Judith H. Parkinson¹ / Raoul Kutil¹ / Jonas Kuppler^{2,3} / Robert R. Junker² / Wolfgang Trutschnig¹ / Arne C. Bathke¹

A Fast and Robust Way to Estimate Overlap of Niches, and Draw Inference

¹ Department of Mathematics, University of Salzburg, Hellbrunner Straße 34, 5020 Salzburg, Austria, E-mail: judith.parkinson@sbg.ac.at, raoulhugo.kutil@sbg.ac.at, Wolfgang.Trutschnig@sbg.ac.at, Arne.Bathke@sbg.ac.at

² Department of Biosciences, University of Salzburg, Hellbrunner Straße 34, 5020 Salzburg, Austria, E-mail: jonas.kuppler@mail.de, Robert.Junker@sbg.ac.at. <http://orcid.org/0000-0003-4409-9367>.

³ Institute of Evolutionary Ecology and Conservation Genomics, University of Ulm, Albert-Einstein-Allee 11 89075 Ulm, Germany, E-mail: jonas.kuppler@mail.de. <http://orcid.org/0000-0003-4409-9367>.

Abstract:

The problem of quantifying the overlap of Hutchinsonian niches has received much attention lately, in particular in quantitative ecology, from where it also originates. However, the niche concept has the potential to also be useful in many other application areas, as for example in economics. We are presenting a fully nonparametric, robust solution to this problem, along with exact shortcut formulas based on rank-statistics, and with a rather intuitive probabilistic interpretation. Furthermore, by deriving the asymptotic sampling distribution of the estimators, we are proposing the first asymptotically valid inference method, providing confidence intervals for the niche overlap. The theoretical considerations are supplemented by simulation studies and a real data example.

Keywords: rank statistics, Asymptotics, Bootstrap, overlap of populations, niches

DOI: 10.1515/ijb-2017-0028

Received: March 31, 2017; **Revised:** February 16, 2018; **Accepted:** May 23, 2018

1 Introduction

Recent methodological progress has reignited interest (see, e.g., [1–6]) in the problem of quantifying the overlap of d -dimensional *Hutchinsonian niches* [7], in the following simply referred to as *H-niches*. These H-niches represent multi-dimensional objects modeled as subsets of an d -dimensional space, where each axis represents an individual trait or environmental factor of interest. In biology, these objects have long been considered to conceptualize and define ecological niches representing the requirements of species to establish and survive in a given habitat. Here, each dimension represents a limiting factor such as temperature, or the availability of resources. Alternatively, H-niches can also be utilized to characterize the “impact niche” of a species where each dimension represents a functional trait. These impact niches or “trait spaces” are being used to work on a broad range of ecological questions. For instance, by comparing the volume and overlap of trait spaces of species within communities, hypotheses how climate or other factors influence species diversity and assembly of communities can be tested. For example, species tend to have similar “trait spaces” if abiotic factors play an important role, while competition between species should limit this similarity [8, 9]. Or, by defining species traits as dimensions that influence resource utilization, it can be explored how species are specialized on certain resources and how several species partition resources that are locally available [10]. While these applications usually focus on species or communities, H-niches or “hypervolumes” can also be utilized to define the states of ecosystems and assess their shifts after environmental change [11]. Besides this ecological focused research, H-niches are becoming increasingly used in other areas of biology, such as animal behavior science or evolution [12, 13] and could be easily be extended to other disciplines, as for example, economics, where one may be interested in the market niche of a company or a product. We refer to the excellent review paper by [14] regarding further details on the theoretical foundation of H-niches, and fields of application.

The geometrical interpretation of the H-niche differs throughout the literature. While Hutchinson’s original concept is a hyper-cube, other authors have interpreted H-niches to be ellipses or minimal convex sets. Each of these interpretations has its benefits and shortcomings, as discussed briefly in [14].

Arne C. Bathke is the corresponding author.

© 2018 Walter de Gruyter GmbH, Berlin/Boston.

This content is free.

In order to evaluate and compare H-niches, their size and their reciprocal overlap are valuable indicators. However, there are (at least) two challenges considering a useful and interpretable quantification of H-niche overlap. First, how to measure overlap of H-niches in one dimension and second, how to extend the concept to multi-dimensional situations? Our goal is to provide such a quantification, focusing mostly on the first question, and without imposing a particular parametric model or distribution family. That is, the approach presented here is fully nonparametric, it can even be applied when some of the measured variables are only ordinal, and it does not depend on the type of scale chosen (e.g., linear, or logarithmic). Further, we aim at devising an inferential tool to not only estimate overlap of hypervolumes, but also to provide a confidence interval for the true (and unknown) overlap. So far, in the vast literature following, adapting, and generalizing Hutchinson's work, no inference procedures can be found yet. The present manuscript attempts to close this gap by developing the appropriate techniques needed to draw inferential conclusions from samples of multivariate data in ecology, economy and beyond.

Let us start with the one-dimensional situation and the goal of quantifying H-niche overlap between two, say, objects. That is, there is one (real-valued) variable of interest, and for each subject under consideration, an observation of this variable can, at least theoretically, be obtained. Also, each subject can be classified into one of exactly two groups. The theoretical distributions of values for the two groups are described by their cumulative distribution functions (cdf), denoted by F and G , respectively.

Language convention: for simplicity of reading, we will not distinguish in the text between the distributions and their cdfs. The respective meaning will be clear from the context. The "range of a distribution F " will be meant to be the interval between $\inf\{t \in \mathbb{R} : F(t) > 0\}$ and $\sup\{t \in \mathbb{R} : F(t) < 1\}$. Further, even though Hutchinson has established the term *d-dimensional hypervolume*, which has caught on in the ecological and biological literature, we prefer to use here *H-niche* instead, due to the connotations evoked by the former term among mathematicians and statisticians. Further, we will assume that F and G are the true underlying population distributions from which samples are drawing representatively. Thus, we will not consider or discuss scenarios where the true distribution and the sampled distribution differ substantially due to bias in the sampling process, or other reasons. We will not explicitly define what the true underlying distribution function is – it may depend on the information of interest. For further discussion on the topic of underlying distributions we refer to [15].

From a statistical point of view, the problem can generally be formulated as quantifying the proportion of values taken by one of the distributions, say F , that is covered by the range of values taken by another distribution, G . It is immediately clear that it would not make sense to limit oneself to the full support of the latter distribution, that is, the full range of values. For one thing, many common families of distributions have unbounded support, so one would basically be restricted to distributions whose densities have compact support. This may not seem crucial, for example in ecology, where one might assume that observations "live" on compact support sets. However, generality of the mathematical model may be an advantage when transferring the method to other fields of application. An arguably more severe problem in using the full support (or its empirical version, the data range) in practice is presented by the fact that the range is a very unrobust, albeit sometimes still not very informative functional [[16], p. 29]. Indeed, the range is so sensitive to outliers that relying on it alone would cause unstable estimation – individual observations can unduly influence it. On the other hand, the situation may also occur where two distributions have the exact same support, but still one of them is concentrated in the center of this support, while the other is more uniformly distributed, indicating that it would be advantageous to incorporate information regarding the full distribution, not only the smallest and largest observation, as in the range.

The recently published nonparametric solution approach by [5] consists of dynamic range boxes where for each quantile $\alpha \in [0, 1]$, it is assessed how much of the central $(1 - \alpha)$ proportion of F , loosely speaking, is covered by the range of the central $(1 - \alpha)$ proportion of G . Integrating over α leads to a rather robust estimator of overlap between the two distributions. As the coverage for each α is between 0 and 1, the integral also takes values between 0 and 1. Outliers only contribute to the calculation for those α -values that are close to 1. When integrating, their influence becomes reduced. The most extreme cases possible are the following. (a) For each α , the range of the central $(1 - \alpha)$ proportion of F is fully included in the corresponding range of G , resulting in an integrated overlap value of 1. (b) None of the central F ranges are included in the corresponding G ranges, resulting in an integrated overlap of 0.

While their method has the advantages of being robust, it also suffers from three disadvantages which are remedied by the approach proposed here.

1. Dynamic range boxes [5] require the calculation of mutual coverages for each α individually, resulting in either computational burden or resorting to an only approximate solution.
2. At each step, the relative proportion of the covered range is calculated as a fraction of the respective interval lengths. This assumes that the measured variables are metric, and the proportion is not invariant

under nonlinear transformations of the variables. For example, sometimes researchers prefer to transform observed values onto a logarithmic scale – a change between linear and logarithmic scales would however change the results obtained using the above method.

3. The sampling distribution of the integral above is not known. Therefore inference (confidence intervals) can not be provided. As mentioned in different guidelines, see [17–19], confidence intervals should be provided along with the estimator, whenever possible.

Figure 1: Graphical illustration of the differences in derivation of the two methods, present paper (left), and *Junker et. al.* (right). This plot shows exemplarily two underlying normal distributions, and evaluation of niche overlap at $\alpha = 5\%$. Both methods integrate over all values of α between 0 and 1.

Figure 1 illustrates the difference between the approach proposed in this paper and the one proposed by [5], which is the method most closely resembling the newly proposed one. Please note that the methods were simplified for the visualization. The red arrows indicate the quantities measured at each α , in order to calculate the niche overlap. The new method then integrates over all α , the precise derivation of this calculation is given later. The method by [5] measures the overlap of the two individual α intervals and averages across the resulting values.

2 New measures of H-niche containment and overlap

We propose the following approach for quantifying H-niche containment and overlap. For simplicity we will start with one-dimensional H-niche containment and overlap before shortly explaining how to expand the method to the d -dimensional case in Section 2.5.

For each quantile α , we consider the central $(1 - \alpha)$ portion of F , as above. However, we now calculate the mass that G assigns to exactly this region. This defines an H-niche containment function $h_2 : [0, 1] \rightarrow [0, 1]$,

$$h_2(\alpha) = G[F^{-1}(1 - \alpha/2)] - G[F^{-1}(\alpha/2)],$$

with F^{-1} the quantile function of F . With h_1 we will denote the analogous function where the roles of F and G are switched. Note that $h_i(\alpha)$, $i = 1, 2$, are monotonically decreasing in $\alpha \in [0, 1]$.

Based on independent and respectively *i.i.d.* samples $X_1, \dots, X_n \sim F$ and $Y_1, \dots, Y_m \sim G$, canonical estimators for $h_i(\alpha)$, $i = 1, 2$, are given by

$$\hat{h}_2(\alpha) = \hat{G}_m[\hat{F}_n^{-1}(1 - \alpha/2)] - \hat{G}_m[\hat{F}_n^{-1}(\alpha/2)],$$

and analogously defined $\hat{h}_1$, where $\hat{F}_n$ and $\hat{G}_m$ are the respective empirical distribution functions for F and G . The empirical distribution functions $\hat{F}_n$ and $\hat{G}_m$ are known to be strongly consistent, uniformly over $t \in \mathbb{R}$, and unbiased at each t .

Clearly, for each $\alpha \in (0, 1)$, $\hat{h}_1(\alpha)$ and $\hat{h}_2(\alpha)$ are invariant under strictly monotone (isotone or antitone) transformations of the observations. That is, if $X_1, \dots, X_n, Y_1, \dots, Y_m$ are replaced by $\tau(X_1), \dots, \tau(X_n), \tau(Y_1), \dots, \tau(Y_m)$ for strictly monotone τ , the quantities $\hat{h}_1(\alpha)$ and $\hat{h}_2(\alpha)$ remain unchanged. Thus, these functions are well equipped to serve as the foundation for a fully nonparametric quantification of H-niche containment and overlap.

