Journal of Informetrics 7 (2013) 542–554

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Correlation between variables subject to an order restriction, with application to scientometric indices[☆]

Miguel A. García-Pérez[a,*], Vicente Núñez-Antón[b]

[a] Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain
[b] Departamento de Econometría y Estadística (E.A. III), Facultad de Ciencias Económicas y Empresariales, Universidad del País Vasco UPV/EHU, Avda. Lehendakari Aguirre 83, 48015 Bilbao, Spain

ARTICLE INFO

Article history:
Received 22 October 2012
Received in revised form 29 January 2013
Accepted 30 January 2013
Available online 24 March 2013

Keywords:
Correlation
Order restrictions
Sampling distribution
Significance testing
Simulation

ABSTRACT

Variables subject to an order restriction, for instance $Y \leq X$, have a bivariate distribution over a non-rectangular joint domain that entails a non-null and potentially large structural relation even if the variables show no association (in the sense that particular ranges of values of X do not co-occur with particular ranges of values of Y). Order restrictions affect a number of scientometric indices (including the h index and its variants) that are routinely subjected to correlational analyses to assess whether they provide redundant information, but these correlations are contaminated by the structural relation. This paper proposes an alternative definition of association between variables subject to an order restriction that eliminates their structural relation and reverts to the conventional definition when applied to variables that are not subject to order restrictions. This alternative definition is illustrated in a number of theoretical cases and it is also applied to empirical data involving scientometric indices subject to an order restriction. A test statistic is also derived which allows testing for the significance of an association between variables subject to an order restriction.

1. Introduction

Many variants of Hirsch's (2005) h index and many alternative indices of a researcher's impact have been proposed whose relative merits have been evaluated using correlational analyses. For instance, Bornmann, Mutz, Hug, & Daniel (2011) reviewed papers reporting correlations between the h index and 37 of its variants in search for evidence that might indicate whether these variants provide added information not carried by the h index itself. The founding assumption of their study was that "a high correlation between the h index and its variants would indicate that the h index variants hardly provide added information to the h index" (Bornmann et al., 2011, p. 346). Also in the same vein, Schreiber, Malesios, and Psarakis (2012) reported the results of an exploratory factor analysis involving the h index, 17 of its variants, and other bibliometric indicators. In both of these studies, correlations were used as indicators of the extent to which variables are associated and not as measures of the strength of a (presumed or established) linear relation between the variables.

[☆] This research was supported by grants PSI2009-08800 (Ministerio de Ciencia e Innovación, Spain), MTM2010-14913 (Ministerio de Ciencia e Innovación–FEDER, Spain), UFI11/03 and US12/09 (Universidad del País Vasco UPV/EHU), and IT-334-07 (UPV/EHU Econometrics Research Group, Departamento de Educación del Gobierno Vasco, Spain).

[*] Corresponding author at: Facultad de Psicología, Universidad Complutense, 28223 Madrid, Spain. Tel.: +34 913 943 061; fax: +34 913 943 189.
E-mail addresses: miguel@psi.ucm.es (M.A. García-Pérez), vicente.nunezanton@ehu.es (V. Núñez-Antón).

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.01.010

The use of correlation as an indicator of strength of association is widespread for assessing the extent to which several variables carry similar information, as is clear in the studies of Bornmann et al. (2011) or Schreiber et al. (2012). But some of the variables involved in these studies have characteristics that spuriously inflate correlations and can make variables that are stochastically independent appear to have some association. Specifically, the relation between the *h* index and some of its variants (denoted *h*' here, to use a single symbol to refer to any of these variants) is subject to an order restriction. For instance, the order restriction $0 \le h' \le h$ holds when h' is the second or any higher component of the multidimensional *h* index of García-Pérez (2009) and the order restriction $h \le h' < h + 1$ holds when h' is the fractional *h* index of Ruane and Tol (2008).

Note that the order restrictions just mentioned are *structural* (e.g., the value of *Y* cannot exceed the value of *X* for theoretical reasons) rather than simply *empirical* (e.g., when the value of *Y* does not happen to exceed the value of *X* in this particular sample). The methods presented in this paper apply to variables subject to *structural order restrictions* (or, for short, *order restrictions*), not to variables that just happen to show an *empirical order relation* in the sample of concern. Rosenberg (2011) reviewed variants of the *h* index (and alternatives to it) and stated the order restrictions that govern their relationships. Order restrictions are also found in many other fields and affect, for instance, variables that have a joint multinomial distribution such as successive scores in answer-until-correct procedures (García-Pérez, 1990; Parks & Yonelinas, 2009; Kellen & Klauer, 2011) or variables that relate to time such as age at diagnosis and years since diagnosis (Zebrack & Chesler, 2001), time since the death of a tree and time from death to fall (Storaunet & Rolstad, 2002), age and duration of smoking (Bain et al., 2004), or age and years since menopause (Rossouw et al., 2007). All variables subject to order restrictions have the additional characteristic that they have a lower bound (which also lies usually at zero); the methods presented in this paper apply to variables bounded low although the lower bound does not need to be zero.

Order restrictions limit the domain of the joint distribution of variables, which is no longer the Cartesian product of the domain of each variable. This characteristic introduces a *structural relation* (and a non-null correlation) that can be very high by conventional measures even when the variables involved show no association in the sense that a researcher actually wants to assess, namely, whether particular ranges of values of *X* co-occur with particular ranges of values of *Y* within its domain. Although this issue will be addressed in depth in subsequent sections of this paper, the consequences of using a conventional measure of correlation with variables subject to an order restriction should be noted. Consider Bornmann et al.'s (2011) study. They found (see their Fig. 1) that most of the *h*-index variants have correlations in excess of 0.8 with *h*, that a good number of them have correlations in excess of 0.9, and that only two variants have correlations with *h* that are sufficiently low so as to consider that they actually “make a non-redundant contribution to the *h* index” (Bornmann et al., 2011, p. 346). Interestingly, the strength of the correlations reported by Bornmann et al. goes hand in hand with the strength of the structural relations imposed by the underlying order restrictions (which are described by Rosenberg, 2011). These range from totally absent (when no order restriction exists and the product-moment correlation actually measures the strength of empirical association) through mild (when an order restriction enforces a non-rectangular but non-slanted joint domain, as will be shown in Fig. 1) to strong (when an order restriction enforces a non-rectangular and highly slanted joint domain, as will be shown in Fig. 3). In the latter two cases, the absence of data in regions of a putative rectangular domain is incorrectly quantified by the product-moment correlation as evidence of empirical association, not as a structural feature of the order restriction. In these cases, the association between the variables must be measured with consideration of the particular shape of their joint domain. Areas of the joint domain that are *structurally* deprived of data should not spuriously inflate measures of the empirical (*functional*) association between the variables, which is solely indicated by how data are distributed over the joint domain. The methods presented in this paper assess this functional association.

Assessment of the association between variables subject to an order restriction thus requires an alternative approach that separates true association from the uninteresting structural relation enforced by the order restriction. This paper presents a procedure that accomplishes this goal. A test statistic is also derived which allows checking for a significant association once the structural relation has been removed. The procedure is illustrated with empirical data.

2. Failure of conventional definitions and measures when order restrictions exist

Two random variables *X* and *Y* are defined to be stochastically independent if and only if their joint density f_{XY} equals the product of their unconditional (marginal) densities f_X and f_Y, that is, if $f_{XY}(x, y) = f_X(x)\,f_Y(y)$. A consequence of this definition is that the correlation between *X* and *Y* is null for stochastically independent variables whose joint distribution has domain $D_X \times D_Y$, where D_X and D_Y are the domains of *X* and *Y*. The correlation between *X* and *Y* is thus often used as an index of association. In some cases (e.g., the bivariate normal distribution), the correlation ρ is also one of the parameters of the distribution, but a specific parameter reflecting the correlation between variables is often not present and, hence, the correlation must be computed from the specific joint distribution of the variables. This is true of many bivariate distributions, and it is also true for the joint distribution of variables subject to order restrictions.

