to pseudo-observations (normalized ranks, b), which are the basis for the empirical checkerboard copula (as estimator of the true underlying dependence structure, c).

**Microbiome [Fig. 5]**

We used a dataset of bacteria associated with surfaces of the plant *Metrosideros polymorpha* (6) (study accession number of sequences at the European Nucleotide Archive: PRJEB7828, [http://www.ebi.ac.uk/ena/data/view/PRJEB7828](http://www.ebi.ac.uk/ena/data/view/PRJEB7828)) and tested for asymmetric relationships qad between pairs of $n = 93$ operational taxonomic units (OTUs) that were observed in at least 75% of all samples ($n = 125$) (Supporting Information 4 – Microbiomes Fig. 5). The quantification of asymmetric dependencies allows a novel definition of key species $S_f$, which influence the abundance of other species $S_j$ but are less influenced by others. Thus, key species $S_i$ influence other species $S_j$ on average to a larger extend than they are dependent on other species $S_i$. Median influence of species $S_i$ is calculated as the median of $l_i$ -values calculated as $I^j_i = q(S_f, S_j) - q(S_j, S_f), 1 \leq j \leq n, j \neq f$. If median ($l_i$) > 0 species $S_i$ influences most other species $S_j$ stronger than vice versa. As mentioned before, $q(S_f, S_j)$ denotes the dependence of $S_j$ on $S_f$ and $q(S_j, S_f)$ denotes the dependence in the other direction (Tab. S2_2, Supporting Information 4 – Microbiomes Fig. 5).

Networks and network indices were calculated using the R-package *igraph* (7). $q$-values and $r$-values below 0.325 were set to 0 in order to restrict network graphs to relatively strong dependencies only. Likewise, only significant ($p < 0.05$) dependencies were included in the networks. Weighted and directed network visualizing asymmetric dependencies $q$ between OTUs confirmed the prominent role of the key-OTUs in shaping the abundances of a number of other OTUs. Most of the seven key-OTUs occupy central positions (Tab. S2_2) indicating their role as network hubs (Tab. S2_2). In contrast, in an undirected and weighted network based on Pearson’s $r^2$, key-OTUs defined by qad occupied peripheral positions (Tab. S2_2). Accordingly, these OTUs were not identified as hubs (Tab. S2_2).
Tab. S2_2 Influence and network positions of OTUs associated with surfaces of the plant *Metrosideros polymorpha*. For each OTU, the median \( l_i \), betweenness, degree, and hub score based on \( q \)-values is given. Betweenness, degree, and hub score are also given based on \( r \)-values. OTUs are ordered by increasing median \( l_i \).

| OTU   | Family       | Genus              | median \( l_i \) | betweenness\_quad | degree\_quad | hub\_score\_quad |
|-------|--------------|--------------------|------------------|-------------------|--------------|-----------------|
| 3820  | Moraxellaceae| Acinetobacter      | -0.02795         | 0                  | 0            | 0               |
| 936   | Oxalobacteraceae| NA                | -0.01985         | 0                  | 47           | 5               |
| 23    | Bradyrhizobiaceae| Bradyrhizobium    | -0.01836         | 241                | 204          | 8               |
| 321   | Microbacteriaceae| Curtobacterium    | -0.01714         | 0                  | 0            | 0               |
| 7     | Caulobacteraceae| NA                | -0.01667         | 25                 | 14           | 6               |
| 12    | Hyphomicrobiaceae| NA                | -0.01654         | 111                | 169          | 6               |
| 76    | Sphingomonadaceae| Sphingomonas      | -0.01640         | 29                 | 5            | 4               |
| 15    | Comamonadaceae| NA                | -0.01584         | 0                  | 1            | 2               |
| 13    | Oxalobacteraceae| NA                | -0.01534         | 189                | 124          | 6               |
| 3     | Weekellsaceae| Chryseobacterium   | -0.01393         | 162                | 0            | 11              |
| 4     | Acetobacteraceae| Gluconobacter     | -0.01217         | 0                  | 0            | 1               |
| 17    | Caulobacteraceae| Phenylobacterium  | -0.01156         | 240                | 14           | 10              |
| 4057  | Moraxellaceae| Acinetobacter      | -0.01059         | 23                 | 661          | 7               |
| 29    | Rhizobiaceae| Agrobacterium      | -0.01026         | 54                 | 13           | 8               |
| 33    | Oxalobacteraceae| Cupriavidus       | -0.01012         | 147                | 19           | 6               |
| 36    | Aurantimonadaceae| NA            | -0.01011         | 0                  | 31           | 3               |
| 71    | Rhodobacteraceae| Paracoccus        | -0.00972         | 0                  | 44           | 1               |
| 26    | Pseudomonadaceae| NA                | -0.00948         | 116                | 3            | 7               |
| 3284  | Oxalobacteraceae| NA                | -0.00899         | 83                 | 3            | 3               |
| 3684  | Sphingomonadaceae| Sphingomonas      | -0.00884         | 0                  | 25           | 3               |
| 157   | Oxalobacteraceae| NA                | -0.00794         | 1                  | 6            | 2               |
| 53    | Aurantimonadaceae| NA                | -0.00741         | 0                  | 50           | 2               |
| 72    | Xanthomonadaceae| Stenotrophomonas | -0.00680         | 117                | 15           | 5               |
| 3493  | Comamonadaceae| NA                | -0.00668         | 499                | 1            | 5               |
| 100   | Sphingomonadaceae| Kaistobacter     | -0.00636         | 0                  | 7            | 3               |
| 22    | Oxalobacteraceae| NA                | -0.00602         | 0                  | 285          | 3               |
| 40    | Caulobacteraceae| Phenyllobacterium| -0.00533         | 32                 | 22           | 3               |
| 146   | Corynebacteriaceae| Corynebacterium  | -0.00510         | 0                  | 33           | 3               |
| 472   | Pseudomonadaceae| Pseudomonas       | -0.00445         | 0                  | 0            | 0               |
| 656   | Caulobacteraceae| NA                | -0.00426         | 0                  | 4            | 2               |
| 35    | Hyphomicrobiaceae| Devosia           | -0.00402         | 342                | 11           | 7               |
| 51    | Nocardiaceae| Rhodococcus        | -0.00396         | 123                | 145          | 6               |
| 2286  | Comamonadaceae| NA                | -0.00389         | 0                  | 0            | 1               |
| 63    | Caulobacteraceae| NA                | -0.00350         | 0                  | 19           | 2               |
| 117   | Caulobacteraceae| NA                | -0.00329         | 0                  | 26           | 1               |
| 2999  | Oxalobacteraceae| NA                | -0.00325         | 2                  | 18           | 2               |
| 83    | Nocardiaceae| NA                | -0.00256         | 30                 | 52           | 5               |
| 3725  | Oxalobacteraceae| NA                | -0.00235         | 54                 | 308          | 3               |
| 2172  | Comamonadaceae| NA                | -0.00163         | 0                  | 8            | 3               |
| 86    | Rhodospirillaceae| Azospirillum      | -0.00149         | 14                 | 14           | 3               |
| 28    | Staphyloccaceae| Staphylococcus    | -0.00127         | 396                | 52           | 4               |
| 80    | Chromatiaceae| Rheinheimera      | -0.00118         | 126                | 22           | 7               |
| 235   | Bacillaceae| Bacillus           | -0.00091         | 33                 | 0            | 4               |
| 19    | Comamonadaceae| Limnohabitans     | -0.00080         | 84                 | 132          | 6               |
| 2093  | Moraxellaceae| Acinetobacter      | -0.00048         | 14                 | 305          | 3               |
| 30    | Caulobacteraceae| NA                | -0.00037         | 0                  | 11           | 2               |
| 3516  | Rhizobiaceae| Agrobacterium      | -0.00006         | 436                | 19           | 4               |
| 5     | Moraxellaceae| Acinetobacter      | 0.00014          | 0                  | 0            | 3               |
| 1886  | Comamonadaceae| NA                | 0.00062          | 368                | 14           | 10              |
| 61    | Methylbacteriaceae| NA            | 0.00063          | 0                  | 5            | 1               |
World development indicators [Fig. 6]

