

Equitability, Interval Estimation, and Statistical Power

Yakir A. Reshef¹, David N. Reshef¹, Pardis C. Sabeti² and Michael Mitzenmacher²

Abstract. Emerging high-dimensional data sets often contain many non-trivial relationships, and, at modern sample sizes, screening these using an independence test can sometimes yield too many relationships to be a useful exploratory approach. We propose a framework to address this limitation centered around a property of measures of dependence called *equitability*. Given some measure of relationship strength, an equitable measure of dependence is one that assigns similar scores to equally strong relationships of different types. We formalize equitability within a semiparametric inferential framework in terms of interval estimates of relationship strength, and we then use the correspondence of these interval estimates to hypothesis tests to show that equitability is equivalent under moderate assumptions to requiring that a measure of dependence yield well-powered tests not only for distinguishing nontrivial relationships from trivial ones but also for distinguishing stronger relationships from weaker ones. We then show that equitability, to the extent it is achieved, implies that a statistic will be well powered to detect all relationships of a certain minimal strength, across different relationship types in a family. Thus, equitability is a strengthening of power against independence that enables exploration of data sets with a small number of strong, interesting relationships and a large number of weaker, less interesting ones.

Key words and phrases: Equitability, measure of dependence, statistical power, independence test, semiparametric inference.

1. INTRODUCTION

Suppose we have a data set that we would like to explore to find associations of interest. A commonly taken approach that makes minimal assumptions about the structure in the data is to compute a measure of dependence, that is, a statistic whose population value is zero exactly in cases of statistical independence, on all

Yakir A. Reshef is Ph.D. candidate, School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: yakir@seas.harvard.edu). David N. Reshef is Ph.D. candidate, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA (e-mail: dnreshef@mit.edu). Pardis C. Sabeti is Professor, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: pardis@broadinstitute.org). Michael Mitzenmacher is Professor, School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: michaelm@eecs.harvard.edu).

¹Co-first author.

²Co-last author.

possible pairs of variables. The score of each variable pair can be evaluated against a null hypothesis of statistical independence, and variable pairs with significant scores can be kept for follow-up (Storey and Tibshirani, 2003, Emilsson et al., 2008, Sun and Zhao, 2014, Rachel Wang, Waterman and Huang, 2014). There is a wealth of measures of dependence from which to choose for this task (Hoeffding, 1948, Breiman and Friedman, 1985, Kraskov, Stögbauer and Grassberger, 2004, Gretton et al., 2005a, Székely, Rizzo and Bakirov, 2007, Székely and Rizzo, 2009, Reshef et al., 2011, Gretton et al., 2012, Heller, Heller and Gorfine, 2013, Sugiyama and Borgwardt, 2013, Heller et al., 2016, Lopez-Paz, Hennig and Schölkopf, 2013, Rachel Wang, Waterman and Huang, 2014, Jiang, Ye and Liu, 2015, Reshef et al., 2016, Zhang, 2016, Wang, Jiang and Liu, 2017, Romano et al., 2018).

While this approach works well in some settings, it can be limited by the size of modern data sets. In particular, as data sets grow in dimensionality and sample size, the above approach often results in lists of significant relationships that are too large to allow for meaningful follow-up of every identified relationship, even after correction for multiple hypothesis testing. For example, in the gene

expression data set analyzed in Heller et al. (2016), several measures of dependence reliably identified, at a false discovery rate of 5%, thousands of significant relationships amounting to between 65 and 75 percent of the variable pairs in the data set. Given the extensive manual effort that is usually necessary to better understand each of these results, further characterizing all of them is impractical.

A tempting way to deal with this challenge is to rank all the variable pairs in a data set according to the test statistic used (or according to p-value) and to examine only a small number of pairs with the most extreme values (Faust and Raes, 2012, Turk-Browne, 2013). However, this idea has a pitfall: while a measure of dependence guarantees nonzero scores to dependent variable pairs, the magnitude of these nonzero scores can depend heavily on the type of dependence in question, thereby skewing the top of the list toward certain types of relationships over others (Faust and Raes, 2012, Sun and Zhao, 2014). For example, if some measure of dependence $\hat{\phi}$ systematically assigns higher scores to, say, linear relationships than to nonlinear relationships, then using $\hat{\phi}$ to rank variable pairs in a large data set could cause noisy linear relationships in the data set to crowd out strong nonlinear relationships from the top of the list. The natural result would be that the human examining the top-ranked relationships would never see the nonlinear relationships, and they would not be discovered (Speed, 2011, Sun and Zhao, 2014).

The consistency guarantee of measures of dependence is therefore not strong enough to solve the data exploration problem posed here. What is needed is a way not just to identify as many relationships of different kinds as possible in a data set, but also to identify a small number of strongest relationships of different kinds.

Here we propose and formally characterize *equitability*, a framework for meeting this goal. In previous work, equitability was informally described as the extent to which a measure of dependence assigns similar scores to equally noisy relationships, regardless of relationship type (Reshef et al., 2011). Given that this informal definition has led to substantial follow-up work (Murrell, Murrell and Murrell, 2016, Reshef et al., 2016, Ding et al., 2017, Wang, Jiang and Liu, 2017, Romano et al., 2018), the concept of equitability merits a unifying framework. In this paper, we therefore formalize equitability in terms of interval estimates of relationship strength and use the correspondence between confidence intervals and hypothesis tests to tie it to the notion of statistical power. Our formalization shows that equitability essentially amounts to an assessment of the degree to which a measure of dependence can be used to perform conservative semiparametric inference based on extremum quantiles. In this sense, it is a natural application of ideas from statistical decision theory to measures of dependence.

Intuitively, our proposal is simply to quantify the extent to which a measure of dependence can be used to estimate an effect size rather than just to reject a null of independence. More formally, given a measure of dependence $\hat{\phi}$, a benchmark set \mathcal{Q} of relationship types, and some quantification Φ of relationship strength defined on \mathcal{Q} , we construct an interval estimate of the relationship strength Φ from the value of $\hat{\phi}$ that is valid over \mathcal{Q} . We then use the sizes of these intervals to quantify the utility of $\hat{\phi}$ as an estimate of effect size on \mathcal{Q} , and we define an equitable statistic to be one that yields narrow interval estimates. As we explain, this property can be viewed as a natural generalization of one of the “fundamental properties” described by Rényi in his framework for measures of dependence (Rényi, 1959). It can also be viewed as a weakening of the notion of consistency of an estimator.

After defining equitability, we connect it to statistical power using a variation on the standard equivalence of interval estimation and hypothesis testing. Specifically, we show that under moderate assumptions an equitable statistic is one that yields tests for distinguishing finely between relationships of two different strengths that may both be nontrivial. This result gives us a way to understand equitability as a natural strengthening of the traditional requirement of power against independence, which asks that a statistic be useful only for detecting deviations from strict independence (i.e., distinguishing zero relationship strength from nonzero relationship strength). As we discuss, this view of equitability is related to the concept of separation rate in the minimax hypothesis testing literature (Baraud, 2002, Fromont, Lerasle and Reynaud-Bouret, 2016, Arias-Castro, Pelletier and Saligrama, 2018).

Finally, motivated by the connection between equitability and power, we define an additional property, the *detection threshold* of an independence test, which is the minimal relationship strength x such that the test is well powered to detect all relationships with strength at least x at some fixed sample size, across different relationship types in \mathcal{Q} . This is analogous to the commonly analyzed notion of testing rate (Ingster, 1987, Lepski and Spokoiny, 1999, Baraud, 2002, Ingster and Suslina, 2003), which has been studied in detail for independence testing both in the statistics literature (Ingster, 1989, Yodé, 2011, Zhang, 2016) as well as the computer science and information theory literature (Paninski, 2008, Acharya, Daskalakis and Kamath, 2015). Traditionally, testing rate for independence testing problems has been defined in terms of some distance (usually total variation distance) between the alternatives in question and independence. Here we define the property generically in terms of an arbitrary notion of relationship strength (e.g., R^2 of a noisy functional relationship) and show that high equitability implies low detection threshold but that the

converse does not hold. Therefore, when equitability is too much to ask, low detection threshold on a broad set of relationships with respect to an interesting measure of relationship strength may be a reasonable surrogate goal.

As additional methods are developed around equitability (Murrell, Murrell and Murrell, 2016, Reshef et al., 2016, Ding et al., 2017, Wang, Jiang and Liu, 2017, Romano et al., 2018), a framework for rigorously thinking about this property is becoming increasingly important. The results we present here provide such a framework, including language that is sufficiently general to accommodate related ideas that have arisen in the literature. For example, the definitions provided here allow us to precisely discuss the alternative definitions of Kinney and Atwal (2014) and to explain the implications and limitations of the results therein, as well as to crystallize and conceptually discuss the power against independence of equitable methods (Simon and Tibshirani, 2012).

Throughout this paper, we attempt whenever possible to use terminology consistent with previously published literature on equitability. However, given the extensive relationship between ideas from semiparametric inference and elements of the equitability framework, we point out several cases in which similar or related ideas can be rephrased in more standard statistical language. We also give concrete examples of how our formalism relates to the analysis of equitability in practice, and we close with an example empirical analysis of the equitability of a few popular measures of dependence. We emphasize, however, that much more extensive empirical analyses have been conducted elsewhere (see Reshef et al., 2016, 2018), and the analyses shown here are intended only to be illustrative.

2. DEFINING EQUITABILITY

2.1 Preliminaries

Suppose we are given a statistic $\hat{\phi}$ taking values in $[0, 1]$ that is a measure of dependence. To formally define what it means for $\hat{\phi}$ to give similar scores to equally noisy relationships of different types, we must specify which relationships we are talking about. Therefore, we assume that there is some set \mathcal{Q} of distributions called *standard relationships*, on which we have a well-defined notion of relationship strength in the form of a scalar-valued functional $\Phi : \mathcal{Q} \rightarrow [0, 1]$ that we call the *property of interest*. The idea is that \mathcal{Q} contains relationships of many different types, and for any distribution $\mathcal{Z} \in \mathcal{Q}$, $\Phi(\mathcal{Z})$ is the way we would ideally quantify the strength of \mathcal{Z} if we knew the distribution \mathcal{Z} . Our goal is then to see, given a sample Z of size n from \mathcal{Z} , how well $\hat{\phi}(Z)$ can be used to estimate $\Phi(\mathcal{Z})$.

