The Annals of Applied Statistics
2018, Vol. 12, No. 1, 123–155
https://doi.org/10.1214/17-AOAS1093

AN EMPIRICAL STUDY OF THE MAXIMAL AND TOTAL INFORMATION COEFFICIENTS AND LEADING MEASURES OF DEPENDENCE

BY DAVID N. RESHEF[*,‡,§,1], YAKIR A. RESHEF[†,‡,1,2], PARDIS C. SABETI[†,**,3] AND MICHAEL MITZENMACHER[†,**,4]

Massachusetts Institute of Technology[*] and Harvard University[†]

In exploratory data analysis, we are often interested in identifying promising pairwise associations for further analysis while filtering out weaker ones. This can be accomplished by computing a measure of dependence on all variable pairs and examining the highest-scoring pairs, provided the measure of dependence used assigns similar scores to equally noisy relationships of different types. This property, called *equitability* and previously formalized, can be used to assess measures of dependence along with the power of their corresponding independence tests and their runtime.

Here we present an empirical evaluation of the equitability, power against independence, and runtime of several leading measures of dependence. These include the two recently introduced and simultaneously computable statistics MIC_e, whose goal is equitability, and TIC_e, whose goal is power against independence.

Regarding equitability, our analysis finds that MIC_e is the most equitable method on functional relationships in most of the settings we considered. Regarding power against independence, we find that TIC_e and Heller and Gorfine's S^{DDP} share state-of-the-art performance, with several other methods achieving excellent power as well. Our analyses also show evidence for a trade-off between power against independence and equitability consistent with recent theoretical work. Our results suggest that a fast and useful strategy for achieving a combination of power against independence and equitability is to filter relationships by TIC_e and then to rank the remaining ones using MIC_e. We confirm our findings on a set of data collected by the World Health Organization.

Received December 2016; revised August 2017.

[‡] Co-first authors.

[§] To whom correspondence should be addressed. E-mail: dnreshef@mit.edu

[**] Co-last authors.

[1] Supported by the Paul and Daisy Soros Foundation.

[2] Supported by award Number T32GM007753 from the National Institute of General Medical Sciences, as well as the National Defense Science and Engineering Graduate Fellowship.

[3] Supported by the Howard Hughes Medical Institute.

[4] Supported in part by NSF Grants CNS-1228598, CCF-1320231, CCF-1535795 and CCF-1563710.

Key words and phrases. Maximal information coefficient, equitability, total information coefficient, statistical power, measures of dependence.

1. Introduction.

Suppose we have a data set with hundreds or thousands of dimensions and we wish to find interesting associations within it to assess further. Consider, for example, the collection of the 356 social, medical, economic, and political indicators measured by the World Health Organization (WHO) about every country in the world. How can we explore this data set to find important relationships, given that we may not be able to anticipate the models governing those relationships?

One natural approach to this problem would be to use regression-based models such as the LASSO [Tibshirani (1996)] or nonparametric regression frameworks [Breiman et al. (1984), Jaakkola and Haussler (1999)]. However, these strategies are limited to detecting relationships with a nontrivial regression function. This is an important limitation because many important relationships, such as relationships involving an unmeasured effect modifier, are not well described by a single function [Algeo and Lyons (2006), Caspi et al. (2003), Clayton and Mayeda (1996), Reshef et al. (2011)].

This shortcoming is partially addressed by *measures of dependence*: statistics whose population value is zero when the variables in question are statistically independent and nonzero in any other circumstance. Measures of dependence guarantee that we will asymptotically detect any deviations from independence in our data, regardless of the form of those relationships. A common, simple way of using a measure of dependence is to compute it for each pair of variables and then to manually examine all variable pairs for which a null hypothesis of independence can be rejected after accounting for multiple testing.

One way to measure the utility of the measure of dependence used in such a strategy is to assess the power of its associated independence test. This is an important goal if there are only a small number of true dependencies in the data, or if our sample size is small enough that only a small number of marginal dependencies can be uncovered. But some high-dimensional data sets contain a very large number of nontrivial relationships, some strong and others weak, and a list of all of them may be too large to allow for detailed model building or for manual follow-up of each identified relationship [Emilsson et al. (2008), Reshef et al. (2015)]. For example, in an analysis carried out in Heller et al. (2016) of a gene expression data set, five different measures of dependence each identified thousands of relationships–comprising over half of the possible relationships in the data set–as significant after multiple testing correction, and as we show in this paper, a similar phenomenon holds for the aforementioned WHO data set. Thus, we may want the statistic we use not only to detect as many of the nontrivial associations as possible, but also to give us a score that is interpretable in terms of relationship strength across a broad range of relevant relationship types. This would allow us to rank relationships by strength and consider a manageable number from the top of the list.

Equitability, introduced in Reshef et al. (2011) and formally defined in Reshef et al. (2015), is a property of measures of dependence that addresses this challenge.

While the formalization of equitability allows for arbitrary definitions of “relevant relationship types” and “relationship strength,” one natural instantiation of equitability is that, when used on functional relationships, the value of an equitable measure of dependence should reflect the coefficient of determination (R^2) with respect to the generating function with as weak a dependence as possible on the particular function in question.

Most measures of dependence do not have high equitability even on functional relationships. (This is understandable, as they are not designed with that goal in mind.) One measure of dependence that has been shown empirically to have good equitability on a broad set of functional relationships is the maximal information coefficient (MIC) [Reshef et al. (2011)]. In Reshef et al. (2016) a new, efficiently computable, consistent estimator of the population MIC, called MIC_e, is introduced, along with a related measure of dependence called the total information coefficient (TIC_e), which is essentially free to compute when MIC_e is computed.

In this paper we compare, under a wide range of settings, the equitability on functional relationships, power, and runtime of MIC_e, TIC_e, and a suite of leading measures of dependence. With regard to equitability, our results show that estimation of the population MIC via MIC_e is more equitable on functional relationships than other methods in a large majority of the settings of noise/marginal distributions and sample size that we tested, though in a few settings the Kraskov mutual information estimator outperforms MIC_e. With regard to power against independence, we find that TIC_e and a related method called S^{DDP} [Heller et al. (2016)] share state-of-the-art performance, and that many other methods including distance correlation [Szekely and Rizzo (2009)] also do quite well. We also characterize a more general power-equitability trade-off that holds across methods, and we present a runtime analysis to characterize the scale of data that each method can analyze. Our full set of simulation analyses of power, equitability, and runtime, including sensitivity analyses and additional sample sizes and models, are available in an online empirical supplement [Reshef et al. (2018b)] that we hope will be a resource to the community.

We close by applying all the methods examined to the WHO data set described above. Our analysis of real data validates the results of our power simulations, reveals empirical relationships among the methods we benchmarked that are consistent with our equitability simulations, and shows that MIC_e and TIC_e detect new relationships of scientific interest that would not be easily found using the other methods we consider here. Taken together, our results suggest that MIC_e can be efficiently used in conjunction with TIC_e to achieve a useful mix of power against independence (by filtering results using TIC_e) and equitability (by using MIC_e on the remaining variable pairs) when exploring a data set with a large number of nontrivial relationships.

2. A review of equitability.

Equitability is a property of measures of dependence introduced in Reshef et al. (2011) and formalized in Reshef et al. (2015).

Because this paper analyzes in depth the equitability of several leading measures of dependence, we present here a brief summary of some basic definitions and results.

There are two equivalent ways to view equitability [Reshef et al. (2015)]. The first states roughly that an equitable measure of dependence gives similar scores to equally noisy relationships of different types [Reshef et al. (2011)]. In this viewpoint, a highly equitable measure of dependence allows us to find, in a sense, the strongest K relationships in our data set for any K. The second view is based on statistical power: an equitable measure of dependence provides good tests for distinguishing between relationships with different, potentially nonzero amounts of noise. That is, a highly equitable measure of dependence allows us in principle to find with high power relationships in our data set with strength at least x_0 for any x_0. (An ordinary measure of dependence, in contrast, only provides such a guarantee for the case $x_0 = 0$.) Equitability includes the usual null hypothesis of statistical independence as a special case.

We present here a formal definition of the second viewpoint followed by an informal description of the first. Let $\mathcal{Q}$ denote a set of distributions, called the *standard relationships*, on which we can state what we mean by relationship strength, and let $\Phi : \mathcal{Q} \to [0, 1]$ be the functional that computes that strength. For example, $\mathcal{Q}$ could be some diverse set of functional relationships with noise added and Φ could be R^2, that is, the coefficient of determination with respect to the generating function. Equitability can then be defined via statistical power as follows.

DEFINITION 2.1 [Reshef et al. (2015)]. Let $\hat{\varphi}$ be a statistic, let $\mathcal{Q}$ be a set of standard relationships, let $\Phi : \mathcal{Q} \to [0, 1]$, and let $0 < \alpha < 1 - \beta < 1$. The statistic $\hat{\varphi}$ is *$1/d$-equitable* with respect to Φ at level α and power $1 - \beta$ if and only if for every $x_0, x_1 \in [0, 1]$ satisfying $x_1 - x_0 > d$, there exists a right-tailed level-α test based on $\hat{\varphi}$ that can distinguish between $H_0 : \Phi(\mathcal{Z}) = x_0$ and $H_1 : \Phi(\mathcal{Z}) = x_1$ with power at least $1 - \beta$.

The smaller d is the better. The best equitability that can be achieved is when $d = 0$; this is called *perfect equitability*, and is typically discussed as a property of the population value of a statistic. In this paper we set $\alpha = \beta = 0.05$ always.

Definition 2.1 is illustrated schematically in Figure 1. When Φ is 0 precisely in cases of statistical independence, equitability can be viewed as a generalization of power against statistical independence on $\mathcal{Q}$. Specifically, when we set $x_0 = 0$, a $1/d$-equitable statistic yields a test that has good power against independence on any alternative hypothesis as extreme or more extreme than $H_1 : \Phi = d$. More generally, the definition says that a $1/d$-equitable statistic allows us, given some threshold x_0 of relationship strength as measured by Φ, to identify with high power the relationships in a data set with strength greater than $x_0 + d$. This is important if our data set has many weak relationships and a smaller number of stronger relationships that we would like to find.

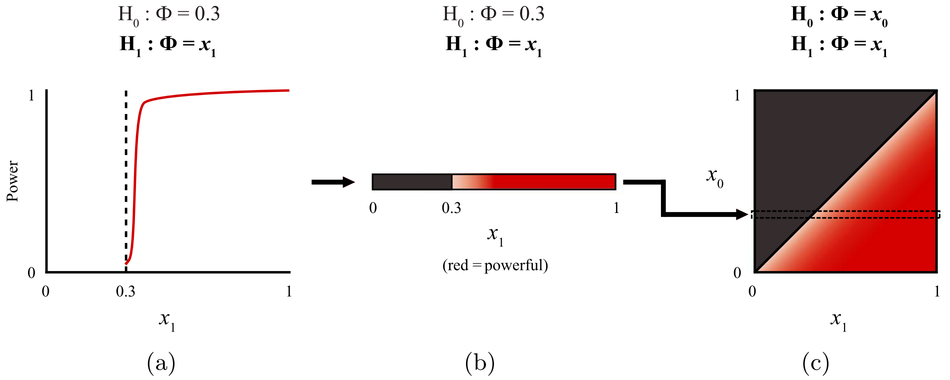

FIG. 1. *Equitability as a generalization of power against independence.* (a) *The power function of a size-α right-tailed test for fixed α, based on a statistic $\hat{\varphi}$, of the null hypothesis $H_0 : \Phi = 0.3$. The curve shows the power of the test as a function of x_1, the value of Φ under the alternative hypothesis.* (b) *The power function can be depicted instead as a heat map.* (c) *Instead of considering just one null hypothesis/critical value, we can consider a set of null hypotheses (with corresponding critical values) of the form $H_0 : \Phi = x_0$ and plot each of the resulting power curves as a heat map. The result is a plot in which the intensity of the color in the coordinate (x_1, x_0) corresponds to the power of a size-α right-tailed test based on $\hat{\varphi}$ at distinguishing $H_1 : \Phi = x_1$ from $H_0 : \Phi = x_0$. A $1/d$-equitable statistic at level α and power $1 - \beta$ is one for which this power surface attains the value $1 - \beta$ within distance d of the diagonal along each row.*

Thus, an analysis of equitability must assess power against many null hypotheses, and because $\mathcal{Q}$ can contain many relationship types (e.g., linear, exponential, etc.), the null and alternative hypotheses analyzed are composite. Of course, though we can formulate the null hypothesis $H_0 : \Phi = x_0$ conceptually and test it in simulations, testing it on real data is challenging because the sampling distribution of the statistic under this null may depend on relationship type. The value of equitability does not stem from the ability to test this null hypothesis, but rather from the fact that ranking relationships based on a statistic that performs well with respect to this null hypothesis in controlled situations is a better idea than ranking them based on a statistic that does not.

