Equitability, mutual information, and the maximal information coefficient

Justin B. Kinney[1] and Gurinder S. Atwal

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

Edited* by David L. Donoho, Stanford University, Stanford, CA, and approved January 21, 2014 (received for review May 24, 2013)

How should one quantify the strength of association between two random variables without bias for relationships of a specific form? Despite its conceptual simplicity, this notion of statistical "equitability" has yet to receive a definitive mathematical formalization. Here we argue that equitability is properly formalized by a self-consistency condition closely related to Data Processing Inequality. Mutual information, a fundamental quantity in information theory, is shown to satisfy this equitability criterion. These findings are at odds with the recent work of Reshef et al. [Reshef DN, et al. (2011) *Science* 334(6062):1518–1524], which proposed an alternative definition of equitability and introduced a new statistic, the "maximal information coefficient" (MIC), said to satisfy equitability in contradistinction to mutual information. These conclusions, however, were supported only with limited simulation evidence, not with mathematical arguments. Upon revisiting these claims, we prove that the mathematical definition of equitability proposed by Reshef et al. cannot be satisfied by any (nontrivial) dependence measure. We also identify artifacts in the reported simulation evidence. When these artifacts are removed, estimates of mutual information are found to be more equitable than estimates of MIC. Mutual information is also observed to have consistently higher statistical power than MIC. We conclude that estimating mutual information provides a natural (and often practical) way to equitably quantify statistical associations in large datasets.

This paper addresses a basic yet unresolved issue in statistics: How should one quantify, from finite data, the association between two continuous variables? Consider the squared Pearson correlation R^2. This statistic is the standard measure of dependence used throughout science and industry. It provides a powerful and meaningful way to quantify dependence when two variables share a linear relationship exhibiting homogenous Gaussian noise. However, as is well known, R^2 values often correlate badly with one's intuitive notion of dependence when relationships are highly nonlinear.

Fig. 1 provides an example of how R^2 can fail to sensibly quantify associations. Fig. 1*A* shows a simulated dataset, representing a noisy monotonic relationship between two variables x and y. This yields a substantial R^2 measure of dependence. However, the R^2 value computed for the nonmonotonic relationship in Fig. 1*B* is not significantly different from zero even though the two relationships shown in Fig. 1 are equally noisy.

It is therefore natural to ask whether one can measure statistical dependencies in a way that assigns "similar scores to equally noisy relationships of different types." This heuristic criterion has been termed "equitability" by Reshef et al. (1, 2), and its importance for the analysis of real-world data has been emphasized by others (3, 4). It has remained unclear, however, how equitability should be defined mathematically. As a result, no dependence measure has yet been proved to have this property.

Here we argue that the heuristic notion of equitability is properly formalized by a self-consistency condition that we call "self-equitability." This criterion arises naturally as a weakened form of the well-known Data Processing Inequality (DPI). All DPI-satisfying dependence measures are thus proved to satisfy self-equitability. Foremost among these is "mutual information," a quantity of central importance in information theory (5, 6). Indeed, mutual information is already widely believed to quantify dependencies without bias for relationships of one type or another. And although it was proposed in the context of modeling communications systems, mutual information has been repeatedly shown to arise naturally in a variety of statistical problems (6–8).

The use of mutual information for quantifying associations in continuous data is unfortunately complicated by the fact that it requires an estimate (explicit or implicit) of the probability distribution underlying the data. How to compute such an estimate that does not bias the resulting mutual information value remains an open problem, one that is particularly acute in the undersampled regime (9, 10). Despite these difficulties, a variety of practical estimation techniques have been developed and tested (11, 12). Indeed, mutual information is now routinely computed on continuous data in many real-world applications (e.g., refs. 13–17).

Unlike R^2, the mutual information values I of the underlying relationships in Fig. 1 *A* and *B* are identical (0.72 bits). This is a consequence of the self-equitability of mutual information. Applying the kth nearest-neighbor (KNN) mutual information estimation algorithm of Kraskov et al. (18) to simulated data drawn from these relationships, we see that the estimated mutual information values agree well with the true underlying values.

However, Reshef et al. claim in their paper (1) that mutual information does not satisfy the heuristic notion of equitability. After formalizing this notion, the authors also introduce a new statistic called the "maximal information coefficient" (MIC), which, they claim, does satisfy their equitability criterion. These results are perhaps surprising, considering that MIC is actually defined as a normalized estimate of mutual information. However, no mathematical arguments were offered for these assertions; they were based solely on the analysis of simulated data.

Here we revisit these claims. First, we prove that the definition of equitability proposed by Reshef et al. is, in fact, impossible for

Significance

Attention has recently focused on a basic yet unresolved problem in statistics: How can one quantify the strength of a statistical association between two variables without bias for relationships of a specific form? Here we propose a way of mathematically formalizing this "equitability" criterion, using core concepts from information theory. This criterion is naturally satisfied by a fundamental information-theoretic measure of dependence called "mutual information." By contrast, a recently introduced dependence measure called the "maximal information coefficient" is seen to violate equitability. We conclude that estimating mutual information provides a natural and practical method for equitably quantifying associations in large datasets.

Author contributions: J.B.K. and G.S.A. designed research, performed research, and wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

Data deposition: All analysis code reported in this paper have been deposited in the SourceForge database at https://sourceforge.net/projects/equitability/.

[1]To whom correspondence should be addressed. E-mail: jkinney@cshl.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1309933111/-/DCSupplemental.

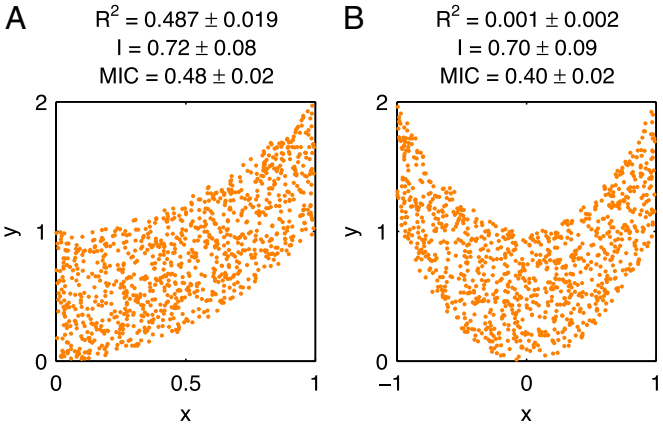

Fig. 1. Illustration of equitability. (*A* and *B*) $N = 1{,}000$ data points simulated for two noisy functional relationships that have the same noise profile but different underlying functions. (*Upper*) Mean ± SD values, computed over 100 replicates, for three statistics: Pearson's R^2, mutual information I (in bits), and MIC. Mutual information was estimated using the KNN algorithm (18) with $k = 1$. The specific relationships simulated are both of the form $y = x^2 + \frac{1}{2} + \eta$, where η is noise drawn uniformly from $(-0.5, 0.5)$ and x is drawn uniformly from one of two intervals, (*A*) $(0,1)$ or (*B*) $(-1,1)$. Both relationships have the same underlying mutual information (0.72 bits).

any (nontrivial) dependence measure to satisfy. MIC is then shown by example to violate various intuitive notions of dependence, including DPI and self-equitability. Upon revisiting the simulations of Reshef et al. (1), we find the evidence offered in support of their claims about equitability to be artifactual. Indeed, random variations in the MIC estimates of ref. 1, which resulted from the small size of the simulated datasets used, are seen to have obscured the inherently nonequitable behavior of MIC. When moderately larger datasets are used, it becomes clear that nonmonotonic relationships have systematically reduced MIC values relative to monotonic ones. The MIC values computed for the relationships in Fig. 1 illustrate this bias. We also find that the nonequitable behavior reported for mutual information by Reshef et al. does not reflect inherent properties of mutual information, but rather resulted from the use of a nonoptimal value for the parameter k in the KNN algorithm of Kraskov et al. (18).

Finally we investigate the power of MIC, the KNN mutual information estimator, and other measures of bivariate dependence. Although the power of MIC was not discussed by Reshef et al. (1), this issue is critical for the kinds of applications described in their paper. Here we find that, when an appropriate value of k is used, KNN estimates of mutual information consistently outperform MIC in tests of statistical power. However, we caution that other nonequitable measures such as "distance correlation" (dCor) (19) and Hoeffding's D (20) may prove to be more powerful on some real-world datasets than the KNN estimator.

In the text that follows, uppercase letters $(X, Y, \ldots)$ are used to denote random variables, lowercase letters $(x, y, \ldots)$ denote specific values for these variables, and tildes $(\tilde{x}, \tilde{y}, \ldots)$ signify bins into which these values fall when histogrammed. A "dependence measure," written $D[X;Y]$, refers to a function of the joint probability distribution $p(X,Y)$, whereas a "dependence statistic," written $D\{x;y\}$, refers to a function computed from finite data $\{x_i, y_i\}_{i=1}^N$ that has been sampled from $p(X,Y)$.