In standard statistical terminology, this is a semiparametric setup in which \mathcal{Q} is a model, Φ is simply a one-

dimensional parameter of interest, and all other parameters are nuisance parameters. We deviate from this terminology here both for continuity with existing literature and to emphasize the fact that $\hat{\phi}$ is not simply an estimator of Φ but rather a measure of dependence whose utility as a (potentially imperfect) estimator of Φ we wish to evaluate. This correspondence of terminology, along with a summary of the other equitability-related terms defined in this section, is listed following the definitions themselves, at the end of Section 2.3.

We keep our exposition generic in order to accommodate variations—both existing (Kinney and Atwal, 2014, Murrell, Murrell and Murrell, 2016, Ding et al., 2017, Wang, Jiang and Liu, 2017) and potential—on the concepts defined here. However, as a motivating example, we often return to the setting in which \mathcal{Q} is a set of noisy functional relationships and Φ is the coefficient of determination (R^2) with respect to the generating function, that is, the squared Pearson correlation between the dependent variable and the generating function evaluated on the independent variable.

2.2 Q-Confidence Intervals

Our approach to defining equitability is to construct from $\hat{\phi}$ an interval estimate of Φ by inverting a certain set of hypothesis tests. The statistic $\hat{\phi}$ will then be equitable if it yields narrow interval estimates of Φ . To construct our interval estimates, we must first describe the acceptance regions of the hypothesis tests that we invert. We do so using a standard construction of acceptance regions in terms of quantiles of a statistic. (In this definition as well as later definitions, we implicitly assume a fixed sample size of n .)

DEFINITION 2.1 (\mathcal{Q} -acceptance region). Let $\hat{\phi}$ be a statistic taking values in $[0, 1]$, and let $x, \alpha \in [0, 1]$. The level- α \mathcal{Q} -acceptance region of $\hat{\phi}$ at x , denoted by $A_\alpha(x)$, is the closed interval $[a, b]$ where a is the minimum $\alpha/2$ quantile of $\hat{\phi}(Z)$ and b is the maximum $1 - \alpha/2$ quantile of $\hat{\phi}(Z)$, with Z being a sample from some $\mathcal{Z} \in \mathcal{Q}$ and the minimum and maximum taken over all \mathcal{Z} satisfying $\Phi(\mathcal{Z}) = x$.

See Figure 1(a) for an illustration. The \mathcal{Q} -acceptance region of $\hat{\phi}$ at x is an acceptance region for one particular test of the null hypothesis $H_0 : \Phi(\mathcal{Z}) = x$ on relationships in \mathcal{Q} . We refer to it as a \mathcal{Q} -acceptance region to emphasize that, although the underlying statistic $\hat{\phi}$ is a measure of dependence that could be applied without assumptions about the underlying data-generating process, the acceptance regions we describe are valid only on \mathcal{Q} .

If there is only one $\mathcal{Z} \in \mathcal{Q}$ satisfying $\Phi(\mathcal{Z}) = x$, the \mathcal{Q} -acceptance region amounts to a central interval of the sampling distribution of $\hat{\phi}$ on \mathcal{Z} . If there is more than one such \mathcal{Z} , the acceptance region expands to include the relevant central intervals of the sampling distributions of $\hat{\phi}$ on

all the distributions \mathcal{Z} in question. For example, when \mathcal{Q} is a set of noisy functional relationships with several different function types and Φ is R^2 , the \mathcal{Q} -acceptance region at x is the smallest interval A such that for any functional relationship $\mathcal{Z} \in \mathcal{Q}$ with $R^2(\mathcal{Z}) = x$, $\hat{\varphi}(\mathcal{Z})$ falls in A with high probability over the sample Z of size n from \mathcal{Z} .

We can now construct interval estimates of Φ in terms of $A_\alpha(x)$ via the standard approach of inversion of hypothesis tests (Casella and Berger, 2002).

DEFINITION 2.2 (\mathcal{Q} -confidence interval). Let $\hat{\varphi}$ be a statistic taking values in $[0, 1]$, and let $y, \alpha \in [0, 1]$. The $(1 - \alpha)$ \mathcal{Q} -confidence interval of $\hat{\varphi}$ at y for Φ , denoted by $I_\alpha(y)$, is the smallest closed interval containing the set

$$\{x \in [0, 1] : y \in A_\alpha(x)\},$$

where $A_\alpha(\cdot)$ denotes level- α \mathcal{Q} -acceptance regions of $\hat{\varphi}$.

See Figure 1(a) for an illustration. The \mathcal{Q} -confidence interval is a conservative confidence interval for the parameter $\Phi(\mathcal{Z})$ at $\hat{\varphi} = y$ induced by the extremum quantiles of $\hat{\varphi}$. In other words, we have the following guarantee about the coverage probability of the \mathcal{Q} -confidence intervals, whose proof is given by the standard argument about the relationship between quantiles and confidence sets (Casella and Berger, 2002).

PROPOSITION 2.1. Let $\hat{\varphi}$ be a statistic taking values in $[0, 1]$, and let $\alpha \in [0, 1]$. For all $\mathcal{Z} \in \mathcal{Q}$,

$$\mathbf{P}(\Phi(\mathcal{Z}) \in I_\alpha(\hat{\varphi}(\mathcal{Z}))) \geq 1 - \alpha,$$

where Z is a sample of size n from \mathcal{Z} .

The definitions just presented have natural nonstochastic counterparts in the large-sample limit, which we defer to Appendix A, that quantify the degree of nonidentifiability induced by φ with respect to Φ on \mathcal{Z} independently of any finite-sample effects. See Figure 1(b) for an illustration.

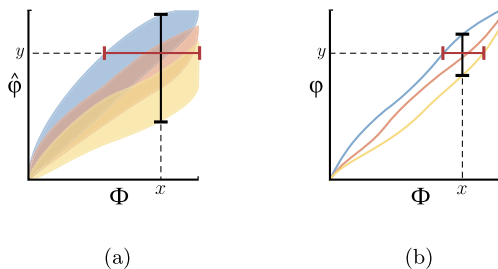


FIG. 1. A schematic illustration of \mathcal{Q} -acceptance regions and \mathcal{Q} -confidence intervals. In both figure parts, \mathcal{Q} consists of noisy relationships of three different types depicted in the three different colors. (a) The relationship between a statistic $\hat{\varphi}$ and Φ on \mathcal{Q} at a finite sample size. The bottom and top boundaries of each shaded region indicate the $(\alpha/2) \cdot 100\%$ and $(1 - \alpha/2) \cdot 100\%$ percentiles of the sampling distribution of $\hat{\varphi}$ for each relationship type at various values of Φ . The vertical interval (in black) is the \mathcal{Q} -acceptance region $A_\alpha(x)$, and the horizontal interval (in red) is the \mathcal{Q} -confidence interval $I_\alpha(y)$. (b) In the large-sample limit, we replace $\hat{\varphi}$ with a population quantity φ .

2.3 Definition of Equitability via \mathcal{Q} -Confidence Intervals

Proposition 2.1 implies that if the \mathcal{Q} -confidence intervals of $\hat{\varphi}$ with respect to Φ are small then $\hat{\varphi}$ will give good interval estimates of Φ . There are many ways to summarize whether the \mathcal{Q} -confidence intervals of $\hat{\varphi}$ are small; the traditional concept of equitability corresponds to worst-case performance.

DEFINITION 2.3 (Equitability³). For $0 \leq d \leq 1$, the statistic $\hat{\varphi}$ is worst-case $1/d$ -equitable with respect to Φ on \mathcal{Q} with confidence $1 - \alpha$ if and only if the width of $I_\alpha(y)$ is at most d for all y .

Equitability of a measure of dependence $\hat{\varphi}$ therefore simply amounts to a uniform bound on the length of a certain set of confidence intervals constructed from $\hat{\varphi}$. Widths of confidence intervals are commonly used to measure accuracy in many inferential frameworks; Definition 2.3 shows that equitability is a natural application of this concept to a specific set of confidence intervals constructed using a measure of dependence in a semi-parametric scheme. We remark that, as in other settings, one could imagine more fine-grained ways to use widths of confidence intervals to quantify equitability according to, for example, some weighting of the distributions in \mathcal{Q} that reflects a belief about the importance or prevalence of various types of relationships; for simplicity, we do not pursue this here.

The corresponding definition for equitability can be made for φ in the large-sample limit as well (see Appendix A). In that setting, it is possible that all the \mathcal{Q} -confidence intervals of φ with respect to Φ have size 0; that is, the value of $\varphi(\mathcal{Z})$ uniquely determines the value of $\Phi(\mathcal{Z})$. The worst-case equitability of φ is then ∞ , and φ is said to be *perfectly equitable*.

We give a summary of equitability-related terms defined in this section in Table 1.

2.4 Examples of- and Results About Equitability

We provide examples, using the vocabulary developed here, of some concrete instantiations of- and results about equitability. We begin with two examples of statistics that are perfectly equitable in the large-sample limit. First, the mutual information (Cover and Thomas, 2006, Csiszár, 2008) is perfectly equitable with respect to the correlation ρ^2 on the set \mathcal{Q} of bivariate normal random variables. This is because for bivariate normals, $1 - 2^{-2I} = \rho^2$, where I denotes mutual information (Linfoot, 1957). Additionally, Theorem 6 of Székely and Rizzo (2009) shows

³Other literature on this topic occasionally uses the word “interpretability” instead of “equitability” and “interpretable intervals” instead of “ \mathcal{Q} -confidence intervals.” These can be considered synonymous.

TABLE 1

A summary of equitability-related terminology. Equitability-related terms are listed on the left with summaries in standard statistical language on the right

Term	Corresponding statistical object
Set of standard relationships (\mathcal{Q})	A type of <i>model</i> consisting of a set of bivariate distributions on which we can define some notion of relationship strength.
Property of interest (Φ)	A <i>parameter</i> corresponding to the notion of relationship strength for our set of standard relationships.
Measure of dependence ($\hat{\phi}$)	A <i>statistic</i> whose population value is zero exactly under statistical independence, and whose utility as an estimator of Φ we wish to assess.
\mathcal{Q} -acceptance region at x	The <i>acceptance region</i> for a specific test of $H_0 : \Phi = x$ constructed using the measure of dependence $\hat{\phi}$ that is valid over \mathcal{Q} .
\mathcal{Q} -confidence interval at y	The (conservative) <i>confidence interval</i> for Φ at $\hat{\phi} = y$ constructed via inversion of above hypothesis tests.
Equitability	The extent to which we have a <i>uniform bound on widths</i> of the above confidence intervals; a tighter bound corresponds to higher equitability.

that for bivariate normals distance correlation is a deterministic and monotonic function of ρ^2 as well. Therefore, distance correlation is also perfectly equitable with respect to ρ^2 on the set of bivariate normals \mathcal{Q} .