The other, equivalent formalization of equitability, measures the degree to which a statistic assigns similar scores to equally noisy relationships of different types. This is done via an object called the *interpretable interval*, which is an interval estimate of Φ constructed from $\hat{\varphi}$. Specifically, the interpretable interval is guaranteed to cover the true value of Φ with high probability for any relationship $\mathcal{Z} \in \mathcal{Q}$. This is illustrated in Figure 2. For more detail, see Appendix A [Reshef et al. (2018a)]. The width of the widest interpretable interval obtained from any value of $\hat{\varphi}$ can be shown to equal the parameter d in Definition 2.1 above. For a formal statement of the equivalence of the two formalizations of equitability proven in Reshef et al. (2015), see Appendix A.

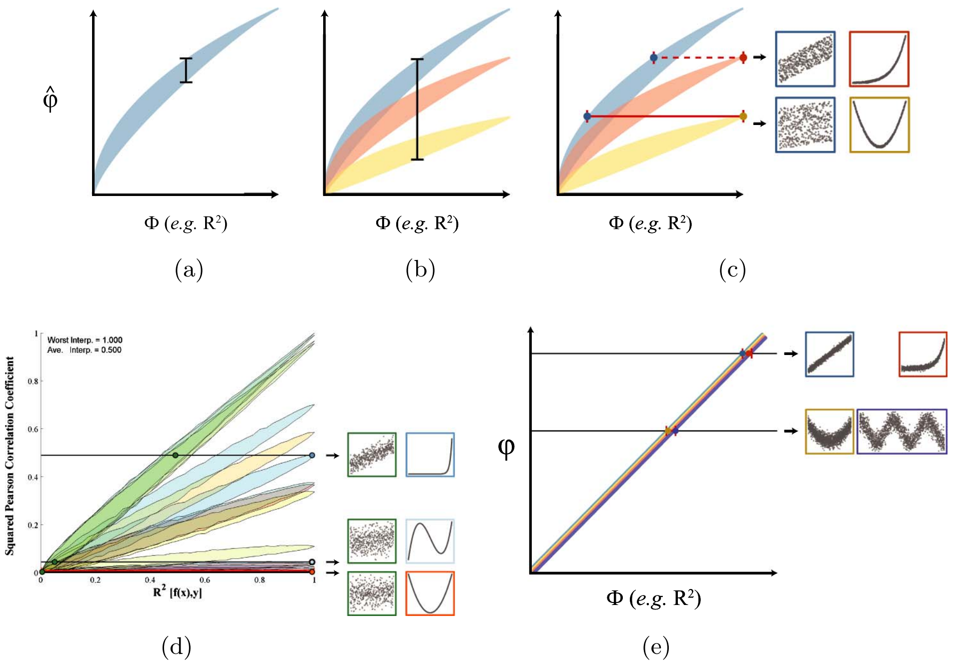

FIG. 2. *An illustration of equitability via interpretable intervals when* Φ *is* R^2. (a) *A plot of central intervals of the sampling distributions of* $\hat{\varphi}$ *against* $R^2(\mathcal{Z})$ *for* $\mathcal{Z} \in \mathcal{Q}$, *when* $\mathcal{Q}$ *consists only of linear relationships with varying amounts of added noise*; *the black interval denotes a central interval corresponding to one particular distribution* $\mathcal{Z} \in \mathcal{Q}$. (b) *The analogous plot in the case where* $\mathcal{Q}$ *contains noisy functional relationships ranging over three different functions*: *linear* (*blue*), *exponential* (*red*), *and parabolic* (*yellow*). *The black interval*, *called the reliable interval*, *is now the smallest interval containing the central intervals for all three relationship types*. (c) *The same plot*, *with interpretable intervals pictured. The interpretable interval at each value of* $\hat{\varphi}$ *is composed of the* R^2 *values whose reliable intervals contain that value of* $\hat{\varphi}$. *The shorter the interpretable intervals*, *the more equitable the statistic*. *The widest interpretable interval is denoted by a solid red line*; *an additional interpretable interval is shown with a dashed red line*. (d) *The analogous plot*, *but with* $\hat{\varphi}$ *set to be the squared sample correlation coefficient* $\hat{\rho}^2$ *and* $\mathcal{Q}$ *equal to the set of noisy functional relationships described in Appendix* C.1, *with* $n = 500$. *The fact that the interpretable intervals of* $\hat{\rho}^2$ *are large indicates that a given* $\hat{\rho}^2$ *value could correspond to relationships with very different* R^2 *values*. (e) *The analogous plot*, *for a hypothetical measure of dependence that achieves perfect equitability in the large-sample limit*. [*Parts* (d) *and* (e) *are reproduced from* Reshef *et al*. (2015).]

2.1. *Equitability on functional relationships*. In evaluating the equitability of measures of dependence, we would like to assess equitability on as broad as possible a set of relationships $\mathcal{Q}$ for which we can define a reasonable measure of relationship strength Φ. In this work we choose to focus on sets of noisy functional relationships as defined in Reshef et al. (2016). Briefly, these are relationships of the form $(X + \varepsilon, f(X) + \varepsilon')$ where f is a function that ranges over some set of functions F, and where ε and ε' are independent of each other and of X, and may be trivial. We make this choice because noisy functional relationships are a broad, easily definable class of relationships commonly found in practical applications that comes with an intuitive and natural measure of relationship strength: R^2, the coefficient of determination with respect to the generating function. Note that this set of standard relationships only includes relationships for which the first coordinate is the independent variable; extending this paradigm to larger sets of noisy functional relationships where this is not the case is a subject of future work.

Because of our focus on noisy functional relationships, in this paper, as in Reshef et al. (2011, 2016), we will typically use "equitability" to mean equitability with respect to R^2 on particular (finite) sets of noisy functional relationships that are representative of a variety of relationship types. Alternative definitions of equitability with other sets $\mathcal{Q}$ and functions Φ have been proposed; these are discussed in detail in Reshef et al. (2015).

3. Equitability analysis.

Having reviewed equitability and how to quantify it, we turn to evaluating the equitability of MIC_e, TIC_e, and several leading measures of dependence. We do so first using interpretable intervals, as in Figure 2, followed by an alternate visualization of the equitability of each measure of dependence using a power analysis, as in Figure 1.

3.1. *Setting up the analysis.*

3.1.1. *Choice of methods to analyze.*

We include in our analysis a collection of methods that is representative of the broad spectrum of approaches prevalent in the field today.

Grid-based methods. The maximal information coefficient and the total information coefficient can be viewed as exploring the space of possible grids that can be drawn on the sampled data, assigning a score to each grid via some metric, and then aggregating the scores. For MIC [Reshef et al. (2011)], the metric is a normalized mutual information score and the aggregation is a supremum. (We remind the reader that MIC is difficult to compute efficiently and so in practice a heuristic approximation called APPROX-MIC is used to compute it that does not explore the space of all possible grids.) MIC_e [Reshef et al. (2016)] is similar to MIC but explores a more restricted set of grids over which an efficient search is possible while retaining the property that its population value is a supremum over all possible grids. (As such, no approximation algorithm is needed for MIC_e.) TIC_e [Reshef et al. (2016)] is like MIC_e except it aggregates by summation. For all of these methods, the parameter α controls the space of grids that is explored; higher α means grids with more cells. For a review of the formal definitions of MIC_e and TIC_e, see Appendix B.

We also include other, more recent grid-based methods. HHG [Heller, Heller and Gorfine (2013)] explores a set of three-by-three grids defined by pairs of data points, uses as its score Pearson's χ^2 test statistic computed on two-by-two contingency tables derived from the three-by-three grids, and aggregates by summation.

Though similar to Hoeffding's D [Hoeffding (1948)] in that it proceeds via two-by-two contingency tables, it differs in the way it constructs the tables, and it is not distribution free whereas Hoeffding's D is. S^{DDP} [Heller et al. (2016)] explores a larger set of grids defined by subsets of the data points, uses nonnormalized mutual information as its score, and also aggregates by summation.[5] Another notable grid-based method introduced recently is dynamic slicing [Jiang, Ye and Liu (2015)], which like the idealized MIC explores all possible grids and aggregates by maximization, but uses as its score a version of mutual information that is regularized according to a prior on the space of possible grids. We did not include dynamic slicing in our comparison, however, because it is formulated only for performing a k-sample test whereas our focus here is on measuring dependence between two continuous random variables.

Mutual information estimation. We compare to a standard mutual information estimator introduced by Kraskov [Kraskov, Stogbauer and Grassberger (2004)]. For convenience, we represent the estimated mutual information values in terms of the squared Linfoot correlation [Linfoot (1957), Speed (2011)], defined by $L^2(X, Y) = 1 - 2^{-2I(X,Y)}$ where $I(X, Y)$ represents the raw mutual information. $L^2(X, Y)$ takes values in [0, 1].

Distance/kernel-based statistics. We compare to the distance correlation (dCor) [Szekely and Rizzo (2009)], a statistic that is defined analogously to ordinary correlation but using a notion of *distance* variance/covariance that is based on pairwise distances between points. The use of distance variance/covariance is a significant advance because in contrast to ordinary variance/covariance it produces an omnibus consistent test that, unlike grid-based approaches, easily generalizes to testing for dependence in higher dimensions. In addition to distance correlation, we compare to the Hilbert–Schmidt information criterion (HSIC) [Gretton et al. (2005, 2008)], a more general statistic defined on reproducing kernel Hilbert spaces of which dCor is a special case [Sejdinovic et al. (2013)].

Correlation-based methods. As an intuitive benchmark for the reader, we include the squared Pearson correlation coefficient (ρ^2). We also include methods that use ρ after computing a nonlinear transformation of the data. Perhaps the best-known one is maximal correlation [Rényi (1959)], which, given random variables X and Y, searches for arbitrary measurable functions f and g such that $\rho(f(X), g(Y))$ is maximized. This is algorithmically hard in general, but the (approximate) method of alternating conditional expectations [Breiman and Friedman (1985)] is widely used and we use it here as well. We also include a more recent related method, the randomized dependence coefficient [Lopez-Paz, Hennig and

[5] Several variations on these statistics are presented in Heller, Heller and Gorfine (2013), Heller et al. (2016). Results for these other methods were generally similar or worse than the ones we display, and we omit them.

Schölkopf (2013)], which applies many random transformations to X and Y and then searches for the linear combinations of the transformed features that maximize the correlation.

Parameter choice. For each of the above methods that is parametrized, we conducted a parameter sweep and present for each sample size the best seen results. Results for all parameter values are in Empirical Supplement 1E.

3.1.2. *Choice of* $\mathcal{Q}$, Φ, *and sample sizes.* As discussed in Section 2.1, we focus here on equitability with respect to R^2 on a set of noisy functional relationships. To ensure robustness, we vary the relationships tested along as many dimensions as possible including relationship type, the type of noise added, marginal distributions, and sample size.

Specifically, we considered 12 different sampling/noise models. Each sampling/noise model was defined by choosing one of four independent-variable marginal distributions (points equidistant or uniformly sampled, along the graph of the function or along the X-axis) and one of three noise distributions (X noise only, Y noise only, or noise in both variables; see Appendix C.4.1). For each sampling/noise model, we created a set of relationships $\mathcal{Q}$ by including between 16 and 21 different functional relationships (see Appendix C.4.2), each with increasing levels of additive Gaussian noise, at four sample size regimes ($n = 250, 500, 5000$, and the infinite data limit).