Recall that the receiver operating characteristic (ROC) curve for two distribution functions F and G is defined as $\text{ROC}(t) = 1 - G[F^{-1}(1 - t)]$. The process

$$\sqrt{n}(\hat{G}_m[\hat{F}_n^{-1}(t)] - G[F^{-1}(t)]), \quad t \in (0, 1),$$

is called the empirical ROC process, for which the following result holds.

Lemma 2.1

Let X and Y be independently distributed according to F and G , with continuous density functions f and g , respectively. If $\frac{n}{m}$ converges to a $\lambda > 0$ for $n \rightarrow \infty$, where n and $m = m(n)$ denote the sample sizes, and if the slope of $G[F^{-1}(t)]$, that is $g[F^{-1}(t)]/f[F^{-1}(t)]$, is bounded on any interval (a, b) with $0 < a < b < 1$, then the empirical ROC process converges weakly to the sum of two tight Gaussian processes. More explicitly, it holds that

$$\begin{aligned} & \sqrt{n}[\hat{G}_m\hat{F}_n^{-1}(t) - GF^{-1}(t)] \\ = & \sqrt{\lambda}B_1^{(n)}[GF^{-1}(t)] + \frac{g[F^{-1}(t)]}{f[F^{-1}(t)]}B_2^{(n)}(t) + o[n^{-1/2}(\log n)^2] \end{aligned} \quad (1)$$

for $t \in [0, 1]$ almost surely, uniformly on $[a, b]$, where $\{B_1^{(n)}, B_2^{(n)}, 0 \leq t \leq 1\}$ are two sequences of independent Brownian Bridges.

Proof

The Lemma is equivalent to Theorem 2.2. [20].

We will denote the limit process of the empirical ROC process by $\mathbb{L}(t)$. The similarity between our H-niche containment function and the ROC process will be exploited in Section 2.2 for the derivation of asymptotic properties of the proposed estimators.

2.1 Asymmetric H-niche containment and its estimation

In this section we will focus on developing an estimator for, roughly speaking, how much of the range of F is contained in the range of G . As a robust measure of H-niche containment, let us now consider the statistical functional

$$\begin{aligned} I_2 &= \int_0^1 h_2(\alpha) d\alpha = \int_0^1 G[F^{-1}(1 - \alpha/2)] d\alpha - \int_0^1 G[F^{-1}(\alpha/2)] d\alpha \\ &= 2 \left(\int_{F^{-1}(1/2)}^\infty G(t) dF(t) - \int_{-\infty}^{F^{-1}(1/2)} G(t) dF(t) \right), \end{aligned}$$

as well as I_1 , with the roles of F and G switched. These integrals basically average over H-niche overlap at the center, as well as at the margins of the individual H-niches – and everywhere in between. They take into account the full distribution information of F , all the while set into corresponding relation with G .

For a (not necessarily continuous) distribution function F let F^{-1} denote the pseudoinverse (quantile function). The following identities are straightforward to verify.

$$\begin{aligned} \lim_{t \rightarrow 0+} F^{-1}(t) &= \inf\{t \in \mathbb{R} : F(t) > 0\} =: t_F \\ \lim_{t \rightarrow 1-} F^{-1}(t) &= \sup\{t \in \mathbb{R} : F(t) < 1\} =: \bar{t}_F \end{aligned}$$

In the sequel we will refer to $\text{Ran}(F) := [t_F, \bar{t}_F]$ as the range of F . Letting μ_G denote the probability measure corresponding to the distribution function G , the quantities I_2 and I_1 can be expressed as

$$I_2 = \int_{(0,1)} \mu_G\left((F^{-1}(\frac{\alpha}{2}), F^{-1}(1 - \frac{\alpha}{2}))\right) d\alpha, \quad (2)$$

$$I_1 = \int_{(0,1)} \mu_F\left((G^{-1}(\frac{\alpha}{2}), G^{-1}(1 - \frac{\alpha}{2}))\right) d\alpha. \quad (2)$$

The subsequent lemma gathers the most important properties of I_1 and I_2 .

Lemma 2.2

Suppose that F and G are distribution functions. Then the following properties hold:

1. If G is continuous and $F = G$, then $I_1 = I_2 = \frac{1}{2}$.
2. $I_2 = 0$ if and only if $\mu_G((\underline{t}_F, \bar{t}_F)) = 0$. If G is continuous then $I_2 = 0$ is equivalent to $\mu_G(\text{Ran}(F)) = 0$.
3. For $I_2 = 1$ it holds $\mu_G([F^{-1}(\frac{1}{2}), F^{-1}(\frac{1}{2}+)]) = 1$. If G is continuous then $I_2 = 1$ is equivalent to $\mu_G((F^{-1}(\frac{1}{2}), F^{-1}(\frac{1}{2}+))) = 1$.
4. I_1 and I_2 are invariant under strictly monotone, continuous transformations, where the same transformation ϕ is being applied to $X \sim F$ and to $Y \sim G$.

Proof

- i. If $F = G$ and G is continuous, considering that $F^{-1} \circ F(t) = t$ holds for all $t \in (0, 1)$, the identity $I_1 = I_2 = \int_{(0,1)} (1 - \alpha) d\alpha = \frac{1}{2}$ follows immediately.
- ii. If $I_2 = 0$ then, using the fact that $\alpha \mapsto \mu_G((F^{-1}(\frac{\alpha}{2}), F^{-1}(1 - \frac{\alpha}{2})))$ is (not necessarily strictly) decreasing, it follows that $\mu_G((F^{-1}(\frac{\alpha}{2}), F^{-1}(1 - \frac{\alpha}{2}))) = 0$ holds for arbitrarily small $\alpha > 0$, implying $\mu_G((\underline{t}_F, \bar{t}_F)) = 0$. The other implication is obvious. In case G is continuous $\mu_G((\underline{t}_F, \bar{t}_F)) = 0$ is equivalent to $\mu_G([\underline{t}_F, \bar{t}_F]) = 0$, which completes the proof of the second assertion.
- iii. If $I_2 = 1$, again using monotonicity together with left-continuity of F^{-1} on $(0, 1)$ and the fact that $\bigcap_{\alpha \in (0,1)} (F^{-1}(\frac{\alpha}{2}), F^{-1}(1 - \frac{\alpha}{2})) \subseteq [F^{-1}(\frac{1}{2}), F^{-1}(\frac{1}{2}+)]$, the desired property $\mu_G([F^{-1}(\frac{1}{2}), F^{-1}(\frac{1}{2}+)]) = 1$ follows immediately. In case G is continuous, $I_2 = 1$ implies $\mu_G((F^{-1}(\frac{1}{2}), F^{-1}(\frac{1}{2}+))) = 1$. Since the reverse implication is trivial the proof of the third assertion is complete.
- iv. Let ϕ be a strictly monotone transformation function and denote the new distribution functions of X and Y with F_ϕ and G_ϕ . There exists an inverse ϕ^{-1} of ϕ . Now $F_\phi(t) = F \circ \phi^{-1}(t)$ and the inverse $F_\phi^{-1}(t) = \phi \circ F^{-1}(t)$, analogue for G .

Then

$$\begin{aligned} \mu_{G_\phi}((F_\phi^{-1}(\frac{\alpha}{2}), F_\phi^{-1}(1 - \frac{\alpha}{2}))) &= G_\phi(\phi \circ F^{-1}(1 - \frac{\alpha}{2})) - G_\phi(\phi \circ F^{-1}(\frac{\alpha}{2})) \\ &= G \circ \phi^{-1} \circ \phi \circ F^{-1}(1 - \frac{\alpha}{2}) - G \circ \phi^{-1} \circ \phi \circ F^{-1}(\frac{\alpha}{2}) \\ &= \mu_G((F^{-1}(\frac{\alpha}{2}), F^{-1}(1 - \frac{\alpha}{2}))) \end{aligned}$$

holds for all $\alpha \in [0, 1]$.

The following examples illustrate that we can not get sharper versions of assertions (i)-(iv) in Lemma 2.2.

Example 2.3

In case of $F = G = \mathbf{1}_{[0, \infty)}$ we have $(\underline{t}_F, \bar{t}_F) = \emptyset$ and $\mu_G(\emptyset) = 0$. It then follows that $I_1 = I_2 = 0$.

Example 2.4

Consider $G = \mathbf{1}_{[1, \infty)}$ and let F denote the distribution function of a uniform distribution on $[0, 1]$. Then $\text{Ran}(F) = [0, 1]$, so $\mu_G(\text{Ran}(F)) = 1$ although $\mu_G([0, 1)) = 0$.

Example 2.5

Consider $G = \mathbf{1}_{[0, \infty)}$ and let F denote the distribution function of a discrete uniform distribution on $\{0, 1\}$. Then $[F^{-1}(\frac{1}{2}), F^{-1}(\frac{1}{2}+)] = [0, 1]$, from which we get $\mu_G([F^{-1}(\frac{1}{2}), F^{-1}(\frac{1}{2}+)]) = 1 < 0 = \mu_G((F^{-1}(\frac{\alpha}{2}), F^{-1}(\frac{\alpha}{2}+)))$.

Intuitively, the expressions I_1 and I_2 can be interpreted as follows. If the range of F is well contained within the range of G then the values of the H-niche containment function h_2 are small, and the integral I_2 is close to zero. In these cases, h_1 takes larger values, and I_1 is closer to one. That is, I_1 can be interpreted as a measure of “how much is F (i.e., its range) contained in G ”, while I_2 quantifies “how much is G contained in F ”. This will be made more precise in Lemma 2.8.

Remark 2.6

In the following, we will always assume $X_i \sim F$ and $Y_j \sim G$ to be independent, absolutely continuous random variables, even if not explicitly mentioned. Further requirements for the upcoming results are stated directly in the individual lemmas and theorems.

Denote the distribution of F below its median as F_1 and above its median as F_2 , such that

$$F_1(t) = \begin{cases} 2F(t), & t < F^{-1}(1/2), \\ 1, & t \geq F^{-1}(1/2), \end{cases} \quad F_2(t) = \begin{cases} 0, & t < F^{-1}(1/2), \\ 2F(t) - 1, & t \geq F^{-1}(1/2). \end{cases} \quad (3)$$

Remark 2.7

The random variables $X^{(1)} \sim F_1$ and $X^{(2)} \sim F_2$ can be constructed from X by conditioning on $X \leq F^{-1}(0.5)$ and $X > F^{-1}(0.5)$, respectively.

Lemma 2.8

Let $X^{(1)} \sim F_1$, $X^{(2)} \sim F_2$, and $Y \sim G$ be three independent, absolutely continuous random variables. Then, I_2 can be written as

$$I_2 = P(Y < X^{(2)}) - P(Y < X^{(1)}) = P(X^{(1)} < Y < X^{(2)}), \quad (4)$$

using the fact that by construction $X^{(1)} < X^{(2)}$

Proof

Compare Chapter 1.4 of [20].