The perfect equitability with respect to ρ^2 on bivariate normals exhibited in both of these examples is one of the “fundamental properties” introduced by Renyi in his framework for thinking about ideal properties of measures of dependence (Rényi, 1959). This property contains a compromise: it guarantees equitability that on the one hand is perfect, but on the other hand applies only on a relatively small set of standard relationships. One goal of equitability is to give us the tools to relax the “perfect” requirement in exchange for the ability to make \mathcal{Q} a larger set, for example, a set of noisy functional relationships. Thus, equitability can be viewed as a generalization of Renyi’s requirement that allows for a tradeoff between the precision with which our statistic tells us about Φ and the set \mathcal{Q} on which it does so.

Renyi’s framework of desiderata for measures of dependence has inspired much follow-up work over the years. For example, Schweizer and Wolff (1981) modified them by weakening several of the invariance requirements and adding a continuity requirement that was satisfied by copula measures. Gretton et al. (2005b) proposed removing the requirement of perfect scores if and only if one variable is a function of the other. Reimherr and Nicolae (2013) proposed a reduced set of axioms focused only on existence, range, and interpretability that allowed for more flexibility in construction of measures of dependence tailored to different areas of application. Móri and Székely (2019) proposed four axioms that emphasized continuity and affine invariance rather than invariance relative to all one-to-one functions of the real line. Our work fits into this continuing conversation about how Renyi’s desiderate should be modified, in our case motivated by the different instances of the data exploration

problem that arise in different fields, each of which may require a different notion of relationship strength but all of which require not just detection but also ranking of relationships of many different kinds.

We next give some examples of—and results about—equitability on noisy functional relationships, as defined below.

DEFINITION 2.4 (Noisy functional relationship).

A random variable distributed over \mathbb{R}^2 is called a noisy functional relationship if and only if it can be written in the form $(X + \varepsilon, f(X) + \varepsilon')$ where $f : [0, 1] \rightarrow \mathbb{R}$, X is a random variable distributed over $[0, 1]$, and ε and ε' are (possibly trivial) random variables independent of each other and of X .

A natural version of equitability to apply to sets of noisy functional relationships is equitability with respect to R^2 . Of course, this definition depends on the set \mathcal{Q} in question. The general approach taken in the literature thus far has been to either (a) fix a set of functions that on the one hand is large enough to be representative of relationships encountered in real data sets and on the other hand is small enough to enable empirical analysis (see, e.g., Reshef et al., 2011, 2018, Kinney and Atwal, 2014, Wang, Jiang and Liu, 2017), as is done when assessing power against independence (see, e.g., Simon and Tibshirani, 2012, Jiang, Ye and Liu, 2015, Heller et al., 2016), or (b) to analyze random sets of relationships drawn from a distribution such as a Gaussian process (Reshef et al., 2016).

As important as the choice of functions to analyze is the choice of marginal distributions and noise model. In past work, we and others have considered several possibilities. The simplest is $X \sim \text{Unif}$, $\varepsilon' \sim \mathcal{N}(0, \sigma^2)$ with σ varying, and $\varepsilon = 0$. Slightly more complex noise models include having ε and ε' be i.i.d. Gaussians, or having

ε be Gaussian and $\varepsilon' = 0$. More complex marginal distributions include having X be distributed in a way that depends on the graph of f , or having it be nonstochastic (Reshef et al., 2011, 2018). Given that we often lack a neat description of the noise or sampling patterns of real data sets, we would ideally like a statistic to be highly equitable on as many different models as possible, and our formalism is designed to be flexible enough to express this.

The larger a noise model is, the harder equitability is to achieve; that is, just as the setting described above in which \mathcal{Q} is the set of bivariate Gaussians is “too easy,” there are settings in which \mathcal{Q} is so large that equitability is “too hard.” This is illustrated by the fact that an impossibility result is known for the following set of relationships, introduced in Kinney and Atwal (2014):

$$\mathcal{Q}_K = \{(X, f(X) + \eta) \mid f : [0, 1] \rightarrow [0, 1], \\ (\eta \perp X) \mid f(X)\}$$

with η representing a random variable that is conditionally independent of X given $f(X)$. This model describes relationships with noise in the second coordinate only, where that noise can depend arbitrarily on the value of $f(X)$ but must be otherwise independent of X .

Kinney and Atwal prove that no nontrivial measure of dependence can be perfectly worst-case equitable with respect to R^2 on the set \mathcal{Q}_K . We note two important limitations of this interesting result, however. The first limitation, pointed out in the technical comment of Murrell, Murrell and Murrell (2014), is that \mathcal{Q}_K is extremely permissive (i.e., large): in particular, the fact that the noise term η can depend arbitrarily on the value of $f(X)$ leads to identifiability issues such as obtaining the noiseless relationship $f(X) = X^2$ as a noisy version of $f(X) = X$. Additionally, since \mathcal{Q}_K is not contained in the other major models considered in, for example, Reshef et al. (2011, 2018), this impossibility result does not imply impossibility for any of those models (Reshef et al., 2014).

An additional limitation of Kinney and Atwal’s result is that it only addresses *perfect* equitability rather than the more general, approximate notion with which we are primarily concerned. While a statistic that is perfectly equitable with respect to R^2 may indeed be difficult or even impossible to achieve for many large models \mathcal{Q} , such impossibility would make *approximate* equitability no less desirable a property. The question thus remains how equitable various measures are, both provably and empirically.

As suggested by the above discussion, the appropriate definitions of \mathcal{Q} and Φ may change from application to application. For instance, rather than using R^2 as the property of interest, one may decide to focus on the discrepancy between the noisy y-values and the corresponding de-noised y-values captured by φ itself, as in the following instantiation of perfect equitability defined in Kinney and Atwal (2014):

DEFINITION 2.5 (Self-equitability (Kinney and Atwal, 2014)⁴). A functional φ is self-equitable if and only if it is symmetric and perfectly equitable on \mathcal{Q}_K with respect to $\Phi(X, f(x) + \eta) = \varphi(f(X), f(X) + \eta)$.

A second possibility is that we might focus on the fraction of deterministic signal in a mixture, as in the following type of equitability, defined in Ding et al. (2017):

DEFINITION 2.6 (Robust equitability (Ding et al., 2017)). Let \mathcal{Q} be the set of all distributions whose copula is of the form $pC_s + (1 - p)\Pi$ for some $0 \leq p \leq 1$, where Π is the independence copula $\Pi(u, v) = uv$ and C_s is a singular copula. A measure of dependence $\hat{\varphi}$ is robust-equitable if it is equitable on \mathcal{Q} with respect to $\Phi(pC_s + (1 - p)\Pi) = p$.

Proving further relationships among these different instantiations of equitability remains an open problem. There are also yet-undefined instantiations that may prove useful if formalized, such as for relationships supported on one-manifolds with additive noise rather than convolution with the independence copula or perhaps even relationships supported on subsets of cells of a predefined grid (Zhang, 2016). In constructing new instantiations, the overarching goal is to have \mathcal{Q} be as large as possible without making it impossible to define a Φ that is appropriate to the question at hand and for which good equitability is achievable.

We emphasize that, although the above discussion considers primarily noisy functional relationships as a simple and illustrative example, nonfunctional relationships can be easily accommodated in our framework, as Definition 2.6 illustrates. For example, one could imagine augmenting the definition of equitability on noisy functional relationships to also require that asymptotically perfect scores be assigned to any union of a finite number of noiseless functional relationships. This would encode, for instance, the intuition that a relationship supported on a noiseless circle is highly interesting and should be discovered. Therefore, while here we focus primarily on the example of noisy functional relationships to elucidate the principles of equitability, noisy functional relationships are not the sole goal of work aimed at achieving equitability.

2.5 Quantifying Equitability: An Example

The formalism above can be used to empirically quantify equitability with respect to R^2 on a specific set of

⁴There is an abuse of notation here because of the identifiability issues with \mathcal{Q}_K discussed above; for example, $f(X) = X^2$ can be a noisy version of $f(X) = X$. Since there can be two identical distributions $Z \in \mathcal{Q}$ corresponding to different functions f , a formal definition would require information about f to be embedded into \mathcal{Q} . If η were restricted to be, for example, mean-zero noise, this modification would not be necessary.

noisy functional relationships. To demonstrate this, we take as an example statistic the sample correlation $\hat{\rho}$. This statistic is of course not a measure of dependence, since its population value can be zero for relationships with non-trivial dependence. We analyze it here solely as an instructional example since it is widely used and behaves intuitively; we provide illustrative analyses of true measures of dependence in Section 5, and refer the reader to Reshef et al. (2016, 2018) for more thorough empirical work on this topic.

Figure 2(a) shows an analysis of the equitability with respect to R^2 of $\hat{\rho}$ at a sample size of $n = 500$ on the set

$$\mathcal{Q} = \{(X, f(X) + \varepsilon'_\sigma) : X \sim \text{Unif}, \varepsilon'_\sigma \sim \mathcal{N}(0, \sigma^2), \\ f \in F, \sigma \in \mathbb{R}_{\geq 0}\},$$

where F is a set of 16 functions analyzed in Reshef et al. (2018). (See Appendix C for details.)

As expected, the \mathcal{Q} -confidence intervals at many values of $\hat{\rho}$ are large. This is because our set of functions F contains many nonlinear functions, and so a given value of $\hat{\rho}$ can be assigned to relationships of different types with very different R^2 values. This is shown by the pairs of thumbnails in the figure, each of which depicts two relationships with the same $\hat{\rho}$ but different values of R^2 . Thus, the analysis confirms that the preference of $\hat{\rho}$ for linear relationships leads it to have poor equitability with respect to R^2 on this set \mathcal{Q} , which contains many nonlinear relationships. In contrast, Figure 2(b) depicts the way this analysis would look for a hypothetical measure of dependence with *perfect* equitability: all the \mathcal{Q} -confidence intervals would have size 0.

2.6 When Is Equitability Useful?

When \mathcal{Q} is so small that there is only one distribution corresponding to every value of Φ , equitability becomes a less rich property. This is because asymptotic monotonicity of $\hat{\phi}$ with respect to Φ is sufficient for perfect equitability in the large-sample limit. In such a scenario, the only obstacle to the equitability of $\hat{\phi}$ is finite-sample effects. For example, on the set \mathcal{Q} of bivariate Gaussians, many measures of dependence are asymptotically perfectly equitable with respect to the correlation.