3.1.3. *Quantification of equitability.* The equitability of each measure of dependence is quantified using $(5\%, 95\%)$-interpretable intervals, as in Figure 2 (see Appendix A). We report both average-case and worst-case equitability in our analyses, and the interval plotted in red on each plot represents the worst-case interpretable interval for that plot.

3.2. *Results and discussion.* Figures 3 and 4 display the results of our analysis for a subset of methods under the noise/sampling model $(x_i + \varepsilon_i, f(x_i) + \varepsilon_i')$, where $\varepsilon_i, \varepsilon_i'$ are i.i.d. Gaussians for all i and the x_i are chosen to make consecutive points $(x_i, f(x_i))$ equidistant along the graph of f. The full results are in Empirical Supplement 1A–F, and are summarized in Tables A4 and A5.

Our results demonstrate that MIC_e is consistently highly equitable for these noise/sample models and sample sizes, and is the only one of the methods examined here to be so. Mutual information shows relatively poorer equitability at $n = 250$ and $n = 500$. While its equitability appears improved at $n = 5000$, this improvement is not robust to variation in noise/sampling model; we discuss the equitability of mutual information in more detail in the following section. Of the remaining schemes, maximum correlation appears to provide the best equitability. This is interesting because on the one hand the squared maximal correlation is bounded from below by R^2, and on the other hand the lack of equitability of

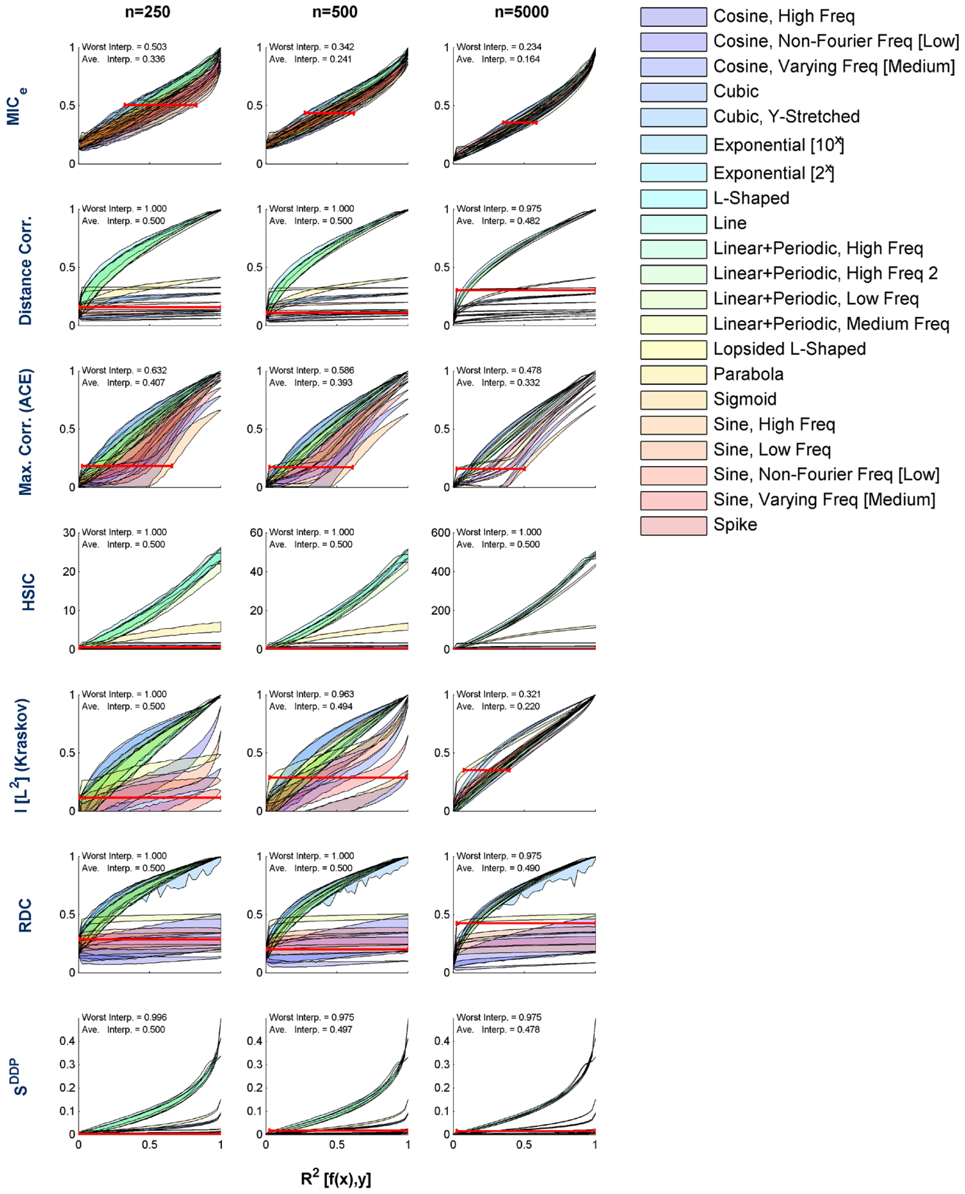

FIG. 3. *The equitability of measures of dependence on a set of noisy functional relationships. (Narrower is more equitable.) The relationships take the form $(X+\varepsilon, f(X)+\varepsilon')$ where ε and ε' are i.i.d. normals of varying amplitude, and relationship strength is quantified by $\Phi = R^2$. The plots were constructed as described in Figure 2. In each plot, the worst-case interpretable interval is indicated by a red line, and both the worst- and average-case equitability are listed. Mutual information, estimated using the Kraskov estimator, is represented using the squared Linfoot correlation. For every parametrized statistic whose parameter meaningfully affects equitability, results are presented at each sample size using parameter settings that maximize worst-case equitability across all 12 of the noise/marginal distributions tested at that sample size.*

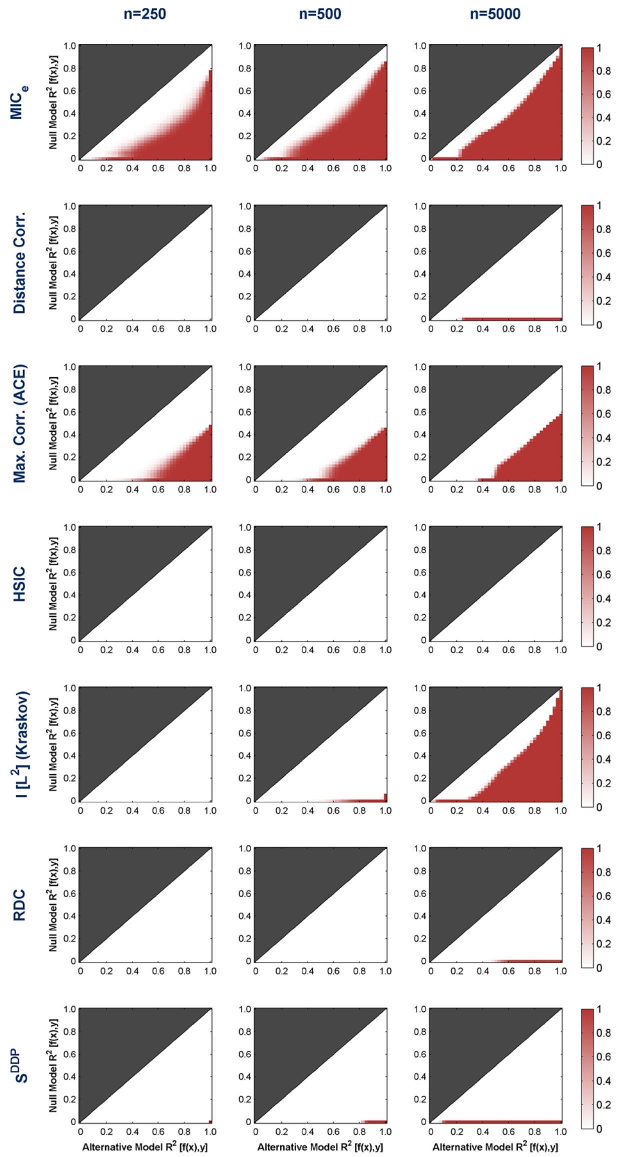

FIG. 4. *The equitability with respect to $\Phi = R^2$ of measures of dependence on the noisy functional relationships analyzed in Figure 3, visualized in terms of power. (Redder is more equitable.) Plots were generated as in Figure 1. The intensity of the pixel at coordinate (x_1, x_0) in each heat map shows the power of a right-tailed test based on the statistic in question at distinguishing the (composite) alternative hypothesis $H_1 : R^2 = x_1$ from the (composite) null hypothesis $H_0 : R^2 = x_0$ with type I error at most $\alpha = 0.05$. Mutual information, estimated using the Kraskov estimator, is represented using the squared Linfoot correlation. For every parametrized statistic whose parameter meaningfully affects equitability, results are presented at each sample size using parameter settings that maximize worst-case equitability across all twelve of noise/marginal distributions tested at that sample size.*

maximal correlation seems to stem from the ACE method returning results below this lower bound. We therefore wonder whether maximal correlation—were it computable exactly—would be highly equitable with respect to R^2.

We comment briefly on the remaining methods: HSIC, distance correlation, RDC, S^{DDP}, TIC$_e$, HHG, and ρ. These methods all display relatively poor equitability over the models $\mathcal{Q}$ tested, with the equitability profiles of both dCor and RDC appearing similar to that of the squared sample correlation (Empirical Supplement 1E). Of course, none of these methods were designed with equitability with respect to R^2 in mind or make claims about equitability with respect to R^2. We note further that each of these methods that converges to some population value is trivially a consistent estimator of that population value and therefore trivially perfectly equitable with respect to that population value. Therefore, if, for example, a practitioner believes that the population value of dCor is the best way to measure relationship strength for a particular application, then the dCor statistic should of course be the statistic of choice. Our results have implications only for cases in which R^2 is considered an appropriate measure of relationship strength against which to benchmark the methods we have evaluated.

Interestingly, Figure 4 also shows poor power to reject a null hypothesis of independence at $n = 250$ and $n = 500$ even for methods like HSIC, dCor, and RDC, which are traditionally considered to have good power against independence. The reason this happens is because equitability measures worst-case power across all relationship types with a given R^2; that is, the alternative hypotheses considered are composite. Correspondingly, the statistics that are not well powered to detect even one of the relationship types analyzed compare unfavorably in this analysis, even if they have good power on a large subset of the relationships.

We note that, for MIC$_e$, the best parameter regime for equitability is different than the best parameter regime for power against independence presented later in this paper. This suggests that there is a trade-off between power against independence and equitability, a theme to which we return in Section 6.

3.2.1. *Comparing the equitability of* MIC$_e$ *and mutual information.* Given the connections between the maximal information coefficient and mutual information, it is natural to ask whether direct estimation of mutual information achieves similar equitability to MIC$_e$. The equitability of mutual information estimation has been assessed previously, most notably in Reshef et al. (2011), Reshef et al. (2013), Kinney and Atwal (2014), and Reshef et al. (2014). The analyses conducted here, which subsume those analyses, show that in general the answer appears to be: at $n = 250$ and $n = 500$, MIC$_e$ outperforms mutual information estimation on all models tested, often by substantial margins; at $n = 5000$, MIC$_e$ outperforms mutual information on all models except for the ones that contain Y noise only, on which mutual information performs better. We present a more detailed breakdown below.

