DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Detecting Meaningful Relationships in Large Data Sets

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:40049997

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Detecting Meaningful Relationships in Large Data Sets

a dissertation presented
by
Yakir A. Reshef
to
The John A. Paulson School of Engineering and Applied Sciences

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Computer Science

Harvard University
Cambridge, Massachusetts
March 2018

©2018 – Yakir A. Reshef
all rights reserved.

Advisors: Professors Ryan Adams and Michael Mitzenmacher Yakir A. Reshef

Detecting Meaningful Relationships in Large Data Sets

Abstract

As data sets grow and algorithms scale, two questions have become central to data-rich science. The first is the *exploration* question: how can we avoid only testing hypotheses consistent with current models and instead find new, unanticipated types of relationships that will extend our understanding? The second is the *interpretation* question: given a robust relationship that has been identified, how can we know whether it proves our hypothesis or whether there are other confounders that are responsible for what we see? In this thesis, we develop a set of tools and theory centered around these two questions.

We begin with the exploration question, considering a common scenario in which researchers compute some statistic on every pair of variables in a high-dimensional data set, rank the variable pairs by their scores, and then examine the top of the resulting list. We formulate a theoretical framework for codifying which properties the statistic in question should have in order for this approach to successfully identify new, interesting relationships. We then introduce a suite of tools aimed at achieving these properties,

show through theoretical analysis and simulations that they do so, and demonstrate their practical utility by using them to discover robust, novel relationships in a data set of social, political, and economic indicators collected by the World Health Organization about every country in the world.

We then turn to the interpretation question, specifically in the context of genome-wide association study (GWAS) data. Interpretation of GWAS data is notoriously difficult because tight correlations between nearby genetic variants, along with the multiple biological functions of each individual variant, mean that identified associations are consistent with many different hypotheses about disease mechanism. We posit a new type of genome-wide pattern that, when present, points to a relatively specific set of biological explanations and is therefore highly scientifically informative. We develop a statistic for confidently identifying this type of pattern, show in simulations that it indeed does so, and apply it to GWAS data spanning tens of diseases and complex traits, identifying both known and novel disease genes across a range of human diseases and traits.

Contents

Acknowledgments

I HAVE been the beneficiary of a tremendous amount of support in putting together the work in this thesis. My advisors, Ryan Adams and Michael Mitzenmacher, have been incredible mentors to me both personally and academically. I also owe a special thanks to Alkes Price, with whom I worked very closely on the genetics portion of this thesis, and Pardis Sabeti, who has been a constant source of scientific guidance and emotional support. I'm grateful to the members of the research groups I've worked in and with — the Harvard Intelligent Probabilistic Systems group, the Price lab, the Sabeti lab, and the CGTA discussion group — for their scientific companionship and camaraderie. I'd like to thank my family members, both immediate and extended, and especially my parents for their unending support and fierce commitment to securing for me a better life than they could possibly imagine. I am deeply thankful to my brother, David, for being a phenomenal and fun collaborator, in science as in all things. And finally I thank my wife, Hilary, whose judgement, smile, intellect, and heart sustained me in innumerable ways over the course of this degree, as they have for most of my life.

1

Introduction

Science is a data-driven endeavor. Accordingly, as methods of data collection have evolved, so have the methods — and goals — of analysis. In the 16th century, the need to summarize repeated navigational measurements inspired the use of the median; in the 19th century, the study of heritable traits inspired the introduction of correlation

and regression; and the 20th century saw our increasing ability to measure, quantify, and compute lead to the maturation of statistics and the birth of machine learning.

The massive size and richness of modern data sets, together with the increased sophistication of tools for analyzing them, now means that researchers are better powered than ever to effectively find patterns in their data. However, this highlights a new challenge: not all patterns are equally interesting. Consider, for example, a simple bivariate association. A large sample size may provide enough evidence to establish that two variables are non-randomly related even though the effect of one on the other is small. Additionally, even if the relationship between the two variables is very strong, it could be confounded by a third variable and therefore be limited in its scientific utility. The central question we consider in this thesis is: in data sets with extensive non-random structure, how can we find the deviations from randomness that are the most scientifically interesting?

We structure our work around two different ways of thinking about what constitutes a scientifically interesting association. The first approach considers bivariate associations and posits that the *strongest* such associations in a data set are likely to be the most scientifically interesting. The challenge for this approach is that defining "strongest" typically involves specifying a relationship type — e.g., linear, exponential, non-functional, etc. — and is therefore in tension with the desire to find interesting relationships of all types. In the first part of this thesis, we develop a theoretical framework for codifying what it would mean to have the best of both of these worlds via a property we

define, called *equitability*, that formalizes the intuitive notion of a statistic giving similar scores to equally noisy relationships of many types. We then introduce and theoretically characterize a family of statistics whose goal is to achieve a high degree of equitability. Finally, we perform an extensive empirical analysis of these statistics as well as others, showing that the statistics we introduce have state-of-the-art equitability and can be efficiently and effectively applied to real data sets.

The second approach for thinking about what constitutes a scientifically interesting hypothesis is related to causal inference. Specifically, this approach supposes that *informative* relationships – that is, relationships that point to highly specific hypotheses – are likely to be the most scientifically interesting. In the second part of this thesis, we instantiate this viewpoint in the context of the *genome-wide association study* (GWAS), a study design in which the genomes of many people with and without disease are collected with the goal of finding genetic variants whose presence correlates with disease status and using them to understand the causes of the disease. Modern GWAS have uncovered tens of thousands of areas in the genomic (called *loci*) in which genetic variation correlates with some disease, but in the vast majority of cases the set of possible biological explanations for a given association is extremely broad. For example, each locus contains many potentially causal genes whose signals are difficult to pull apart due to correlations among genetic variants, so it is often unclear which gene is responsible for the disease. And even when a causal gene is identified, it is often unclear via which biological process the gene causes the disease since a single gene can play many different

roles in different tissues of the body or in response to different environmental stimuli. We address this challenge by developing and applying a statistical method for analysis of GWAS data that looks for certain types of patterns that, when they hold across the entire genome rather than at just one locus, point to relatively specific biological mechanisms. In doing so, we provide a step forward toward understanding the molecular basis of complex diseases.

1.1 Overview

1.1.1 Detecting strong relationships

Suppose we have a data set such as the collection of the 356 medical, social, economic, and political indicators measured by the World Health Organization (WHO) about every country in the world. In the first half of this thesis, we address the following *data exploration* question: how can we find the new and interesting relationships in such a rich data set without pre-specifying the types of relationships we are looking for (e.g., linear, exponential, non-functional, etc.)? Even this moderately sized data set contains approximately $64,000$ potential pairwise relationships, far too many for manual examination, and so a computational solution is needed. A common approach among practitioners is to evaluate some statistic on many candidate variable pairs in a data set, sort the variable pairs from highest-scoring to lowest, and manually examine all the pairs above a threshold score.

An appealing class of statistics for this approach is *measures of dependence*, i.e., statistics that are guaranteed in the large-sample limit to be zero in cases of statistical independence and positive otherwise. Measures of dependence are attractive because they promise, asymptotically at least, to detect any deviation from statistical independence and so ensure that no non-trivial relationships will be missed Additionally, there is a long line of fruitful research on such statistics and, consequently, a broad array of methods from which to choose. The utility of a measure of dependence $\hat{\varphi}$ is often assessed by their power against independence, i.e., the power of independence testing based on $\hat{\varphi}$ to detect various types of non-trivial relationships at finite sample sizes.

Our work in this thesis stems from the observation that, while power against independence is an important goal for data sets that have very few non-trivial relationships, or only very weak relationships that are difficult to detect, there are many contexts in which it is not the right goal for addressing the data exploration question. This is because modern measures of dependence are good enough — and modern data sets large enough — that the number of relationships declared statistically significant by a measure of dependence often greatly exceeds the number of relationships that can be explored further by the researcher. For example, biological data sets often contain many non-trivial relationships, but further corroborating any one of them may take extensive manual lab work or a study on human or animal subjects.

In this case, one tempting "fix" is to restrict manual follow-up to a few relationships with the highest values of $\hat{\varphi}$, but the guarantee provided by measures of dependence

says nothing about the rank of the detected relationships, and indeed this approach can skew the direction of follow-up work. For example, many measures of dependence $\hat{\varphi}$ systematically assign higher scores to linear relationships than to non-linear ones. If this is the case, then relatively noisy linear relationships might crowd out strong non-linear relationships from the top-scoring set and prevent the latter from being discovered by the researcher.

In Chapter 2, we formally take up this problem and introduce a new way of assessing measures of dependence, called *equitability*. Intuitively, an equitable statistic is one that, for some measure of relationship strength, assigns similar scores to equally strong relationships regardless of relationship type. (For example, one instantiation of equitability would be to require that on noisy functional relationships a measure of dependence assign similar scores to relationships with the same R^2.) After formalizing this intuition, we prove that under moderate assumptions it is equivalent to requiring that a measure of dependence yield well powered tests not only for distinguishing non-trivial relationships from trivial ones but also for distinguishing stronger relationships from weaker ones. This result suggests a trade-off between equitability and power against independence. 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, a strictly weaker property that we call *low detection threshold*. Thus, when equitability is not achievable, low detection threshold may be a reasonable surrogate.

In Chapter 3, we define and theoretically characterize two new statistics that together yield an efficient approach for obtaining both equitability and power. To do this, we first introduce a new measure of dependence in the large-sample limit, the population maximal information coefficient, and prove three equivalent ways that it can be viewed, including as a canonical "smoothing" of mutual information. We then introduce an efficiently computable consistent estimator of the population maximal information coefficient, and we empirically establish its equitability on a large class of noisy functional relationships including relationships generated by randomly drawn functions. This new statistic has better bias/variance properties and better runtime complexity than previous heuristic approaches. Next, we derive a second, related statistic, the total information coefficient, whose computation is a trivial side-product of our algorithm and whose goal is powerful independence testing rather than equitability. We prove that this statistic yields a consistent independence test and show in simulations that the test has good power against independence. Taken together, the results in this chapter suggest that these two statistics are a valuable pair of tools for exploratory data analysis.

In Chapter 4, we introduce a framework for rigorous empirical evaluation of the power, equitability, and runtime of the statistics we have introduced together with several other leading measures of dependence. Our framework examines many different classes of relationships, noise models, and sample sizes and measures each of our desiderata across different potential parameter settings for each method. Our results generally confirm state-of-the-art performance for the statistics introduced in Chapter 3, and suggest that

a fast and useful strategy for data exploration is to first filter relationships using the total information coefficient and then to rank the remaining ones using the maximal information coefficient. We close with an analysis of the WHO data set described above that demonstrates the utility of this approach.

1.1.2 Detecting informative relationships

In the latter part of this thesis, we turn from the data exploration regime, in which we seek to learn the broadest contours of the dependence structure of a data set, to that of *data interpretation*. That is, if we have a well characterized data set in which many robust relationships have already been found, how can we figure out which relationships in it are maximally scientifically informative? This problem is especially salient in the field of GWAS: modern GWAS data sets, which have been generated over the last decade through tremendous expense and effort, now together contain tens of thousands of robust, reproducible associations between genomic regions and diseases. Ideally, these data would teach us which genes cause common disease, and what biological processes those genes modify in order to do so (e.g., mutations that damage the *CFTR* gene cause cystic fibrosis by hurting the CFTR protein's role in production of sweat, digestive fluids, and mucus). However, despite the wealth of discovered statistical associations, the vast majority of these relationships have not yet been successfully translated into concrete understanding of the biological underpinnings of disease.

Part of the reason that relationships arising from GWAS have been challenging to

interpret is that each relationship has a large number of potential explanations. This arises from two distinct phenomena. The first, called *linkage disequilibrium*, is the presence of strong correlation among nearby genetic variants, sometimes in very long stretches. Linkage disequilibrium makes it difficult to state which variant in a region is responsible for a given association signal. The second challenge is *pleiotropy*, the tendency of a genetic variant to affect many different biological processes rather than just one. Pleiotropy makes it difficult to interpret the meaning of a genetic signal even if linkage disequilibrium is effectively dealt with. For example, a genetic variant that is confidently associated with a disease may affect the expression of two different nearby genes, only one of which is actually causal for the disease in question.

These considerations motivate us to ask in Chapter 5 whether GWAS data sets contain relationships that have highly specific explanations, i.e., relationships that can only be explained by a relatively narrow set of hypotheses. We do this by extending a common analysis strategy called *GWAS enrichment analysis*, in which investigators specify a subset of the genome corresponding to a biological process of interest and ask whether genetic variants in that subset of the genome tend to be important, in aggregate, for some disease of interest. For example, a biological experiment might nominate certain genes as being important for a process such as inflammation, and a researcher might then check whether variants in and around those genes are "enriched" for variants that correlate with an autoimmune disease.

GWAS enrichment analysis is a valuable, tool, but the interpretation of a result arising

from such an analysis can be difficult. This is because biological processes often either localize near- (linkage disequilibrium) or with- (pleiotropy) each other in the genome. Therefore, if an un-modeled or un-measured biological process is important for disease, this might lead to signals for other processes that are located near it in the genome. For example, the inflammation gene set described above might be a subset of the set of genes expressed in immune cells. Since genes expressed in immune cells are generally enriched for disease signal for auto-immune disease, a naive analysis may see an inflammation enrichment that is simply a side effect of the more generic immune-cell enrichment.

Our work is based on the idea that in some cases, we can determine not just whether a genetic variant is important for a process or not, but also whether it *promotes* or *hinders* that process. This allows us to be choosier in our search: if we see that variants that promote inflammation tend to increase disease risk across the genome, this is no longer attributable to a generic enrichment of genes expressed in immune cells. The set of potential confounders is restricted to biological process that are affected in the same direction as inflammation by the genetic variants in question. Therefore, if we can estimate the directional effect of variants on some biological process, we can use this to find relationships with much more specific explanations.

With this motivation and these data in mind, we develop a statistical method, signed LD profile regression, for taking GWAS data together with signed information relating genetic variants to a biological process, and estimating the effect of the process on disease. After proving the consistency of our statistic for estimating this effect, we assess

it in detailed simulations with real genetic data and simulated phenotypes, showing that it indeed is not confounded by co-localization between the process in question and other processes that are important for the trait. We first apply our method to GWAS data in which the phenotypes are molecular traits, such as gene expression levels, and recover biological programs that are known to, e.g., activate or repress gene expression in different tissues in the body. Some of these results are known, though to our knowledge they have not previously been demonstrated directly using human genetic data, and several are new and intriguing. We next apply our method to GWAS data about diseases and complex traits, where we discover several new relationships that strengthen and extend the current understanding of diseases including intellectual disability, systemic lupus erythematosus, and Crohn's disease. Our method provides a new, orthogonal way of interpreting GWAS data that can improve our ability to translate genetic associations into disease mechanisms.

2

Equitability, interval estimation, and statistical power

Emerging high-dimensional data sets often contain many non-trivial relationships, and, at modern sample sizes, screening these using an independence test can yield too

many relationships to be a useful exploratory approach. In this chapter, 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 in terms of interval estimates of relationship strength, and then show that under moderate assumptions it is equivalent to requiring that a measure of dependence yield well powered tests not only for distinguishing non-trivial 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.*

2.1 Background

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, i.e., a statistic whose population value

The material in this chapter is adapted from a manuscript posted to arXiv as "Equitability, interval estimation, and statistical power" by Yakir Reshef, David Reshef*, *et al.*[133] that as of this writing is in submission at *Statistical Science*.

is zero exactly in cases of statistical independence, on all 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[149,35]. When faced with this task, there is a wealth of measures of dependence from which to choose[60,14,79,52,154,153,121,51,58,152,59,86,66,130,166,135].

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.[59], 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[39,162]. However, this idea has a pitfall: while a measure of dependence guarantees non-zero scores to dependent variable pairs, the magnitude of these non-zero 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[39]. For example, if some measure of dependence $\hat{\varphi}$ systematically assigns higher scores to, say, linear relationships than to non-linear relationships, then using $\hat{\varphi}$ to rank variable pairs in a large data set could cause noisy linear relationships in the data set to crowd out strong non-linear relationships from the top of the list. The natural result would be that the human examining the top-ranked relationships would never see the non-linear relationships, and they would not be discovered[146].

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.

In this chapter we propose and formally characterize *equitability*, a framework for meeting this goal. While 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[121], here we formalize this notion in the language of estimation theory and tie it to the theory of hypothesis testing.

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 $\hat{\varphi}$, a benchmark set $\boldsymbol{\mathcal{Q}}$ of relationship types, and some quantification Φ of relationship strength defined on $\boldsymbol{\mathcal{Q}}$, we construct an interval estimate of the relationship strength Φ from the value of $\hat{\varphi}$ that is valid for any relationship in $\boldsymbol{\mathcal{Q}}$. We then propose using the sizes of these intervals to

quantify the utility of $\hat{\varphi}$ 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 Renyi in his framework for measures of dependence[120].

After defining equitability, we connect it to notions of statistical power using the 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 non-trivial. This result gives us a way to understand equitability as a natural strengthening of the traditional requirement of power against independence, in which we ask only that our statistic be useful for detecting deviations from strict independence (i.e., distinguishing zero relationship strength from non-zero relationship strength).

Finally, motivated by the connection between equitability and power, we define a new 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 $\mathcal{Q}$. We 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[99,130,32,166,135], a framework

for rigorously thinking about this property is becoming increasingly important. The results we present in this chapter 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 and explain the limitations of the alternative definitions and accompanying results of Kinney & Atwal[76], as well as to crystallize and conceptually discuss the power against independence of equitable methods[144].

Throughout this chapter, we give concrete examples of how our formalism relates to the analysis of equitability in practice, and we close with a demonstration of an example empirical analysis of the equitability of a few popular measures of dependence. This analysis is meant to be illustrative, with more empirical work appearing in Chapter 3 and especially in Chapter 4.

2.2 Defining equitability

2.2.1 Preliminaries

Suppose we are given a statistic $\hat{\varphi}$ taking values in $[0, 1]$ that is a measure of dependence. To formally define what it means for $\hat{\varphi}$ 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 $\boldsymbol{\mathcal{Q}}$ of distributions called *standard relationships*, on which we have a well defined notion of relationship strength in the form of a scalar-

valued functional $\Phi : \boldsymbol{\mathcal{Q}} \to [0, 1]$ that we call the *property of interest*. The idea is that $\boldsymbol{\mathcal{Q}}$ contains relationships of many different types, and for any distribution $\mathcal{Z} \in \boldsymbol{\mathcal{Q}}$, $\Phi(\mathcal{Z})$ is the way we would ideally quantify the strength of $\mathcal{Z}$ if we had knowledge of the distribution $\mathcal{Z}$. Our goal is then, given a sample Z of size n from $\mathcal{Z}$, to quantify how well $\hat{\varphi}(Z)$ can be used to draw inferences about $\Phi(\mathcal{Z})$.

We keep our exposition generic in order to accommodate variations – both existing[76,99,32,166] and potential – on the concepts defined here. However, as a motivating example, we often return to the setting in which $\boldsymbol{\mathcal{Q}}$ is a set of noisy functional relationships and Φ is the coefficient of determination (R^2) with respect to the generating function, i.e., the squared Pearson correlation between the dependent variable and the generating function evaluated on the independent variable.

2.2.2 Q-CONFIDENCE INTERVALS

Our approach to defining equitability is to construct from $\hat{\varphi}$ an interval estimate of Φ by inverting a certain set of hypothesis tests. The statistic $\hat{\varphi}$ 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. (In this definition as well as later definitions, we implicitly assume a fixed sample size of n.)

Definition 2.2.1 ($\boldsymbol{\mathcal{Q}}$-acceptance region)**.** Let $\hat{\varphi}$ be a statistic taking values in $[0, 1]$, and let $x, \alpha \in [0, 1]$. The level-α $\boldsymbol{\mathcal{Q}}$-acceptance region of $\hat{\varphi}$ at x, denoted by $R_{\alpha}^{\hat{\varphi}}(x)$, is

the smallest closed interval A with the property that, for all $\mathcal{Z} \in \boldsymbol{\mathcal{Q}}$ with $\Phi(\mathcal{Z}) = x$, we have

$$\mathbf{P}\left(\hat{\varphi}(Z) < \min A\right) < \alpha/2 \quad \text{and} \quad \mathbf{P}\left(\hat{\varphi}(Z) > \max A\right) < \alpha/2$$

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

See Figure 2.1a for an illustration. The $\boldsymbol{\mathcal{Q}}$-acceptance region of $\hat{\varphi}$ at x is an acceptance region for one particular test of the null hypothesis $H_0 : \Phi(\mathcal{Z}) = x$ on relationships in $\boldsymbol{\mathcal{Q}}$. If there is only one $\mathcal{Z} \in \boldsymbol{\mathcal{Q}}$ satisfying $\Phi(\mathcal{Z}) = x$, it amounts to a central interval of the sampling distribution of $\hat{\varphi}$ 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 $\hat{\varphi}$ on all the distributions $\mathcal{Z}$ in question. For example, when $\boldsymbol{\mathcal{Q}}$ is a set of noisy functional relationships with several different function types and Φ is R^2, the equitable acceptance region at x is the smallest interval A such that for any functional relationship $\mathcal{Z} \in \boldsymbol{\mathcal{Q}}$ with $R^2(\mathcal{Z}) = x$, $\hat{\varphi}(Z)$ 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 $R^{\hat{\varphi}}_{\alpha}(x)$ via the standard approach of inversion of hypothesis tests[17].

Definition 2.2.2 ($\boldsymbol{\mathcal{Q}}$-confidence interval)**.** Let $\hat{\varphi}$ be a statistic taking values in $[0, 1]$, and let $y, \alpha \in [0, 1]$. The $(1 - \alpha)$ $\boldsymbol{\mathcal{Q}}$-confidence interval of $\hat{\varphi}$ at y, denoted by $I^{\hat{\varphi}}_{\alpha}(y)$, is

the smallest closed interval containing the set

$$\left\{x \in [0,1] : y \in R_\alpha^{\hat{\varphi}}(x)\right\}.$$

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

See Figure 2.1a for an illustration. Our construction gives us the following guarantee about the coverage probability of the $\boldsymbol{\mathcal{Q}}$-confidence intervals, whose proof we omit.

Proposition 2.2.3. *Let $\hat{\varphi}$ be a statistic taking values in $[0,1]$, and let $\alpha \in [0,1]$. For all $x \in [0,1]$ and for all $\mathcal{Z} \in \boldsymbol{\mathcal{Q}}$,*

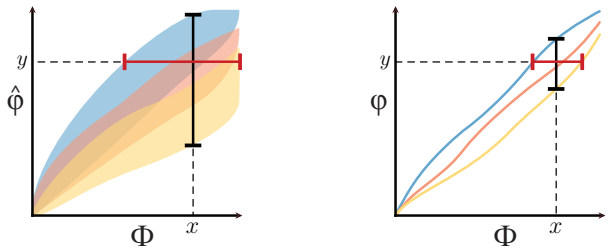
$$\mathbf{P}\left(\Phi(\mathcal{Z}) \in I_\alpha^{\hat{\varphi}}(\hat{\varphi}(Z))\right) \geq 1 - \alpha$$

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

The definitions just presented have natural non-stochastic counterparts in the large-sample limit, which we omit, that quantify the degree of non-identifiability induced by φ with respect to Φ on $\mathcal{Z}$ independently of any finite-sample effects. See Figure 2.1b for an illustration.

2.2.3 Definition of equitability via Q-confidence intervals

Proposition 2.2.3 implies that if the $\boldsymbol{\mathcal{Q}}$-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



(a) (b)

Figure 2.1: 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 $R_\alpha^{\hat{\varphi}}(x)$, and the horizontal interval (in red) is the $\mathcal{Q}$-confidence interval $I_\alpha^{\hat{\varphi}}(y)$. (b) In the large-sample limit, we replace $\hat{\varphi}$ with a population quantity φ.

whether the $\mathcal{Q}$-confidence intervals of $\hat{\varphi}$ are small; the traditional concept of equitability corresponds to worst-case performance.

Definition 2.2.4 (Equitability[†])**.** For $0 \leq d \leq 1$, the statistic $\hat{\varphi}$ is worst-case $1/d$-*equitable* with respect to Φ on $\mathcal{Q}$ at y with confidence $1-\alpha$ if and only if the width of $I_\alpha^{\hat{\varphi}}(y)$ is at most d for all y.

One could imagine more fine-grained ways to quantify equitability according to, for example, some prior over 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.

[†]Other 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.

The corresponding definition for equitability can be made for φ in the large-sample limit as well. 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*.

2.2.4 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[23,25] 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[84]. Additionally, Theorem 6 of[153] shows that for bivariate normals distance correlation is a deterministic 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[120]. 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, e.g., 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.

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

Definition 2.2.5 (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 $f : [0, 1] \to \mathbb{R}$, 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.[121],[126], Kinney & Atwal[76], Wang et al.[166]), as is done when assessing power against independence (see, e.g., Simon & Tibshirani[144], Jiang et al.[66], Heller et al.[59]), or b) to analyze random sets of relationships drawn from a distribution such as a Gaussian process[130].

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 $X \sim$ Unif, $\varepsilon' \sim \mathcal{N}(0, \sigma^2)$ with σ varying, and $\varepsilon = 0$. Slightly more complex noise models include having ε and ε' 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 non-stochastic[121,126]. Given that we often lack a neat description of the noise in 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 handle general models that include arbitrary such variations.

The larger a noise model is, the harder equitability is to achieve; that is, just as the setting described above in which $\boldsymbol{\mathcal{Q}}$ is the set of bivariate Gaussians is "too easy", there are settings in which $\boldsymbol{\mathcal{Q}}$ 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 & Atwal[76].

$$\boldsymbol{\mathcal{Q}}_K = \left\{ (X, f(X) + \eta) \;\middle|\; f : [0,1] \to [0,1], (\eta \perp X) | f(X) \right\}$$

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 non-trivial measure of dependence can be perfectly worst-case equitable with respect to R^2 on the set $\boldsymbol{\mathcal{Q}}_K$. We note two important limitations of this interesting result, however. The first limitation, pointed out in the technical comment of Murrell et al.[98], is that $\boldsymbol{\mathcal{Q}}_K$ 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 $\boldsymbol{\mathcal{Q}}_K$ is not contained in the other major models considered in, e.g.,[121] and[126], this impossibility result does not imply impossibility for any of those models[123].

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 $\boldsymbol{\mathcal{Q}}$, 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 $\boldsymbol{\mathcal{Q}}$ and Φ may change from application to application. For instance, instead of functional relationships one may be interested in relationships supported on one-manifolds, with added noise. Or perhaps instead of R^2 one may decide to focus on the mutual information between the sampled y-values and the corresponding de-noised y-values, as in a different variant

of perfect equitability defined in Kinney & Atwal[76], or on the fraction of deterministic signal in a mixture, as in the type of equitability defined in Ding et al.[32]. In each case the overarching goal should be to have $\boldsymbol{\mathcal{Q}}$ 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.

2.2.5 Quantifying equitability: an example

The formalism above can be used to empirically quantify equitability with respect to R^2 on a specific set of 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 2.5, and refer the reader to Reshef et al.[130],[126] for more thorough empirical work on this topic.

Figure 2.2a shows an analysis of the equitability with respect to R^2 of $\hat{\rho}$ at a sample size of $n = 500$ on the set

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

where F is a set of 16 functions analyzed in[126]. (See Appendix A.2 for details.)

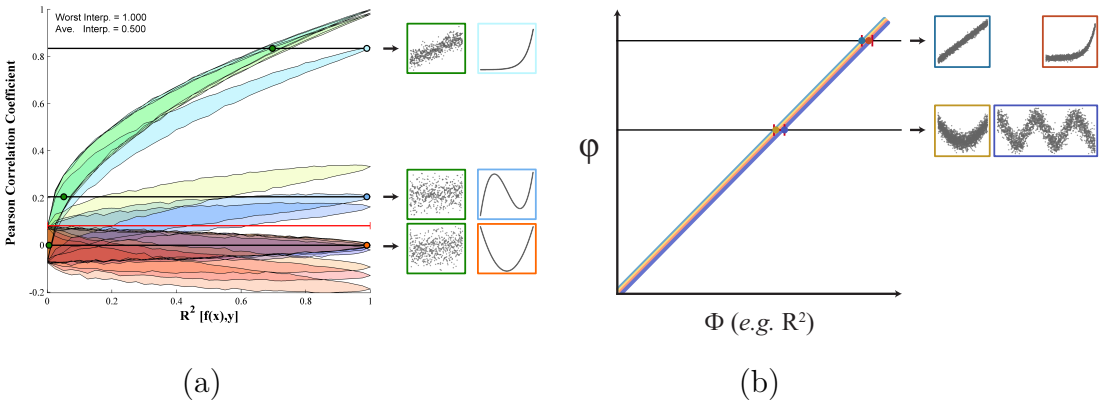


Figure 2.2: Examples of equitable and non-equitable behavior on a set of noisy functional relationships. (a) The equitability with respect to R^2 of the Pearson correlation coefficient $\hat{\rho}$ over the set $\mathcal{Q}$ of relationships described in Section 2.2.5, with $n = 500$. Each shaded region 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. (b) A hypothetical population quantity φ that achieves *perfect equitability* in the large-sample limit. Thumbnails are shown for sample relationships that have the same φ. See Appendix A.2 for a legend of the function types used.

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 non-linear 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 non-linear relationships. In contrast, Figure 2.2b 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.

2.2.6 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 $\hat{\varphi}$ 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 $\hat{\varphi}$ 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 $\hat{\varphi}$ on different relationship types in $\mathcal{Q}$.

Regardless of the size of $\mathcal{Q}$ though, equitability is fundamentally meant for a situation in which simply estimating Φ directly is undesirable. (In fact, if $\hat{\varphi}$ is a consistent estimator of Φ on $\mathcal{Q}$, it is trivially perfectly equitable in the large-sample limit.) This is because in data exploration we typically require that $\hat{\varphi}$ be a measure of dependence in order to obtain a minimal robustness guarantee, and this requirement makes it very difficult to make $\hat{\varphi}$ a consistent estimator of Φ on a large set $\mathcal{Q}$. For instance, suppose $\mathcal{Q}$ is a set of noisy functional relationships and $\Phi = R^2$. Here, computing the sample R^2 relative to a non-parametric estimate of the generating function will be asymptotically

perfectly equitable. However, this approach is undesirable for data exploration because of its lack of robustness, as exemplified by the fact that it would assign a score of zero to, e.g., a circular relationship since the regression function of that relationship is constant. Therefore, we are left with the problem of finding the next-best thing: a measure of dependence $\hat{\varphi}$ whose values have a clear, if approximate, interpretation in terms of Φ. Equitability supplies us with a way of talking about how well $\hat{\varphi}$ does in this regard.

2.3 Equitability and statistical power

2.3.1 Intuition for connection between equitability and power

Given our construction of $\boldsymbol{\mathcal{Q}}$-confidence intervals via 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 re-casts equitability as a strengthening of power against statistical independence on $\boldsymbol{\mathcal{Q}}$ 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 $\boldsymbol{\mathcal{Q}}$-acceptance region $R_\alpha(x_0)$ is an acceptance region of a two-sided level-α test of $H_0 : \Phi(\mathcal{Z}) = x_0$. Focusing for intuition

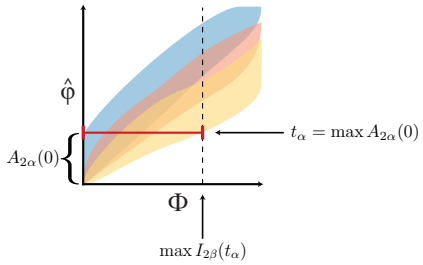


Figure 2.3: An illustration of the connection between equitability and power.

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 2.3, the answer can be stated in terms of the $\mathcal{Q}$-acceptance regions and the $\mathcal{Q}$-confidence intervals of $\hat{\varphi}$.

Specifically, if t_α is the maximal element of $R_{2\alpha}(0)$, then the minimal value of Φ at which a right-tailed test based on $\hat{\varphi}$ will achieve power $1-\beta$ is $\Phi = \max I_{2\beta}(t_\alpha)$, i.e., 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 2.3.1.