Remark 2.9

When reformulating Lemma 2.8 for general, not necessarily continuous random variables, the normalized version of the distribution function, $F(x) = \frac{1}{2}[P(X \leq x) + P(X < x)]$ (as given in Definition 1.3 of [21]), should be used, resulting in

$$I_2 = P(Y < X^{(2)}) - P(Y < X^{(1)}) + \frac{1}{2}P(Y = X^{(2)}) - \frac{1}{2}P(Y = X^{(1)}).$$

In analogy to F_1 and F_2 , define distribution functions G_1 and G_2 based on G . Let $Y^{(1)} \sim G_1$, $Y^{(2)} \sim G_2$ be random variables, independent of $X^{(1)}$, $X^{(2)}$. Using these definitions, we can rearrange I_2 to

$$\frac{1}{2}[P(Y^{(1)} < X^{(2)}) + P(Y^{(2)} < X^{(2)}) - P(Y^{(1)} < X^{(1)}) - P(Y^{(2)} < X^{(1)})]. \quad (5)$$

Lemma 2.10

If F and G are continuous distribution functions then $I_1 + I_2 \in [0, 1]$

Proof

Due to the way the random variables are defined, one of the two probabilities will always be 1, while the other ranges between $[0, 1]$.

$$i=1, 2, P(Y^{(i)} < X^{(i)}) + P(X^{(i)} < Y^{(i)}) = 1$$

Expressions (4) and (5) provide a rather intuitive probabilistic interpretation of I_2 (and analogously for I_1).

Using the canonical estimators $\hat{F}_n$ and $\hat{G}_m$ for the cdfs, we can construct the following consistent plugin estimators for I_2 (and I_1).

$$\hat{I}_2 = 2 \left(\int_{\hat{F}_n^{-1}(1/2)}^{\infty} \hat{G}_m(t) d\hat{F}_n(t) - \int_{-\infty}^{\hat{F}_n^{-1}(1/2)} \hat{G}_m(t) d\hat{F}_n(t) \right) \quad (6)$$

The functional I_1 is estimated analogously with the roles of $\hat{F}_n$ and $\hat{G}_m$ switched. Although the estimator in eq. (6) looks rather complex at first sight, there exists a surprisingly simple shortcut formula in order to evaluate the integral. This formula is based on rank statistics, as follows.

Rank all $m + n$ observations, in case of ties use midranks. Sort the X -sample, and without loss of generality, assume that $X_1 \leq X_2 \leq \dots \leq X_n$ (to simplify notation). Then, all the X -observations below their median are $X_1, \dots, X_K$ where K is the largest integer less than or equal to $(n + 1)/2$. Denote their ranks by $R_1^{X<}, \dots, R_K^{X<}$, and denote the ranks of the remaining X -sample by $R_{K+1}^{X>}, \dots, R_n^{X>}$. Finally, let

$$R^{X<} = \sum_{i=1}^K R_i^{X<}, \quad R^{X>} = \sum_{i=K+1}^n R_i^{X>}.$$

Lemma 2.11

For $R^{X>}$ and $R^{X<}$ defined as above, $\hat{I}_2$ can be rearranged to

$$\frac{2}{mn} (R^{X>} - R^{X<}) + \frac{1}{2}c, \quad (7)$$

with $c = -n/m$ for n even and $c \approx -n/m$ for n odd, and n and m large.

Proof

The proof of this is given in the Appendix, Proof A.

Remark 2.12

Note, that the expression of c in the Lemma above only holds for samples without ties. The result for samples with ties is given in the Appendix, eq. (8) after Proof A.

The expression in (7) allows for a quick and direct estimation of I_2 , and, analogously, of I_1 . This is a major computational advantage, potentially also improving numerical stability, compared to the approach by [5], which required evaluations of an H-niche containment function similar to h_1 and h_2 , but for several different values of α separately.

2.2 Asymptotic sampling distribution, and inference

In this section, we will establish the sampling distribution of the estimator defined in eq. (6), using the relation to the empirical ROC process mentioned in the beginning of Section 2. To this end, let $\mathbb{D} := D[0, 1]$ be the Skorokhod space, that is, the space of all càdlàg functions defined on $[0, 1]$ (see Chapter VI of [21]), and $\mathbb{E} := [0, 1]$. Define the map

$$\varphi : \mathbb{D} \rightarrow \mathbb{E}, \quad \varphi([f(t)]_{t \in [0,1]}) = \int_0^1 f\left(1 - \frac{\alpha}{2}\right) d\alpha - \int_0^1 f\left(\frac{\alpha}{2}\right) d\alpha.$$

Note that φ is a continuous linear map. The topology here is the Skorokhod topology, as given in Theorem 1.14 of Chapter VI of [21].

Theorem 2.13

Let the map φ be defined as above. Then it holds that $\sqrt{n}(\varphi(\hat{G}_m \circ \hat{F}_n^{-1}) - \varphi(G \circ F^{-1}))$ converges weakly to $\varphi'_{G \circ F^{-1}}(\mathbb{L})$, where $\mathbb{L}$ is the limit process of Lemma 2.1. Further, $\varphi'_{G \circ F^{-1}}(\mathbb{L})$ is asymptotically normally distributed with expectation zero.

Proof

The proof can be found in the appendix, Proof B.

Remark 2.14

In general $\hat{G}_m \circ \hat{F}_n^{-1}$ is not a càdlàg function. However it is a monotone decreasing step function that can simply be transformed into a càdlàg function with identical values for ϕ .

Remark 2.15

An alternative way to derive the normality of $\hat{I}_2$ in eq. (6) could be pursued by rewriting it as a finite sum of random variables, namely

$$\frac{n}{2} \hat{I}_2 = \sum_{i=K+1}^n \hat{G}_m(X_i) - \sum_{i=1}^K \hat{G}_m(X_i),$$

with slight modification in case of ties among the X_i . After establishing an asymptotic equivalence result with respect to the similarly defined sum

$$\sum_{i=K+1}^n G(X_i) - \sum_{i=1}^K G(X_i),$$

asymptotic normality can be established using the Lyapunov central limit theorem.

We have decided to elaborate more on the approach that uses the similarity to the ROC process, as we think that this relation may open up the path for further methodological developments. Thus, we intended to demonstrate its potential usefulness. But the next steps also show some limits when using the empirical process theory in practice.

Inference requires a valid variance estimation for the sampling distribution. However, considering the weak convergence of the empirical ROC process to the sum of two Brownian Bridges, we easily realize that calculating the resulting variance based on result (1) will be rather complex. Deriving a consistent empirical estimator for

the variance based on this result alone would be even more complicated, as it would require density estimation for both distributions. Yet, as our goal is to construct confidence intervals, we will need an alternative approach for finding a consistent estimator of the variance. In our case the bootstrap supplies a feasible and justified alternative.

Theorem 2.16

Let $\hat{G}_m^*$ and $\hat{F}_n^*$ be the empirical distribution functions of the bootstrapped sample and $\hat{I}_2^*$ the estimator of I_2 based on the bootstrapped sample. Then it holds that $\sqrt{n}[\varphi(\hat{G}_m^* \hat{F}_n^{*-1}) - \varphi(\hat{G}_m \hat{F}_n^{-1})]$ is asymptotically normally distributed with identical limit distribution as $\sqrt{n}[\varphi(\hat{G}_m \hat{F}_n^{-1}) - \varphi(GF^{-1})]$.

Proof

The proof is given in the Appendix, Proof C.

Now, we can give a confidence interval based solely on the empirical quantiles of the bootstrap process. Through simulation of $\hat{I}_2^*$ we can obtain an approximate α -quantile of $\sqrt{n}(\hat{I}_2^* - \hat{I}_2)$, denoted by $\hat{z}_\alpha^*$.

Theorem 2.17

The asymptotic bootstrap confidence interval at level $1 - \alpha$ for the above defined estimator $\hat{I}_2$ is given by

$$[\hat{I}_2 + \frac{1}{\sqrt{n}} \hat{z}_{\frac{\alpha}{2}}^*, \hat{I}_2 + \frac{1}{\sqrt{n}} \hat{z}_{1-\frac{\alpha}{2}}^*].$$

Further, as our process converges to a normal distribution, we can construct asymptotic confidence intervals by applying Slutsky's Lemma. We therefore can derive a symmetric confidence interval directly from the variance of the bootstrapped data.

Theorem 2.18

A second confidence interval at level $1 - \alpha$ for the estimator $\hat{I}_2$ is given by

$$[\hat{I}_2 - \frac{1}{\sqrt{n}} z[\text{Var}(\hat{I}_2^*)]^{1/2}, \hat{I}_2 + \frac{1}{\sqrt{n}} z[\text{Var}(\hat{I}_2^*)]^{1/2}],$$

where z denotes the $(1 - \alpha/2)$ -quantile of the standard normal distribution.

Obviously, the same properties hold for I_1 . Observe that the first confidence interval is not necessarily symmetric as the absolute values of the two bootstrap quantiles may differ, in contrast to the quantiles of the normal distribution. Note that in cases where the estimator is close to 0 or 1 this might yield better results, as the boundaries of the confidence interval based on the empirical quantiles never exceed the interval $[0, 1]$, while the approach using the variance estimator could exceed it.

For certain special cases, a simplified direct variance estimator can be constructed.

Lemma 2.19

Let F and G be continuous distribution functions. Reconsider the random variables $X^{(1)}, X^{(2)}, Y^{(1)}, Y^{(2)}$ defined in Section 2.1. Assume $G^{-1}(\frac{1}{2}) = F^{-1}(\frac{1}{2})$ holds true, that is, the distributions F and G are assumed to have the same median. In that case, I_2 , as given in eq. (5), further simplifies to

$$I_2 = \frac{1}{2} [1 + P(X^{(1)} < Y^{(1)}) - P(X^{(2)} < Y^{(2)})],$$

and analogously for I_1 , which immediately yields $I_1 + I_2 = 1$.

Proof

By definition of the random variables every realization of $X^{(1)}$ and $Y^{(1)}$ is contained in the interval $(-\infty, F^{-1}(\frac{1}{2})]$. Similarly, $X^{(2)}, Y^{(2)}$ can only take values in $(F^{-1}(\frac{1}{2}), \infty)$. It then follows directly that $P(X^{(1)} < Y^{(2)}) = P(Y^{(1)} < X^{(2)}) = 1$ and $P(X^{(2)} < Y^{(1)}) = P(Y^{(2)} < X^{(1)}) = 0$. Entering these results into eq. (5) leads directly to the statement above.

As the random variables involved in estimating $P(X^{(1)} < Y^{(1)})$ and $P(X^{(2)} < Y^{(2)})$ are independent under the assumption of equal medians, we can construct a variance estimator for these cases using a simpler approach. We only require variance estimators for the canonical estimators of these two probabilities, namely for $\int_{-\infty}^{\infty} \hat{F}_1(t) d\hat{G}_1(t)$ and $\int_{-\infty}^{\infty} \hat{F}_2(t) d\hat{G}_2(t)$. To this end, we can use the procedure developed by [22]. Without loss of generality, assume again, as in the description before eq. (7), that the X -observations below their median are $X_1, \dots, X_K$, while those above their median are $X_{K+1}, \dots, X_n$. Analogously, sort the Y -sample to $Y_1, \dots, Y_L$ and $Y_{L+1}, \dots, Y_m$. Now, the samples of X and Y observations which are below (above) their respective medians

are compared with each other. For $i = 1, \dots, K$, denote by $R_i^{X(<)}$ the rank of X_i among the combined sample $X_1, \dots, X_K, Y_1, \dots, Y_L$,

and by $\bar{R}^{X(<)} = \frac{1}{K} \sum_{i=1}^K R_i^{X(<)}$ their average. Analogously, define $\bar{R}^{Y(<)}, \bar{R}^{X(>)},$ and $\bar{R}^{Y(>)}$. Finally, the internal ranks, where observations are ranked only within each of these four groups, are denoted by $R_i^{(X<)}, R_i^{(Y<)}, R_i^{(X>)},$ and $R_i^{(Y>)}$, respectively.