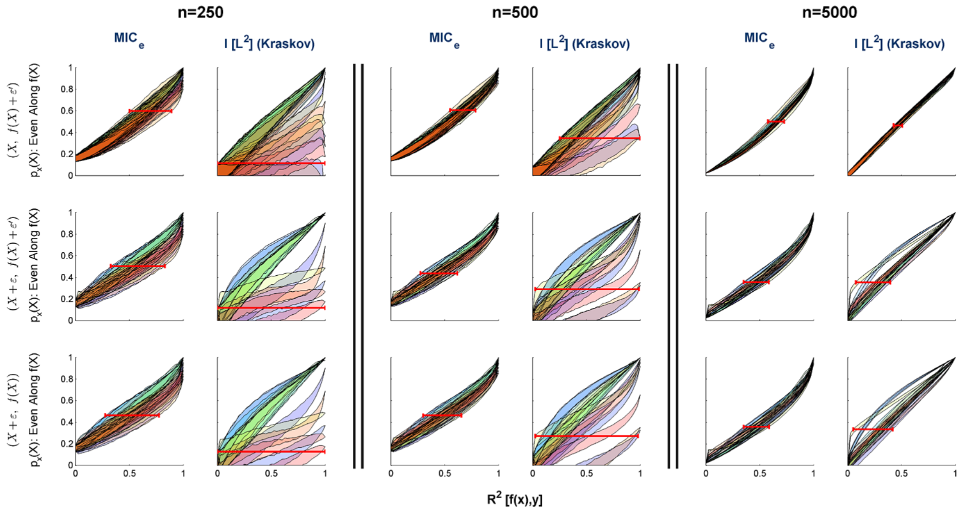

FIG. 5. *A comparison of the equitability of* MIC$_e$ *and mutual information estimation under three noise models. (Narrower is more equitable.) Plots are analogous to those in Figure* 3. *As in that figure, results for both statistics are presented for each sample size using parameter settings that maximize equitability across all twelve of the noise/marginal distributions tested at that sample size. For versions of this analysis using additional independent variable marginal distributions, see Empirical Supplement* 1*C*.

Effect of model choice on equitability. Figure 5 demonstrates the relative robustness to model choice of the equitability of MIC$_e$ compared to that of the Kraskov mutual information estimator. At each sample size, the equitability of MIC$_e$ is fairly stable with respect to the variations in noise models and independent variable marginal distributions tested. In contrast, it seems that mutual information estimation's equitability relies on the noise model containing no X noise.

Effect of sample size on equitability. Estimating mutual information from finite samples is a challenging problem that has inspired many sophisticated methods [Kraskov, Stogbauer and Grassberger (2004), Moon, Rajagopalan and Lall (1995), Paninski (2003)], and indeed our analyses demonstrate strong finite-sample effects on the equitability of mutual information estimation. MIC$_e$ suffers much less from this problem: for $n = 250$ and $n = 500$, MIC$_e$ has both superior worst-case and average-case equitability over mutual information estimation (using $k = 1$, 6, 10, and 20 in the Kraskov estimator) in every model $\mathcal{Q}$ tested, in most cases by substantial margins. This is intuitively consistent with the fact that the population value of MIC$_e$ is uniformly continuous as a functional while mutual information is not [Reshef et al. (2016)].

Equitability in the large-sample limit. To disentangle finite-sample effects from properties of the population values of the statistics in question, we also compared the equitability of the population value of MIC$_e$ (called MIC$_*$) and the population value of mutual information (Figure A4). Results were essentially the same as those for $n = 5000$, implying that neither MIC_* nor mutual information is worst-case perfectly equitable with respect to R^2 over the sets $\mathcal{Q}$ examined. This is not surprising given the broad range of relationships, noise models, and independent variable marginal distributions tested.

Relationship to equitability analysis from Kinney and Atwal (2014). A more limited empirical analysis of the equitability of MIC and mutual information estimation was presented in Kinney and Atwal (2014). There, the authors examined the equitability of MIC and mutual information estimation at a large sample size ($n = 5000$) and under one choice of $\mathcal{Q}$ (the same as in Figure 3, only with no noise in the first coordinate). From this, they concluded that mutual information estimation was more equitable than MIC. This empirical argument was accompanied by a theoretical result exhibiting a family of relationships on which no measure of dependence can be perfectly equitable with respect to R^2, and a statement that this impossibility result implies that previous claims [Reshef et al. (2011)] about the equitability of MIC were incorrect.

Since its publication, Kinney and Atwal (2014) has been the subject of two published technical comments [Murrell, Murrell and Murrell (2014), Reshef et al. (2014)] describing its main limitations, which are threefold. First, the central proof of the impossibility of equitability with respect to R^2 in Kinney and Atwal (2014) applies only to *perfect* equitability, and says nothing about the achievability of the more general (approximate) notion with which we are primarily interested and regarding which we have previously made claims about MIC. That is, even if no method is perfectly equitable with respect to R^2, some methods can be more equitable with respect to R^2 than others, and the question remains which methods come meaningfully close to the ideal [Reshef et al. (2014)]. Second, the impossibility result relies crucially on a nonidentifiable noise model $\mathcal{Q}$ in which, for example, a noiseless parabola can be obtained as a "noisy" linear relationship [Murrell, Murrell and Murrell (2014)]. Third, though mutual information indeed outperforms MIC under the specific sample size and noise model chosen in Kinney and Atwal (2014), this is not the case in general [Reshef et al. (2014)]. As our analysis here importantly establishes, this empirical point remains true even when we further expand the set of noise models and sample sizes under consideration.

3.2.2. *Sensitivity of analysis to choice of functions.* One potential question about the equitability analyses performed here is whether they are sensitive to the particular choice of functions analyzed. This is justified given that the current theoretical understanding of the maximal set of functions on which we should expect MIC_e (or any method) to behave equitably is quite limited, and given that one can construct functions, such as a step function, for which all three of the methods that show nontrivial equitability in the above analysis provably perform very nonequitably. (See Appendix J for a proof.) However, analyses we have conducted in separate work suggest that our results appear robust over a wide range of "probable" function types. Specifically, in Reshef et al. (2016) we conducted equitability analyses similar to the ones above but on a set of 160 functions chosen at random from Gaussian process distributions with radial basis function kernels of different bandwidths. Results were similar, with MIC_e attaining the best equitability, followed by mutual information estimation, and then maximal correlation.

3.2.3. *Nonfunctional relationships.* Equitability as we have applied it here is only defined for noisy functional relationships. However, in previous work [Reshef et al. (2011)] we showed empirically in the case of MIC that reasonable equitability with respect to R^2 can translate into reasonable behavior on several different nonfunctional relationships, with the MIC of those relationships degrading intuitively as noise is added [see Figures 2G, S5, and S6 of Reshef et al. (2011)]. We also proved that the population MIC (and therefore also the population MIC_e) of superpositions of noiseless never-constant functional relationships is 1 [see Theorem 4 of Reshef et al. (2011)]. More in-depth empirical and theoretical examination of this aspect of MIC and MIC_e is an important direction of future work.

4. Statistical power analysis. There are many settings that call simply for testing for *any* deviation from independence rather than relationship ranking. These settings require a measure of dependence that yields tests with high power against a null hypothesis of statistical independence.

Here, we turn to assessing the power against independence of the above statistics. Such analyses have been done previously, most notably by Simon and Tibshirani [Simon and Tibshirani (2012)]. Our analysis expands upon the power analysis performed by Simon and Tibshirani in three key ways. First, for each of the statistics we analyze that has a free parameter, we perform a parameter sweep to understand the power of the corresponding tests as a function of that parameter. Second, we analyze a larger set of methods and a greater variety of sample sizes. Finally, we consider several ways to aggregate information across noise levels and across function types to get a more general picture of which methods have the highest overall power.

4.1. *Setting up the analysis.* We analyze all methods listed in Section 3, and we perform parameter sweeps for every method and report best-seen results as in that section. We use the set of relationships and noise model (uniform independent-variable marginal, Gaussian noise in the second coordinate only) chosen by Simon and Tibshirani [Simon and Tibshirani (2012)]. For consistency with the sample sizes used throughout this work, we show results for $n = 500$; results for all analyses using $n = 100$ are similar and are provided in the empirical supplement.

We first compute power curves for each relationship type and each method, having performed parameter sweeps to choose optimal parameters for each method as a function of sample size only (see Appendix C.2). The parameter sweeps themselves, which characterize power against independence as a function of statistic parameters, are presented in Figures A6 and A7.

To compare power across methods, we need to aggregate information across relationship types as well as across alternative hypotheses. The first way we do this is to integrate under the power curve of each relationship type and average across relationship types, using limits of integration defined via R^2 for consistency across relationship types (see Appendix C.2.1). The second way we aggregate this information is to compute, for each method and each function type, the R^2 at which 50% power is reached, and then average this quantity across function type.

4.2. *Results.* The full power curves for individual relationship types and methods are displayed in Figure 6, and the aggregated results comparing overall power across methods are shown in Figure 7. We discuss several aspects of these below.

4.2.1. *Power on specific relationships.* In Figure 6, no method clearly dominates; different methods have good power for different relationship types. For example, distance correlation and HSIC are relatively better powered to detect linear dependence than MIC_e and TIC_e, but are relatively worse at detecting most of the other forms of dependence tested. In contrast, S^{DDP} appears to have a similar profile to that of TIC_e. This is interesting because S^{DDP} is closely related to the maximal and total information coefficients in that it too is an aggregation via summation of mutual information scores taken over many different grids.

However, choice of parameter values is an important determinant of power, and unsurprisingly, the optimal parameter choices used here cause the power of tests based on several of the statistics included in this analysis to be substantially better than previously reported [Gorfine, Heller and Heller (2012), Jiang, Ye and Liu (2015), Kinney and Atwal (2014), Lopez-Paz, Hennig and Schölkopf (2013), Simon and Tibshirani (2012)]. In particular, MIC with optimal parameters (black line) performs substantially better than MIC with what were previously the default, equitability oriented parameters. This performance gain is achieved by a wide range of parameter settings comprising a regime suited for independence testing (Figure A6). Importantly, it is preserved on an independent validation set of randomly chosen noisy functional relationships (Figure A1), indicating that it is not idiosyncratic to the particular relationships employed in this analysis. We have therefore updated our software to allow users to choose between the parameters that optimize power or the parameters that optimize equitability.

4.2.2. *Average power across relationship types.* The two rankings displayed in Figure 7, while robust to sample size and thresholds used, are different from each other, and are sensitive to choices such as inclusion/exclusion of certain function types (Empirical Supplement 2A and 2B). However, there are some general patterns that seem consistent. First, state-of-the-art performance is always achieved