2.3.2 Equivalent view 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 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 than uncertain intervals are $\boldsymbol{\mathcal{Q}}$-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 2.3.1. Fix $\alpha, x_0 \in [0,1]$, and let $T_\alpha^{x_0}$ be the most permissive level-α right-tailed test based on $\hat{\varphi}$ of the (possibly composite) null hypothesis $H_0 : \Phi(\mathcal{Z}) = x_0$. For $x_1 \in [0,1]$, define

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

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 $K_\alpha^{x_0} : [0,1] \to [0,1]$ the *level-α power function* associated to $\hat{\varphi}$ 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 $\boldsymbol{\mathcal{Q}}$. That is, $\mathcal{Z}$ can be one of 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 $\hat{\varphi}$ at distinguishing various non-zero values of Φ from statistical independence across

the different relationship types in $\boldsymbol{\mathcal{Q}}$. 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 $\hat{\varphi}$ against the null hypothesis of $\Phi = 0$, and that this information can be equivalently viewed in terms of $\boldsymbol{\mathcal{Q}}$-confidence intervals. Specifically, we can recover the equitability of $\hat{\varphi}$ 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 $\hat{\varphi}$ that will allow us to do this.

Definition 2.3.2. 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 $\boldsymbol{\mathcal{Q}}$-confidence intervals and vice versa.

Theorem 2.3.1. *Fix a set $\boldsymbol{\mathcal{Q}} \subset \boldsymbol{\mathcal{P}}$, a function $\Phi : \boldsymbol{\mathcal{Q}} \to [0, 1]$, and $0 < \alpha < 1/2$. Let $\hat{\varphi}$ be a statistic with the property that $\max R_{2\alpha}(x)$ is a strictly increasing function of x. Then for all $d \geq 0$, the following are equivalent.*

1. *$\hat{\varphi}$ 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 $\hat{\varphi}$ 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 $1 - \alpha$.*

This characterization 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, $\Phi(\mathcal{Z}) = 0.8$.

2.3.3 Quantifying equitability via statistical power

Theorem 2.3.1 gives us an alternative to measuring equitability via lengths of $\boldsymbol{\mathcal{Q}}$-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 $\hat{\varphi}$ at distinguishing $H_0 : \Phi = x_0$ from $H_1 : \Phi = x_1$. This process is illustrated schematically in Figure 2.4. In that figure, good equitability corresponds to high power on pairs (x_1, x_0) such that $x_1 - x_0$ is small, and a redder triangle denotes better equitability.

2.3.4 Equitability is stronger than power against independence

Theorem 2.3.1 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 $\boldsymbol{\mathcal{Q}}$ 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

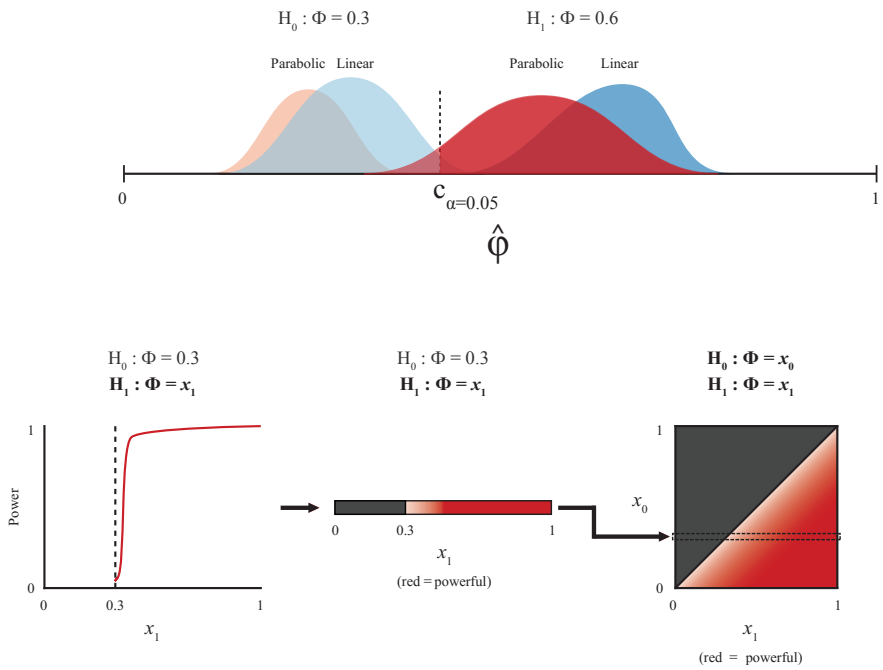


Figure 2.4: A schematic illustration of assessment of equitability via statistical power. (Top) The sampling distributions of a test statistic $\hat{\varphi}$ when a data set contains only four relationships: a parabolic and a linear relationship, each with either $\Phi = 0.3$ or $\Phi = 0.6$. The dashed 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 $\hat{\varphi}$ of the null hypothesis $H_0 : \Phi = 0.3$. The curve shows the power of the test as a function of x_1, the value of Φ 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 $\hat{\varphi}$ 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.

may have reduced to a single null hypothesis in the case of statistical independence, they very often yield composite null hypotheses once we allow Φ to be non-zero.

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[144], 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.[126] 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[59] of the gene expression data set discussed earlier in this chapter, as well as in a similar analysis of a global health data set in Chapter 4, 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.

2.4 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 $\hat{\varphi}$ 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 low detection threshold as follows.

Definition 2.4.1 (Detection threshold)**.** A statistic $\hat{\varphi}$ has a $(1-\beta)$-*detection threshold* of d at level α with respect to Φ on $\boldsymbol{\mathcal{Q}}$ if there exists a level-α right-tailed test based on $\hat{\varphi}$ 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 $\mathcal{Z} \in \boldsymbol{\mathcal{Q}}$ with $\Phi(\mathcal{Z}) > d$.

The connection between equitability and low detection threshold is then a straightforward corollary of Theorem 2.3.1.

Corollary 2.4.2. *Fix some* $0 < \alpha < 1$*, let* $\hat{\varphi}$ *be worst-case* $1/d$*-equitable with respect to* Φ *on* $\boldsymbol{\mathcal{Q}}$ *with confidence* $1-2\alpha$*, and assume that* $\max R_{2\alpha}(\cdot)$ *is a strictly increasing function. Then* $\hat{\varphi}$ *has a* $(1-\alpha)$*-detection threshold of* d *at level* α *with respect to* Φ *on* $\boldsymbol{\mathcal{Q}}$*.*

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 $\boldsymbol{\mathcal{Q}}$ using $\hat{\varphi}$ and strictly weaker than equitability of $\hat{\varphi}$ on $\boldsymbol{\mathcal{Q}}$.

A concrete way to see the utility of low detection threshold is to imagine that we pre-filter 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.[126], 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.

2.5 Quantification of equitability of measures of dependence

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[130], distance correlation[154,153,65], and Linfoot-transformed mutual information[84,23] as estimated using the Kraskov estimator[79].

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 A.2, and X_f is the set of n x-values that result in the points $(x_i, f(x_i))$ being equally spaced along the graph of f. Results are shown in Figure 2.5.

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.[126] and Reshef et al.[130].

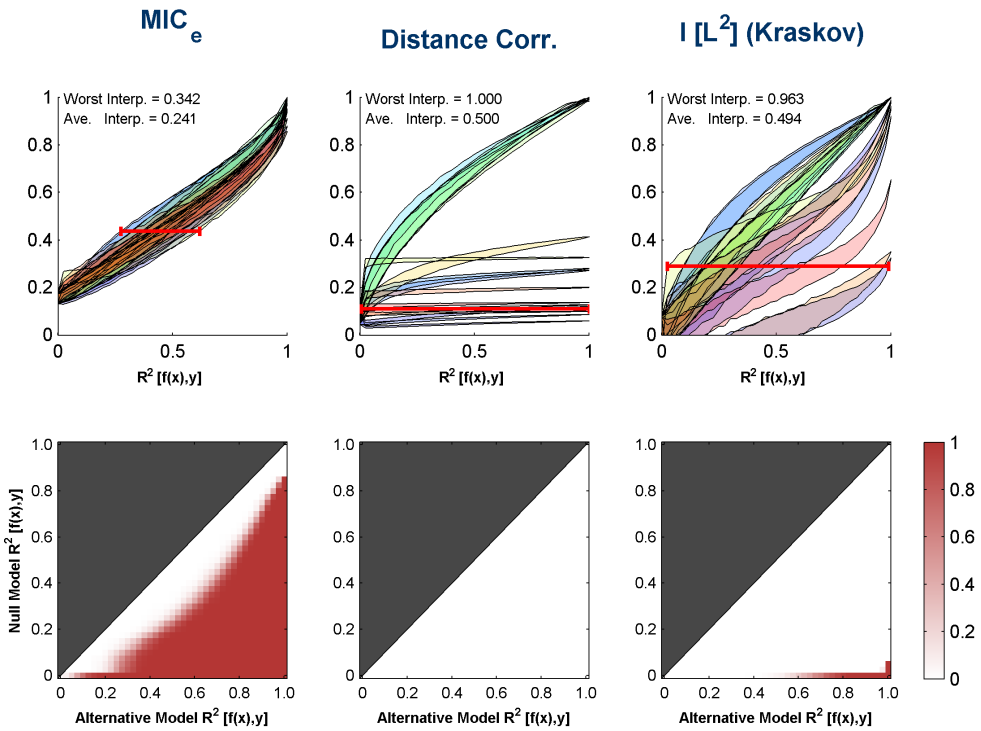


Figure 2.5: An analysis of the equitability with respect to R^2 of three measures of dependence on a set of functional relationships. The set of relationships used is described in Section 2.5. Each column contains results for the indicated measure of dependence. (Top) The analysis visualized via $\mathcal{Q}$-confidence intervals as in Figure 2.2. *[Narrower is more equitable.]* The worst-case and average-case widths of the 0.1 $\mathcal{Q}$-confidence intervals for the statistic in question are indicated. (Bottom) The same analysis visualized via statistical power as in Figure 2.4. *[Redder is more equitable.]*

2.6 Conclusion

2.6.1 Two dual motivations for equitability

There are two different ways to motivate equitability. The first is to begin with a measure of dependence $\hat{\varphi}$ and to observe that, though $\hat{\varphi}$ will asymptotically allow us to detect all deviations from independence in a data set, it need not tell us anything about the strength of those relationships. Since it often happens that we detect many more

relationships than can be realistically followed up, it would be desirable to have $\hat{\varphi}$ tell us something not just about the presence or absence of a relationship, but also about relationship strength as defined by Φ on at least a partial set of "standard relationships" $\boldsymbol{\mathcal{Q}}$.

The second way is to suppose that $\hat{\varphi}$ is a consistent estimator of Φ on $\boldsymbol{\mathcal{Q}}$ and to ask "what is the minimal requirement we can add to ensure that $\hat{\varphi}$ is robust to detecting relationships outside of $\boldsymbol{\mathcal{Q}}$?" Perhaps the weakest stipulation we can impose is that the population value φ of our statistic be non-zero in cases of non-trivial dependence of any sort. That is, we want $\hat{\varphi}$ to be a measure of dependence as well.

Both of these scenarios would be resolved by a measure of dependence that is also a consistent estimator of Φ. However, in many interesting cases it is hard to construct a statistic satisfying both properties: for instance, if $\boldsymbol{\mathcal{Q}}$ 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 non-parametric estimate of the generating function will be a consistent estimator of Φ, but will 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.

This motivates us to ask whether, despite the difficulty of simultaneously estimating Φ consistently and retaining the properties of a measure of dependence, we can at least seek an approximate version of this ideal. Doing so, however, necessitates a weaker requirement than consistent estimation. This is what leads us to equitability. Equitability allows us to seek statistics that have the robustness of measures of dependence but that

also, via their relationship to a property of interest Φ, give values that have a clear, if approximate, interpretation and can therefore be used to rank relationships.

2.6.2 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 good equitability is achieved? Are there systematic ways of obtaining equitable behavior via a learning framework as has been done, e.g., for causation in[87]? 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, each with its own way 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 an 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.

Acknowledgements

We would like to acknowledge E Airoldi, T Broderick, H Finucane, A Gelman, M Gorfine, R Heller, J Huggins, T Jaakkola, J Mueller, J Tenenbaum, and R Tibshirani for constructive conversations and useful feedback.

3

Measuring dependence powerfully and equitably

The concept of equitability is only useful to the extent that we can exhibit measures of dependence with meaningful levels of equitability that still have reasonable power

to detect non-trivial relationships to begin with. In this chapter, we define and theoretically characterize two new statistics that together yield an efficient approach for obtaining both power and equitability. To do this, we first introduce a new population measure of dependence and show three equivalent ways that it can be viewed, including as a canonical "smoothing" of mutual information. We then introduce an efficiently computable consistent estimator of our population measure of dependence, and we empirically establish its equitability on a large class of noisy functional relationships. This new statistic has better bias/variance properties and better runtime complexity than a previous heuristic approach. Next, we derive a second, related statistic whose computation is a trivial side-product of our algorithm and whose goal is powerful independence testing rather than equitability. We prove that this statistic yields a consistent independence test and show in simulations that the test has good power against independence. Taken together, our results suggest that these two statistics are a valuable pair of tools for exploratory data analysis.*

3.1 Introduction

In Chapter 2, we introduced a new way of assessing a measure of dependence, called *equitability*[121]. Informally, an equitable statistic is one that, for some measure of relationship strength, assigns similar scores to equally strong relationships regardless of

*The material in this chapter is adapted from a manuscript published in the November 2016 edition of the *Journal of Machine Learning Research* as "Measuring dependence powerfully and equitably" by Yakir Reshef*, David Reshef*, *et al.*[131] (* = co-first author)

relationship type. For instance, we may want our measure of dependence to also have the property that on noisy functional relationships it assigns similar scores to relationships with the same R^2, i.e., the squared Pearson correlation between the observed y-values and the x-values passed through the underlying function in question[121]. Or, alternatively, we may want the value of our statistic to tell us about the proportion of points coming from the deterministic component of a mixture containing part signal and part uniform noise[33]. Defining measures of dependence that achieve good equitability with respect to interesting measures of relationship strength is a new and challenging problem, with a number of different formalizations. (See, e.g., Reshef et al.[132] and Ding & Li[33], as well as Kinney & Atwal[76] along with associated technical comments Reshef et al.[123] and Murrell et al.[98].)

In this chapter, we introduce and theoretically characterize two new measures of dependence that we empirically show to have good equitability with respect to R^2 and power against independence, respectively. We begin by introducing a new population measure of dependence called MIC_*. Given a pair of jointly distributed random variables (X, Y), $\text{MIC}_*(X, Y)$ is the supremum, over all finite grids G imposed on the support of (X, Y), of the mutual information of the discrete distribution induced by (X, Y) on the cells of G, subject to a regularization based on the resolution of G. We prove three results, each of which gives a different way that this population quantity can be viewed.

1. MIC_* is the population value of the maximal information coefficient (MIC), a statistic introduced in Reshef et al.[121] that is empirically highly equitable with respect to R^2 on a large class of noisy functional relationships. Simple corollaries

of this result simplify and strengthen many of the theoretical results proven in Reshef et al.[121] about MIC.

2. MIC_* is a minimal smoothing of mutual information, in the sense that the regularization in the definition of MIC_* renders it uniformly continuous as a function of random variables with respect to statistical distance, and no "smaller" regularization achieves continuity. This result yields as a corollary that mutual information by itself is not continuous with respect to statistical distance.

3. MIC_* is the supremum of an infinite sequence defined in terms of optimal (one-dimensional) partitions of the marginal distributions of (X, Y) rather than optimal (two-dimensional) grids imposed on the joint distribution. This characterization greatly simplifies computation.

After proving these three results, we leverage them to introduce efficient algorithms both for approximating MIC_* in practice and for estimating it consistently from a finite sample. We first provide an efficient algorithm that in many cases allows for computation to arbitrary precision of the MIC_* of a pair of random variables whose joint density is known. We then introduce a statistic, called MIC_e, that we prove is a consistent estimator of MIC_*. In contrast to the MIC statistic from Reshef et al.[121], for which no efficient algorithm is known and a heuristic algorithm is used in practice, MIC_e is efficiently computable. It has a better runtime complexity than the heuristic algorithm currently in use for computing the original MIC statistic, and is orders of magnitude faster in practice.

With a consistent and fast estimator for MIC_* in hand, we turn to empirical analysis of its performance. Specifically, we show through simulation that MIC_e has better bias/variance properties than the heuristic algorithm used in Reshef et al.[121] for com-

puting MIC, which has no theoretical convergence guarantees. Our analysis also reveals that the main parameter of MIC_e can be used to tune statistical performance toward either stronger or weaker relationships in general. After studying the bias/variance properties of MIC_e, we then demonstrate via simulation that it outperforms currently available methods in terms of equitability with respect to R^2 on a broad set of noisy functional relationships. We show this performance advantage both on the set of functional relationships analyzed in Reshef et al.[121] as well as on a large set of randomly chosen noisy functional relationships.

We choose in this chapter to analyze equitability specifically with respect to R^2, rather than some other notion of relationship strength, because R^2 on noisy functional relationships is a simple measure with broad familiarity and intuitive interpretation among practitioners. Of course, it is also important to develop measures of dependence that are equitable with respect to notions of relationship strength besides R^2 or on families of relationships besides noisy functional relationships; however, our focus here remains on the "simple" case of R^2 on noisy functional relationships.

Importantly, we note that although there are methods for directly estimating the R^2 of a noisy functional relationship via nonparametric regression (see, e.g., Cleveland & Devlin[22], Stone[148]), those methods are not applicable in the context of equitability because they are not measures of dependence. That is, because non-parametric regression methods *assume* a functional form for the relationship in question, they can give trivial scores to non-functional relationships, even in the large-sample limit. (A simple exam-

ple of this is a uniform distribution over a circle, whose regression function is constant.) In contrast, a *measure of dependence* is guaranteed never to make this "mistake". A measure of dependence that is equitable with respect to R^2 can therefore be viewed either as an "upgraded" measure of dependence that also comes with some of the interpretability properties of non-parametric regression, or as an "upgraded" approximate non-parametric regression method that also has the robustness properties of a measure of dependence.

The main strength of MIC_e is equitability rather than power to reject a null hypothesis of independence. In some settings, though, it may be more important to focus on good power against independence. We therefore introduce here a statistic closely related to MIC_e called the total information coefficient and denoted TIC_e. We prove the consistency of testing for independence using TIC_e, and show via simulations that it achieves excellent power in practice, performing comparably to or better than current methods on an index suite of relationships from Simon & Tibshirani[144]. Because TIC_e arises naturally as a side-product of the computation of MIC_e, it is available "for free" once MIC_e has been computed. This leads us to propose a data analysis strategy consisting of first using TIC_e to filter out non-significant relationships, and then ranking the remaining ones using the simultaneously computed values of MIC_e.

We focus primarily on theory and illustrative empirical work in this chapter, deferring more detailed empirical studies to Chapter 4, in which we explore in detail the empirical performance of the methods introduced here. That chapter compares MIC_e and TIC_e to

several leading measures of dependence[79,153,58,59,52,14,86] on a broad range of relationship types under many different sampling and noise models, finding that the equitability with respect to R^2 of MIC_e and the power of independence testing using TIC_e are both state-of-the-art on the relationships examined. It also shows that these methods can be computed very fast in practice.

Taken together, our results shed significant light on the theory behind the maximal information coefficient, and suggest that TIC_e and MIC_e are a useful pair of methods for data exploration. Specifically, they point to joint use of these two statistics to filter and then rank relationships as a fast, practical way to explore large data sets by measuring dependence both powerfully and equitably.

3.2 Preliminaries

We work extensively in this chapter with grids and discrete distributions over their cells. Given a grid G and a point (x, y), we define the function $\text{row}_G(y)$ to be the row of G containing y and we define $\text{col}_G(x)$ analogously. For a pair (X, Y) of jointly distributed random variables, we write $(X, Y)|_G$ to denote $(\text{col}_G(X), \text{row}_G(Y))$, and we use $I((X, Y)|_G)$ to denote the discrete mutual information[23,26,25] between $\text{col}_G(X)$ and $\text{row}_G(Y)$. Given a finite sample D from the distribution of (X, Y), we sometimes use D to refer both to the set of points in the sample as well as to a point chosen uniformly at random from D. In the latter case, it will then make sense to talk about, e.g., $D|_G$

and $I(D|_G)$.

For natural numbers k and ℓ, we use $G(k,\ell)$ to denote the set of all k-by-ℓ grids (possibly with empty rows/columns). A grid G is an equipartition of (X,Y) if all the rows of $(X,Y)|_G$ have the same probability mass, and all the columns do as well. We also use the term equipartition in the analogous way for one-dimensional partitions into just rows or columns. For a one-dimensional partition P into rows and a one-dimensional partition Q into columns, we write (P,Q) to refer to the grid constructed from these two partitions. When a partition P can be obtained from a partition P' by addition of separators alone, we write $P' \subset P$.

Finally, let us establish some notation for infinite matrices. We use m^∞ to denote the space of infinite matrices equipped with the supremum norm. Given a matrix $A \in m^\infty$, we often examine only the k,ℓ-th entries of A for which $k\ell \leq i$ for some i. Thus, for $i \in \mathbb{Z}^+$, we define the projection $r_i : m^\infty \to m^\infty$ via

$$r_i(A)_{k,\ell} = \begin{cases} A_{k,\ell} & k\ell \leq i \\ 0 & k\ell > i \end{cases}.$$

Unless noted otherwise, all logarithms are to base 2.

3.3 The population maximal information coefficient MIC_*

In this section, we define and characterize the population maximal information coefficient MIC_*. We begin by defining the population quantity $\text{MIC}_*(X, Y)$ for a pair of jointly distributed random variables (X, Y). We then show three different ways to characterize this population quantity: first, as the large-sample limit of the statistic MIC from Reshef et al.[121]; second, as a minimally smoothed version of mutual information; and third, as the supremum of an infinite sequence defined in terms of optimal one-dimensional partitions of the marginal distributions of (X, Y). We conclude the section by showing how the third characterization leads to an efficient approach for approximating MIC_* in practice from the density of (X, Y).

3.3.1 Defining MIC_*

The population maximal information coefficient can be defined in several equivalent ways, as we will see later. For now, we begin with the simplest definition.

Definition 3.3.1. Let (X, Y) be jointly distributed random variables. The *population maximal information coefficient* (MIC_*) of (X, Y) is defined by

$$\text{MIC}_*(X, Y) = \sup_G \frac{I((X, Y)|_G)}{\log \|G\|}$$

where $\|G\|$ denotes the minimum of the number of rows of G and the number of columns

of G.

Given that $I(X,Y) = \sup_G I((X,Y)|_G)$ (see, e.g., Chapter 8 of Cover & Thomas[23]), this can be viewed as a regularized version of mutual information that penalizes complicated grids and ensures that the result falls between zero and one.

Before we continue, we state one simple equivalent definition of MIC_* that is useful for the results in this section. This definition views MIC_* as the supremum of a matrix called the *population characteristic matrix*, defined below.

Definition 3.3.2. Let (X,Y) be jointly distributed random variables. Let

$$I^*((X,Y),k,\ell) = \max_{G \in G(k,\ell)} I((X,Y)|_G).$$

The *population characteristic matrix* of (X,Y), denoted by $M(X,Y)$, is defined by

$$M(X,Y)_{k,\ell} = \frac{I^*((X,Y),k,\ell)}{\log \min\{k,\ell\}}$$

for $k,\ell > 1$.

It is easy to see the following:

Proposition 3.3.3. *Let (X,Y) be jointly distributed random variables. We have*

$$MIC_*(X,Y) = \sup M(X,Y)$$

where $M(X,Y)$ is the population characteristic matrix of (X,Y).

The population characteristic matrix is so named because just as MIC_*, the supremum of this matrix, captures a sense of relationship strength, other properties of this matrix correspond to different properties of relationships. For instance, later in this chapter we introduce an additional property of the characteristic matrix, the total information coefficient, that is useful for testing for the presence or absence of a relationship rather than quantifying relationship strength.

3.3.2 First alternate characterization: MIC_* is the population value of MIC

With MIC_* defined, we now state our first alternate characterization of it, as the large-sample limit of the statistic MIC introduced in Reshef et al.[121]. We begin by first reproducing a description of MIC from Reshef et al.[121], via the two definitions below.

Definition 3.3.4 (Reshef et al.[121])**.** Let $D \subset \mathbb{R}^2$ be a set of ordered pairs. The *sample characteristic matrix* $\widehat{M}(D)$ of D is defined by

$$\widehat{M}(D)_{k,\ell} = \frac{I^*(D,k,\ell)}{\log \min\{k,\ell\}}.$$

Definition 3.3.5 (Reshef et al.[121])**.** Let $D \subset \mathbb{R}^2$ be a set of n ordered pairs, and let

$B : \mathbb{Z}^+ \to \mathbb{Z}^+$. We define

$$\mathrm{MIC}_B(D) = \max_{k\ell \le B(n)} \widehat{M}(D)_{k,\ell}.$$

where the function $B(n)$ is specified by the user. In Reshef et al.[121], it was suggested that $B(n)$ be chosen to be n^α for some constant α in the range of 0.5 to 0.8. (The statistics we introduce later will have an analogous parameter; see Section 3.4.4.)

We show the following result about convergence of functions of the sample characteristic matrix to their population counterparts, a consequence of which is the convergence of MIC to MIC_*. (In the theorem statement below, recall that m^∞ is the space of infinite matrices equipped with the supremum norm, and given a matrix A the projection r_i zeros out all the entries $A_{k,\ell}$ for which $k\ell > i$.)

Theorem 3.3.1. *Let $f : m^\infty \to \mathbb{R}$ be uniformly continuous, and assume that $f \circ r_i \to f$ pointwise. Then for every random variable (X, Y), we have*

$$\left(f \circ r_{B(n)}\right)\left(\widehat{M}(D_n)\right) \to f(M(X, Y))$$

in probability where D_n is a sample of size n from the distribution of (X, Y), provided $\omega(1) < B(n) \le O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.

Proof. See Section B.1. □

Since the supremum of a matrix is uniformly continuous as a function on m^∞ and

can be realized as the limit of maxima of larger and larger segments of the matrix, this theorem yields our claim about MIC_* as a corollary.

Corollary 3.3.6. *MIC_B is a consistent estimator of MIC_* provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.*

Though Theorem 3.3.1 is proven in Section B.1, we provide here some intuition for why it should hold as well as a description of the obstacles that must be overcome in the proof.

For concreteness, suppose f is the supremum function. To see why the theorem should hold, fix a random variable (X, Y) and let D be a sample of size n from its distribution. It is known that for a fixed grid G $I(D|_G)$ is a consistent estimator of $I((X,Y)|_G)$[136,108]. We might therefore expect $I^*(D, k, \ell)$ to be a consistent estimator of $I^*((X,Y), k, \ell)$ as well. And if $I^*(D, k, \ell)$ is a consistent estimator of $I^*((X,Y), k, \ell)$, then we might expect the maximum of the sample characteristic matrix (which just consists of normalized I^* terms) to be a consistent estimator of the supremum of the true characteristic matrix.

These intuitions turn out to be true, but there are two reasons they are non-trivial to prove. First, consistency for I^* does not follow from abstract considerations since the supremum of an infinite set of estimators is not necessarily a consistent estimator of the supremum of the estimands.[†] Second, consistency of I^* alone does not suffice to show

[†] If $\hat{\theta}_1, \ldots, \hat{\theta}_k$ is a finite set of estimators, then a union bound shows that the random variable $(\hat{\theta}_1(D), \ldots, \hat{\theta}_k(D))$ converges in probability to $(\theta_1, \ldots, \theta_k)$ with respect to the supremum metric.

that the maximum of the sample characteristic matrix converges to MIC_*. In particular, if $B(n)$ grows too quickly, and the convergence of $I^*(D, k, \ell)$ to $I^*((X, Y), k, \ell)$ is slow, inflated values of MIC can result. To see this, notice that if $B(n) = \infty$ then MIC = 1 for uniformly generated noise at any finite sample size, even though each individual entry of the sample characteristic matrix converges to its true value eventually.

The technical heart of the proof is overcoming these obstacles by using the dependencies between the quantities $I(D|_G)$ for different grids G to not only show the consistency of $I^*(D, k, \ell)$ but then to quantify how quickly $I^*(D, k, \ell)$ converges to $I^*((X, Y), k, \ell)$.

3.3.3 Second alternate characterization: MIC_* is a minimally smoothed mutual information

We now describe a second equivalent view of MIC_*. Recall that for a pair of jointly distributed random variables (X, Y), we defined $\text{MIC}_*(X, Y)$ as

$$\text{MIC}_*(X, Y) = \sup_G \frac{I((X, Y)|_G)}{\log \|G\|}$$

where $\|G\|$ denotes the minimum of the number of rows of G and the number of columns of G. As we discussed in Section 3.3.1, the mutual information $I(X, Y)$ is also a supre-

The continuous mapping theorem then gives the desired result. However, if the set of estimators is infinite, the union bound cannot be employed. And indeed, if we let $\theta_1 = \cdots = \theta_k = 0$, and let $\hat{\theta}_i(D_n) = i/n$ deterministically, then each $\hat{\theta}_i$ is a consistent estimator of θ_i, but since the set $\{\hat{\theta}_1(D_n), \hat{\theta}_2(D_n), \ldots\} = \{1/n, 2/n, \ldots\}$ is unbounded, $\sup_i \hat{\theta}_i(D_n) = \infty$ for every n.

mum, namely

$$I(X,Y) = \sup_G I((X,Y)|_G).$$

and so MIC_* can be viewed as a regularized version of I. It is natural to ask whether the regularization in the definition of MIC_* has any smoothing effect on I. In this sub-section we show first that it does, in the sense that MIC_* is uniformly continuous as a function of random variables with respect to the metric of statistical distance,[‡] and second that the regularization by $\log \|G\|$ is in some sense the minimal one necessary for achieving any sort of continuity. As a corollary, we obtain that I by itself is not continuous as a function of random variables with respect to the metric of statistical distance. This provides a view of MIC_* as a canonical smoothing of I that yields continuity.

Formally, let $\boldsymbol{\mathcal{P}}(\mathbb{R}^2)$ denote the space of random variables supported on $\mathbb{R}^2$ equipped with the metric of statistical distance. Our first claim is that as a function defined on $\boldsymbol{\mathcal{P}}(\mathbb{R}^2)$, MIC_* is uniformly continuous. We prove this claim by establishing a stronger result: the uniform continuity of the characteristic matrix $M(X,Y)$. Specifically, by showing that the family of maps corresponding to each individual entry of the characteristic matrix is uniformly equicontinuous, we obtain the following result.

Theorem 3.3.2. *The map from $\boldsymbol{\mathcal{P}}(\mathbb{R}^2)$ to m^∞ defined by $(X,Y) \mapsto M(X,Y)$ is uni-*

[‡] Recall that the statistical distance between random variables A and B is defined as $\sup_T |\mathbf{P}(A \in T) - \mathbf{P}(B \in T)|$. When A and B have probability density functions or probability mass functions, this equals one-half of the L^1 distance between those functions.

formly continuous.

Proof. See Appendix B.2. □

Since the supremum is a uniformly continuous function on m^∞, Theorem 3.3.2 yields the following corollary.

Corollary 3.3.7. *The map $(X, Y) \mapsto MIC_*(X, Y)$ is uniformly continuous.*

Similar corollaries exist for any uniformly continuous function of the characteristic matrix.

Interestingly, Theorem 3.3.2 relies crucially on the normalization in the definition of the characteristic matrix. This is not a coincidence: as the following proposition shows, any normalization that is meaningfully smaller than the one in the definition of the characteristic matrix will cause the matrix to contain a discontinuity as a function on $\boldsymbol{\mathcal{P}}(\mathbb{R}^2)$.

Proposition 3.3.8. *For some function $N(k, \ell)$, let M^N be the characteristic matrix with normalization N, i.e.,*

$$M^N(X, Y)_{k,\ell} = \frac{I^*((X, Y), k, \ell)}{N(k, \ell)}.$$

If $N(k, \ell) = o(\log \min\{k, \ell\})$ along some infinite path in $\mathbb{N} \times \mathbb{N}$, then M^N and $\sup M^N$ are not continuous as functions of $\mathcal{P}([0, 1] \times [0, 1]) \subset \boldsymbol{\mathcal{P}}(\mathbb{R}^2)$.