Order restrictions create singular problems for the assessment of stochastic independence and the interpretation of correlation coefficients. Structural relations arise in these cases because the domain of the joint distribution of such variables is a non-rectangular region of $D_X \times D_Y$: The joint density of *X* and *Y* cannot equal the product of the densities of *X* and *Y* and it is instead structurally null within some region of $D_X \times D_Y$. Then, the correlation between *X* and *Y* is mostly determined by this structural relation and it is sometimes only minimally affected by the peculiarities of the joint distribution of *X* and *Y* within the restricted domain, whose analysis would reveal whether or not the variables are associated. Variables subject to

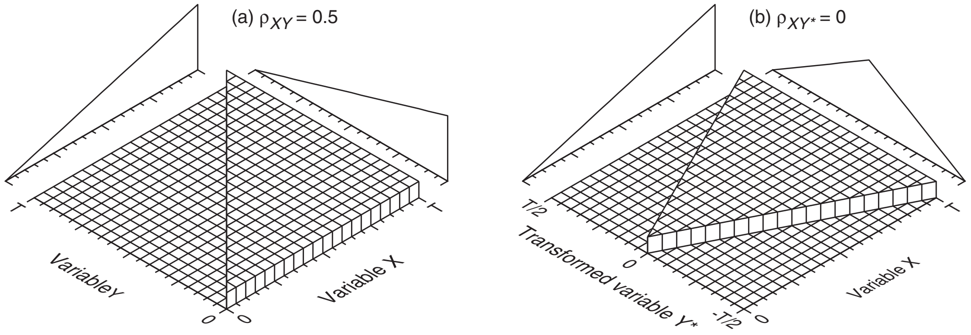

Fig. 1. (a) Variables subject to an order restriction and with a uniform joint distribution, yielding a structural correlation of 0.5 despite the lack of any real association (i.e., the distribution of Y is uniform at all values of X and only the range of Y varies with X as a consequence of the order restriction). Projection planes show the marginal distribution of each variable. (b) Transformation that reshapes the joint domain so that the structural correlation is removed and the computed correlation is 0.

an order restriction generally have a lower bound (e.g., all of the variables mentioned in the Introduction have a lower bound at 0) and some of them also have a finite upper bound, but the presence of finite or infinite upper bounds is inconsequential for our purposes here.

To illustrate the problem that order restrictions cause for assessments of stochastic independence, consider two variables that are subject to the order restriction $Y \leq X$. To simplify the presentation, and without loss of generality, also consider that X is bounded on $[0,T]$ so that the domain of the joint distribution of X and Y is the region of $[0,T] \times [0,T]$ below the diagonal (i.e., the joint density is structurally null above the diagonal). The lower bound could certainly differ from 0 without affecting our argument, and the upper bound T could be arbitrarily large and could even be removed also without affecting our argument. Consider also that X and Y have triangular marginal densities $f_X(x) = 2x/T^2$ and $f_Y(y) = 2(T - y)/T^2$, with $f_{Y|X}(y|x) = 1/x$ so that the joint probability density is $f_{XY}(x, y) = 2/T^2$ and, thus, it is uniform over the joint domain (see Fig. 1a). These variables do not have any association beyond that implied by their structural relation: Their joint distribution is uniform. Yet, $f_{XY}(x, y) \neq f_X(x)f_Y(y)$, implying a lack of stochastic independence by the conventional definition. Simple computations show that $\mu_X \equiv E(X) = 2T/3$, $\mu_Y \equiv E(Y) = T/3$, $\mu_{11} \equiv E(XY) = T^2/4$, $\sigma_X^2 \equiv var(X) = T^2/18$, $\sigma_Y^2 \equiv var(Y) = T^2/18$, and $\sigma_{XY} \equiv \text{cov}(X, Y) = T^2/36$. Thus, $\rho_{XY} \equiv \text{corr}(X, Y) = \sigma_{XY}/\sigma_X\ \sigma_Y = 0.5$, which is defined as the *structural correlation* under the order restriction, that is, the correlation that two variables displaying no association within the domain of their joint distribution will have as a result of this particular order restriction. It is also apparent from the triangular shape of the joint domain in Fig. 1a that correlation coefficients under alternative joint distributions for X and Y will hardly ever be negative or substantially smaller than 0.5. These characteristics cannot be used to interpret conventional correlation as a measure of association in these cases because the range of possible correlation coefficients varies with the type of order restriction (see Fig. 4).

A related problem is that conventional significance tests for the correlation coefficient cannot distinguish structural correlation from actual association. Thus, these tests will almost always reject the null hypothesis on the basis of a structural relation that is known beforehand and, thus, uninteresting. Alternative definitions and measures of association are thus needed. These definitions should also reduce to the conventional definitions for variables not subject to order restrictions, and they should also lend themselves to significance tests.

3. An alternative definition of stochastic independence under order restrictions

Let X and Y be random variables with distributions on $[0,\infty)$ and subject to a general order restriction of the type $g_{\text{inf}}(x) \leq Y \leq g_{\text{sup}}(x)$, for arbitrary functions g_{inf} and g_{sup} satisfying $g_{\text{inf}} \leq g_{\text{sup}}$. The simple order restriction in Fig. 1a implies $g_{\text{inf}}(x) = 0$ and $g_{\text{sup}}(x) = x$, and this expression can also accommodate the absence of order restrictions if g_{inf} and g_{sup} are both independent of X.

A definition of stochastic independence between variables subject to order restrictions must bypass the property $f_{XY}(x, y) = f_X(x)f_Y(y)$, which is never attainable for the reasons stated before. The underlying property $f_{Y|X}(y|x) = f_Y(y)$ of stochastically independent variables is also not attainable under order restrictions because the conditional distributions are bounded on $[g_{\text{inf}}(x), g_{\text{sup}}(x)]$ whereas the unconditional distribution is bounded on $[0,\infty)$. A final property of (unrestricted) independent variables that is convenient for use with variables subject to an order restriction is $E(Y|X) = E(Y)$. With unrestricted variables, this property implies that $E(Y|X)$ as a function of X describes a horizontal line, which is the reference line for assessments of stochastic independence with unrestricted variables. In these circumstances, positive (negative) association implies that $E(Y|X)$ tends to increase (decrease) with X, although not necessarily linearly.

Under an order restriction, the shape of the joint domain of X and Y is not rectangular. Yet, the preceding property can be adapted to set a suitable reference line (not necessarily straight) for the definition of stochastic independence with restricted variables and for distinguishing positive from negative association. Specifically, consider the line describing how

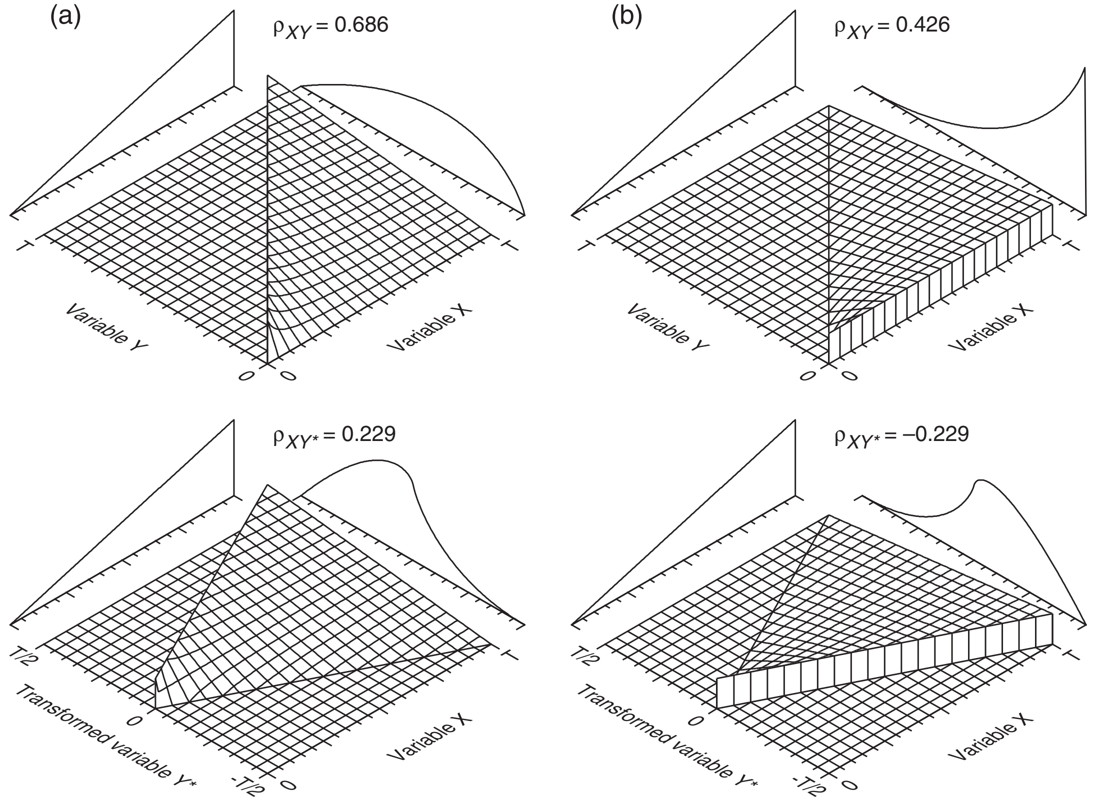

Fig. 2. Variables subject to an order restriction in the original domain (top row) and after the transformation that reshapes their joint domain (bottom row). Projection planes show the marginal distribution of each variable. (a) Positive association between the variables, whose magnitude is assessed in the reshaped domain. (b) Negative association between the variables, whose magnitude and sign is assessed in the reshaped domain.

the midpoint of the range of Y changes as a function of X under the order restriction. This line is $m(x) = (g_{\text{inf}}(x) + g_{\text{sup}}(x))/2$, which partitions the joint domain into upper and lower halves of equal area. Then, define X and Y to be independent if $E(Y|X)$ as a function of X varies as m does. For the case in Fig. 1a, where $m(x) = x/2$, X and Y are independent because $E(Y|X) = x/2$ also.