We retrieved World Development Indicators (WDI) for the year 2015 from the database provided by the World Bank (http://databank.worldbank.org, last accessed July 2017). The indicators were restricted to those that were available for at least 100 countries. Additionally, we excluded those WDIs with too many ties to allow for a precise estimation of the empirical checkerboard copula. This data filtering resulted in a total of $n = 450$ WDIs of $n = 179$ countries included in the analysis. For each indicator pair $X$ and $Y$, we quantified asymmetric dependence and as well as the asymmetry in dependence (Supplementary material 4). Asymmetry in all indicator pairs ranged from $|a| = 0$ to $|a| = 0.317$; 99.8% of all pairs had an asymmetry below $a = 0.2016$. The indicators birth rate and death rate that are moderately dependent on each other ($qad$: mean $q$
= 0.42, \( p < 0.001 \)), nevertheless the dependence is significantly asymmetric (\( qad: a = 0.2016, p < 0.001 \); Fig. S2_6).

**Fig. S2_6** Boxplot of asymmetry \(|a|\) values quantified for all indicator pairs possible. The absolute value of the asymmetry of the indicator pair ‘birth rate’ and ‘death rate’ is indicated by the pink arrow.

Death rate is well predictable by birth rate (\( q(birth \ rate, \ death \ rate) = 0.53, p < 0.001 \)) whereas variability in birth rate is not as strongly reduced by the knowledge of death rate (\( q(death \ rate, \ birth \ rate) = 0.32, p < 0.001 \); Fig. S2_7).
The roughly ‘U’ shaped relationship between birth and death rate deserved further data exploration – we found the gross domestic product per capita (GDP) to be an underlying factor potentially explaining the non-linear relationship. GDP is strongly and symmetrically associated to birth rate ($q_{ad}: \text{mean } q = 0.67, \ p < 0.001; \ a = 0.02, \ p = 0.59$) and weakly and symmetrically to death rate ($q_{ad}: \text{mean } q = 0.31, \ p < 0.001; \ a = 0.03, \ p = 0.39$). Roughly speaking, Fig. 4 in the main text suggests that in countries with a GDP below the mean GDP of all countries, birth rate and death rate are positively associated, in countries with a GDP above the mean GDP a negative association is suggested. This pattern in supported by $q_{ad}$: In countries with a GDP below the mean GDP of all countries birth and death rate show are symmetrically associated to each other ($q_{ad}: \text{mean } q = 0.48, \ p < 0.001; \ a = 0.06, \ p = 0.19$; Fig. S2_8), the same is true for the same association in countries with a GDP above the mean GDP of all countries ($q_{ad}: \text{mean } q = 0.53, \ p < 0.001; \ a = 0.08, \ p = 0.08$; Fig. S2_9).
Fig. S2_8 Empirical checkerboard copula for the indicator pair ‘birth rate’ (X1) and ‘death rate’ (X2) for countries with a GDP below the mean GDP of all countries. For each vertical or horizontal stripe, the mass distributions sums up to 1. Therefore, the checkerboard copula can be used to predict X2 given the knowledge of X1 and vice versa.
Fig. S2_9  Empirical checkerboard copula for the indicator pair ‘birth rate’ ($X_1$) and ‘death rate’ ($X_2$) for countries with a GDP above the mean GDP of all countries. For each vertical or horizontal stripe, the mass distribution sums up to 1. Therefore, the checkerboard copula can be used to predict $X_2$ given the knowledge of $X_1$ and vice versa.