Proof. See Appendix B.3. □

The above proposition implies that the "smoothing" that MIC_* applies to mutual information is necessary in some sense. In particular, one corollary of the proposition is that mutual information with no smoothing will contain a disconuity.

Corollary 3.3.9. *Mutual information is not continuous on* $\mathcal{P}([0,1] \times [0,1]) \subset \boldsymbol{\mathcal{P}}(\mathbb{R}^2)$.

Proof. Mutual information is the supremum of M^N with $N \equiv 1$. □

The same result can also be shown for the squared Linfoot correlation[146,84], which equals $1 - 2^{-2I}$ where I represents mutual information. Thus, though the Linfoot correlation smoothes the mutual information enough to cause it to lie in the unit interval, it does not smooth the mutual information sufficiently to cause it to be continuous.

As we remarked previously, these results, when contrasted with the uniform continuity of MIC_*, allow us to view the latter as a canonical "minimally smoothed" version of mutual information that is uniformly continuous. This view gives a meaningful interpretation to the normalization used in MIC_*. Understanding MIC_* as having smoothness properties not shared by mutual information also suggests that estimators of MIC_* may have better statistical properties than estimators of ordinary mutual information. This is consistent with a recent hardness-of-estimation result for mutual information in Ding & Li[33] and we show in Chapter 4 that it also borne out empirically.

3.3.4 Third alternate characterization: MIC_* is the supremum of the boundary of the characteristic matrix

We now show the third alternate view of MIC_*: that it can be equivalently defined as the supremum over a *boundary* of the characteristic matrix rather than as a supremum over all of the entries of the matrix. This characterization of MIC_* will serve as the foundation both for our approach to approximating $\text{MIC}_*(X, Y)$ as well as the new estimator of MIC_* that we introduce later in this chapter.

We begin by defining what we mean by the boundary of the characteristic matrix. Our definition rests on the following observation.

Proposition 3.3.10. *Let M be a population characteristic matrix. Then for $\ell \geq k$, $M_{k,\ell} \leq M_{k,\ell+1}$.*

Proof. Let (X, Y) be the random variable in question. Since we can always let a row/column be empty, we know that $I^*((X, Y), k, \ell) \leq I^*((X, Y), k, \ell + 1)$. And since $\ell, \ell + 1 \geq k$, we know that $M_{k,\ell} = I^*((X, Y), k, \ell)/\log k \leq I^*((X, Y), k, \ell + 1)/\log k = M_{k,\ell+1}$. □

Since the entries of the characteristic matrix are bounded, the monotone convergence theorem then gives the following corollary. In the corollary and henceforth, we let $M_{k,\uparrow} = \lim_{\ell\to\infty} M_{k,\ell}$ and define $M_{\uparrow,\ell}$ similarly.

Corollary 3.3.11. *Let M be a population characteristic matrix. Then $M_{k,\uparrow}$ exists, is finite, and equals* $\sup_{\ell \geq k} M_{k,\ell}$. *The same is true for $M_{\uparrow,\ell}$.*

The above corollary allows us to define the *boundary* of the characteristic matrix.

Definition 3.3.12. Let M be a population characteristic matrix. The *boundary* of M is the set

$$\partial M = \{M_{k,\uparrow} : 1 < k < \infty\} \bigcup \{M_{\uparrow,\ell} : 1 < \ell < \infty\}.$$

The theorem below then gives a relationship between the boundary of the characteristic matrix and MIC_*.

Theorem 3.3.3. *Let (X, Y) be a random variable. We have*

$$MIC_*(X, Y) = \sup \partial M(X, Y)$$

where $M(X, Y)$ is the population characteristic matrix of (X, Y).

Proof. The following argument shows that every entry of M is at most $\sup \partial M$: fix a pair (k, ℓ) and notice that either $k \leq \ell$, in which case $M_{k,\ell} \leq M_{k,\uparrow}$, or $\ell \leq k$, in which case $M_{k,\ell} \leq M_{\uparrow,\ell}$. Thus, $\mathrm{MIC}_* \leq \sup\{M_{\uparrow,\ell}\} \cup \{M_{k,\uparrow}\} = \sup \partial M$.

On the other hand, Corollary 3.3.11 shows that each element of ∂M is a supremum over some elements of M. Therefore, $\sup \partial M$, being a supremum over suprema of elements of M, cannot exceed $\sup M = \mathrm{MIC}_*$. □

3.3.5 Approximating MIC_* in practice

The importance of the characterization in Theorem 3.3.3 from the previous sub-section is computational. Specifically, elements of the boundary of the characteristic matrix can be expressed in terms of a maximization over (one-dimensional) partitions rather than (two-dimensional) grids, the former being much quicker to compute exactly. This is stated in the theorem below.

Theorem 3.3.4. *Let M be a population characteristic matrix. Then $M_{k,\uparrow}$ equals*

$$\max_{P \in P(k)} \frac{I(X, Y|_P)}{\log k}$$

where $P(k)$ denotes the set of all partitions of size at most k.

Proof. See Appendix B.4. □

To formally state how this will help us from an algorithmic standpoint, we note that Theorems 3.3.3 and 3.3.4 above together give the following corollary.

Corollary 3.3.13. *Let (X, Y) be a random variable, and let $\mathbb{P}$ be the set of finite-size partitions. Then*

$$MIC_*(X, Y) = \sup \left\{ \frac{I(X, Y|_P)}{\log |P|} : P \in \mathbb{P} \right\} \bigcup \left\{ \frac{I(X|_P, Y)}{\log |P|} : P \in \mathbb{P} \right\}$$

where $|P|$ is the number of bins in the partition P.

We can exploit the fact that the expressions in the above corollary involve maximization only over one-dimensional partitions rather than two-dimensional grids to give an algorithm for computing elements of the boundary of the characteristic matrix to arbitrary precision, and by extension an approach to approximating MIC_* in practice. To do so, we utilize as a subroutine a dynamic programming algorithm from Reshef et al.[121] called OptimizeXAxis. Before continuing, we therefore give a brief overview of that algorithm.

Overview of OptimizeXAxis *algorithm from Reshef et al.*[121]. The OptimizeXAxis algorithm takes as input a set D of n data points, a fixed partition into columns[§] Q of size ℓ, a "master" partition into rows Π, and a number k. The algorithm returns, for $2 \leq i \leq k$, the partition into rows $P_i \subset \Pi$ that maximizes the mutual information of $D|_{(P_i,Q)}$ among all sub-partitions of Π of size at most i. The algorithm works by exploiting the fact that, conditioned on the location y of the top-most line of P_i, the optimization of the rest of P_i can be formulated as a sub-problem that depends only on the data points below y. The algorithm uses dynamic programming to store and reuse solutions to these subproblems, resulting in a runtime of $O(|\Pi|^2 k\ell)$. If a black-box algorithm is used to compute each required mutual information in time at most T, then the runtime of the algorithm can be shown to be $O(Tk|\Pi|)$.

[§] Despite its name, the OptimizeXAxis algorithm can be used to optimize a partition of either axis. In our description of the algorithm here, we choose to describe the algorithm as it would work for optimizing a partition of the *y-axis* rather than the x-axis. This is for notational coherence of this chapter only.

The following theorem shows that the theory developed about the boundary of the characteristic matrix, together with OptimizeXAxis, yields an efficient algorithm for computing entries of the boundary to arbitrary precision.

Theorem 3.3.5. *Given a random variable (X, Y), $M_{k,\uparrow}$ (resp. $M_{\uparrow,\ell}$) is computable to within an additive error of $O(k\varepsilon \log(1/(k\varepsilon))) + E$ (resp. $O(\ell\varepsilon \log(1/(\ell\varepsilon))) + E$) in time $O(kT(E)/\varepsilon)$ (resp. $O(\ell T(E)/\varepsilon)$), where $T(E)$ is the time required to numerically compute the mutual information of a continuous distribution to within an additive error of E.*

Proof. See Appendix B.5. □

The algorithm proposed in Theorem 3.3.5 gives us a polynomial-time method for computing any finite subset of the boundary ∂M of the population characteristic matrix $M(X, Y)$ of a random variable (X, Y). Thus, if we have some k_0, ℓ_0 such that the maximum of the finite subset $\{M_{k,\uparrow}, M_{\uparrow,\ell} : k \leq k_0, \ell \leq \ell_0\}$ of ∂M will be ε-close to the supremum of the entire set ∂M, we can compute $\text{MIC}_*(X, Y)$ to within an error of ε. Though we usually do not have precise knowledge of k_0 and ℓ_0, for many distributions it is often easy to make very conservative educated guesses for them, in which case this algorithm allows us to approximate $\text{MIC}_*(X, Y)$ very well in practice.

Being able to compute $\text{MIC}_*(X, Y)$ to arbitrary precision in some cases has two main advantages. The first advantage is that it allows us to assess in simulations the large-sample properties of MIC_* independent of any estimator. In Chapter 4, we use

this to show that MIC_* achieves high equitability with respect to R^2 on a set of noisy functional relationshipsm thereby confirming that statistically efficient estimation of MIC_* is a worthwhile goal.

The second advantage is that we can empirically assess the bias, variance, and expected squared error of estimators of MIC_* by taking a distribution, computing MIC_*, and then comparing the result to estimates of it based on finite samples. In the next section, we introduce a new estimator MIC_e of MIC_* and carry out such an analysis to compare its statistical properties to those of the statistic MIC from Reshef et al.[121].

3.4 Estimating MIC_* with MIC_e

As we have shown, MIC_* is the population value of the statistic MIC introduced in Reshef et al.[121]. However, though consistent, the statistic MIC is not known to be efficiently computable and in Reshef et al.[121] a heuristic approximation algorithm called Approx-MIC was computed instead. In this section, we leverage the theory we have developed here to introduce a new estimator of MIC_* that is both consistent and efficiently computable. The new estimator, called MIC_e, has better runtime complexity even than the heuristic Approx-MIC algorithm, and runs orders of magnitude faster in practice.

The estimator MIC_e is based on one of the alternate characterizations of MIC_* proven in the previous section. Namely, if MIC_* can be viewed as the supremum of the *bound-*

ary of the characteristic matrix rather than of the entire matrix, then only the boundary of the matrix must be accurately estimated in order to estimate MIC_*. This has the advantage that, whereas computing individual entries of the sample characteristic matrix involves finding optimal (two-dimensional) grids, estimating entries of the boundary requires us only to find optimal (one-dimensional) partitions. While the former problem is computationally difficult, the latter can be solved using the dynamic programming algorithm from Reshef et al.[121] that we also employed in Section 3.3.5 to compute MIC_* to arbitrary precision in the large-sample limit.

We formalize this idea via a new object called the *equicharacteristic matrix*, which we denote by $[M]$. The difference between $[M]$ and the characteristic matrix M is as follows: while the k, ℓ-th entry of M is computed from the maximal achievable mutual information using any k-by-ℓ grid, the k, ℓ-th entry of $[M]$ is computed from the maximal achievable mutual information using any k-by-ℓ grid that equipartitions the dimension with more rows/columns. (See Figure 3.1.) Despite this difference, as the equipartition in question gets finer and finer it becomes indistinguishable from an optimal partition of the same size. This intuition can be formalized to show that the boundary of $[M]$ equals the boundary of M, and therefore that $\sup[M] = \sup M = \text{MIC}_*$. It will then follow that estimating $[M]$ and taking the supremum—as we did with M in the case of MIC—yields a consistent estimate of MIC_*.

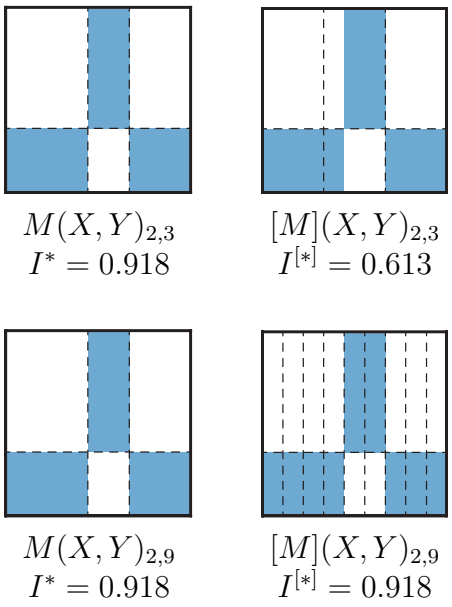


Figure 3.1: A schematic illustrating the difference between the characteristic matrix M and the equicharacteristic matrix $[M]$. (Top) When restricted to 2 rows and 3 columns, the characteristic matrix M is computed from the optimal 2-by-3 grid. In contrast, the equicharacteristic matrix $[M]$ still optimizes the smaller partition of size 2 but is restricted to have the larger partition be an equipartition of size 3. This results in a lower mutual information of 0.613. (Bottom) When 9 columns are allowed instead of 3, the grid found by the characteristic matrix does not change, since the grid with 3 columns was already optimal. However, now the equicharacteristic matrix uses an equipartition into columns of size 9, whose resolution is able to fully capture the dependence between X and Y.

3.4.1 The equicharacteristic matrix

We now define the equicharacteristic matrix and show that its supremum is indeed MIC_*. To do so, we first define a version of I^* that equipartitions the dimension with more rows/columns. Note that in the definition, brackets are used to indicate the presence of an equipartition.

Definition 3.4.1. Let (X, Y) be jointly distributed random variables. Define

$$I^*\left((X,Y),k,[\ell]\right) = \max_{G \in G(k,[\ell])} I\left((X,Y)|_G\right)$$

where $G(k,[\ell])$ is the set of k-by-ℓ grids whose y-axis partition is an equipartition of size ℓ. Define $I^*((X,Y),[k],\ell)$ analogously.

Define $I^{[*]}((X,Y),k,\ell)$ to equal $I^*((X,Y),k,[\ell])$ if $k < \ell$ and $I^*((X,Y),[k],\ell)$ otherwise.

We now define the equicharacteristic matrix in terms of $I^{[*]}$. In the definition below, we continue our convention of using brackets to denote the presence of equipartitions.

Definition 3.4.2. Let (X, Y) be jointly distributed random variables. The *population equicharacteristic matrix* of (X, Y), denoted by $[M](X, Y)$, is defined by

$$[M](X,Y)_{k,\ell} = \frac{I^{[*]}((X,Y),k,\ell)}{\log \min\{k,\ell\}}$$

for $k, \ell > 1$.

The boundary of the equicharacteristic matrix can be defined via a limit in the same way as the characteristic matrix. We then have the following theorem.

Theorem 3.4.1. *Let (X, Y) be jointly distributed random variables. Then $\partial[M] = \partial M$.*

Proof. See Appendix B.6. □

Since every entry of the equicharacteristic matrix is dominated by some entry on its boundary, the equivalence of $\partial[M]$ and ∂M yields the following corollary as a simple consequence.

Corollary 3.4.3. *Let (X, Y) be jointly distributed random variables. Then* $\sup[M](X, Y) = MIC_*(X, Y)$.

3.4.2 The estimator MIC_e

With the equicharacteristic matrix defined, we can now define our new estimator MIC_e in terms of the sample equicharacteristic matrix, analogously to the way we defined MIC in terms of the sample characteristic matrix.

Definition 3.4.4. Let $D \subset \mathbb{R}^2$ be a set of ordered pairs. The *sample equicharacteristic matrix* $\widehat{[M]}(D)$ of D is defined by

$$\widehat{[M]}(D)_{k,\ell} = \frac{I^{[*]}(D, k, \ell)}{\log \min\{k, \ell\}}.$$

Definition 3.4.5. Let $D \subset \mathbb{R}^2$ be a set of n ordered pairs, and let $B : \mathbb{Z}^+ \to \mathbb{Z}^+$. We define

$$\mathrm{MIC}_{e,B}(D) = \max_{k\ell \leq B(n)} \widehat{[M]}(D)_{k,\ell}.$$

With the equivalence between the boundary of the characteristic matrix and that of the equicharacteristic matrix established, it is straightforward to show that MIC_e is a

consistent estimator of MIC_* via arguments similar to those we applied in the case of MIC. (See Appendix B.7.) Specifically, we show the following theorem, an analogue of Theorem 3.3.1.

Theorem 3.4.2. *Let $f : m^\infty \to \mathbb{R}$ be uniformly continuous, and assume that $f \circ r_i \to f$ pointwise. Then for every random variable (X, Y), we have*

$$\left(f \circ r_{B(n)}\right)\left(\widehat{[M]}(D_n)\right) \to f([M](X, Y))$$

in probability where D_n is a sample of size n from the distribution of (X, Y), provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.

By setting $f([M]) = \sup[M]$, we then obtain as a corollary the consistency of MIC_e.

Corollary 3.4.6. *$\mathrm{MIC}_{e,B}$ is a consistent estimator of MIC_* provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.*

As with the statistic MIC, the statistic MIC_e requires the user to specify a function $B(n)$ to use. While the theory suggests that any function of the form $B(n) = n^\alpha$ suffices provided $0 < \alpha < 1$, different values of α may yield different finite-sample properties. We study the empirical performance of MIC_e for different choices of $B(n)$ in Section 3.4.4 and point the reader to specific recommendations for practical use in Section 3.4.4.

3.4.3 Computing MIC_e

Both MIC and MIC_e are consistent estimators of MIC_*. The difference between them is that while MIC can currently be computed efficiently only via a heuristic approximation, MIC_e can be computed exactly, very efficiently, via an approach similar to the one used for approximating MIC_* involving the OptimizeXAxis subroutine. We now describe the details of this approach.

Recall that, given a fixed x-axis partition Q into ℓ columns, a set of n data points, a "master" y-axis partition Π, and a number k, the OptimizeXAxis subroutine finds, for every $2 \leq i \leq k$, a y-axis partition $P_i \subset \Pi$ of size at most i that maximizes the mutual information induced by the grid (P_i, Q). The algorithm does this in time $O(|\Pi|^2 k\ell)$. (For more discussion of OptimizeXAxis, see Section 3.3.5)

In the pair of theorems below, we show two ways that OptimizeXAxis can be used to compute MIC_e efficiently. In the proofs of both theorems, we neglect issues of divisibility, e.g., we often write $B/2$ rather than $\lfloor B/2 \rfloor$. This does not affect the results.

Theorem 3.4.3. *There exists an algorithm* Equichar *that, given a sample D of size n and some $B \in \mathbb{Z}^+$, computes the portion $r_{B(n)}(\widehat{[M]}(D))$ of the sample equicharacteristic matrix in time $O(n^2 B^2)$, which equals $O(n^{4-2\varepsilon})$ for $B(n) = O(n^{1-\varepsilon})$ with $\varepsilon > 0$.*

Proof. We describe the algorithm and simultaneously bound its runtime. We do so only for the k, ℓ-th entries of $\widehat{[M]}(D)$ satisfying $k \leq \ell, k\ell \leq B$. This suffices, since by symmetry computing the rest of the required entries at most doubles the runtime.

To compute $\widehat{[M]}(D)_{k,\ell}$ with $k \leq \ell$, we must fix an equipartition into ℓ columns on the x-axis and then find the optimal partition of the y-axis of size at most k. If we set the master partition Π of the OptimizeXAxis algorithm to be an equipartition into rows of size n, then it performs precisely the required optimization. Moreover, for fixed ℓ it can carry out the optimization simultaneously for all of the pairs $\{(2,\ell),\ldots,(B/\ell,\ell)\}$ in time $O(|\Pi|^2(B/\ell)\ell) = O(n^2B)$. For fixed ℓ, this set contains all the pairs (k,ℓ) satisfying $k \leq \ell, kl \leq B$. Therefore, to compute all the required entries of $\widehat{[M]}(D)$ we need only apply this algorithm for each $\ell = 2,\ldots,B/2$. Doing so gives a runtime of $O(n^2B^2)$. $\square$

The algorithm above, while polynomial-time, is nonetheless not efficient enough for use in practice. However, a simple modification solves this problem without affecting the consistency of the resulting estimates. The modification hinges on the fact that OptimizeXAxis can use master partitions Π besides the equipartition of size n that we used above. Spefically, setting Π in the above algorithm to be an equipartition into ck "clumps", where k is the size of the largest optimal partition being sought, speeds up the computation significantly. This modification gives a slightly different statistic, but one that has all of the theoretical properties of MIC_e—namely, consistent estimation of MIC_* and efficient exact computation. These properties are formalized in the following theorem.

Theorem 3.4.4. *Let (X,Y) be a pair of jointly distributed random variables, and let D_n be a sample of size n from the distribution of (X,Y). For every $c \geq 1$, there exists*

a matrix $\{\widehat{M}\}^c(D_n)$ *such that*

1. *The function*
$$\widetilde{MIC}_{e,B}(\cdot) = \max_{k\ell \leq B(n)} \{\widehat{M}\}^c(\cdot)_{k,\ell}$$
is a consistent estimator of MIC_* *provided* $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ *for some* $\varepsilon > 0$.

2. *There exists an algorithm* EquicharClump *for computing* $r_B(\{\widehat{M}\}^c(D_n))$ *in time* $O(n + B^{5/2})$, *which equals* $O(n + n^{5(1-\varepsilon)/2})$ *when* $B(n) = O(n^{1-\varepsilon})$.

Proof. See Appendix B.8. □

For an analysis of the effect of the parameter c in the above theorem on the results of the EquicharClump algorithm, see Appendix B.8.3.

Setting $\varepsilon = 0.6$ in the above theorem yields the following corollary.

Corollary 3.4.7. *MIC_* can be estimated consistently in linear time.*

Of course, at low sample sizes, setting $\varepsilon = 0.6$ would be undesirable. However, we show empirically in Chapter 4 that at large sample sizes this strategy works very well on typical relationships.

We remark that the EquicharClump algorithm given above is asymptotically faster even than the heuristic Approx-MIC algorithm used to calculate MIC in practice, which runs in time $O(B(n)^4)$. As we demonstrate in Chapter 4, this difference translates into a substantial difference in runtimes for similar performance at a range of realistic

sample sizes, ranging from a 30-fold speedup at $n = 500$ to over a 350-fold speedup at $n = 10,000$.

For readability, in the rest of this chapter we do not distinguish between the two versions of MIC_e computed by the EQUICHAR and EQUICHARCLUMP algorithms described above. Wherever we present simulation data about MIC_e in simulations though, we use the version of the statistic computed by EQUICHARCLUMP.

3.4.4 BIAS/VARIANCE CHARACTERIZATION OF MIC_e

The algorithm we presented in Section 3.3.5 for computing MIC_* to arbitrary precision in some cases allows us to examine the bias/variance properties of estimators of MIC_*. Here, we use it to examine the bias and variance of both MIC as computed by the heuristic APPROX-MIC algorithm from Reshef et al.[121], and MIC_e as computed by the EQUICHARCLUMP algorithm given above. To do this, we performed a simulation analysis on the following set of relationships

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

where ε_σ and ε'_σ are i.i.d., F is the set of 16 functions analyzed in Reshef et al.[121], and X_f is the set of n x-values that result in the points $(x_i, f(x_i))$ being equally spaced along the graph of f.

For each relationship $\mathcal{Z} \in \mathcal{Q}$ that we examined, we used the algorithm from The-

orem 3.3.5 with very conservative values of k_0 and ℓ_0 to compute MIC_*. We then simulated 500 independent samples from $\mathcal{Z}$, each of size $n = 500$, and computed both Approx-MIC and MIC_e on each one to obtain estimates of the sampling distributions of the two statistics. From each of the two sampling distributions, we estimated the bias and variance of either statistic on $\mathcal{Z}$. We then analyzed the bias, variance, and expected squared error of the two statistics as a function of relationship strength, which we quantified using the coefficient of determination (R^2) with respect to the generating function.

The results, presented in Figure 3.2, are interesting for two reasons. First, they demonstrate that for a typical usage parameter of $B(n) = n^{0.6}$, MIC_e performs substantially better than Approx-MIC overall. Specifically, the median of the expected squared error of MIC_e across the set F of functions is uniformly lower across R^2 values than that of Approx-MIC. When average expected squared error is used instead of median, MIC_e still performs better on all but the strongest of relationships (R^2 above $\sim$0.9). The superior performance of MIC_e is consistent with the fact that we have theoretical guarantees about its statistical properties whereas Approx-MIC is a heuristic.

Second, the results show that different values of the exponent in $B(n) = n^\alpha$ give good performance in different signal-to-noise regimes due to a bias-variance trade-off represented by this parameter. We expand on this phenomenon and discuss its implications for choosing α in practice below.

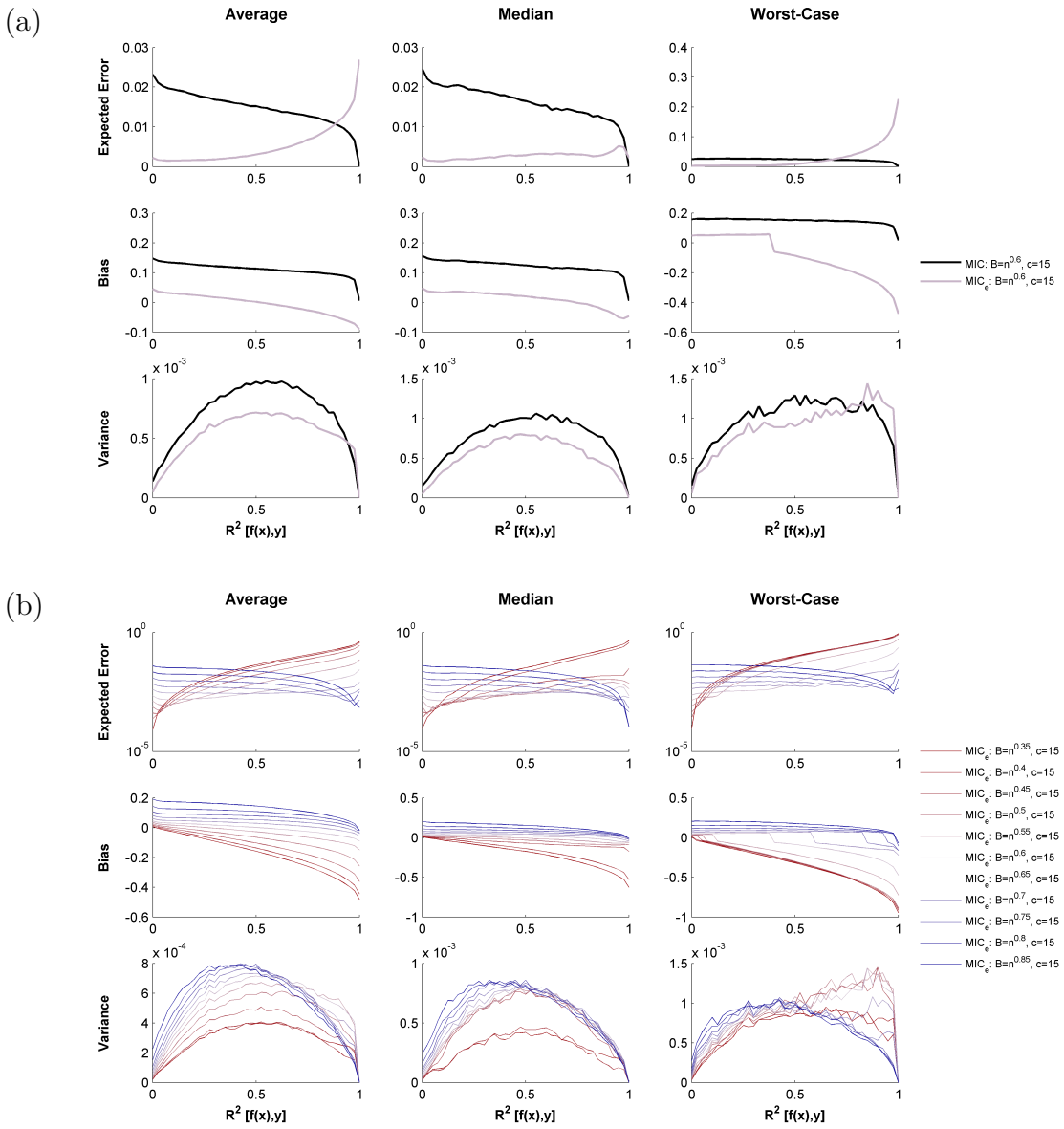


Figure 3.2: Bias/variance characterization of ApproxMIC and MIC$_e$. Each plot shows expected squared error, bias, or variance across the set of noisy functional relationships described in Section 3.4.4 as a function of the R^2 of the relationships. The results are aggregated across the 16 function types analyzed by either the average, median, or worst result at every value of R^2. (a) A comparison between MIC$_e$ (light purple) and MIC as computed via the heuristic ApproxMIC algorithm (black), at a typical usage parameter. (b) Performance of MIC$_e$ with $B(n) = n^{\alpha}$ for various values of α.

Choosing $B(n)$

Large values of α lead to increased expected error in lower-signal regimes (low R^2) through both a positive bias in those regimes and a general increase in variance that predominantly affects those regimes. On the other hand, small values of α lead to an increased expected error in higher-signal regimes (high R^2) by leading to a negative bias in those regimes and by shifting the variance of the estimator toward those regimes. In other words, lower values of α are better suited for detecting weaker signals, while higher values of α are better suited for distinguishing among stronger signals. This is consistent with the results seen in Chapter 4, which show that low values of α cause MIC_e to yield better powered independence tests while high values of α cause MIC_e to have better equitability.

Chapter 4 provides simple, empirical recommendations about appropriate values of α for different settings. Those recommendations are formulated by choosing a set of representative relationships (e.g., a set of noisy functional relationships), as well as a "ground truth" population quantity Φ (e.g., R^2) that can be used to quantify the strength of each of those relationships, and then assessing which values of α maximize the equitability of MIC_e with respect to Φ at a given sample size. This approach is applied to an analysis of real data from the World Health Organization in Chapter 4, and the parameters chosen for that analysis are the ones used for all subsequent analyses in this chapter.

We remark that if the goal of the user is only detection of non-trivial relationships rather discovery of the strongest such relationships, α can also be chosen in a more straightforward manner: the user can subsample a small random set of relationships on which to compare the power of MIC_e for different values of α. Those relationships can then be discarded and the rest of the relationships analyzed with the optimal value of α. However, if the user's primary goal is power against independence, the statistic TIC_e introduced in Section 3.5 should be used with this strategy rather than MIC_e.

3.4.5 Equitability of MIC_e

As mentioned previously, one of the main motivations for the introduction of MIC was equitability, the extent to which a measure of dependence usefully captures some notion of relationship strength on some set of standard relationships. We therefore carried out an empirical analysis of the equitability of MIC_e with respect to R^2 and compared its performance to distance correlation[154,153], mutual information estimation[79], and maximal correlation estimation[14].

We began by assessing equitability on the set of relationships $\mathcal{Q}$ defined above, a set that has been analyzed in previous work[121,127,76]. The results, shown in Figure 3.3, confirm the superior equitability of the new estimator MIC_e on this set of relationships.

