APPLICATION

qad: An R-package to detect asymmetric and directed dependence in bivariate samples

Florian Griessenberger1 | Wolfgang Trutschnig1 | Robert R. Junker2,3

1Department of Mathematics, University of Salzburg, Salzburg, Austria
2Department of Environment and Biodiversity, University of Salzburg, Salzburg, Austria
3Evolutionary Ecology of Plants, Department of Biology, Philipps-University Marburg, Marburg, Germany

Correspondence
Robert R. Junker
Email: robert.junker@uni-marburg.de

Funding information
Austrian Science Fund, Grant/Award Number: Y 1102 B29; WISS 2025 - IDA-lab Salzburg, Grant/Award Number: 20204-WISS/225/197-2019 and 20102-F1901166-KZP

Handling Editor: Timothée Poisot

Abstract
1. Correlations belong to the standard repertoire of ecologists for quantifying the strength of dependence between two random variables. Classical dependence measures are usually not capable of detecting non-monotonic or non-functional dependencies. Furthermore, they completely fail to detect asymmetry and direction in dependence, which exist in many situations and should not be ignored.

2. In this paper, we present qad (short for quantification of asymmetric dependence), a nonparametric statistical method to quantify directed and asymmetric dependence of bivariate samples. Qad is applicable in general (e.g. linear, non-linear, or non-monotonic) situations, is sensitive to noise in data, exhibits a good small sample performance, detects asymmetry in dependence, shows high power in testing for independence, requires no assumptions regarding the underlying distribution of the data and reliably quantifies the information gain/predictability of quantity Y given knowledge of quantity X, and vice versa (i.e. q(X,Y) ≠ q(Y,X)).

3. Here, we briefly recall the methodology underlying qad, introduce the functions of the R-package qad, which returns estimates for the measures q(X,Y) denoting the directed dependence of Y on X (or, equivalently, the influence of X on Y), q(Y,X) the directed dependence of X on Y, a(X,Y) := q(X,Y) − q(Y,X) the asymmetry in dependence. Furthermore, qad can be used to predict Y given knowledge of X, and vice versa. Additionally, we compare empirical performance of qad with that of seven other well established measures and demonstrate the applicability of qad on ecological datasets.

4. We illustrate that direction and asymmetry in dependence are universal properties of bivariate associations. Qad thus provides additional information gain and avoids model bias and will therefore advance and facilitate the understanding of ecological systems.

KEYWORDS
association, asymmetry, copula, correlation, directed dependence, qad

1 | INTRODUCTION

Although the number of available statistical tools is continuously increasing, classical measures such as correlations often remain the first choice for quantifying the dependence between two random variables (Anderson et al., 2021; Bolt et al., 2021). Usually, the decision for a specific correlation method is based on the models’ underlying assumptions on the data, for example, Pearson’s r should...
be used for continuous data, whereas Spearman’s $\rho$ is advised for data on the ordinal scale. Both just mentioned dependence measures, however, provide information on different aspects of bivariate distributions: Pearson’s $r$ quantifies how linear a relationship is, whereas Spearman’s $\rho$ measures the extent of monotonicity. Additional insight may be gained by considering other less frequently applied or less well-known dependence measures. Examples are distance correlation ($dCor$; Székely et al., 2007), which is implemented in the R-package energy or the information-theoretic-based maximal information coefficient (MIC; Reshef et al., 2011; R-package minerva). Very recent developments are the asymmetric dependence measures $xicor$ (Chatterjee, 2021) and quantification of asymmetric dependence ($qad$; Junker et al., 2021).

In recent years, the usefulness of symmetric dependence measures for inferring the structure of complex systems or causality in bi-variate associations has been debated and potential biases have been discussed (see, for instance, Zhang et al. (2015), Wang and Huang (2014), Okimoto (2008), Hirano and Takemoto (2019)). Thus, the concept of asymmetry/direction in dependence, which exists in most situations, should not be ignored in data analysis. Whereas in a linear setting, the dependence between two variables $X$ and $Y$ is indeed symmetric (Figure 1a) in the sense that $Y$ can be equally well predicted by knowing $X$ as vice versa, the situation, however, is different in more complex relationships. For instance, for a two-dimensional sample in the form of a parabola (Figure 1b) or a sinusoidal curve (Figure 1c), the dependence structure is clearly asymmetric. In these cases, knowing the value of the variable $X$ strongly improves the predictability of $Y$, whereas in the other direction, the information gain is significantly smaller. As an example, consider the year of deglaciation along a glacier forefield and plant diversity (Junker et al., 2020). Naturally, the year of deglaciation has a strong influence on plant diversity (not vice versa), and this directed dependence structure is clearly captured by $qad$ (Figure 1d). Especially, in cases where no a priori knowledge about the causal relationship is available, directional dependence is a useful measure for exploring and estimating the association between two random variables in a more detailed and more realistic way than classical (symmetric) dependence measures. On top, $qad$ will provide more detailed insights into the structure of communities and functional linkages between organisms or individuals and may thus assist network inference. The limitations of standard methods (e.g. Spearman’s correlation coefficient) in network inference have been recently pointed out (Coenen & Weitz, 2018) and directed and asymmetric approaches have been demanded (Ambland & Michel, 2011; Carr et al., 2019; Karmon & Piplel, 2016).