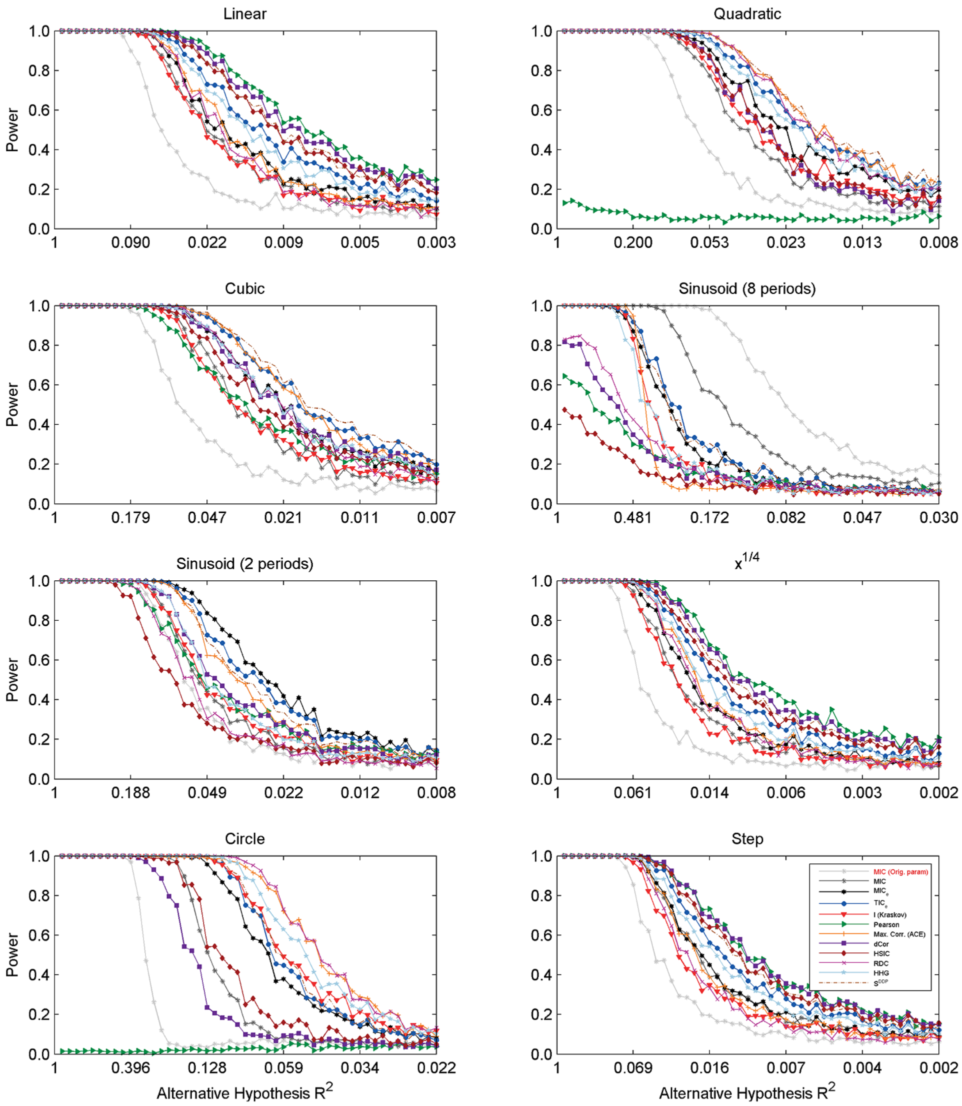

FIG. 6. *Power of independence testing using the measures of dependence examined, on the relationships in Simon and Tibshirani* (2012), *at* 50 *noise levels with linearly increasing magnitude for each relationship and* $n = 500$. *To enable comparison of power regimes across relationships, the x-axis of each plot lists* R^2 *rather than noise magnitude. For each statistic that has a parameter, an optimal value for the parameter was chosen using the parameter sweeps in Figure* A6. (*For a version with* $n = 100$ *see Empirical Supplement* 2A.)

by either S^{DDP} or TIC_e, depending on how power is quantified. This provides evidence that the basic approach of aggregating mutual information scores over a large set of grids, whether via the characteristic matrix or other statistics, is a fun-

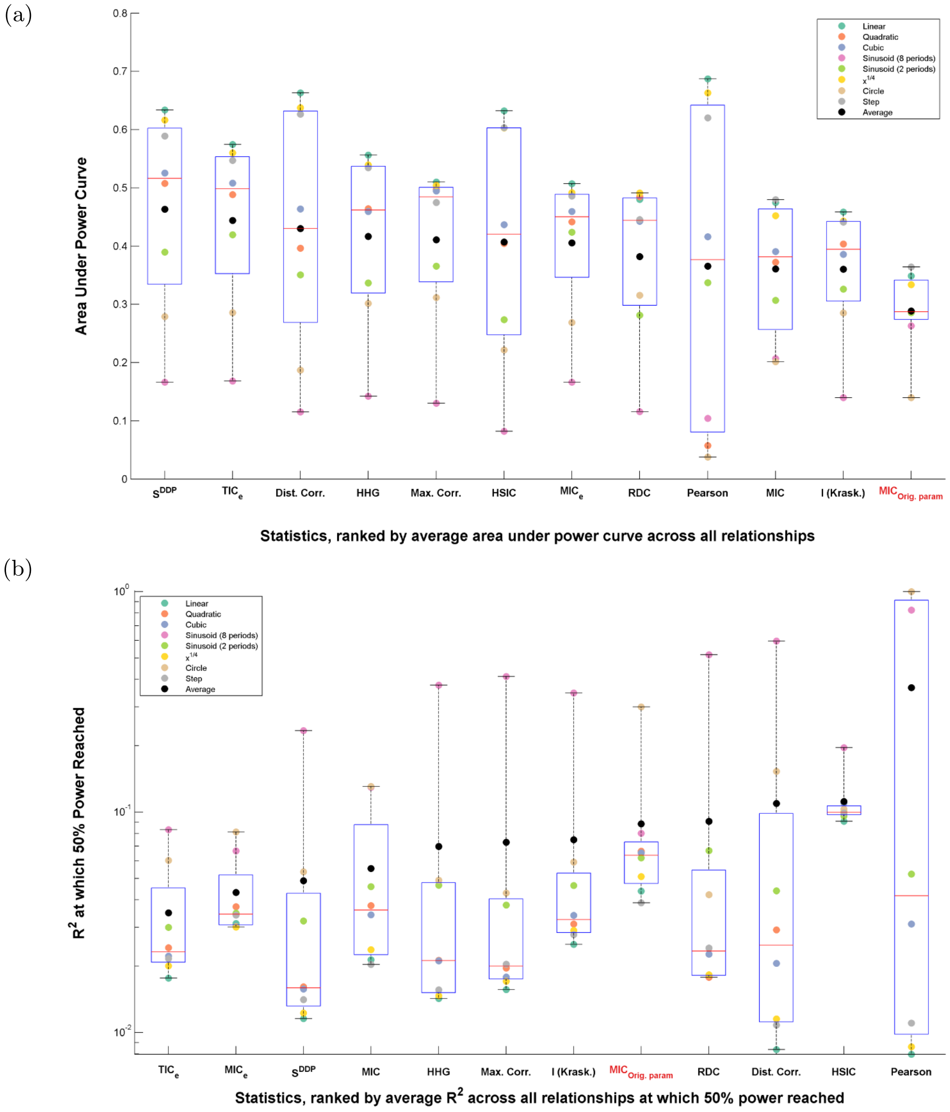

FIG. 7. *Measures of dependence ranked by the power of their corresponding independence tests. For each measure of dependence and each relationship type, power was quantified using* (a) *the area under the power curve (higher is more powerful), or* (b) *the minimal* R^2 *at which at least* 50% *power is achieved (lower is more powerful). The collection of these scores across relationship types is then plotted for each method along with quartiles. Optimal parameter values for each test statistic were chosen to maximize average performance across relationship types*; *see* (a) *Figure* A6, *or* (b) *Figure* A7. *The* MIC *statistic from Reshef et al.* (2011) *with the parameters used in Simon and Tibshirani* (2012) *is labeled in red. The sample size is* $n = 500$; *results are similar with* $n = 100$ *and, for* (b), *with power thresholds besides* 50%. (*See Empirical Supplement* 2B.)

damentally promising avenue for thinking about dependence. Additionally, when power is quantified by computing the area under the power curve, distance correlation also does quite well, thus highlighting the value of the by-now well established paradigm of energy statistics for relationship detection.

Second, the power of independence testing using MIC_e, with parameters suited for power against independence rather than equitability, is not far from the state of the art, though its relative performance depends on the method of quantification. In particular, its power is comparable to- and usually higher than that of its predecessor MIC [Reshef et al. (2011)], which estimates the same population quantity (MIC_*), even when the latter also has optimally chosen parameters. This demonstrates that the improved bias/variance properties of MIC_e relative to MIC [Reshef et al. (2016)] indeed translate into an improvement in power.

Another observation arising from Figure 7(b) is that if we computed for each method the R^2 at which 50% power is reached for *all* function types tested simultaneously—that is, the maximum over function types of R^2 at which 50% power is reached, instead of the average over function types—then the rankings of the methods would be quite different. We will return to this in Section 7, where we argue for the utility of this way of assessing performance.

We remark that since the parameters chosen for each parametrized method were optimized for the function suite we analyzed, one may ask to what extent these results would generalize to relationship types beyond the ones considered here. To assess this, we also conducted the same power analysis on an independent set of 160 relationships consisting of randomly chosen functions with noise added, using the parameter settings resulting from our parameter sweep of the fixed set of relationships. Results were similar. (See Appendix C.3.)

5. Runtime analysis.

Computational efficiency is often desirable when evaluating dependence, and here we assess the runtimes associated with the set of measures of dependence examined.

5.1. *Setting up the analysis.*

Since the runtime of MIC_e/TIC_e depends on parameter choice, results for MIC_e are presented for parameter settings recommended for maximizing equitability, maximizing power against independence, and attaining "reasonable equitability". The third set of parameters was computed by searching at each sample size for the parameters that resulted in the fastest runtime while still yielding 80% of the best observed equitability at that sample size. All the parameters used for MIC_e/TIC_e in this analysis are detailed in Table A9.

The only other method whose runtime is affected by its parameter was S^{DDP}. Since at the sample size regimes we tested only three parameter settings led to practical runtimes for S^{DDP}, we have included all three. For statistics whose runtimes did not depend on parameter choice, defaults were used (see Appendix G.3).

TABLE 1

Average runtimes, in seconds, of algorithms for computing measures of dependence over 100 *trials of uniformly distributed, independent samples at a range of sample sizes. Results for* MIC_e *are presented for three sample-size-dependent parameter settings that optimize for maximal power against independence* ([*P*]), 99% *of optimal equitability* ([*E*]), *and* 80% *of optimal equitability* (*fast equitability,* [*FE*]). *For a list of the parameters used in each of these settings, see Table* A9. TIC_e *is omitted because its runtime is very similar to that of* MIC_e [*P*]

n	ρ^2	**Max. Corr.**	**RDC**	**dCor**	**HSIC**	**HHG**	$\mathbf{I}_{\mathbf{(Kraskov)}}$
50	0.0001	0.0004	0.0015	0.0010	0.0016	0.0017	0.0096
100	0.0001	0.0005	0.0014	0.0014	0.0032	0.0063	0.0100
500	0.0001	0.0014	0.0023	0.0504	0.0847	0.2185	0.0122
1000	0.0002	0.0025	0.0035	0.3518	0.4886	1.0956	0.0150
5000	0.0002	0.0119	0.0129	6.1402	6.5975	34.0171	0.0427
10,000	0.0002	0.0239	0.0251	25.9859	25.7333	465.3222	0.0927

n	**MIC**	$\mathbf{MIC}_e$ **[E]**	$\mathbf{MIC}_e$ **[FE]**	$\mathbf{MIC}_e$ **[P]**	$S^{\mathrm{DDP}}_{m=2}$	$S^{\mathrm{DDP}}_{m=3}$	$S^{\mathrm{DDP}}_{m=4}$
50	0.0015	0.0021	0.0009	0.0004	0.0018	0.0010	0.0094
100	0.0061	0.0052	0.0012	0.0005	0.0022	0.0023	0.0861
500	0.2187	0.1630	0.0079	0.0018	0.0035	0.0529	14.2690
1000	0.9628	0.1992	0.0172	0.0037	0.0050	0.2122	121.7311
5000	18.7627	0.3398	0.0974	0.0195	0.0574	5.7464	1.72×10^4
10,000	66.2238	0.6835	0.1819	0.0398	0.2154	23.4473	1.40×10^5

5.2. *Results.* The results of our runtime analysis, found in Table 1, have several salient features. First, there is a clear set of fastest methods: maximum correlation, RDC, MIC_e (with any of the three parameter settings tested), TIC_e (which has identical runtime to MIC_e and so is omitted from Table 1), mutual information, and S^{DDP} with $m = 2$ (a parameter setting that was not chosen by our parameter sweeps due to its worse power; see Figures A6 and A7). Each of these methods takes under a second to compute at a sample size of 10,000, while the remaining methods all take over 20 seconds.

Second, MIC_e with all three of the parameter settings given is substantially faster than the previously introduced MIC statistic from Reshef et al. (2011) run using default parameters. This matches the theoretical analysis in Reshef et al. (2016), which shows that the complexity of the search procedure in MIC_e is $O(n^{2.5\alpha})$ whereas the complexity of the search procedure in the APPROX-MIC algorithm used to compute MIC is $O(n^{4\alpha})$.