To assess equitability more objectively without relying on a manually curated set of functions, we then analyzed 160 random functions drawn from a Gaussian process distribution with a radial basis function kernel with one of eight possible bandwidths

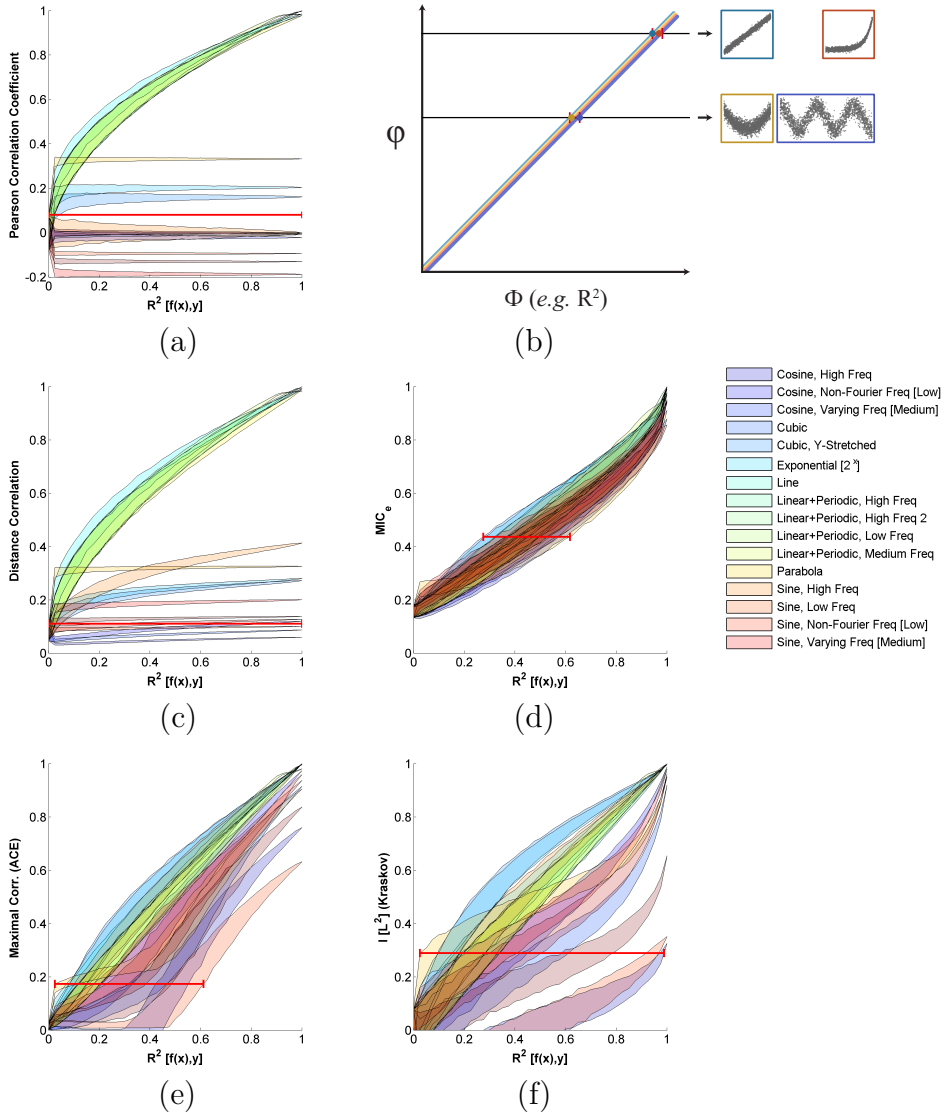


Figure 3.3: Equitability with respect to R^2 on a set of noisy functional relationships of (a) the Pearson correlation coefficient, (b) a hypothetical measure of dependence φ with perfect equitability, (c) distance correlation, (d) MIC_e, (e) maximal correlation estimation, and (f) mutual information estimation. For each relationship, a shaded region denotes estimated 5th and 95th percentile values of the sampling distribution of the statistic in question on that relationship at every R^2. The resulting plot shows which values of R^2 correspond to a given value of each statistic. The red interval on each plot indicates the widest range of R^2 values corresponding to any one value of the statistic; the narrower the red interval, the higher the equitability.

in the set $\{0.01, 0.025, 0.05, 0.1, 0.2, 0.25, 0.5, 1\}$ to represent a range of possible relationship complexities. The results, shown in Figure 3.4, show that MIC_e outperforms existing methods in terms of equitability with respect to R^2 on these functions as well. Appendix Figure B.9 shows a version of this analysis under a different noise model that yields the same conclusion. We also examined the effect of outlier relationships on our results by repeatedly subsampling random subsets of 20 functions from this large set of relationships and measuring the equitability of each method on average over the subsets; results were similar.

One feature of the performance of MIC_e on these randomly chosen relationships that is demonstrated in Figure 3.4 is that it appears minimally sensitive to the bandwidth of the Gaussian process from which a given relationship is drawn. This puts it in contrast to, e.g., mutual information estimation, which shows a pronounced sensitivity to this parameter that prevents it from being highly equitable when relationships with different bandwidths are present in the same data set.

In Chapter 4, we perform more in-depth analyses of the equitability with respect to R^2 of MIC_e, MIC, and the four measures of dependence described above as well as the Hilbert-Schmidt independence criterion (HSIC)[52,53], the Heller-Heller-Gorfine (HHG) test[58], the data-derived partitions (DDP) test[59], and the randomized dependence coefficient (RDC)[86]. These analyses consider a range of sample sizes, noise models, marginal distributions, and parameter settings. They conclude that, in terms of equitability with respect to R^2 on the sets of noisy functional relationships studied, a) MIC_e uniformly

outperforms MIC, and b) MIC_e outperforms all the methods tested in the large majority of settings examined. Appendix Figure B.7 contains a reproduction of a representative equitability analysis from that paper for the reader's reference.

3.5 The total information coefficient

So far we have presented results about estimators of the population maximal information coefficient, a quantity for which equitability is the primary motivation. We now introduce and analyze a new measure of dependence, the *total information coefficient* (TIC). In contrast to the maximal information coefficient, the total information coefficient is designed not for equitability but rather as a test statistic for testing a null hypothesis of independence.

We begin by giving some intuition. Recall that the maximal information coefficient is the supremum of the characteristic matrix. While estimating the supremum of this matrix has many advantages, this estimation involves taking a maximum over many estimates of individual entries of the characteristic matrix. Since maxima of sets of random variables tend to become large as the number of variables grows, one can imagine that this procedure may lead to an undesirable positive bias in the case of statistical independence, when the population characteristic matrix equals 0. This might be detrimental for independence testing, when the sampling distribution of a statistic under a null hypothesis of independence is crucial.

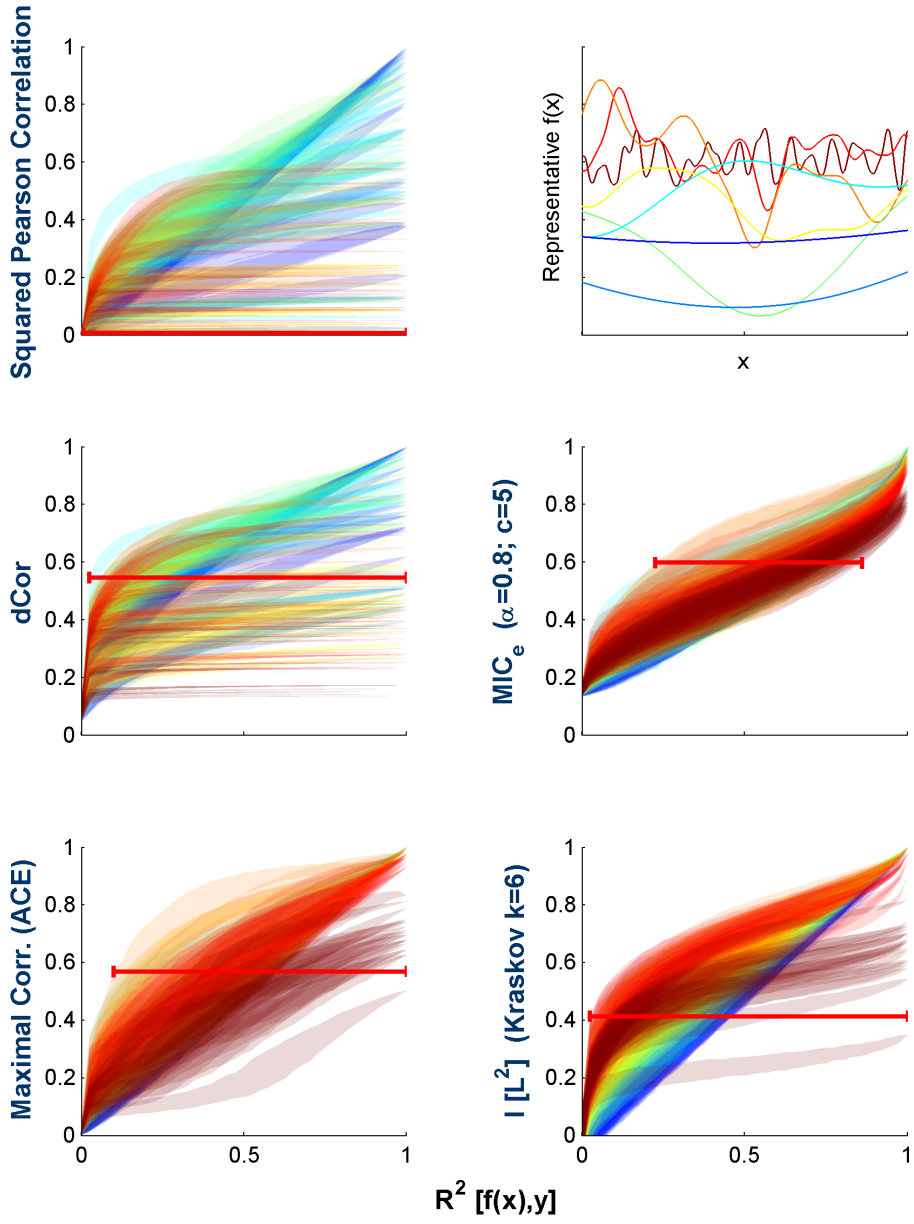


Figure 3.4: Equitability of methods examined on functions randomly drawn from a Gaussian process distribution. Each method is assessed as in Figure 3.3. Each shaded region corresponds to one relationship, and the regions are colored by the bandwidth of the Gaussian process from which they were sampled. Sample relationships for each bandwidth are shown in the top right with matching colors.

The intuition behind the total information coefficient is that if we instead consider a more stable property, such as the sum of the entries in the characteristic matrix, we might expect to obtain a statistic with a smaller bias in the case of independence and therefore better power. Stated differently, if our only goal is to distinguish any dependence at all from complete noise, then disregarding all of the sample characteristic matrix except for its maximal value may throw away useful signal, and the total information coefficient avoids this by summing all the entries.

We remark that in Reshef et al.[121] it is suggested that other properties of the characteristic matrix may allow us to measure other aspects of a given relationship besides its strength, and several such properties were defined. The total information coefficient fits within this conceptual framework.

In this section we define the total information coefficient in the case of both the characteristic matrix (TIC) and the equicharacteristic matrix (TIC_e). We then prove that both TIC and TIC_e yield independence tests that are consistent against all dependent alternatives. (As in the case of MIC and MIC_e, TIC_e is more easily computable than TIC.) Finally, we present a simulation study of the power of independence testing based on TIC_e on an index set of relationships chosen in Simon & Tibshirani[144], showing that TIC_e outperforms other common measures of dependence on many of the relationships and closely matches their performance on the rest.

3.5.1 Definition and consistency of the total information coefficient

We begin by defining the two versions of the total information coefficient. In the definition below, recall that $\widehat{M}$ denotes a sample characteristic matrix whereas $\widehat{[M]}$ denotes a sample equicharacteristic matrix.

Definition 3.5.1. Let $D \subset \mathbb{R}^2$ be a set of n ordered pairs, and let $B : \mathbb{Z}^+ \to \mathbb{Z}^+$. We define

$$\mathrm{TIC}_B(D) = \sum_{k\ell \leq B(n)} \widehat{M}(D)_{k,\ell}$$

and

$$\mathrm{TIC}_{e,B}(D) = \sum_{k\ell \leq B(n)} \widehat{[M]}(D)_{k,\ell}.$$

To show that these two statistics lead to consistent independence tests, we must take a step back and analyze the behavior of the analogous population quantities.

Definition 3.5.2. For a matrix A and a positive number B, the *B-partial sum* of A, denoted by $S_B(A)$, is

$$S_B(A) = \sum_{k\ell \leq B} A_{k,\ell}.$$

When A is an (equi)characteristic matrix, $S_B(A)$ is the sum over all entries corresponding to grids with at most B total cells. Thus, if $\widehat{M}(D)$ is a sample characteristic matrix of a sample D, $S_B(\widehat{M}(D)) = \mathrm{TIC}_B(D)$, and the same holds for $S_B(\widehat{[M]}(D))$ and

$\mathrm{TIC}_{e,B}(D)$.

It is clear that if X and Y are statistically independent random variables, then both the characteristic matrix $M(X,Y)$ and the equicharacteristic matrix $[M](X,Y)$ are identically 0, so that $S_B(M(X,Y)) = S_B([M](X,Y)) = 0$ for all B. However, we are also interested in how these quantities behave when X and Y are dependent. The following pair of propositions helps us understand this. The first proposition shows a lower bound on the values of entries in both $M(X,Y)$ and $[M](X,Y)$. The second proposition translates this into an asymptotic characterization of how quickly $S_B(M)$ and $S_B([M])$ grow as functions of B. These two propositions are the technical heart of why the total information coefficient yields a consistent independence test.

Proposition 3.5.3. *Let (X,Y) be a pair of jointly distributed random variables. If X and Y are statistically independent, then $M(X,Y) \equiv [M](X,Y) \equiv 0$. If not, then there exists some $a > 0$ and some integer $\ell_0 \geq 2$ such that*

$$M(X,Y)_{k,\ell}, [M](X,Y)_{k,\ell} \geq \frac{a}{\log \min\{k,\ell\}}$$

either for all $k \geq \ell \geq \ell_0$, or for all $\ell \geq k \geq \ell_0$.

Proof. See Appendix B.11.1 □

Proposition 3.5.4. *Let (X,Y) be a pair of jointly distributed random variables. If X and Y are statistically independent, then $S_B(M(X,Y)) = S_B([M](X,Y)) = 0$ for all*

$B > 0$. *If not, then* $S_B(M(X,Y))$ *and* $S_B([M](X,Y))$ *are both* $\Omega(B \log\log B)$.

Proof. See Appendix B.11.2 □

The propositions above, together with reasoning analogous to the convergence arguments presented earlier, can be used to show the main result of this section, namely that the statistics TIC and TIC_e yield consistent independence tests.

Theorem 3.5.1. *The statistics* TIC_B *and* $TIC_{e,B}$ *yield consistent right-tailed tests of independence, provided* $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ *for some* $\varepsilon > 0$.

Proof. See Appendix B.11.3. □

In practice, we often use the EquicharClump algorithm (see Section 3.4.3) to compute the equicharacteristic matrix from which we calculate TIC_e. This algorithm does not compute the sample equicharacteristic matrix exactly. However, as in the case of MIC_e, the use of the algorithm does not affect the theoretical properties of the statistic. This is proven in Appendix B.8.

3.5.2 Power of independence tests based on TIC_e

With the consistency of independence tests based on TIC and TIC_e established, we turn now to empirical evaluation of the power of independence testing based on TIC_e as computed using the EquicharClump algorithm.

To evaluate the power of TIC_e-based tests, we reproduced the analysis performed in Simon & Tibshirani[144]. Namely, we considered the set of relationships they analyzed, defined by

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

where F is a set of functions specified in Simon & Tibshirani[144]. (NB: one of the relationships is a circle, which we treat as a union of two half-circles.)

For each relationship $\mathcal{Z}$ in this set that we examined, we simulated a null hypothesis of independence with the same marginal distributions, and generated $1,000$ independent samples, each with a sample size of $n = 500$, from both $\mathcal{Z}$ and from the null distribution. These were used to estimate the power of the size-α right-tailed independence test based on each statistic being evaluated. Following Simon and Tibshirani, we compared TIC_e to the distance correlation[154,153], the original maximal information coefficient[121] as approximated using ApproxMIC, and to the Pearson correlation. (Though it is not a measure of dependence, the Pearson correlation was presumably included by Simon and Tibshirani as an intuitive benchmark for what is achievable under a linear model.) We also compared to MIC_e using identical parameters to those of TIC_e to examine whether the summation performed by TIC_e is better than maximization when all other things are equal. Note that we do not compare to methods of analyzing contingency tables, such as Pearson's chi-squared test. This is because our data are real-valued rather than

discrete, and so contingency-based methods are not applicable. However, when data are discrete, those methods can be very well powered.

The results of our analysis are presented in Figure 3.5. First, the figure shows that TIC_e compares quite favorably with distance correlation, a method considered to have state-of-the-art power[144]. Specifically, TIC_e uniformly outperforms distance correlation on 5 of the 8 relationship types examined, and performs comparably to it on the other three relationship types. We remark that distance correlation has many advantages over TIC_e, including the fact that it easily generalizes to higher-dimensional relationships and comes with an elegant and comprehensive theoretical framework.

The analysis also shows that TIC_e outperforms the original maximal information coefficient by a very large margin, and outperforms MIC_e as well, supporting the intuition that the summation performed by the former can indeed lead to substantial gains in power against independence over the maximization performed by the latter. (We note that in both Simon and Tibshirani's analysis and in this one, the original maximal information coefficient was run with default parameters that were optimized for equitability rather than power against independence. When run with different parameters, its power improves substantially, though it still does not match the power of MIC_e. See Appendix Figure B.8 and Chapter 4.)

In Chapter 4 we expand on this analysis, conducting an in-depth evaluation of the the power against independence of the tests described above as well as tests based on mutual information estimation[79], maximal correlation estimation[14], HSIC[52,53], HHG[58],

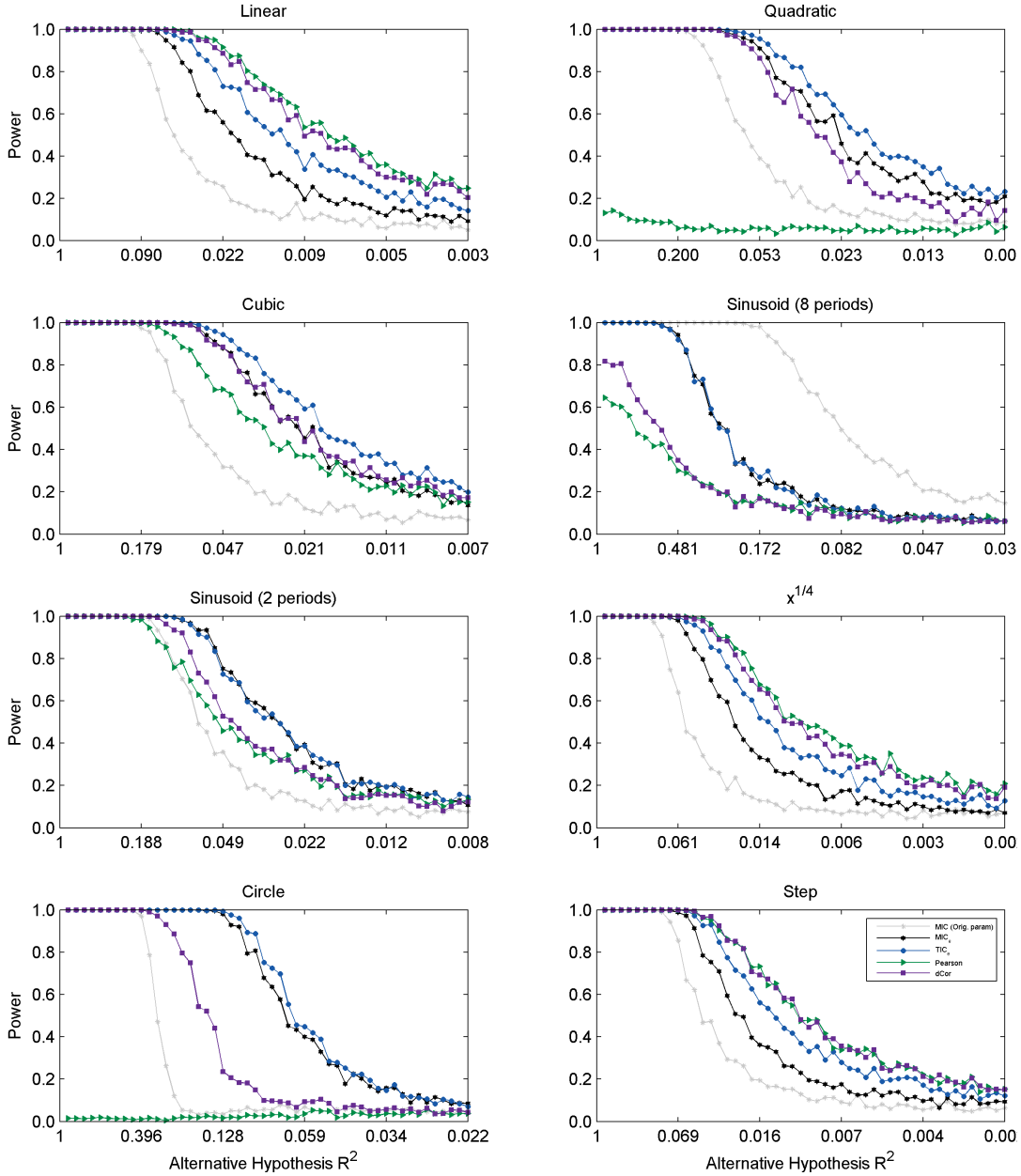


Figure 3.5: Comparison of power of independence testing based on TIC_e (blue) to MIC with default parameters (gray), MIC_e with the same parameters as TIC_e (black), distance correlation (purple), and the Pearson correlation coefficient (green) across several alternative hypothesis relationship types chosen by Simon & Tibshirani[144]. The relationships analyzed are described in Section 3.5.2.

DDP[59], and RDC[86]. These analyses consider a range of sample sizes and parameter settings, as well as a variety of ways of quantifying power across different alternative hypothesis relationship types and noise levels. They conclude that in most settings TIC_e either outperforms all the methods tested or performs comparably to the best ones. Appendix Figure B.8 contains a reproduction of one detailed set of power curves from the main analysis in that paper for the reader's reference.

3.6 Discussion

As high-dimensional data sets become increasingly common, data exploration requires not only statistics that can accurately detect a large number of non-trivial relationships in a data set, but also ones that can identify a smaller number of strongest relationships. The former property is achieved by measures of dependence that yield independence tests with high power; the latter is achieved by measures of dependence that are equitable with respect to some measure of relationship strength. In this chapter, we introduced two related measures of dependence that achieve these two goals, respectively, through the following theoretical contributions.

- A new population measure of dependence, MIC_*, that we proved can be viewed in three different ways: as the population value of the maximal information coefficient (MIC) from Reshef et al.[121], as a "minimal smoothing" of mutual information that makes it uniformly continuous, or as the supremum of an infinite sequence defined in terms of optimal partitions of one marginal at a time of a given joint distribution.
- An efficient approach for approximating the MIC_* of a given joint distribution.

- A statistic MIC_e that is a consistent estimator of MIC_*, is efficiently computable, and has good equitability with respect to R^2 both on a manually chosen set of noisy functional relationships as well as on a set of randomly chosen noisy functional relationships.
- The total information coefficient (TIC_e), a statistic that arises as a trivial side-product of the computation of MIC_e and yields a consistent and powerful independence test.

Though we presented here some empirical results for MIC_*, MIC_e, and TIC_e, our focus was on theoretical considerations; the performance of these methods is analyzed in detail in Chapter 4. In that chapter we show that on a large set of noisy functional relationships with varying noise and sampling properties, the asymptotic equitability with respect to R^2 of MIC_* is quite high and the equitability with respect to R^2 of MIC_e is state-of-the-art. We also show that the power of the independence test based on TIC_e is state-of-the-art across a wide variety of dependent alternative hypotheses. Finally, we demonstrate that the algorithms presented here allow for MIC_e and TIC_e to be computed simultaneously very quickly, enabling analysis of extremely large data sets using both statistics together.

Our contributions are of both theoretical and practical importance for several reasons. First, our characterization of MIC_* as the large-sample limit of MIC sheds light on the latter statistic. For example, while MIC is parametrized, MIC_* is not. Knowing that MIC converges in probability to MIC_* tells us that this parametrization is statistical only: it controls the bias/variance properties of the statistic, but not its asymptotic behavior.

Second, the normalization in the definition of MIC, while empirically seen to yield good performance, had previously not been theoretically understood. Our result that this normalization is the minimal smoothing necessary to make mutual information uniformly continuous provides for the first time a lens through which the normalization is canonical. In doing so, it constitutes an initial step toward understanding the role of the normalization in the performance of MIC_* and MIC. The uniform continuity of MIC_* and the lack of continuity of ordinary mutual information also suggest that estimation of the former may be easier in some sense than estimation of the latter. This is consonant with a recent result concerning difficulty of estimation of mutual information shown in Ding & Li[33]. It is also borne out empirically by the substantial finite-sample bias and variance we observe in Chapter 4 of the Kraskov mutual information estimator[79] compared to MIC_e.

Third, our alternate characterization of MIC_* in terms of one-dimensional optimization over partitions rather than two-dimensional optimization over grids enhances our understanding of how to efficiently compute it in the large-sample limit and estimate it from finite samples using MIC_e. This is a significant improvement over the previous state of affairs, in which the statistic MIC could only be approximated heuristically, with even the heuristic approximation being orders of magnitude slower than the results in this chapter now allow.

Finally, the introduction of the total information coefficient provides evidence that the basic approach of considering the set of normalized mutual information values achievable

by applying different grids to a joint distribution is of fundamental value in characterizing dependence. Interestingly, a statistic introduced in Heller et al.[59] follows a similar approach by considering the (non-normalized) sum of the mutual information values achieved by all possible finite grids. Consistent with our demonstration here that an aggregative grid-based approach works well, that statistic also achieves excellent power. (TIC_e is compared to the statistic from Heller et al.[59] in Chapter 4.)

Taken together, our results point to joint use of the statistics MIC_e and TIC_e as a theoretically grounded, computationally efficient, and highly practical approach to data exploration. Specifically, since the two statistics can be computed simultaneously with little extra cost beyond that of computing either individually, we propose computing both of them on all variable pairs in a data set, using TIC_e to filter out non-significant associations, and then using MIC_e to rank the remaining variable pairs. Such a strategy would have the advantage of leveraging the state-of-the-art power of TIC_e to substantially reduce the multiple-testing burden on MIC_e, while utilizing the latter statistic's state-of-the-art equitability to effectively rank relationships for follow-up by the practitioner.

Our results, while useful, nevertheless have limitations that warrant exploration in future work. First, for a sample D from the distribution of some random (X, Y), all of the sample quantities we define here use the naive estimate $I(D|_G)$ of the quantity $I((X, Y)|_G)$ for various grids G. There is a long and fruitful line of work on more sophisticated estimators of the discrete mutual information Paninski[108] whose use instead of

$I(D|_G)$ could improve the statistics introduced here. Second, our approach to approximating the MIC_* of a given joint density consists of computing a finite subset of an infinite set whose supremum we seek to calculate. However, the choice of how large a finite set we should compute in order to approximate the supremum to a given precision remains heuristic. Finally, though empirical characterization of the equitability of MIC_e on representative sets of relationships is important and promising, we are still missing a theoretical characterization of its equitability in the large-sample limit. A clear theoretical demarcation of the set of relationships on which MIC_* achieves good equitability with respect to R^2, and an understanding of why that is, would greatly advance our understanding of both MIC_* and equitability.

3.7 Acknowledgements

We would like to acknowledge R Adams, E Airoldi, T Broderick, A Gelman, M Gorfine, R Heller, J Huggins, T Jaakkola, J Mueller, J Tenenbaum, and R Tibshirani for constructive conversations and useful feedback.

4

An empirical study of the maximal and total information coefficients and leading measures of dependence

Having introduced both equitability as a concept and a two novel measures of dependence, MIC_e and TIC_e, we now turn to in-depth empirical evaluation. Specifically, we evaluate the equitability, power against independence, and runtime of several leading measures of dependence including MIC_e and TIC_e.

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

to filter relationships by TIC_e and then to rank the remaining ones using MIC_e. We confirm our findings on a set of data collected by the World Health Organization.*

4.1 Introduction

Though Chapter 2 and Chapter 3 introduce valuable theory and conduct proof-of-concept empirical analyses, both equitability and power against independence are highly dependent on many different aspects of the data-generating process: which trend(s) govern(s) the relationships in question? What type of noise is added to the data? What is the sample size? What are the marginal distributions of the variables in question? Runtime, while generally simpler to analyze, in some cases also depends on these questions. A proper understanding of the relative merits of different measures of dependence therefore requires thorough empirical exploration of these questions.

In this chapter, we therefore compare under a wide range of settings the equitability on functional relationships, power, and runtime of MIC_e, TIC_e, and a suite of leading measures of dependence. With regard to equitability, our results show that estimation of the population MIC via MIC_e is more equitable on functional relationships than other methods in a large majority of the settings of noise/marginal distributions and sample

*The material in this chapter is adapted from a manuscript published in the March 2018 edition of the *Annals of Applied Statistics* as "An empirical study of the maximal and total information coefficients and leading measures of dependence" by David Reshef*, Yakir Reshef*, *et al.*[128] (* = co-first author), as well as a technical comment published in the August 2014 edition of the *Proceedings of the National Academy of Sciences* by David Reshef*, Yakir Reshef*, *et al.*[124] (* = co-first author)

size that we tested, though in a few settings the Kraskov mutual information estimator outperforms MIC_e. With regard to power against independence, we find that TIC_e and a related method called S^{DDP}[59] share state-of-the-art performance, and that many other methods including distance correlation[153] also do quite well. We also characterize a more general power-equitability trade-off that holds across methods, and we present a runtime analysis to characterize the scale of data that each method can analyze. Our full set of simulation analyses of power, equitability, and runtime, including sensitivity analyses and additional sample sizes and models, are available in an online empirical supplement that we hope will be a resource to the community (see Appendix C.1).

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

4.2 Equitability analysis

We begin by evaluating the equitability of MIC_e, TIC_e, and several leading measures of dependence. We do so first using $\mathcal{Q}$-confidence intervals, followed by an alternate visualization of the equitability of each measure of dependence using a power analysis.

4.2.1 Setting up the analysis

Choice of methods to analyze

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

Grid-based methods The maximal information coefficient and the total information coefficient can be viewed as exploring the space of possible grids that can be drawn on the sampled data, assigning a score to each grid via some metric, and then aggregating the scores. For MIC[121], the metric is a normalized mutual information score and the aggregation is a supremum. (We remind the reader that MIC is difficult to compute efficiently and so in practice a heuristic approximation called APPROX-MIC is used to compute it that does not explore the space of all possible grids.) MIC_e, introduced in Chapter 3 of this thesis, is similar to MIC but explores a more restricted set of grids over which an efficient search is possible while retaining the property that its population value is a supremum over all possible grids. (As such, no approximation algorithm is

needed for MIC_e.) TIC_e, also introduced in Chapter 3 of this thesis, is like MIC_e except it aggregates by summation. For all of these methods, the parameter α controls the space of grids that is explored; higher α means grids with more cells.

We also include other, more recent grid-based methods. HHG[58] explores a set of three-by-three grids defined by pairs of data points, uses as its score Pearson's χ^2 test statistic computed on two-by-two contingency tables derived from the three-by-three grids, and aggregates by summation. Though similar to Hoeffding's D[60] in that it proceeds via two-by-two contingency tables, it differs in the way it constructs the tables, and it is not distribution free whereas Hoeffding's D is. S^{DDP}[59] explores a larger set of grids defined by subsets of the data points, uses non-normalized mutual information as its score, and also aggregates by summation.[†] Another notable grid-based method introduced recently is dynamic slicing[66], which like the idealized MIC explores all possible grids and aggregates by maximization, but uses as its score a version of mutual information that is regularized according to a prior on the space of possible grids. We did not include dynamic slicing in our comparison, however, because it is formulated only for performing a k-sample test whereas our focus here is on measuring dependence between two continuous random variables.

MUTUAL INFORMATION ESTIMATION We compare to a standard mutual information estimator introduced by Kraskov[79]. For convenience, we represent the estimated

[†]Several variations on these statistics are presented in Heller et al.[58],[59]. Results for these other methods were generally similar or worse than the ones we display, and we omit them.

mutual information values in terms of the squared Linfoot correlation[146,84], defined by $L^2(X,Y) = 1 - 2^{-2I(X,Y)}$ where $I(X,Y)$ represents the raw mutual information. $L^2(X,Y)$ takes values in $[0,1]$.

Distance/kernel-based statistics We compare to the distance correlation (dCor)[153], a statistic that is defined analogously to ordinary correlation but using a notion of *distance* variance/covariance that is based on pairwise distances between points. The use of distance variance/covariance is a significant advance because in contrast to ordinary variance/covariance it produces an omnibus consistent test that, unlike grid-based approaches, easily generalizes to testing for dependence in higher dimensions. In addition to distance correlation, we compare to the Hilbert-Schmidt Information Criterion (HSIC)[52,54], a more general statistic defined on reproducing kernel Hilbert spaces of which dCor is a special case[140].