Results

R^2-Equitability. In their paper, Reshef et al. (1) suggest the following definition of equitability. This makes use of the squared Pearson correlation measure $R^2[\cdot]$, so for clarity we call this criterion "R^2-equitability."

Definition 1. *A dependence measure $D[X;Y]$ is R^2-equitable if and only if, when evaluated on a joint probability distribution $p(X,Y)$ that corresponds to a noisy functional relationship between two real random variables X and Y, the following relation holds*:

$$D[X;Y] = g\big(R^2[f(X);Y]\big). \qquad [1]$$

Here, g is a function that does not depend on $p(X,Y)$ and f is the function defining the noisy functional relationship, i.e.,

$$Y = f(X) + \eta, \qquad [2]$$

for some random variable η. The noise term η may depend on $f(X)$ as long as η has no additional dependence on X, i.e., as long as $X \leftrightarrow f(X) \leftrightarrow \eta$ is a Markov chain.[†]

Heuristically this means that, by computing the measure $D[X;Y]$ from knowledge of $p(X,Y)$, one can discern the strength of the noise η, as quantified by $1 - R^2[f(X);Y]$, without knowing the underlying function f. Of course this definition depends strongly on what properties the noise η is allowed to have. In their simulations, Reshef et al. (1) considered only uniform homoscedastic noise: η was drawn uniformly from some symmetric interval $[-a, a]$. Here we consider a much broader class of heteroscedastic noise: η may depend arbitrarily on $f(X)$, and $p(\eta \mid f(X))$ may have arbitrary functional form.

Our first result is this: No nontrivial dependence measure can satisfy R^2-equitability. This is due to the fact that the function f in Eq. **2** is not uniquely specified by $p(X,Y)$. For example, consider the simple relationship $Y = X + \eta$. For every invertible function h there also exists a valid noise term ξ such that $Y = h(X) + \xi$ (*SI Text*, *Theorem 1*). R^2-equitability then requires $D[X;Y] = g(R^2[X;Y]) = g(R^2[h(X);Y])$. However, $R^2[X;Y]$ is not invariant under invertible transformations of X. The function g must therefore be constant, implying that $D[X;Y]$ does not depend on $p(X,Y)$ and is therefore trivial.

Self-Equitability and Data Processing Inequality. Because R^2-equitability cannot be satisfied by any (interesting) dependence measure, it cannot be adopted as a useful mathematical formalization of Reshef et al.'s heuristic (1). Instead we propose formalizing the notion of equitability as an invariance property we term self-equitability, which is defined as follows.

Definition 2. *A dependence measure $D[X;Y]$ is self-equitable if and only if it is symmetric $(D[X;Y] = D[Y;X])$ and satisfies*

$$D[X;Y] = D[f(X);Y], \qquad [3]$$

whenever f is a deterministic function, X and Y are variables of any type, and $X \leftrightarrow f(X) \leftrightarrow Y$ forms a Markov chain.

The intuition behind this definition is similar to that behind Eq. **1**, but instead of using R^2 to quantify the noise in the relationship we use D itself. An important advantage of this definition is that the Y variable can be of any type, e.g., categorical, multidimensional, or non-Abelian. By contrast, the definition of R^2-equitability requires that Y and $f(X)$ must be real numbers.

Self-equitability also employs a more general definition of "noisy relationship" than does R^2-equitability: Instead of positing additive noise as in Eq. **2**, one simply assumes that Y depends on X only through the value of $f(X)$. This is formalized by the Markov chain condition $X \leftrightarrow f(X) \leftrightarrow Y$. As a result, any self-equitable measure $D[X;Y]$ must be invariant under arbitrary invertible transformations of X or Y (*SI Text*, *Theorem 2*). Self-equitability also has a close connection to DPI, a fundamental criterion in information theory (6) that we briefly restate here.

Definition 3. *A dependence measure $D[X;Y]$ satisfies DPI if and only if*

[†]The Markov chain condition $X \leftrightarrow f(X) \leftrightarrow \eta$ means that $p(\eta \mid f(X), X) = p(\eta \mid f(X))$. Chapter 2 of ref. 6 gives a good introduction to Markov chains relevant to this discussion.

$$D[X;Z] \le D[Y;Z], \quad [4]$$

whenever the random variables X, Y, Z form a Markov chain $X \leftrightarrow Y \leftrightarrow Z$.

DPI formalizes our intuitive notion that information is generally lost, and is never gained, when transmitted through a noisy communications channel. For instance, consider a game of telephone involving three children, and let the variables X, Y, and Z represent the words spoken by the first, the second, and the third child, respectively. The criterion in Eq. **4** is satisfied only if the measure D upholds our intuition that the words spoken by the third child will be more strongly dependent on those said by the second child (as quantified by $D[Y;Z]$) than on those said by the first child (quantified by $D[X;Z]$).

It is readily shown that all DPI-satisfying dependence measures are self-equitable (*SI Text*, *Theorem 3*). Moreover, many dependence measures do satisfy DPI (*SI Text*, *Theorem 4*). This begs the question of whether there are any self-equitable measures that do not satisfy DPI. The answer is technically "yes": For example, if $D[X;Y]$ satisfies DPI, then a new measure defined as $D'[X;Y] = -D[X;Y]$ will be self-equitable but will not satisfy DPI. However, DPI enforces an important heuristic that self-equitability does not, namely that adding noise should not increase the strength of a dependency. So although self-equitable measures that violate DPI do exist, there is good reason to require that sensible measures also satisfy DPI.

Mutual Information. Among DPI-satisfying dependence measures, mutual information is particularly meaningful. Mutual information rigorously quantifies, in units known as "bits," how much information the value of one variable reveals about the value of another. This has important and well-known consequences in information theory (6). Perhaps less well known, however, is the natural role that mutual information plays in the statistical analysis of data, a topic we now touch upon briefly.

The mutual information between two random variables X and Y is defined in terms of their joint probability distribution $p(X,Y)$ as

$$I[X;Y] = \int dx\, dy\, p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}. \quad [5]$$

$I[X;Y]$ is always nonnegative and $I[X;Y] = 0$ only when $p(X,Y) = p(X)p(Y)$. Thus, mutual information will be greater than zero when X and Y exhibit any mutual dependence, regardless of how nonlinear that dependence is. Moreover, the stronger the mutual dependence is, the larger the value of $I[X;Y]$. In the limit where Y is a (nonconstant) deterministic function of X (over a continuous domain), $I[X;Y] = \infty$.

Mutual information is intimately connected to the statistical problem of detecting dependencies. From Eq. **5** we see that, for data drawn from the distribution $p(X,Y)$, $I[X;Y]$ quantifies the expected per-datum log-likelihood ratio of the data coming from $p(X,Y)$ as opposed to $p(X)p(Y)$. Thus, $1/I[X;Y]$ is the typical amount of data one needs to collect to get a twofold increase in the posterior probability of the true hypothesis relative to the null hypothesis [i.e., that $p(X,Y) = p(X)p(Y)$]. Moreover, the Neyman–Pearson lemma (21) tells us that this log-likelihood ratio, $\sum_i \log_2[p(x_i,y_i)/p(x_i)p(y_i)]$, has the maximal possible statistical power for such a test. The mutual information $I[X;Y]$ therefore provides a tight upper bound on how well any test of dependence can perform on data drawn from $p(X,Y)$.

Accurately estimating mutual information from finite continuous data, however, is nontrivial. The difficulty lies in estimating the joint distribution $p(X,Y)$ from a finite sample of N data points $\{x_i, y_i\}_{i=1}^N$. The simplest approach is to "bin" the data—to superimpose a rectangular grid on the x,y scatter plot and then assign each continuous x value (or y value) to the column bin $\tilde{x}$ (or row bin $\tilde{y}$) into which it falls. Mutual information can then be estimated from the data as

$$I_{\text{naive}}\{x;y\} = \sum_{\tilde{x},\tilde{y}} \hat{p}(\tilde{x},\tilde{y}) \log_2 \frac{\hat{p}(\tilde{x},\tilde{y})}{\hat{p}(\tilde{x})\hat{p}(\tilde{y})}, \quad [6]$$

where $\hat{p}(\tilde{x},\tilde{y})$ is the fraction of data points falling into bin $(\tilde{x},\tilde{y})$. Estimates of mutual information that rely on this simple binning procedure are commonly called "naive" estimates (22). The problem with such naive estimates is that they systematically overestimate $I[X;Y]$. As was mentioned above, this has long been recognized as a problem and significant attention has been devoted to developing alternative methods that do not systematically overestimate mutual information. We emphasize, however, that the problem of estimating mutual information becomes easy in the large data limit, because $p(X,Y)$ can be determined to arbitrary accuracy as $N \to \infty$.