However, this differs from the motivating data exploration scenario we consider, in which \mathcal{Q} contains many different relationship types and there are multiple different relationships corresponding to a given value of Φ . Here, equitability can be hindered either by finite-sample effects, or by the differences in the asymptotic behavior of $\hat{\phi}$ on different relationship types in \mathcal{Q} .

Regardless of the size of \mathcal{Q} though, equitability is fundamentally meant to be applied to measures of dependence rather than to bespoke estimators of various quantities Φ . (In fact, if $\hat{\phi}$ is a consistent estimator of Φ on \mathcal{Q} , it is trivially asymptotically perfectly equitable.) This

is because in data exploration we typically require that $\hat{\phi}$ be a measure of dependence in order to obtain a minimal guarantee about not missing relationships of unanticipated types, and this requirement typically conflicts with the goal of making $\hat{\phi}$ a consistent estimator of Φ on a large set \mathcal{Q} . For instance, if \mathcal{Q} is a set of noisy functional relationships and Φ is R^2 , then on the one hand computing the sample R^2 with respect to a nonparametric estimate of the regression function will be a consistent estimator of Φ but will miss other interesting relationships that happen to be nonfunctional (e.g., it would give a score of 0 to a circle). And on the other hand, no measure of dependence is known also to be a consistent estimator of R^2 on noisy functional relationships.

In a setting such as this, it is reasonable to seek the next-best thing: a measure of dependence $\hat{\phi}$ whose values have an *approximate* interpretation in terms of R^2 . Equitability supplies us with a way of talking about how well $\hat{\phi}$ does in this regard. In this sense, equitability can be viewed as a weakening of the requirement of consistency: a statistic can, for example, be asymptotically $1/d$ -equitable for some $d > 0$ without being asymptotically perfectly equitable. That is, although the statistic is not a consistent estimator of Φ , it still has the property that its population value gives us information about the value of Φ to within an accuracy of d .

3. EQUITABILITY AND STATISTICAL POWER

3.1 Intuition for Connection Between Equitability and Power

Given our construction of \mathcal{Q} -confidence intervals via the standard technique of inversion of a set of hypothesis tests, it is natural to ask whether there is any connection between equitability and the power of those tests with respect to specific alternatives. We answer this question by showing that equitability can be equivalently formulated in terms of power with respect to a family of null hypotheses corresponding to different relationship strengths. This result recasts equitability as a strengthening of power against statistical independence on \mathcal{Q} and gives a second formal definition of equitability that is easily quantifiable using standard power analysis.

Before stating the formal relationship between equitability and power, let us first state intuitively why it should hold. Recall that the \mathcal{Q} -acceptance region $A_\alpha(x_0)$ is an acceptance region of a two-sided level- α test of $H_0 : \Phi(\mathcal{Z}) = x_0$. Focusing for intuition on $x_0 = 0$, we can ask: what is the minimal $x_1 > 0$ such that a right-tailed level- α test of $H_0 : \Phi = 0$ will have power at least $1 - \beta$ on $H_1 : \Phi = x_1$? As shown graphically in Figure 3, in which $\max A_{2\alpha}(\cdot)$ is an increasing function (and $\alpha = \beta$ for simplicity), the answer can in some cases be stated in terms of the \mathcal{Q} -acceptance regions and the \mathcal{Q} -confidence intervals of $\hat{\phi}$.

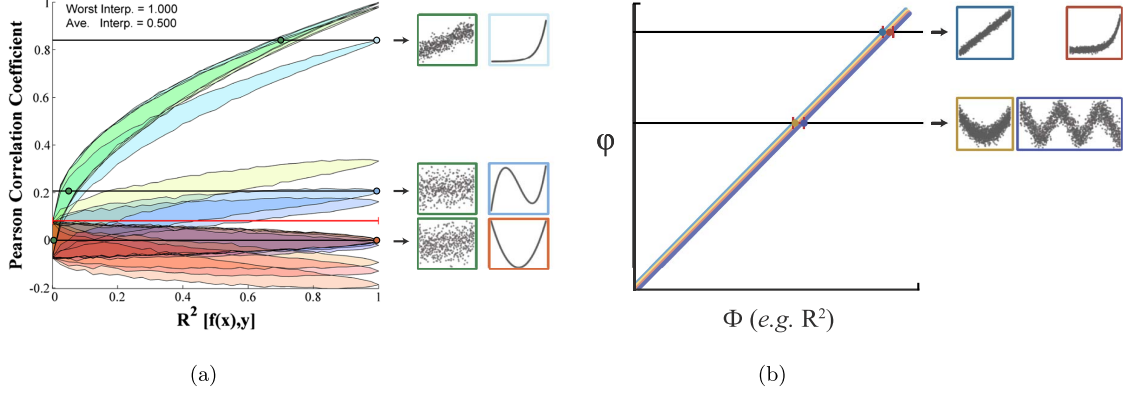


FIG. 2. Examples of equitable and nonequitable behavior on a set of noisy functional relationships. (a) The equitability with respect to R^2 of the Pearson correlation coefficient $\hat{\rho}$ over the set \mathcal{Q} of relationships described in Section 2.5, with $n = 500$. Each shaded region is an estimated 90% central interval of the sampling distribution of $\hat{\rho}$ for a given relationship at a given R^2 . The pairs of thumbnails show relationships with the same $\hat{\rho}$ but different R^2 values. The largest \mathcal{Q} -confidence interval is indicated by a red line. The worst-case and average-case widths of the \mathcal{Q} -confidence intervals are given numerically in the top-left of the plot. (b) A hypothetical population quantity ϕ that achieves the ideal of perfect equitability in the large-sample limit (i.e., the lines corresponding to different relationship types are exactly coincident). This ideal is an illustrative theoretical idea but is not attainable in practice in most interesting cases. Thumbnails are shown for sample relationships that have the same ϕ . See Appendix C for a legend of the function types used.

Specifically, if t_α is the maximal element of $A_{2\alpha}(0)$, then the minimal value of Φ at which a right-tailed test based on $\hat{\phi}$ will achieve power $1 - \beta$ is $\Phi = \max I_{2\beta}(t_\alpha)$, that is, the maximal element of the $(1 - 2\beta)$ \mathcal{Q} -confidence interval at t_α . So if the statistic is highly equitable at t_α , then we will be able to achieve high power against very small departures from the null hypothesis of independence. That is, good equitability on \mathcal{Q} implies good power against independence on \mathcal{Q} . This reasoning holds for null hypotheses beyond independence, and in the converse direction as well, as we state in Theorem 3.1.

3.2 An Equivalent Characterization of Equitability in Terms of Power

To be able to state our result, we need to formally describe how equitability would be formulated in terms of

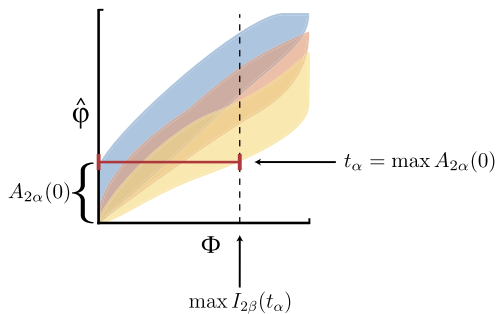


FIG. 3. An illustration of the connection between equitability and power. In this example, we ask for the minimal $x > 0$ that allows a right-tailed level- α test based on $\hat{\phi}$ to achieve power $1 - \beta$ in distinguishing between $H_0 : \Phi = 0$ and $H_1 : \Phi = x$. (For simplicity, the \mathcal{Q} -acceptance regions and \mathcal{Q} -confidence intervals pictured are for the case that $\alpha = \beta$.)

power. This requires two definitions. The first is a definition of a power function that parametrizes the space of possible alternative hypotheses specifically by the property of interest. The second is a definition of a property of this power function called its uncertain interval. It will turn out later that uncertain intervals are \mathcal{Q} -confidence intervals and vice versa. In the definition below, the most permissive member of a set of right-tailed tests based on the same statistic is the one with the smallest critical value.

DEFINITION 3.1. Fix $\alpha, x_0 \in [0, 1]$, and let $T_\alpha^{x_0}$ be the most permissive level- α right-tailed test based on $\hat{\phi}$ of the (possibly composite) null hypothesis $H_0 : \Phi(\mathcal{Z}) = x_0$. For $x_1 \in [0, 1]$, define

$$K_\alpha^{x_0}(x_1) = \inf_{\substack{\mathcal{Z} \in \mathcal{Q} \\ \Phi(\mathcal{Z}) = x_1}} \mathbf{P}(T_\alpha^{x_0}(\mathcal{Z}) \text{ rejects}),$$

where Z is a sample of size n from \mathcal{Z} . That is, $K_\alpha^{x_0}(x_1)$ is the power of $T_\alpha^{x_0}$ with respect to the composite alternative hypothesis $H_1 : \Phi = x_1$.

We call the function $K_\alpha^{x_0} : [0, 1] \rightarrow [0, 1]$ the level- α power function associated to $\hat{\phi}$ at x_0 with respect to Φ .

Note that in the above definition our null and alternative hypotheses may be composite since they are based on Φ and not on a complete parametrization of \mathcal{Q} . That is, \mathcal{Q} can contain several distributions with $\Phi(\mathcal{Z}) = x_0$ or $\Phi(\mathcal{Z}) = x_1$ respectively.

Under the assumption that $\Phi(\mathcal{Z}) = 0$ if and only if \mathcal{Z} represents statistical independence, the power function K_α^0 gives the power of optimal level- α right-tailed tests based on $\hat{\phi}$ at distinguishing various nonzero values of

Φ from statistical independence across the different relationship types in \mathcal{Q} . One way to view the main result of this section is that the set of power functions at values of x_0 besides 0 contains much more information than just the power of right-tailed tests based on $\hat{\phi}$ against the null hypothesis of $\Phi = 0$, and that this information can be equivalently viewed in terms of \mathcal{Q} -confidence intervals. Specifically, we can recover the equitability of $\hat{\phi}$ at every $y \in [0, 1]$ by considering its power functions at values of x_0 beyond 0.

Let us now define the precise aspect of the power functions associated to $\hat{\phi}$ that will allow us to do this.

DEFINITION 3.2. The *uncertain set* of a power function $K_\alpha^{x_0}$ is the set $\{x_1 \geq x_0 : K_\alpha^{x_0}(x_1) < 1 - \alpha\}$.