Correlation-based methods As an intuitive benchmark for the reader, we include the squared Pearson correlation coefficient (ρ^2). We also include methods that use ρ after computing a non-linear transformation of the data. Perhaps the best-known one is maximal correlation[120], which given random variables X and Y searches for arbitrary measurable functions f and g such that $\rho(f(X), g(Y))$ is maximized. This is algorithmically hard in general, but the (approximate) method of alternating conditional expectations[14] is widely used and we use it here as well. We also include a more

recent related method, the randomized dependence coefficient[86], which applies many random transformations to X and Y and then searches for the linear combinations of the transformed features that maximize the correlation.

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

Choice of $\mathcal{Q}$, Φ, and sample sizes

We focus here on equitability with respect to R^2 on a set of noisy functional relationships. To ensure robustness, we vary the relationships tested along as many dimensions as possible including relationship type, the type of noise added, marginal distributions, and sample size.

Specifically, we considered 12 different sampling/noise models. Each sampling/noise model was defined by choosing one of four independent-variable marginal distributions (points equidistant or uniformly sampled, along the graph of the function or along the X-axis) and one of three noise distributions (X noise only, Y noise only, or noise in both variables; see Appendix C.2.4). For each sampling/noise model, we created a set of relationships $\mathcal{Q}$ by including between 16 and 21 different functional relationships (see Appendix C.2.4), each with increasing levels of additive Gaussian noise, at four sample size regimes ($n = 250, 500, 5,000$, and the infinite data limit).

Quantification of equitability

The equitability of each measure of dependence is quantified using level-0.05 $\mathcal{Q}$-confidence intervals (see Chapter 2). We report both average-case and worst-case equitability in our analyses, and the interval plotted in red on each plot represents the largest $\mathcal{Q}$-confidence interval for that plot.

4.2.2 Results and discussion

Figures 4.1 and 4.2 display the results of our analysis for a subset of methods under the noise/sampling model $(x_i + \varepsilon_i, f(x_i) + \varepsilon_i')$, where $\varepsilon_i, \varepsilon_i'$ are i.i.d. Gaussians for all i and the x_i are chosen to make consecutive points $(x_i, f(x_i))$ equidistant along the graph of f. The full results are in Empirical Supplement 1A-F, and are summarized in Tables C.3 and C.4.

Our results demonstrate that MIC_e is consistently highly equitable for these noise/sample models and sample sizes, and is the only one of the methods examined here to be so. Mutual information shows relatively poorer equitability at $n = 250$ and $n = 500$. While its equitability appears improved at $n = 5,000$, this improvement is not robust to variation in noise/sampling model; we discuss the equitability of mutual information in more detail in the following section. Of the remaining schemes, maximum correlation appears to provide the best equitability. This is interesting because on the one hand the squared maximal correlation is bounded from below by R^2, and on the other hand the lack of

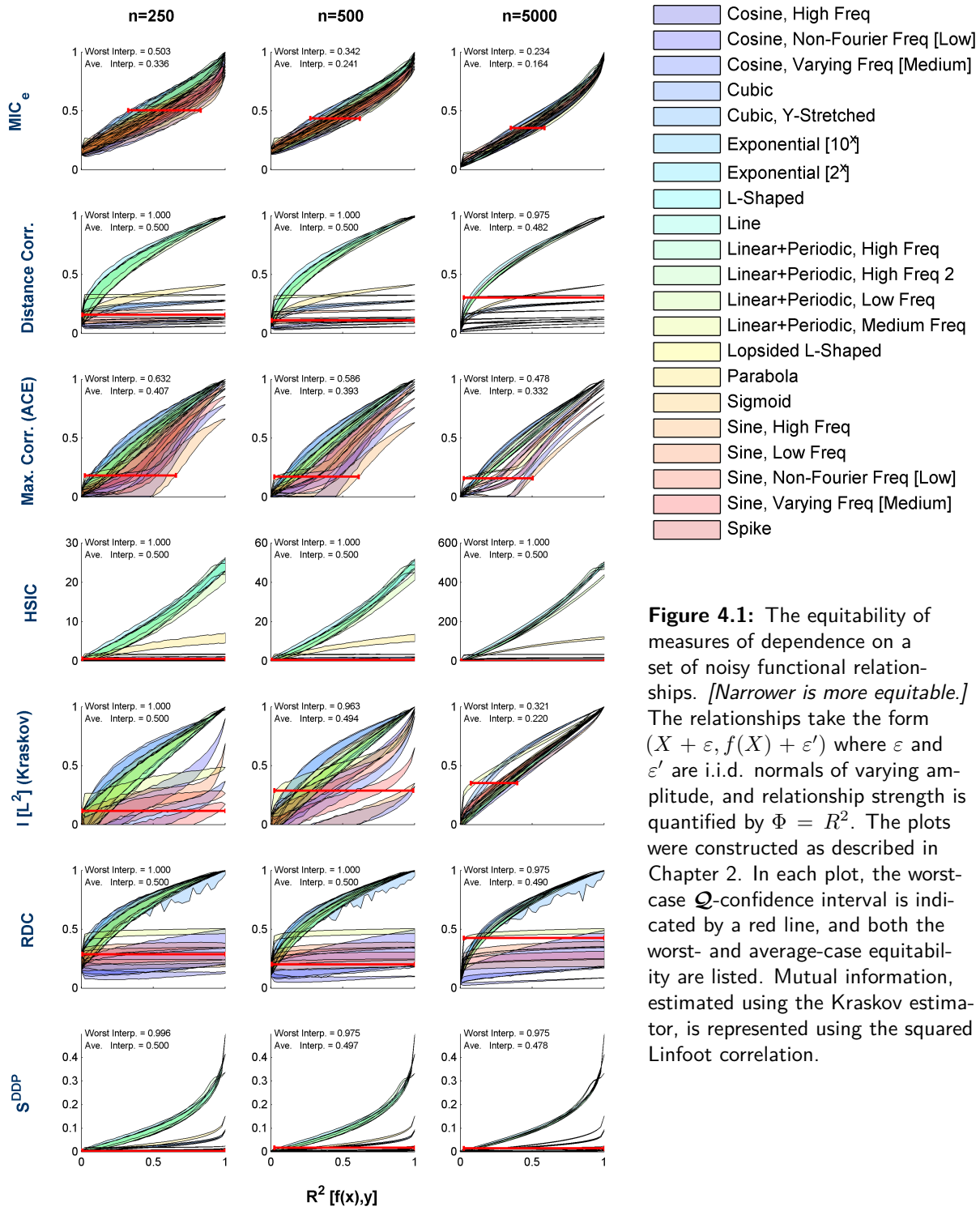


Figure 4.1: The equitability of measures of dependence on a set of noisy functional relationships. *[Narrower is more equitable.]* The relationships take the form $(X + \varepsilon, f(X) + \varepsilon')$ where ε and ε' are i.i.d. normals of varying amplitude, and relationship strength is quantified by $\Phi = R^2$. The plots were constructed as described in Chapter 2. In each plot, the worst-case $\mathcal{Q}$-confidence interval is indicated by a red line, and both the worst- and average-case equitability are listed. Mutual information, estimated using the Kraskov estimator, is represented using the squared Linfoot correlation.

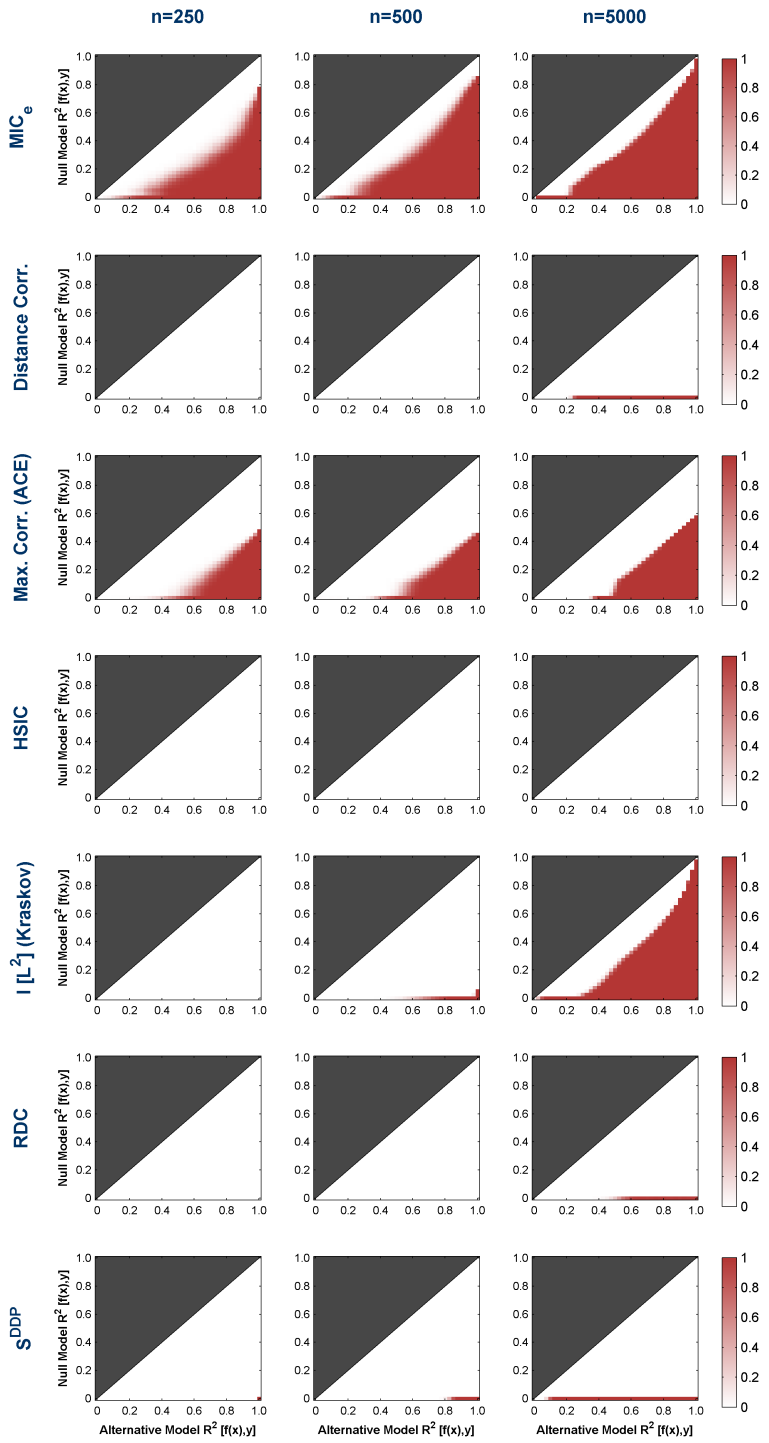


Figure 4.2: The equitability with respect to $\Phi = R^2$ of measures of dependence on the noisy functional relationships analyzed in Figure 4.1, visualized in terms of power. *[Redder is more equitable.]* Plots were generated as in Chapter 2. The intensity of the pixel at coordinate (x_1, x_0) in each heat map shows the power of a right-tailed test based on the statistic in question at distinguishing the (composite) alternative hypothesis $H_1 : R^2 = x_1$ from the (composite) null hypothesis $H_0 : R^2 = x_0$ with type I error at most $\alpha = 0.05$. Mutual information, estimated using the Kraskov estimator, is represented using the squared Linfoot correlation. For every parametrized statistic whose parameter meaningfully affects equitability, results are presented at each sample size using parameter settings that maximize worst-case equitability across all twelve of noise/marginal distributions tested at that sample size.

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

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

Interestingly, Figure 4.2 also shows poor power to reject a null hypothesis of independence at $n = 250$ and $n = 500$ even for methods like HSIC, dCor, and RDC, which are traditionally considered to have good power against independence. The reason this happens is because equitability measures worst-case power across all relationship types with a given R^2; that is, the alternative hypotheses considered are composite. Corre-

spondingly, the statistics that are not well powered to detect even one of the relationship types analyzed compare unfavorably in this analysis, even if they have good power on a large subset of the relationships.

We note that for MIC_e, the best parameter regime for equitability is different than the best parameter regime for power against independence presented later in this chapter. This suggests that there is a trade-off between power against independence and equitability, a theme to which we return in Section 4.5.

Comparing the equitability of MIC_e and mutual information

Given the connections between the maximal information coefficient and mutual information, it is natural to ask whether direct estimation of mutual information achieves similar equitability to MIC_e. The equitability of mutual information estimation has been assessed previously, most notably in Reshef et al.[121], Reshef et al.[122], Kinney & Atwal[76], and Reshef et al.[123]. The analyses conducted here, which subsume those analyses, show that in general the answer appears to be: at $n = 250$ and $n = 500$, MIC_e outperforms mutual information estimation on all models tested, often by substantial margins; at $n = 5,000$, MIC_e outperforms mutual information on all models except for the ones that contain Y noise only, on which mutual information performs better. We present a more detailed breakdown below.

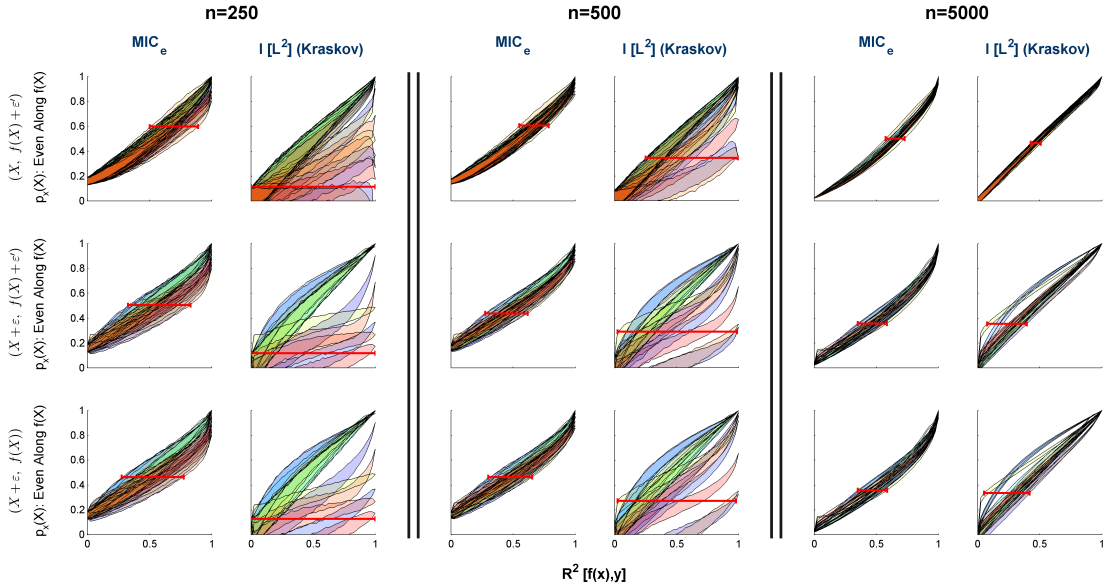


Figure 4.3: A comparison of the equitability of MIC_e and mutual information estimation under three noise models. *[Narrower is more equitable.]* Plots are analogous to those in Figure 4.1. As in that figure, results for both statistics are presented for each sample size using parameter settings that maximize equitability across all twelve of the noise/marginal distributions tested at that sample size. For versions of this analysis using additional independent variable marginal distributions, see Empirical Supplement 1C.

EFFECT OF MODEL CHOICE ON EQUITABILITY Figure 4.3 demonstrates the relative robustness to model choice of the equitability of MIC_e compared to that of the Kraskov mutual information estimator. At each sample size, the equitability of MIC_e is fairly stable with respect to the variations in noise models and independent variable marginal distributions tested. In contrast, it seems that mutual information estimation's equitability relies on the noise model containing no X noise.

EFFECT OF SAMPLE SIZE ON EQUITABILITY Estimating mutual information from finite samples is a challenging problem that has inspired many sophisticated methods[108,95,79],

and indeed our analyses demonstrate strong finite-sample effects on the equitability of mutual information estimation. MIC_e suffers much less from this problem: for $n = 250$ and $n = 500$, MIC_e has both superior worst-case and average-case equitability over mutual information estimation (using $k = 1$, 6, 10, and 20 in the Kraskov estimator) in every model $\mathcal{Q}$ tested, in most cases by substantial margins. This is intuitively consistent with the fact, proven in Chapter 3, that the population value of MIC_e is uniformly continuous as a functional while mutual information is not.

Equitability in the large-sample limit To disentangle finite-sample effects from properties of the population values of the statistics in question, we also compared the equitability of the population value of MIC_e (called MIC_*) and the population value of mutual information (Figure C.4). Results were essentially the same as those for $n = 5,000$, implying that neither MIC_* nor mutual information is worst-case perfectly equitable with respect to R^2 over the sets $\mathcal{Q}$ examined. This is not surprising given the broad range of relationships, noise models, and independent variable marginal distributions tested.

Relationship to equitability analysis from Kinney & Atwal[76] A more limited empirical analysis of the equitability of MIC and mutual information estimation was presented in Kinney & Atwal[76]. There, the authors examined the equitability of MIC and mutual information estimation at a large sample size ($n = 5,000$) and under

one choice of $\mathcal{Q}$ (the same as in Figure 4.1, only with no noise in the first coordinate). From this, they concluded that mutual information estimation was more equitable than MIC. This empirical argument was accompanied by a theoretical result exhibiting a family of relationships on which no measure of dependence can be perfectly equitable with respect to R^2, and a statement that this impossibility result implies that previous claims[121] about the equitability of MIC were incorrect.

Since its publication, Kinney & Atwal[76] has been the subject of two published technical comments[123,98] describing its main limitations, which are threefold. First, the central proof of the impossibility of equitability with respect to R^2 in Kinney & Atwal[76] applies only to *perfect* equitability, and says nothing about the achievability of the more general (approximate) notion with which we are primarily interested and regarding which we have previously made claims about MIC. That is, even if no method is perfectly equitable with respect to R^2, some methods can be more equitable with respect to R^2 than others, and the question remains which methods come meaningfully close to the ideal[123]. Second, the impossibility result relies crucially on a non-identifiable noise model $\mathcal{Q}$ in which, e.g., a noiseless parabola can be obtained as a "noisy" linear relationship[98]. Third, though mutual information indeed outperforms MIC under the specific sample size and noise model chosen in Kinney & Atwal[76], this is not the case in general[123]. As our analysis here importantly establishes, this empirical point remains true even when we further expand the set of noise models and sample sizes under consideration.

Sensitivity of analysis to choice of functions

One potential question about the equitability analyses performed here is whether they are sensitive to the particular choice of functions analyzed. This is justified given that the current theoretical understanding of the maximal set of functions on which we should expect MIC_e (or any method) to behave equitably is quite limited, and given that one can construct functions, such as a step function, for which all three of the methods that show non-trivial equitability in the above analysis provably perform very non-equitably. (See Appendix C.9 for a proof.) However, additional analyses in Chapter 3 of this thesis suggest that our results appear robust over a wide range of "probable" function types. Specifically, we conducted in that chapter equitability analyses similar to the ones above but on a set of 160 functions chosen at random from Gaussian process distributions with radial basis function kernels of different bandwidths. Results were similar, with MIC_e attaining the best equitability, followed by mutual information estimation and then maximal correlation.

Non-functional relationships

Equitability as we have applied it here is only defined for noisy functional relationships. However, in previous work[121] we showed empirically in the case of MIC that reasonable equitability with respect to R^2 can translate into reasonable behavior on several different non-functional relationships, with the MIC of those relationships degrading intuitively

as noise is added (see Figures 2G, S5, and S6 of Reshef et al.[121]). We also proved that the population MIC (and therefore also the population MIC_e) of superpositions of noiseless never-constant functional relationships is 1 (see Theorem 4 of Reshef et al.[121]). More in-depth empirical and theoretical examination of this aspect of MIC and MIC_e is an important direction of future work.

4.3 Statistical power analysis

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

Here, we turn to assessing the power against independence of the above statistics. Such analyses have been done previously, most notably by Simon and Tibshirani[144]. Our analysis expands upon the power analysis performed by Simon and Tibshirani in three key ways. First, for each of the statistics we analyze that has a free parameter, we perform a parameter sweep to understand the power of the corresponding tests as a function of that parameter. Second, we analyze a larger set of methods and a greater variety of sample sizes. Finally, we consider several ways to aggregate information across noise levels and across function types to get a more general picture of which methods have the highest overall power.

4.3.1 Setting up the analysis

We analyze all methods listed in Section 4.2, and we perform parameter sweeps for every method and report best-seen results as in that section. We use the set of relationships and noise model (uniform independent-variable marginal, Gaussian noise in the second coordinate only) chosen by Simon and Tibshirani[144]. For consistency with the sample sizes used throughout this work, we show results for $n = 500$; results for all analyses using $n = 100$ are similar and are provided in the empirical supplement.

We first compute power curves for each relationship type and each method, having performed parameter sweeps to choose optimal parameters for each method as a function of sample size only (see Appendix C.2.2). The parameter sweeps themselves, which characterize power against independence as a function of statistic parameters, are presented in Figures C.6 and C.7.

To compare power across methods, we need to aggregate information across relationship types as well as across alternative hypotheses. The first way we do this is to integrate under the power curve of each relationship type and average across relationship types, using limits of integration defined via R^2 for consistency across relationship types (see Appendix C.2.2). The second way we aggregate this information is to to compute, for each method and each function type, the R^2 at which 50% power is reached, and then average this quantity across function type.

4.3.2 Results

The full power curves for individual relationship types and methods are displayed in Figure 4.4, and the aggregated results comparing overall power across methods are shown in Figure 4.5. We discuss several aspects of these below.

Power on specific relationships

In Figure 4.4 no method clearly dominates; different methods have good power for different relationship types. For example, distance correlation and HSIC are relatively better powered to detect linear dependence than MIC_e and TIC_e, but are relatively worse at detecting most of the other forms of dependence tested. In contrast, S^{DDP} appears to have a similar profile to that of TIC_e. This is interesting because S^{DDP} is closely related to the maximal and total information coefficients in that it too is an aggregation via summation of mutual information scores taken over many different grids.

However, choice of parameter values is an important determinant of power, and unsurprisingly, the optimal parameter choices used here cause the power of tests based on several of the statistics included in this analysis to be substantially better than previously reported[144,49,86,76,66]. In particular, MIC with optimal parameters (black line) performs substantially better than MIC with what were previously the default, equitability-oriented parameters. This performance gain is achieved by a wide range

of parameter settings comprising a regime suited for independence testing (Figure C.6). Importantly, it is preserved on an independent validation set of randomly chosen noisy functional relationships (Figure C.1), indicating that it is not idiosyncratic to the particular relationships employed in this analysis. We have therefore updated our software to allow users to choose between the parameters that optimize power or the parameters that optimize equitability.

Average power across relationship types

The two rankings displayed in Figure 4.5, while robust to sample size and thresholds used, are different from each other, and are sensitive to choices such as inclusion/exclusion of certain function types (Empirical Supplement 2A and 2B). However, there are some general patterns that seem consistent. First, state-of-the-art performance is always achieved by either S^{DDP} or TIC_e, depending on how power is quantified. This provides evidence that the basic approach of aggregating mutual information scores over a large set of grids, whether via the characteristic matrix or other statistics, is a fundamentally promising avenue for thinking about dependence. Additionally, when power is quantified by computing the area under the power curve, distance correlation also does quite well, thus highlighting the value of the by-now well established paradigm of energy statistics for relationship detection.

Second, the power of independence testing using MIC_e, with parameters suited for power against independence rather than equitability, is not far from the state of the art,

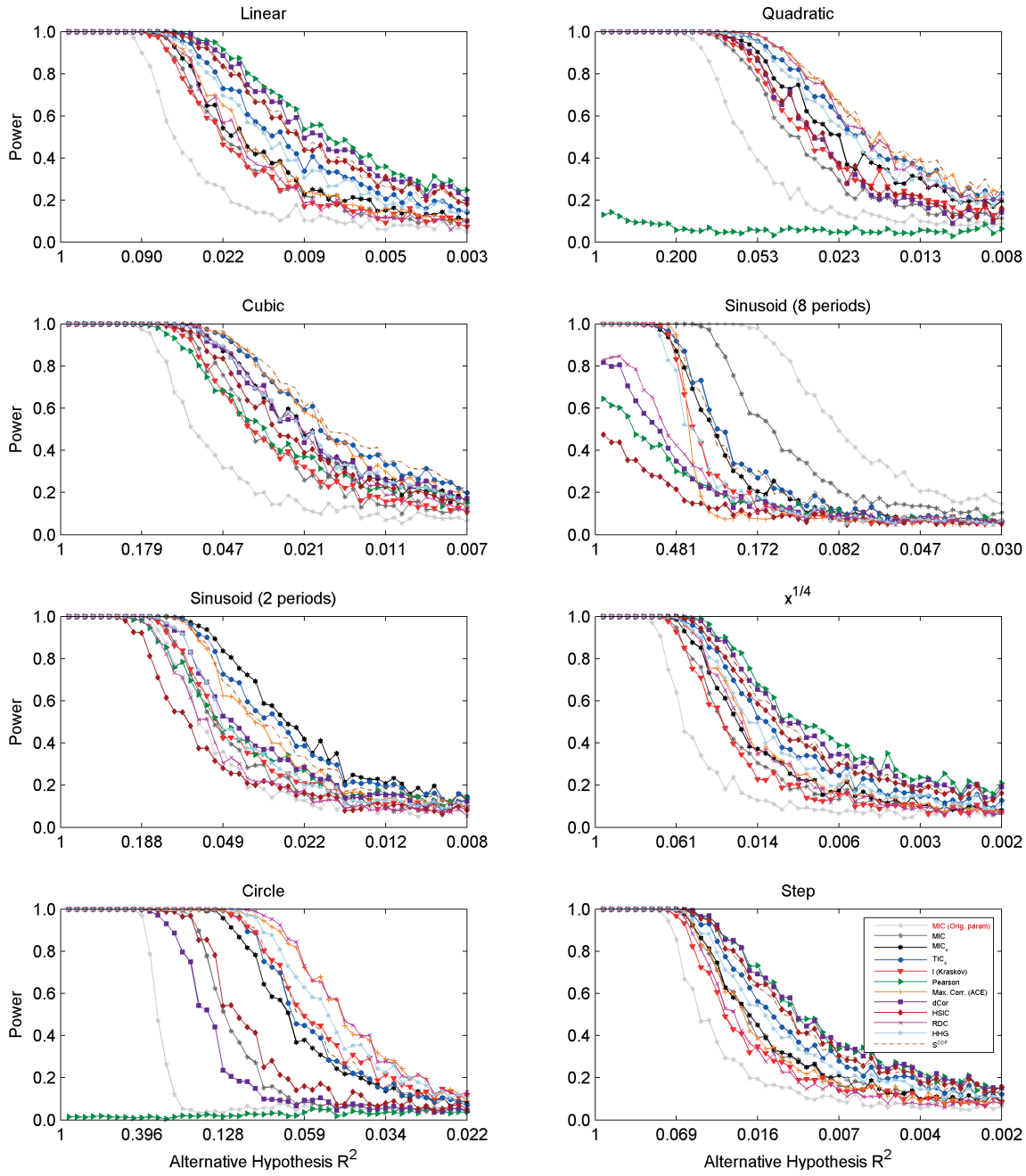
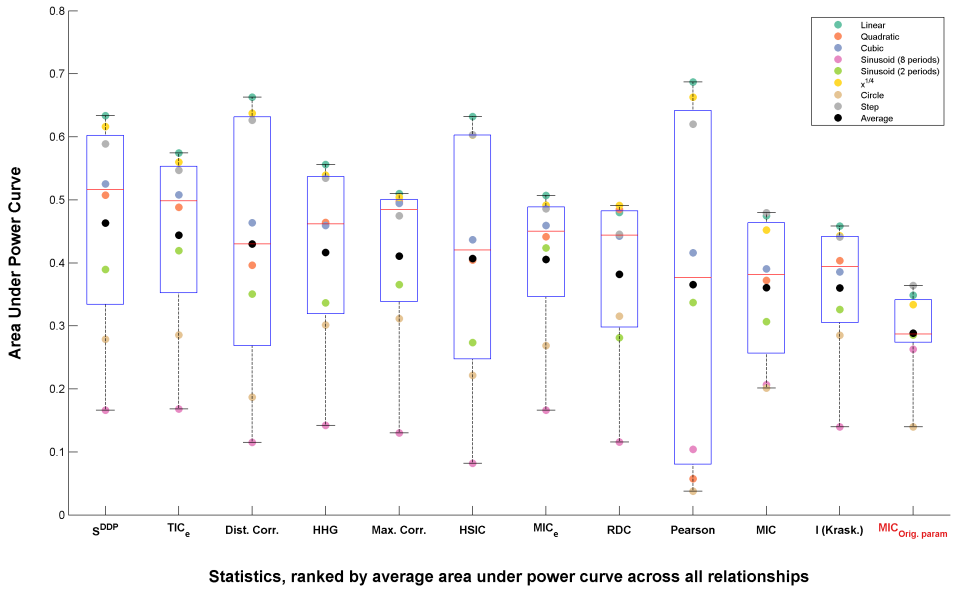


Figure 4.4: Power of independence testing using the measures of dependence examined, on the relationships in Simon & Tibshirani [144], at 50 noise levels with linearly increasing magnitude for each relationship and $n = 500$. For each statistic that has a parameter, an optimal value for the parameter was chosen using the parameter sweeps in Figure C.6. (For a version with $n = 100$ see Empirical Supplement 2A.)

(a)



(b)

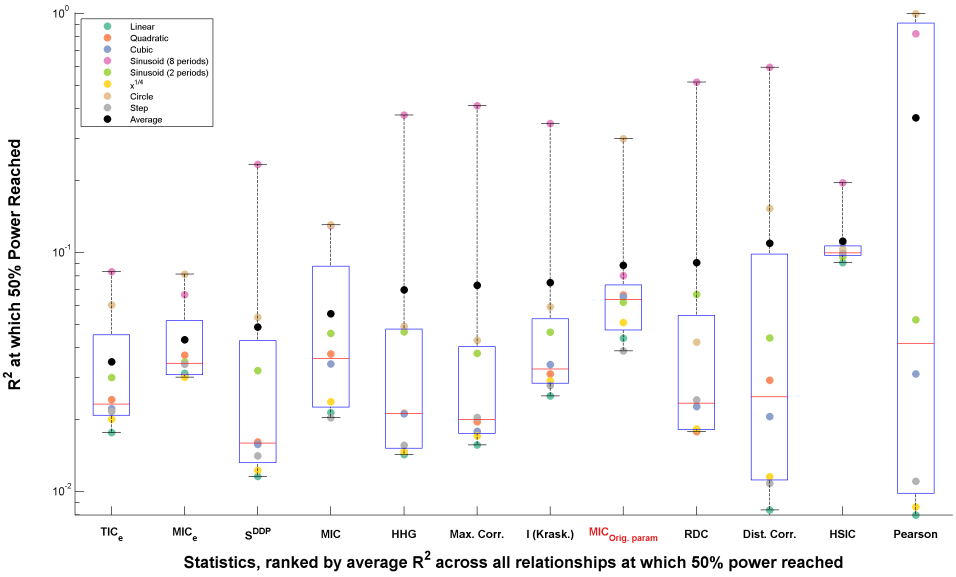


Figure 4.5: Measures of dependence ranked by the power of their corresponding independence tests. For each measure of dependence and each relationship type, power was quantified using (a) the area under the power curve *[higher is more powerful]*, or (b) the minimal R^2 at which at least 50% power is achieved *[lower is more powerful]*. The collection of these scores across relationship types is then plotted for each method along with quartiles. Optimal parameter values were chosen to maximize average performance across relationships; see (a) Figure C.6, or (b) Figure C.7. The MIC statistic from Reshef et al.[121] with the parameters used in Simon & Tibshirani[144] is labeled in red. The sample size is $n = 500$; results are similar with $n = 100$ and, for (b), with power thresholds besides 50%. (See Empirical Supplement 2B.)

though its relative performance depends on the method of quantification. In particular its power is comparable to- and usually higher than that of its predecessor MIC[121], which estimates the same population quantity (MIC_*), even when the latter also has optimally chosen parameters. This demonstrates that the improved bias/variance properties of MIC_e relative to MIC shown in Chapter 3 indeed translate into an improvement in power.