Linear models do not find a significant association between birth and death rate ($F_{1,170} = 0.44, p = 0.51$), however, death rate is negatively associated to GDP ($F_{1,170} = 8.84, p = 0.003$) and the interaction between birth rate and GDP significantly affects death rate ($F_{1,170} = 12.02, p < 0.001$).

References

1. D. N. Reshef et al., Detecting novel associations in large data sets. Science **334**, 1518 (Dec 16, 2011).
2. J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, B. Scholkopf, Distinguishing cause from effect using observational data: methods and benchmarks. J Mach Learn Res **17**,  (2016).
3. D. Dua, E. Karra Taniskidou. (Irvine, CA: University of California, School of Information and Computer Science, 2017).
4. D. N. Karger et al., Data Descriptor: Climatologies at high resolution for the earth’s land surface areas. Sci Data **4**,  (Sep 5, 2017).
5. D. N. Karger et al., Data from: Climatologies at high resolution for the earth’s land surface areas. *Dryad Digital Repository* [https://doi.org/10.5061/dryad.kd1d4](https://doi.org/10.5061/dryad.kd1d4), (2017).

6. R. R. Junker, A. Keller, Microhabitat heterogeneity across leaves and flower organs promotes bacterial diversity. *FEMS Microbiol Ecol* 91: fiv097, (2015).