Here, we present the method $qad$, a nonparametric and directed, hence asymmetric, measure of dependence, which is publicly available in the free software environment R (Griessenberger et al., 2021; Junker et al., 2021). $qad$ returns estimates for the measures $q(X, Y)$ denoting the directed dependence of $Y$ on $X$ (or, equivalently, the influence of $X$ on $Y$), $q(Y, X)$ the directed dependence of $X$ on $Y$ and $a(X, Y) = q(X, Y) − q(Y, X)$ the asymmetry in dependence. The measure $a(X, Y)$ for asymmetry in dependence can be interpreted as the difference of the predictability of $Y$ given knowledge on $X$ and the predictability of $X$ given knowledge on $Y$. In this paper, we first describe the methodology of $qad$ and demonstrate the application of the R-package $qad$. Furthermore, we compare the empirical performance of $qad$ with existing publicly available dependence measures and highlight the information gain by considering asymmetry and direction in dependence. A complementary R-shiny app is available as Supporting Information (https://r-qad.shinyapps.io/quantification_of_dependence/) facilitating the interpretation and comparison of the results and performance returned by $qad$ and other dependence measures. An application of $qad$ to real world data concludes the paper. We hope that this introduction to $qad$ and the executed comparative analyses as well as the resources provided will be helpful for ecologists and researchers from other disciplines.

![Figure 1](https://example.com/figure1.png)

**Figure 1.** (a–c) Samples of size $n = 50$ drawn from (a) symmetric/undirected as well as (b, c) asymmetric/directed dependence structures. (d) Depicts real-world data representing plant diversity as a function of the estimated year of deglaciation at $n = 140$ studied plots. In the symmetric setting (a), the knowledge of $X$ provides roughly as much information on $Y$ as vice versa, whereas in the asymmetric and the real-world data setting (c–d) knowing the value of $X$ allows to predict the value of $Y$ much better than vice versa. Asymmetry in dependence is detected by the dependence measure $qad(q(X, Y)$ and $q(Y, X))$, whereby Pearson’s $r$ and Spearman’s $\rho$ are not capable of taking into account asymmetry in dependence: (a) $r(X, Y) = 0.989$, $\rho(X, Y) = 0.985$, $q(X, Y) = q(Y, X) = 0.861$, (b) $r(X, Y) = 0.051$, $\rho(X, Y) = 0.021$, $q(X, Y) = 0.797$ and $q(Y, X) = 0.460$, (c) $r(X, Y) = −0.092$, $\rho(X, Y) = −0.167$, $q(X, Y) = 0.667$ and $q(Y, X) = 0.311$ and (d) $r(X, Y) = −0.206$, $\rho(X, Y) = −0.274$ and $q(X, Y) = 0.478$ whereas $q(Y, X) = 0.320$. 
2 | BRIEF METHODOLOGICAL DESCRIPTION OF THE COPULA-BASED DEPENDENCE MEASURE qad