The Maximal Information Coefficient. In contrast to mutual information, Reshef et al. (1) define MIC as a statistic, not as a dependence measure. At the heart of this definition is a naive mutual information estimate $I_{\text{MIC}}\{x;y\}$ computed using a data-dependent binning scheme. Let n_X and n_Y, respectively, denote the number of bins imposed on the x and y axes. The MIC binning scheme is chosen so that (*i*) the total number of bins $n_X n_Y$ does not exceed some user-specified value B and (*ii*) the value of the ratio

$$MIC\{x;y\} = \frac{I_{\text{MIC}}\{x;y\}}{Z_{\text{MIC}}}, \quad [7]$$

where $Z_{\text{MIC}} = \log_2(\min(n_X, n_Y))$, is maximized. The ratio in Eq. **7**, computed using this data-dependent binning scheme, is how MIC is defined. Note that, because I_{MIC} is bounded above by Z_{MIC}, MIC values will always fall between 0 and 1. We note that $B = N^{0.6}$ (1) and $B = N^{0.55}$ (2) have been advocated, although no mathematical rationale for these choices has been presented.

In essence the MIC statistic $MIC\{x;y\}$ is defined as a naive mutual information estimate $I_{\text{MIC}}\{x;y\}$, computed using a constrained adaptive binning scheme and divided by a data-dependent normalization factor Z_{MIC}. However, in practice this statistic often cannot be computed exactly because the definition of MIC requires a maximization step over all possible binning schemes, a computationally intractable problem even for modestly sized datasets. Rather, a computational estimate of MIC is typically required. Except where noted otherwise, MIC values reported in this paper were computed using the software provided by Reshef et al. (1).

Note that when only two bins are used on either the x or the y axis in the MIC binning scheme, $Z_{\text{MIC}} = 1$. In such cases the MIC statistic is identical to the underlying mutual information estimate I_{MIC}. We point this out because a large majority of the MIC computations reported below produced $Z_{\text{MIC}} = 1$. Indeed it appears that, except for highly structured relationships, MIC typically reduces to the naive mutual information estimate I_{MIC} (*SI Text*).[‡]

Analytic Examples. To illustrate the differing properties of mutual information and MIC, we first compare the exact behavior of these dependence measures on simple example relationships $p(X,Y)$.[§] We begin by noting that MIC is completely insensitive to certain types of noise. This is illustrated in Fig. 2 *A–C*, which provides examples of how adding noise at all values of X will decrease $I[X;Y]$ but not necessarily decrease $MIC[X;Y]$. This pathological behavior results from the binning scheme used in

[‡]As of this writing, code for the MIC estimation software described by Reshef et al. in ref. 1 has not been made public. We were therefore unable to extract the I_{MIC} values computed by this software. Instead, I_{MIC} values were extracted from the open-source MIC estimator of Albanese et al. (23).

[§]Here we define the dependence measure $MIC[X;Y]$ as the value of the statistic $MIC\{x;y\}$ in the $N \to \infty$ limit.

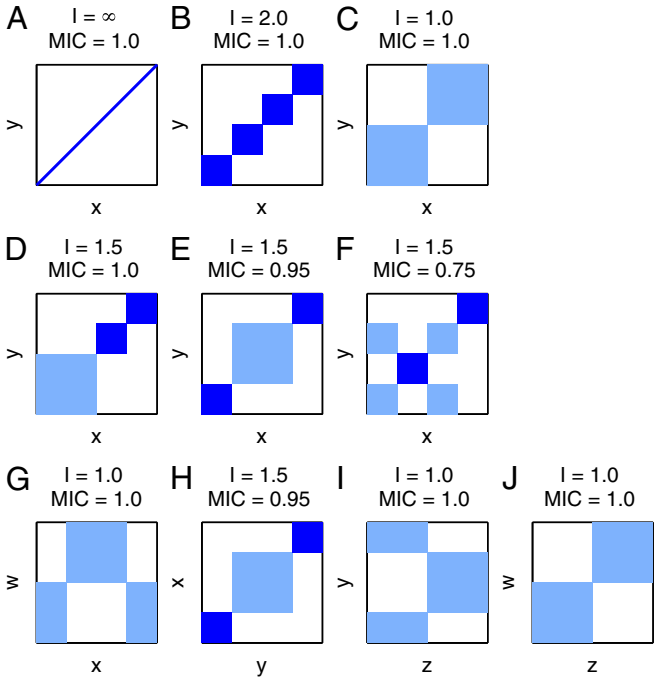

Fig. 2. MIC violates multiple notions of dependence that mutual information upholds. (*A–J*) Example relationships between two variables with indicated mutual information values (*I*, shown in bits) and MIC values. These values were computed analytically and checked using simulated data (Fig. S1). Dark blue blocks represent twice the probability density of light blue blocks. (*A–C*) Adding noise everywhere to the relationship in *A* diminishes mutual information but not necessarily MIC. (*D–F*) Relationships related by invertible nonmonotonic transformations of *X* and *Y*. Mutual information is invariant under these transformations but MIC is not. (*G–J*) Convolving the relationships shown in *G–I* along the chain $W \leftrightarrow X \leftrightarrow Y \leftrightarrow Z$ produces the relationship shown in *J*. In this case MIC violates DPI because $MIC[W;Z] > MIC[X;Y]$. Mutual information satisfies DPI here because $I[W;Z] < I[X;Y]$.

the definition of MIC: If all data points can be partitioned into two opposing quadrants of a 2×2 grid (half the data in each), a relationship will be assigned $MIC[X;Y] = 1$ regardless of the structure of the data within the two quadrants. Mutual information, by contrast, has no such limitations on its resolution.

Furthermore, $MIC[X;Y]$ is not invariant under nonmonotonic transformations of X or Y. Mutual information, by contrast, is invariant under such transformations. This is illustrated in Fig. 2 *D–F*. Such reparameterization invariance is a necessary attribute of any dependence measure that satisfies self-equitability or DPI (*SI Text*, *Theorem 2*). Fig. 2 *G–J* provides an explicit example of how the noninvariance of MIC causes DPI to be violated, whereas Fig. S2 shows how noninvariance can lead to violation of self-equitability.

Equitability Tests Using Simulated Data. The key claim made by Reshef et al. (1) in arguing for the use of MIC as a dependence measure has two parts. First, MIC is said to satisfy not just the heuristic notion of equitability, but also the mathematical criterion of R^2-equitability (Eq. **1**). Second, Reshef et al. (1) argue that mutual information does not satisfy R^2-equitability. In essence, the central claim made in ref. 1 is that the binning scheme and normalization procedure that transform mutual information into MIC are necessary for equitability. As mentioned in the Introduction, however, no mathematical arguments were made for these claims; these assertions were supported entirely through the analysis of limited simulated data.

We now revisit this simulation evidence. To argue that MIC is R^2-equitable, Reshef et al. simulated data for various noisy functional relationships of the form $Y = f(X) + \eta$. A total of 250, 500, or 1,000 data points were generated for each dataset; see Table S1 for details. $MIC\{x;y\}$ was computed for each data set and was plotted against $1 - R^2\{f(x);y\}$, which was used to quantify the inherent noise in each simulation.

Were MIC to satisfy R^2-equitability, plots of MIC against this measure of noise would fall along the same curve regardless of the function f used for each relationship. At first glance Fig. 3*A*, which is a reproduction of figure 2B of ref. 1, suggests that this may be the case. These MIC values exhibit some dispersion, of course, but this is presumed in ref. 1 to result from the finite size of the simulated datasets, not any inherent f-dependent bias of MIC.