Third, analysis of large data sets is possible using MIC_e and TIC_e. For example, computing both TIC_e with parameters optimized for power and MIC_e with parameters chosen to achieve 80% of the best achievable equitability can be done on a sample size of 5000 in 97 milliseconds. For a data set with $n = 5000$ consisting of

1000 variables, this translates into a total runtime of 16 minutes to compute both statistics for all variable pairs using 50 processing nodes.

We note one interesting feature of the runtime of MIC_e. Since estimating MIC_* involves a search procedure, runtimes for estimating it are substantially faster when data contain less noise; as such, the runtimes on statistically independent data presented in Table 1 represent worst-case performance. When run on data drawn from a noiseless linear relationship at the same sample sizes, MIC_e ran 5%-75% faster. The runtime of S^{DDP} exhibited a similar phenomenon, but the runtimes of the other methods were insensitive to the level of structure present and did not exhibit this effect.

We emphasize that our results represent a snapshot based on currently available implementations. Just as MIC_e has provided an improvement over APPROX-MIC, and just as estimating distance correlation has recently been shown to be estimable in time $O(n \log n)$ rather than $O(n^2)$ (not benchmarked here; see Remark A1), we expect that with time algorithmic improvements will allow for more efficient computation of some of the newer methods analyzed here.

6. The power-equitability trade-off.

For several methods, the parameter regimes that maximize power are different from the parameter regimes that maximize equitability. This suggests that there may be a trade-off between these two objectives that is being captured by the choice of parameter setting [Reshef et al. (2013)]. Such a trade-off seems plausible given the equivalence proven in Reshef et al. (2015) between equitability and power against a range of null hypotheses corresponding to different relationship strengths. Since equitability is about simultaneously achieving high power against many null hypotheses, it is reasonable that to attain this objective we have to give up some of the power we previously had against the specific null hypothesis of independence. Here we show empirically that such a trade-off does indeed exist for each of the parametrized methods we consider.

6.1. *Demonstrating the power-equitability trade-off.*

For each statistic under consideration we plotted worst-case equitability against average power at a sample size of 500 while varying the statistic's parameter if it had one. The results are displayed in Figure 8.

Figure 8 shows that every parametrized method with a nontrivial level of equitability does indeed exhibit a power-equitability trade-off on the sets of relationships considered in this paper. In the case of MIC_e, the trade-off is captured by the parameter α, which controls the maximal grid resolution used by the statistic. This is consistent with the bias-variance analysis in Reshef et al. (2016), which showed that low values of α lead to better performance in the low-signal regime while larger values of α lead to better performance in mid-to-high-signal regimes. It is also consistent with the intuition that disallowing high-resolution grids may

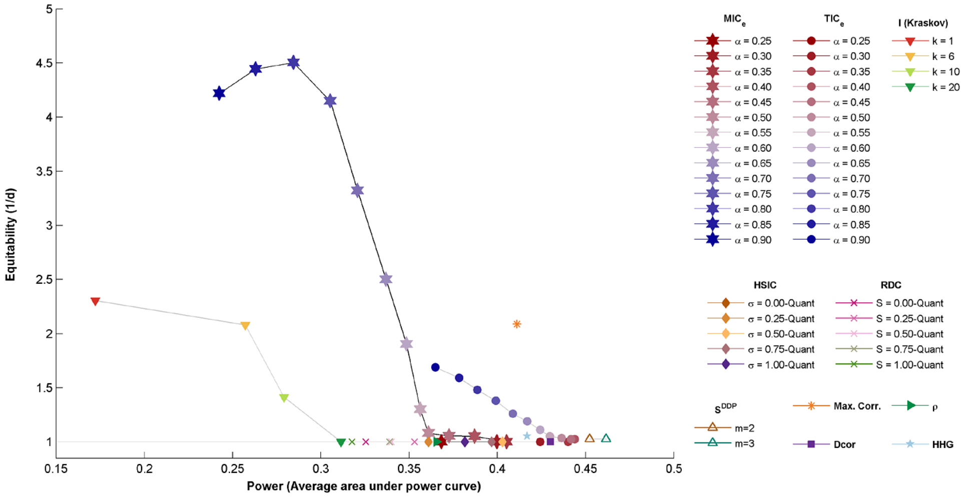

FIG. 8. *The trade-off between equitability and power against statistical independence across methods. For each method, average power as quantified in Figure* 7(*a*) *is plotted against the worst-case equitability under the analogous noise/sampling model, with* $n = 500$. *For every parametrized method, a point is plotted for each assessed value of the parameter in question. Since each coordinate is strictly preferable to all coordinates below and to the left of it, there is a Pareto "power-equitability" front. The methods with points along this front are* MIC_e, *maximal correlation*, TIC_e, *and* S^{DDP}.

increase power against independence but will allow only coarse-grained distinguishability among distributions, while allowing high-resolution grids might enable distinguishing between distributions that may be more similar to each other.

Figure 8 is also a useful summary of how the different methods we considered compare to each other along these two dimensions (for this sample size and set of relationships). Specifically, if one point is both above and to the right of another then it is strictly preferable. Thus, the figure shows a Pareto front of methods that offer optimal performance with respect to power against independence and equitability. This front includes MIC_e, maximal correlation, TIC_e, and S^{DDP}. When power is assessed as in Figure 7(b) instead of Figure 7(a), the Pareto front includes only MIC_e and TIC_e (see Empirical Supplement 3).

7. Practical suggestions. To choose which method to use in a data analysis, we must consider our goals.

In many situations, such as when sample sizes are small enough or relationships noisy enough that any method will identify only a small number of relationships, we want to maximize the number of relationships identified by a method. In these cases, it will be desirable to use a method with high power to detect the relationship types that are most common in the dataset. So if the dataset contains a large number of linear relationships and a small number of sinusoidal relationships, then the best choice of method will be one that has high power to detect linear relationships, even if it has lower power on sinusoidal relationships.

The situation we have chosen to focus on for this work is different: we are interested in the situation in which many methods will return a large number of relationships, and so the number of relationships detected is less important than their relative ranking. This does happen in practice, for example in the gene expression analysis of Heller et al. (2016), in which several methods identified over half of the thousands of relationships in the data set as significant, as well as in the analysis of the WHO data set conducted in Section 8 of this paper. In such cases, increasing the proportion of variable pairs identified as significant seems less important for scientific inquiry than having a meaningful way to prioritize the detected relationships for follow-up.

Thus, a promising strategy for exploratory data analysis is: first, to compute a statistic designed to identify a large number of significant relationships of all kinds, and then second, to compute an equitable statistic on all significant relationships, ensuring a ranking that is meaningful. For this approach to be fruitful, the statistic used in the first step must have high power on a wide range of relationship types; otherwise, the first step will eliminate many relationships that would otherwise be ranked as highly interesting in the second step. In other words, the statistic used in the first step should perform well with respect to the third quantification of power discussed at the end of Section 4: minimum R^2 at which 50% power is reached for all function types tested. The R^2 at which this is achieved is called the *detection threshold* of the method; a method that does well with respect to this quantification of power has a *low detection threshold* [Reshef et al. (2015)]. As described in Reshef et al. (2015), low detection threshold is related to equitability: an equitable statistic provably has a low detection threshold on its set of standard relationships, whereas the converse is not true.

Figure 7(b) shows that MIC_e and TIC_e both have lower detection thresholds than the other methods considered here. This phenomenon is robust to choice of power threshold (see Empirical Supplement 2B) and holds over a range of parameter settings (see Figure A7). Because their detection thresholds are very similar and TIC_e has better power than MIC_e on almost every function type, we propose to use TIC_e for a "first-pass" filtering of the relationships, and then the more equitable MIC_e to rank significant relationships.

Detection threshold is sensitive to the relationship set in question, and different relationship sets may lead to different conclusions. For instance, at the parameter setting shown in Figure 7(b), if the higher-frequency sinusoid is removed from the set of functions, S^{DDP} achieves a lower detection threshold than TIC_e. If the parameters for all methods are optimized for the new, smaller set of functions, the performance of TIC_e matches and sometimes exceeds that of S^{DDP} (Empirical Supplement 2B), but choosing parameters in this way may be difficult in practice. Therefore, for analyzing a data set where even the most interesting relationships are relatively simple, our results suggest that S^{DDP} may provide a good first-pass filter. However, for analyzing a data set in which the types of relationships present are unknown or diverse, as is our focus here, our results suggest that TIC_e is less likely than S^{DDP} to exclude relationships that might later be ranked as very interesting by MIC_e. We note parenthetically that for larger sample sizes, the increased runtime of S^{DDP} may present an additional challenge.

Using TIC_e for the first-pass filtering step has the advantage that computing MIC_e and TIC_e simultaneously is not more computationally expensive than computing just one of them. This is true even though the value of the parameter α of TIC_e that leads to optimal power against independence is not equal to the value of α used for optimal equitability of MIC_e, since computing either statistic with a given value of α also yields the values of that statistic for all lower values of α. In most situations, we expect that the value of α desired for MIC_e will be greater than that desired for TIC_e since the former will be run with equitability in mind, and so TIC_e will be a trivial side product of the computation of MIC_e.

When choosing parameters we recommend using the parameters for TIC_e that maximize power and the parameters for MIC_e that maximize equitability. These are the defaults in our software. For a discussion of alternative ways to choose parameters, see Appendix I.

8. Analysis of WHO data. To test the conclusions of our simulations on real data, we analyzed the aforementioned set of 356 social, medical, economic, and political indicators measured by the WHO in different countries. We chose to analyze this data set because previous analyses [Reshef et al. (2011)] have shown it to contain many linear relationships but also interesting nonlinear relationships. These include, for example, a relationship between obesity and income per person that consists of one trend among Pacific island nations, where female obesity is a sign of status [Gill et al. (2002)], and a separate trend in the rest of the world. Here we analyzed the 49,286 potential pairwise relationships in this data set with $n \geq 50$ using the parameter settings determined by the simulations from Sections 3 and 4. (See Appendix G.4 for details.)

We first conducted a standard power analysis, asking how many nontrivial relationships the methods under consideration identified in this data set (Table 2). Strikingly, most methods identified over 15,000 relationships as significant at level 0.05 after Bonferroni correction. When a false discovery rate of 0.05 was used instead, these methods discovered at least 30,000 relationships. The combination of MIC_e and TIC_e proposed in the previous section detected 34,465 relationships. In comparison, the most powerful method, HHG, detected 36,338 relationships. The large number of relationships detected by most of the methods underscores the need for a principled way of exploring large data sets that is more fine-grained than testing for deviations from independence.

We next turned to assessing equitability. Equitability is difficult to analyze directly here since we do not have a ground truth: we do not know which relationships in the data set are in our $\mathcal{Q}$ and which are not, and we cannot directly compute a population quantity of interest. However, we can still indirectly learn about

TABLE 2

The performance of each of the statistics on the WHO data set. Jaccard indices were computed using the top thousand relationships ranked by each method. (*Higher Jaccard distance indicates less similarity.*)

Statistic	**# (%) rejections, FWER ≤ 0.05**	**# (%) rejections, FDR ≤ 0.05**	**% of top 1k rels. with $\vert\rho\vert < 0.85$**	**Avg. Jaccard to other statistics**
$\text{MIC}_e/\text{TIC}_e$	17,630 (36%)	34,465 (70%)	29.9%	52.7%
dCor	17,783 (36%)	34,992 (71%)	1.7%	35.5%
MaxCor	4324 (9%)	29,042 (59%)	24.0%	43.3%
HSIC	17,524 (36%)	35,052 (71%)	16.1%	47.2%
Kraskov	15,326 (31%)	30,477 (62%)	7.1%	36.2%
RDC	3577 (7%)	23,086 (47%)	26.2%	45.9%
S^{DDP}	18,721 (38%)	35,582 (72%)	5.5%	34.7%
HHG	18,891 (38%)	36,338 (74%)	20.3%	48.7%
Sq. Pearson	17,073 (35%)	33,202 (67%)	0.0%	35.4%

equitability by checking for behaviors that we would expect an equitable statistic to exhibit.