Lemma 2.20

Define

$$s_{X1}^2 = \frac{1}{L^2(K-1)} \sum_{i=1}^K (R_i^{X(<)} - R_i^{(X<)} - \bar{R}^{X(<)} + \frac{K+1}{2})^2,$$

$$s_{X2}^2 = \frac{1}{(m-L)^2(n-K-1)} \sum_{j=K+1}^n (R_j^{X(>)} - R_j^{(X>)} - \bar{R}^{X(>)} + \frac{n-K+1}{2})^2,$$

and s_{Y1}^2 and s_{Y2}^2 analogously. Further the same requirements as in Lemma 2.19 shall hold. A consistent variance estimator for $\hat{I}_2$, the estimator of the simplified I_2 , is given by

$$s_2^2 = \frac{L+K}{2} \cdot \left(\frac{s_{X1}^2}{K} + \frac{s_{Y1}^2}{L} \right) + \frac{n+m-L-K}{2} \cdot \left(\frac{s_{X2}^2}{n-K} + \frac{s_{Y2}^2}{m-L} \right).$$

Proof

This Lemma follows directly from the results of [22], as well as the decomposition of $\hat{I}_2$ given in eq. (5).

Note that this is the same variance estimator as for $\hat{I}_1$.

Theorem 2.21

The confidence interval at level $1 - \alpha$ for the simplified estimator $\hat{I}_2$ is then simply given by

$$[\hat{I}_2 - \frac{1}{\sqrt{n}} z_{1-\alpha/2} s_2, \hat{I}_2 + \frac{1}{\sqrt{n}} z_{1-\alpha/2} s_2],$$

where $z_{1-\alpha/2}$ is again the $(1 - \alpha/2)$ -quantile of the standard normal distribution.

Remark 2.22

Clearly, in practice, one may not know whether the assumption of equal medians is fulfilled, thus, the calculations for the simplified method may have limited application. So, why mention it then? Apart from the intriguing mathematical fact, a variance estimation in case of equal medians could be used in the construction of a hypothesis test. Additionally, we hope that the techniques used in the simplified variance estimation may be useful to other situations, in particular since simulations (see below) showed that one in general obtains narrower confidence intervals with the simplified method.

2.3 Symmetric H-niche overlap and its estimation

The concept of H-niche overlap describes the common share of two H-niches, or of their respective distribution ranges. The asymmetric containment index derived in the previous section provides a basis for defining a straightforward symmetric H-niche overlap index, namely as $O(F, G) = 4I_1I_2$, where the factor 4 arises from our desire to standardize the total overlap to 1 for two identical distributions. Using notation as in the intuitive interpretation above we can write

$O(F, G) = 4P(X^{(1)} < Y < X^{(2)})P(Y^{(1)} < X < Y^{(2)})$. Thus, the symmetric H-niche overlap index $O(F, G)$ has the following properties, now formulated in terms of H-niches.

Similar to Lemma 2.2 we will give the most important properties of $O(F, G)$ in the subsequent lemma.

Lemma 2.23

Suppose that F and G are distribution functions, not necessarily continuous. Then the following properties hold:

1. In case of identical H-niches, $F = G$, and G continuous we have $O(F, G) = 1$.
2. $O(F, G) = 0$ if and only if $\mu_G((t_F, \bar{t}_F)) = 0$ or $\mu_F((t_G, \bar{t}_G)) = 0$. If F and G are continuous then $O(F, G) = 0$ is equivalent to $\mu_G(\text{Ran}(F)) = 0$, and simultaneously $\mu_F(\text{Ran}(G)) = 0$.

3. The overlap index $O(F, G)$ is invariant under strictly monotone transformations, where the same transformation ϕ is being applied to $X \sim F$ and to $Y \sim G$.

Proof

(i) This follows directly from Lemma 2.2.

(ii) If $O(F, G) = 0$ then either I_1 or I_2 must be zero. By Lemma 2.2 this is the case if either $\mu_G((\underline{t}_F, \bar{t}_F)) = 0$ or $\mu_F((\underline{t}_G, \bar{t}_G)) = 0$. Obviously the other implication holds, too. The proof for the continuous case is equivalent. Note that, due to continuity, if one of the asymmetric overlaps is zero, the other will be zero as well.

(iii) This follows again directly from Lemma 2.2.

Example 2.24

Consider $F = 1_{[0, \infty)}$ and let $G = U[0, 2]$. Then $(\underline{t}_F, \bar{t}_F) = \emptyset$, from which we get $\mu_G((\underline{t}_F, \bar{t}_F)) = 0$. On the other hand, $\mu_F((\underline{t}_G, \bar{t}_G)) = \mu_F((0, 2)) = 1 = \mu_F([G^{-1}(\frac{1}{2}), G^{-1}(\frac{1}{2}+)])$.

As stated in the previous lemma, identical H-niches yield a symmetric H-niche overlap of 1. On the other hand, if one H-niche is well contained within the range of the other, then one of the factors, and thus the product $O(F, G)$ will be close to zero. By “well contained”, we mean that most of the weight of the one niche is centered in a small area around the median of the other, as, for example, in the scenario of Example 2.24. A canonical and consistent estimator is given by $\hat{O}(F, G) = 4\hat{I}_1\hat{I}_2$, where the shortcut formula (7) for $\hat{I}_2$, and its analog for $\hat{I}_1$, can be used.

2.4 Asymptotic distribution

In deriving confidence intervals for the overlap, we can not assume independence of the two random variables $\hat{I}_1$ and $\hat{I}_2$. Therefore, the properties of the individual convergences do not suffice here. As a consequence, we can only provide a conservative confidence interval for the general case, since estimation of the joint asymptotic distribution is rather complex. A conservative $(1 - \alpha)$ -interval can be constructed based on individual $(1 - \alpha/2)$ -intervals for I_1 and I_2 , which yields a simultaneous $(1 - \alpha)$ -region for (I_1, I_2) , and thus a straightforward $(1 - \alpha)$ -interval for $4I_1I_2$. We will restrict the confidence interval obtained by this method to the interval $[0, 1]$, as the true overlap only ranges in this interval.

Note that for an important special case, namely under the assumption of equal medians, $G^{-1}(\frac{1}{2}) = F^{-1}(\frac{1}{2})$, a confidence interval for $O(F, G)$ can simply be constructed using a confidence interval for I_2 and the fact that under the given assumption, $O(F, G) = 4I_2(1 - I_2) =: \xi(I_2)$. That is, the confidence interval for I_2 , namely $\mathcal{J}_2 = [I_{2L}, I_{2U}]$, is simply transformed to $\xi(\mathcal{J}_2)$. If $\mathcal{J}_2$ does not contain $1/2$, the interval for $O(F, G)$ is directly obtained by transforming the limits I_{2L}, I_{2U} . If $1/2$ is contained in the confidence interval, the upper limit of the confidence interval for $O(F, G)$ is set to 1.

2.5 Extension to the d -dimensional case

In most applications of H-niche overlap, we not only have one trait, but many. Therefore, it is indispensable to extend our approach to the multi-dimensional case. In order to maintain certain properties for our d -dimensional H-niche overlap, we propose to define the d -dimensional asymmetric H-niche containment similar to [5], as

$$I_2^{(d)} = \sqrt[d]{\prod_{i=1}^d I_{2,i}}$$

where $I_{2,i}$, $i = 1, \dots, d$, is the one-dimensional asymmetric H-niche containment of component i . The asymmetric H-niche containment I_1 is defined analogously.

It then follows that the d -dimensional symmetric H-niche containment is given by

$$O^{(d)}(F, G) = 4I_1^{(d)}I_2^{(d)} = 4\sqrt[d]{\prod_{i=1}^d I_{1,i}}\sqrt[d]{\prod_{i=1}^d I_{2,i}} = \sqrt[d]{\prod_{i=1}^d O(F, G)_i},$$

where $O(F, G)_i$, $i = 1, \dots, d$, is the one-dimensional symmetric H-niche containment of component i . Here, F and G are d -dimensional distribution functions. For the d -dimensional H-niche containment $O^{(d)}(F, G)$ the following properties hold.

Lemma 2.25

Suppose that F and G are distribution functions, not necessarily continuous. Then the following properties hold

1. In case of identical H -niches, $F = G$, and if F and G are continuous, we have $O^{(d)}(F, G) = 1$.
2. $O^{(d)}(F, G) = 0$ if and only if for at least one component $i \in \{1, \dots, d\}$ either $\mu_{G_i}((t_{F_i}, \bar{t}_{F_i})) = 0$ or $\mu_{F_i}((t_{G_i}, \bar{t}_{G_i})) = 0$.

Consistent estimators are given through the estimators in the one-dimensional case. Confidence intervals can be obtained by transforming the boundaries of the one-dimensional intervals, similar to the transformations of the actual estimators.

Remark 2.26

If the correlation matrices between the traits are rather dissimilar between the two samples, one may be able to construct examples, where the d -dimensional version of the H -niche will fail to describe the “true” relation between the two samples. For example, consider two bivariate samples where one has positively correlated traits, and the other has negatively correlated ones, while the marginal distributions are identical. In such a case, we would obtain an overlap close to one, while an intuitively “true” value should be closer to zero. This problem will require further research.

3 Simulation

In order to investigate the small sample properties of the estimator and its confidence intervals, we have performed simulation studies in R (R version 3.2.3, R Core Team, 2017). The first subsection considers asymmetric H -niche containment, while the second subsection focuses on the symmetric overlap.

3.1 Asymmetric H -Niche Containment

Table 1: Simulated ECP of the simplified rank-based confidence interval (upper row) and bootstrap confidence interval (lower row) for I_2 with different combinations of distributions F and G with equal medians and sample sizes 10, 20, ..., 90.

sample sizes	10	20	30	40	50	60	70	80	90
N(0,1)	0.924	0.954	0.945	0.936	0.940	0.946	0.955	0.948	0.942
x N(0,2)	0.931	0.951	0.961	0.941	0.951	0.939	0.964	0.944	0.935
N(0,1)	0.948	0.973	0.970	0.959	0.966	0.956	0.963	0.961	0.946
x N(0,5)	0.847	0.934	0.944	0.939	0.929	0.940	0.950	0.944	0.942
N(0,1)	0.880	0.926	0.933	0.942	0.943	0.943	0.947	0.936	0.940
x U(-0.5,0.5)	0.943	0.951	0.952	0.955	0.944	0.959	0.957	0.958	0.956
U(-0.5,0.5)	0.921	0.924	0.935	0.938	0.949	0.935	0.945	0.941	0.946
x N(0,1)	0.932	0.938	0.955	0.959	0.948	0.950	0.940	0.944	0.956
U(-0.5,0.5)	0.934	0.955	0.932	0.944	0.938	0.945	0.945	0.942	0.948
x U(-0.5,0.5)	0.904	0.948	0.936	0.945	0.953	0.936	0.940	0.949	0.935
B(2,3)	0.886	0.912	0.932	0.920	0.937	0.924	0.935	0.939	0.930
x B(3,2)-0.2	0.939	0.959	0.956	0.964	0.946	0.947	0.958	0.963	0.949
B(2,3)	0.866	0.910	0.897	0.924	0.926	0.930	0.913	0.914	0.934
x B(4,1)-0.4	0.936	0.951	0.937	0.939	0.942	0.947	0.961	0.952	0.955
N(0.615,1)	0.707	0.766	0.820	0.806	0.827	0.829	0.852	0.849	0.841
x B(3,2)	0.896	0.901	0.898	0.909	0.904	0.891	0.913	0.900	0.926
B(3,2)	0.957	0.964	0.966	0.952	0.954	0.938	0.972	0.956	0.951
x N(0.615,1)	0.869	0.911	0.926	0.933	0.947	0.937	0.943	0.935	0.942
U(0,1)	0.866	0.897	0.905	0.920	0.931	0.937	0.940	0.941	0.936

x B(3,2)- 0.113	0.924	0.931	0.944	0.949	0.944	0.945	0.950	0.948	0.949
B(3,2)- 0.113	0.925	0.929	0.935	0.939	0.938	0.939	0.957	0.951	0.957
x U(0,1)	0.930	0.949	0.950	0.960	0.946	0.950	0.936	0.946	0.947

The simulations are split into two parts, where one shows the appropriateness of the confidence intervals with respect to the considered sample size, and the other demonstrates when the simplified version of the confidence interval can or cannot be used.