Another observation arising from Figure 4.5b is that if we computed for each method the R^2 at which 50% power is reached for *all* function types tested simultaneously—i.e., the maximum over function types of R^2 at which 50% power is reached, instead of the average over function types—then the rankings of the methods would be quite different. We will return to this in Section 4.6, where we argue for the utility of this way of assessing performance.

We remark that since the parameters chosen for each parametrized method were optimized for the function suite we analyzed, one may ask to what extent these results would generalize to relationship types beyond the ones considered here. To assess this, we also conducted the same power analysis on an independent set of 160 relationships consisting of randomly chosen functions with noise added, using the parameter settings resulting from our parameter sweep of the fixed set of relationships. Results were similar. (See Appendix C.2.3.)

4.4 Runtime analysis

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

4.4.1 Setting up the analysis

Since the runtime of MIC_e/TIC_e depends on parameter choice, results for MIC_e are presented for parameter settings recommended for maximizing equitability, maximizing power against independence, and attaining "reasonable equitability". The third set of parameters was computed by searching at each sample size for the parameters that resulted in the fastest runtime while still yielding 80% of the best observed equitability at that sample size. All the parameters used for MIC_e/TIC_e in this analysis are detailed in Table C.8.

The only other method whose runtime is affected by its parameter was S^{DDP}. Since at the sample size regimes we tested only three parameter settings led to practical runtimes for S^{DDP}, we have included all three. For statistics whose runtimes did not depend on parameter choice, defaults were used (see Appendix C.6.3).

4.4.2 Results

The results of our runtime analysis, found in Table 4.1, have several salient features. First, there is a clear set of fastest methods: maximum correlation, RDC, MIC_e (with

n	ρ^2	Max. Corr.	RDC	dCor	HSIC	HHG	$\text{I}_{(\text{Kraskov})}$
50	0.0001	0.0004	0.0015	0.0010	0.0016	0.0017	0.0096
100	0.0001	0.0005	0.0014	0.0014	0.0032	0.0063	0.0100
500	0.0001	0.0014	0.0023	0.0504	0.0847	0.2185	0.0122
1,000	0.0002	0.0025	0.0035	0.3518	0.4886	1.0956	0.0150
5,000	0.0002	0.0119	0.0129	6.1402	6.5975	34.0171	0.0427
10,000	0.0002	0.0239	0.0251	25.9859	25.7333	465.3222	0.0927

n	MIC	MIC_e [E]	MIC_e [FE]	MIC_e [P]	${S^{DDP}}_{m=2}$	${S^{DDP}}_{m=3}$	${S^{DDP}}_{m=4}$
50	0.0015	0.0021	0.0009	0.0004	0.0018	0.0010	0.0094
100	0.0061	0.0052	0.0012	0.0005	0.0022	0.0023	0.0861
500	0.2187	0.1630	0.0079	0.0018	0.0035	0.0529	14.2690
1,000	0.9628	0.1992	0.0172	0.0037	0.0050	0.2122	121.7311
5,000	18.7627	0.3398	0.0974	0.0195	0.0574	5.7464	1.72×10^4
10,000	66.2238	0.6835	0.1819	0.0398	0.2154	23.4473	1.40×10^5

Table 4.1: Average runtimes, in seconds, of algorithms for computing measures of dependence over 100 trials of uniformly distributed, independent samples at a range of sample sizes. Results for MIC_e are presented for three sample-size-dependent parameter settings that optimize for maximal power against independence ([P]), 99% of optimal equitability ([E]), and 80% of optimal equitability (fast equitability, [FE]). For a list of the parameters used in each of these settings, see Table C.8. TIC_e is omitted because its runtime is very similar to that of MIC_e [P].

any of the three parameter settings tested), TIC_e (which has identical runtime to MIC_e and so is omitted from Table 4.1), mutual information, and S^{DDP} with $m = 2$ (a parameter setting that was not chosen by our parameter sweeps due to its worse power; see Figures C.6 and C.7). Each of these methods takes under a second to compute at a sample size of 10,000, while the remaining methods all take over 20 seconds.

Second, MIC_e with all three of the parameter settings given is substantially faster than the previously introduced MIC statistic from Reshef et al.[121] run using default parameters. This matches the theoretical analysis in Chapter 3, which shows that the complexity of the search procedure in MIC_e is $O(n^{2.5\alpha})$ whereas the complexity of the search procedure in the APPROX-MIC algorithm used to compute MIC is $O(n^{4\alpha})$.

Third, analysis of large data sets is possible using MIC_e and TIC_e. For example, computing both TIC_e with parameters optimized for power and MIC_e with parameters chosen to achieve 80% of the best achievable equitability can be done on a sample size of $5,000$ in 97 milliseconds. For a data set with $n = 5,000$ consisting of $1,000$ variables, this translates into a total runtime of 16 minutes to compute both statistics for all variable pairs using 50 processing nodes.

We note one interesting feature of the runtime of MIC_e. Since estimating MIC_* involves a search procedure, runtimes for estimating it are substantially faster when data contain less noise; as such, the runtimes on statistically independent data presented in Table 4.1 represent worst-case performance. When run on data drawn from a noiseless linear relationship at the same sample sizes, MIC_e ran 5%-75% faster. The runtime of S^{DDP} exhibited a similar phenomenon, but the runtimes of the other methods were insensitive to the level of structure present and did not exhibit this effect.

We emphasize that our results represent a snapshot based on currently available implementations. Just as MIC_e has provided an improvement over APPROX-MIC, and just as estimating distance correlation has recently been shown to be estimable in time $O(n \log n)$ rather than $O(n^2)$ (not benchmarked here; see Remark C.6.1), we expect that with time algorithmic improvements will allow for more efficient computation of some of the newer methods analyzed here.

4.5 The power-equitability trade-off

For several methods, the parameter regimes that maximize power are different from the parameter regimes that maximize equitability. This suggests that there may be a trade-off between these two objectives that is being captured by the choice of parameter setting[122]. Such a trade-off seems plausible given the equivalence proven in Chapter 2 between equitability and power against a range of null hypotheses corresponding to different relationship strengths. Since equitability is about simultaneously achieving high power against many null hypotheses, it is reasonable that to attain this objective we have to give up some of the power we previously had against the specific null hypothesis of independence. Here we show empirically that such a trade-off does indeed exist for each of the parametrized methods we consider.

4.5.1 Demonstrating the power-equitability trade-off

For each statistic under consideration we plotted worst-case equitability against average power at a sample size of 500 while varying the statistic's parameter if it had one. The results are displayed in Figure 4.6.

Figure 4.6 shows that every parametrized method with a non-trivial level of equitability does indeed exhibit a power-equitability trade-off on the sets of relationships considered in this chapter. In the case of MIC_e, the trade-off is captured by the parameter α, which controls the maximal grid resolution used by the statistic. This is

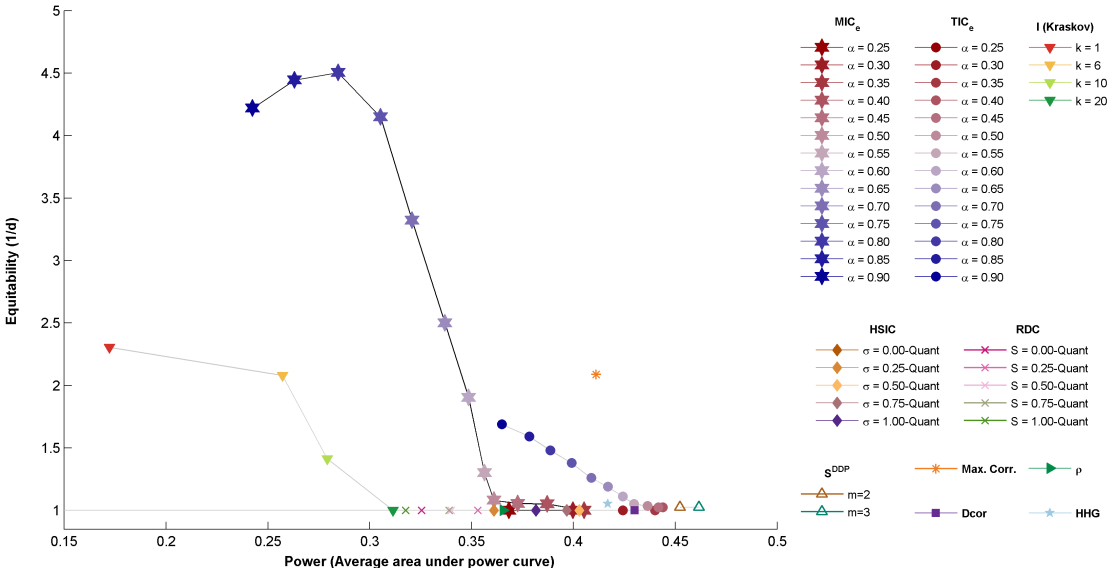


Figure 4.6: The trade-off between equitability and power against statistical independence across methods. For each method, average power as quantified in Figure 4.5a is plotted against the worst-case equitability under the analogous noise/sampling model, with $n = 500$. For every parametrized method, a point is plotted for each assessed value of the parameter in question. Since each coordinate is strictly preferable to all coordinates below and to the left of it, there is a Pareto "power-equitability" front. The methods with points along this front are MIC_e, maximal correlation, TIC_e, and S^{DDP}.

consistent with the bias-variance analysis in Chapter 3, which showed that low values of α lead to better performance in the low-signal regime while larger values of α lead to better performance in mid-to-high-signal regimes. It is also consistent with the intuition that disallowing high-resolution grids may increase power against independence but will allow only coarse-grained distinguishability among distributions, while allowing high-resolution grids might enable distinguishing between distributions that may be more similar to each other.

Figure 4.6 is also a useful summary of how the different methods we considered compare to each other along these two dimensions (for this sample size and set of

relationships). Specifically, if one point is both above and to the right of another then it is strictly preferable. Thus, the figure shows a Pareto front of methods that offer optimal performance with respect to power against independence and equitability. This front includes MIC_e, maximal correlation, TIC_e, and S^{DDP}. When power is assessed as in Figure 4.5b instead of Figure 4.5a, the Pareto front includes only MIC_e and TIC_e (see Empirical Supplement 3).

4.6 Practical suggestions

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

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

The situation we have chosen to focus on for this work is different: we are interested in the situation in which many methods will return a large number of relationships, and so the number of relationships detected is less important than their relative ranking. This

does happen in practice, for example in the gene expression analysis of Heller et al.[59], in which several methods identified over half of the thousands of relationships in the data set as significant, as well as in the analysis of the WHO data set conducted in Section 4.7 of this chapter. In such cases increasing the proportion of variable pairs identified as significant seems less important for scientific inquiry than having a meaningful way to prioritize the detected relationships for follow-up.

Thus, a promising strategy for exploratory data analysis is: first, to compute a statistic designed to identify a large number of significant relationships of all kinds, and then second, to compute an equitable statistic on all significant relationships, ensuring a ranking that is meaningful. For this approach to be fruitful, the statistic used in the first step must have high power on a wide range of relationship types; otherwise, the first step will eliminate many relationships that would otherwise be ranked as highly interesting in the second step. In other words, the statistic used in the first step should perform well with respect to the third quantification of power discussed at the end of Section 4.3: minimum R^2 at which 50% power is reached for all function types tested. Recall that in Chapter 2 of this thesis we defined the R^2 at which this is achieved as the *detection threshold* of the method; a method that does well with respect to this quantification of power has a *low detection threshold*. As described in Chapter 2, low detection threshold is related to equitability: an equitable statistic provably has a low detection threshold on its set of standard relationships, whereas the converse is not true.

Figure 4.5b shows that MIC_e and TIC_e both have lower detection thresholds than

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

Detection threshold is sensitive to the relationship set in question, and different relationship sets may lead to different conclusions. For instance, at the parameter setting shown in Figure 4.5b, if the higher-frequency sinusoid is removed from the set of functions, S^{DDP} achieves a lower detection threshold than TIC_e. If the parameters for all methods are optimized for the new, smaller set of functions, the performance of TIC_e matches and sometimes exceeds that of S^{DDP} (Empirical Supplement 2B), but choosing parameters in this way may be difficult in practice. Therefore, for analyzing a data set where even the most interesting relationships are relatively simple, our results suggest that S^{DDP} may provide a good first-pass filter. However, for analyzing a data set in which the types of relationships present are unknown or diverse, as is our focus here, our results suggest that TIC_e is less likely than S^{DDP} to exclude relationships that might later be ranked as very interesting by MIC_e. We note parenthetically that for larger sample sizes, the increased runtime of S^{DDP} may present an additional challenge.

Using TIC_e for the first-pass filtering step has the advantage that computing MIC_e and TIC_e simultaneously is not more computationally expensive than computing just

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

When choosing parameters we recommend using the parameters for TIC_e that maximize power and the parameters for MIC_e that maximize equitability. These are the defaults in our software. For a discussion of alternative ways to choose parameters, see Appendix C.8.

4.7 Analysis of WHO data

To test the conclusions of our simulations on real data, we analyzed the aforementioned set of 356 social, medical, economic, and political indicators measured by the WHO in different countries. We chose to analyze this data set because previous analyses[121] have shown it to contain many linear relationships but also interesting non-linear relationships. These include, e.g., a relationship between obesity and income per person that consists of one trend among Pacific island nations, where female obesity is a sign of status[47], and a separate trend in the rest of the world. Here we analyzed the $49,286$

Statistic	# (%) rejections, FWER ≤ 0.05	# (%) rejections, FDR ≤ 0.05	% of top 1k rels. with $\lvert\rho\rvert < 0.85$	Avg. Jaccard to other statistics
$\text{MIC}_e/\text{TIC}_e$	17,630 (36%)	34,465 (70%)	29.9%	52.7%
dCor	17,783 (36%)	34,992 (71%)	1.7%	35.5%
MaxCor	4,324 (9%)	29,042 (59%)	24.0%	43.3%
HSIC	17,524 (36%)	35,052 (71%)	16.1%	47.2%
Kraskov	15,326 (31%)	30,477 (62%)	7.1%	36.2%
RDC	3,577 (7%)	23,086 (47%)	26.2%	45.9%
S^{DDP}	18,721 (38%)	35,582 (72%)	5.5%	34.7%
HHG	18,891 (38%)	36,338 (74%)	20.3%	48.7%
Sq. Pearson	17,073 (35%)	33,202 (67%)	0.0%	35.4%

Table 4.2: The performance of each of the statistics on the WHO data set. Jaccard indices were computed using the top thousand relationships ranked by each method. *[Higher Jaccard distance indicates less similarity.]*

potential pairwise relationships in this data set with $n \geq 50$ using the parameter settings determined by the simulations from Sections 4.2 and 4.3. (See Appendix C.6.4 for details.)

We first conducted a standard power analysis, asking how many non-trivial relationships the methods under consideration identified in this data set (Table 4.2). Strikingly, most methods identified over $15,000$ relationships as significant at level 0.05 after Bonferroni correction. When a false discovery rate of 0.05 was used instead, these methods discovered at least $30,000$ relationships. The combination of MIC_e and TIC_e proposed in the previous section detected $34,465$ relationships. In comparison, the most powerful method, HHG, detected $36,338$ relationships. The large number of relationships detected by most of the methods underscores the need for a principled way of exploring large data sets that is more fine-grained than testing for deviations from independence.

We next turned to assessing equitability. Equitability is difficult to analyze directly

here since we do not have a ground truth: we do not know which relationships in the data set are in our $\mathcal{Q}$ and which are not, and we cannot directly compute a population quantity of interest. However, we can still indirectly learn about equitability by checking for behaviors that we would expect an equitable statistic to exhibit.

For example, the equitability plots in Figure 4.1 show that most of the non-equitable statistics tend to give higher scores to linear and monotonic relationships. This leads to the hypothesis that in a data set that contains some complex relationships, a more equitable statistic will be better able to rank these complex relationships highly, rather than below a large number of linear relationships. And indeed, the fraction of the top $1,000$ relationships as ranked by $\text{MIC}_e/\text{TIC}_e$ with $|\rho| < 0.85$ was 29.9%, the most of any of the statistics tested. (See Table 4.2.) The two next-best-performing methods by this metric were RDC and maximal correlation, which achieved 26.2% and 24% respectively. This behavior is consistent with the non-trivial levels of equitability shown by maximal correlation in our simulations along with the theoretical parallels between RDC and maximal correlation (see below). Of the six methods besides $\text{MIC}_e/\text{TIC}_e$ that detected a very large number of relationships (rejection rate $\geq 30\%$ after Bonferroni correction), HHG was closest in performance, identifying 20.3% relationships that were not strongly linear, about two-thirds the amount identified by $\text{MIC}_e/\text{TIC}_e$.

The relationships ranked highly by $\text{MIC}_e/\text{TIC}_e$ contain results of potential scientific interest. These include relationships previously detected in Reshef et al.[121] such as the aforementioned relationship between income per person and obesity ($p \leq 6.0 \times$

10^{-7}), a highly non-linear relationship between number of physicians and deaths due to HIV/AIDS ($p \leq 6.0 \times 10^{-7}$), and others (see Table C.9). Our analysis here further identified several previously unreported relationships that would not easily be found using the other methods we assessed. For example, of the top 500 relationships as ranked by $\text{MIC}_e/\text{TIC}_e$, 33 were ranked 1,000th or worse by all eight of the other methods, including: a strongly non-linear relationship whereby adult male mortality rate is much higher among countries with per capita oil consumption below a certain threshold ($p \leq 6.0 \times 10^{-7}$, rank by $\text{MIC}_e/\text{TIC}_e$: 209, best rank by any other statistic: $1,510$); a non-linear but monotonic relationship between percent of the population below the poverty line and children per woman ($p \leq 6.0 \times 10^{-7}$, rank by $\text{MIC}_e/\text{TIC}_e$: 374, best rank by any other statistic: $1,227$); and a relationship between incidence of Ceasarian sections and government expenditure on health, in which there is a weak monotonic trend among most countries except for a small group of Northwestern European countries together with the United States that cluster away from the trend with a markedly higher expenditure on health ($p \leq 6.0 \times 10^{-7}$, rank by $\text{MIC}_e/\text{TIC}_e$: 464, best rank by any other statistic: $1,129$); For plots, see Figure C.9.We emphasize that our goal here is to establish that the relationships ranked highly by $\text{MIC}_e/\text{TIC}_e$ are of interest, but this does not preclude other methods finding interesting relationships that are not as highly ranked by MIC_eTIC_e; in general, we expect that most methods will rank some interesting relationships highly that are not as highly ranked by other methods.

The analyses above suggest that a) $\text{MIC}_e/\text{TIC}_e$ have a reduced preference for linear

relationships, thus making finding non-linear relationships easier, and b) more generally, $\text{MIC}_e/\text{TIC}_e$ give high ranks to potentially interesting relationships that would not be found using other statistics. This motivates us to ask systematically whether $\text{MIC}_e/\text{TIC}_e$ are more different from the rest of the methods tested than those methods are from each other in terms of highly ranked relationships. To examine this, we compared every pair of methods using the Jaccard distance between the top $1,000$ relationships identified by each method. (The Jaccard distance is a metric on sets defined by $J(A, B) = 1 - |A \cap B|/|A \cup B|$.) We found that the top-ranked relationships by $\text{MIC}_e/\text{TIC}_e$ were the most different from those of the other statistics in that they had the highest average Jaccard distance from the top-ranked relationships of the other statistics (52.7%; Table 4.2). These results were robust to the number of top relationships examined (see Empirical Supplement 5A). Consistent with our non-linearity analysis, HHG again came the closest in performance to $\text{MIC}_e/\text{TIC}_e$ among the statistics with extremely good power, with an average Jaccard distance of 48.7% to the rest of the statistics.

To gain a broader view of the behavior of the methods tested, we also created a dendrogram from these Jaccard distances using agglomerative hierarchical clustering. This recapitulated our findings, showing $\text{MIC}_e/\text{TIC}_e$ as the farthest away from any other single method. More generally, we believe it provides a valuable way to understand relationships between these measures of dependence. For instance, it shows distance correlation as similar to the squared Pearson correlation coefficient (in terms of rela-

tionship ranking, not power against independence), a fact that is consistent with our simulations. Additionally, the statistic closest to maximal correlation is RDC, which makes sense since RDC can be interpreted as an attempt to maximize correlation using linear combinations of random functions of the two variables in question. Finally, this dendrogram paired HSIC and HHG as similar, suggesting a hypothesis that there may be an as-yet uncharacterized aspect of dependence that these two statistics both capture.

We lastly plotted the scores of the five methods that detected the most relationships against each other for all the relationships in the data set. This is shown in Figure 4.7b; for all methods, see Empirical Supplement 5B.

4.8 Discussion

In this chapter, we presented an in-depth empirical evaluation of the equitability, power against independence, and runtime of several leading measures of dependence, including the two new statistics MIC_e and TIC_e introduced in Chapter 3. Our aims were to give an accessible exposition of equitability and its relationship to power against independence, provide the community with a comprehensive side-by-side comparison of existing methods, and evaluate the new statistics against the existing state of the art. Our main findings were as follows.

(1) Equitability. MIC_e, the estimator of the population MIC introduced in Chapter 3, generally has superior and more robust equitability with respect to R^2 than other measures of dependence. In some specific settings (models with no X noise and $n = 5,000$), mutual information estimation achieves superior equitability in our experiments, but its equitability is otherwise highly variable and often poor, particularly at lower sample sizes. Maximal correlation achieves some degree of equitability over the models examined, but all other statistics tested have very poor equitability.

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

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

power against independence than has been reported in previous analyses.

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

(4) Power/equitability tradeoff. The parameter α in the estimator MIC_e corresponds to a trade-off between power against independence and equitability that is consistent with the characterization of equitability given in Chapter 2. Lower values of α lead to higher power against a null of independence at the expense of power against null hypotheses representing weak relationship strength (i.e., equitability), while higher values of α lead to better equitability at the expense of power against independence. Other parameterized methods display a similar trade-off.

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

The fact that many measures of dependence performed similarly in our analysis of power against independence and had tens of thousands of rejections in our analyses of real data suggests that for some settings power against independence may not be where

the true challenge lies, and that we ought to demand more of measures of dependence in those settings. Equitability is one attempt to formulate a more ambitious goal, as is the concept of low detection threshold introduced in Chapter 2 of this thesis and discussed here, but there may well be other possibilities. Of course there are instances, such as detection of higher-dimensional relationships, in which even just power against independence is very difficult to achieve, and many of the methods evaluated here are quite useful in that setting.

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

As methodological work on measures of dependence continues, we expect and hope that methods with improved performance by each of the metrics assessed here will be developed, and already since the conclusion of this study there have been interesting and enlightening advances to note. For instance, an improved algorithm for estimating distance correlation is now known that runs much faster than the one benchmarked by

us[64]; the advances used in that algorithm could potentially be leveraged to improve other measures of dependence that rely on quantities computed between pairs of points. Similarly, a new measure of dependence called G^2 has recently been shown to achieve substantial levels of equitability[166]. This method, like MIC_e, is partition-based and uses a dynamic programming algorithm to optimize the choice of partition, providing further evidence of the utility of these concepts as we try to understand what about MIC_e is essential to its performance and what is ancillary.

While the results presented here make a compelling case for the use of MIC_e and TIC_e and provide insight into the trade-offs between different measures of dependence, there are some important limitations for both the new statistics and the comparisons we performed. First, we evaluated here only equitability with respect to R^2 on noisy functional relationships, whereas the definition we give of equitability explicitly acknowledges the possibility of using other properties of interest besides R^2 and standard relationships that are not noisy functional relationships. We feel that R^2 is an important measure of relationship strength that is intuitive and familiar to many practitioners, but equitability with respect to other properties of interest (see, e.g., Ding & Li[33]) merits study as well, and the methods tested here may perform much better or worse when their equitability is evaluated with respect to other properties of interest.

We observe that more general versions of equitability can be considered without abandoning the notion of R^2 on noisy functional relationships. For example, we could add only *noiseless* versions of non-functional relationships such as a circle to our existing

set of standard relationships, and then define the property of interest to equal 1 on those relationships. This has the virtue of encoding a strong intuition about the importance of non-functional relationships without requiring a stringent assumption about exactly how *noisy* non-functional relationships should be scored. Since the original motivation for the maximal information coefficient stems from its ability to detect non-functional relationships as well, assessing equitability with respect to a criterion such as this one is an interesting avenue of future inquiry.

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

An additional limitation of the present work is that, though an attempt at comprehensiveness was made, we did limit our scope to the set of noisy functional relationships in Reshef et al.[121] for equitability and the relationships introduced in Simon & Tibshirani[144] for power against independence, along with corresponding randomly chosen relationships in each setting. While we feel each of these suites of relationships provides reasonable insight into the performance of the methods in question on a broad set of realistic relationship types, there do for instance exist relationships, such as a step function, that when added to these suites provably result in poor equitability for all the

methods tested (see Appendix C.9), and we believe that the same is true for the power analyses. Characterizing those relationships theoretically and empirically in the settings of both equitability and power against independence is vital for fully understanding the strengths and weaknesses of each of these methods. This is an important direction for future work for which the analyses of random functions in Chapter 2 and here are only a first step. We note that as we try to understand what constitutes an appropriate set of standard relationships, it would be useful not just to better characterize performance of various sets, but also to have a way of evaluating to extent to which a given set of standard relationships "matches" a real data set that is being analyzed. Such a metric would provide valuable empirical guidance to this avenue of investigation.

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

Acknowledgments

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

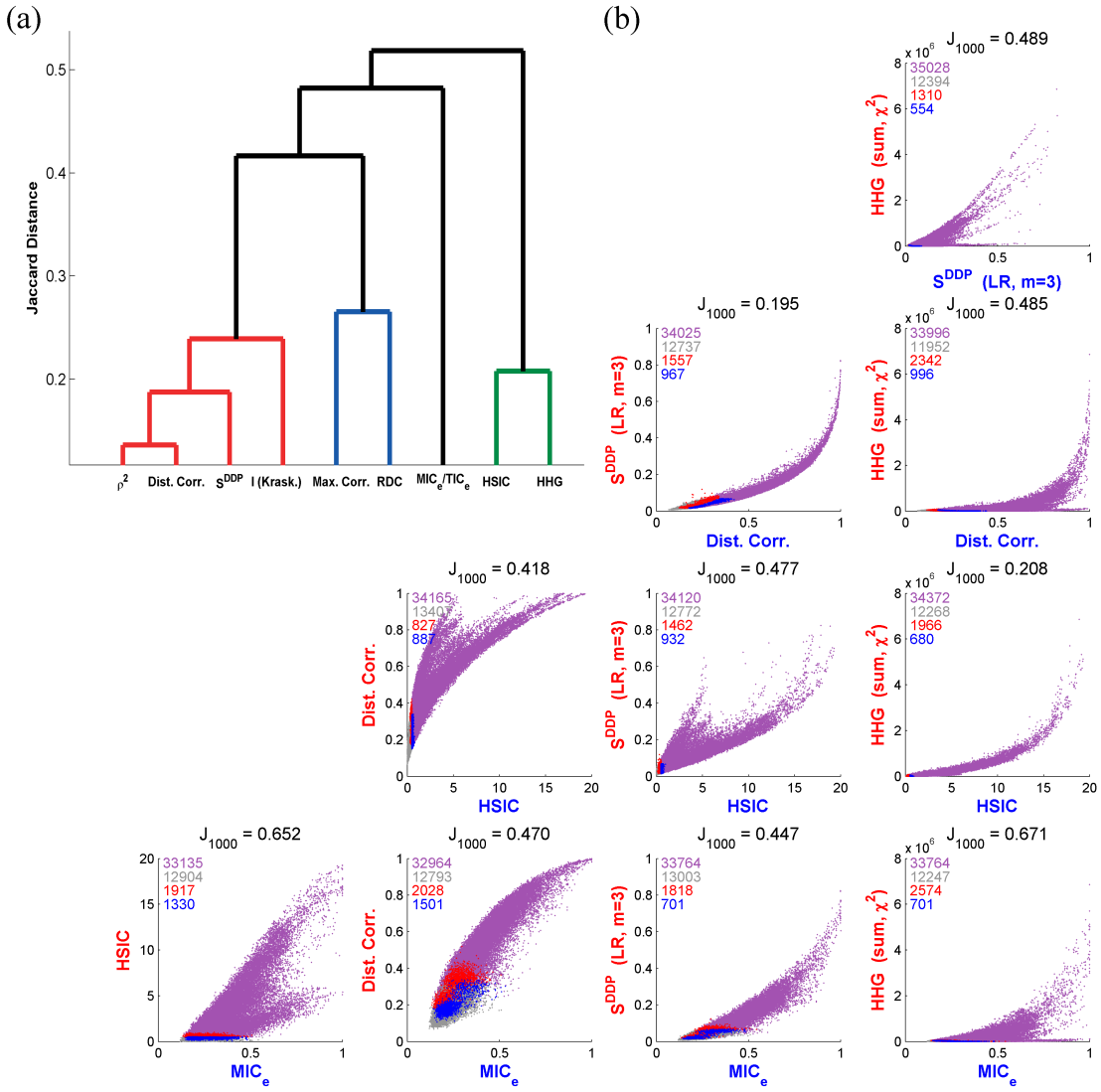


Figure 4.7: Pairwise comparison of the statistics on the WHO data set. (a) A dendrogram computed using the Jaccard distances between the top $1,000$ relationships identified by each statistic. (b) For each pair of statistics, a plot of one statistic's score against the other's across all the variable pairs in the data set. Purple, red/blue, and grey points denote relationships declared significant using both statistics, one statistic but not the other, and neither statistic, respectively. Numbers indicate number of dots of each respective color. J_{1000} indicates Jaccard distance between the top $1,000$ relationships ranked by the two statistics. For $\text{MIC}_e/\text{TIC}_e$, significance was determined using TIC_e and the plotted scores are the MIC_e scores; analogously, for mutual information estimation, significance was assessed using parameters optimized for independence testing and the plotted scores were computed using parameters optimized for equitability.

5

Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk

While detecting strong relationships is important, our ultimate interest is translating a detected relationship into a scientific insight. We therefore turn in this chapter to considering the detection of scientifically informative relationships.

What do we mean by "informative"? One operational definition might be that a relationship is informative to the extent that it is not susceptible to confounding. For example, a relationship between a drug treatment and disease outcome detected in a randomized controlled trial is considered by the medical establishment to be highly informative because it is immune to confounding. In contrast, an observational relationship between, say, years and education and longevity is considered uninformative because it can be confounded by any of several markers of socioeconomic status. There are intermediate levels of informativeness: for instance, the observation that melanoma

occurs more often on the left side of the body in the United States[111] has been used as evidence that sun exposure is a risk factor for melanoma. This relationship, though not the result of a genuine perturbational experiment, is considered relatively informative because the space of potential confounders – which would have to be correlated with both melanoma status and body side – seems quite small.

Because it requires thinking about potential confounders, determining how to detect informative relationships is necessarily a field-specific notion that requires domain expertise. In this chapter, we concretize and apply these ideas in the context of large-scale genetics efforts that have gained popularity in recent years. Specifically, we describe a criterion for informativeness of a certain kind for tying diseases to biological processes using genetic data. We then introduce a method for detecting associations that are informative by this criterion and validate the method in simulations. Finally, we use the method to analyze several different real data sets and discover instances of these associations that lead to specific biological hypotheses about disease mechanism.*

5.1 Background

Because of the biological nature of this chapter, we begin with a brief overview of necessary background before introducing the problem in more detail.

*The material in this chapter is adapted from a manuscript posted to bioRxiv as "Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk" by Yakir Reshef *et al.*[129] that as of this writing is under revision at *Nature Genetics*.

5.1.1 The genome-wide association study

One field in which observational relationships are considered highly informative is genetics: if genetic variation in a certain part of the genome, known as a *locus*, is associated with disease (after controlling for ancestry), it is highly likely that this association indicates a causal link between some variant at that locus and disease status. A common method for discovering such loci is the *genome-wide association study* (GWAS), in which many people with and without a disease are genotyped at many different genomic locations, and the measured variants are correlated one by one with disease status to discover disease-associated variants. Due to a combination of ethical, political, and logistical constraints, individual-level data arising from such studies are often not shared; rather, what is shared are summary statistics comprised of marginal correlations between each variant and disease status. Modern GWAS sample sizes are now in the tens to hundreds of thousands, and many reproducible associations have been discovered in this way.