However, as Fig. 3*B* shows, substantial f-dependent bias in the values of MIC become evident when the number of simulated data points is increased to 5,000. This bias is particularly strong for noise values between 0.6 and 0.8. To understand the source

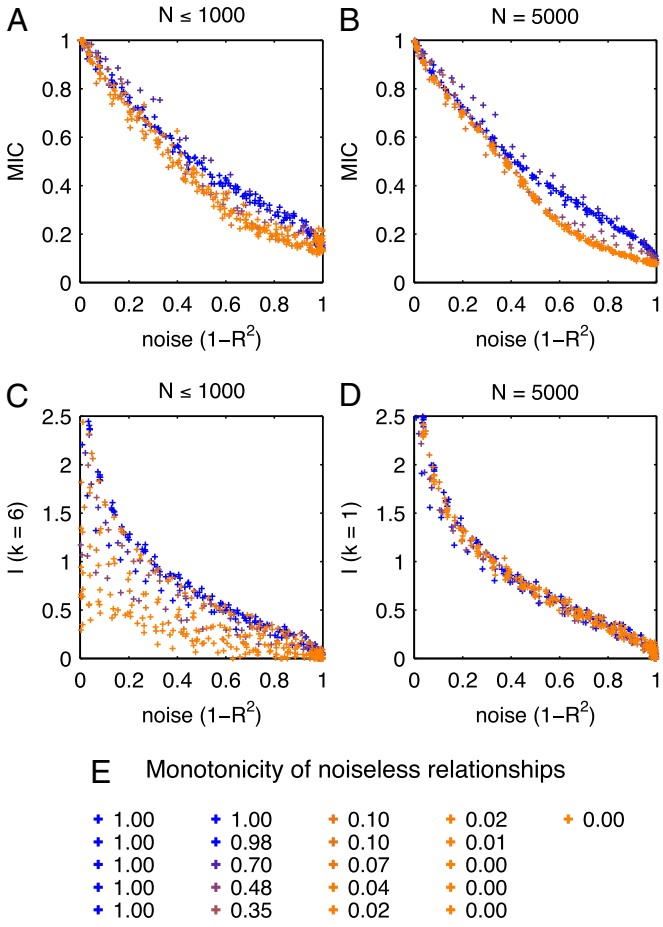

Fig. 3. Reexamination of the R^2-equitability tests reported by Reshef et al. (1). MIC values and mutual information values were computed for datasets simulated as described in figure 2 B–F of ref. 1. Specifically, each simulated relationship is of the form $Y = f(X) + \eta$. Twenty-one different functions f and twenty-four different amplitudes for the noise η were used. Details are provided in Table S1. MIC and mutual information values are plotted against the inherent noise in each relationship, as quantified by $1 - R^2\{f(x);y\}$. (*A*) Reproduction of figure 2B of ref. 1. $MIC\{x;y\}$ was calculated on datasets comprising 250, 500, or 1,000 data points, depending on f. (*B*) Same as *A* but using datasets comprising 5,000 data points each. (*C*) Reproduction of figure 2D of ref. 1. Mutual information values $I\{x;y\}$ were computed (in bits) on the datasets from *A*, using the KNN estimator with smoothing parameter $k = 6$. (*D*) KNN estimates of mutual information, made using $k = 1$, computed for the datasets from *B*. (*E*) Each point plotted in *A–D* is colored (as indicated here) according to the monotonicity of f, which is quantified using the squared Spearman rank correlation between X and $f(X)$ (Fig. S3).

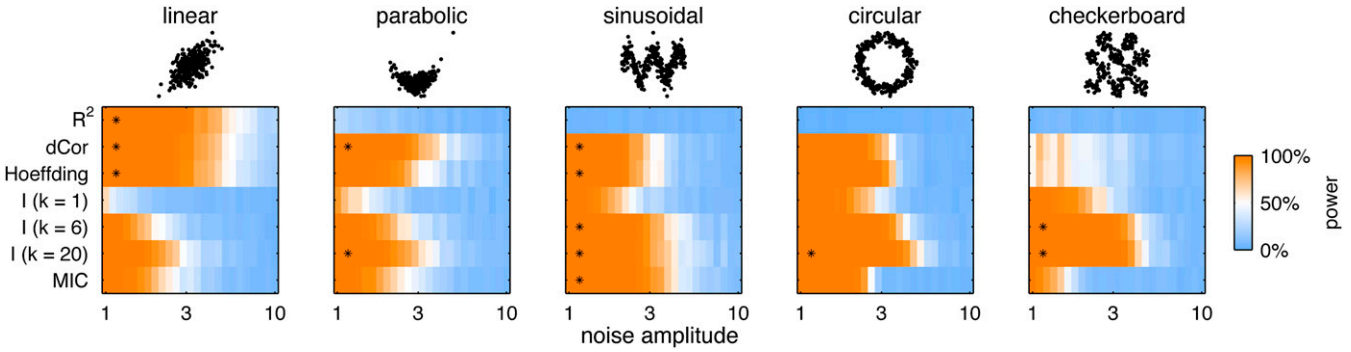

Fig. 4. Assessment of statistical power. Heat maps show power values computed for R^2; dCor (19); Hoeffding's D (20); KNN estimates of mutual information, using $k = 1$, 6, or 20; and MIC. Full power curves are shown in Fig. S6. Simulated datasets comprising 320 data points each were generated for each of five relationship types (linear, parabolic, sinusoidal, circular, or checkerboard), using additive noise that varied in amplitude over a 10-fold range; see Table S2 for simulation details. Asterisks indicate, for each relationship type, the statistics that have either the maximal noise-at-50%-power or a noise-at-50%-power that lies within 25% of this maximum. The scatter plot above each heat map shows an example dataset having noise of unit amplitude.

of this bias, we colored each plotted point according to the monotonicity of the function f used in the corresponding simulation. We observe that MIC assigns systematically higher scores to monotonic relationships (colored in blue) than to nonmonotonic relationships (colored in orange). Relationships of intermediate monotonicity (purple) fall in between. This bias of MIC for monotonic relationships is further seen in analogous tests of self-equitability (Fig. S4*A*).

MIC is therefore seen, in practice, to violate R^2-equitability, the criterion adopted by Reshef et al. (1). However, this nonequitable behavior of MIC is obscured in figure 2B of ref. 1 by two factors. First, scatter due to the small size of the simulated datasets obscures the f-dependent bias of MIC. Second, the nonsystematic coloring scheme used in figure 2B of ref. 1 masks the bias that becomes apparent with the coloring scheme used here.

To argue that mutual information violates their equitability criterion, Reshef et al. (1) estimated the mutual information in each simulated dataset and then plotted these estimates $I\{x;y\}$ against noise, again quantified by $1-R^2\{f(x);y\}$. These results, initially reported in figure 2D of ref. 1, are reproduced here in Fig. 3*C*. At first glance, Fig. 3*C* suggests a bias of mutual information for monotonic functions that is significantly worse than the bias exhibited by MIC. However, these observations are artifacts resulting from two factors.

First, Reshef et al. (1) did not compute the true mutual information of the underlying relationship; rather, they estimated it using the KNN algorithm of Kraskov et al. (18). This algorithm estimates mutual information based on the distance between kth nearest-neighbor data points. In essence, k is a smoothing parameter: Low values of k will give estimates of mutual information with high variance but low bias, whereas high values of k will lessen this variance but increase bias. Second, the bias due to large values of k is exacerbated in small datasets relative to large datasets. If claims about the inherent bias of mutual information are to be supported using simulations, it is imperative that mutual information be estimated on datasets that are sufficiently large for this estimator-specific bias to be negligible.

We therefore replicated the analysis in figure 2D of ref. 1, but simulated 5,000 data points per relationship and used the KNN mutual information estimator with $k=1$ instead of $k=6$. The results of this computation are shown in Fig. 3*D*. Here we see nearly all of the nonequitable behavior cited in ref. 1 is eliminated; this observation holds in the large data limit (Fig. S4*D*).

Of course mutual information does not exactly satisfy R^2-equitability because no meaningful dependence measure does. However, mutual information does satisfy self-equitability, and Fig. S4*E* shows that the self-equitable behavior of mutual information is seen to hold approximately for KNN estimates made on the simulated data from Fig. 3*D*. Increasing values of k reduce the self-equitability of the KNN algorithm (Fig. S4 *E*–*G*).

Statistical Power. Simon and Tibshirani (24) have stressed the importance of statistical power for measures of bivariate association. In this context, "power" refers to the probability that a statistic, when evaluated on data exhibiting a true dependence between X and Y, will yield a value that is significantly different from that for data in which X and Y are independent. MIC was observed (24) to have substantially less power than a statistic called dCor (19), but KNN mutual information estimates were not tested. We therefore investigated whether the statistical power of KNN mutual information estimates could compete with dCor, MIC, and other non–self-equitable dependence measures.

Fig. 4 presents the results of statistical power comparisons performed for various statistics on relationships of five different types.[¶] As expected, R^2 was observed to have optimal power on the linear relationship, but essentially negligible power on the other (mirror symmetric) relationships. dCor and Hoeffding's D (20) performed similarly to one another, exhibiting nearly the same power as R^2 on the linear relationship and retaining substantial power on all but the checkerboard relationship.

Power calculations were also performed for the KNN mutual information estimator using $k=1$, 6, and 20. KNN estimates computed with $k=20$ exhibited the most statistical power of these three; indeed, such estimates exhibited optimal or near-optimal statistical power on all but the linear relationship.

However, R^2, dCor, and Hoeffding's D performed substantially better on the linear relationship (Fig. S6). This is important to note because the linear relationship is likely to be more representative of many real-world datasets than are the other four relationships tested. The KNN mutual information estimator also has the important disadvantage of requiring the user to specify k without any mathematical guidelines for doing so. The choices of k used in our simulations were arbitrary, and, as shown, these choices can greatly affect the power and equitability of one's mutual information estimates.