For example, the equitability plots in Figure 3 show that most of the nonequitable statistics tend to give higher scores to linear and monotonic relationships. This leads to the hypothesis that in a data set that contains some complex relationships, a more equitable statistic will be better able to rank these complex relationships highly, rather than below a large number of linear relationships. And indeed, the fraction of the top 1000 relationships as ranked by $\text{MIC}_e/\text{TIC}_e$ with $|\rho| < 0.85$ was 29.9%, the most of any of the statistics tested. (See Table 2.) The two next-best-performing methods by this metric were RDC and maximal correlation, which achieved 26.2% and 24% respectively. This behavior is consistent with the nontrivial levels of equitability shown by maximal correlation in our simulations along with the theoretical parallels between RDC and maximal correlation (see below). Of the six methods besides $\text{MIC}_e/\text{TIC}_e$ that detected a very large number of relationships (rejection rate $\geq 30\%$ after Bonferroni correction), HHG was closest in performance, identifying 20.3% relationships that were not strongly linear, about two-thirds the amount identified by $\text{MIC}_e/\text{TIC}_e$.

The relationships ranked highly by $\text{MIC}_e/\text{TIC}_e$ contain results of potential scientific interest. These include relationships previously detected in Reshef et al. (2011), such as the aforementioned relationship between income per person and obesity ($p \leq 6.0 \times 10^{-7}$), a highly nonlinear relationship between number of physicians and deaths due to HIV/AIDS ($p \leq 6.0 \times 10^{-7}$), and others (see Table A10). Our analysis here further identified several previously unreported relationships that would not easily be found using the other methods we assessed. For example, of the top 500 relationships as ranked by $\text{MIC}_e/\text{TIC}_e$, 33 were ranked 1000th or worse by all eight of the other methods, including: a strongly nonlinear relationship whereby adult male mortality rate is much higher among countries with per capita oil consumption below a certain threshold ($p \leq 6.0 \times 10^{-7}$, rank by $\mathrm{MIC}_e/\mathrm{TIC}_e$: 209, best rank by any other statistic: 1510); a nonlinear but monotonic relationship between percent of the population below the poverty line and children per woman ($p \leq 6.0 \times 10^{-7}$, rank by $\mathrm{MIC}_e/\mathrm{TIC}_e$: 374, best rank by any other statistic: 1227); and a relationship between incidence of Ceasarian sections and government expenditure on health, in which there is a weak monotonic trend among most countries except for a small group of Northwestern European countries together with the United States that cluster away from the trend with a markedly higher expenditure on health ($p \leq 6.0 \times 10^{-7}$, rank by $\mathrm{MIC}_e/\mathrm{TIC}_e$: 464, best rank by any other statistic: 1129); For plots, see Figure A9. We emphasize that our goal here is to establish that the relationships ranked highly by $\mathrm{MIC}_e/\mathrm{TIC}_e$ are of interest, but this does not preclude other methods finding interesting relationships that are not as highly ranked by $\mathrm{MIC}_e\mathrm{TIC}_e$; in general, we expect that most methods will rank some interesting relationships highly that are not as highly ranked by other methods.

The analyses above suggest that (a) $\mathrm{MIC}_e/\mathrm{TIC}_e$ have a reduced preference for linear relationships, thus making finding nonlinear relationships easier, and (b) more generally, $\mathrm{MIC}_e/\mathrm{TIC}_e$ give high ranks to potentially interesting relationships that would not be found using other statistics. This motivates us to ask systematically whether $\mathrm{MIC}_e/\mathrm{TIC}_e$ are more different from the rest of the methods tested than those methods are from each other in terms of highly ranked relationships. To examine this, we compared every pair of methods using the Jaccard distance between the top 1000 relationships identified by each method. [The Jaccard distance is a metric on sets defined by $J(A, B) = 1 - |A \cap B|/|A \cup B|$.] We found that the top-ranked relationships by MIC$_e$/TIC$_e$ were the most different from those of the other statistics in that they had the highest average Jaccard distance from the top-ranked relationships of the other statistics (52.7%; Table 2). These results were robust to the number of top relationships examined (see Empirical Supplement 5A). Consistent with our nonlinearity analysis, HHG again came the closest in performance to $\mathrm{MIC}_e/\mathrm{TIC}_e$ among the statistics with extremely good power, with an average Jaccard distance of 48.7% to the rest of the statistics.

To gain a broader view of the behavior of the methods tested, we also created a dendrogram from these Jaccard distances using agglomerative hierarchical clustering. This recapitulated our findings, showing MIC$_e$/TIC$_e$ as the farthest away from any other single method. More generally, we believe it provides a valuable way to understand relationships between these measures of dependence. For instance, it shows distance correlation as similar to the squared Pearson correlation coefficient (in terms of relationship ranking, not power against independence), a fact that is consistent with our simulations. Additionally, the statistic closest to maximal correlation is RDC, which makes sense since RDC can be interpreted as an attempt to maximize correlation using linear combinations of random functions of the two

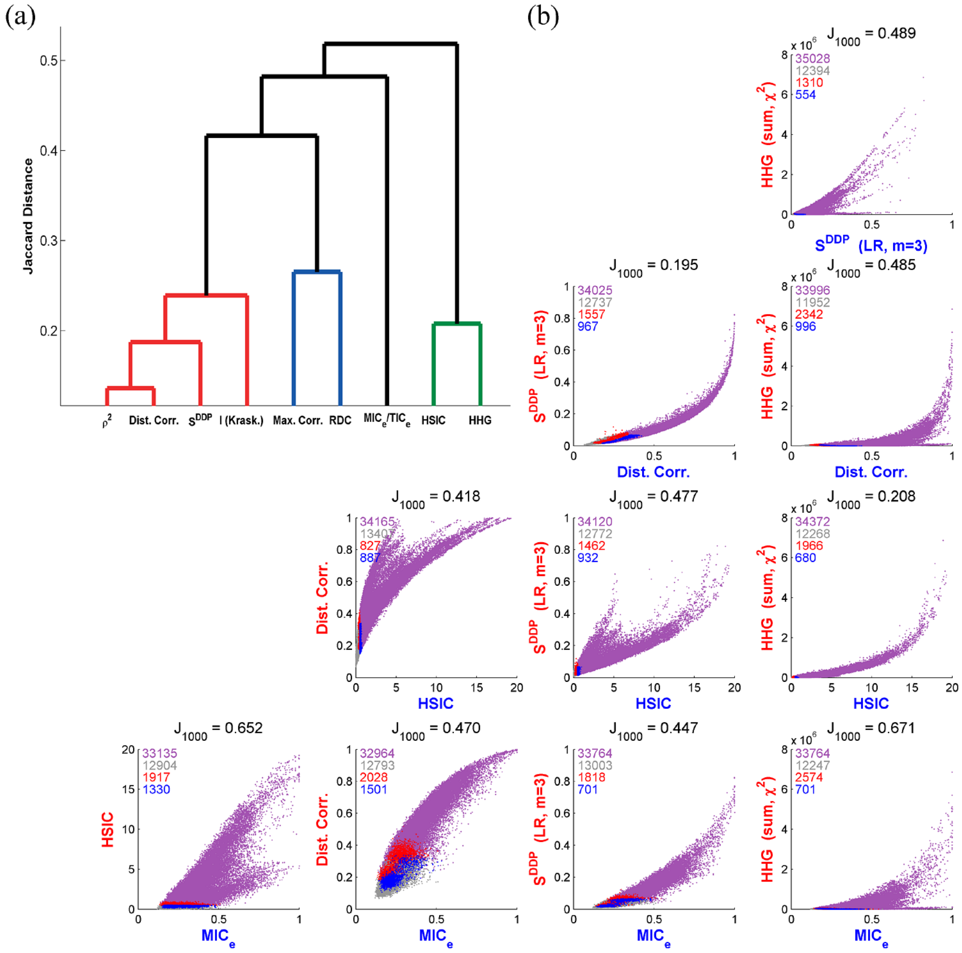

FIG. 9. *Pairwise comparison of the statistics on the WHO data set*. (a) *A dendrogram computed using the Jaccard distances between the top* 1000 *relationships identified by each statistic*. (b) *For each pair of statistics*, *a plot of one statistic's score against the other's across all the variable pairs in the data set*. *Purple*, *red/blue*, *and grey points denote relationships declared significant using both statistics*, *one statistic but not the other, and neither statistic*, *respectively*. *Numbers indicate number of dots of each respective color*. J_{1000} *indicates Jaccard distance between the top* 1000 *relationships ranked by the two statistics*. *For* $\mathrm{MIC}_e/\mathrm{TIC}_e$, *significance was determined using* TIC_e *and the plotted scores are the* MIC_e *scores*; *analogously*, *for mutual information estimation*, *significance was assessed using parameters optimized for independence testing and the plotted scores were computed using parameters optimized for equitability*.

variables in question. Finally, this dendrogram paired HSIC and HHG as similar, suggesting a hypothesis that there may be an as-yet uncharacterized aspect of dependence that these two statistics both capture.

We lastly plotted the scores of the five methods that detected the most relationships against each other for all the relationships in the data set. This is shown in Figure 9b; for all methods, see Empirical Supplement 5B.

9. Conclusion. In this paper, we presented an in-depth empirical evaluation of the equitability, power against independence, and runtime of several leading measures of dependence, including two new statistics introduced in Reshef et al. (2016). Our aims were to give an accessible exposition of equitability and its relationship to power against independence, provide the community with a comprehensive side-by-side comparison of existing methods, and evaluate the new statistics against the existing state of the art. Our main findings were as follows.

(1) *Equitability*. MIC_e, the estimator of the population MIC introduced in Reshef et al. (2016), generally has superior and more robust equitability with respect to R^2 than other measures of dependence. In some specific settings (models with no X noise and $n = 5000$), mutual information estimation achieves superior equitability in our experiments, but its equitability is otherwise highly variable and often poor, particularly at lower sample sizes. Maximal correlation achieves some degree of equitability over the models examined, but all other statistics tested have very poor equitability.

More generally, the analyses presented here demonstrate that equitability with respect to R^2 is achievable to a significant extent, at least on the relationships tested here. However, while the noise models, marginal distributions, and functions used were chosen to be representative of real-world relationships, they by no means form a large enough set to allow us to make claims about the performance of these methods in general. Given this state of affairs, a better theoretical understanding of MIC_e and also of equitability–with respect to R^2 and otherwise–is crucial for allowing us to determine when and to what extent equitability can be achieved.

(2) *Power against independence*. TIC_e and S^{DDP} had the best power against independence, outperforming each other by different metrics. Distance correlation, MIC_e, maximal correlation, HSIC, RDC, and HHG also had good power against independence. The power against independence of TIC_e and MIC_e was more robust than other methods to alternative hypothesis relationship type. When a different parameter setting from the equitability-oriented default is used, the original statistic MIC has substantially higher power against independence than has been reported in previous analyses.

(3) *Runtime.* MIC_e and TIC_e, each of which can be trivially computed once the other has been obtained, have runtimes that allow them to be run together even on large samples in reasonable time. This runtime compares favorably with that of other complex measures of dependence. The fastest measures of dependence were maximal correlation and the randomized dependence coefficient. There is a large variety of runtimes across the measures of dependence examined.

(4) *Power/equitability tradeoff.* The parameter α in the estimator MIC_e corresponds to a trade-off between power against independence and equitability that is

consistent with the characterization of equitability given in Reshef et al. (2015). Lower values of α lead to higher power against a null of independence at the expense of power against null hypotheses representing weak relationship strength (i.e., equitability), while higher values of α lead to better equitability at the expense of power against independence. Other parameterized methods display a similar trade-off.

(5) *Practical suggestions.* For exploration of data sets with unknown or potentially diverse relationship types, we recommend first using TIC_e to filter to only significant relationships, and then MIC_e to rank the relationships. This approach combines power, equitability, and speed, and performs well on the real data set we analyzed.

The fact that many measures of dependence performed similarly in our analysis of power against independence and had tens of thousands of rejections in our analyses of real data suggests that for some settings power against independence may not be where the true challenge lies, and that we ought to demand more of measures of dependence in those settings. Equitability is one attempt to formulate a more ambitious goal, as is the concept of low detection threshold introduced in Reshef et al. (2015) and discussed here, but there may well be other possibilities. Of course there are instances, such as detection of higher-dimensional relationships, in which even just power against independence is very difficult to achieve, and many of the methods evaluated here are quite useful in that setting.