Unfortunately, despite the existence of many disease-associated loci, the biological mechanism(s) driving most loci are difficult to work out. This is because each locus contains numerous genetic variants that are correlated with each other in the population, making it difficult to identify the causal one, and because causal variants usually fall outside of genes, complicating the process of identifying the gene through which a causal variant acts. For this reason, one popular paradigm for analyzing the results of GWAS

is *enrichment analysis*, a process whereby a researcher checks whether a certain subset of the genome corresponding to a biological process of interest contains excess genetic signal. For instance, one might check whether genetic variants lying in or near a set of genes known to be involved in inflammation tend to have larger-magnitude effects than variants in the rest of the genome.

GWAS enrichment analysis has been very useful for learning the broad contours of disease: for instance, GWAS enrichment analysis of genomic regions known to be active in different cell types has been very successful in determining which cell types are relevant for a given disease[160,40,38]. However, the informativeness of relationships arising from enrichment analysis for further dissection of disease mechanism is limited. This is because different biological processes often co-localize with or near each other, and so an enrichment can be confounded by co-localization with other, potentially unknown processes. For example, the aforementioned inflammation gene set might be enriched for genetic signal purely because inflammation genes tend to be expressed in immune cells, and genes expressed in immune cells are generically enriched for auto-immune disease variants. In this chapter, we propose that this confounding would be less likely if we observed not ony that variants that are important for inflammation are important for disease, but rather that variants that *increase* inflammation tend to *increase* disease risk on average across the genome.

5.1.2 Transcription factors

In general it can be difficult to obtain information on whether variants increase or decrease a biological process. For this reason, we focus in this chapter on one case in particular: transcription factor binding.

A *transcription factor* is a DNA-binding protein whose binding influences the expression levels of nearby genes. Transcription factors are biologically fundamental: by turning on and off large sets of genes at different times, they are the "software" that makes our the different cell types in our body, which all have the same "hardware", behave so differently from each other. There is therefore great interest in understanding the relationship of transcription factors to human disease.

The most important fact that we will use about transcription factors in this chapter is that whether or not a transcription factor binds to a stretch of DNA is determined, in part, by the content of the DNA sequence. Because of this, there has been a proliferation of methods that are able to predict from DNA sequence alone whether a transcription factor will bind to a genomic region of interest[72] and what the consequences of that binding will be[71]. This gives us a reasonable way to estimate the impact of genetic variants on the binding of a transcription factor: simply run both alleles of the variant, in their genomic context, through a predictor that can quantify the strength of binding of the transcription factor in the two situations, and compare the results. The crux of our work is develop a method that allows us to construct such "signed annotations" of

every common variant in the genome for many transcription factors and then to ask using GWAS data whether those annotations are correlated with the (unobserved) true causal effects of those variants on a phenotype of interest.

5.2 Introduction

Mechanistic interpretation of GWAS data sets has become a central challenge for efforts to learn about the biological underpinnings of disease. One successful paradigm for such efforts has been GWAS enrichment, in which a genome annotation containing SNPs that affect some biological process is shown to be enriched for GWAS signal[90,160,40,13,184]. However, there are instances in which experimental data allow us not only to identify SNPs that affect a biological process, but also to predict which SNP alleles promote the process and which SNP alleles hinder it, thereby enabling us to assess whether there is a systematic relationship between SNP alleles' direction of effect on the process and their direction of effect on a trait. Transcription factor (TF) binding, which plays a major role in human disease[69,89,116], represents an important case in which such signed functional annotations are available: because TFs have a tendency to bind to specific DNA sequences, it is possible to estimate whether the sequence change introduced by a SNP allele will increase or decrease binding of a TF[114,81,183,2,134,179,72].

Detecting genome-wide directional effects of TF binding on disease would constitute a significant advance in terms of both evidence for causality and understanding

of biological mechanism. Regarding causality, this is because directional effects are not confounded by simple co-localization in the genome (e.g., of TF binding sites with other regulatory elements), and thus provide stronger evidence for causality than is available using unsigned enrichment methods. Regarding biological mechanism, it is currently unknown whether disease-associated TFs affect only a few disease genes or whether transcriptional programs comprising many target genes are responsible for TF associations; a genome-wide directional effect implies the latter model (see Discussion).

Here we introduce a new method, signed LD profile (SLDP) regression, for quantifying the genome-wide directional effect of a signed functional annotation on polygenic disease risk, and apply it in conjunction with 382 annotations each reflecting predicted binding of a particular TF in a particular cell line. Our method requires only GWAS summary statistics[110], accounts for linkage disequilibrium and untyped causal SNPs, and is computationally efficient.

We show via theory and simulations that our method is well-powered and is well-calibrated even when TF binding sites co-localize with other enriched regulatory elements, which can confound unsigned enrichment methods[114,40]. We apply our method to 12 molecular traits and recover many known relationships including positive associations between gene expression and genome-wide binding of RNA polymerase II, NF-κB, and several ETS family members, as well as between known chromatin modifiers and their respective chromatin marks. Finally, we apply our method to 46 diseases and complex traits (average $N = 289,617$) and identify 77 significant associations at per-

trait FDR $< 5\%$, representing 12 independent signals. Our results include a positive association between educational attainment and genome-wide binding of BCL11A, consistent with recent work linking *BCL11A* hemizygosity to intellectual disability; a negative association between lupus risk and genome-wide binding of CTCF, which has been shown to suppress myeloid differentiation; and a positive association between Crohn's disease (CD) risk and genome-wide binding of IRF1, an immune regulator that lies inside a CD GWAS locus and has eQTLs that increase CD risk. Our method provides a new way to leverage functional data to draw inferences about causal mechanisms of disease.

5.3 Overview of methods

Our method for quantifying directional effects of signed functional annotations on disease risk, signed LD profile regression, relies on the fact that the signed marginal association of a SNP to disease includes signed contributions from all SNPs tagged by that SNP. Given a signed functional annotation with a directional linear effect on disease risk, the vector of marginal SNP effects on disease risk will therefore be proportional (in expectation) to a vector quantifying each SNP's aggregate tagging of the signed annotation, which we call the *signed LD profile* of the annotation. Thus, our method detects directional effects by assessing whether the vector of marginal SNP effects and the signed LD profile are systematically correlated genome-wide.

More precisely, under a polygenic model[174] in which true causal SNP effects are correlated with a signed functional annotation, we show that

$$E(\hat{\alpha}|v) = r_f \sqrt{h_g^2} Rv \tag{5.1}$$

where $\hat{\alpha}$ is the vector of marginal correlations between SNP alleles and a trait, v is the signed functional annotation (re-scaled to norm 1), R is the LD matrix, h_g^2 is the SNP-heritability of the trait, and r_f is the correlation between the vector v and the vector of true causal effects of each SNP, which we call the *functional correlation*. (The value of r_f^2 cannot exceed the proportion of SNP-heritability explained by SNPs with non-zero values of v.) Equation 5.1, together with an estimate of h_g^2, allows us to estimate r_f by regressing $\hat{\alpha}$ on the signed LD profile Rv of v. We assess statistical significance by randomly flipping the signs of entries of v, with consecutive SNPs being flipped together in large blocks (e.g., ~ 300 blocks total), to obtain a null distribution and corresponding P-values and false discovery rates (FDRs). To improve power, we use generalized least-squares regression, incorporating weights to account for the fact that SNPs in linkage disequilibrium (LD) provide redundant information due to their correlated values of $\hat{\alpha}$. We remove the major histocompatibility complex (MHC) region from all analyses due to its unusual LD patterns. We perform a multiple regression that includes a "signed background model" quantifying directional effects of minor alleles in five equally sized minor allele frequency (MAF) bins, which could reflect confounding due to genome-wide

negative selection or population stratification. We note that signed LD profile regression requires signed effect size estimates $\hat{\alpha}$ and quantifies directional effects, in contrast to stratified LD score regression[40], which analyzes unsigned χ^2 statistics and quantifies unsigned heritability enrichment. Details of the method are described in Section 5.8 and Appendix D; we have released open-source software implementing the method (see URLs).

We applied signed LD profile regression using a set of 382 signed annotations v, each quantifying the predicted effects of SNP alleles on binding of a particular TF in a particular cell line. We constructed the annotations by training a sequence-based neural network predictor of ChIP-seq peak calls, using the Basset software[72], on the results of 382 TF binding ChIP-seq experiments from ENCODE[157] and comparing the neural network's predictions for the major and minor allele of each SNP in the ChIP-seq peaks. The 382 experiments spanned 75 distinct TFs and 84 distinct cell lines. The resulting annotations were sparse, with only 0.2% of SNPs having nonzero entries on average (see Section 5.8 and Table D.1).

5.4 Simulations

We performed simulations with real genotypes, simulated phenotypes, and the 382 signed TF binding annotations to assess null calibration, robustness to confounding, and power. All simulations used well-imputed genome-wide genotypes from the GERA

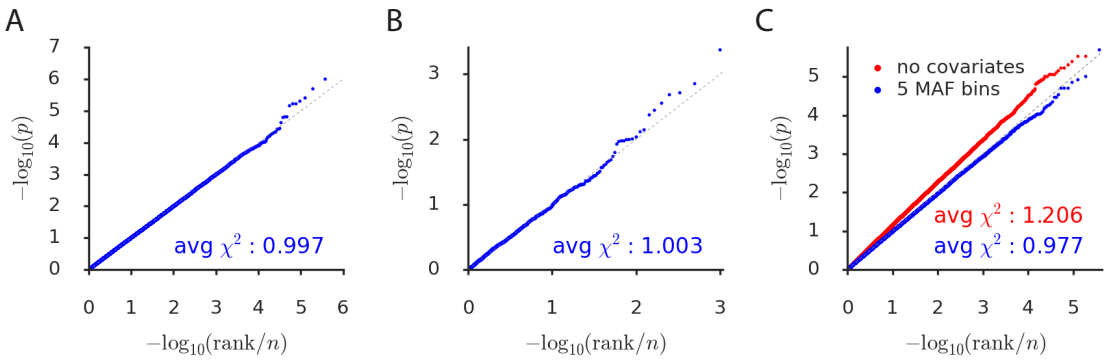


Figure 5.1: Simulations assessing null calibration. We report null calibration (q-q plots of $-\log_{10}$ P-values) in simulations of (a) no enrichment, (b) unsigned enrichment, and (c) directional effects of minor alleles. The q-q plots are based on (a) 382 annotations $\times 1,000$ simulations $= 382,000$, (b) $1,000$, and (c) two sets of $382 \times 1,000 = 382,000$ P-values. A 5-MAF-bin signed background model is included in all cases except for the red points in part (c), which are computed with no covariates. We also report the average χ^2 statistic corresponding to each set of P-values. Numerical results are reported in Table D.2.

cohort[6], corresponding to $M = 2.7$ million SNPs and $N = 47,360$ individuals of European ancestry. We simulated traits using normally distributed causal effect sizes (with annotation-dependent mean and variance in some cases), with $h_g^2 = 0.5$. Further details of the simulations are provided in Section 5.8.

We first performed null simulations involving a heritable trait with no unsigned enrichment or directional relationship to any of the 382 annotations. In 1,000 independent simulations, we applied signed LD profile regression to test each of the 382 annotations for a directional effect. The resulting P-values were well-calibrated (see Figure 5.1a and Table D.2). Analyses of the P-value distribution for each annotation in turn confirmed correct calibration for these annotations (see Figure D.1a).

We next performed null simulations involving a trait with unsigned enrichment but

no directional effects; these simulations were designed to mimic unsigned genomic confounding in which the binding sites of some TF lie in or near regulatory regions that are enriched for heritability for reasons other than binding of that TF. In 1,000 independent simulations, we randomly selected an annotation, simulated a trait in which the annotation had a 20x unsigned enrichment[40] (but no directional effect), and applied signed LD profile regression to test the annotation for a directional effect. We again observed well-calibrated P-values (see Figure 5.1b). It is notable that our method is well-calibrated even though it has no knowledge of the unsigned genomic confounder; this contrasts with unsigned enrichment approaches such as heritability partitioning, in which unsigned genomic confounders must be carefully accounted for and modeled[40].

We next performed null simulations to assess whether our method remains well-calibrated in the presence of confounding due to genome-wide directional effects of minor alleles on both disease risk and TF binding, which could arise due to genome-wide negative selection or population stratification. We simulated a trait for which 10% of heritability is explained by directional effects of minor alleles in the bottom fifth of the MAF spectrum (roughly MAF$<$ 5%). In 1,000 independent simulations, we applied signed LD profile regression to test each of the 382 annotations for a directional effect. P-values were well-calibrated for the default version of the method, which conditions on the 5-MAF-bin signed background model, but were not well-calibrated without conditioning on this model (see Figure 5.1c). (We note that this represents a best-case scenario in which the background model exactly matches the confounding being sim-

ulated, up to differences in MAF between the reference panel and the GWAS sample, and we caution that our method may not be appropriate for annotations with much stronger correlations to minor alleles than the annotations that we analyze here; see Figure D.1b.) The incorrect calibration that we observe when we do not include our signed background model could potentially be explained by genome-wide negative selection against decreased TF binding[4]. Indeed, most of our annotations show a small but highly significant bias of minor alleles toward decreasing TF binding (see Figure D.2) that is consistent with this explanation; however, it is also possible that this is a result of our procedure for constructing the annotations, and we do not explore it further in this work.

Finally, we performed causal simulations with true directional effects to assess the power and establish unbiasedness of signed LD profile regression. At default parameter settings, the method is well-powered to detect directional effects corresponding to a functional correlation of 2-6% (see Figure 5.2a and Table D.3), similar to values observed in analyses of real traits (see below). Notably, the power of the method is improved dramatically by our use of generalized least-squares to account for redundant information (see Figure 5.2a). Our method is also much more powerful than a naive method that regresses the vector of GWAS summary statistics on the annotation rather than its signed LD profile, an approach that does not model untyped causal SNPs in linkage disequilibrium with typed SNPs (see Figure D.3). The power of our method increases with sample size and SNP-heritability (see Figure D.4), and is only minimally

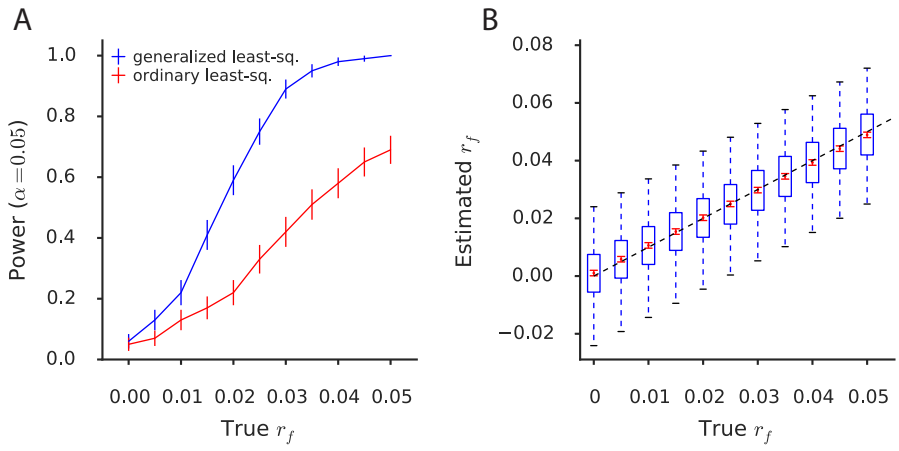


Figure 5.2: Simulations assessing power, bias, and variance. (a) Power curves under simulation scenarios comparing signed LD profile regression using generalized least-squares (i.e., weighting) to an ordinary (i.e., unweighted) regression of the summary statistics on the signed LD profile. Error bars indicate standard errors of power estimates. (b) Assessment of bias and variance of the signed LD profile regression estimate of r_f at realistic sample size $(47,360)$ and heritability (0.5), across a range of values of the true r_f. Blue box and whisker plots depict the sampling distribution of the statistic, while the red dots indicate the estimated sample mean and the red error bars indicate the standard error around this estimate. Numerical results are reported in Table D.3.

affected by within-Europe reference panel mismatch (see Figure D.5). In all instances, our method produced either unbiased or nearly unbiased estimates of functional correlation and related quantities (see Figure 5.2b and Figure D.6).

5.5 Analysis of molecular traits

TF binding is known to affect gene expression and other molecular traits[36]. We therefore applied signed LD profile regression to 12 molecular traits with an average sample size of $N = 149$. We first analyzed cis-eQTL data based on RNA-seq experiments in three blood cell types from the BLUEPRINT consortium[19] (see Section 5.8). For each

cell type, we collapsed eQTL summary statistics across 15,023-17,081 genes into a single vector of summary statistics for aggregate expression by summing, for each SNP, the marginal effect sizes of that SNP for the expression of all nearby genes (within 500kb). This is equivalent to analyzing one gene at a time and then performing a meta-analysis that accounts for linkage among nearby genes; it is also roughly equivalent to analyzing eQTL summary statistics for the sum of expression values of all genes, with each gene normalized to mean zero and variance one in the population (see Section 5.8 and Table D.4).

We tested each of the 382 TF binding annotations for a directional effect on aggregate expression in each of the three blood cell types. We detected a total of 92 significant associations at a per-trait FDR of 5% (see Figure 5.3a and Table D.5a; P-values from 5×10^{-6} to 1.0×10^{-2}). All 92 associations were positive, implying that greater binding of these TFs leads to greater aggregate expression and matching the known tendency of TF binding to promote rather than repress transcription for many TFs[36].

Many of the associations we detected recapitulate known aspects of transcriptional regulation. For example, associated TF binding annotations included RNA polymerase II in many cell lines, along with other members of the transcription pre-initiation complex (PIC) such as TATA-associated Factor 1 (TAF1) and TATA Binding Protein (TBP). There were also associations for TFs unrelated to the PIC but known to have activating activity, such as the ETS family members GABPA, ELF1, and PU.1[141], as well as the

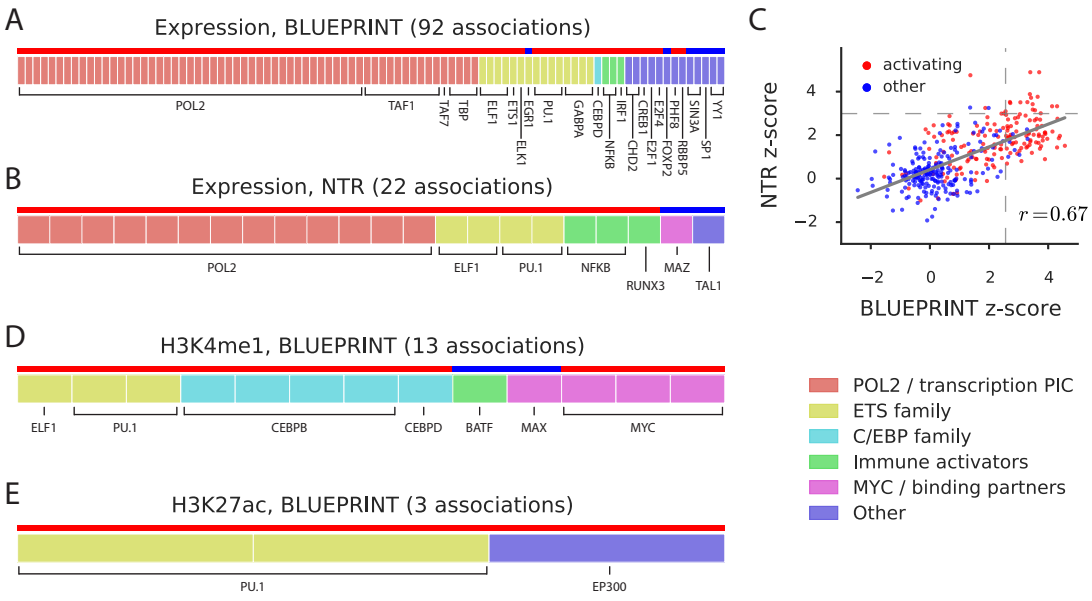


Figure 5.3: Analysis of molecular traits using signed LD profile regression. Each segmented bar in (a,b,d,e) represents the set of significant annotations at a per-trait FDR of 5% for the indicated traits, with each annotation corresponding to a particular TF profiled in a particular cell line. Results in (a,d,e) are aggregated across the 3 BLUEPRINT cell types. The stripe above each segmented bar is colored red for UniProt activating TFs (see main text) and blue for other TFs. (c) z-scores from the analyses of expression in the NTR data set and neutrophil expression in the BLUEPRINT data set, respectively, for each of the 382 annotations tested; red and blue again indicate UniProt activating TFs and other TFs, respectively. Dashed lines represent significance thresholds for 5% FDR. Numerical results are reported in Table D.5.

immune-related transcriptional activators IRF1 and NF-κB family member RELA[57,75]. Overall, the vast majority of the positive associations (85 out of 92) involved known "activating" TFs, defined as TFs with activating activity but not repressing activity in UniProt[158] (compared with 45% of all 382 annotations; $P = 3.1 \times 10^{-7}$ for difference using one-sided binomial test; see Figure 5.3a and Section 5.8). 52 of the 92 associations replicated (same direction of effect with nominal $P < 0.05$) in an independent set of

whole-blood eQTL summary statistics based on expression array experiments from the Netherlands Twin Registry (NTR)[168], including all of the examples mentioned above except IRF1 (see Figure 5.3b and Table D.5b). Across all 382 annotations analyzed, we observed a correlation of $r = 0.67$ between z-scores for signed annotation effects in the BLUEPRINT neutrophil and NTR data sets (see Figure 5.3c and Table D.5c).

We next conducted a similar analysis using histone QTL (H3K27me1 and H3K27ac) and methylation QTL from the BLUEPRINT data set. We detected 16 significant associations at a per-trait FDR of 5%, all of which were positive, including 13 for H3K27me1 QTL (see Figure 5.3d and Table D.5d; P-values from $\leq 10^{-6}$ to 4.3×10^{-4}), 3 for H3K27ac QTL (see Figure 5.3e and Table D.5e; P-values from 1.2×10^{-5} to 2.1×10^{-4}), and 0 for methylation QTL. Many of the detected associations recover known aspects of histone mark biology. For example, TFs associated to H3K4me1 included PU.1 and CEBPB, both of which act to increase H3K4me1 in blood cells and play strong roles in differentiation of those cell types[80,12,163,21], and binding of MYC, which has a known role as a chromatin modifier[88,3], including of H3K4 methylation[115]. We also observed an association between EP300 binding and H3K27ac, matching the fact that EP300 is a lysine acetyltransferase with a well-documented role in creation and maintenance of this mark[104]. Finally, while we did not find significant associations to methylation QTL at FDR$<$ 5%, we found 40 results at FDR$<$ 10%, almost all of which were negative associations between CTCF binding and methylation that are consistent with the literature on the negative relationship between CTCF binding and this epigenetic mark[113,165,91] (see

Table D.5f). In our analysis of the activating marks H3K4me1 and H3K27ac, signed LD profile regression again distinguished between activating and repressing TFs: of the 239 positive associations and 19 negative associations at a nominal significance threshold of $P < 0.05$ (chosen due to limited number of FDR$< 5\%$ associations) across all three cell types, 85% of the positive associations corresponded to activating TFs[158] (compared with 45% for all annotations; one-sided binomial $P = 7.4 \times 10^{-8}$ for difference); only 26% of the negative associations had this property (one-sided binomial $P = 7.0 \times 10^{-2}$ vs. 45%, $P = 4.6 \times 10^{-5}$ vs. 85%).

5.6 Analysis of 46 diseases and complex traits

We applied signed LD profile regression to 46 diseases and complex traits with an average sample size of 289,617, including 16 traits with publicly available summary statistics and 30 UK Biobank trants for which we have publicly released summary statistics computed using BOLT-LMM[85] (see URLs and Table D.6). We ran signed LD profile regression using each of our 382 TF annotations for each of these traits. We detected 77 significant associations at a per-trait FDR of 5%, spanning six diseases and complex traits (see Figure 5.4 and Table D.7a). (Following standard practice, we report per-trait FDR, but we estimated the global FDR of this procedure to be 9.4%, which is larger than the per-trait FDR of 5%; see Section 5.8). The 77 significant associations represent 12 independent signals after pruning correlated annotations (Table 5.1; see Section 5.8).

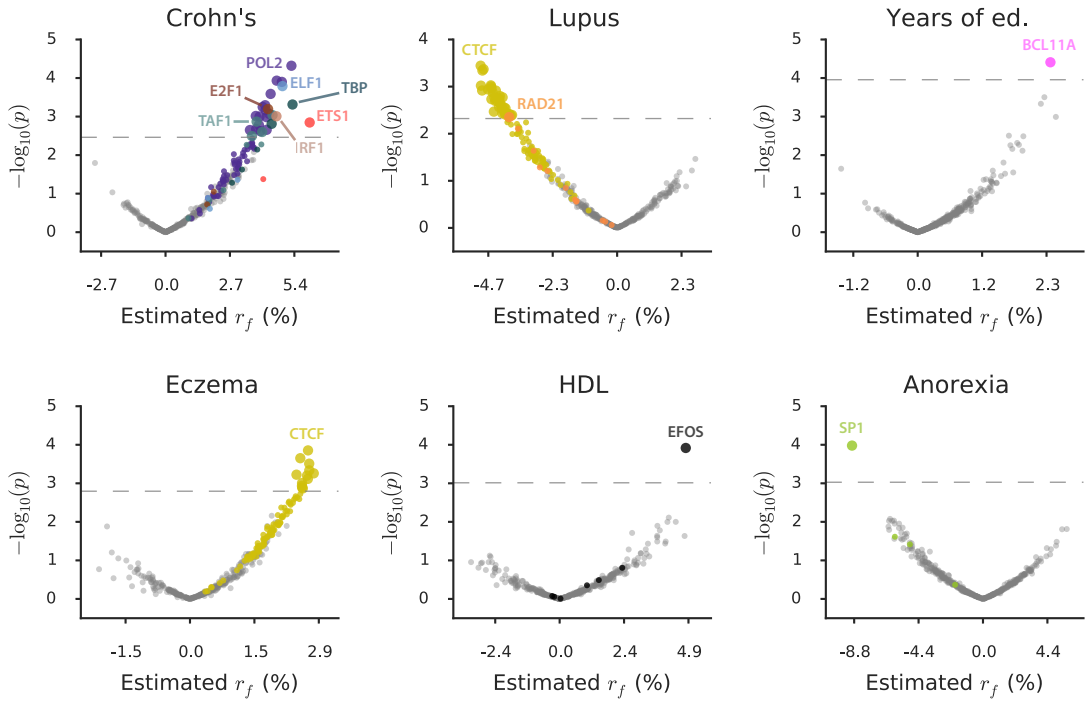


Figure 5.4: Analysis of diseases and complex traits using signed LD profile regression. For each disease or complex trait with at least one significant result, we plot $-\log_{10}(p)$ against estimated effect size for each of the 382 annotations analyzed. Points are colored by TF identity, with TFs with no significant associations for the trait colored in gray. Larger points denote significant results. The number of significant results for each trait is: Crohn's, 26; Lupus, 36; Years of education, 1; Eczema, 12; HDL, 1; Anorexia, 1. Numerical results are reported in Table D.7a.

To verify empirically that our results were not driven by directional effects of minor alleles, we re-analyzed our data using 382 annotations defined using the same set of SNPs with non-zero effects but with the directionality of effect determined by minor allele coding rather than predicted TF binding, for SNPs in the bottom quintile of the MAF spectrum. This analysis yielded only 4 significant associations at per-trait FDR$<5\%$. (Due to the small number of associations relative to the number of traits,

Trait	Top TF (num)	Top cell line	r_f	p	q
Years of ed.	BCL11A (1)	GM12878 (LCL)	2.4%	3.9×10^{-5}	1.5×10^{-2}
Crohn's	POL2* (20)	GM18951 (LCL)	5.3%	4.8×10^{-5}	1.5×10^{-2}
Anorexia	SP1 (1)	HEPG2 (hepatocyte)	-8.9%	1.1×10^{-4}	4.0×10^{-2}
HDL	FOS (1)	K562 (myeloid)	4.8%	1.2×10^{-4}	4.6×10^{-2}
Eczema	CTCF (12)	MCF7 (mammary)	2.7%	1.4×10^{-4}	3.4×10^{-2}
Crohn's	ELF1 (1)	GM12878 (LCL)	4.9%	1.6×10^{-4}	1.5×10^{-2}
Crohn's	POL2 (1)	U87 (glioblast)	4.4%	2.6×10^{-4}	1.5×10^{-2}
Lupus	CTCF** (36)	K562 (myeloid)	-5.0%	3.6×10^{-4}	4.4×10^{-2}
Crohn's	TBP (1)	HEPG2 (hepatocyte)	5.4%	4.9×10^{-4}	1.5×10^{-2}
Crohn's	E2F1 (1)	HELAS3 (cervical epithelium)	4.3%	6.4×10^{-4}	2.7×10^{-2}
Crohn's	IRF1 (1)	K562 (myeloid)	4.7%	9.8×10^{-4}	1.5×10^{-2}
Crohn's	ETS1 (1)	K562 (myeloid)	6.1%	1.4×10^{-3}	1.5×10^{-2}

Table 5.1: Independent associations from analysis of diseases and complex traits using signed LD profile regression. For each of 12 independent associations at per-trait FDR $< 5\%$ after pruning correlated annotations ($R^2 \geq 0.25$), we report the associated trait; the TF corresponding to the most significant annotation and the total number of correlated annotations that produced a significant result; the cell line corresponding to the most significant annotation; and the estimate of the functional correlation r_f, the P-value, and the per-trait q-value for the most significant annotation. Linked TFs also producing significant associations include (*) TAF1, TBP, and (**) RAD21.

this corresponds to a global FDR of 92.9% after accounting for 46 traits.) None of these 4 minor-allele associations overlapped with our set of 77 significant associations (see Section 5.8 and Table D.7b). We also examined, for each annotation, the estimated covariance between the GWAS summary statistics and the signed LD profile in each of 300 independent genomic blocks, finding agreement with the genome-wide direction of association in 59% of the blocks on average across our 12 independent associations, and in 85% of the blocks with estimated covariances of large magnitude (see Figure D.7).

Many of our results are supported by orthogonal genetic and non-genetic evidence and extend our understanding of the associated traits; we highlight three in particular. Our most significant result is a positive association between genome-wide binding of BCL11A in LCLs and years of education (see Figure 5.5a and Table D.8). This result aligns with

existing common and rare variant signals: *BCL11A* was one of the top genes identified in a GWAS of educational attainment[105] and *de novo* missense and loss-of-function mutations in *BCL11A* cause intellectual disability in a dosage-dependent manner[29,31]. (Additionally, our fine-mapping of the *BCL11A* GWAS locus[61] identified a putatively causal SNP in an intron of the *BCL11A* gene; see Table D.9.) *BCL11A* has also been shown to be the causal gene for a microdeletion syndrome characterized by cognitive impairment[8,44]. Recent experimental studies showing that heterozygous knock-out of *Bcl11a* in mice leads to microcephaly and cognitive impairment[31] have further confirmed the causal role of BCL11A in cognitive function, with directionality consistent with our result. This association thus represents a case in which our method provides stronger evidence for a causal association than previously available from common variant data, and establishes that BCL11A causes intellectual disability via a genome-wide mechanism involving binding throughout the genome—and presumably the modulation of a transcriptional program relevant to brain function or development—rather than regulation of one key disease gene (see Discussion).