To begin with, combinations of distributions F, G were analyzed, where both distributions had the same median, and each was normal, uniform, or Beta. Note that the simplified rank-based confidence interval formula requires equal medians. The sample sizes for each of the distributions were increased from 10 to 100 in steps of 10 and from 100 to 1,000 in steps of 100. For each combination of distributions 1,000 data sets were generated. The number of bootstrap runs per data set was set to 1,000.

Table 2: Simulated ECP of the simplified rank-based confidence interval (upper row) and bootstrap confidence interval (lower row) for I_2 with different combinations of distributions F and G with equal medians and sample sizes 100, 200, ..., 1000.

sample size	100	200	300	400	500	600	700	800	900	1000
N(0,1)	0.95	0.94	0.95	0.96	0.94	0.95	0.95	0.96	0.95	0.94
x N(0,2)	0.95	0.96	0.94	0.94	0.96	0.94	0.94	0.94	0.95	0.96
N(0,1)	0.96	0.95	0.96	0.95	0.94	0.95	0.94	0.96	0.95	0.93
x N(0,5)	0.94	0.94	0.95	0.93	0.95	0.95	0.94	0.95	0.94	0.95
N(0,1)	0.95	0.95	0.95	0.95	0.94	0.96	0.95	0.94	0.95	0.96
x U(- 0.5,0.5)	0.95	0.95	0.95	0.95	0.95	0.96	0.95	0.95	0.95	0.95
U(- 0.5,0.5)	0.94	0.95	0.96	0.95	0.95	0.94	0.95	0.93	0.95	0.95
x N(0,1)	0.95	0.95	0.95	0.96	0.95	0.94	0.95	0.96	0.94	0.94
U(- 0.5,0.5)	0.95	0.94	0.95	0.94	0.95	0.95	0.94	0.94	0.96	0.95
x U(- 0.5,0.5)	0.95	0.96	0.96	0.96	0.96	0.93	0.96	0.96	0.95	0.94
B(2,3)	0.94	0.94	0.95	0.94	0.95	0.95	0.94	0.94	0.95	0.95
x B(3,2)- 0.227	0.95	0.96	0.95	0.95	0.94	0.95	0.95	0.94	0.94	0.96
B(2,3)	0.94	0.93	0.93	0.92	0.92	0.94	0.94	0.92	0.93	0.94
x B(4,1)- 0.453	0.96	0.96	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.94
N(0.615,1)	0.88	0.90	0.90	0.93	0.93	0.92	0.94	0.93	0.93	0.95
x B(3,2)	0.93	0.93	0.93	0.94	0.94	0.95	0.95	0.95	0.95	0.93
B(3,2)	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.94
x N(0.615,1)	0.94	0.95	0.95	0.94	0.94	0.94	0.95	0.94	0.96	0.94
U(0,1)	0.93	0.94	0.94	0.96	0.95	0.94	0.95	0.94	0.95	0.95
x B(3,2) -0.113	0.95	0.94	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95
B(3,2) - 0.113	0.93	0.95	0.93	0.95	0.93	0.95	0.95	0.95	0.96	0.94
x U(0,1)	0.94	0.94	0.94	0.94	0.95	0.96	0.95	0.95	0.95	0.95

Simulated expected coverage probabilities (ECP) are shown in Table 1 and Table 2, and average confidence interval lengths in Table 3.

The true overlaps of I_2 (and I_1) were calculated by the following R-function, here given exemplarily for a Beta- and a Normal-distribution.

```
#N(a1, b1)
a1 = 0.615; b1 = 1
#B(a2, b2)
a2 = 3; b2 = 2
```

```

True_I_2 <- function(a1, b1, a2, b2){
  integ <- function(x){pbeta(x, a2, b2)*dnorm(x, a1, b1)}
  int_gtm = integrate(integ, qnorm(1/2, a1, b1), 1)
  int_stm = integrate(integ, 0, qnorm(1/2, a1, b1))
  overlap = 2 * (int_gtm$value - int_stm$value)
  return(overlap)
}

```

Both methods (bootstrap and simplified) generally maintained the nominal confidence level well for sample sizes of at least 40. For small sample sizes, it depended on the underlying distribution which of the methods yielded results that were closer to the nominal level. For combinations of symmetric distributions, both methods performed similarly.

Table 3: Average length of the simulated rank-based confidence interval (upper row) and bootstrap confidence interval (lower row) for I_2 with different combinations of distributions F and G with equal medians and sample sizes 10, 20, ..., 90.

sample size	10	20	30	40	50	60	70	80	90
N(0,1)	0.458	0.333	0.269	0.232	0.206	0.188	0.173	0.162	0.153
x N(0,2)	0.458	0.326	0.267	0.230	0.205	0.186	0.173	0.162	0.153
N(0,1)	0.315	0.246	0.205	0.177	0.158	0.144	0.131	0.123	0.115
x N(0,5)	0.283	0.219	0.184	0.158	0.141	0.129	0.121	0.114	0.106
N(0,1)	0.505	0.353	0.286	0.248	0.221	0.201	0.186	0.174	0.164
x U(-0.5,0.5)	0.613	0.407	0.319	0.271	0.238	0.214	0.197	0.183	0.172
U(-0.5,0.5)	0.497	0.354	0.287	0.247	0.221	0.202	0.186	0.175	0.164
x N(0,1)	0.505	0.355	0.291	0.251	0.223	0.203	0.188	0.175	0.165
U(-0.5,0.5)	0.426	0.319	0.258	0.223	0.199	0.181	0.168	0.156	0.147
x U(-0.5,0.5)	0.420	0.308	0.251	0.218	0.197	0.178	0.164	0.154	0.145
B(2,3)	0.511	0.363	0.294	0.254	0.226	0.206	0.191	0.179	0.168
x B(3,2)-0.2	0.566	0.392	0.312	0.267	0.238	0.216	0.199	0.185	0.174
B(2,3)	0.497	0.352	0.283	0.246	0.220	0.199	0.185	0.173	0.163
x B(4,1)-0.4	0.617	0.420	0.329	0.281	0.249	0.226	0.208	0.193	0.180
N(0.615,1)	0.393	0.277	0.218	0.184	0.162	0.147	0.134	0.126	0.117
x B(3,2)	0.730	0.493	0.391	0.326	0.281	0.251	0.226	0.203	0.189
B(3,2)	0.320	0.255	0.210	0.183	0.161	0.146	0.135	0.125	0.117
x N(0.615,1)	0.294	0.224	0.185	0.161	0.146	0.135	0.124	0.117	0.110
U(0,1)	0.494	0.340	0.277	0.239	0.213	0.194	0.180	0.168	0.158
x B(3,2)-0.113	0.629	0.414	0.326	0.275	0.242	0.216	0.197	0.183	0.172
B(3,2)-0.113	0.477	0.341	0.277	0.238	0.213	0.194	0.180	0.168	0.158
x U(0,1)	0.487	0.343	0.279	0.242	0.215	0.195	0.181	0.169	0.159

Figure 2: Simulated performance of estimation and confidence interval methods. True values, estimates, and average confidence interval boundaries (left), average confidence interval length (middle, blue - simplified calculation, red - bootstrap), and simulated expected coverage probability (right) for the two distributions $F = N(0.615, 1)$ and $G = B(3, 2)$, with respect to sample size.

When using two skewed distributions, such as the Beta-distribution, we noticed slight undercoverage of the simplified method, in particular for small to moderate sample sizes. The bootstrap generally maintained the nominal level well in these simulations. When F was skewed and G symmetric, both methods performed similarly, with small sample advantages of the simplified version, whereas for F symmetric and G skewed, the simplified method exhibited some undercoverage for small to moderate sample sizes.

Figure 3: Simulated performance of estimation and confidence interval methods. True values, estimates, and average confidence interval boundaries (left), average confidence interval length (middle, blue - simplified calculation, red - bootstrap), and simulated expected coverage probability (right) for the two distributions $F = B(3, 2)$ and $G = N(0.615, 1)$, with respect to sample size.

These similarities and differences are also illustrated in Figure 2 and Figure 3 for combinations of a skewed and a symmetric distribution. The plot of the boundaries in Figure 3 shows the similarities of the two confidence interval methods, with a slightly narrower bootstrap confidence interval, but at the expense of lower coverage than the rank-based interval in this situation, for sample sizes below 100.

Figure 4: Simulated performance of estimation and confidence interval methods. True values, estimates, and average confidence interval boundaries (left), average confidence interval length (middle, blue - simplified calculation, red - bootstrap), and simulated expected coverage probability (right) for the two distributions $F = N(0, 1)$ and $G = N(m, 1)$, with respect to increasing difference m between means.

As to the second part, we have already mentioned in Section 2.2 that the simplified variance estimator is only valid under the assumption of equal medians. The authors conjectured that it might still be used for small deviations from this assumption. In order to investigate the conditions under which it might become inappropriate, we conducted simulations using increasing location shifts between both distributions. All results are based on 1,000 simulation runs for each configuration, a sample size of 1,000, and, in the comparative analysis of the bootstrap interval performance, 1,000 bootstrap runs.

Figure 4, and Table 4 demonstrate the effect on ECP when the equal median assumption for the simplified confidence interval method is violated. Here, two normal distributions with different variances were used. The distribution function F was always standard normal, while G was a normal distribution with means $m = 0, 0.1, \dots, 4.9, 5$ and variances 1, 2, and 5. The simplified method exhibited increasing undercoverage, in particular for smaller variances ($\sigma^2 = 1$), as the difference in the true medians between both distributions increased. For a difference of 0.4 between the means, coverage was still about 90%, however, it already dropped to 70% for a difference of 1 between the means, dropping further as the mean differences increased. This is caused by the

variance estimator underestimating the true variance, leading to confidence intervals which are too short. For larger variances, the performance of the simplified version was better. In general, it would be a useful research task to devise a bias correction for the simplified rank-based variance estimator.

Figure 5: Plot of the confidence intervals with the true value and the estimator (left), average length of the confidence interval (middle, blue - simplified calculation, red - bootstrap), and simulated expected coverage probability (right) for the two distributions $F = U(-0.5, 0.5)$ and $G = U(m - 0.5, m + 0.5)$ w.r.t. the increase of difference m between the means.

While it seems that in distributions with infinite support and higher variance, the rank-based variance estimator can cope with rather small deviations from the equal median assumption, as seen in Table 4, the same does not seem to hold when working with distributions with finite support, as illustrated in Figure 5. Here, two uniform distribution were considered, each with a range of 1. In an almost linear manner the ECP of the simplified calculation dropped to zero up to $m = 0.5$, as for $m \geq 0.5$ the range of G can not possibly include the median of F anymore, and vice versa. Thus, the probabilities $P(X_1 < Y_1)$ and $P(X_2 < Y_2)$ are both equal to 1 in this scenario, and the resulting variance estimator is simply zero, rendering the simplified method not usable. However, up to $m = 0.2$, the simulated ECP was still higher than 80%. These results, together with the true values for I_2 , are also contained in Table 5. The simulated ECP dropped to zero for combinations of $U(-0.5, 0.5)$ and $U(m - 0.5, m + 0.5)$ when we reached $m = 0.5$. In comparison, the bootstrap confidence interval performed well until m reached 1, where the true overlap equals zero and thus also the variance estimator becomes zero. As in general one does not know whether the assumption of equal medians is valid or not, and one may only have the results of a hypothesis test for equal medians, it is important to note the influence that minor deviations from the assumptions have on the results.