Commonly used approaches to quantify the strength of associations between two variables such as correlation or regression capture only a fraction of the information that is contained in the data. In contrast, copulas contain full information about associations and are therefore frequently applied on finances and other disciplines (Ghosh et al., 2020). In fact, in the bivariate case, copulas are two-dimensional distribution functions restricted to the unit square with uniformly distributed univariate marginals. The theorem of Sklar (see Nelsen (2007)) allows to split the joint distribution function \( H \) of the random vector \((X,Y)\) into the dependence structure \( C \) and the marginal distributions \( F \) and \( G \), that is, \( H(x,y) = C(F(x),G(y)) \) for every \((x,y) \in \mathbb{R}^2\). The aforementioned dependence structure \( C \) is exactly the copula. Since copulas are scale-invariant (see again Nelsen (2007)), it is natural to study scale-free dependence measures on a copula basis. For more background on copulas and their application in dependence modelling, we refer to the books of Nelsen (2007) and Durante and Sempi (2015). The copula-based dependence measure \( qad \), originally introduced as \( \zeta_1 \) in Trutschnig (2011), is defined as a type of distance between the conditional distribution functions of the copula \( C \) underlying the random vector \((X,Y)\) and the uniform distribution representing independence of \( X \) and \( Y \). In other words, \( qad \) measures how much the dependence structure of \((X,Y)\) differs from independence. Contrary to many other approaches, \( qad \) is able to detect both complete dependence (i.e. \( Y \) is a function of \( X \)) as well as independence. The method works as follows: Given a two-dimensional sample \((x_1,y_1), \ldots, (x_n,y_n)\) of size \( n \) from the random vector \((X,Y)\) (see Figure 2a), the normalized ranks of the sample are calculated first (i.e. we get values of the form \((i/n, j/n)\) for \( i,j \in \{1, \ldots, n\} \)). Then the so-called empirical copula \( \hat{E}_n \) is computed (see Figure 2b). As next step, the empirical copula is aggregated to the empirical checkerboard copula (two-dimensional histogram in the copula setting). In fact, the masses of the small squares (empirical copula) are summed up to the larger \( N \times N \) squares, whereby the resolution \( N \) depends on the sample size \( n \) (see Figure 2c,d). Note that by default the resolution of the empirical checkerboard copula is proportional to the square root of the sample size; thus, as for any statistical method, \( qad \) results become more reliable as the sample size increases. We recommend a sample size of no smaller than \( n = 16 \), resulting in a resolution of \( N = 4 \). Finally, the conditional distribution functions of the checkerboard copula are compared with the distribution function of the uniform distribution on the unit interval (in the sense that the area between the graphs is calculated). This step is conducted both for the vertical strips (to calculate the influence of \( X \) on \( Y \)) and the horizontal strips, see, for instance, Figure 2e,f. Computing the sum of all areas and normalizing appropriately with the constant \( 3 \) (see Junker et al. (2021)) yields the two directed \( qad \)-values \( q(X,Y) \in [0,1] \), quantifying the influence of \( X \) on \( Y \) and \( q(Y,X) \in [0,1] \) denoting the influence of \( Y \) on \( X \). High values indicate strong associations, whereas low values describe weakly dependent random variables. Note that for dependence measures which are strictly positive (e.g. \( qad \)), deviation from 0 in the case of independence is to be expected. As example, a value of \( q(X,Y) = 0.2 \) is common for independent random variables \( X \) and \( Y \). Thus, the value of \( q(X,Y) \) alone is clearly insufficient for deciding if, or if not, the sample is likely to come from independent random variables. Therefore, overcoming this problem, a permutation test is implemented in the R package qad to obtain a \( p \)-value for \( q(X,Y) \) and \( q(Y,X) \) in testing for independence, that is, testing the hypothesis \( H_0: q(X,Y) = 0 = q(Y,X) \). Therefore, non-significant \( qad \) values (\( p \)-value >0.05) indicate no dependence. This allows to interpret the obtained values and puts them into perspective.

Furthermore, if we have \( q(X,Y) > q(Y,X) \), then the \( qad \) estimator informs us that the variable \( X \) provides more information about \( Y \) than vice versa. The same holds for the reverse direction. This information is also gathered in the measure for asymmetry, which is computed as \( a(X,Y) := q(X,Y) - q(Y,X) \) and can therefore attain values within the interval \((-1,1)\). Additionally, as a rank-based quantity \( qad \) is robust to outliers and invariant with respect to monotone transformations, for instance, log-transformations.

3 | APPLICATION OF THE R PACKAGE qad

The package \textit{qad} is implemented in the software R (R Development Core Team, 2020) and is publicly available on CRAN (https://cran.r-project.org/web/packages/qad/index.html). The development version of \textit{qad} is accessible via GitHub (https://github.com/griefl/qad). In the following, we briefly sketch the main functions of the package. Additionally, each function contains examples in the description, which are called via the R-help function (e.g. \texttt{?qad}). The following code snippets, which are applied on the data depicted in Figure 1d, sketch the application of \textit{qad}.

3.1 | Calculating the directed dependence measure \( q \)

Given bivariate observations \((x_1,y_1), \ldots, (x_n,y_n)\) of size \( n \) the function \texttt{qad(...)} computes the dependence values \( q(X,Y), q(Y,X) \), the maximum dependence (i.e. \( \max(q(X,Y), q(Y,X)) \)), and the asymmetry in dependence \( a(X,Y) \). The implemented method \texttt{qad(...)} requires two numeric vectors containing the observations of the sample, or, alternatively, accepts a numeric data frame of the form \texttt{data.frame(sample_X, sample_Y)}. The optional argument \texttt{p.value} (default is \texttt{TRUE}) allows to calculate \( p \)-values (based on permutations with \texttt{nperm runs}) for \( q(X,Y) \) and \( q(Y,X) \). A \( p \)-value below 0.05 strengthens the hypothesis that \( X \) and \( Y \) are not independent. The output of \texttt{qad} shows the dependence values and their respective \( p \)-values as well as further descriptive statistics, for example, sample size and the number of unique ranks, which are essential in calculating the resolution of the underlying empirical checkerboard copula. The checkerboard resolution is adjustable through the parameter \texttt{resolution}, however, since the output strongly depends on the resolution, we highly recommend to use the default setting (\texttt{resolution = NULL}),
FIGURE 2  Illustration of the methodology of qad. (a) Sample of size \( n = 40 \) drawn from a slightly noisy U-shaped function. (b) Empirical copula and normalized ranks (points). Note that the masses are uniform on each squares and that, by construction of the empirical copula, the upper right corner of the squares are the normalized ranks. (c) Empirical copula and the checkerboard grid with resolution \( N = 6 \). (d) Checkerboard aggregation. (e) Distance between the conditional distribution functions of the checkerboard copula and the uniform distribution representing independence, for vertical strips (magenta area depicting the distance for one strip) and (f) for horizontal strips.
which uses the optimal choice (optimal in the sense that the estimator performs well independent of the underlying dependence structure; Junker et al., 2021).

```r
library(qad)
# load data
data(“PlantDiversity”)
q <- qad(PlantDiversity, p.value = TRUE, nperm = 10000)
## (Sketch of the output):
Data: x1:=deglacYear, x2:=div
Sample Size: 140
Number of unique ranks: x1:116; x2: 139; (x1,x2): 140
Resolution: 10 x 10
Dependence measures:    q  p.value
q(x1,x2)  0.478  0.000
q(x2,x1)  0.320  0.006
max.dependence  0.478  0.000
asymmetry   0.157
```