7. G. Csardi, T. Nepusz, The igraph software package for complex network research. *InterJournal Complex Systems*, 1695 (2006).
| Country | Birth | Death | GDP  |
|--------|-------|-------|------|
| AFG    | 33.314| 8.034 | 1861.12433 |
|AGO     | 44.998| 13.424| 6634.40227  |
| ALB    | 13.632| 7.463 | 11504.7482  |
| ARE    | 10.592| 1.641 | 70245.9325  |
| ARG    | 17.359| 7.55  | 20337.7161  |
| ARM    | 13.008| 9.316 | 8709.541099 |
| ATG    | 16.202| 6.087 | 21415.708   |
| AUS    | 12.7  | 6.6   | 46475.9278  |
| AUT    | 9.8   | 9.6   | 49419.3301  |
| AZE    | 17.2  | 5.7   | 17779.7737  |
| BDI    | 43.497| 11.066| 796.994415  |
| BEL    | 10.9  | 9.8   | 45608.4293  |
| BEN    | 35.592| 9.244 | 2115.7956   |
| BFA    | 13.632| 7.463 | 11504.7482  |
| BGD    | 19.455| 5.35  | 3335.33726  |
| BGR    | 9.2   | 15.3  | 18248.8319  |
| BHR    | 14.205| 2.367 | 47333.848   |
| BHS    | 15.044| 6.234 | 23072.8147  |
| BIH    | 8.818 | 10.69 | 11687.0708  |
| BHR    | 12.5  | 12.6  | 18344.8161  |
| BLZ    | 22.758| 5.568 | 8583.1383   |
| BOL    | 23.591| 7.359 | 6954.30177  |
| BRA    | 14.513| 6.163 | 15615.3431  |
| BRN    | 15.691| 3.024 | 79429      |
| BTN    | 17.404| 6.218 | 8236.35647  |
| BWA    | 24.57 | 7.536 | 16350.4743  |
| CAF    | 33.435| 14.088| 666.958172  |
| CAN    | 10.9  | 7.5   | 44204.9466  |
| CHE    | 10.2  | 8.1   | 62499.6372  |
| CHL    | 13.048| 5.137 | 23579.4623  |
| CHN    | 12.1  | 7.1   | 14448.2656  |
| CIV    | 36.895| 13.323| 3461.60381  |
| CMR    | 36.233| 11.197| 3184.81054  |
| COD    | 41.527| 10.966| 799.076704  |
| COG    | 36.028| 8.308 | 5901.6826   |
| COL    | 15.483| 5.93  | 13825.9213  |
| COM    | 33.433| 7.38  | 1504.52598  |
| CPV    | 21.108| 5.344 | 6301.69607  |
| CRI    | 14.524| 4.889 | 15879.5968  |
| CYP    | 11.206| 6.897 | 31539.5268  |
| CZE    | 10.5  | 10.5  | 33743.1909  |
| DEU    | 9.2   | 11.3  | 47998.8627  |
| DJI    | 24.721| 8.561 | 3342.47733  |
| DMA    | NA    | NA    | 10792.9615  |
| DNK    | 10.2  | 9.2   | 48980.7917  |
| DOM    | 20.506| 6.087 | 14237.0579  |
| DZA    | 23.692| 5.114 | 14612.6804  |
| ECU    | 20.487| 5.119 | 11474.1372  |
| EGY    | 27.298| 6.066 | 10749.0946  |
| ESP    | 9.2   | 9.2   | 34696.3362  |
| EST    | 10.6  | 11.6  | 28946.7899  |
| ETH    | 31.906| 7.198 | 1632.34501  |
| FIN    | 10.1  | 9.6   | 42275.2399  |
| FJI    | 19.688| 6.982 | 9323.15657  |
| FRA    | 12    | 9     | 41178.1418  |
| GAB    | 29.718| 8.391 | 17926.4291  |
| GBR    | 11.9  | 9.3   | 41767.2897  |
| GEO    | 13.359| 11.533| 9609.32603  |
| GHA    | 32.262| 8.821 | 4184.04756  |
| GIN    | 36.409| 9.678 | 1260.68464  |
| Country | GMB | 41.698 | 8.58  | 1691.00974 |
|---------|-----|--------|-------|------------|
| GNB     |     | 36.647 | 11.9  | 1516.48439 |
| GNQ     |     | 34.626 | 10.594| 29000.9177 |
| GRC     |     | 8.5    | 11.2  | 26357.938  |
| GRD     |     | 19.0   | 7.17  | 13558.7991 |
| GTM     |     | 26.752 | 5.339 | 7764.77359 |
| GUY     |     | 19.288 | 8.242 | 7520.03423 |
| HKG     |     | 8.2    | 6.3   | 56951.9637 |
| HND     |     | 20.881 | 5.015 | 4590.23694 |
| HRV     |     | 8.9    | 12.9  | 22488.7462 |
| HTI     |     | 24.578 | 8.629 | 1758.11173 |
| HUN     |     | 9.4    | 13.4  | 26436.2078 |
| IDN     |     | 19.579 | 7.167 | 11038.7916 |
| IND     |     | 19.658 | 7.306 | 6126.52266 |
| IRL     |     | 14.2   | 6.4   | 67974.1633 |
| IRN     |     | 17.084 | 4.607 | 10464.4389 |
| IRQ     |     | 34.233 | 5.112 | 15895.2266 |
| ISL     |     | 12.5   | 6.6   | 47690.3816 |
| ISR     |     | 21.3   | 5.3   | 36545.699  |
| ITA     |     | 8.6    | 10.7  | 37255.17   |
| JAM     |     | 17.23  | 6.914 | 8630.07386 |
| JOR     |     | 26.512 | 3.839 | 9040.64198 |
| JPN     |     | 7.9    | 10.2  | 40686.0244 |
| KAZ     |     | 22.66  | 7.48  | 25044.878  |
| KEN     |     | 34.084 | 8.025 | 3019.19363 |
| KGZ     |     | 27.4   | 5.8   | 3447.17155 |
| KHM     |     | 23.776 | 6.038 | 3503.97435 |
| KOR     |     | 8.6    | 5.4   | 34421.5796 |
| KWT     |     | 19.559 | 2.523 | 73817.0753 |
| LAO     |     | 26.27  | 6.629 | 5785.93143 |
| LBN     |     | 15.397 | 4.547 | 13934.5054 |
| LBR     |     | 34.613 | 8.52  | 836.07557  |
| LCA     |     | 14.933 | 7.257 | 11368.5283 |
| LKA     |     | 15.601 | 6.844 | 11777.8718 |
| LSO     |     | 28.461 | 14.693| 2956.22201 |
| LUK     |     | 10.8   | 14.1  | 28936.2736 |
| LUX     |     | 10.7   | 10.7  | 104206.109 |
| LVA     |     | 11.1   | 14.4  | 24919.4788 |
| MAC     |     | 11.681 | 4.816 | 107024.284 |
| MAR     |     | 20.38  | 5.682 | 7757.25318 |
| MDA     |     | 10.566 | 11.363| 5054.0422  |
| MDG     |     | 34.223 | 6.52  | 1465.42355 |
| MDV     |     | 20.804 | 3.761 | 12770.2713 |
| MEX     |     | 18.47  | 4.814 | 17244.3658 |
| MGD     |     | 11.257 | 9.492 | 14022.7729 |
| MLD     |     | 42.908 | 10.118| 2043.45923 |
| MLT     |     | 10     | 8     | 35880.6148 |
| MMR     |     | 17.49  | 8.263 | 5399.36051 |
| MNE     |     | 11.441 | 10.035| 16183.0827 |
| MNG     |     | 23.429 | 6.087 | 12147.9422 |
| MOZ     |     | 38.81  | 11.235| 1190.60139 |
| MRT     |     | 33.002 | 7.861 | 3834.74526 |
| MUS     |     | 10.1   | 7.7   | 20085.1709 |
| MWI     |     | 38.542 | 7.471 | 1159.17299 |
| MYS     |     | 16.789 | 4.983 | 26606.3433 |
| NAM     |     | 29.414 | 7.033 | 10554.1592 |
| NER     |     | 49.211 | 8.819 | 955.58681  |
| NGA     |     | 39.125 | 12.704| 6037.69743 |
| NIC     |     | 19.926 | 4.753 | 5282.03008 |
| NLD     |     | 10     | 8.7   | 49546.9564 |
| NOR     |     | 11.4   | 7.8   | 62053.2129 |
| Country | NPL | NZL | OMN | PAK | PAN | PER | PHL | PNG | POL | PRT | PRY | PSE | QAT | ROU | RUS | RW | SAU | SDN | SEN | SGP | SLB | SLE | SLV | SOM | SRY | SRB | STP | SUR | SVK | SVN | SWE | SWZ | SYC | SYR | TCD | TGO | THA | TJK | TLS | TON | TTO | TUN | TUR | TZA | UGA | UKR | URY | USA | UZB | VCT | VEN | VNM | VUT | WSM | XXX | YEM | ZAF | ZMB | ZWE |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 20.216 | 6.339 | 2449.8233 |
| 13.27  | 6.87  | 37948.9196 |
| 19.322 | 2.666 | 42737.1283 |
| 28.845 | 7.339 | 4998.75825 |
| 19.124 | 5.044 | 22012.538  |
| 19.6   | 5.619 | 12529.2282 |
| 23.316 | 6.773 | 7319.57126 |
| 28.151 | 7.649 | NA       |
| 9.7    | 10.4  | 26855.7747 |
| 8.3    | 10.5  | 29687.7904 |
| 21.167 | 5.693 | 9198.50408 |
| 32.363 | 3.582 | 2865.80511 |
| 11.806 | 1.486 | 127500.756 |
| 9.3    | 13.2  | 22070.5305 |
| 13.3   | 13    | 23702.8445 |
| 31.163 | 6.833 | 1826.95569 |
| 19.688 | 3.418 | 54007.0362 |
| 32.645 | 7.662 | 4568.08615 |
| 37.522 | 5.99  | 2445.40838 |
| 9.7    | 4.8   | 86128.1725 |
| 29.298 | 5.7   | 2186.35699 |
| 35.445 | 13.268| 1401.24809 |
| 17.175 | 6.833 | 8352.97116 |
| 43.373 | 11.807| NA       |
| 9.3    | 14.6  | 14111.9386 |
| 33.532 | 6.814 | 3132.83955 |
| 17.946 | 7.461 | 15722.6476 |
| 10.3   | 9.9   | 29907.0594 |
| 10     | 9.6   | 31964.654 |
| 11.7   | 9.3   | 47823.298 |
| 29.372 | 14.485| 8575.05739 |
| 17     | 7.5   | 27177.1741 |
| 22.509 | 5.715 | NA       |
| 44.792 | 13.799| 2180.17983 |
| 35.068 | 8.699 | 1438.24836 |
| 10.532 | 8.028 | 16222.9718 |
| 30.225 | 5.568 | 2811.51371 |
| 37.097 | 6.809 | 2290.34727 |
| 24.331 | 5.985 | 5524.73679 |
| 13.979 | 9.507 | 33308.4225 |
| 17.94  | 6.593 | 11445.6928 |
| 16.489 | 5.735 | 24054.1778 |
| 38.535 | 6.682 | 2652.19566 |
| 42.528 | 9.324 | 1802.08621 |
| 10.7   | 14.9  | 7948.1421 |
| 14.148 | 9.344 | 21115.1266 |
| 12.4   | 8.2   | 56207.0368 |
| 23.5   | 4.9   | 6069.20979 |
| 15.776 | 7.127 | 11140.3499 |
| 19.268 | 5.567 | NA       |
| 16.941 | 5.847 | 6034.25995 |
| 26.167 | 4.718 | 2988.47194 |
| 25.097 | 5.334 | 5918.61003 |
| 17.1   | 7.9465 | 99066 |
| 31.889 | 6.829 | 2811.99556 |
| 20.415 | 12.409| 13229.6246 |
| 39.672 | 8.671 | 3861.98923 |
| 34.521 | 9.086 | 2013.16712 |
library(qad)