Using correlation as a measure of stochastic independence (association) between X and Y under this definition requires two steps. First, reshape the domain by defining $Y^* = Y - m(x)$. With this transformation, the joint domain of X and Y^* becomes $-(g_{\text{sup}}(x) - g_{\text{inf}}(x))/2 \leq Y^* \leq (g_{\text{sup}}(x) - g_{\text{inf}}(x))/2$ and the transformation thus makes this joint domain bilaterally symmetric about $Y^* = 0$ (see Fig. 1b). This transformation does not render a rectangular joint domain for X and Y^* but it achieves the sufficient goal of removing the structural relation by reshaping the joint domain to be bilaterally symmetric about $Y^* = 0$. An additional property of this transformation is that the reference line $Y = m(x)$ in the original domain becomes $Y^* = 0$ in the transformed domain, a horizontal line like that which holds for unrestricted variables. Once the joint domain is reshaped, the second step consists of measuring the association between X and Y as the product-moment correlation between X and Y^*. With empirical data, this implies applying the transformation to obtain data for Y^* from the original data in Y. For the case in Fig. 1, this yields $\mu_{Y^*} = 0$, $\mu_{11^*} \equiv E(XY^*) = 0$, $\sigma^2_{Y^*} = T^2/24$, $\sigma_{XY^*} = 0$, and $\rho_{XY^*} = 0$. Given the uniform joint density of X and Y in Fig. 1a, a null correlation seems more adequate as a measure of association between X and Y than the (structural) correlation of 0.5 in the original domain. And note that all of this reverts to the conventional approach in the absence of order restrictions: For variables bounded on $[0,T]$, with T arbitrarily large in case the variables are unbounded high, lack of an order restriction implies $g_{\text{inf}}(x) = 0$ and $g_{\text{sup}}(x) = T$ so that $Y^* = Y - T/2$ and $\rho_{XY^*} = \rho_{XY}$ because only an inconsequential linear transformation of Y is involved.

To illustrate, Fig. 2 shows two additional examples involving different true associations between variables with a common structural relation determined by the order restriction $Y \leq X$, with X also bounded on $[0,T]$. Hence, $m(x) = x/2$ again. In both cases, X has the marginal triangular density $f_X(x) = 2x/T^2$. In Fig. 2a, $f_{Y|X}(y|x) = 2y/x^2$ so that $f_{XY}(x, y) = 4y/xT^2$ and $f_Y(y) = -4y\ln(y/T)/T^2$. Thus, $E(Y|X) = 2x/3$, which increases with X faster than m does. The correlation between X and Y is $\rho_{XY} = \sqrt{8/17} \approx 0.686$, slightly in excess of the structural correlation of 0.5. Instead, $\rho_{XY^*} = 1/\sqrt{19} \approx 0.229$. In Fig. 2b, on the other hand, $f_{Y|X}(y|x) = 2(x - y)/x^2$ so that $f_{XY}(x, y) = 4(y - x)/xT^2$ and $f_Y(y) = 4(T - y + y\ln(y/T))/T^2$. Now $E(Y|X) = x/3$, which increases with X slower than m does. Here, $\rho_{XY} = \sqrt{2/11} \approx 0.426$, still positive and only slightly lower than the structural correlation of 0.5. Instead, $\rho_{XY^*} = -1/\sqrt{19} \approx -0.229$.

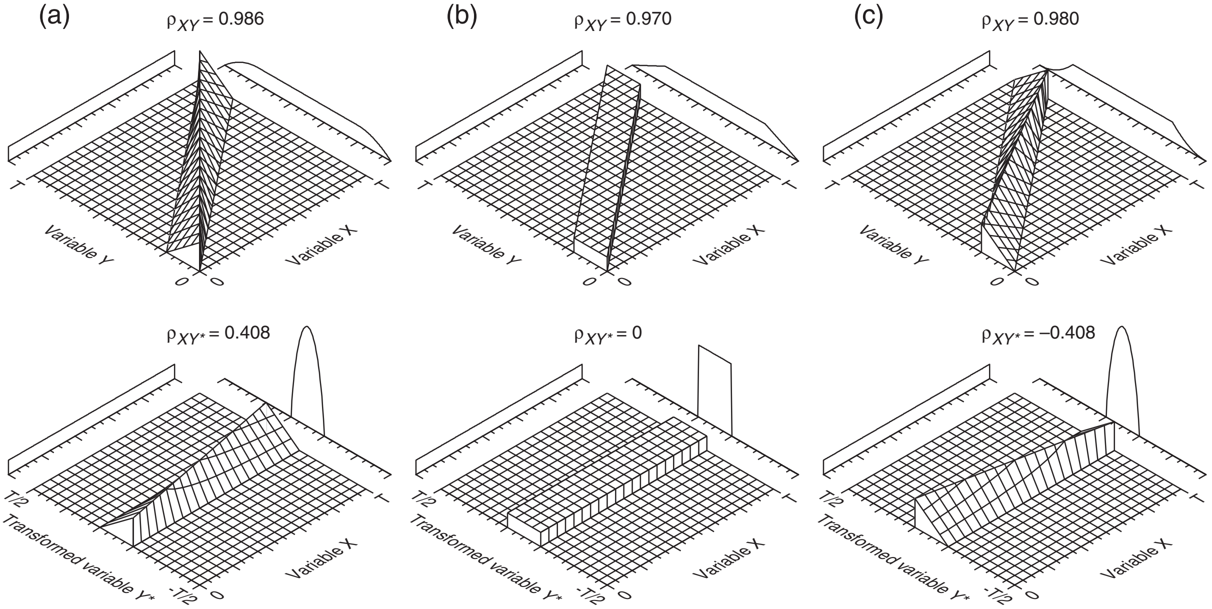

Fig. 3. Variables subject to an alternative order restriction in the original domain (top row) and in the reshaped domain (bottom row). Projection planes show the marginal distribution of each variable. The structural correlation is positive and high in the original domain and the isolated correlation between the variables can only be assessed in the reshaped domain. (a) Positive association, (b) null association and (c) negative association.

The sign of the association between X and Y estimated through the correlation between X and Y^* is perhaps better understood on consideration of the order restriction illustrated in Fig. 3, where $g_{\text{inf}}(x) = 4x/5$ and $g_{\text{sup}}(x) = (4x + T)/5$ so that $m(x) = 4x/5 + T/10$. Order restrictions such that the lower and upper bounds of Y both vary with X also affect some h index variants, as discussed in the Introduction. As seen in the top panels of Fig. 3, the shape of the joint domain makes any possible correlation between X and Y measured in the original domain essentially positive and very high. Yet, the form of the joint density of X and Y within this narrowly slanted domain may actually imply positive, null, or negative associations that are only apparent in the reshaped domain (bottom panels of Fig. 3). The marginal distribution of X is uniform and given by $f_X(x) = 1/T$ in the three sample cases of Fig. 3, which only differ as to the form of the conditional distribution of Y. In Fig. 3a, $f_{Y|X}(y|x)$ has a triangular distribution with $a = g_{\text{inf}}(x) = 4x/5$, $b = g_{\text{sup}}(x) = (4x + T)/5$, and $c = x$ (see Appendix A). Then, $E(Y|X) = 13x/15 + T/15$, implying a positive association because $E(Y|X)$ increases with X faster than m does. In Fig. 3b, $f_{Y|X}(y|x) = 5/T$, a uniform distribution for which $E(Y|X) = 4x/5 + T/10 = m(x)$, thus implying no association. Finally, in Fig. 3c, $f_{Y|X}(y|x)$ has a triangular distribution with $a = g_{\text{inf}}(x) = 4x/5$, $b = g_{\text{sup}}(x) = (4x + T)/5$, and $c = 3x/5 + T/5$. Then, $E(Y|X) = 11x/15 + 2T/15$ increases with X slower than m does and implies a negative association.