Furthermore, the function qad returns an object of class ‘qad’, that allows the application of the generic functions coef(), summary() and plot(). The plot function generates a two-dimensional histogram (heatmap) visualizing the empirical checkerboard copula. The colour of each square corresponds to the density of the normalized ranks (the so-called pseudo-observations). The checkerboard plot helps to understand the type of the dependence structure underlying the variables X and Y. Setting the optional parameter copula to FALSE yields a two-dimensional histogram of the unscaled (raw) data. In our example, we obtain significant q-values \( q(x_1,x_2) = 0.478, p < 0.001 \) and \( q(x_2,x_1) = 0.320, p < 0.01 \), which indicate evidently an asymmetric setting (\( a = 0.157 \)). The additional plots underline the findings and insinuate a slightly inverted U-shaped pattern.

### 3.2 Using qad as a prediction tool

As a by-product of the checkerboard approach, the random variable Y given \( X = x \) and \( X \) given \( Y = y \) can be predicted for every \( x \in \text{Range}(X) \) and \( y \in \text{Range}(Y) \). This additional feature is implemented in the R-function `predict.qad()`. Note that prediction is possible only within the range of measured \( X \) and \( Y \) values; since qad is calculated independently of a parametric regression function, no extrapolation is possible. In contrast to regression methods and many machine learning algorithms, qad does not return point estimates, but probabilities that values of \( Y \) fall in a given range given \( X \) (or vice versa). The function `predict.qad()` requires three arguments: a ‘qad’ object, the conditioning variable and a vector of x-values. Then the function returns the probabilities of the event that \( Y \) falls into the interval \( I_j \) given \( X = x \), or vice versa. Thereby the intervals \( I_j \) are calculated as the retransformed intervals defining the checkerboard grid, that is, for every

```r
#show dependence values q and p.values
coeff(q)
#show total output of qad
summary(q)
#density plot of the emp. checkerboard copula and rescaled histogram
plot(q, density = TRUE, copula = TRUE) # (left panel)
plot(q, density = TRUE, copula = FALSE) # (right panel)
```
3.3 | Multivariate application of qad

Given a multivariate distribution with more than two variables, the function `pairwise.qad(...)` can be applied to quantify all pairwise dependencies and allows an interpretation similar to that of a correlation matrix. The method `pairwise.qad(...)` requires an \( n \times d \)-dimensional numeric matrix, or alternatively, a data.frame of the form `data.frame(sample_X1, sample_X2, ... , sample_Xd)`, describing the observations of a \( d \)-dimensional random vector. Note, that \( p \)-value correction should be applied in multiple testing. To this end, the parameter `p.adjust.method` in the function `pairwise.qad(...)` can be used to select a suitable correction method. Among other details, the main output of `pairwise.qad(...)` is a data.frame containing all pairwise dependencies and corresponding (adjusted) \( p \)-values, which may be readily visualized by `heatmap.qad(...`). Optional parameters allow to select between the directed dependence measures or the asymmetry values and to highlight all significant pairs.

```r
# simulate a four-dimensional sample of size 100
x1 <- runif(100); x2 <- x1^2 + rnorm(100, 0, 0.1);
x3 <- runif(100); x4 <- x3 * x2

# calculate all pairwise qad-values
fit <- pairwise.qad(cbind(x1, x2, x3, x4), p.value = TRUE, p.adjust.method = "fdr")

# visualize the pairwise qad values and highlight significant pairs
heatmap.qad(fit, select = "dependence", significance = TRUE)
```

4 | PERFORMANCE AND COMPARISON OF qad WITH OTHER DEPENDENCE MEASURES

The main features of `qad` compared with seven other well established and in R available dependence measures are summarized in Table 1 and also discussed in Supplementary Information 3. For each measure, we provide information on whether it allows for linear, monotonic or general dependence estimation, whether it is scale-invariant, whether the estimator returns a value in \([0,1]\) and whether it captures asymmetry in dependence. Dependence measures that capture the functional relationship between \(X\) and \(Y\) (see Figure 3a,b,d–f). Note that `qad` returned dependence values slightly smaller than 1 in functional settings without noise (see Figure 3a,b,d–f), which is directly caused by the checkerboard binning. It is guaranteed, however, that asymptotically `qad` attains the maximum value 1 in these settings. Therefore, a direct comparison of two `qad` values has to be done always on consideration of the sample size. Unlike commonly used measures of association like Pearson's \(r\), Spearman's \(\rho\) or more recent measures such as distance correlation and \(MIC\), which are symmetric measures by construction, `qad` (as well as `xicor`) indicated asymmetry in dependence in settings in which (on average) more information on \(Y\) could be obtained by knowing the value of \(X\) than...
TABLE 1