MIC, computed using $B=N^{0.6}$, was observed to have relatively low statistical power on all but the sinusoidal relationship. This is consistent with the findings of ref. 24. Interestingly, MIC actually exhibited less statistical power than the mutual information estimate I_{MIC} on which it is based (Figs. S5 and S6). This argues that the normalization procedure in Eq. **7** may actually reduce the statistical utility of MIC.

We note that the power of the KNN estimator increased substantially with k, particularly on the simpler relationships, whereas the self-equitability of the KNN estimator was observed to decrease with increasing k (Fig. S4 *E*–*G*). This trade-off between power and equitability, observed for the KNN estimator,

[¶]These five relationships were chosen to span a wide range of possible qualitative forms; they should not be interpreted as being equally representative of real data.

appears to reflect the bias vs. variance trade-off well known in statistics. Indeed, for a statistic to be powerful it must have low variance, but systematic bias in the values of the statistic is irrelevant. By contrast, our definition of equitability is a statement about the bias of a dependence measure, not the variance of its estimators.

Discussion

We have argued that equitability, a heuristic property for dependence measures that was proposed by Reshef et al. (1), is properly formalized by self-equitability, a self-consistency condition closely related to DPI. This extends the notion of equitability, defined originally for measures of association between one-dimensional variables only, to measures of association between variables of all types and dimensionality. All DPI-satisfying measures are found to be self-equitable, and among these mutual information is particularly useful due to its fundamental meaning in information theory and statistics (6–8).

Not all statistical problems call for a self-equitable measure of dependence. For instance, if data are limited and noise is known to be approximately Gaussian, R^2 (which is not self-equitable) can be a much more useful statistic than estimates of mutual information. On the other hand, when data are plentiful and noise properties are unknown a priori, mutual information has important theoretical advantages (8). Although substantial difficulties with estimating mutual information on continuous data remain, such estimates have proved useful in a variety of real-world problems in neuroscience (14, 15, 25), molecular biology (16, 17, 26–28), medical imaging (29), and signal processing (13).

In our tests of equitability, the vast majority of MIC estimates were actually identical to the naive mutual information estimate I_{MIC}. Moreover, the statistical power of MIC is noticeably reduced relative to I_{MIC} in situations where the denominator Z_{MIC} in Eq. **7** fluctuates (Figs. S5 and S6). This suggests that the normalization procedure at the heart of MIC actually decreases MIC's statistical utility.

We briefly note that the difficulty of estimating mutual information has been cited as a reason for using MIC instead (3). However, MIC is actually much harder to estimate than mutual information due to the definition of MIC requiring that all possible binning schemes for each dataset be tested. Consistent with this we have found the MIC estimator from ref. 1 to be orders of magnitude slower than the mutual information estimator of ref. 18.

In addition to its fundamental role in information theory, mutual information is thus seen to naturally solve the problem of equitably quantifying statistical associations between pairs of variables. Unfortunately, reliably estimating mutual information from finite continuous data remains a significant and unresolved problem. Still, there is software (such as the KNN estimator) that can allow one to estimate mutual information well enough for many practical purposes. Taken together, these results suggest that mutual information is a natural and potentially powerful tool for making sense of the large datasets proliferating across disciplines, both in science and in industry.

Materials and Methods

MIC was estimated using the "MINE" suite of ref. 1 or the "minepy" package of ref. 23 as described. Mutual information was estimated using the KNN estimator of ref. 18. Simulations and analysis were performed using custom Matlab scripts; details are given in *SI Text*. Source code for all of the analysis and simulations reported here is available at https://sourceforge.net/projects/equitability/.

ACKNOWLEDGMENTS. We thank David Donoho, Bud Mishra, Swagatam Mukhopadhyay, and Bruce Stillman for their helpful feedback. This work was supported by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

1. Reshef DN, et al. (2011) Detecting novel associations in large data sets. *Science* 334(6062):1518–1524.
2. Reshef DN, Reshef Y, Mitzenmacher M, Sabeti P (2013) Equitability analysis of the maximal information coefficient with comparisons. arXiv:1301.6314v1 [cs.LG].
3. Speed T (2011) Mathematics. A correlation for the 21st century. *Science* 334(6062): 1502–1503.
4. Anonymous (2012) Finding correlations in big data. *Nat Biotechnol* 30(4):334–335.
5. Shannon CE, Weaver W (1949) *The Mathematical Theory of Communication.* (Univ of Illinois, Urbana, IL).
6. Cover TM, Thomas JA (1991) *Elements of Information Theory* (Wiley, New York).
7. Kullback S (1959) *Information Theory and Statistics* (Dover, Mineola, NY).
8. Kinney JB, Atwal GS (2013) Parametric inference in the large data limit using maximally informative models. *Neural Comput*, 10.1162/NECO_a_00568.
9. Miller G (1955) Note on the bias of information estimates. *Information Theory in Psychology II-B*, ed Quastler H (Free Press, Glencoe, IL), pp 95–100.
10. Treves A, Panzeri S (1995) The upward bias in measures of information derived from limited data samples. *Neural Comput* 7(2):399–407.
11. Khan S, et al. (2007) Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys Rev E Stat Nonlin Soft Matter Phys* 76(2 Pt 2):026209.
12. Panzeri S, Senatore R, Montemurro MA, Petersen RS (2007) Correcting for the sampling bias problem in spike train information measures. *J Neurophysiol* 98(3): 1064–1072.
13. Hyvärinen A, Oja E (2000) Independent component analysis: Algorithms and applications. *Neural Netw* 13(4–5):411–430.
14. Sharpee T, Rust NC, Bialek W (2004) Analyzing neural responses to natural signals: Maximally informative dimensions. *Neural Comput* 16(2):223–250.
15. Sharpee TO, et al. (2006) Adaptive filtering enhances information transmission in visual cortex. *Nature* 439(7079):936–942.
16. Kinney JB, Tkacik G, Callan CG, Jr. (2007) Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA* 104(2):501–506.
17. Kinney JB, Murugan A, Callan CG, Jr., Cox EC (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci USA* 107(20):9158–9163.
18. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. *Phys Rev E Stat Nonlin Soft Matter Phys* 69(6 Pt 2):066138.
19. Szekely G, Rizzo M (2009) Brownian distance covariance. *Ann Appl Stat* 3(4): 1236–1265.
20. Hoeffding W (1948) A non-parametric test of independence. *Ann Math Stat* 19(4): 546–557.
21. Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc A* 231:289–337.
22. Paninski L (2003) Estimation of entropy and mutual information. *Neural Comput* 15(6):1191–1253.
23. Albanese D, et al. (2013) Minerva and minepy: A C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics* 29(3):407–408.
24. Simon N, Tibshirani R (2011) Comment on 'Detecting novel associations in large data sets' by Reshef et al., Science Dec 16, 2011. arXiv:1401.7645.
25. Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W (1997) *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA).
26. Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* 28(2):337–350.
27. Goodarzi H, et al. (2012) Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* 485(7397):264–268.
28. Margolin AA, et al. (2006) ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl 1):S7.
29. Pluim JPW, Maintz JBA, Viergever MA (2003) Mutual-information-based registration of medical images: A survey. *IEEE Trans Med Imaging* 22(8):986–1004.

Supporting Information

Kinney and Atwal 10.1073/pnas.1309933111

SI Text

Theorem 1. *Let X and η be (not necessarily independent) random variables and define $Y = X + \eta$. For any invertible function h, the random variable $\xi = Y - h(X)$ obeys the Markov chain condition $X \leftrightarrow h(X) \leftrightarrow \xi$.*

Proof: This follows trivially from the fact that, because h is invertible, $p(\xi\,|h(X),X) = p(\xi\,|h(X))$, which is what it means for $X \leftrightarrow h(X) \leftrightarrow \xi$ to be a Markov chain. Note that the definition of ξ is actually irrelevant here. □

Theorem 2. *Every self-equitable dependence measure $D[X;Y]$ is invariant under all invertible transformations of X and Y.*

Proof: If $X \leftrightarrow f(X) \leftrightarrow Y$ is a Markov chain and the functions h_1 and h_2 are invertible, then

$$X \leftrightarrow h_1(X) \leftrightarrow X \leftrightarrow f(X) \leftrightarrow Y \leftrightarrow h_2(Y) \leftrightarrow Y \qquad \text{[S1]}$$

is also a Markov chain. Extracting various Markov subchains and invoking Eq. **3** of the main text, we see that for every self-equitable dependence measure D,

$$X \leftrightarrow h_1(X) \leftrightarrow Y \Rightarrow D[X;Y] = D[h_1(X);Y], \qquad \text{[S2]}$$

$$Y \leftrightarrow h_2(Y) \leftrightarrow h_1(X) \Rightarrow D[Y;h_1(X)] = D[h_2(Y);h_1(X)]. \qquad \text{[S3]}$$

Invoking the symmetry of D we get $D[X;Y] = D[h_1(X);h_2(Y)]$, which is what we aimed to prove. □

Theorem 3. *Every DPI-satisfying dependence measure $D[X;Y]$ is self-equitable.*

Proof: If $X \leftrightarrow f(X) \leftrightarrow Y$ is a Markov chain, so is

$$f(X) \leftrightarrow X \leftrightarrow f(X) \leftrightarrow Y. \qquad \text{[S4]}$$

Extracting Markov subchains and using Eq. **4** of the main text, we see than for any DPI-satisfying measure $D[X;Y]$,

$$X \leftrightarrow f(X) \leftrightarrow Y \Rightarrow D[X;Y] \le D[f(X);Y] \qquad \text{[S5]}$$

$$f(X) \leftrightarrow X \leftrightarrow Y \Rightarrow D[f(X);Y] \le D[X;Y] \qquad \text{[S6]}$$

and so $D[X;Y] = D[f(X);Y]$, proving that D is self-equitable. □

Theorem 4. *Every dependence measure $I_F[X;Y]$ that can be written as*

$$I_F[X;Y] \equiv \int dx\,dy\,p(x)p(y)F\left(\frac{p(x,y)}{p(x)p(y)}\right), \qquad \text{[S7]}$$

where F is a convex function on the nonnegative real numbers, satisfies DPI. Note that such measures of dependence are called "F-information" measures (1).