# Function to generate the quadratic function
quadratic <- function(n,a){
  x <- seq(-1,1, length.out = n)
  y <- x^2 + runif(n, -a, a)
  A <- data.frame(x=x, y=y)
  A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)) ,y=(A$y-min(A$y))/(max(A$y)-min(A$y)))
  return(A)
}

rt <- quadratic(1000, 0.01) # quadratic(sample size, noise)
plot(rt$x, rt$y, col = rgb(255/255, 120/255, 180/255, 0.55), pch = 16, cex = 2)
plot(rank(rt$x)/length(rt$x), rank(rt$y)/length(rt$x), col = rgb(255/255, 120/255, 180/255, 0.55), pch = 16, cex = 2)
qr <- qad(rt$x,rt$y, print=TRUE, permutation = FALSE)
plot(qr, copula = TRUE, density = FALSE)

# Function to generate the linear function
linear <- function(n,a){
  x <- seq(0,1, length.out = n)
  y <- x + runif(n, -a, a)
  A <- data.frame(x=x, y=y)
  A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)) ,y=(A$y-min(A$y))/(max(A$y)-min(A$y)))
  return(A)
}

xs <- function(n,a){
  x1 <- seq(0,1, length.out = n/2)
  y1 <- x1 + runif(n, -a, a)
  x2 <- seq(0,1, length.out = n/2)
  y2 <- 1-x2 + runif(n, -a, a)
  x <- c(x1, x2)
  y <- c(y1, y2)
  A <- data.frame(x=x, y=y)
  A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)) ,y=(A$y-min(A$y))/(max(A$y)-min(A$y)))
  return(A)
# two parallel lines

tpl <- function(n,a){
x <- seq(0,1, length.out = n)
y <- c(seq(0,1,length.out = n/2), seq(0,1,length.out = n/2))
x <- x + runif(length(x), -a, a)
y <- y + runif(length(y), -a, a)

A <- data.frame(x=x, y=y)
A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)), y=(A$y-min(A$y))/(max(A$y)-min(A$y)))
return(A)
}