Comparison of features of eight well-established dependence measures: \(dCor\) (described here and in Junker et al. (2021)), distance correlation \(dCor\) (Székely et al. (2007)), maximal information coefficient \(MIC\) (Reshef et al. (2011)), robust copula dependence \(RCD\) (Ding et al. (2017)), randomized dependence coefficient \(rdc\) (Lopez-Paz et al.), \(MICe\) or \(XICOR\) (Chatterjee, 2015) and the commonly used correlation measures by Pearson and Spearman. For each of the eight dependence measures considered here, important properties affecting the functionality and the interpretation of the measures are listed.

| R-function | Asymmetry | Directional | Scale invariance | Estimator in \([0,1]\) |
|------------|-----------|-------------|-----------------|------------------|
| \(dCor\)   | ✓         | ✓           | ✓               | ✓                |
| \(MIC\)    | ✓         | ✓           | ✓               | ✓                |
| Pearson's \(r\) | ✓         | ✓           | ✓               | ✓                |
| \(qad\)    | ✓         | ✓           | ✓               | ✓                |
| \(RCD\)    | ✓         | ✓           | ✓               | ✓                |
| Spearman's \(\rho\) | ✓         | ✓           | ✓               | ✓                |
| \(XICOR\)  | ✓         | ✓           | ✓               | ✓                |

An absolute value is considered which (in this case) essential to assure comparability with the other measures.

5 | APPLYING qad ON ECOLOGICAL DATA

We tested the \(qad\)-package on a dataset of microbiota and additional environmental metadata publicly available at http://ocean-microbiome.embl.de/companion.html (Albanese et al., 2018; de Vargas et al., 2015; Sunagawa et al., 2015; Villar et al., 2015). More precisely, we used the aggregated version of the annotated 16S \(\alpha\)-tags OTU count table, available in the additional materials of Albanese et al. (2018) and conducted a similar analysis. We computed all pairwise \(q\)-values across the relative abundances of genera with less than 10% ties and the environmental variables (mean depth, mean salinity, mean temperature and mean oxygen level), resulting in 94 variables and \(n = 115\) samples and compared the \(qad\) results with the values of Pearson’s and Spearman’s correlation coefficient, and illustrated the information gain provided by \(qad\) over the classical symmetric methods by the number of detected relationships and some specific examples. Since directly comparing dependence values of different measures is not reasonable, we considered the significant relationships detected by each of the measures. As usual, we used \(\alpha = 0.05\) and considered the false discovery rate as procedure for the multiple testing correction.

Overall, the measure \(qad\) returned 2907 significant relationships, whereas Spearman’s \(\rho\) (2564) and Pearson’s \(r\) (1729) found substantially fewer significant pairs. Furthermore, the classical measures \(r\) and \(\rho\) assigned relatively low dependence scores to many relationships that were highly ranked by the measure \(qad\) (see Figure 5a,f). This again results from the fact that the classical measures fail to detect many nonlinear and non-monotonic dependence structures. We depicted several pairs of variables attaining a high \(qad\) value but at the same time a low Pearson and Spearman correlations to demonstrate the major differences in the information gain between symmetric and an asymmetric measure of dependence. For instance, \(qad\) detected a significant asymmetric...
dependence between the variable Methylophilaceae-OM43 clade (variable X) and a *Sphingomonas* strain (variable Y), whereas Pearson's correlation returned a non-significant dependence. The scatterplot depicted in Figure 5d reveals an inverted U-shaped pattern of the data points, that is, knowing the relative abundance of Methylophilaceae-OM43 clade is more informative for the prediction of the *Sphingomonas* strain than vice versa. Moreover, *qad* also picked up a highly asymmetric dependence structure between Alteromonadaceae-SAR92 clade and a *Marinosciillum* strain. The detected dependence structure can be revealed by a log-transformed scatterplot (Figure 5e). Note that *qad* is scale-invariant and hence invariant with respect to log-transformation of samples. We obtained similar results, for example, for the variables Methylophilaceae-OM43 clade and Alcaligenaceae-MWH-UniP1, see Figure 5i, and the variables Methylophilaceae-OM43 clade and mean temperature in °C, depicted in Figure 5j. Additionally, Pearson's *r* reacts very sensitive to outliers (see, for instance, Figure 5c), which explains that there are several highly ranked relationships found by Pearson's correlation but ignored by *qad* or Spearman's correlation.
Spearman’s
Pearson’s
setting, we recommend classical measures of association such as in advance that the data originate from a linear or a monotonic statistical tool exhibits limitations in specific settings. If it is known – here outperforms all other methods in full generality, every sta
tically expands our ability to detect more complex, and potentially asymmetric relationships between organisms and their environ
ture also in nonlinear settings, whereas classical measures only capture linear and monotonic associations. In most real-world situations no, or almost no, prior knowledge about the inter
dependence of variables is available. Aiming at an objective estimate of the strength of dependence, it is therefore unavoidable to work with measures not relying on distributional assumptions. Considering non-monotonic and non-functional relationships naturally expands our ability to detect more complex, and potentially asymmetric relationships between organisms and their environment. We demonstrated that neither of the methods discussed here outperforms all other methods in full generality, every sta
tistical tool exhibits limitations in specific settings. If it is known in advance that the data originate from a linear or a monotonic setting, we recommend classical measures of association such as Pearson’s \( r \), Spearman’s \( \rho \) or dCor. These measures are well established and show greater power in these settings than other methods. In most situations, however, wrongly imposing linearity/monotonicity without prior knowledge may lead to wrong conclusions. We therefore recommend the use of \( qad \) for quantifying pairwise dependencies in the general case. We showed that \( qad \) is powerful in detecting dependence and provides reliable and easily interpretable results.