Our result is then that uncertain sets are \mathcal{Q} -confidence intervals and vice versa.

THEOREM 3.1. Fix a set \mathcal{Q} of distributions, a function $\Phi : \mathcal{Q} \rightarrow [0, 1]$, and $0 < \alpha < 1/2$. Let $\hat{\phi}$ be a statistic with the property that $\max A_{2\alpha}(x)$ is a strictly increasing function of x . Then for all $d \geq 0$, the following are equivalent.

1. $\hat{\phi}$ is worst-case $1/d$ -equitable with respect to Φ with confidence $1 - 2\alpha$.
2. For every $x_0, x_1 \in [0, 1]$ satisfying $x_1 - x_0 > d$, there exists a level- α right-tailed test based on $\hat{\phi}$ that can distinguish between $H_0 : \Phi(\mathcal{Z}) \leq x_0$ and $H_1 : \Phi(\mathcal{Z}) \geq x_1$ with power at least $1 - \alpha$.

The proof of Theorem 3.1, which we defer to Appendix B, is similar to the well-known construction of hypothesis tests from confidence intervals (Casella and Berger, 2002). The main difference is that the usual construction only yields guarantees about the type I error of the resulting tests, whereas here we also provide guarantees about their power on specific alternatives. This is the reason for the monotonicity assumption in the theorem statement.

The characterization of equitability provided by Theorem 3.1 clarifies that the concept of equitability is fundamentally about being able to distinguish not just signal ($\Phi > 0$) from no signal ($\Phi = 0$) but also stronger signal ($\Phi = x_1$) from weaker signal ($\Phi = x_0$), and being able to do so across relationships of different types. This makes sense when a data set contains an overwhelming number of heterogeneous relationships that exhibit, say, $\Phi(\mathcal{Z}) = 0.3$ and that we would like to ignore because they are not as interesting as the small number of relationships with, say, $\Phi(\mathcal{Z}) = 0.8$.

Another advantage of this characterization is that it demonstrates that equitability is related to existing statistical concepts. For instance, the estimation theory literature describes a notion of uniform consistency of an estimator (Yatracos, 1985), which is a guarantee that a statistic

not only converges to a desired population value but does so at a rate that is uniformly bounded across all possible distributions in the model. Equitability, by allowing for fine-grained distinguishability between distributions with difference values of Φ , can be viewed in the asymptotic setting as providing a guarantee that could be translated into an ‘‘approximate’’ uniform consistency for estimation of Φ on \mathcal{Q} . Additionally, the literature on minimax hypothesis testing includes the notion of separation rate (or separation radius in the nonasymptotic setting), which is the minimal distance between two distributions under some metric such that a level- α two-sample test is guaranteed a certain power at distinguishing samples drawn from the distributions (Baraud, 2002). The focus in that setting is to prove minimax separation rates for various two-sample testing problems (Baraud, 2002, Fromont, Lerasle and Reynaud-Bouret, 2016, Arias-Castro, Pelletier and Saligrama, 2018). Equitability, in contrast, is motivated by finding important relationships of all kinds in large-scale data sets via a statistic that can usefully rank the relationships in the data set. This requires assessing the performance of a measure of dependence with respect to the specific distance metric implied by a given one-dimensional notion of relationship strength. Thus, equitability is in a sense a one-dimensional analogue of separation radius.

3.3 Quantifying Equitability via Statistical Power

Theorem 3.1 gives us an alternative to measuring equitability via lengths of \mathcal{Q} -confidence intervals. For every $x_0 \in [0, 1]$ and for every $x_1 > x_0$, we can estimate the power of right-tailed tests based on $\hat{\phi}$ at distinguishing $H_0 : \Phi = x_0$ from $H_1 : \Phi = x_1$. This process is illustrated schematically in Figure 4. In that figure, good equitability corresponds to high power on pairs (x_1, x_0) even when $x_1 - x_0$ is relatively small, and a redder triangle denotes better equitability.

3.4 Equitability Is Stronger than Power Against Independence

Theorem 3.1 shows that equitability is more stringent than the conventional notion of power against independence in three ways.

1. Instead of just one null hypothesis (i.e., $H_0 : \Phi(\mathcal{Z}) = 0$), there are many possible null hypotheses $H_0 : \Phi(\mathcal{Z}) = x_0$ for different values of x_0 .
2. Each of the new null hypotheses can be composite since \mathcal{Q} can contain relationships of many different types (e.g., noisy linear, noisy sinusoidal, and noisy parabolic). Whereas for many measures of dependence all of these relationships may have reduced to a single null hypothesis in the case of statistical independence, they often yield composite null hypotheses once we allow Φ to be nonzero.

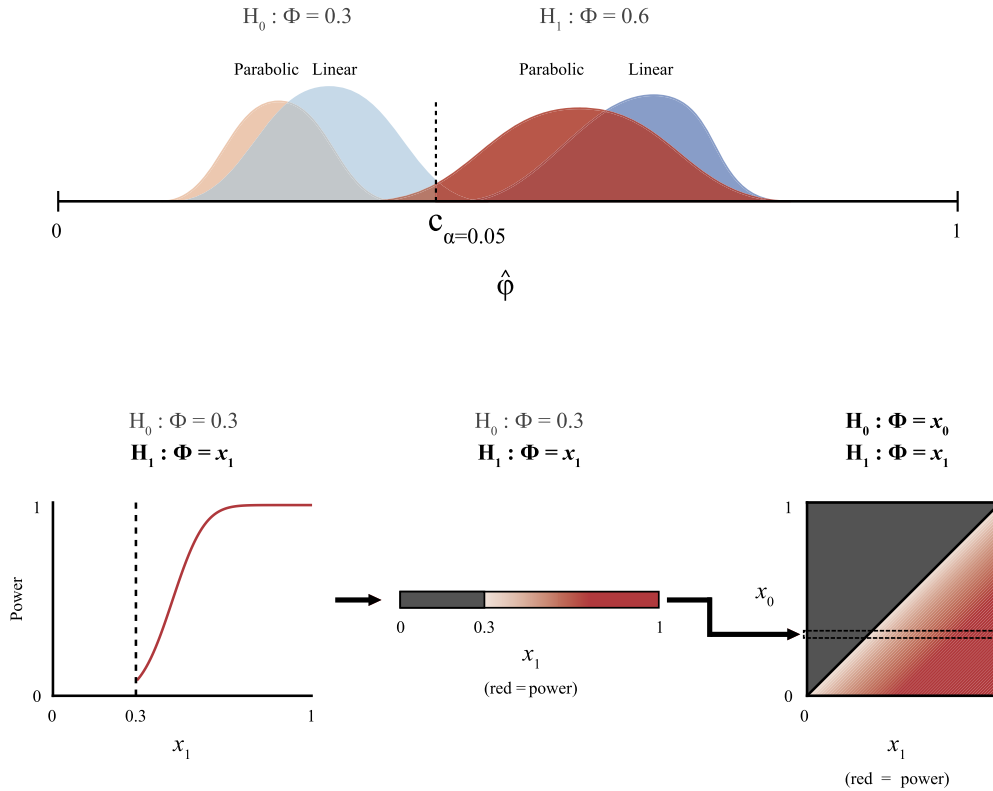


FIG. 4. A schematic illustration of assessment of equitability via statistical power. (Top) The sampling distributions of a test statistic $\hat{\phi}$ when a data set contains only four relationships: a parabolic and a linear relationship, each with either $\Phi = 0.3$ or $\Phi = 0.6$. The dashed line represents the critical value of the most permissive level- α right-tailed test of $H_0 : \Phi = 0.3$. (Bottom left) The power function of the most permissive level- α right-tailed test based on a statistic $\hat{\phi}$ 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 that defines the alternative hypothesis. (Bottom middle) The power function can be depicted instead as a heat map. (Bottom right) Instead of considering just one null hypothesis, we consider a set of null hypotheses (with corresponding critical values) of the form $H_0 : \Phi = x_0$ and plot each of corresponding power curve as a heat map. The result is a plot in which the intensity in the coordinate (x_1, x_0) corresponds to the power of the size- α right-tailed test based on $\hat{\phi}$ at distinguishing $H_1 : \Phi = x_1$ from $H_0 : \Phi = x_0$. A statistic is $1/d$ -equitable with confidence $1 - 2\alpha$ if this power surface attains the value $1 - \alpha$ within distance d of the diagonal along each row. The power curves and heatmap in this figure are schematic and correspond only approximately to the hypothetical distributions shown.

3. The alternative hypotheses are also composite, since each one similarly consists of several different relationship types with the same Φ . Whereas conventional analysis of power against independence considers only one alternative at a time, here we require that tests simultaneously have good power on sets of alternatives with the same Φ .

The understanding that equitability corresponds to power against a much larger set of null hypotheses suggests, via “no free lunch”-type considerations (Simon and Tibshirani, 2012), that if we want to achieve higher power against this larger set of null hypotheses, we may need to give up some power against independence. And indeed, in Reshef et al. (2018) we demonstrate empirically that such a trade-off does seem to exist for several measures of dependence. However, there are situations in which this trade-off is worth making. For instance, in the analysis by Heller et al. (2016) of the gene expression data set discussed earlier in this paper, as well as in a similar analysis of a global health data set (Reshef et al., 2018), several

measures of dependence each detect thousands of significant relationships after correction for multiple hypothesis testing. In such settings it may be worthwhile to sacrifice some power against independence to obtain more information about how to choose among the large number of relationships being detected.

4. EQUITABILITY IMPLIES LOW DETECTION THRESHOLD

The primary motivation given for equitability is that often data sets contain so many relationships that we are not interested in all deviations from independence but rather only in the strongest few relationships. However, there are many data sets in which, due to low sample size, multiple-testing considerations, or relative lack of structure in the data, very few relationships pass significance. Alternatively, there are also settings in which equitability is too ambitious even at large sample sizes. In such settings, we may indeed be interested in simply detecting deviations from independence rather than ranking them by strength.

In this situation, there is still cause for concern about the effect of our choice of test statistic $\hat{\phi}$ on our results. For instance, it is easy to imagine that, despite asymptotic guarantees, an independence test will suffer from low power even on strong relationships of a certain type at a finite sample size n because the test statistic systematically assigns lower scores to relationships of that type. To avoid this, we might want a guarantee that, at a sample size of n , the test has a given amount of power in detecting relationships whose strength as measured by Φ is above a certain threshold, across a broad range of relationship types. This would ensure that, even if we cannot rank relationships by strength, we at least will not miss important relationships as a result of the statistic we use.