To further document the consequences of structural relations on conventional measures of association and how our procedure handles these cases, simulation studies were carried out to explore the range of values attainable by ρ_{XY^*} in comparison to those of ρ_{XY}. Samples of 500 paired observations in X and Y were drawn from each of 1000 different joint distributions all of which were subject to the order restriction $Y \leq X$. True association was again manipulated by altering the way in which the conditional distribution of Y varies with X. Without loss of generality, X was bounded on $[0,T]$ with $T = 20$. In one set of simulations, X had the triangular density $f_X(x) = 2x/T^2$; in another set, X had the uniform density $f_X(x) = 1/T$. In both sets of simulations Y had a conditional distribution on $[0,X]$ given by a scaled beta density with parameters v and w that varied randomly across replications such that v and w were independent from one another and uniformly distributed on $[0.5,10]$. Pseudo-random variates from these distributions were drawn using NAG subroutines (Numerical Algorithms Group, 1999). For each sample, correlation coefficients r_{xy} and r_{xy^*} were computed and a scatter plot of the resultant values is shown in Fig. 4a for the triangular distribution of X and in Fig. 4b for the uniform distribution of X. A somewhat tight but not deterministic relation can be observed between r_{xy} and r_{xy^*} and, interestingly, r_{xy^*} covers adequately the range $[-1,1]$ whereas r_{xy} is always positive and in excess of 0.2 (Fig. 4a) or 0.3 (Fig. 4b). Positive values for r_{xy} arise always from the order restriction $Y \leq X$, even though X and Y were indeed negatively associated in many cases.

In another simulation, order restrictions had the form in Fig. 3 with conditional distributions for Y given by two-sided power distributions (van Dorp and Kotz, 2002; see Appendix A) that yielded true associations that varied across replications from very strong and negative to very strong and positive. In this case (results not shown), r_{xy^*} also covered adequately the range $[-1,1]$ whereas r_{xy} was invariably greater than 0.965.

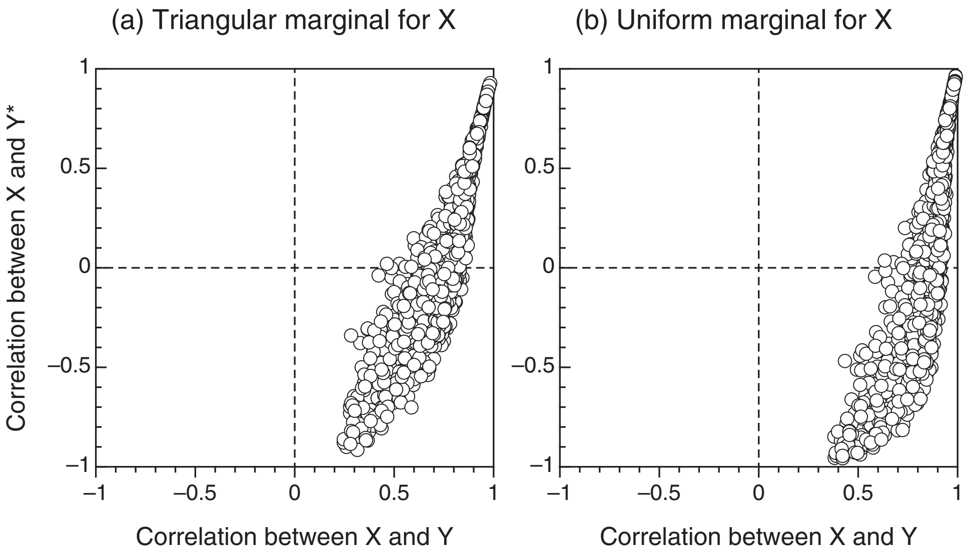

Fig. 4. Scatter plot of correlation in the reshaped domain against correlation in the original domain. Each data point comes from one of the 1000 samples in the simulation, each sample of size 500.

4. A significance test for the null hypothesis $H_0 : \rho_{XY^*} = 0$

Measuring the association between variables subject to an order restriction is only the first step toward establishing a significant association. A number of statistics are available for testing the null hypothesis $H_0 : \rho_{XY} = 0$ with unrestricted variables and these have been shown to be robust to violations of bivariate normality (e.g., Edgell & Noon, 1984; Hayes, 1996; Kowalski, 1972; Kraemer, 1980; Subrahmaniam & Gajjar, 1980; van den Brink, 1988; Zimmerman, Zumbo, & Williams, 2003). Yet, with variables subject to an order restriction and in the XY^* space defined above, the accuracy of these test statistics is uncertain because of the bounded and non-rectangular domain.

Simulations were thus conducted to investigate the validity of the conventional statistic

$$R = \frac{r_{xy^*}\sqrt{n-2}}{\sqrt{1-r_{xy^*}^2}} \tag{1}$$

used in the domain of X and Y^*. In these simulations, X was again bounded on [0,T] with T = 20. Two pairs of conditions were considered. In the first pair, the order restriction $Y \le X$ in Fig. 1 was used; in the second pair, the order restriction $4x/5 \le Y \le (4x+T)/5$ in Fig. 3 was used (see Fig. 5a). Thus, in the reshaped domain, X was always bounded on [0,T] whereas Y^* was bounded on [−X/2,X/2] in the first pair and on [−T/10,T/10] in the second pair (see Fig. 5b). Note that the argument and the validity of the simulations hold for arbitrarily large T (even if it goes to infinity). The two conditions within each pair differed only in that the marginal density of X was either a triangular distribution with a = 0, b = T, and c = T, or a triangular distribution with a = 0, b = T, and c = T/2 (see Fig. 5c). The conditional distribution of Y was always triangular with a = 0, b = x, and c = x/2 for the first pair of conditions and with a = 4x/5, b = (4x + T)/5, and c = (8x + T)/10 for the second pair (see Fig. 5d). In each condition, 50,000 samples of n paired observations were drawn from the applicable joint distribution of X and Y, with $n \in \{20, 40, 60, 80, 100, 120, 140, 160, 180, 200\}$. For each sample, data in Y were first transformed to render data in Y^*, then the statistic in Eq. (1) was computed, and the proportion of rejections across the 50,000 samples was determined for two-sided tests with $\alpha \in \{0.01, 0.05, 0.10\}$. Fig. 5e shows the distribution of the test statistic R across the 50,000 samples with n = 200 in each of the four simulation conditions, along with the asymptotic t distribution with n − 2 degrees of freedom. Fig. 5f shows the empirical size of the test as a function of sample size in each simulation condition.

Clearly, the conventional statistic in Eq. (1) is overly inaccurate to test $H_0 : \rho_{XY^*} = 0$ when the order restriction renders a non-rectangular joint domain in XY^* space (left half of Fig. 5). In these cases, the statistic can yield conservative or liberal tests according to the form of the joint distribution. In contrast, when the order restriction renders a rectangular joint domain in XY^* space (right half of Fig. 3), the conventional statistic is still accurate, providing further evidence of its robustness to bounded domains and non-normal distributions.