Another important property of bivariate associations is asymmetry and direction in dependence in the sense that predictability of quantity \( Y \) given knowledge of quantity \( X \) is not the same as vice versa. Considering direction and asymmetry in dependence facilitates the detection and extraction of patterns from ecological data
tests the detection and extraction of patterns from ecological data
tes the testing of refined hypotheses. For instance, correlation analysis
researching for relationships between the abundance of pairs of taxa is usually performed as basis for network inference, which, in turn, facilitates the interpretation of, for example, 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 in such a way that a given taxon is facilitating or inhibiting the growth of another taxon without being affected itself by the other taxon (e.g. commensalism, amensalism). As shown before, conventional correlation analysis neither detects directed relationships nor discriminates between directed and mutual relationships and is therefore of limited value for the interpretation of community dynamics. We are aware of only two methods that are able to quantify directed dependence, namely \( qad \) (Junker et al., 2021) and xicor (Chatterjee, 2021). We have shown that \( qad \) has a higher overall
power in detecting deviation from independence, especially in very noisy datasets \( qad \) performs better than xicor. The power de
iciency of xicor is also discussed in Shi et al. (2022). Furthermore, the implemented estimator in \( qad \) always attains positive values, whereas xicor can attain negative values, which is hard to interpret.

**6 | CONCLUSION**

Our theoretical and real-world examples demonstrate that the measure \( qad \) is able to quantify and indicate the extent of depend
dence also in nonlinear settings, whereas classical measures only capture linear and monotonic associations. In most real-world situations no, or almost no, prior knowledge about the inter
dependence of variables is available. Aiming at an objective estimate of the strength of dependence, it is therefore unavoidable to work with measures not relying on distributional assumptions. Considering non-monotonic and non-functional relationships naturally expands our ability to detect more complex, and potentially asymmetric relationships between organisms and their environment. We demonstrated that neither of the methods discussed here outperforms all other methods in full generality, every statistical tool exhibits limitations in specific settings. If it is known in advance that the data originate from a linear or a monotonic setting, we recommend classical measures of association such as Pearson’s \( r \), Spearman’s \( \rho \) or dCor. These measures are well established and show greater power in these settings than other methods. In most situations, however, wrongly imposing linearity/monotonicity without prior knowledge may lead to wrong conclusions. We therefore recommend the use of \( qad \) for quantifying pairwise dependencies in the general case. We showed that \( qad \) is powerful in detecting dependence and provides reliable and easily interpretable results.

Another important property of bivariate associations is asymmetry and direction in dependence in the sense that predictability of quantity \( Y \) given knowledge of quantity \( X \) is not the same as vice versa. Considering direction and asymmetry in dependence facilitates the detection and extraction of patterns from ecological data
tests the detection and extraction of patterns from ecological data
tes the testing of refined hypotheses. For instance, correlation analysis
researching for relationships between the abundance of pairs of taxa is usually performed as basis for network inference, which, in turn, facilitates the interpretation of, for example, 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 in such a way that a given taxon is facilitating or inhibiting the growth of another taxon without being affected itself by the other taxon (e.g. commensalism, amensalism). As shown before, conventional correlation analysis neither detects directed relationships nor discriminates between directed and mutual relationships and is therefore of limited value for the interpretation of community dynamics. We are aware of only two methods that are able to quantify directed dependence, namely \( qad \) (Junker et al., 2021) and xicor (Chatterjee, 2021). We have shown that \( qad \) has a higher overall
power in detecting deviation from independence, especially in very noisy datasets \( qad \) performs better than xicor. The power de
iciency of xicor is also discussed in Shi et al. (2022). Furthermore, the implemented estimator in \( qad \) always attains positive values, whereas xicor can attain negative values, which is hard to interpret.

**FIGURE 4** Power analysis for different dependence measures. (a–c) Three (noisy) relationships considered in statistical power analysis testing for independence (further results are given in Supplementary Information 3). Empirical power is illustrated for a relationship with vertically added noise following a normal distribution with mean 0 and standard deviation 1. The underlying noise-free relationship is depicted in the left top corner. In each setting, the sample size increases from left \( n = 10 \) to right \( n = 500 \) with increments of 10.