There is a simple connection between equitability as defined above and this desideratum, which we call *low detection threshold*. In particular, we show via the alternate characterization of equitability proven in the previous section that low detection threshold is a straightforward consequence of high equitability. Since the converse does not hold, low detection threshold may be a reasonable criterion to use in situations in which equitability is too much to ask.

Given a set \mathcal{Q} of standard relationships, and a property of interest Φ , we define detection threshold as follows.

DEFINITION 4.1 (Detection threshold). A statistic $\hat{\phi}$ has a $(1 - \beta)$ -detection threshold of d at level α with respect to Φ on \mathcal{Q} if there exists a level- α right-tailed test based on $\hat{\phi}$ of the null hypothesis $H_0 : \Phi(\mathcal{Z}) = 0$ whose power on $H_1 : \mathcal{Z}$ at a sample size of n is at least $1 - \beta$ for all $\mathcal{Z} \in \mathcal{Q}$ with $\Phi(\mathcal{Z}) > d$.

Just as equitability is analogous to the notion of separation rate in the context of minimax two-sample testing, detection threshold is likewise analogous the one-sample version of this idea, known as the testing rate. This is the minimal distance between an alternative and the null under some metric such that a level- α test is guaranteed a certain power on that alternative. There is a long and fruitful line of work proving minimax testing rates for various hypothesis testing problems, including nonparametric ones (Ingster, 1987, Lepski and Spokoiny, 1999, Baraud, 2002, Ingster and Suslina, 2003) and even for independence testing (Ingster, 1989, Paninski, 2008, Yodé, 2011, Acharya, Daskalakis and Kamath, 2015, Zhang, 2016). However, in the context of independence testing, testing rate is typically defined in the specific case in which relationship strength is measured by total variation distance. For example, in (Zhang, 2016), a minimax testing rate result is proven for the max BET test in terms of statistical distance from independence, and the fact that this rate is uniformly bounded away from zero on the family of distributions in question is referred to as uniform consistency of the max BET hypothesis test. Our

definition is a natural generalization of this notion that allows for different instantiations to involve different quantifications of relationship strength.

The connection between equitability and low detection threshold is a straightforward corollary of Theorem 3.1.

COROLLARY 4.1. Fix some $0 < \alpha < 1$, let $\hat{\phi}$ be worst-case $1/d$ -equitable with respect to Φ on \mathcal{Q} with confidence $1 - 2\alpha$, and assume that $\max A_{2\alpha}(\cdot)$ is a strictly increasing function. Then $\hat{\phi}$ has a $(1 - \alpha)$ -detection threshold of d at level α with respect to Φ on \mathcal{Q} .

Assume that Φ has the property that it is zero precisely in cases of statistical independence. Then it is easy to see that low detection threshold is an intermediate property that is strictly stronger than asymptotic consistency of independence testing on \mathcal{Q} using $\hat{\phi}$ and strictly weaker than equitability of $\hat{\phi}$ on \mathcal{Q} .

A concrete way to see the utility of low detection threshold is to imagine that we prefilter our data set using some independence test before conducting a more fine-grained analysis with a second statistic. In that case, low detection threshold ensures that we will not “throw out” important relationships prematurely just because of their relationship type. In Reshef et al. (2018), we propose precisely such a scheme, and we analyze the detection threshold of the preliminary test in question to argue that the scheme will perform well.

5. EXAMPLE OF QUANTIFICATION OF EQUITABILITY IN PRACTICE

To concretize the preceding theory, we exhibit an analysis of the equitability on a set of noisy functional relationships of some commonly used methods: the maximal information coefficient as estimated by a new estimator⁵ MIC_e introduced in Reshef et al. (2016), distance correlation (Székely, Rizzo and Bakirov, 2007, Székely and Rizzo, 2009, Huo and Székely, 2016), and Linfoot-transformed mutual information (Linfoot, 1957, Cover and Thomas, 2006) as estimated using the Kraskov estimator (Kraskov, Stögbauer and Grassberger, 2004).

In this analysis, we use $\Phi = R^2$ as our property of interest, $n = 500$ as our sample size, and

$$\mathcal{Q} = \{(x + \varepsilon_\sigma, f(x) + \varepsilon'_\sigma) : x \in X_f, \varepsilon_\sigma, \varepsilon'_\sigma \sim \mathcal{N}(0, \sigma^2), f \in F, \sigma \in \mathbb{R}_{\geq 0}\},$$

where ε_σ and ε'_σ are i.i.d., F is the set of functions in Appendix C, and X_f is the set of n x-values that result

⁵The interested reader may wish to read about MIC_e in the reference provided; however, MIC_e is not the focus of this paper. For the purposes of this paper it can be treated as a black box being used to demonstrate how one would evaluate the equitability of an arbitrary statistic.

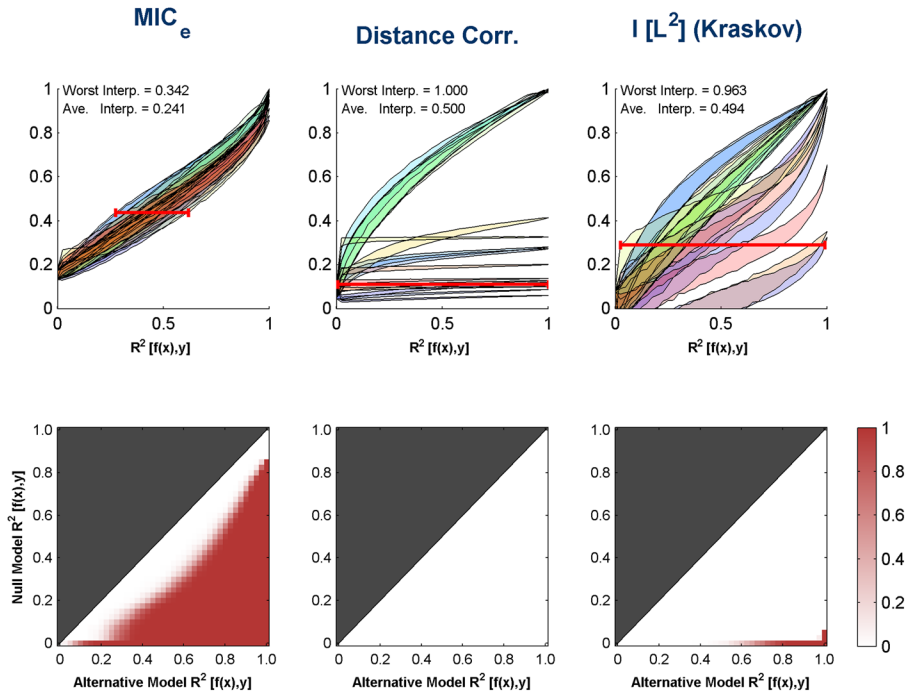


FIG. 5. An analysis of the equitability with respect to R^2 of three measures of dependence on a set of functional relationships. The set of relationships used is described in Section 5. Each column contains results for the indicated measure of dependence. (Top) The analysis visualized via Q -confidence intervals as in Figure 2. [Narrower is more equitable.] The worst-case and average-case widths of the 0.1 Q -confidence intervals for the statistic in question given numerically in the top-left of each plot. (Bottom) The same analysis visualized via statistical power as in Figure 4. [Redder is more equitable.]

in the points $(x_i, f(x_i))$ being equally spaced along the graph of f . Results are shown in Figure 5.

We emphasize that this analysis is intended only as a demonstrative example; for an in-depth empirical evaluation of a comprehensive set of methods under many different settings and with randomly drawn functions, see Reshef et al. (2016, 2018). We remark that, though equitability is an approximate quantity, improving the equitability of existing methods is a worthwhile goal; to that end, exploration of why current methods fail to achieve perfect equitability, say, with respect to R^2 is a valuable avenue of future work.

6. CONCLUSION

In this paper, we formalized and developed the theory of equitability in three ways. We first defined the equitability of a statistic as a uniform bound on the length of a certain set of interval estimates of relationship strength constructed using that statistic; under this view, equitability amounts to an application of ideas from statistical decision theory to assess the extent to which a measure of dependence can be used to perform conservative semi-parametric inference based on extremum quantities. Second, we showed that this formalization of equitability can be equivalently stated in terms of power to distinguish different degrees of (possibly nontrivial) relationship strength from each other; this stands in contrast to

the way that measures of dependence have conventionally been judged, which is only by their power at distinguishing nontrivial signal from statistical independence. Third, we showed that equitability implies the strictly weaker property of a statistic yielding independence tests with a guaranteed minimal power to detect relationships whose strength passes a certain threshold, across a range of relationship types. This property, which we call low detection threshold in the context of measures of dependence, is a natural weaker criterion that one could aim for when equitability proves difficult to achieve.

Our formalization and its results serve three primary purposes. The first is to provide a framework for rigorous discussion and exploration of equitability and related concepts. The second is to clarify the relationship of equitability to central statistical concepts such as confidence and statistical power. The third is to show that equitability and the language developed around it can help us to both formulate and achieve other useful desiderata for measures of dependence.

These connections provide a framework for thinking about the utility of both current and future measure of dependence for exploratory data analysis. Power against independence, the lens through which measures of dependence are currently most often evaluated, is appropriate in many settings in which very few significant relationships are expected, or in which we want to know whether one

specific relationship is nontrivial or not. However, in situations in which most measures of dependence already identify a large number of relationships, a rigorous theory of equitability will allow us to begin to assess when we can glean more information from a given measure of dependence than just the binary result of an independence test.

6.1 Future Work

There is much left to understand about equitability. For instance, to what extent is it achievable for different properties of interest? What are natural and useful properties of interest for sets \mathcal{Q} besides noisy functional relationships? For common statistics, can we obtain a theoretical characterization of the sets \mathcal{Q} and properties Φ for which those statistics achieve good equitability? Are there systematic ways of obtaining equitable behavior via a learning framework as has been done, for example, for causation in Lopez-Paz et al. (2015)? These questions all deserve attention.

Equitability as framed here is certainly not the only goal to which we should strive in developing new measures of dependence. As data sets not only grow in size but also become more varied, there will undoubtedly develop new and interesting use-cases for measures of dependence that will come with new ways of assessing success. Notwithstanding which particular modes of assessment are used, it is important that we formulate and explore concepts that are stronger than power against independence, at least in the bivariate setting. Equitability provides one approach to coping with the changing nature of data exploration. But more generally we can and should ask more of measures of dependence, and this is only one of many possibilities for doing so.