Simulating data from a normal (with infinite support) and a uniform distribution (with finite support), we additionally observed the following. The simulated ECP depended on which of the two asymmetric H-niche estimators we were calculating. This has implications for estimating the symmetric overlap with confidence, as only one of the two individual confidence intervals needs to fail, in order to render their product estimator unreliable. On the other hand, the bootstrap confidence interval obtained roughly the same ECP no matter whether $\hat{I}_1$ or $\hat{I}_2$ was estimated, as can be seen in Table 6. Confidence intervals for the symmetric overlap were specifically assessed in additional simulations discussed below.

Table 4: True value of I_2 and simulated ECP of rank-based and bootstrap confidence intervals for normal distributions, with respect to increasing difference m between means.

m	$N(0,1)$ I_2	x rank	$N(m,1)$ boot	$N(0,1)$ I_2	x rank	$N(m,2)$ boot	$N(0,1)$ I_2	x rank	$N(m,5)$ boot
0.0	0.5000	0.95	0.96	0.2952	0.95	0.96	0.1257	0.95	0.96
0.2	0.4937	0.94	0.96	0.2939	0.94	0.96	0.1256	0.95	0.96
0.4	0.4752	0.90	0.96	0.2901	0.95	0.96	0.1253	0.96	0.94
0.6	0.4460	0.86	0.96	0.2839	0.94	0.95	0.1248	0.96	0.94
0.8	0.4082	0.79	0.96	0.2755	0.93	0.96	0.1241	0.97	0.95
1.0	0.3645	0.71	0.95	0.2650	0.92	0.93	0.1232	0.97	0.94
1.3	0.2939	0.59	0.96	0.2460	0.90	0.94	0.1216	0.97	0.95
1.7	0.2030	0.44	0.95	0.2162	0.87	0.94	0.1188	0.98	0.95
2.0	0.1449	0.33	0.95	0.1918	0.85	0.95	0.1162	0.99	0.96
2.3	0.0985	0.33	0.94	0.1670	0.82	0.95	0.1134	0.99	0.94
2.7	0.0547	0.20	0.94	0.1347	0.79	0.94	0.1090	0.99	0.96
3.0	0.0333	0.14	0.94	0.1121	0.76	0.94	0.1055	0.99	0.94
3.3	0.0194	0.09	0.93	0.0915	0.70	0.95	0.1016	0.99	0.95
3.7	0.0088	0.04	0.94	0.0678	0.66	0.95	0.0962	0.99	0.95
4.0	0.0047	0.02	0.91	0.0529	0.63	0.96	0.0920	0.99	0.94

4.3	0.0024	0.00	0.90	0.0405	0.60	0.94	0.0877	0.99	0.95
4.7	0.0009	0.00	0.88	0.0276	0.57	0.95	0.0817	0.99	0.95
5.0	0.0004	0.00	0.84	0.0202	0.51	0.94	0.0772	0.99	0.95

In situations where the overlap between both distributions was small, the bootstrap also showed undercoverage due to underestimation of the estimator's variance. This was the case for true values of I_1 or I_2 close to zero, roughly less than $2/\text{sample size}$. However, here the absolute difference between estimates and true values was still small, but the confidence intervals did not capture the true parameter due to their minuscule lengths. This is not surprising. Indeed, when the overlap between two distributions is close to zero, the bootstrap may yield two resamples which are fully separated in the sense that the largest observation in one sample is smaller than then smallest observation in the other. All those resamples would produce the same estimated zero for the overlap, with overall very little variation among the bootstrap samples, eventually yielding a variance estimator very close to zero.

Table 5: True value of I_2 and simulated ECP of rank-based and bootstrap confidence intervals for uniform distributions, with respect to increasing difference m between means.

U(-0.5,0.5)		x	U(m- 0.5,m+0.5)	U(- 0.5,0.5)	x	U(m- 1,m+1)
m	I_2	rank	boot	I_2	rank	boot
0.0	0.500	0.939	0.946	0.250	0.957	0.955
0.1	0.490	0.907	0.944	0.250	0.936	0.952
0.2	0.460	0.794	0.942	0.250	0.918	0.937
0.3	0.410	0.579	0.949	0.250	0.926	0.954
0.4	0.340	0.233	0.931	0.250	0.894	0.943
0.5	0.250	0.012	0.952	0.250	0.836	0.945
0.6	0.160	0.000	0.956	0.245	0.716	0.946
0.7	0.090	0.000	0.937	0.230	0.547	0.953
0.8	0.040	0.000	0.951	0.205	0.338	0.949
0.9	0.010	0.000	0.946	0.170	0.153	0.953
1.0	0.000	0.000	0.000	0.125	0.006	0.955
1.1	0.000	0.000	0.000	0.080	0.000	0.950

Table 6: True value of I_1 and I_2 and simulated ECP of rank-based and bootstrap confidence intervals for $F = U(-0.5, 0.5)$ and $G = N(m, 1)$, with respect to increasing difference m between means.

m	I_1	rank	boot	I_2	rank	boot
0.0	0.8046	0.94	0.95	0.1954	0.96	0.95
0.2	0.7765	0.84	0.94	0.1917	0.95	0.95
0.4	0.6930	0.65	0.95	0.1810	0.96	0.96
0.6	0.5646	0.58	0.95	0.1644	0.96	0.95
0.8	0.4425	0.58	0.94	0.1437	0.95	0.95
1.0	0.3370	0.57	0.96	0.1209	0.95	0.95
1.3	0.2119	0.59	0.94	0.0869	0.92	0.95
1.7	0.1025	0.56	0.95	0.0488	0.88	0.94
2.0	0.0536	0.54	0.95	0.0287	0.83	0.94
2.3	0.0270	0.53	0.92	0.0154	0.78	0.94
2.7	0.0094	0.47	0.92	0.0059	0.63	0.93
3.0	0.0039	0.36	0.86	0.0026	0.46	0.88
3.2	0.0021	0.24	0.79	0.0014	0.35	0.83
3.4	0.0010	0.13	0.66	0.0008	0.20	0.70
3.6	0.0005	0.03	0.54	0.0004	0.10	0.55
3.8	0.0003	0.01	0.33	0.0002	0.03	0.37
4.0	0.0001	0.00	0.20	0.0001	0.01	0.21

Concluding, we can say that our estimators and their confidence intervals appear to perform quite well. The simplified rank-based confidence interval should only be used when deviation from its equal median assumption is not strong. And, for highly skewed data, larger sample sizes than for symmetric data are needed to obtain fitting ECPs.

3.2 Symmetric overlap

In simulating the performance of the proposed confidence interval methods for the symmetric H-niche overlap, we used three combinations of distributions, namely $N(0, 1)$ and $N(0, 2)$, $N(0, 1)$ and $U(-1, 1)$, as well as $N(0.615, 1)$ and $B(3, 2)$. For these distributions and sample sizes 10, 20, ..., 150 we simulated 1,000 data sets each. For each data set, we calculated the estimators $\hat{I}_1$ and $\hat{I}_2$ along with their corresponding confidence intervals, and the resulting symmetric H-niche overlap estimator and its confidence intervals. The bootstrap confidence intervals are again based on 1,000 bootstrap runs each.

We used individual asymptotic 95% confidence intervals for I_1 and I_2 . As can be seen in Table 7, we obtained high ECPs even for sample sizes as small as 10. However the average length of the confidence intervals was too high as to actually provide any useful information about the true parameter value. Even for sample sizes of 150 each, we still had an average length of at least 1/3 for the three combinations. This would imply a need for larger sample sizes to actually obtain informatively precise confidence intervals, but such large samples may not always be available in practice.

Even though the length of the rank based confidence intervals is lower than for the bootstrap, without any significant loss of ECP, it still does not seem reasonable to use the confidence interval for small sample sizes as the contained information content is too low.

Table 7: Simulated ECP and average length of the confidence intervals for the symmetric H-niche overlap based on 95%-confidence intervals for I_1 and I_2

N(0,1) x N(0,2)	10	20	30	40	50	60	70	80	90	100	150
Rank ECP	0.93	0.98	0.97	0.98	0.99	0.99	0.99	0.99	0.99	1.00	0.99
Length	0.86	0.71	0.61	0.55	0.51	0.48	0.46	0.44	0.42	0.41	0.36
Boot. ECP	0.97	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Length	0.95	0.82	0.71	0.65	0.60	0.56	0.53	0.51	0.48	0.47	0.41
N(0,1) x U(-1,1)	10	20	30	40	50	60	70	80	90	100	150
Rank ECP	0.94	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Length	0.84	0.65	0.55	0.49	0.45	0.41	0.39	0.37	0.35	0.33	0.29
Boot. ECP	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Length	0.94	0.78	0.66	0.59	0.53	0.50	0.46	0.44	0.42	0.40	0.33
N(0.615,1) x B(3,2)	10	20	30	40	50	60	70	80	90	100	150
Rank ECP	0.94	0.97	0.96	0.96	0.95	0.96	0.97	0.96	0.96	0.96	0.97
Length	0.83	0.76	0.66	0.58	0.52	0.47	0.44	0.41	0.38	0.36	0.29
Boot. ECP	0.91	0.94	0.95	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98
Length	0.84	0.80	0.71	0.64	0.59	0.54	0.51	0.48	0.45	0.43	0.35

4 Application to ecological data

To visualize our results on a real data set we chose the data set *Finch2* from the R-package *dynRB* ([5]). This package implements the method developed by [5]. The data set contains nine morphological different traits for two species, *Geospiza fuliginosa pavula* and *Geospiza fortis fortis* that describe the similarity of both species. These two finch species are endemic to the Galápagos Islands. They are more commonly known as Darwin’s finches, along with some other species. There are 81 observations for *G. fuliginosa pavula* and 22 for *G. fortis fortis*. Out of the nine traits, we chose four, one of which only had a small overlap (tarsus length), two had a more pronounced overlap (body length and wing length), and one with complete containment of one species’ observed distribution within the other’s (tail length). Figure 6 depicts the trait distributions for both species.

Figure 6: Four of the nine traits of two species of finches, *Geospiza fuliginosa pavula* (red) and *Geospiza fortis fortis* (blue).

In this example, we analyzed “impact niches” (see Introduction). One of the main features of the ecological system of the Galápagos Islands is that it is solitary. Thus, analyzing traits of two species living in this ecosystem and having similar abiotic factors is of interest in ecology.

For each of the chosen traits we calculated both H-niche containment estimators $\hat{I}_1$ and $\hat{I}_2$, as well as their 95% confidence intervals. We used both, the simplified rank-based method, and the bootstrap, for the confidence interval calculations even though the equal median assumption does not appear to hold for any of the traits. However, as we saw in the simulations of the previous section, it may still yield reasonable results. Further we calculated the values corresponding closely to I_1 and I_2 given by [5] using the R-package *dynRB*. Table 8 shows the estimates and the confidence intervals. Here, in all cases, the rank-based confidence interval is well contained in the respective bootstrap confidence interval.

Table 8: Estimators of I_1 and I_2 for four traits of the finch data, along with their confidence intervals, and closely corresponding estimators given by the R-package *dynRB*.