# two lines:
two.lines <- function(n=2000, a=0.05){
x <- runif(n,0,1)
#y <- rnorm(n,0,a)
y <- runif(n,-a,a)
A <- data.frame(x=x, y=y)
A$case <- sample(c(0,1), n, replace=TRUE)
A$u <- 0
A$v <- 0

phi <- c(pi/20,pi/4)
Rot0 <- matrix(c(cos(phi[1]),sin(phi[1]),-sin(phi[1]),cos(phi[1])), nrow=2)
Rot1 <- matrix(c(cos(phi[2]),sin(phi[2]),-sin(phi[2]),cos(phi[2])), nrow=2)

for(i in 1:n){
z <- Rot0 %*% c(A$x[i], A$y[i])
if(A$case[i]==1){z <- Rot1 %*% c(A$x[i], A$y[i])}
A[i,4:5] <- as.vector(z)
}
A <- subset(A, select = c(u,v))
names(A) <- c("x", "y")
A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)), y=(A$y-min(A$y))/(max(A$y)-min(A$y)))
return(A)
}

# Non-coexistence
non.coexistence <- function(n,a){
A <- data.frame(x=runif(n), y=runif(n))
A <- subset(A, A$x<=a | A$y<=a)
A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)), y=(A$y-min(A$y))/(max(A$y)-min(A$y)))
return(A)
}

# quadratic
quadratic <- function(n,a){
x <- seq(-1,1, length.out = n)
y <- x^2 + runif(n, -a, a)
A <- data.frame(x=x, y=y)
A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)), y=(A$y-min(A$y))/(max(A$y)-min(A$y)))
return(A)
}

# sinus
sinus <- function(n,a){
x <- seq(-8,8, length.out = n)
y <- sin(x) + runif(n, -a, a)
A <- data.frame(x=x, y=y)
A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)), y=(A$y-min(A$y))/(max(A$y)-min(A$y)))
return(A)
}
# Torus
torus <- function(n,a){
  r <- sqrt(runif(n,1-a,1+a))
  phi <- runif(n,0,2*pi)
  A <- data.frame(x=r*cos(phi),y=r*sin(phi))
  A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)) ,y=(A$y-min(A$y))/(max(A$y)-
  min(A$y)))
  return(A)
}

cbs <- function(n,a){
  #x1 <- seq(0,1, length.out = 16)
  #yp <- seq(0,1, length.out = 4)
  #y1 <- c(sample(yp, 4, replace = FALSE), sample(yp, 4, replace = FALSE), sample(yp, 4,
  replace = FALSE), sample(yp, 4, replace = FALSE))
  x1 <- c(0.00000000, 0.06666667, 0.13333333, 0.20000000, 0.26666667, 0.33333333,
  0.40000000, 0.46666667, 0.53333333, 0.60000000, 0.66666667, 0.73333333, 0.80000000,
  0.86666667, 0.93333333, 1.00000000)
  y1 <- c(0.6666667, 0.0000000, 1.0000000, 0.3333333, 0.6666667, 0.3333333, 0.0000000,
  1.0000000, 0.6666667, 0.3333333, 0.0000000, 1.0000000, 0.6666667, 0.3333333, 1.0000000,
  0.0000000)
  x2 <- rep(x1, round(n/length(x1), 0))
  y2 <- rep(y1, round(n/length(y1), 0))
  x <- x2 + runif(length(x2), -a, a)
  y <- y2 + runif(length(y2), -a, a)

  A <- data.frame(x=x,y=y)
  A <- data.frame(x=(A$x-min(A$x))/(max(A$x)-min(A$x)) ,y=(A$y-min(A$y))/(max(A$y)-
  min(A$y)))
  return(A)
}

### loop [example: Torus]
n = 1000
noise <- seq(0,1, length.out = 100)
simnum <- 10
sto_Pr <- rep(NA, (simnum * length(noise)))
sto_Sr <- rep(NA, (simnum * length(noise)))
sto_MIC <- rep(NA, (simnum * length(noise)))
sto_x1x2 <- rep(NA, (simnum * length(noise)))
sto_x2x1 <- rep(NA, (simnum * length(noise)))
sto_asy <- rep(NA, (simnum * length(noise)))
noi <- rep(NA, (simnum * length(noise)))
pos <- 0
for (i in 1:length(noise)){
  ai <- noise[i]
  for (j in 1:simnum){
    pos <- pos + 1
    rt <- torus(n, ai)
    sto_Pr[pos] <- cor(rt$x,rt$y)
    sto_Sr[pos] <- cor(rt$x,rt$y, method = "spearman")
    mr <- mine(rt$x,rt$y)
    sto_MIC[pos] <- mr$MIC
    qr <- qad(rt$x,rt$y, print=FALSE)
    sto_x1x2[pos] <- qr$results$coef[1]
    sto_x2x1[pos] <- qr$results$coef[2]
    sto_asy[pos] <- qr$results$coef[4]
    noi[pos] <- ai
  }
  print(pos)
}
source("http://janhove.github.io/RCode/plot_r.R")

rV <- seq(-0.999, 0.999, length.out = 200)
x1x2 <- matrix(nrow = length(rV), ncol = 16)
x2x1 <- matrix(nrow = length(rV), ncol = 16)
asy <- matrix(nrow = length(rV), ncol = 16)
MIC <- matrix(nrow = length(rV), ncol = 16)
rho <- matrix(nrow = length(rV), ncol = 16)
rf <- matrix(nrow = length(rV), ncol = 16)
for(i in 1:length(rV)){
  rd <- plot_r(r = rV[i], n = 500, showdata = TRUE)
  for(j in 1:16){
    data <- as.data.frame(rd$data[j])
    qr <- qad(data[,1], data[,2])
    x1x2[i, j] <- qr$results$coef[1]
    x2x1[i, j] <- qr$results$coef[2]
    asy[i, j] <- qr$results$coef[4]
    mr <- mine(data[,1], data[,2])
    MIC[i, j] <- mr$MIC
    rf[i, j] <- rV[i]
    sr <- cor.test(data[,1], data[,2], method = "spearman")
    rho[i, j] <- as.numeric(sr$estimate)
  }
}

setwd("folder containing data")

m <- read.table("Results_BioClim_Raster_Coords.txt", header = T) ## data are provided in SIX
m <- m[3:ncol(m)]