Proof: Let $X \leftrightarrow Y \leftrightarrow Z$ be a Markov chain, so that $p(x,z) = \int dy\,p(x\,|y)p(y,z)$ and $p(x) = \int dy\,p(x\,|y)p(y)$. Then

$$I_F[X;Z] = \int dx\,dz\,p(x)p(z)F\left(\frac{p(x,z)}{p(x)p(z)}\right) \qquad \text{[S8]}$$

$$= \int dx\,dy\,dz\,p(x\,|y)p(y)p(z)F\left(\frac{\int dy\,p(x\,|y)p(y,z)}{\int dy\,p(x\,|y)p(y)p(z)}\right) \qquad \text{[S9]}$$

$$\le \int dx\,dy\,dz\,p(x\,|y)p(y)p(z)F\left(\frac{p(y,z)}{p(y)p(z)}\right) \qquad \text{[S10]}$$

$$= \int dy\,dz\,p(y)p(z)F\left(\frac{p(y,z)}{p(y)p(z)}\right) \qquad \text{[S11]}$$

$$= I_F[Y;Z], \qquad \text{[S12]}$$

where in Eq. **S10** we have used Jensen's inequality (2). This proves that $I_F[X;Z] \le I_F[Y;Z]$ and thus that I_F satisfies DPI. □

Noise in the X Variable. The supplemental material of ref. 3 as well as follow-up work (4) has suggested that R^2-equitability should actually be extended to include cases where noise is present in both the X and Y variables. This is formalized by replacing Eq. **1** with

$$D[X + \zeta;Y] = g\left(R^2[f(X);Y]\right), \qquad \text{[S13]}$$

where ζ is an additional noise term. It is clear, however, that this more general definition does not permit nontrivial solutions. This stems from the fact that the left side of Eq. **S13** depends on ζ whereas the right side does not. Any measure $D[X + \zeta;Y]$ would therefore have to be invariant to the amount of noise ζ in the X coordinate. For example, consider $f(X) = X$, $\eta = 0$ and assume the noise ζ is very large and does not depend on X. Eq. **S13** would require $D[X;X] = D[X + \zeta;X] \approx D[\zeta;X]$, implying the value assigned by D to the identity relationship must be the same as the value assigned to a relationship in which the two variables in question are independent.

MIC and I_{MIC} Are Usually Identical. As mentioned in the main text, we were unable to extract I_{MIC} values from the MIC estimator of Reshef et al. (3) because the source code has not, as of this writing, been made publicly available. However, we were able to extract I_{MIC}, n_X, and n_Y values from the MIC estimator of Albanese et al. (5).

For over 99% of the MIC computations shown Fig. 3 *A* and *B* of the main text, the MIC algorithm chose binning schemes having only two bins on either the x or the y axis. As mentioned above, $MIC\{x;y\} = I_{\mathrm{MIC}}\{x;y\}$ in such cases. It therefore appears that the primary tests of equitability reported by Reshef et al. (3) provide virtually no information about how the normalization procedure in the definition of MIC affects MIC's equitability.

The common occurrence of $MIC\{x;y\} = I_{\mathrm{MIC}}\{x;y\}$ is also reflected in Fig. S5. On the linear, parabolic, and sinusoidal relationships we found $Z_{\mathrm{MIC}} = 1$ (corresponding to $\min(n_X,n_Y) = 2$) for every one of the 37,500 relationships tested. Thus, I_{MIC} and MIC exhibited identical power profiles. However, on the circular and checkerboard relationships we consistently observed $Z_{\mathrm{MIC}} > 1$ (i.e., $\min(n_X,n_Y) > 2$) when noise was sufficiently low; Z_{MIC} dropped back down to 1 as noise increased.

Tellingly, I_{MIC} actually outperformed MIC in tests of statistical power. It appears that right at the noise level at which highly structured relationships become difficult to discern, the binning scheme chosen by MIC starts to fluctuate. The normalization factor Z_{MIC} therefore fluctuates, causing MIC values to become

noisier—and thus less powerful—than the underlying mutual information estimate I_{MIC}. This suggests that the normalization procedure in Eq. **7** of the main text, which transforms I_{MIC} into MIC, may actually be harmful for quantifying dependencies.

Computational Methods. The KNN mutual information estimator of Kraskov et al. (6) was downloaded from http://bsp.teithe.gr/members/downloads/Milca.html. Except for k, which we varied depending on context, the KNN estimator was run with default settings.

Except where stated otherwise, MIC estimates were computed with the software described by Reshef et al. (3), downloaded from www.exploredata.net, using default parameters (including $\alpha = 0.6$). All other MIC values were computed using a version of the estimator of Albanese et al. (5) (with parameters $\alpha = 0.6$, $c = 15$) that was edited to also report the values of I_{MIC}, n_X, and n_Y.

All other analysis was performed in Matlab. Our Matlab scripts, along with our edited version of the Albanese et al. MIC estimator, are available at https://sourceforge.net/projects/equitability/.

1. Csiszár I, Shields PC (2004) *Information Theory and Statistics: A Tutorial* (Now Publishers, Hanover, MA).
2. Cover TM, Thomas JA (1991) *Elements of Information Theory* (Wiley, New York).
3. Reshef DN, et al. (2011) Detecting novel associations in large data sets. *Science* 334(6062):1518–1524.
4. Reshef DN, Reshef Y, Mitzenmacher M, Sabeti P (2013) Equitability analysis of the maximal information coefficient, with comparisons. arXiv:1301.6314v1 [cs.LG].
5. Albanese D, et al. (2013) Minerva and minepy: A C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics* 29(3):407–408.
6. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. *Phys Rev E Stat Nonlin Soft Matter Phys* 69(6 Pt 2):066138.

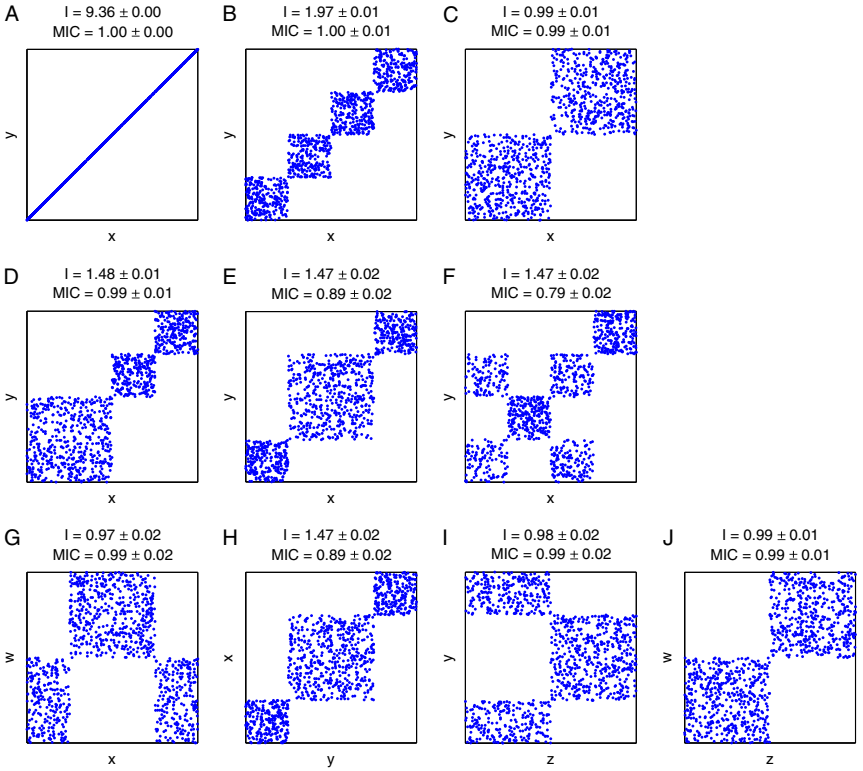

Fig. S1. (*A–J*) Estimates of mutual information and MIC computed for $N = 1{,}000$ data points sampled from the joint probability distributions shown in Fig. 2. Mutual information estimates were computed using the KNN estimator of Kraskov et al. (6) with $k = 1$. MIC estimates were computed using the estimator of Reshef et al. (3) with default parameters. Values for both mutual information and MIC are reported as mean ± SD over 100 replicates.