	$\hat{I}_1$	Rank CI	Boot. CI	<i>dynRB</i>	$\hat{I}_2$	Rank CI	Boot. CI	<i>dynRB</i>
Body L.	0.17	(0.12, 0.22)	(0.09, 0.26)	0.131	0.18	(0.13, 0.23)	(0.05, 0.32)	0.092
Wing L.	0.04	(0.02, 0.05)	(0.00, 0.09)	0.016	0.04	(0.02, 0.05)	(0.01, 0.08)	0.013
Tail L.	0.47	(0.37, 0.57)	(0.32, 0.60)	0.487	0.38	(0.28, 0.49)	(0.25, 0.52)	0.352
Tarsus L.	0.05	(0.03, 0.07)	(0.01, 0.10)	0.020	0.06	(0.04, 0.08)	(0.02, 0.12)	0.018

While for three of the traits the two asymmetric H-niche containment estimators were almost the same, indicating a symmetric overlap of the data, for tail length, there was a difference of 0.09 between both directions. As shown in Figure 6, the observed support interval for *G. fortis fortis* completely covers the corresponding interval of *G. fuliginosa pavula*. And as I_1 is the probability that a randomly chosen observation from the species *G. fuliginosa pavula* lies between two randomly chosen observations of *G. fortis fortis*, it makes sense that it is clearly higher than I_2 , since the support of *G. fuliginosa pavula* does not cover the support of *G. fortis fortis*. Comparing our results with those of *dynRB*, we see similar relations and magnitudes, and would reach the same qualitative conclusions regarding the two species being compared. However, both methods do not provide the same numerical answers. Considering that both use different definitions of asymmetric and symmetric H-niche overlap, with different mathematical interpretations, this is not a surprise. The new approach features, for example, the advantages of a direct and intuitive probabilistic interpretation, and a confidence statement.

Table 9: One-dimensional H-niche overlap of four traits and the corresponding confidence intervals for the finch data.

	sym. NO	Rank CI	Boot. CI
Body L.	0.118	(0.059, 0.196)	(0.023, 0.309)

Wing L.	0.006	(0.002, 0.012)	(0.000, 0.031)
Tail L.	0.720	(0.411, 1.000)	(0.334, 1.000)
Tarsus L.	0.014	(0.005, 0.028)	(0.002, 0.044)

Figure 7: Scatterplot of four of the nine traits for two species of finches, *Geospiza fuliginosa parvula* (red) and *Geospiza fortis fortis* (blue).

Table 10: 2-dimensional symmetric H-niche overlap for pairwise traits.

	Wing L.	Tail L.	Tarsus L.
Body L.	0.027	0.291	0.041
Wing L.	-	0.067	0.009
Tail L.	-	-	0.101

In this data example, the 4-dimensional symmetric H-niche overlap is given by 5.22%, which seems reasonable as the one-dimensional symmetric overlap of two of the traits is rather low, as seen in Table 9. The confidence interval is (0.0233, 0.0918) using the simplified method, and (0.0187, 0.1575) using the bootstrap method. The 2-dimensional symmetric H-niche overlaps are given in Table 10. The resulting values appear reasonable in light of the scatterplots shown in Figure 7.

5 Conclusion

There has been much renewed interest in quantifying overlap between Hutchinsonian niches [1–6], driven in large part by ecologists who have been studying conditions for species to establish and survive in a given habitat, with applications, for example, in the study of native versus invasive species [23]. However, 60 years after the seminal work by [7], no global, unique definition of *H-niche overlap* has existed in a mathematical sense. Different authors have defined this concept differently, motivated by particular research questions (see [24–26]).

In addition, as of now, none of the established methods allows statistically valid inferential statements (confidence intervals) yet, even though it has become a requirement in many fields to supply confidence statements along with effect size estimates, whenever possible. The present manuscript attempts to close this gap. This is based on a definition of H-niche overlap that can be expressed in simple terms using intuitive probabilities, but also has connections to the ROC process and to rank statistics. These connections are advantageous in deriving asymptotic results and in providing shortcuts for the calculations. The proposed method is rather robust, and due to its invariance properties, it can be used for symmetric, skewed and heavy-tailed distributions alike. For

ordinal and binary variables, the authors conjecture that the method will also work, but a different mathematical proof needs to be provided. Nominal variables are outside of the scope of this approach, since it is not possible to assign ranks to nominal responses, while the proposed method utilizes ranks.

Specifically, two confidence interval methods are proposed. One is based on ranks, it can be calculated quickly using an explicit formula, but it requires an assumption on the distributions being compared. The other is based on the bootstrap, it can be used in general, based on the asymptotic theory shown, but it is obviously numerically more demanding. Simulations have demonstrated good performance of the proposed methods under their respective assumptions. Particularly for symmetric distributions, they work well also for small to moderate sample sizes. Thus, the proposed methods are also suitable for real data sets with low to moderate numbers of observations and thus well applicable for analyses of niches in ecological and other data sets.

Comparing the newly proposed method with previously proposed approaches, some similarities exist with the idea of [6] whose approach however was designed for data following multivariate normal distributions, while the presently proposed method does not make any distributional assumptions. Similar to our interpretation of Hutchinsonian niche overlap, the interpretation of their overlap factor is the probability that an individual from species A is found in the niche of B, which [6] defined to be the central 95%-region of the distribution of species B. They do allow for different choices of α . However, depending on the choice of α , it is then possible to obtain any value between 0 and 1 as an overlap factor, supplemented by heuristic confidence intervals based on Monte-Carlo simulations. Our new method, as well as [5], uses every possible α and combines them into an aggregated index, making the method also statistically robust.

The closest resemblance of the newly proposed method is indeed with the approach by [5], and the new method can be seen as a direct improvement of this approach, rendering a more robust estimation process, with simpler formulas, and supplementing the possibility to provide valid inferential conclusions, in addition to estimates for asymmetric and symmetric Hutchinsonian niche overlap.

Funding

This article was funded by the Austrian Science Fund (10.13039/501100002428), Grant Number: I 2697-N31.

Appendix

A Proof

Let X and Y be continuously distributed random variables. As in Section 2.1, assume for simplicity of notation that all the X -observations below or identical to their median $\hat{F}_n^{-1}(1/2)$ have indices $1, \dots, K$. Note that $K = n/2$ for n even and $K = (n+1)/2$ for n odd. In general, we will use the notations of Section 2.1 and 2.2.

First, we can rewrite the integral

$$\int_{-\infty}^{\hat{F}_n^{-1}(1/2)} \hat{G}_m d\hat{F}_n$$

as a sum

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^K \hat{G}_m(X_i) &= \frac{1}{n} \sum_{i=1}^K \frac{1}{m} \sum_{j=1}^m \mathbf{1}\{Y_j < X_i\} \\ &= \frac{1}{mn} \sum_{i=1}^K \left[\sum_{j=1}^m \mathbf{1}\{Y_j < X_i\} + \sum_{k=1}^n \mathbf{1}\{X_k < X_i\} \right] \\ &\quad - \frac{1}{mn} \sum_{i=1}^K \sum_{k=1}^n \mathbf{1}\{X_k < X_i\}. \end{aligned}$$

Obviously, the sum $\sum_{k=1}^n \mathbf{1}\{X_k < X_i\}$ is just the internal rank of X_i within the sample $X_1, \dots, X_n$, denoted by $R_{X,i}^{(X)}$, minus one half. Analogously, $\sum_{j=1}^m \mathbf{1}\{Y_j < X_i\} + \sum_{k=1}^n \mathbf{1}\{X_k < X_i\}$ is the rank of X_i in all observations, namely $R_{X,i}$.

Consequently, we obtain

$$\int_{-\infty}^{\hat{F}_n^{-1}(1/2)} \hat{G}_m d\hat{F}_n = \frac{1}{mn} \sum_{i=1}^K (R_{X,i}) - \frac{1}{mn} \sum_{i=1}^K (R_{X,i}^{(X)}) = \frac{1}{mn} \left(\sum_{i=1}^K R_{X,i} - \sum_{i=1}^K R_{X,i}^{(X)} \right).$$

The second sum over the ranks adds up to $\frac{K(K+1)}{2}$ if we do not have any ties in the data set. Therefore, we have

$$\int_{-\infty}^{\hat{F}_n^{-1}(1/2)} \hat{G}_m d\hat{F}_n = \frac{1}{mn} \left[R^{X<} - \frac{K(K+1)}{2} \right].$$

Similarly, we obtain

$$\begin{aligned} \int_{\hat{F}_n^{-1}(1/2)}^{\infty} \hat{G}_m d\hat{F}_n &= \frac{1}{mn} \sum_{i=K+1}^n R_{X,i} - \frac{1}{mn} \sum_{i=K+1}^n R_{X,i}^{(X)} \\ &= \frac{1}{mn} \left[R^{X>} - \frac{n(n+1)}{2} + \frac{K(K+1)}{2} \right]. \end{aligned}$$

Substituting these into eq. (6) yields

$$\begin{aligned} \hat{I}_2 &= 2 \left(\int_{\hat{F}_n^{-1}(1/2)}^{\infty} \hat{G}_m(t) d\hat{F}_n(t) - \int_{-\infty}^{\hat{F}_n^{-1}(1/2)} \hat{G}_m(t) d\hat{F}_n(t) \right) \\ &= \frac{2}{mn} (R^{X>} - R^{X<}) + \frac{2}{mn} \left(-\frac{n(n+1)}{2} + 2\frac{K(K+1)}{2} \right) \\ &=: \frac{2}{mn} (R^{X>} - R^{X<}) + \frac{1}{2} \cdot c, \end{aligned}$$

where c is given by $-n/m$ for n even and by $-(n+1)/m + 3(n+1)/(nm)$ for n odd. For large m and large odd n , c is approximately $-n/m$.

A.1 Proof A adapted for data sets with ties

For data sets with ties, the sum over the ranks of observations below the median does not necessarily add up to $K(K+1)/2$ when ties occur around the median. We therefore can not simplify the terms $\sum_{i=1}^K R_{X,i}^{(X)}$ in Proof A.

However, it holds as before that $\sum_{i=1}^n R_{X,i}^{(X)} = n(n+1)/2$. It follows that in case of ties, $\hat{I}_2$ can be written as

$$\frac{2}{mn} \left((R^{X>} - R^{X<}) - \frac{n(n+1)}{2} + 2 \sum_{i=1}^K R_{X,i}^{(X)} \right). \quad (8)$$

B Proof

In this proof, the notation, if not specified otherwise, is the same as in [27]. As $\hat{G}_m \circ \hat{F}_n^{-1} \rightsquigarrow G \circ F^{-1}$ and φ a continuous linear map at every point of $\mathbb{D}$, we can apply the continuous mapping theorem, Theorem 1.3.6 of [27]. It follows that

$$\sqrt{n}(\varphi(\hat{G}_m \circ \hat{F}_n^{-1}) - \varphi(G \circ F^{-1})) \rightsquigarrow \varphi'_{G \circ F^{-1}}(\mathbb{L}).$$

As the limit process $\mathbb{L}$ is a tight Borel measurable map and φ is a continuous linear map, it follows with Theorem 3.9.8 of [27] that $\varphi'_{G \circ F^{-1}}(\mathbb{L})$ is normally distributed.