APPENDIX A: NONSTOCHASTIC DEFINITION OF EQUITABILITY

The concepts of \mathcal{Q} -acceptance region and \mathcal{Q} -confidence interval can be defined in the large-sample limit as follows.

DEFINITION A.1 (\mathcal{Q} -acceptance region in the large-sample limit). Let $\varphi : \mathcal{Q} \rightarrow [0, 1]$ be a functional. For $x \in [0, 1]$, the \mathcal{Q} -acceptance region of φ at x , denoted by $A(x)$, is the smallest closed interval containing the set $\varphi(\Phi^{-1}(\{x\}))$.

DEFINITION A.2 (\mathcal{Q} -confidence interval in the large-sample limit). Let $\varphi : \mathcal{Q} \rightarrow [0, 1]$ be a functional. For $y \in [0, 1]$, the \mathcal{Q} -confidence interval of φ at y , denoted by $I(y)$, is the smallest closed interval containing the set $\{x : y \in A(x)\}$.

Equitability is then straightforward to define in the large-sample limit as well.

DEFINITION A.3 (Equitability in the large-sample limit). For $0 \leq d \leq 1$, the functional φ is worst-case $1/d$ -equitable with respect to Φ on \mathcal{Q} if and only if the width of $I(y)$ is at most d for all y .

This definition of equitability quantifies the degree of nonidentifiability induced by φ with respect to Φ on \mathcal{Z} independently of any finite-sample effects.

As mentioned in Section 2, the case of $d = 0$ is referred to as *perfect equitability*. One special case of perfect equitability is when \mathcal{Q} is the set of all bivariate Gaussians and Φ is the squared correlation. In this case, perfect equitability reduces to one of Renyi's fundamental properties of measures of dependence. One trivial way to achieve perfect equitability is to set Φ to be the population value of $\hat{\varphi}$. However, this is not the typical case in which equitability is discussed, as equitability is strictly weaker than consistency of an estimator; see Section 2.6 for details.

APPENDIX B: SUPPLEMENTARY PROOFS

B.1 Proof of Theorem 3.1

Our proof of the alternate characterization of equitability in terms of power requires two short lemmas. The first shows a connection between the maximum element of a \mathcal{Q} -acceptance region and the minimal element of a \mathcal{Q} -confidence interval, namely that these two operations are inverses of each other.

LEMMA B.1. *Given a statistic $\hat{\varphi}$, a property of interest Φ , and some $\alpha \in [0, 1]$, define $f(x) = \max A_\alpha(x)$ and $g(y) = \min I_\alpha(y)$. If f is strictly increasing, then f and g are inverses of each other.*

PROOF. Let $y = f(x) = \max A_\alpha(x)$. By definition, $y \in A_\alpha(x)$, and so $x \in I_\alpha(y)$, which means that $\min I_\alpha(y) \leq x$. On the other hand, for all $x' < x$, $A_\alpha(x') < A_\alpha(x) = y$ by assumption, and so $y \notin A_\alpha(x')$, which means $x' \notin I_\alpha(y)$. \square

The second lemma gives the connection between \mathcal{Q} -acceptance regions and hypothesis testing that we will exploit in our proof.

LEMMA B.2. *Fix a statistic $\hat{\varphi}$, a property of interest Φ , and some $\alpha, x_0 \in [0, 1]$. The most permissive level- $(\alpha/2)$ right-tailed test based on $\hat{\varphi}$ of the null hypothesis $H_0 : \Phi(\mathcal{Z}) = x_0$ has critical value $\max A_\alpha(x_0)$.*

PROOF. We seek the smallest critical value that yields a level- $(\alpha/2)$ test. This would be the supremum, over all \mathcal{Z} with $\Phi(\mathcal{Z}) = x_0$, of the $(1 - \alpha/2) \cdot 100\%$ value of the sampling distribution of $\hat{\varphi}$ when applied to \mathcal{Z} . By definition this is $\max A_\alpha(x_0)$. \square

Theorem 3.1 can then be seen to follow from the proposition below.

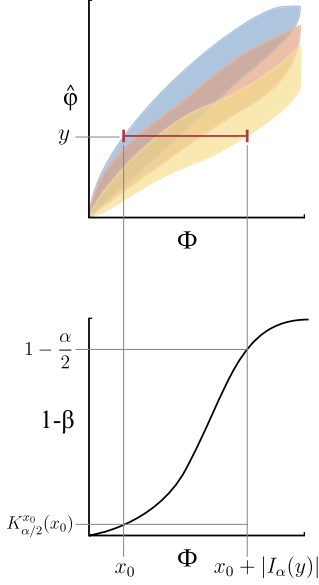


FIG. 6. *The relationship between equitability and power, as in Proposition B.1. The top plot is the same as the one in Figure 1(a), with the indicated interval denoting the \mathcal{Q} -confidence interval $I_\alpha(y)$. The bottom plot is a plot of the power function $K_{\alpha/2}^{x_0}(x)$, with the y-axis indicating statistical power. (Notice that because the null and alternative hypotheses are composite, $K_{\alpha/2}^{x_0}(x_0)$ need not equal $\alpha/2$; in general it may be lower.)*

PROPOSITION B.1. *Fix $0 < \alpha < 1$, and suppose $\hat{\phi}$ is a statistic with the property that $\max A_\alpha(x)$ is a strictly increasing function of x . Then for $y \in [0, 1]$, the interval $I_\alpha(y)$ equals the closure of the uncertain set of $K_{\alpha/2}^{x_0}$ for $x_0 = \min I_\alpha(y)$. Equivalently, for $x_0 \in [0, 1]$, the closure of the uncertain set of $K_{\alpha/2}^{x_0}$ equals $I_\alpha(y)$ for $y = \max A_\alpha(x_0)$.*

An illustration of this proposition and its proof is shown in Figure 6.

PROOF OF PROPOSITION B.1. The equivalence of the two statements follows from Lemma B.1, which states that $y = \max A_\alpha(x_0)$ if and only if $x_0 = \min I_\alpha(y)$. We therefore prove only the first statement, namely that $I_\alpha(y)$ is the uncertain set of $K_{\alpha/2}^{x_0}$ for $x_0 = \min I_\alpha(y)$.

Let U be the uncertain set of $K_{\alpha/2}^{x_0}$. We prove the claim by showing first that $\inf U = \min I_\alpha(y)$, and then that $\sup U = \max I_\alpha(y)$.

To see that $\inf U = \min I_\alpha(y)$, we simply observe that because $\alpha/2 < 1/2$, we have $K_{\alpha/2}^{x_0}(x_0) \leq \alpha/2 < 1 - \alpha/2$, which means that U is nonempty, and so by construction its infimum is x_0 , which we have assumed equals $\min I_\alpha(y)$.

Let us now show that $\sup U \geq \max I_\alpha(y)$: by the definition of the \mathcal{Q} -confidence interval, we can find x arbitrarily close to $\max I_\alpha(y)$ from below such that $y \in A_\alpha(x)$. But this means that there exists some \mathcal{Z} with $\Phi(\mathcal{Z}) = x$ such that if Z is a sample of size n from \mathcal{Z} then

$$\mathbf{P}(\hat{\phi}(Z) < y) \geq \frac{\alpha}{2}$$

that is,

$$\mathbf{P}(\hat{\phi}(Z) \geq y) < 1 - \frac{\alpha}{2}.$$

But since as we already noted $y = \max A_\alpha(x_0)$, Lemma B.2 tells us that it is the critical value of the most permissive level- $(\alpha/2)$ right-tailed test of $H_0 : \Phi(\mathcal{Z}) = x_0$. Therefore, $K_{\alpha/2}^{x_0}(x) < 1 - \alpha/2$, meaning that $x \in U$.

It remains only to show that $\sup U \leq \max I_\alpha(y)$. To do so, we note that $y \notin A_\alpha(x)$ for all $x > \max I_\alpha(y)$. This implies that either $y > \max A_\alpha(x)$ or $y < \min A_\alpha(x)$. However, since $y \in A_\alpha(x_0)$ and $\max A_\alpha(\cdot)$ is an increasing function, no $x > x_0$ can have $y > \max A_\alpha(x)$. Thus, the only option remaining is that $y < \min A_\alpha(x)$. This means that if Z is a sample of size n from any \mathcal{Z} with $\Phi(\mathcal{Z}) = x > \max I_\alpha(y)$, then

$$\mathbf{P}(\hat{\phi}(Z) < y) < \frac{\alpha}{2}$$

that is,

$$\mathbf{P}(\hat{\phi}(Z) \geq y) \geq 1 - \frac{\alpha}{2}.$$

As above, this implies that $K_{\alpha/2}^{x_0}(x) \geq 1 - \alpha/2$, which means that $x \notin U$, as desired. \square

APPENDIX C: DETAILS OF EMPIRICAL ANALYSES

















C.1 Example Quantification of Equitability in Figure 2

To evaluate the equitability of $\hat{\rho}$ in this context, we generate, for each function $f \in F$ and for 41 noise levels chosen for each function to correspond to R^2 values uniformly spaced in $[0, 1]$, 500 independent samples of size $n = 500$ from the relationship $Z_{f,\sigma} = (X, f(X) + \varepsilon'_\sigma)$. We then evaluate $\hat{\rho}$ on each sample to estimate the 5th and 95th percentiles of the sampling distribution of $\hat{\rho}$ on $Z_{f,\sigma}$. By taking, for each σ , the maximal 95th percentile value and the minimal 5th percentile value across all $f \in F$, we obtain estimates of the level-0.1 \mathcal{Q} -acceptance region at each noise level. From the \mathcal{Q} -acceptance regions we can then construct \mathcal{Q} -confidence intervals, and the equitability of $\hat{\rho}$ is the reciprocal of the length of the largest of those intervals.

C.2 Functions Analysed in Figures 2 and 5

Below is the legend showing which function types correspond to the colors in each of Figures 2 and 5. The functions used are the same as the ones in the equitability anal-

yses of Reshef et al. (2018).

	Cosine, High Freq
	Cosine, Non-Fourier Freq [Low]
	Cosine, Varying Freq [Medium]
	Cubic
	Cubic, Y-Stretched
	Exponential [2^{-x}]
	Line
	Linear+Periodic, High Freq
	Linear+Periodic, High Freq 2
	Linear+Periodic, Low Freq
	Linear+Periodic, Medium Freq
	Parabola
	Sine, High Freq
	Sine, Low Freq
	Sine, Non-Fourier Freq [Low]
	Sine, Varying Freq [Medium]

The legend for Figures 2 and 5.