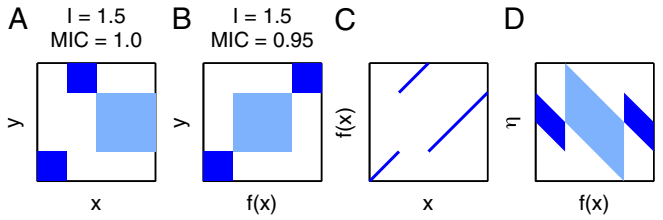

Fig. S2. MIC violates self-equitability. The relationships in *A* and *B* result from $Y = f(X) + \eta$ with function f and noise term η respectively defined in *C* and *D*. Note that the noise term η, as required, depends on $f(X)$ but not otherwise on X. MIC is seen to violate self-equitability (Eq. **3**) because $MIC[X;Y] \neq MIC[f(X);Y]$.

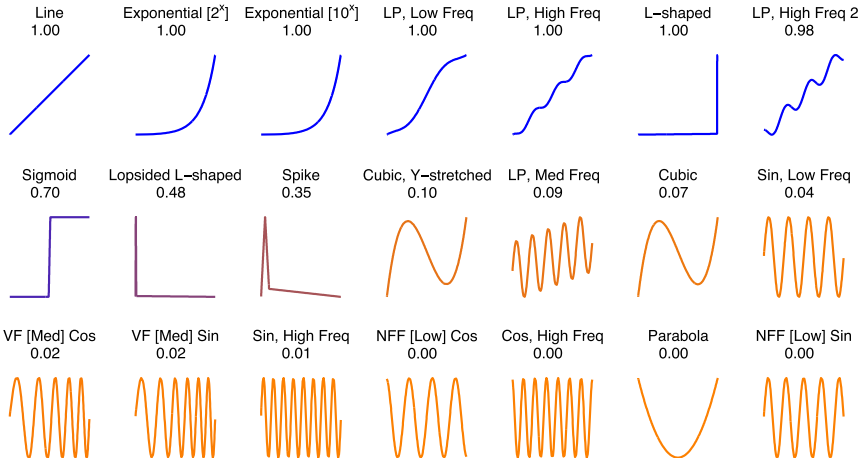

Fig. S3. Functions f used in the simulations described in Fig. 3 and Fig. S4 and listed in Table S1. Functions are colored according to each one's (indicated) monotonicity (Fig. 3*E* and Fig. S4*H*).

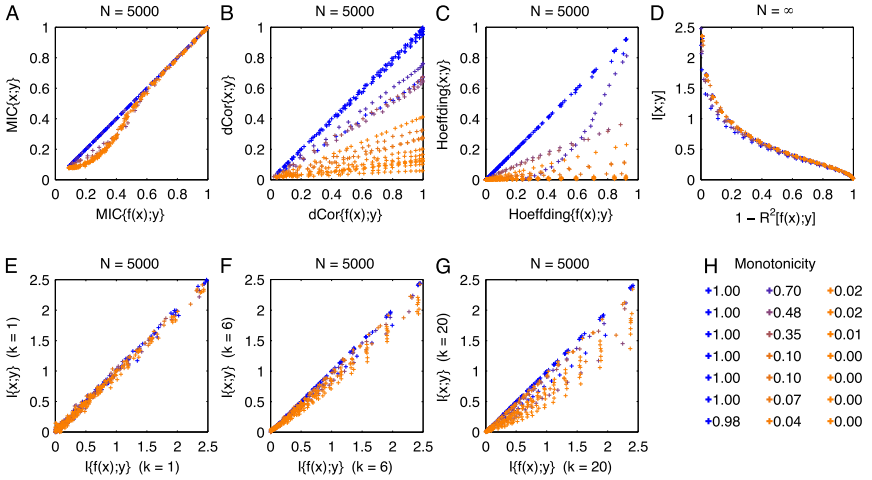

Fig. S4. Additional equitability tests of MIC, dCor, Hoeffding's D, and the KNN mutual information estimator. Noisy relationships were simulated as in Fig. 2 *B* and *D*; see Table S1 for details. (*A*) The MIC estimator of Reshef et al. (1) violates self-equitability on the data simulated for Fig. 2*B* of the main text: Specifically, $MIC\{x;y\} < MIC\{f(x);y\}$ for sufficiently nonmonotonic functions f. This downward bias of MIC is particularly pronounced for $MIC\{f(x);y\}$ values between 0.2 and 0.4. (*B* and *C*) A downward bias for nonmonotonic functions, indicative of non–self-equitability, is also seen in similar tests of (*B*) dCor (2) and (*C*) Hoeffding's D (3). (*D*) Extrapolation of Fig. 2*D* of the main text to datasets of infinite size. Here, mutual information values were computed semianalytically using $I[X;Y] = H[Y] - H[\eta]$ where H is entropy (4). (*E*–*G*) Mutual information estimates $I\{x;y\}$ returned by the KNN algorithm using (*E*) $k = 1$, (*F*) $k = 6$, and (*G*) $k = 20$. Increasing values of k result in decreasing adherence to self-equitability. This is most apparent for nonmonotonic functions, which are more readily blurred by smoothing than are monotonic functions. (*H*) Correspondence between color and monotonicity; same as Fig. 3*E*.

1. Reshef DN, et al. (2011) Detecting novel associations in large data sets. *Science* 334(6062):1518–1524.
2. Szekely G, Rizzo M (2009) Brownian distance covariance. *Ann Appl Stat* 3(4):1236–1265.
3. Hoeffding W (1948) A non-parametric test of independence. *Ann Math Stat* 19(4):546–557.
4. Cover TM, Thomas JA (1991) *Elements of Information Theory* (Wiley, New York).

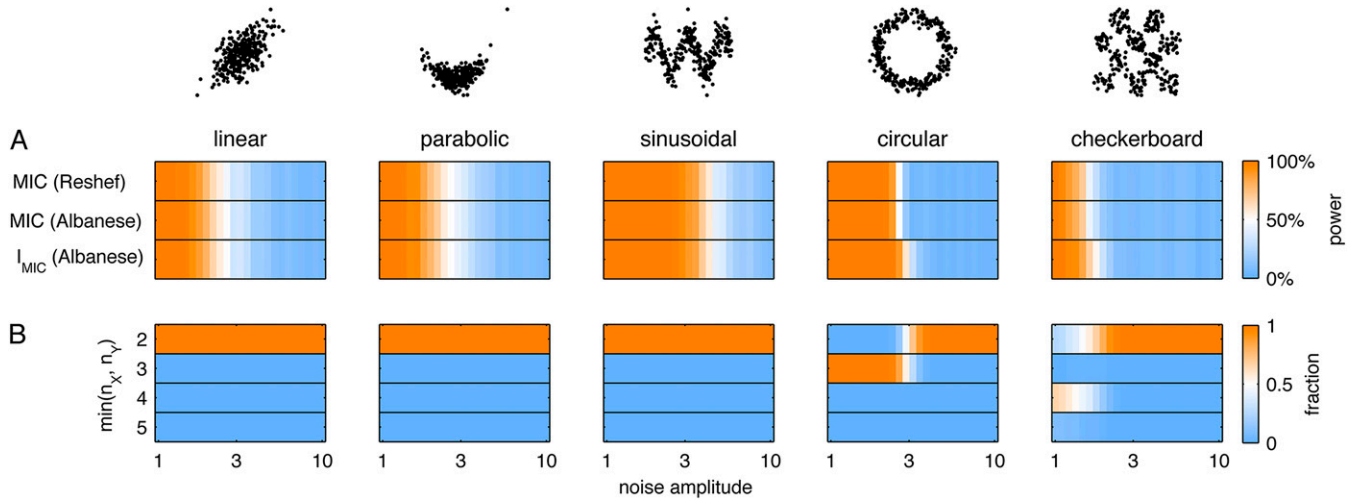

Fig. S5. Power comparisons for I_{MIC} and two MIC estimators. (*A*) Power, estimated as in Fig. 4, for the MIC estimator of Reshef et al. (3), for the MIC estimator of Albanese et al. (5), and for the corresponding I_{MIC} values computed by the Albanese et al. estimator. (*B*) Each heat map shows the fraction of simulations, at each noise level, for which the Albanese et al. MIC estimator used an $n_X \times n_Y$ binning scheme with the indicated value of $\min(n_X, n_Y)$. This value determines the normalization constant $Z_{\mathrm{MIC}} = \log_2(\min(n_X, n_Y))$.