We still need to show that the expectation of our normal distributed process $\varphi'_{GF^{-1}}(\mathbb{L})$ is zero. Let B_1 and B_2 be two independent Brownian bridges. Consider

$$\begin{aligned} E[\varphi'_{GF^{-1}}(\mathbb{L})] &= E \left[\int_0^1 \mathbb{L} \left(1 - \frac{\alpha}{2} \right) - \mathbb{L} \left(\frac{\alpha}{2} \right) d\alpha \right] \\ &= E \left[\int_0^1 \sqrt{\lambda} B_1 \left(GF^{-1} \left(1 - \frac{\alpha}{2} \right) \right) - \sqrt{\lambda} B_1 \left(GF^{-1} \left(\frac{\alpha}{2} \right) \right) d\alpha \right] \\ &\quad + E \left[\int_0^1 \frac{g(F^{-1}(1-\alpha/2))}{f(F^{-1}(1-\alpha/2))} B_2 \left(1 - \frac{\alpha}{2} \right) - \frac{g(F^{-1}(\alpha/2))}{f(F^{-1}(\alpha/2))} B_2 \left(\frac{\alpha}{2} \right) d\alpha \right]. \end{aligned}$$

To calculate the expectation of the integral, we will apply the Fubini-Tonelli Theorem. First we however show that the requirements are fulfilled. As the proof is similar for both summands, we will only show it explicitly for one of them. We can limit $\left| \sqrt{\lambda} B_1(GF^{-1}(1-\alpha/2)) - \sqrt{\lambda} B_1(GF^{-1}(\alpha/2)) \right|$ by $\sqrt{\lambda} \sup_{0 \leq t \leq 1} B_1(t) - \sqrt{\lambda} \inf_{0 \leq t \leq 1} B_1(t)$. Now this term has the same distribution as $\sqrt{\lambda} \sup_{0 \leq t \leq 1} e(t)$, where $e(t)$ is a Brownian excursion, see [28]. It therefore holds that

$$\begin{aligned} &\int_0^1 E \left[\left| \sqrt{\lambda} B_1(GF^{-1}(1-\alpha/2)) - \sqrt{\lambda} B_1(GF^{-1}(\alpha/2)) \right| \right] d\alpha \\ &\leq \int_0^1 \sqrt{\lambda} E \left[\sup_{0 \leq t \leq 1} e(t) \right] d\alpha = \sqrt{\lambda} E \left[\sup_{0 \leq t \leq 1} e(t) \right]. \end{aligned}$$

As given in Corollary 3.2 of [29], the expectation of a Brownian excursion is given by $\sqrt{\pi/2}$. As further $\lambda \in (0, 1)$ we can conclude that the term above is finite and that we can apply the Fubini-Tonelli Theorem. For the second summand, recall that the slope of $G(F^{-1}(t))$ is bounded. Switching expectation and integral, we obtain that the expectation of the process is given by

$$\begin{aligned} &\int_0^1 \sqrt{\lambda} E \left[B_1(GF^{-1}(1-\alpha/2)) \right] - \sqrt{\lambda} E \left[B_1(GF^{-1}(\alpha/2)) \right] d\alpha \\ &+ \int_0^1 \frac{g(F^{-1}(1-\alpha/2))}{f(F^{-1}(1-\alpha/2))} E \left[B_2(1-\alpha/2) \right] - \frac{g(F^{-1}(\alpha/2))}{f(F^{-1}(\alpha/2))} E \left[B_2(\alpha/2) \right] d\alpha. \end{aligned}$$

Since B_1 and B_2 are Brownian Bridges, they have expectation zero for all time points. It follows that $E[\varphi'_{GF^{-1}}(\mathbb{L})]$ is zero.

Alternatively, the result can also be concluded from the fact that each of the components of I_2 in eq. (5) can be estimated in an unbiased manner by a corresponding component of $\hat{I}_2$ in eq. (6).

C Proof

In order to apply Theorem 3.9.11 of [27], define $Y = (Y_1, \dots, Y_m)' \in \mathbb{R}^m$, $X = (X_1, \dots, X_n)' \in \mathbb{R}^n$, and $m = m(n)$. Further, let $Pf = G(F^{-1})$, $Y_j \sim G$ $j = 1, \dots, m$, and $X_i \sim F$, $i = 1, \dots, n$. To show that the requirements of Theorem of 3.9.11 of [27] are fulfilled, we define the class $\mathcal{F}$ as the set of functions

$$f_n(X, Y, t, i) = \hat{G}_m(X_i) \mathbf{1}\{\hat{F}_n(X_i) \leq t\} \prod_{j=1, j \neq i}^n [1 - \mathbf{1}\{\hat{F}_n(X_j) \leq t\} \mathbf{1}\{X_j < X_i\}]$$

with $t \in [0, 1]$, $\hat{G}_m(X_i) = \frac{1}{m} \sum_{j=1}^m \mathbf{1}\{Y_j \leq X_i\}$, $\hat{F}_n(X_i) = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{X_j \leq X_i\}$, and $i = 1, \dots, n$. Observe that this is simply the empirical counterpart to G evaluated at the t -quantile of the empirical counterpart to F . As shown in Example 2.5.4 of [27], the set of all indicator functions of cells in $\mathbb{R}$ is a Donsker class. From this we can directly conclude that the individual sets of functions $\mathbf{1}\{X_j < X_i\}$, $\hat{G}_m(X_i)$, $\hat{F}_n(X_i)$, and $\mathbf{1}\{\hat{F}_n(X_i) \leq t\}$ are Donsker classes. Theorem 2.10.8 of [27] states that the product of two Donsker classes is again a Donsker class. Further, it can easily be seen that the set of functions given by $1 - \mathbf{1}\{\hat{F}_n(X_j) \leq t\} \mathbf{1}\{X_j < X_i\}$ is also a Donsker class.

Altogether we can conclude that $\mathcal{F} = \{f_n(X, Y, t, i) : X \in \mathbb{R}^n, Y \in \mathbb{R}^m, t \in [0, 1], i = 1, \dots, n\}$ is a Donsker class and hence

$$\sqrt{n} (\mathbb{P}_n f_n - Pf)(t) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n f_n(X, Y, t, i) - G(F^{-1}(t)) \right) \rightsquigarrow \mathbb{G}$$

a tight random element, as seen in equation (2.1.1) of [27]. As we now have a Donsker class, we can apply Theorem 3.6.1 and obtain that the upper part of equation 3.9.9 of [27] holds in outer probability. Additionally, the lower equation holds as well, since the empirical process and the bootstrap are measurable. The Hadamard-differentiable function is given by φ defined in the beginning of Section 2.2. With this, we have shown that the requirements are indeed fulfilled, and we can apply Theorem 3.9.11 of [27], which completes the proof.

References

- [1] Blonder B. Pushing past boundaries for trait hypervolumes: a response to carmona et al. *Trends Ecol Evol.* 2016;31:665–667.
- [2] Blonder B, Lamanna C, Violle C, and Enquist BJ. The n-dimensional hypervolume. *Global Ecol Biogeogr.* 2014;23:595–609.
- [3] Carmona CP, de Bello F, Mason NW, Lepš J. The density awakens: a reply to blonder. *Trends Ecol Evol.* 2016;31:667–669.
- [4] Carmona CP, de Bello F, Mason NW, Lepš J. Traits without borders: integrating functional diversity across scales. *Trends Ecol Evol.* 2016;31:382–394.
- [5] Junker RR, Kuppler J, Bathke AC, Schreyer M, Trutschnig W. Dynamic range boxes a robust nonparametric approach to quantify size and overlap of n dimensional hypervolumes. *Meth Ecol Evol.* 2016;7:1503–1513.
- [6] Swanson HK, Lysy M, Power M, Stasko AD, Johnson JD, Reist JD. A new probabilistic method for quantifying n-dimensional ecological niches and niche overlap. *Ecology* 2015;96:318–324.
- [7] Hutchinson G. Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 1957.
- [8] Lamanna C, Blonder B, Violle C, Kraft N, Sandel B, Šímová I, Donoghue JC, Svenning JC, McGill BJ, Boyle B, Buzzard V. Functional trait space and the latitudinal diversity gradient. *Proc Nat Acad Sci USA.* 2014;111:13745–13750.
- [9] Li Y, Shipley B, Price JN, Dantas VD, Tamme R, Westoby M, Siefert A, Schamp BS, Spasojevic M, Jung V, Laughlin DC. Habitat filtering determines the functional niche occupancy of plant communities worldwide. *J Ecol.* 2017. doi: 10.1111/1365-2745.12802.
- [10] Kuppler J, Hoefers M, Trutschnig W, Bathke A, Eiben J, Daehler C, Junker R. Exotic flower visitors exploit large floral trait spaces resulting in asymmetric resource partitioning with native visitors. *Func Ecol.* 2017. doi: 10.1111/1365-2435.12932.
- [11] Barros C, Thuiller W, Georges D, Boulangeat I, and Münkemüller T. N-dimensional hypervolumes to study stability of complex ecosystems. *Ecol Lett.* 2016;19:729–742.
- [12] Lichenstein J, Wright C, McEwen B, Pinter-Wollman N, Pruitt J. The multidimensional behavioural hypervolumes to two interactions species predict their space use and survival. *Anim Behav.* 2017;132:129–136.
- [13] Nunes L, Pearson R. A null biogeographical test for assessing ecological niche evolution. *J Biogeogr.* 2017;44:1331–1343.
- [14] Blonder B. Hypervolume concepts in niche- and trait-based ecology. *Ecography* 2017;40:001–013.
- [15] Drake JM. Range bagging: a new method for ecological niche modelling from presence-only data. *J R Soc Interface.* 2015;12:20150086.
- [16] Isotalo J. . Basics of Statistics. CreateSpace Independent Publishing Platform, 2014.
- [17] ICH Expert Working Group. Statistical principles for clinical trials. ICH Harmonised Tripartite Guideline, 1998.
- [18] Schulz KF, Altman DG, Moher D. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Med.* 2010;8:18.
- [19] Vandembroucke JP, Von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (strobe): explanation and elaboration. *PLoS Med.* 2007;4:e297.
- [20] Hsieh FY, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann Stat.* 1996;24:25–40.
- [21] Brunner E, Munzel U. Nichtparametrische Datenanalyse. Berlin Heidelberg: Springer Spektrum, 2013.
- [22] Jean J, Shiryaev A. Limit theorems for stochastic processes, 2nd ed. Berlin Heidelberg New York: Springer-Verlag, 2003.
- [23] Konietschke F. Simultane konfidenzintervalle für nicht parametrische relative kontrasteffekte. Ph.D. thesis, University of Göttingen, 2009.
- [24] Junker RR, Daehler CC, Dötterl S, Keller A, Blüthgen N. Hawaiian ant-flower networks: nectar-thieving ants prefer undefended native over introduced plants with floral defenses. *Ecol Monogr.* 2011;81 (2): 295–311.
- [25] McInerny GJ, Etienne RS. Ditch the niche—is the niche a useful concept in ecology or species distribution modelling?. *J Biogeogr.* 2012;39:2096–2102.
- [26] McInerny GJ, Etienne RS. Pitch the niche—taking responsibility for the concepts we use in ecology and species distribution modelling. *J Biogeogr.* 2012;39:2112–2118.
- [27] McInerny GJ, Etienne RS. Stitch the niche—a practical philosophy and visual schematic for the niche concept. *J Biogeogr.* 2012;39:2103–2111.
- [28] van der Vaart W, Wellner JA. Weak convergence and empirical processes. New York: Springer Verlag, 1996.
- [29] Vervaat W. A relation between brownian bridge and brownian excursion. *Ann Probab.* 1979;7:143–149.
- [30] Durrett R, Iglehart D. Functionals of brownian meander and brownian excursion. *Ann Probab.* 1977;5:130–135.
- [31] Billingsley P. Convergence of probability measures. New York: John Wiley and Sons Inc, 1999.