In search for a dependable test statistic, we used the delta method to obtain an asymptotically normal distribution for r_{xy^*} while making no assumptions other than $\rho_{XY^*} = 0$. The derivation is in Appendix B, which shows that the mean of the distribution of r_{xy^*} is zero and its variance is

$$\sigma_r^2 = \frac{\mu_X^2\sigma_{Y^*}^2 + \mu_{Y^*}^2\sigma_X^2 - 2\mu_X^2\mu_{Y^*}^2 + 6\mu_X\mu_{Y^*}\mu_{11} - 2\mu_X\mu_{12} - 2\mu_{Y^*}\mu_{21} + \mu_{22} - \mu_{11}^2}{n\sigma_X^2\sigma_{Y^*}^2}, \tag{2}$$

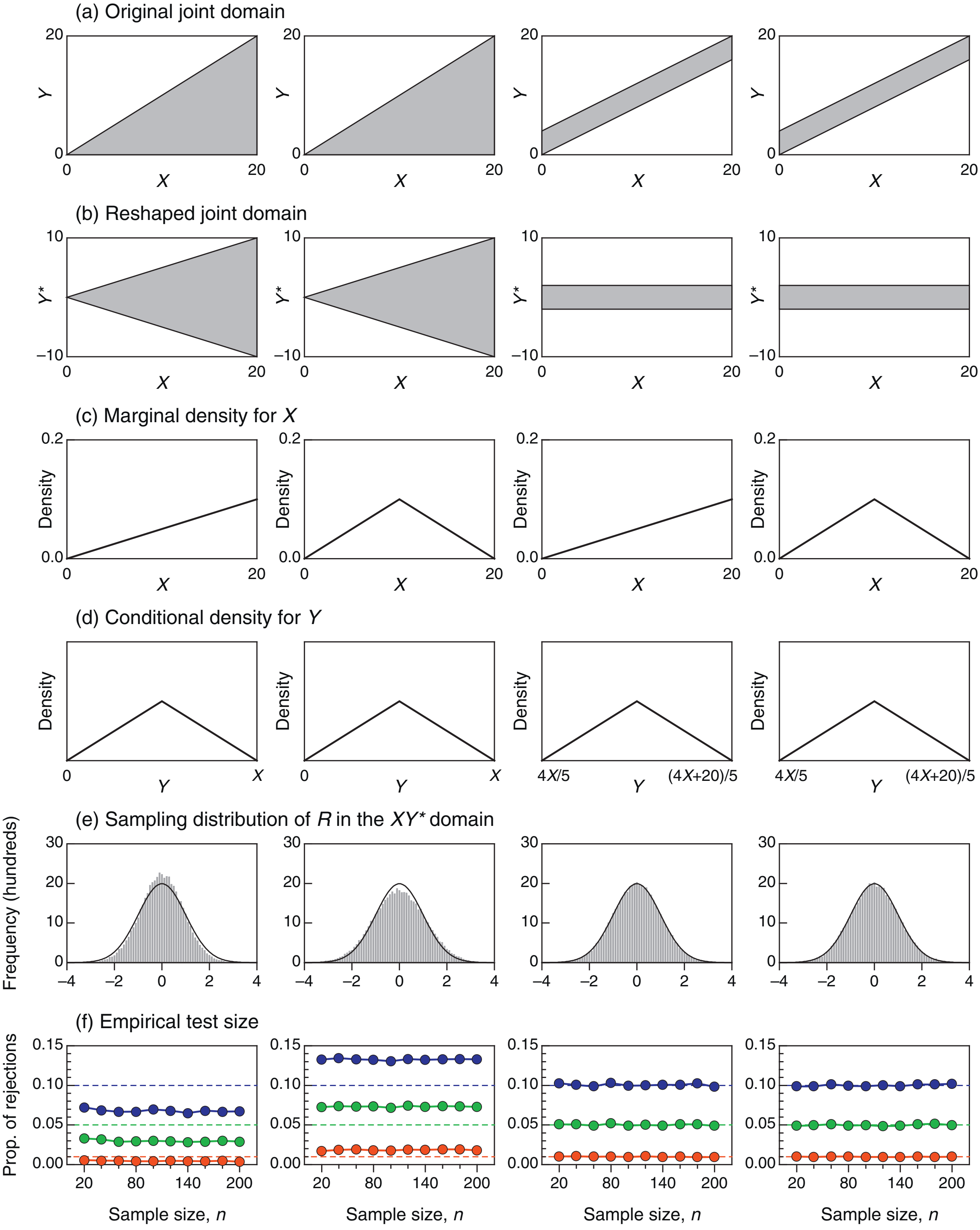

Fig. 5. Simulation results on the accuracy of the conventional test statistic in Eq. (1) under four conditions: two types of order restriction crossed with two forms of marginal distribution for X. (a) Order restriction in the joint XY domain. (b) Reshaped joint domain. (c) Marginal distribution of X. (d) Conditional distribution of Y in the original domain. (e) Empirical sampling distribution of the test statistic for $n = 200$ (histogram, based on 50,000 replications) and theoretical t distribution with $n - 2$ degrees of freedom (continuous curve). (f) Accuracy of the test as a function of sample size at nominal test sizes $\alpha \in \{0.01, 0.05, 0.10\}$ (red, green, and blue strands). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

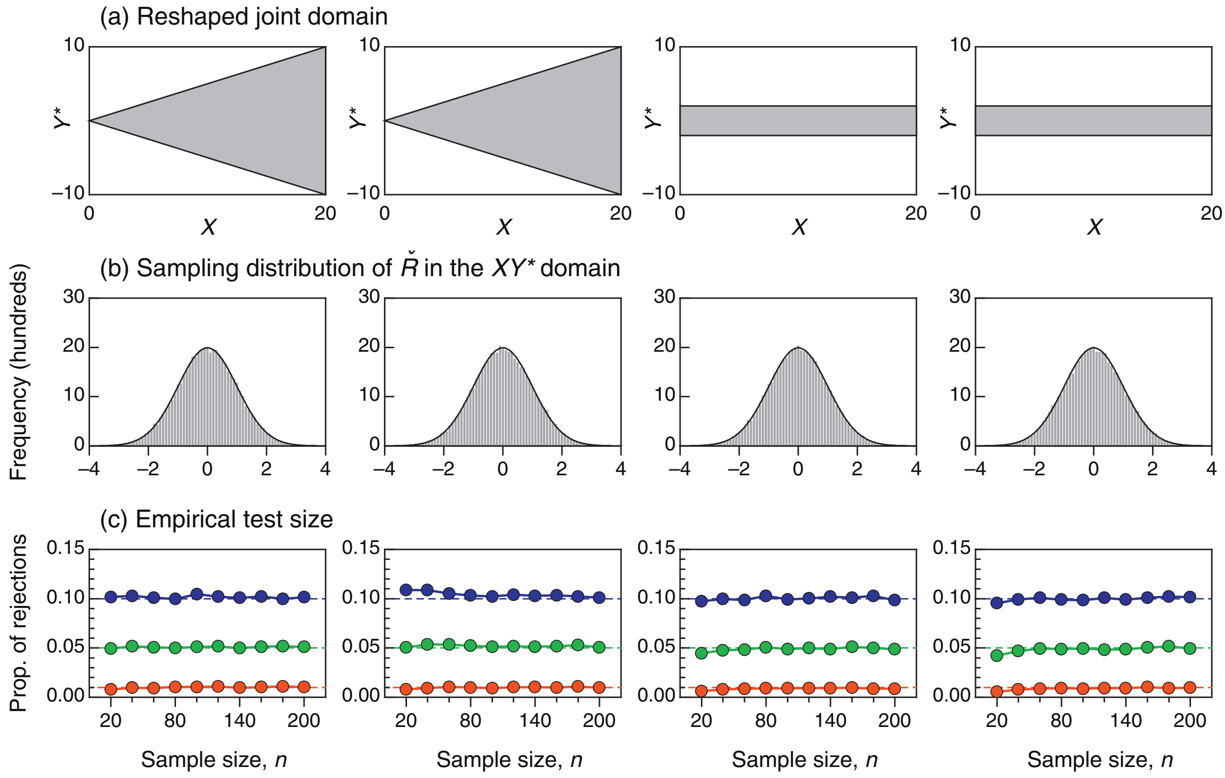

Fig. 6. Simulation results on the accuracy of the proposed test statistic in Eq. (5) under the same conditions as in Fig. 5. Only the reshaped domain is plotted in (a) for reference; the original domain, the marginal distributions of X, and the conditional distributions of Y remain as illustrated in Fig. 5. (b) Empirical sampling distribution of the test statistic for $n = 200$ (histogram, also based on 50,000 replications) and theoretical t distribution with $n - 2$ degrees of freedom (continuous curve). (c) Accuracy of the test as a function of sample size, using the same conventions as in Fig. 5. Note that the test statistic is accurate also when the joint domain in XY^* space does not have a rectangular shape.

where $\mu_{jk} = E(X^j Y^{*k})$. We then repeated the simulation but now using

$$\overset{\vee}{R} = \frac{r_{xy^*}}{\sigma_r} \tag{3}$$

as a test statistic, where σ_r was computed using the known distributions of X and Y^*.Use of $\overset{\vee}{R}$ in Eq. (3) yielded accurate tests regardless of the shape of the XY^* domain (results not shown). However, the form of the distributions of X and Y will be unknown in practical applications and, then, the information needed to compute σ_r will not be available. We thus considered estimating σ_r^2 from the data through

$$s_r^2 = \frac{\bar{X}^2 s_{y^*}^2 + \bar{Y}^{*2} s_x^2 - 2\bar{X}^2 \bar{Y}^{*2} + 6\bar{X}\bar{Y}^* \hat{\mu}_{11} - 2\bar{X}\hat{\mu}_{12} - 2\bar{Y}^* \hat{\mu}_{21} + \hat{\mu}_{22} - \hat{\mu}_{11}^2}{(n-2) s_x^2 s_{y^*}^2}, \tag{4}$$

where $\mu_{jk} = \frac{1}{n}\sum_{i=1}^{n} X_i^j Y_i^{*k}$, and then tested the null hypothesis with the modified statistic

$$\overset{\vee}{R} = \frac{r_{XY^*}}{s_r} \tag{5}$$

The factor $n - 2$ in the denominator of Eq. (4) is an approximation derived from extensive simulations which also showed that the test statistic in Eq. (5) seems to have a t distribution with $n - 2$ degrees of freedom. Fig. 6 shows results of application of this test statistic in a replication of the simulation study whose results were reported in Fig. 5. Clearly, the test statistic in Eq. (5) is accurate and insensitive to the shape of the joint domain in XY^* space.