The comprehensiveness of our results provides significant understanding of the comparative performance of various measures. To our knowledge, our analyses are the most exhaustive to date in that they evaluate a large swath of measures of dependence side-by-side along a number of dimensions (equitability, power against independence, and runtime); over a wide range of models, relationship types, and sample sizes; and with parameter sweeps for each individual statistic in each analysis. Our hope is that the full set of results, which are included in bulk in the empirical supplement, will be a resource to the community that facilitates a precise discussion of the trade-offs and assumptions associated with each measure of dependence in various settings.

As methodological work on measures of dependence continues, we expect and hope that methods with improved performance by each of the metrics assessed here will be developed, and already since the conclusion of this study there have been interesting and enlightening advances to note. For instance, an improved algorithm for estimating distance correlation is now known that runs much faster than the one benchmarked by us [Huo and Szekely (2014)]; the advances used in that algorithm could potentially be leveraged to improve other measures of dependence that rely on quantities computed between pairs of points. Similarly, a new measure of dependence called G^2 has recently been shown to achieve substantial levels of equitability [Wang, Jiang and Liu (2017)]. This method, like MIC_e, is partition-based and uses a dynamic programming algorithm to optimize the choice of partition, providing further evidence of the utility of these concepts as we try to understand what about MIC_e is essential to its performance and what is ancillary.

While the results presented here make a compelling case for the use of MIC_e and TIC_e and provide insight into the trade-offs between different measures of dependence, there are some important limitations for both the new statistics and the comparisons we performed. First, in this paper we evaluated only equitability with respect to R^2 on noisy functional relationships, whereas the definition we give of equitability explicitly acknowledges the possibility of using other properties of interest besides R^2 and standard relationships that are not noisy functional relationships. We feel that R^2 is an important measure of relationship strength that is intuitive and familiar to many practitioners, but equitability with respect to other properties of interest [see, e.g., Ding and Li (2013)] merits study as well, and the methods tested here may perform much better or worse when their equitability is evaluated with respect to other properties of interest.

We observe that more general versions of equitability can be considered without abandoning the notion of R^2 on noisy functional relationships. For example, we could add only *noiseless* versions of nonfunctional relationships, such as a circle, to our existing set of standard relationships, and then define the property of interest to equal 1 on those relationships. This has the virtue of encoding a strong intuition about the importance of nonfunctional relationships without requiring a stringent assumption about exactly how *noisy* nonfunctional relationships should be scored. Since the original motivation for the maximal information coefficient stems from its ability to detect nonfunctional relationships as well, assessing equitability with respect to a criterion such as this one is an interesting avenue of future inquiry.

There are other classes of relationships to consider from the perspective of statistical power as well. For instance, we assessed power primarily on functional relationships with noise added uniformly to the distribution in question. However, one family of relationships that may exhibit qualitatively different behavior is relationships with local dependence, for which the performance of aggregative methods such as TIC_e and S^{DDP} may be quite different.

An additional limitation of the present work is that, though an attempt at comprehensiveness was made, we did limit our scope to the set of noisy functional relationships in Reshef et al. (2011) for equitability and the relationships introduced in Simon and Tibshirani (2012) for power against independence, along with corresponding randomly chosen relationships in each setting. While we feel each of these suites of relationships provides reasonable insight into the performance of the methods in question on a broad set of realistic relationship types, there do, for instance, exist relationships, such as a step function, that when added to these suites provably result in poor equitability for all the methods tested (see Appendix J), and we believe that the same is true for the power analyses. Characterizing those relationships theoretically and empirically in the settings of both equitability and power against independence is vital for fully understanding the strengths and weaknesses of each of these methods. This is an important direction for future work for which the analyses of random functions in Reshef et al. (2016) and here are only a first step. We note that as we try to understand what constitutes an appropriate set of standard relationships, it would be useful not just to better characterize performance of various sets, but also to have a way of evaluating the extent to which a given set of standard relationships “matches” a real data set that is being analyzed. Such a metric would provide valuable empirical guidance to this avenue of investigation.

Measures of dependence are useful in a variety of settings and identifying which measures of dependence provide superior performance in the face of different objectives, assumptions, and constraints is critical. For each separate goal, we must understand both which measure of dependence is most appropriate and also which parameter regimes lead to the best performance. Such an understanding provides insight into the inherent trade-offs of different methods, allowing us to navigate the landscape of measures of dependence effectively and-ultimately-to better understand our data.

Acknowledgments. The authors would like to acknowledge R. Adams, E. Airoldi, K. Arnold, H. Finucane, A. Gelman, M. Gorfine, A. Gretton, T. Hashimoto, R. Heller, J. Hernández-Lobato, J. Huggins, T. Jaakkola, A. Miller, J. Mueller, A. Narayanan, G. Szekely, J. Tenenbaum, and R. Tibshirani for constructive conversations and useful feedback. The authors would also like to acknowledge the anonymous referees for their useful comments.

The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

Software implementation of MIC_e and TIC_e is available at http://exploredata.net.

SUPPLEMENTARY MATERIAL

Appendix: Supplementary methods and figures (DOI: 10.1214/17-AOAS1093SUPPA; .pdf). Details of the methods, parameter choices, and supplemental figures referenced in the main text.

Empirical Supplement: Full results of all analyses (DOI: 10.1214/17-AOAS1093SUPPB; .zip). The full set of results for all analyses presented, as well as additional, complementary analyses.

REFERENCES

Algeo, T. J. and Lyons, T. W. (2006). Mo–total organic carbon covariation in modern anoxic marine environments: Implications for analysis of paleoredox and paleohydrographic conditions. *Paleoceanography* **21** PA1016.

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CASPI, A., SUGDEN, K., MOFFITT, T. E., TAYLOR, A., CRAIG, I. W., HARRINGTON, H., MCCLAY, J., MILL, J., MARTIN, J., BRAITHWAITE, A. and POULTON, R. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science* **301** 386–389.

CLAYTON, R. N. and MAYEDA, T. K. (1996). Oxygen isotope studies of achondrites. *Geochim. Cosmochim. Acta* **60** 1999–2017.

DING, A. A. and LI, Y. (2013). Copula correlation: An equitable dependence measure and extension of pearson's correlation. Preprint. Available at arXiv:1312.7214.

EMILSSON, V., THORLEIFSSON, G., ZHANG, B., LEONARDSON, A. S., ZINK, F., ZHU, J., CARLSON, S., HELGASON, A., BRAGI WALTERS, G., GUNNARSDOTTIR, S. et al. (2008). Genetics of gene expression and its effect on disease. *Nature* **452** 423–428.

GILL, T. ET AL. (2002). Obesity in the pacific: Too big to ignore. World Health Organization Regional Office for the Western Pacific, Secretariat of the Pacific Community.

GORFINE, M., HELLER, R. and HELLER, Y. (2012). Comment on "Detecting novel associations in large data sets." Unpublished. Available at http://www.math.tau.ac.il/~ruheller/Papers/science6.pdf.

GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory* 63–77. Springer, Berlin.

GRETTON, A., FUKUMIZU, K., TEO, C. H., LE, S., SCHÖLKOPF, B. and SMOLA, A. J. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* 585–592.

HELLER, R., HELLER, Y. and GORFINE, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100** 503–510. MR3068450

HELLER, R., HELLER, Y., KAUFMAN, S., BRILL, B. and GORFINE, M. (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *J. Mach. Learn. Res.* **17** 1–54.

HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* 546–557.

HUO, X. and SZEKELY, G. J. (2014). Fast computing for distance covariance. Preprint. Available at arXiv:1410.1503.

JAAKKOLA, T. S. and HAUSSLER, D. (1999). Probabilistic kernel regression models. In *AISTATS*.

JIANG, B., YE, C. and LIU, J. S. (2015). Nonparametric k-sample tests via dynamic slicing. *J. Amer. Statist. Assoc.* **110** 642–653.

KINNEY, J. B. and ATWAL, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. USA* **111** 3354–3359. MR3200177

KRASKOV, A., STOGBAUER, H. and GRASSBERGER, P. (2004). Estimating mutual information. *Phys. Rev. E* **69** 066138.

LINFOOT, E. H. (1957). An informational measure of correlation. *Inf. Control* **1** 85–89.

LOPEZ-PAZ, D., HENNIG, P. and SCHÖLKOPF, B. (2013). The randomized dependence coefficient. In *Advances in Neural Information Processing Systems* 1–9.

MOON, Y.-I., RAJAGOPALAN, B. and LALL, U. (1995). Estimation of mutual information using kernel density estimators. *Phys. Rev. E* **52** 2318–2321.

MURRELL, B., MURRELL, D. and MURRELL, H. (2014). R2-equitability is satisfiable. *Proc. Natl. Acad. Sci. USA* **111** E2160–E2160. Available at http://www.pnas.org/content/early/2014/04/29/1403623111.

PANINSKI, L. (2003). Estimation of entropy and mutual information. *Neural Comput.* **15** 1191–1253.

RÉNYI, A. (1959). On measures of dependence. *Acta Math. Hungar.* **10** 441–451.

RESHEF, D. N., RESHEF, Y. A., SABETI, P. C. and MITZENMACHER, M. (2018a). Appendix to "An empirical study of the maximal and total information coefficients and leading measures of dependence." DOI:10.1214/17-AOAS1093SUPPA.

RESHEF, D. N., RESHEF, Y. A., SABETI, P. C. and MITZENMACHER, M. (2018b). Supplement to "An empirical study of the maximal and total information coefficients and leading measures of dependence." DOI:10.1214/17-AOAS1093SUPPB.

RESHEF, D. N., RESHEF, Y. A., FINUCANE, H. K., GROSSMAN, S. R., MCVEAN, G., TURNBAUGH, P. J., LANDER, E. S., MITZENMACHER, M. and SABETI, P. C. (2011). Detecting novel associations in large data sets. *Science* **334** 1518–1524.

RESHEF, D., RESHEF, Y., MITZENMACHER, M. and SABETI, P. (2013). Equitability analysis of the maximal information coefficient, with comparisons. Preprint. Available at arXiv:1301.6314.

RESHEF, D. N., RESHEF, Y. A., MITZENMACHER, M. and SABETI, P. C. (2014). Cleaning up the record on the maximal information coefficient and equitability. *Proc. Natl. Acad. Sci. USA* **111** E3362–E3363. Available at http://www.pnas.org/content/early/2014/08/07/1408920111.

RESHEF, Y. A., RESHEF, D. N., SABETI, P. C. and MITZENMACHER, M. (2015). Equitability, interval estimation, and statistical power. Available at arXiv:1505.02212.

RESHEF, Y. A., RESHEF, D. N., FINUCANE, H. K., SABETI, P. C. and MITZENMACHER, M. (2016). Measuring dependence powerfully and equitably. *J. Mach. Learn. Res.* **17** Paper No. 212, 63. MR3595146

SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. and FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41** 2263–2291. MR3127866

SIMON, N. and TIBSHIRANI, R. (2012). Comment on "Detecting novel associations in large data sets". Unpublished. Available at http://statweb.stanford.edu/ tibs/reshef/comment.pdf.

SPEED, T. (2011). A correlation for the 21st century. *Science* **334** 1502–1503.

SZEKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

WANG, X., JIANG, B. and LIU, J. S. (2017). Generalized R-squared for detecting dependence. *Biometrika* **104** 129–139. MR3626486

D. N. RESHEF
DEPARTMENT OF COMPUTER SCIENCE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: dnreshef@mit.edu

Y. A. RESHEF
M. MITZENMACHER
SCHOOL OF ENGINEERING AND APPLIED SCIENCES
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: yakir@seas.harvard.edu
michaelm@eecs.harvard.edu

P. C. SABETI
DEPARTMENT OF ORGANISMIC AND EVOLUTIONARY BIOLOGY
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138
USA
E-MAIL: pardis@broadinstitute.org