**FIGURE 5** Application of Pearson’s \( r \), Spearman’s \( \rho \) and \( qad \) to a subset of the Tara Oceans dataset. (a, f) Hex bin plot of \( qad \) versus Pearson’s \( r \) (Spearman’s \( \rho \)) for all pairwise relationships; colour code corresponds to count numbers per hexagonal bin. (b, g) Venn diagrams depicting the number of significant relationships across all pairwise associations. (c) Scatterplot of selected pairwise relationships significant w.r.t. Pearson’s \( r \) but not w.r.t. \( qad \). The outlier in the top right corner strongly determines the high correlation value, whereas \( qad \) is robust to outliers. (d, e, i, j) Scatterplots of selected pairwise relationships (highly asymmetric) which are not significant w.r.t. Pearson’s \( r \) (or Spearman’s \( \rho \)) but highly significant w.r.t. the measure \( qad \). In some cases, transforming the axis reveals the underlying dependence structure (e.g. Figure e). (h) Scatterplot depicting a slightly monotonic relationship, detected by both Spearman’s \( \rho \) and \( qad \). (c–e, h–j) The colour of the points corresponds to the coloured stars in the hex bin plot depicting the two \( q \)-values as well as the Pearson (Spearman) correlation. A complete list of the concrete dependence values is provided in the Supplementary Information 3.
In very large datasets, however, *xicor* is more efficient with respect to runtime due to the fact that it uses a $p$-value based on asymptotic theory, whereas *qad* runs a permutation test.

An additional feature of the R-package *qad* is that it provides user-friendly outputs and a number of additional features that facilitate the interpretation of the results as well as functions to use *qad* as a prediction tool.

We conclude that the interpretation of ecological data may be strongly biased by the choice of statistical approaches quantifying dependence between two random variables. The acknowledgement and adequate handling of asymmetry, a universal property of bivariate associations, is an important step towards additional information gain and the avoidance of model bias for small, medium and large datasets, and will advance and allow for a deeper understanding of ecological systems.

**AUTHOR CONTRIBUTIONS**

Florian Griessenberger, Robert R. Junker and Wolfgang Trutschnig designed the study; Florian Griessenberger analysed the data; Florian Griessenberger, Robert R. Junker and Wolfgang Trutschnig wrote the manuscript.

**ACKNOWLEDGEMENTS**

This study was funded by the Austrian Science Fund (FWF, Y 1102 B29) granted to RRJ. Moreover, the first and the second authors gratefully acknowledge the support of the WISS 2025 project ‘IDA-lab Salzburg’ (20204-WISS/225/197-2019 and 20102-F1901166-KZP).

**CONFLICT OF INTEREST**

The authors declare no conflict of interest.

**PEER REVIEW**

The peer review history for this article is available at [https://publons.com/publon/10.1111/2041-210X.13951](https://publons.com/publon/10.1111/2041-210X.13951).

**DATA AVAILABILITY STATEMENT**

All data and supplementary code used in the study can be found at other sources (mentioned at the corresponding paragraphs). The *qad* package is available for the R programming language and can be downloaded at [https://cran.r-project.org/web/packages/qad/index.html](https://cran.r-project.org/web/packages/qad/index.html). This paper describes the latest CRAN-version of *qad* (v.1.0.2). To install the package, run `install.packages("qad")`. The development version of *qad* is available on GitHub ([https://github.com/griefl/qad](https://github.com/griefl/qad)) and can be installed by running `devtools::install_github("griefl/qad", dependencies = TRUE, build_vignettes = TRUE)`. Code stored at github. com is also archived on Zenodo ([Griessenberger et al., 2022](https://doi.org/10.5281/zenodo.6816606), [Griessenberger et al., 2022](https://doi.org/10.5281/zenodo.6816606)) and [Zenodo](https://doi.org/10.5281/zenodo.6816606). Code presented in Supplementary Information 1 and 2 can be found at Mendeley Data [Junker et al., 2022](https://doi.org/10.17632/wxSydhsrj.1), (v1.0.2) (v1.0.2). [Zenodo](https://doi.org/10.5281/zenodo.6816606).

**REFERENCES**

Albanese, D., Riccadonna, S., Donati, C., & Franceschi, P. (2018). A practical tool for maximal information coefficient analysis. *Gigascience*, 7, giy032.

Amblard, P. O., & Michel, O. J. J. (2011). On directed information theory and Granger causality graphs. *Journal of Computational Neuroscience*, 30, 7–16.

Anderson, S. C., Elsen, P. R., Hughes, B. B., Tonietto, R. K., Bietz, M. C., Gill, D. A., Holgerson, M. A., Kuebbing, S. E., McDonough MacKenzie, C., Meek, M. H., & Verissimo, D. (2021). Trends in ecology and conservation over eight decades. *Frontiers in Ecology and the Environment*, 19, 274–282.

Bolt, T., Nomi, J. S., Bzdok, D., & Uddin, L. Q. (2021). Educating the future generation of researchers: A cross-disciplinary survey of trends in analysis methods. *PloS Biology*, 19, e3001313.