5. Empirical illustration

Three empirical applications of the procedure just described are used to illustrate it in the context of assessing potentially redundant information provided by variants of the h index. As discussed in the Introduction, there is concern as to whether variants of the h index provide additional information not included in the h index itself, and correlational methods are typically used to investigate this issue (e.g., Bornmann et al., 2011; Schreiber et al., 2012). Because such variants, extensions, or alternatives are often subject to order restrictions with respect to the h index, these analyses are contaminated by structural relations.

Our first illustration assesses the strength of association between the conventional h index and a variant denoted h_1^+ by Ruane and Tol (2008), which represents a fractional increase over h. Thus, $h \leq h_1^+ < h + 1$, yielding a joint domain analogous to that illustrated in Fig. 3 with $g_{\text{inf}}(x) = x$ and $g_{\text{sup}}(x) = x + 1$ under the general form of the order restriction. Ruane and Tol (2008) reported a Spearman rank correlation of 0.92 between h and h_1^+, which is unsurprisingly high given the narrow and positively slanted domain of their joint distribution. Also, h is integer-valued by definition and, thus, a high and positive rank correlation is also expected because the fractional increase that h_1^+ entails cannot essentially alter the ranking: h_1^+ only potentially unties cases that are tied in h. The product-moment correlation between h and h_1^+ for data reported in Table 1 of Ruane and Tol (2008) is 0.987; with the procedure described here, the correlation is instead −0.388 ($\overset{\vee}{R} = -0.842$, $p = 0.427$, two-sided). Because the reshaped domain is rectangular under this order restriction, the conventional statistic in Eq. (1) would also have been appropriate, which yields an analogous result: $R = -1.115$, with a two-sided $p = 0.302$. Removal of the strong structural relation between h and h_1^+ thus reveals that these indices have a weak negative association for these data: The higher is h, the smaller is the fractional increase that h_1^+ brings. But the association is not significant and, hence, h_1^+ provides unique information.

In our second illustration, the strength of association was re-assessed between the three first components of the multidimensional h index, using data reported by García-Pérez (2009). The first component, h_1, is the conventional h index and the two other components, h_2 and h_3, are analogous h indices computed out the tail of the citation curve. These components are subject to the order restriction $0 \leq h_{i+1} \leq h_i$, yielding a joint domain analogous to that in Fig. 1 with $g_{\text{inf}}(x) = 0$ and $g_{\text{sup}}(x) = x$ under the general form of the order restriction. Raw data from 204 cases and tabulated scatter plots of the relations between the three components were given in Fig. 2 of García-Pérez (2009), for which product-moment correlations were reported to be 0.862 (h_1 and h_2), 0.862 (h_2 and h_3), and 0.780 (h_1 and h_3), which were deemed naturally high and positive given the order restriction. A re-analysis with the procedure described here renders correlations of −0.137 (h_1 and h_2; $\overset{\vee}{R} = -1.634$, $p = 0.104$, two-sided), 0.407 (h_2 and h_3; $\overset{\vee}{R} = 3.576$, $p < 0.001$, two-sided), and −0.543 (h_1 and h_3; $\overset{\vee}{R} = -3.168$, $p = 0.002$, two-sided). Thus, when the structural relation is eliminated, associations between components are smaller than initially claimed, or negative rather than positive. Of interest are the negative correlations between h_1 and h_2 and between h_1 and h_3. What this means is that as h_1 increases, h_2 and, particularly, h_3 increase at a slower rate than the order restriction allows. These negative correlations and the ensuing interpretations do more justice to the characteristics of the data (see García-Pérez, 2009, his Fig. 2) than the original but flawed positive correlations and the interpretation that h_2 and h_3 tend to be higher as h_1 increases (which is a tautology given the order restriction).

The third illustration additionally addresses the issue of how association should be measured when correlations need to be computed for a set of variables only some of which are affected by order restrictions, as in the studies of Bornmann et al. (2011) or Schreiber et al. (2012). The answer should be obvious: The association between variables not affected by order restrictions should be measured conventionally, whereas that between variables affected by order restrictions should be measured with the procedure described here (which may require the use of different functions g_{inf} and g_{sup} for each particular pair). Consider the two-sided extension of the h index developed by García-Pérez (2012), a multidimensional extension in which the scalar h (denoted h_0 in the two-sided index) is flanked by k positive-indexed components analogous to those that make up the multidimensional extension of García-Pérez (2009) and also by k negative-indexed components analogously computed up the top of the citation curve. Fig. 7 shows scatter plots of the components of the two-sided index ($k = 4$) computed for the sample of 80 researchers in the study of García-Pérez (2012). Regions structurally deprived of data due to order restrictions are grayed out in panels involving h_0 and in panels pairing components on the same side (both negative- or positive-indexed); no order restriction holds for components across sides (i.e., one negative- and the other positive-indexed). The latter type of panel in Fig. 7 displays the value of the conventional product-moment correlation in an inset and the pattern of these values reveals that association decreases with the distance between components (i.e., h_{-1} and h_1 are more strongly associated than h_{-4} and h_4); the former type of panel displays in an inset the value of the inadequate product-moment correlation (top) and the value of the corrected correlation (bottom) computed as described in this paper. Because the order restriction implies a positive structural relation, only the corrected measures of association can identify associations that are actually null, negative, or weakly positive.

As an additional illustration of the interpretation of the measure of association proposed in this paper, consider the scatter of data around the reference line in each panel of Fig. 7 and the pattern of correlations between h_0 and components on the positive side of the two-sided index (lower part of the fifth column in Fig. 7). Note that the inadequate conventional correlation is naturally always relatively high and positive, ranging from 0.787 (correlation between h_0 and h_1) to 0.636 (correlation between h_0 and h_4). Yet, within the restricted domain (white region in each panel), data points lie around the reference line of null association in the case of h_0 and h_1 (in agreement with a corrected correlation of 0.165) but they progressively fall further below the reference line in the panels pairing h_0 with h_2, h_3, and h_4, in agreement with corrected correlations that are negative and increasingly higher (i.e., as the X variable increases, the Y variable increases more slowly than the order restriction allows). Analogous considerations apply across the remaining panels of Fig. 7, but note that the role of the X and Y variables are reversed in the panels whose vertical axis is labeled from h_{-3} to h_0 (panels in the top left