C.3 Parameters Used in Figure 5

In the analysis of the equitability of MIC_e , distance correlation, and mutual information, the following parameter choices were made: for MIC_e , $\alpha = 0.8$ and $c = 5$ were used; for distance correlation no parameter is required; and for mutual information estimation via the Kraskov estimator, $k = 6$ was used. The parameters chosen were the ones that maximize overall equitability in the detailed analyses performed in Reshef et al. (2018). For mutual information, the choice of $k = 6$ (out of the parameters tested: $k = 1, 6, 10, 20$) also maximizes equitability on the specific set \mathcal{Q} that is analyzed in Figure 5.

ACKNOWLEDGEMENTS

The authors would like to acknowledge R. Adams, E. Airoidi, T. Broderick, H. Finucane, A. Gelman, M. Gorfine, R. Heller, J. Huggins, T. Jaakkola, J. Mueller, J. Tenenbaum, R. Tibshirani, the anonymous referees, and the Editor and Associate Editor for constructive conversations and useful feedback.

YAR and DNR were supported by the Paul and Daisy Soros Fellowship. YAR was supported by award No. T32GM007753 from the National Institute of General Medical Sciences and the National Defense Science and Engineering Graduate Fellowship. PCS was supported by the Howard Hughes Medical Institute. MM was supported in part by NSF Grants CCF-1563710 and CCF-1535795.

REFERENCES

- ACHARYA, J., DASKALAKIS, C. and KAMATH, G. (2015). Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems* 3591–3599.
- ARIAS-CASTRO, E., PELLETIER, B. and SALIGRAMA, V. (2018). Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *J. Nonparametr. Stat.* **30** 448–471. [MR3794401](https://doi.org/10.1080/10485252.2018.1435875) <https://doi.org/10.1080/10485252.2018.1435875>
- BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. [MR1935648](https://doi.org/10.1080/10485252.2018.1435875)
- BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619. [MR0803258](https://doi.org/10.1080/01621459.2014.920257)
- CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference, Vol. 2. The Wadsworth & Brooks/Cole Statistics/Probability Series*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA. [MR1051420](https://doi.org/10.1080/01621459.2014.920257)
- COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley Interscience, Hoboken, NJ. [MR2239987](https://doi.org/10.1080/01621459.2014.920257)
- CSISZÁR, I. (2008). Axiomatic characterizations of information measures. *Entropy* **10** 261–273.
- DING, A. A., DY, J. G., LI, Y. and CHANG, Y. (2017). A robust-equitability measure for feature ranking and selection. *J. Mach. Learn. Res.* **18** Paper No. 71, 46. [MR3714234](https://doi.org/10.1080/01621459.2014.920257)
- EMILSSON, V., THORLEIFSSON, G., ZHANG, B., LEONARDSON, A. S., ZINK, F., ZHU, J., CARLSON, S., HELGASON, A., WALTERS, G. B. et al. (2008). Genetics of gene expression and its effect on disease. *Nature* **452** 423–428.
- FAUST, K. and RAES, J. (2012). Microbial interactions: From networks to models. *Nat. Rev., Microbiol.* **10** 538–550. <https://doi.org/10.1038/nrmicro2832>
- FROMONT, M., LERASLE, M. and REYNAUD-BOURET, P. (2016). Family-wise separation rates for multiple testing. *Ann. Statist.* **44** 2533–2563. [MR3576553](https://doi.org/10.1214/15-AOS1418) <https://doi.org/10.1214/15-AOS1418>
- GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005a). Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory. Lecture Notes in Computer Science* **3734** 63–77. Springer, Berlin. [MR2255909](https://doi.org/10.1007/11564089_7) https://doi.org/10.1007/11564089_7
- GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. and SCHÖLKOPF, B. (2005b). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** 2075–2129. [MR2249882](https://doi.org/10.1080/01621459.2014.920257)
- GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. [MR2913716](https://doi.org/10.1080/01621459.2014.920257)
- HELLER, R., HELLER, Y. and GORFINE, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100** 503–510. [MR3068450](https://doi.org/10.1093/biomet/ass070) <https://doi.org/10.1093/biomet/ass070>
- 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** Paper No. 29, 54. [MR3491123](https://doi.org/10.1080/01621459.2014.920257)
- HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* **19** 546–557. [MR0029139](https://doi.org/10.1214/aoms/1177730150) <https://doi.org/10.1214/aoms/1177730150>
- HUO, X. and SZÉKELY, G. J. (2016). Fast computing for distance covariance. *Technometrics* **58** 435–447. [MR3556612](https://doi.org/10.1080/00401706.2015.1054435) <https://doi.org/10.1080/00401706.2015.1054435>
- INGSTER, Y. I. (1989). Asymptotic minimax testing of independence hypothesis. *J. Sov. Math.* **44** 466–476.
- INGSTER, Y. I. (1987). Asymptotically minimax testing of nonparametric hypotheses. In *Probability Theory and Mathematical Statistics, Vol. 1 (Vilnius, 1985)* 553–574. VNU Sci. Press, Utrecht. [MR0901514](https://doi.org/10.1080/01621459.2014.920257)
- INGSTER, Y. I. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. [MR1991446](https://doi.org/10.1007/978-0-387-21580-8) <https://doi.org/10.1007/978-0-387-21580-8>
- JIANG, B., YE, C. and LIU, J. S. (2015). Nonparametric K -sample tests via dynamic slicing. *J. Amer. Statist. Assoc.* **110** 642–653. [MR3367254](https://doi.org/10.1080/01621459.2014.920257) <https://doi.org/10.1080/01621459.2014.920257>
- 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 <https://doi.org/10.1073/pnas.1309933111>
- KRASKOV, A., STÖGBAUER, H. and GRASSBERGER, P. (2004). Estimating mutual information. *Phys. Rev. E* (3) **69** 066138, 16. MR2096503 <https://doi.org/10.1103/PhysRevE.69.066138>
- LEPSKI, O. V. and SPOKOINY, V. G. (1999). Minimax nonparametric hypothesis testing: The case of an inhomogeneous alternative. *Bernoulli* **5** 333–358. MR1681702 <https://doi.org/10.2307/3318439>
- LINFOOT, E. H. (1957). An informational measure of correlation. *Inf. Control* **1** 85–89. MR0092706
- LOPEZ-PAZ, D., HENNIG, P. and SCHÖLKOPF, B. (2013). The randomized dependence coefficient. In *Advances in Neural Information Processing Systems* 1–9.
- LOPEZ-PAZ, D., MUANDET, K., SCHÖLKOPF, B. and TOLSTIKHIN, I. (2015). Towards a learning theory of causation. In *International Conference on Machine Learning (ICML)*.
- MÓRI, T. F. and SZÉKELY, G. J. (2019). Four simple axioms of dependence measures. *Metrika* **82** 1–16. MR3897521 <https://doi.org/10.1007/s00184-018-0670-3>
- MURRELL, B., MURRELL, D. and MURRELL, H. (2014). R2-equitability is satisfiable. *Proc. Natl. Acad. Sci. USA* **111** E2160–E2160.
- MURRELL, B., MURRELL, D. and MURRELL, H. (2016). Discovering general multidimensional associations. *PLoS ONE* **11** e0151551. <https://doi.org/10.1371/journal.pone.0151551>
- PANINSKI, L. (2008). A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory* **54** 4750–4755. MR2591136 <https://doi.org/10.1109/TIT.2008.928987>
- WANG, Y. X. R., WATERMAN, M. S. and HUANG, H. (2014). Gene coexpression measures in large heterogeneous samples using count statistics. *Proc. Natl. Acad. Sci. USA* **111** 16371–16376.
- REIMHERR, M. and NICOLAE, D. L. (2013). On quantifying dependence: A framework for developing interpretable measures. *Statist. Sci.* **28** 116–130. MR3075341 <https://doi.org/10.1214/12-STS405>
- RÉNYI, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hung.* **10** 441–451. MR0115203 <https://doi.org/10.1007/BF02024507>
- 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. 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.
- 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
- RESHEF, D. N., RESHEF, Y. A., SABETI, P. C. and MITZENMACHER, M. (2018). An empirical study of the maximal and total information coefficients and leading measures of dependence. *Ann. Appl. Stat.* **12** 123–155. MR3773388 <https://doi.org/10.1214/17-AOAS1093>
- ROMANO, S., VINH, N. X., VERSPOOR, K. and BAILEY, J. (2018). The randomized information coefficient: Assessing dependencies in noisy data. *Mach. Learn.* **107** 509–549. MR3761295 <https://doi.org/10.1007/s10994-017-5664-2>
- SCHWEIZER, B. and WOLFF, E. F. (1981). On nonparametric measures of dependence for random variables. *Ann. Statist.* **9** 879–885. MR0619291
- SIMON, N. and TIBSHIRANI, R. (2012). Comment on “Detecting novel associations in large data sets.” Unpublished.
- SPEED, T. (2011). A correlation for the 21st century. *Science* **334** 1502–1503.
- STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100** 9440–9445. MR1994856 <https://doi.org/10.1073/pnas.1530509100>
- SUGIYAMA, M. and BORGWARDT, K. M. (2013). Measuring statistical dependence via the mutual information dimension. In *The International Joint Conferences on Artificial Intelligence (IJCAI)* 1692–1698. AAAI Press, Menlo Park, CA.
- SUN, N. and ZHAO, H. (2014). Putting things in order. *Proc. Natl. Acad. Sci. USA* **111** 16236–16237.
- SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127 <https://doi.org/10.1214/09-AOAS312>
- SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665 <https://doi.org/10.1214/009053607000000505>
- TURK-BROWNE, N. B. (2013). Functional interactions as big data in the human brain. *Science* **342** 580–584.
- WANG, X., JIANG, B. and LIU, J. S. (2017). Generalized R-squared for detecting dependence. *Biometrika* **104** 129–139. MR3626486 <https://doi.org/10.1093/biomet/asw071>
- YATRACOS, Y. G. (1985). On the existence of uniformly consistent estimates. *Proc. Amer. Math. Soc.* **94** 479–486. MR0787899 <https://doi.org/10.2307/2045240>
- YODÉ, A. F. (2011). Adaptive minimax test of independence. *Math. Methods Statist.* **20** 246–268. MR2908761 <https://doi.org/10.3103/S1066530711030069>
- ZHANG, K. (2016). Bet on independence. Preprint. Available at [arXiv:1610.05246](https://arxiv.org/abs/1610.05246).