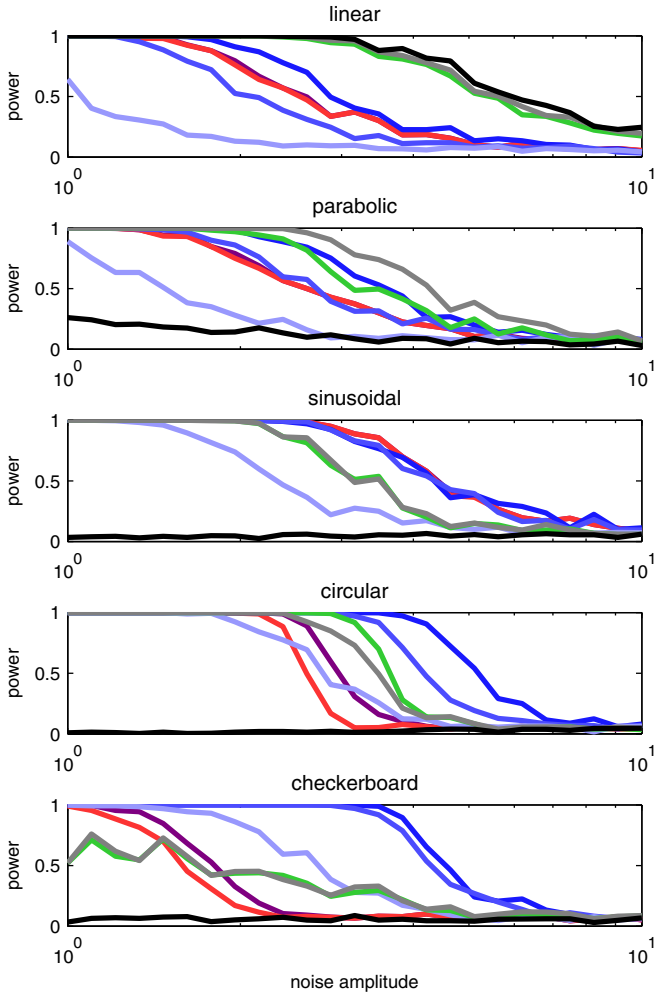

Fig. S6. Power curves corresponding to the heat maps shown in Fig. 4 and Fig. S5. For each of the relationships tested (linear, parabolic, sinusoidal, circular, and checkerboard), power vs. noise is plotted for eight different statistics: R^2 (black); dCor (gray); Hoeffding's D (green); MIC computed using the algorithm of Reshef et al. (3) (red); I_{MIC} computed using the modified algorithm of Albanese et al. (5) (purple); and mutual information computed using the KNN estimator with $k = 1$ (light blue), $k = 6$ (medium blue), or $k = 20$ (dark blue).

Table S1. Noisy functional relationships simulated for Fig. 3 and Fig. S4, ordered according to monotonicity as described in the Fig. 3 legend

Description	Monotonicity	N	X	$f(X)$
Line	1.00	1,000	[0, 1]	X
Exponential [2^X]	1.00	1,000	[0, 10]	2^X
Exponential [10^X]	1.00	1,000	[0, 10]	10^X
LP, low frequency	1.00	1,000	[0, 1]	$\frac{1}{5}\sin(4(2X-1))+\frac{11}{10}(2X-1)$
LP, high frequency	1.00	1,000	[0, 1]	$\frac{1}{10}\sin(10.6(2X-1))+\frac{11}{10}(2X-1)$
L shaped	1.00	1,000	[0, 1]	$\begin{cases} X/99 & \text{if } X \le \frac{99}{100} \\ 99X-98 & \text{if } X > \frac{99}{100} \end{cases}$
LP, high frequency 2	0.98	1,000	[0, 1]	$\frac{1}{5}\sin(10.6(2X-1))+\frac{11}{10}(2X-1)$
Sigmoid	0.70	1,000	[0, 1]	$\begin{cases} 0 & \text{if } X < \frac{49}{100} \\ 50(X-\frac{1}{2})+\frac{1}{2} & \text{if } \frac{49}{100} \le X < \frac{51}{100} \\ 1 & \text{if } X \ge \frac{51}{100} \end{cases}$
Lopsided L shaped	0.48	500*	[0, 1]	$\begin{cases} 200X & \text{if } X < \frac{1}{200} \\ -198X+\frac{199}{100} & \text{if } \frac{1}{200} \le X < \frac{1}{100} \\ -\frac{X}{99}+\frac{1}{99} & \text{if } X \ge \frac{1}{100} \end{cases}$
Spike	0.35	1,000	[0, 1]	$\begin{cases} 20X & \text{if } X < \frac{1}{20} \\ -18X+\frac{19}{10} & \text{if } \frac{1}{20} \le X < \frac{1}{10} \\ -\frac{X}{9}+\frac{1}{9} & \text{if } X \ge \frac{1}{10} \end{cases}$
Cubic, Y stretched	0.10	1,000	[−1.3, 1.1]	$41(4X^3+X^2-4X)$
LP, medium frequency	0.09	500*	[0, 1]	$\sin(10\pi X)+X$
Cubic	0.07	1,000	[−1.3, 1.1]	$4X^3+X^2-4X$
Sin, low frequency	0.04	250	[0, 1]	$\sin(8\pi X)$
VF [med] cos	0.02	1,000	[0, 1]	$\cos(5\pi X(1+X))$
VF [med] sin	0.02	500*	[0, 1]	$\cos(6\pi X(1+X))$
Sin, high frequency	0.01	1,000	[0, 1]	$\sin(16\pi X)$
NFF [low] cos	0.00	250*	[0, 1]	$\cos(7\pi X)$
Cos, high frequency	0.00	500*	[0, 1]	$\cos(14\pi X)$
Parabola	0.00	1,000	$[-\frac{1}{2}, \frac{1}{2}]$	$4X^2$
NFF [low] sin	0.00	1,000	[0, 1]	$\sin(9\pi X)$

These relationships correspond to the relationships simulated by Reshef et al. (3) for Fig. 2 *B–F*, as described in table S3 of ref. 3. Here we have abbreviated function names as follows: LP, linear + periodic; VF, varying frequency; and NFF, non-Fourier frequency. For each relationship listed, data $\{x_i,y_i\}_{i=1}^N$ were simulated, where $Y = f(X)+\eta$, X ranged over the indicated domain, f was one of the 21 functions shown, and η was uniform noise of amplitude equal to the range of $f(X)$ times one of these 24 relative amplitudes: 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.15, 1.3, 1.45, 1.60, 1.8, 2.1, 2.5, 3.1, 4.5, 6, 8, 10, or 20. We note that, as in the simulations Reshef et al. (3) used for their figure 2, X values were not distributed uniformly within the allowed range but rather were chosen so that the points $(X,f(X))$ were spaced evenly along the length of each curve in the x,y plane. This, for instance, is the reason the monotonicities of the "Cubic" and "Cubic, Y stretched" relationships are slightly different.

*Reshef et al. (3) also simulated these relationships using $N = 1{,}000$ data points; we performed simulations only with the number of data points indicated.

Table S2. The *X*,*Y* relationships simulated for the power calculations of Fig. 4 and Figs. S5 and S6

	Random variable	
Relationship type	X	Y
Linear	ξ	$\frac{2}{3}X + a\eta$
Parabolic	ξ	$\frac{2}{3}X^2 + a\eta$
Sinusoidal	$\frac{5}{2}\theta$	$2\cos(X) + a\eta$
Circular	$10\cos(\theta) + a\xi$	$10\sin(\theta) + a\eta$
Checkerboard	$10X_0 + a\xi$	$10Y_0 + a\eta$

Five hundred trial datasets were generated for each of these relationships at each of twenty-five different noise amplitudes distributed logarithmically between 1 and 10. For each dataset, statistics were computed on the "true" data $\{x_i, y_i\}_{i=1}^{N}$ as well as on "null" data, for which the indexes i on the y values had been randomly permuted. The power of each statistic was defined as the fraction of true datasets yielding a statistic value greater than 95% of the values yielded by the corresponding null datasets. ξ and η are random numbers drawn from the normal distribution $\mathcal{N}(0,1)$. θ is a random number drawn uniformly from the interval $[-\pi,\pi)$. (X_0, Y_0) is a pair of random numbers drawn uniformly from the solid squares of a 4×5 checkerboard, where each square has sides of length 1. a is the "noise amplitude" shown in Fig. 4 and Figs. S5 and S6.