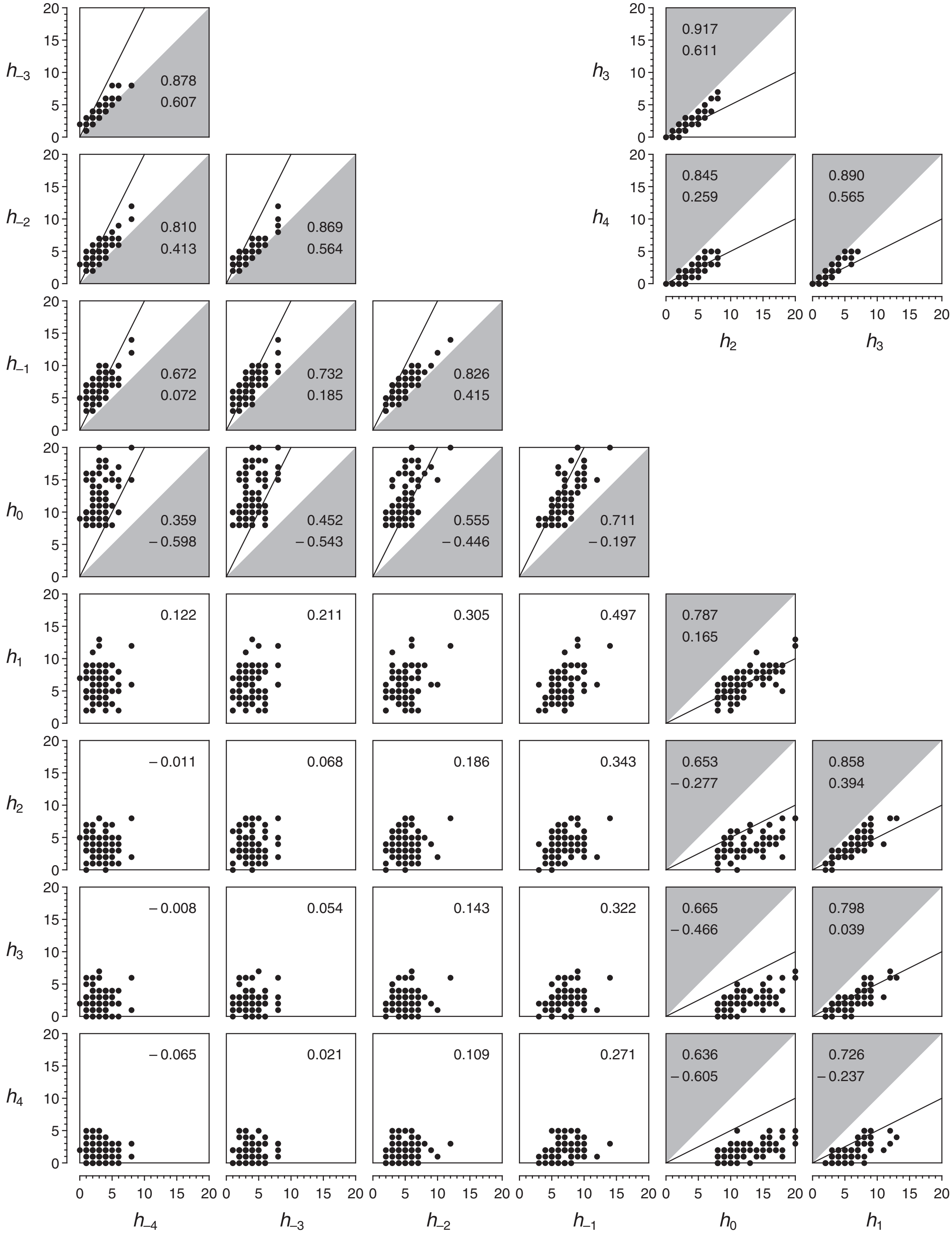

Fig. 7. Scatter plots showing empirical relations between all pairs of components of the two-sided h index of García-Pérez (2012), with component indices ranging from −4 to 4 (component indexed 0 is the conventional h) and using data (circles in each panel; $n = 80$) from that study. Grayed-out regions in some panels reflect areas structurally deprived of data due to order restrictions; in those cases, the reference line $m(x)$ is also displayed and note that the

part of Fig. 7). What this means is that the variable along the vertical axis is regarded as the X variable whereas that along the horizontal axis is regarded as the Y variable so that the order restriction satisfies the general form $g_{\text{inf}}(x) \leq Y \leq g_{\text{sup}}(x)$ implicit in our procedure.

6. Conclusion

Assessment of the association between variants of the h index (or other scientometric indices) is hampered because these variables are affected by order restrictions (see Rosenberg, 2011) that largely determine the value of conventional indices of association. The procedure described here solves this problem and essentially involves a symmetrization of the joint domain by transforming Y into Y^* so that the transformed joint domain is symmetrical about $Y^* = 0$, with the explicit definition that $\rho_{XY^*} = 0$ reflects stochastic independence (i.e., absence of association) between X and Y besides their structural relation. These definitions revert to the usual definitions of stochastic independence and correlation in cases that no order restriction exists. Variables subject to an order restriction typically have a lower bound on their domain, and the procedure described here is applicable whether the upper bound is finite or infinite because the Y variable will in both cases be bounded low and high. This procedure works for restrictions of the general form $g_{\text{inf}}(x) \leq Y \leq g_{\text{sup}}(x)$, involving arbitrary functions satisfying $g_{\text{inf}} \leq g_{\text{sup}}$. This requires that X and Y variables are adequately designated so that this form of order restriction holds and that the transformation is applied to the actual Y variable regardless of how the data may eventually be plotted, as illustrated in Fig. 7.

Examples have been given in which the true association between sample scientometric indices subject to order restrictions was uncovered through removal of their structural relation. In these cases, what appeared to be high and positive associations according to conventional analyses turned into negative, null, or positive associations when the positive structural relation was removed. Simulation results reported in Section 3 show that the conventional correlation r_{xy} is generally much higher than the corrected correlation r_{xy^*} that measures the true association between variables subject to the order restrictions that govern scientometric indices, all of which are of the type $g_{\text{inf}}(x) \leq Y \leq g_{\text{sup}}(x)$ (see Rosenberg, 2011). Although removal of a structural relation does not imply that the true association will not be still high and positive occasionally, correlation analyses will be more dependable if structural relations are removed before conclusions are raised about the redundancy of alternative scientometric indices. The conclusions of Bornmann et al. (2011) and Schreiber et al. (2012) regarding associations between scientometric indices should thus be understood as reflecting mostly structural relations for pairs of indices affected by an order restriction and reflecting true associations for pairs not affected by order restrictions. A more accurate picture of the actual redundancy of information provided by indices affected by order restrictions can only be obtained when structural relations are removed and true association is measured with the procedure described here.

Appendix A. Triangular and two-sided power distributions

The two-sided power (TSP) distribution of van Dorp and Kotz (2002) for a bounded variable on [a, b] is defined as

$$f(x) = \begin{cases} 0 & \text{if } x < 1; \\ \dfrac{n}{b-a}\left(\dfrac{x-a}{m-a}\right)^{n-1} & \text{if } a \leq x \leq m; \\ \dfrac{n}{b-a}\left(\dfrac{b-x}{b-m}\right)^{n-1} & \text{if } m < x \leq b; \\ 0 & \text{if } x > b. \end{cases} \tag{A1}$$

The conventional triangular distribution of a random variable bounded on [a, b] and given by

$$f(x) = \begin{cases} 0 & \text{if } x < 1 \\ \dfrac{2(x-a)}{(b-a)(c-a)} & \text{if } a \leq x \leq c \\ \dfrac{2(b-x)}{(b-a)(b-c)} & \text{if } c < x \leq b \\ 0 & \text{if } x > b \end{cases} \tag{A2}$$

is straightforwardly seen to be a TSP distribution with $n = 2$ and $m = c$.

roles of X and Y are reversed in panels whose vertical axis is labeled from h_{-3} to h_0 (panels at the top left). Insets show the magnitude of the conventional product-moment correlation (the only value displayed in cases not affected by order restrictions, or the top value in cases affected by them) and the corrected magnitude (bottom value in panels displaying two values).

Appendix B. Distribution of r_{xy} by application of the delta method

Let X and Y be random variables with an arbitrary joint distribution with moments $\mu_{jk} = E(X^j Y^k)$ and correlation ρ. Also let (X_i, Y_i) with $1 \le i \le n$ be a set of n paired observations in X and Y. Let

$$\begin{bmatrix} m_X \\ m_Y \\ m_{X^2} \\ m_{Y^2} \\ m_{XY} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} \sum X_i \\ \sum Y_i \\ \sum X_i^2 \\ \sum Y_i^2 \\ \sum X_i Y_i \end{bmatrix} \tag{B1}$$

so that $s_X^2 = m_{X^2} - m_X^2$, $s_Y^2 = m_{Y^2} - m_Y^2$, and $s_{XY} = m_{XY} - m_X m_Y$. By the central limit theorem,

$$\sqrt{n} \left(\begin{bmatrix} m_X \\ m_Y \\ m_{X^2} \\ m_{Y^2} \\ m_{XY} \end{bmatrix} - \begin{bmatrix} \mu_{10} \\ \mu_{01} \\ \mu_{20} \\ \mu_{02} \\ \mu_{11} \end{bmatrix} \right) \tag{B2}$$

has asymptotically a multivariate normal distribution with zero mean and covariance matrix

$$\Sigma = \begin{bmatrix} \operatorname{var}(X) & \operatorname{cov}(X, Y) & \operatorname{cov}\left(X, X^2\right) & \operatorname{cov}\left(X, Y^2\right) & \operatorname{cov}(X, XY) \\ \operatorname{cov}(Y, X) & \operatorname{var}(Y) & \operatorname{cov}\left(Y, X^2\right) & \operatorname{cov}\left(Y, Y^2\right) & \operatorname{cov}(Y, XY) \\ \operatorname{cov}\left(X^2, X\right) & \operatorname{cov}\left(X^2, Y\right) & \operatorname{var}\left(X^2\right) & \operatorname{cov}\left(X^2, Y^2\right) & \operatorname{cov}\left(X^2, XY\right) \\ \operatorname{cov}\left(Y^2, X\right) & \operatorname{cov}\left(Y^2, Y\right) & \operatorname{cov}\left(Y^2, X^2\right) & \operatorname{var}\left(Y^2\right) & \operatorname{cov}\left(Y^2, XY\right) \\ \operatorname{cov}(XY, X) & \operatorname{cov}(XY, Y) & \operatorname{cov}\left(XY, Y^2\right) & \operatorname{cov}\left(XY, Y^2\right) & \operatorname{var}(XY) \end{bmatrix}. \tag{B3}$$