Carr, A., Diener, C., Baliga, N. S., & Gibbons, S. M. (2019). Use and abuse of correlation analyses in microbial ecology. *The ISME Journal*, 13, 2647–2655.

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, 116(s36), 2009–2022. [https://doi.org/10.1080/01621459.2020.1758115](https://doi.org/10.1080/01621459.2020.1758115)

Coenen, A. R., & Weitz, J. S. (2018). Limitations of correlation-based inference in complex virus-microbe communities. *mSystems*, 3, e00084–e00018.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logaeres, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Collin, S., Aurý, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., … Karsenti, E. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348, 1261605-1–1261605-11.

Ding, A. A., Dy, J. G., Li, Y., & Chang, Y. (2017). A robust-equitable measure for feature ranking and selection. *The Journal of Machine Learning Research*, 18, 2394–2439.

Durante, F., & Sempi, C. (2015). Principles of copula theory (1st ed.). Chapman and Hall/CRC.

Ghosh, S., Sheppard, L. W., Holder, M. T., Loecke, T. D., Reid, P. C., Bever, J. D., & Reuman, D. C. (2020). Chapter eleven – Cupolas and their potential for ecology. In A. J. Dumbrell, E. C. Turner, & T. M. Fayle (Eds.), *Tropical ecosystems in the 21st century* (pp. 409–468). Academic Press.

Griessenberger, F., Junker, R. R., Petzel, V., & Trutschnig, W. (2021). *qad*: Quantification of asymmetric dependence. R package version 1.0.0. The Comprehensive R Archive Network.

Griessenberger, F., Junker, R. R., Petzel, V., & Trutschnig, W. (2022). *qad* v1.0.2 (v1.0.2). Zenodo. [https://doi.org/10.5281/zenodo.6816606](https://doi.org/10.5281/zenodo.6816606)

Hirano, H., & Takemoto, K. (2019). Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinformatics*, 20, 329.

Junker, R. R., Griessenberger, F., & Trutschnig, W. (2021). Estimating scale-invariant directed dependence of bivariate distributions. *Computational Statistics & Data Analysis*, 153, 107058.

Junker, R. R., Griessenberger, F., & Trutschnig, W. (2022). "code: qad: An R-package to detect asymmetric and directed dependence in bivariate samples". V1. [Mendeley Data](https://doi.org/10.17632/wxSydhsrj.1)

Junker, R. R., Hanusch, M., He, X., Ruiz-Hernández, V., Otto, J. C., Kraushaar, S., Bauch, K., Griessenberger, F., Ohler, L. M., & Trutschnig, W. (2020). Ödenwinkel: An Alpine platform for observational and experimental research on the emergence of multidiversity and ecosystem complexity. *Web Ecology*, 20, 95–106.
Methods in Ecology and Evolution

GRIESENBERGER ET AL.

Griesenberger, F., Trutschnig, W., & Junker, R. R. (2022). qad: An R-package to detect asymmetric and directed dependence in bivariate samples. Methods in Ecology and Evolution, 13, 2138–2149. https://doi.org/10.1111/2041-210X.13951

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Griesenberger, F., Trutschnig, W., & Junker, R. R. (2022). qad: An R-package to detect asymmetric and directed dependence in bivariate samples. Methods in Ecology and Evolution, 13, 2138–2149. https://doi.org/10.1111/2041-210X.13951

Karmon, A., & Pilpel, Y. (2016). Biological causal links on physiological and evolutionary time scales. eLife, 5, e14424.

Lopez-Paz, D., Henig, P., & Schölkopf, B. (2013). The randomized dependence coefficient. Advances in Neural Information Processing Systems, 1–9.

Nelsen, R. B. (2007). An introduction to copulas. Springer Science & Business Media.

Okimoto, T. (2008). New evidence of asymmetric dependence structures in international equity markets. Journal of Financial and Quantitative Analysis, 43, 787–815.

R Development Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large data sets. Science, 334, 1518–1524.

Shi, H., Drton, M., & Han, F. (2022). On the power of Chatterjee’s rank correlation. Biometrika, 109, 317–333.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., ... Bork, P. (2015). Structure and function of the global ocean microbiome. Science, 348, 1261359-1–1261359-9.

Szekely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35, 2769–2794.

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

Villar, E., Farrant, G. K., Follows, M., Garczarek, L., Speich, S., Audic, S., Bittner, L., Blanke, B., Brum, J. R., Brunet, C., Casotti, R., Chase, A., Dolan, J. R., d’Ortenzio, F., Gattuso, J.-P., Grima, N., Guidi, L., Hill, C. N., Jahn, O., ... Ludicorne, D. (2015). Environmental characteristics of Agulhas rings affect interocean plankton transport. Science, 348, 1261447-1–1261447-11.

Wang, Y. X. R., & Huang, H. (2014). Review on statistical methods for gene network reconstruction using expression data. Journal of Theoretical Biology, 362, 53–61.

Zhang, Z., Yan, C., Krebs, C. J., & Stenseth, N. C. (2015). Ecological non-monotonicity and its effects on complexity and stability of populations, communities and ecosystems. Ecological Modelling, 312, 374–384.