### quantification of asymmetric dependence (q) for all variable pairs

pwq <- pairwise.qad(m, permutation = FALSE)
heatmap.qad(pwq, select = "dependence", fontsize = 2) + theme(axis.text.x = element_text(angle=90, hjust=0, vjust=1))
heatmap.qad(pwq, select = "asymmetry", fontsize = 2) + theme(axis.text.x = element_text(angle=90, hjust=0, vjust=1))

### quantification of symmetric dependence (Pearson’s r^2 and r) for all variable pairs

r2 <- as.matrix(cor(m, method="pearson")^2)
diag(r2) <- NA
r <- as.matrix(cor(m, method="pearson"))
diag(r) <- NA

### Association between annual mean temperature [°C] and annual Precipitation [mm]

plot((m$AMT_R/10), m$AP_R, col=rgb(16/255, 78/255, 139/255, 0.15), cex = 1.5, pch = 16, log="y", xlab = "Annual mean temperature [°C]", ylab = "Annual Precipitation [mm]")
plot(rank((m$AMT_R/10)), rank(m$AP_R), col=rgb(16/255, 78/255, 139/255, 0.15), cex = 1.5, pch = 16, xlab = "Annual mean temperature [°C]", ylab = "Annual Precipitation [mm]"
### Asymmetric dependence between annual mean temperature [°C] and annual Precipitation [mm]

empirical checkerboard copula is shown, which can be used as prediction tool

```r
r <- qad(m$AMT_R/10, log(m$AP_R+1), print=TRUE, permutation = TRUE, nperm = 100)
plot(r, copula = FALSE, density = FALSE)
```

#Microbiome [Fig. 5] #

```r
list.files()
m<-read.table("microbiome.txt", header=T)
m <- m[,2:ncol(m)]

### remove columns with too many non-unique values

```r
unique.prop <- rep(NA, ncol(m))
prop.single.value <- rep(NA, ncol(m))
```

for(n in 1:ncol(m)){
  ```r
  unique.prop[n] <- length(unique(m[,n])) / nrow(m)
  uV <- unique(m[,n])
  ```

  ```r
  numV <- rep(NA, length(uV))
  for(i in 1:length(uV)){
    numV[i] <- length(m[,n] [m[,n] == uV[i]])
  }
  ```

  ```r
  prop.single.value[n] <- max(numV, na.rm = T)/nrow(m)
  ```

m <- m[,prop.single.value < 0.25]

### quantification of asymmetric dependence (q) for all variable pairs

```r
pwq <- pairwise.qad(m)
```

### quantification of symmetric dependence (Pearson’s r^2 and r) for all variable pairs

```r
r2 <- as.matrix(cor(m, method="pearson")^2)
diag(r2) <- NA
```

```r
r <- as.matrix(cor(m, method="pearson"))
diag(r) <- NA
```

### identification of key-species

```r
AdepM <- matrix(NA, ncol(pwq$q), nrow(pwq$q))
colnames(AdepM) <- colnames(pwq$q)
AdepMean <- rep(NA, ncol(pwq$q))
AdepMedian <- rep(NA, ncol(pwq$q))
AdepQ2.5 <- rep(NA, ncol(pwq$q))
AdepMIN <- rep(NA, ncol(pwq$q))
AdepMAX <- rep(NA, ncol(pwq$q))
AdepSD <- rep(NA, ncol(pwq$q))
DepMean <- rep(NA, ncol(pwq$q))
DetMean <- rep(NA, ncol(pwq$q))
ColName <- rep(NA, ncol(pwq$q))
```

for(i in 1:ncol(pwq$q)){
  ```r
  adt <- as.numeric(pwq$q[i,]) - pwq$q[,i]
  AdepM[,i] <- adt
  AdepMean[i] <- mean(adt, na.rm = TRUE)
  AdepMedian[i] <- median(adt, na.rm = TRUE)
  AdepQ2.5[i] <- as.numeric(t.test(adt, mu = 0)$conf.int[1])
  ```