To obtain the distribution of $(s_X^2, s_Y^2, s_{XY})^T$, define $g : \mathbb{R}^5 \to \mathbb{R}^3$ such that $g \begin{pmatrix} m_X \\ m_Y \\ m_{X^2} \\ m_{Y^2} \\ m_{XY} \end{pmatrix} = \begin{pmatrix} m_{X^2} - m_X^2 \\ m_{Y^2} - m_Y^2 \\ m_{XY} - m_X m_Y \end{pmatrix}$. The Jacobian of this transformation is $\dot{g} \begin{pmatrix} m_X \\ m_Y \\ m_{X^2} \\ m_{Y^2} \\ m_{XY} \end{pmatrix} = \begin{bmatrix} -2m_X & 0 & 1 & 0 & 0 \\ 0 & -2m_Y & 0 & 1 & 0 \\ -m_Y & -m_X & 0 & 0 & 1 \end{bmatrix}$. Then, $\sqrt{n} \left[\begin{pmatrix} s_X^2 \\ s_Y^2 \\ s_{XY} \end{pmatrix} - \begin{pmatrix} \sigma_X^2 \\ \sigma_Y^2 \\ \sigma_{XY} \end{pmatrix} \right]$ has asymptotically a multivariate normal distribution with zero mean and covariance matrix

$$\Sigma^* = \dot{g} \begin{pmatrix} \mu_{10} \\ \mu_{01} \\ \sigma_X^2 \\ \sigma_Y^2 \\ \sigma_{XY} \end{pmatrix} \Sigma \left[\dot{g} \begin{pmatrix} \mu_{10} \\ \mu_{01} \\ \sigma_X^2 \\ \sigma_Y^2 \\ \sigma_{XY} \end{pmatrix} \right]^T \begin{bmatrix} \operatorname{var}\left(X^2\right) & \operatorname{cov}\left(X^2, Y^2\right) & \operatorname{cov}\left(X^2, XY\right) \\ \operatorname{cov}\left(Y^2, X^2\right) & \operatorname{var}\left(Y^2\right) & \operatorname{cov}\left(Y^2, XY\right) \\ \operatorname{cov}\left(XY, X^2\right) & \operatorname{cov}\left(XY, Y^2\right) & \operatorname{var}(XY) \end{bmatrix}. \tag{B4}$$

Now let $h(a, b, c) = c/\sqrt{ab}$ so that $h(s_X^2, s_Y^2, s_{XY}) = r_{xy}$. The Jacobian of the transformation $h : \mathbb{R}^3 \to \mathbb{R}$ is $\dot{h}\begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{bmatrix} \dfrac{-c}{2\sqrt{a^3b}} \\ \dfrac{-c}{2\sqrt{ab^3}} \\ \dfrac{1}{\sqrt{ab}} \end{bmatrix}$

so that $A = \dot{h}\begin{pmatrix} \sigma_X^2 \\ \sigma_Y^2 \\ \sigma_{XY} \end{pmatrix} = \begin{bmatrix} \dfrac{-\rho}{2\sigma_X^2} \\ \dfrac{-\rho}{2\sigma_Y^2} \\ \dfrac{1}{\sigma_X\sigma_Y} \end{bmatrix}$. Then, $\sqrt{n}(r_{xy} - \rho)$ is asymptotically normally distributed with mean 0 and variance $\sigma^2 = A^T$ $\Sigma^* A$. Assuming $\rho = 0$, the variance of this distribution can easily be shown to be

$$\sigma^2 = \frac{\mu_X^2\sigma_Y^2 + \mu_Y^2\sigma_X^2 - 2\mu_X^2\mu_Y^2 + 6\mu_X\mu_Y\mu_{11} - 2\mu_X\mu_{12} - 2\mu_Y\mu_{21} + \mu_{22} - \mu_{11}^2}{\sigma_X^2\sigma_Y^2} \tag{B5}$$

so that r_{xy} is asymptotically normally distributed with zero mean and variance $\sigma_r^2 = \sigma^2/n$.

References

Bain, C., Feskanich, D., Speizer, F. E., Thun, M., Hertzmark, E., Rosner, B. A., et al. (2004). Lung cancer rates in men and women with comparable histories of smoking. *Journal of the National Cancer Institute*, *96*, 826–834.

Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H.-D. (2011). A multilevel meta-analysis of studies reporting correlations between the *h* index and 37 different *h* index variants. *Journal of Informetrics*, *5*, 346–359.

Edgell, S. E., & Noon, S. M. (1984). Effect of violation of normality on the *t* test of the correlation coefficient. *Psychological Bulletin*, *95*, 576–583.

García-Pérez, M. A. (1990). A comparison of two models of performance in objective tests: Finite states versus continuous distributions. *British Journal of Mathematical and Statistical Psychology*, *43*, 73–91.

García-Pérez, M. A. (2009). A multidimensional extension to Hirsch's *h*-index. *Scientometrics*, *81*, 779–785.

García-Pérez, M. A. (2012). An extension of the *h* index that covers the tail and the top of the citation curve and allows ranking researchers with similar *h*. *Journal of Informetrics*, *6*, 689–699.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing H_0: $\rho = 0$. *Psychological Methods*, *1*, 184–198.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 16569–16572.

Kellen, D., & Klauer, K. C. (2011). Evaluating models of recognition memory using first- and second-choice responses. *Journal of Mathematical Psychology*, *55*, 251–266.

Kowalski, C. J. (1972). On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Applied Statistics*, *21*, 1–12.

Kraemer, H. C. (1980). Robustness of the distribution theory of the product moment correlation coefficient. *Journal of Educational Statistics*, *5*, 115–128.

Numerical Algorithms Group. (1999). *NAG Fortran Library Manual, Mark 19*. Oxford: Author.

Parks, C. M., & Yonelinas, A. P. (2009). Evidence for a memory threshold in second-choice recognition memory responses. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 11515–11519.

Rosenberg, M. S. (2011). A biologist's guide to impact factors. Retrieved 27.12.12, from http://www.rosenberglab.net/Pubs/Rosenberg2011_ImpactFactor.pdf

Rossouw, J. E., Prentice, R. L., Manson, J. E., Wu, L., Barad, D., Barnabei, V. M., et al. (2007). Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause. *Journal of the American Medical Association*, *297*, 1465–1477.

Ruane, F., & Tol, R. S. J. (2008). Rational (successive) *h*-indices: An application to economics in the Republic or Ireland. *Scientometrics*, *75*, 395–405.

Schreiber, M., Malesios, C. C., & Psarakis, S. (2012). Exploratory factor analysis of the Hirsch index, 17 *h*-type variants, and some traditional bibliometric indicators. *Journal of Informetrics*, *6*, 347–358.

Storaunet, K. O., & Rolstad, J. (2002). Time since death and fall of Norway spruce logs in old-growth and selectively cut boreal forest. *Canadian Journal of Forest Research*, *32*, 1801–1812.

Subrahmaniam, K., & Gajjar, A. V. (1980). Robustness to non-normality of some transformations of the sample correlation-coefficient. *Journal of Multivariate Analysis*, *10*, 60–77.

van den Brink, W. P. (1988). The robustness of the *t* test of the correlation coefficient and the need for simulation studies. *British Journal of Mathematical and Statistical Psychology*, *41*, 251–256.

van Dorp, J. R., & Kotz, S. (2002). A novel extension of the triangular distribution and its parameter estimation. *The Statistician*, *51*, 63–79.

Zebrack, B. J., & Chesler, M. (2001). Health-related worries, self-image, and life outlooks of long-term survivors of childhood cancer. *Health and Social Work*, *26*, 245–256.

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicologica*, *24*, 133–158.