```r
```
### Summary

\[
\begin{align*}
\text{AdepQ97.5}[i] & \leftarrow \text{as.numeric}(\text{t.test}(\text{adt}, \text{mu} = 0)\text{\$conf.int}[2]) \\
\text{AdepMIN}[i] & \leftarrow \text{min}(\text{adt}, \text{na.rm} = \text{TRUE}) \\
\text{AdepMAX}[i] & \leftarrow \text{max}(\text{adt}, \text{na.rm} = \text{TRUE}) \\
\text{AdepSD}[i] & \leftarrow \text{sd}(\text{adt}, \text{na.rm} = \text{TRUE}) \\
\text{DepMean}[i] & \leftarrow \text{mean}(\text{as.numeric}(\text{pwq$q[,i]}) \text{, na.rm=TRUE}) \\
\text{DetMean}[i] & \leftarrow \text{mean}(\text{as.numeric}(\text{pwq$q[i,,])}, \text{na.rm=TRUE}) \\
\text{ColName}[i] & \leftarrow \text{colnames}(\text{pwq$q})[i]
\end{align*}
\]

\[
\text{Summary} \leftarrow \text{as.data.frame}(\text{cbind}(\text{AdepMean}, \text{AdepMedian}, \text{AdepQ2.5}, \text{AdepQ97.5}, \text{AdepMIN}, \text{AdepMAX}, \text{AdepSD}, \text{DepMean}, \text{DetMean}))
\]

\[
\text{row.names}(\text{Summary}) \leftarrow \text{ColName}
\]

### Fig. 5a

\[
\text{SummaryO} \leftarrow \text{Summary[ order( } \text{Summary[,2]}) , ] } \quad \# 1 = \text{mean}; 2 = \text{median}
\]

\[
\text{AdepMO} \leftarrow \text{AdepM[, row.names(SummaryO)]}
\]

\[
iI \leftarrow \text{cut( SummaryO$AdepMedian, breaks = seq( min(SummaryO$AdepMedian), max(SummaryO$AdepMedian), len = 500), include.lowest = TRUE)}
\]

\[
\text{colors}_1 \leftarrow \text{colorRampPalette(c(rgb(80/255, 235/255, 150/255, 1),rgb(64/255,161/255, 255/255, 1), col=rgb(255/255, 120/255, 180/255, 1))}(499)[ii]
\]

\[
\text{boxplot}(\text{AdepMO, las = 2, cex.axis = 0.7, ylim=c(min(\text{AdepMO, na.rm=TRUE}), 0.5), col = colors}_1, \text{range = 0, xaxt = "n"})
\]

\[
\text{xnames} \leftarrow \text{sub("OTU_", "", row.names(SummaryO))}
\]

\[
\text{axis}(1,at=1:ncol(\text{AdepMO}),labels=xnames, las = 2, cex.axis = 0.65)
\]

\[
\text{points(seq(-1,ncol(\text{AdepMO})+2), rep(0, ncol(\text{AdepMO})+4), type="l", lty =1, col = "black", lwd = 1.5)
\]

\[
\text{points(seq(1,ncol(\text{AdepMO}),1), SummaryO$DepMean, type="l", lty =1, col = rgb(80/255, 235/255, 150/255, 1), lwd = 3)
\]

\[
\text{points(seq(1,ncol(\text{AdepMO}),1), SummaryO$DetMean, type="l", lty =2, col=rgb(255/255, 120/255, 180/255, 1), lwd = 3)
\]

---

# World development indicators ; Fig. 6

--

---

\#

\#

\#

# setwd("folder containing data") #

\m <- \text{read.table("WDI\_data.txt", header = T)} \# data are provided in SIX

\text{attach(m)}

\text{head(m)}

### quantification of asymmetric dependence (q)

\[
r \leftarrow \text{qad}(\text{birth, death, permutation = TRUE, nperm = 100, DoParallel = TRUE, print=TRUE})
\]

\[
\text{plot}(r, \text{copula = FALSE, margins = TRUE})
\]

\[
r \leftarrow \text{qad}(\text{birth, log(gdp+1)}, \text{permutation = TRUE, nperm = 100, DoParallel = TRUE, print=TRUE})
\]

\[
\text{plot}(r, \text{copula = FALSE, margins = TRUE})
\]

\[
r \leftarrow \text{qad}(\text{death, log(gdp+1)}, \text{permutation = TRUE, nperm = 100, DoParallel = TRUE, print=TRUE})
\]

\[
\text{plot}(r, \text{copula = FALSE, margins = TRUE})
\]

\[
\text{mean}_gdp \leftarrow \text{mean}(\text{log(gdp+1)}, \text{na.rm = TRUE})
\]

\[
r \leftarrow \text{qad}(\text{birth[log(gdp+1) < mean}_gdp, death[log(gdp+1) < mean}_gdp], \text{permutation = TRUE, nperm = 100, DoParallel = TRUE, print=TRUE})
\]

\[
\text{plot}(r, \text{copula = FALSE, margins = TRUE})
\]

\[
r \leftarrow \text{qad}(\text{birth[log(gdp+1) > mean}_gdp, death[log(gdp+1) > mean}_gdp], \text{permutation = TRUE, nperm = 100, DoParallel = TRUE, print=TRUE})
\]

\[
\text{plot}(r, \text{copula = FALSE, margins = TRUE})
\]

### Fig. 4
ii <- cut(log(gdp+1), include.lowest = TRUE, breaks = 300)

colT <- colorRampPalette(c(rgb(80/255, 235/255, 150/255, 1),rgb(64/255,161/255, 255/255, 1),
col=rgb(255/255, 120/255, 180/255, 1), rgb(247/255,110/255, 94/255, 1)))(300)[ii]

plot(birth, death, xlab = "Birth_rate_crude_per_1000_people",ylab =
"Death_rate_crude_per_1000_people", pch = 16,cex = 2, col= colT)
points(birth, death, xlab = "Birth_rate_crude_per_1000_people",ylab =
"Death_rate_crude_per_1000_people", pch = 1,cex = 2, col= "black")
text(birth, death+0.3, labels=country, cex= 0.7)

###linear model

ml <- aov(death ~ birth*log(gdp+1))
summary(ml)