We also detected a negative association between genome-wide binding of CCCTC-binding factor (CTCF) in the myeloid cell line K562 and risk of systemic lupus erythematosus (SLE) (see Figure 5.5b), accompanied by similar associations for CTCF and cohesin subunit RAD21 (a CTCF binding partner) in several other cell lines. This finding is consistent with several SLE risk loci at which either fine-mapped causal SNPs have been found to modify CTCF binding experimentally[182] and bioinformatically[118], or at

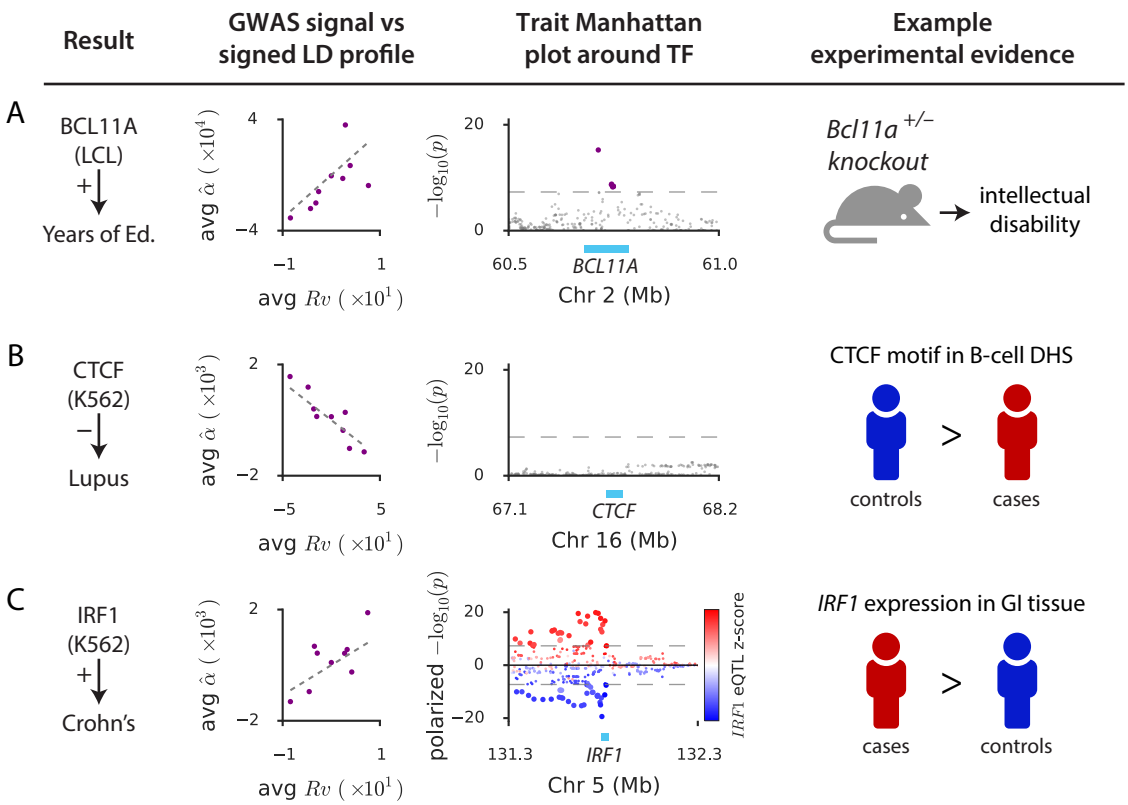


Figure 5.5: Genetic and non-genetic evidence for three TF binding-complex trait associations. For each of (a) BCL11A-years of education, (b) CTCF-lupus, (c) IRF1-Crohn's disease associations, we display plots of the marginal correlation $\hat{\alpha}$ of SNP to trait versus the signed LD profile Rv of the annotation in question, with SNPs collapsed into bins of $4,000$ SNPs and a larger bin around $Rv = 0$; Manhattan plots of the trait GWAS signal near the associated TF; and example experimental evidence from the literature. For Crohn's disease, the GWAS signal is polarized by direction of effect on disease and points are colored by direction and magnitude of association of each SNP to expression of *IRF1*. Additional experimental evidence relevant to each association is summarized in the main text. GI: gastrointestinal. Numerical results are reported in Table D.8.

which risk SNPs have been found to be in LD with SNPs modifying CTCF binding[155,138]. Additionally, CTCF has been shown experimentally to slow the rate of myeloid differentiation[159,106] and is involved in the regulation of 5-hydroxymethylcytosine (5-hmC), an epigenetic modification that is increased in promoters of immune-related genes in

CD4+ T cells of patients with SLE relative to controls[181]. Finally, CTCF motifs are overrepresented among DNA regions that are more accessible in B cells from healthy controls relative to B cells from SLE patients[137], consistent with the negative sign of the association arising from our study. We do not observe a GWAS signal for SLE at the *CTCF* locus. This could be because of the small sample size of the SLE GWAS, and/or because the *CTCF* gene is under strong selective constraint: its probability of loss-of-function intolerance (pLI) is estimated by the Exome Aggregation Consortium[82] to be the maximal value of 1.00, greater than 99.9% of genes. The association between CTCF binding and SLE therefore demonstrates the possibility of using signed LD profile regression to uncover aspects of disease mechanism that are difficult to directly observe in GWAS due to selective pressures on the underlying genes.

We also highlight a positive association between genome-wide binding of Interferon Regulatory Factor 1 (IRF1) in the myeloid cell line K562 and Crohn's disease (CD) (see Figure 5.5c). *IRF1* is located inside the *IBD5* locus, a 250kb region associated with CD and inflammatory bowel disease in multiple GWAS[42,67]; haplotypes containing *IRF1* variants have been shown to be more strongly correlated with CD risk than haplotypes containing variants in nearby genes[63]; CD risk SNPs have been shown to co-localize with *IRF1* alternative splicing QTLs[19]; and *IRF1* is more highly expressed in CD gastrointestinal tissue biopsies relative to control tissue[63]. In a recent large-scale fine-mapping study[62], the causal signal at the *IBD5* locus was narrowed down to a set of 8 SNPs spanning 35kb and lying 15kb away from *IRF1*. However, despite this

resolution, it remains unclear from the locus alone what the causal mechanism is: the study suggested that rs2188962, which received 0.59 of the posterior probability of being causal, could function via an eQTL effect on *SLC22A5* in immune and gut epithelial cells, but rs2188962 is also an eQTL for *IRF1* in blood[168], and we determined that the TWAS approach[56] assigns highly significant scores to both genes ($p \leq 4.0 \times 10^{-14}$ for *IRF1* and 3.17×10^{-18} for *SLC22A5*). In this context, our result therefore provides genome-wide evidence for a genuine causal link between *IRF1* and CD that, unlike the single-locus approaches, is not susceptible to pleiotropy and allelic heterogeneity near the *IRF1* gene (see Discussion). We note that the direction of effect inferred by our method agrees with the positive sign of the TWAS association between *IRF1* and CD, as expected in the case of a causal relationship.

We provide additional discussion of our other results in Appendix D.

5.7 Discussion

We have introduced a method, signed LD profile regression, for identifying genome-wide directional effects of signed functional annotations on diseases and complex traits. We applied this method, in conjunction with 382 annotations describing predicted effects of SNPs on TF binding, to 12 molecular traits (average $N = 149$) and 46 diseases and complex traits (average $N = 289,617$). In our analysis of molecular traits, our method recovered classical aspects of transcriptional regulation, including the pro-transcriptional

effect of RNA polymerase and activating TFs such as NFκB, as well as relationships between several chromatin modifiers and their respective chromatin marks; to our knowledge, these relationships have not previously been demonstrated using eQTL data. Our analysis of complex traits yielded 77 TF-trait associations, corresponding to 12 independent associations. Some of our results, such as the positive association between IRF1 binding and Crohn's disease, provide strong causal hypotheses to explain long-standing GWAS associations; others, such as the positive association between BCL11A binding and educational attainment, provide mechanistic interpretation for a top GWAS locus for which orthogonal genetic evidence, such as rare variant and knock-out studies, already existed; and still others, such as the negative relationship between CTCF binding and SLE, have experimental support but had not previously been observed from GWAS data, possibly due to strong evolutionary constraint on some TFs. We note that although we constructed our predicted TF binding annotations using the neural-network predictor Basset[72], there exist many other effective methods for making such signed predictions[114,81,183,2,179,178].

Our method differs from unsigned GWAS enrichment methods by assessing whether there is a systematic genome-wide correlation between a signed functional annotation and the (signed) true causal effects of SNPs on disease, rather than assessing whether a set of SNPs have large effects on a disease without regard to the directions of those effects. A substantial advantage of this approach is reduced susceptibility to confounding: for example, an unsigned GWAS enrichment for binding of an immune TF could indicate

a causal role for that TF in the associated disease, or could instead be a side effect of a generic enrichment among cell-type specific regulatory elements in immune cells[40]. In contrast, if alleles that increase binding of the TF tend to increase disease risk and alleles that decrease binding of the TF tend to decrease disease risk, the set of potential confounders is smaller because a confounding process has not only to co-localize in the genome with binding of the TF but also to have the property that alleles that increase the process have a consistent directional effect on binding of the TF.

When applied to TF binding, our method enables stronger statements about causality and mechanism than were previously possible with genome-wide methods. Regarding causality, this is because a consistent directional effect throughout the genome of SNPs predicted to affect binding due to sequence change supports stronger causal statements than i) single-locus methods, which are susceptible to pleiotropy and allelic heterogeneity[56], ii) unsigned heritability enrichment methods, which can be confounded by co-localization in the genome of TF binding sites with other enriched regulatory elements as described above[40], and iii) genetic correlation and Mendelian randomization (MR), which can be confounded by reverse causality and pleiotropic effects[15,28,164] and which scale poorly because they require TF ChIP-seq in many individuals for every TF/cell-type pair studied. The reason that our method is not confounded by reverse causality is that each of our annotations is produced in a cell population that is isogenic and therefore does not have variance in genetic liability for any trait. In other words, our annotations provide ideal instrumental variables for the effect of TF binding on

the trait of interest because they are created not by naively correlating SNPs with TF binding but rather by examining the effect of each SNP on local DNA sequence.

Regarding mechanism, our method sheds light on the question of whether TFs affect traits via coordinated regulation of gene expression throughout the genome[93] (a "genome-wide" model) or via regulation of one or a small number of key disease genes[30] (a "local" model). Since the associations we find involve a consistent net direction of effect of TF binding on a trait throughout the genome, they cannot be explained by a local model and therefore represent evidence for the existence of transcriptional programs and their relevance to complex traits. This is of basic interest, but it also has therapeutic relevance: if a TF causally affects a trait but the TF is not druggable due to its nuclear localization or large DNA- and protein-binding domains[41,78], then the local model suggests targeting a downstream gene, whereas the genome-wide model instead suggests targeting an upstream regulator since the causal link between TF and trait is mediated through a large number of downstream genes. (We emphasize that a significant result for our method does not imply that all binding events of the TF in question affect disease via activation of a single transcriptional program; rather, it implies that there exists a program that is widespread enough that we observe its effect on disease in a large number of locations in the genome; see Figure D.7.)

Our method could be used to link disease to biological processes beyond TF binding. For example, sequence-based models can also produce signed predictions of DNase I hypersensitivity[81,183,72], histone modifications[183,72], splicing[169,2], and transcription ini-

tiation[70]. Additionally, massively parallel assays and CRISPR screens are increasingly yielding high-resolution experimental information about the effects of genetic variation on gene expression[156,36,34,43] as well as cellular processes such as growth[24,161,119] and inflammation[109]. Finally, perturbational differential expression experiments can yield signed predictions for the relationships of genes to a variety of biological processes such as drug response[151], immune stimuli[48], and many others[83]. Though converting such data to signed functional annotations will require care, doing so could allow us to leverage them to make detailed statements about disease mechanism.

We note several limitations of signed LD profile regression. First, though our results are less susceptible to confounding due to their signed nature, they are not immune to it: in particular, our method cannot distinguish between two TFs that are close binding partners and thus share sequence motifs. Second, although we have shown our method to be robust in a wide range of scenarios, we cannot rule out the possibility of un-modeled directional effects of minor alleles on both trait and TF binding as a confounder; however, our empirical analysis of real traits with minor-allele-based signed annotations suggests that directional effects of minor alleles are very unlikely to explain our results (see Table D.7b). Third, our method is not well-powered to detect instances in which a TF affects trait in different directions via multiple heterogeneous programs. Fourth, the effect sizes of the associations to diseases and complex traits that we report are small in terms of the estimated values of r_f, which range from 2.4% to 8.9% (see Table D.7a), although signals of this size for predicted TF binding could be indicative of

much stronger relationships, e.g., with true TF binding, TF expression, TF phosphorylation, or TF binding in specific subsets of the genome. We further note that the magnitude of the signals that we detect is commensurate with the very small number of SNPs in our annotations, together with the fact that r_f^2 is bounded by the proportion of SNP-heritability explained by those SNPs (see Table D.7c). Fifth, though we detected many significant associations overall, there were many traits, such as schizophrenia, height, and blood cell traits, for which we did not detect any significant associations using our TF annotations. We believe that this limitation is partially due to the set of TF ChIP-seq annotations available through the ENCODE project, and in particular to the bias of those experiments toward core regulatory proteins such as RNA polymerase II and CTCF as well as their use of cell lines rather than primary tissue samples; we expect this to become clearer as more diverse functional data sets become available.

Despite these limitations, signed LD profile regression is a powerful new way to leverage functional genomics data to draw causal and mechanistic conclusions from GWAS about both diseases and underlying cellular processes.

5.8 Methodological details

5.8.1 Signed LD profile regression

Model and estimands

Let M be the number of SNPs in the genome. We assume a linear model:

$$y|\beta, x \sim \mathcal{N}(x^T\beta, \sigma_e^2) \tag{5.2}$$

where $x \in \mathbb{R}^M$ and $y \in \mathbb{R}$ are the standardized genotype vector and phenotype, respectively, of a randomly chosen individual from some population, $\beta \in \mathbb{R}^M$ is a vector of true causal effects of each SNP on phenotype, and σ_e^2 represents environmental noise. Given a signed functional annotation $v \in \mathbb{R}^M$, we then model

$$\beta|v \sim [\mu v, \sigma^2 I] \tag{5.3}$$

where the scalar μ represents the genome-wide directional effect of v on β, σ^2 represents other sources of heritability unrelated to v, and the notation $[\cdot, \cdot]$ is used to specify the mean and covariance of the distribution without specifying any higher moments.

Though we can estimate μ, its value depends on the units of the annotation and the heritability of the trait. Because of this, we focus instead on the *functional correlation* r_f,

which re-scales μ to be dimensionless and is defined as

$$r_f := \text{corr}(x^T\beta, x^T v) = \mu\sqrt{\frac{v^T R v}{h_g^2}} \tag{5.4}$$

where $h_g^2 = \text{var}(x^T\beta)$ is the SNP-heritability of the phenotype and $R = E(xx^T) \in \mathbb{R}^{M\times M}$ is the (signed) population LD matrix of the genotypes. (Note that r_f can also be defined as a correlation between β and v; this definition is approximately equivalent in expectation under our random effects model, provided $v^T R v \approx |v|^2$.) We additionally estimate $h_v^2 = r_f^2 h_g^2$, the total phenotypic variance explained by the signed contribution of v to β, as well as $h_v^2/h_g^2 = r_f^2$. For annotations with small support, these quantities are expected to be small in magnitude. To see this, notice that h_v^2 cannot exceed the total (unsigned) phenotypic variance explained by SNPs with non-zero values of v. It follows that r_f^2 cannot exceed the proportion of (unsigned) SNP-heritability explained by SNPs with non-zero values of v. For more detail on the model and estimands, see Appendix D.

Main derivation

Let $X \in \mathbb{R}^{N\times M}$ be the genotype matrix in a GWAS of N individuals, with standardized columns, and let $Y \in \mathbb{R}^N$ be the phenotype vector. In Appendix D, we show that under

the above model the following identity approximately holds:

$$\hat{\alpha}|v \sim \left[\mu Rv, \sigma^2 R^2 + \frac{R}{N}\right] \tag{5.5}$$

where $\hat{\alpha} := X^T Y/N$ is a vector whose m-th entry contains the marginal correlation of SNP m to the phenotype and $R \in \mathbb{R}^{M \times M}$ is the population LD matrix. Equation 5.1 from the main text can be derived from Equation 5.4 by re-scaling v so that $v^T Rv = 1$, then substituting for μ.

We call Rv the *signed LD profile* of v. Equation 5.5, together with central limit theorem considerations, implies that it is nearly optimal to estimate μ by regressing $\hat{\alpha}$ on the signed LD profile using generalized least-squares with $\Omega := \sigma^2 R^2 + R/N$ as the inverse weight matrix. It can be shown that if a) all causal SNPs are typed, b) sample size is infinite, and c) R is invertible, this method is equivalent to estimating β via $R^{-1}\hat{\alpha}$ and then regressing this estimate on v to obtain μ, which is the optimal approach in that setting. Note that because we generate P-values for hypothesis testing empirically (see below), we are guaranteed that our generalized least-squares scheme will remain well-calibrated even if our estimate of the matrix Ω is inaccurate due to, e.g., mis-match between the reference panel and the study population. Once we have estimated μ, we re-scale this estimate to yield an estimate of r_f and other estimands of interest. For more detail on derivations and computational considerations, see Appendix D.

Null hypothesis testing

To test the null hypothesis $H_0 : \mu = 0$ (or, equivalently, $H_0 : r_f = 0$), we split the genome into approximately 300 blocks of approximately the same size with the block boundaries constrained to fall on estimated recombination hotspots[10]. We then define the null distribution of our statistic as the distribution arising from independently multiplying v by an independent random sign for each block. We perform this empirical sign-flipping many times to obtain an approximation of the null distribution and corresponding P-values. Our use of sign-flipping ensures that any true positives found by our method are the result of genuine first-moment effects; if in contrast we estimated standard errors using least-squares theory or a re-sampling method such as the jackknife or bootstrap, our method might inappropriately reject the null hypothesis only because the variance of β is higher in parts of the genome where Rv is large in magnitude. This would make our method susceptible to confounding due to unsigned enrichments, as might arise from the co-localization of TF binding sites with enriched regulatory elements such as enhancer regions. Additionally, the fact that we flip the signs of SNPs in each block together ensures that our null distribution preserves any potential relationship of our annotation to the LD structure of the genome. In choosing how many blocks to use for this procedure, we took into account that i) the fewer blocks we use the fewer assumptions we make about LD structure and the faster we can compute P-values, and ii) the more blocks we use the higher the precision of the P-values that we can obtain.

Our choice to use 300 blocks is a compromise between these two considerations.

Controlling for covariates and the signed background model

Given a signed covariate $u \in \mathbb{R}^M$, we can perform inference on the signed effect of v conditional on u by first regressing Ru out of $\hat{\alpha}$ and out of Rv using the generalized least-squares method outlined above, and then proceeding as usual with the residuals of $\hat{\alpha}$ and Rv. This can be done simultaneously for multiple covariates u.

Unless stated otherwise, all analyses in this paper are done controlling in this fashion for a "signed background model" consisting of 5 annotations $u^1, \ldots, u^5$, defined by

$$u_m^i = \mathbf{1}\left\{\text{MAF}_m \text{ is in } i\text{-th quintile}\right\} \sqrt{2\text{MAF}_m(1-\text{MAF}_m)^{1+\alpha_s}} \tag{5.6}$$

where MAF_m is the minor allele frequency of SNP m and α_s is a parameter describing the MAF-dependence of the signed effect of minor alleles on phenotype. Based on the literature on MAF-dependence of the unsigned effects $\text{var}(\beta_m)$, we set $\alpha_s = -0.3$[139].

5.8.2 382 TF annotations

We downloaded every ChIP-seq and DNase I hypersensitivity experiment in ENCODE and trained the sequence-based predictor of peak presence/absence, Basset[72], to jointly predict each downloaded track on a set of held-out genomic segments. (We included tracks other than TF binding tracks because training predictions using all tracks slightly

improved prediction accuracy for the TF binding tracks.) After training the joint predictor, we retained the predictions for every TF binding track for which a) the set of ChIP-seq peaks spanned at least 5,000 SNPs in our 1000G reference panel, and b) Basset's estimated area under the precision-recall curve was at least 0.3. This yielded a set of 382 TF ChIP-seq experiments. For each experiment, we constructed an annotation via

$$v_m = \mathbf{1}\{m \in C\}(P_m^a - P_m^A) \tag{5.7}$$

where C is the set of SNPs in the ChIP-seq peaks arising from the experiment, P_m^a is the Basset prediction for the 1,000 base-pair sequence around SNP m when the minor allele is placed at SNP m, and P_m^A is the Basset prediction for the 1,000 base-pair sequence around SNP m when the major allele is placed at SNP m. (We always used the minor allele as the reference allele in both our TF binding annotations and our GWAS summary statistics.)

5.8.3 Simulations

All simulations were carried out using real genotypes from the GERA cohort[6] ($N = 47,360$). The set of $M = 2.7$ million causal SNPs was defined as the set of very well imputed SNPs (INFO ≥ 0.97) that had very low missingness ($< 0.5\%$), non-negligible MAF (MAF $\geq 0.1\%$) in the GERA data set, and were represented in our 1000G Phase 3 European reference panel[1,16].

Null simulations

For the simulations in Figure 5.1a, we simulated 1,000 independent null phenotypes with the architecture $\beta_m \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = h_g^2/M$ and $h_g^2 = 0.5$. For each phenotype, we computed GWAS summary statistics using plink2[18] (see URLs), adjusting for 3 principal components as well as GERA chip type as covariates. For each of our 382 TF annotations, we then ran signed LD profile regression on each of these 1,000 phenotypes, yielding a set of 382,000 P-values. For the simulations in Figure 5.1b, we simulated 1,000 independent traits in which each trait had an unsigned enrichment for a randomly chosen annotation: after choosing an annotation v, we set $\beta_m \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 + \tau^2 \mathbf{1}\{v_m \neq 0\})$ where σ^2 and τ^2 were set to achieve $h_g^2 = 0.5$ and a 20x unsigned enrichment for the SNPs with non-zero values of v. We then computed summary statistics as above and ran signed LD profile regression to assess v for a genome-wide directional effect. This procedure yielded 1,000 P-values. For the simulations in Figure 5.1c, we simulated 1,000 independent phenotypes with a directional effect of minor alleles: we set $\beta_m \overset{iid}{\sim} \mathcal{N}(\mu u_m^1, \sigma^2)$ where u_m^1 is non-zero if SNP m is in the bottom quintile of the MAF spectrum of the GERA sample and 0 otherwise, as in the signed background model. We set μ such that 10% of heritability would be explained by this directional effect, and then set σ^2 to achieve $h_g^2 = 0.5$. We then computed summary statistics as above and ran signed LD profile regression to assess for a directional effect of each of our 382 annotations on each of the 1,000 phenotypes, yielding a set of 382,000 P-values. Finally, we repeated

the same computation but running signed LD profile regression without the 5-MAF-bin signed background model to obtain an additional set of 382,000 P-values.

Causal simulations

For the simulations in Figure 5.2, we fixed a representative annotation v (binding of IRF4 in GM12878), and simulated traits using $\beta_m \overset{iid}{\sim} \mathcal{N}(\mu v_m, \sigma^2)$, with μ set to achieve $r_f = \{0, 0.005, 0.01, \ldots, 0.05\}$ and σ^2 set to achieve $h_g^2 = 0.5$ in each case. For each value of r_f, we simulated 100 independent traits, computed summary statistics using plink2, and then ran each of the methods under consideration using the annotation v.

5.8.4 Analysis of molecular traits

We downloaded BLUEPRINT consortium QTL data for gene expression, H3K4me1, H3K27ac, and methylation in three different blood cell types with sample sizes of $N = 158, 165,$ and 125 for monocytes, neutrophils, and T cells, respectively[19] (see Table D.5 and URLs). For each of the 3 gene expression traits, we constructed one summary statistics vector $\hat{\alpha}$ by setting

$$\hat{\alpha}_m = \sum_{k \in G_m} \hat{\alpha}_m^{(k)} \tag{5.8}$$

where G_m is the set of all genes within 500kb of SNP m, and $\hat{\alpha}_m^{(k)}$ is the marginal correlation of SNP m to the expression of gene k. Assuming a) infinite sample size, and

b) zero correlation between every SNP and any gene not cis to that SNP, this procedure is equivalent up to a scalar to performing a GWAS of total relative expression. To see this, let $y^{(k)}$ denote expression of gene k after standardization to mean zero and unit variance in the population, and let γ be the GWAS summary statistics arising from a GWAS of total relative expression $\sum_k y^{(k)}$ at infinite sample size. By linearity, we have

$$\gamma_m \propto \sum_k \alpha_m^{(k)} \tag{5.9}$$

$$= \sum_{k \in G_m} \alpha_m^{(k)} + \sum_{k \notin G_m} \alpha_m^{(k)} \tag{5.10}$$

$$= \sum_{k \in G_m} \alpha_m^{(k)} \tag{5.11}$$

$$= \alpha_m \tag{5.12}$$

where α_m^k denotes the large-sample limit of $\hat{\alpha}_m^{(k)}$ and α denotes the large-sample limit of $\hat{\alpha}$ defined in Equation 5.8.

Applying the same procedure to the two histone marks and to methylation in addition to gene expression yielded a total of 12 sets of summary statistics (see Table D.4). We ran signed LD profile regression using each of our 382 TF annotations for each of these 12 traits. We obtained results at FDR$< 5\%$ using the Benjamini-Hochberg procedure[9] within each of the 12 traits (see discussion of Benjamini-Hochberg versus other alternatives below), and reported the union of significant results across cell types for each trait.

For our replication analysis, we used expression array-based whole blood eQTL data from the NTR[168], which we obtained by downloading the set of TWAS weights[56] computed for that data set (see URLs). We then proceeded as above.

Enrichment analysis for activating TFs

For each TF represented in our annotations, we queried the UniProt database[158] to establish whether the TF was annotated as having activating activity and, separately, whether it was annotated as having repressing activity. We then defined as "activating" any TF with the former but not the latter. To estimate whether the set of significant positive signed LD profile associations with gene expression were enriched for activating TFs compared to the set of annotations as a whole, we conducted a one-sided binomial test. To account for the correlated nature of our annotations, we assumed independence only among distinct TFs but not among distinct cell lines for the same TF. We used the same scheme to test for enrichment and depletion of activating TFs among the positive and negative associations detected by signed LD profile regression in our analysis of histone marks.

5.8.5 Analysis of 46 diseases and complex traits

We applied signed LD profile regression to 46 diseases and complex traits with an average sample size of 289,617, including 16 traits with publicly available summary statistics and 30 UK Biobank trants for which we have publicly released summary statistics computed

using BOLT-LMM[85] (see URLs and Table D.6). We ran signed LD profile regression using each of our 382 TF annotations for each of these traits. We obtained results at per-trait FDR$< 5\%$ using the Benjamini-Hochberg procedure[9]. We chose to use the Benjamini-Hochberg procedure rather than more sophisticated procedures such as the Storey-Tibshirani procedure[150] because the latter procedure, while more powerful, is more difficult to analyze in a multi-trait setting (see below) and controls FDR more noisily when applied in situations with only hundreds (rather than thousands) of tests.

5.8.6 Estimation of global FDR for complex trait analysis

When many traits are analyzed, per-trait FDR control does not imply global FDR control. This is because in the case of a completely null trait, the guarantee of FDR control does not imply that there will never be any rejections but rather only that there will be a non-zero number of rejections at most 5% of the time. Therefore, if enough null traits are analyzed the set of results may be contaminated by these spurious findings. In the case of independent tests (i.e., uncorrelated annotations) with FDR controlled by the Benjamini-Hochberg procedure, this can be taken into account[176] and the global FDR can be approximated using the formula

$$q = \frac{q_\ell(D+T)}{D+1} \tag{5.13}$$

where q is the estimated global FDR, q_ℓ is the per-trait FDR, D is the observed total number of discoveries at per-trait FDR q_ℓ, and T is the number of traits. This correction is based on the intuition that for a null trait with independent tests, the Benjamini-Hochberg procedure behaves very similarly to a Bonferroni correction, and so the expected number of rejections per null trait is approximately q_ℓ, and the expected number of rejections for T null traits would be approximately $q_\ell T$.

Applying this correction to our results yields a global FDR estimate of 7.9%. However, since our annotations are dependent, this estimate can be anti-conservative. To see this, imagine a null trait with 100 perfectly correlated tests. The Benjamini-Hochberg procedure will give more than zero rejections only 5% of the time, but whenever it rejects it will yield 100 rejections rather than 1. Therefore, the expected number of rejections is not 0.05 but rather 5. We heuristically corrected for this using the intuition that under dependent tests, the expected number of false discoveries in a null stratum is not q_ℓ but rather q_ℓ times the number of tests conducted per single "independent" test. We estimated the number of independent tests as in the GWAS literature, by simulating 1,000 independent null traits with a heritability of 0.5, testing each trait against our 382 annotations, and asking for what S we see at least one p-value $\leq 0.05/S$ in approximately 5% of the 1,000 null traits. This procedure gave us $S = 250$. We then estimated the global FDR using the equation

$$q = \frac{q_\ell(D + 382T/S)}{D + 1}. \tag{5.14}$$

This yielded the reported global FDR of 9.4%.

5.8.7 Pruning 77 significant associations to 12 independent signals

To prune our set of 77 significant associations to a set of approximately independent results, we used the following iterative greedy approach for each trait: we chose the pair of associations whose annotations had the most strongly correlated signed LD profiles, removed the annotation with the less significant p-value, and repeated until no annotations in the result set had signed LD profiles that were correlated at $R^2 > 0.25$. We used correlation between signed LD profiles rather than between the annotations themselves because, since our method regresses the summary statistics on the signed LD profile rather than the raw annotation, correlation between signed LD profiles most accurately represents the correlation between the test statistics for the two annotations.

5.8.8 Analysis of diseases and complex traits with annotations corresponding to directional effects of minor alleles

We constructed an alternate set of 382 annotations as follows. For each of the 382 ChIP-seq experiments represented by a set of peaks C, we set

$$v_m = \mathbf{1}\{m \in C\} u_m^1 \tag{5.15}$$

where u^1 is the signed background annotation corresponding to SNPs in the bottom quintile of the MAF spectrum. We then used signed LD profile regression to test for association between each of these 382 annotations and each of our 46 traits, assessing significance as above.

5.8.9 Data availability

We have released all genome annotations we analyzed, as well as regression weight matrices for our 1000 genomes reference panel, at http://data.broadinstitute.org/alkesgroup/SLDP/.

5.8.10 Code availability

Open-source software implementing our approach is available at http://www.github.com/yakirr/sldp.

5.8.11 URLs

Signed LD profile regression: open-source software is available at http://www.github.com/yakirr/sldp

Plink2: https://www.cog-genomics.org/plink2/

BLUEPRINT consortium data: ftp://ftp.ebi.ac.uk/pub/databases/blueprint/blueprint_Epivar/qtl_as/QTL_RESULTS/

TWAS weights for NTR data: https://data.broadinstitute.org/alkesgroup/FUSION/WGT/NTR.BLOOD.RNAARR.tar.bz2

5.9 Acknowledgements

We would like to acknowledge C de Boer, L Dicker, J Engreitz, N Friedman, X Liu, M Mitzenmacher, J Perry, S Reilly, D Reshef, S Raychaudhuri, A Schoech, P Sabeti, R Tewhey, P Turley, and the CGTA discussion group for helpful discussions.

A

Supplementary information for Chapter 2

A.1 Proof of Theorem 2.3.1

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 A.1.1. *Given a statistic $\hat{\varphi}$, a property of interest Φ, and some $\alpha \in [0,1]$, define $f(x) = \max R_\alpha(x)$ and $g(y) = \min I_\alpha(y)$. If f is strictly increasing, then f and g are inverses of each other.*

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

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

Lemma A.1.2. *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 $\max R_\alpha(x_0)$.*

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 $\max R_\alpha(x_0)$. □

Theorem 2.3.1 can then be seen to follow from the proposition below.

Proposition A.1.3. *Fix $0 < \alpha < 1$, and suppose $\hat{\varphi}$ is a statistic with the property that $\max R_\alpha(x)$ 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 $K^{x_0}_{\alpha/2}$ for $x_0 = \min I_\alpha(y)$. Equivalently, for $x_0 \in [0,1]$, the closure of the uncertain set of $K^{x_0}_{\alpha/2}$ equals $I_\alpha(y)$ for $y = \max R_\alpha(x_0)$.*

An illustration of this proposition and its proof is shown in Figure A.1.

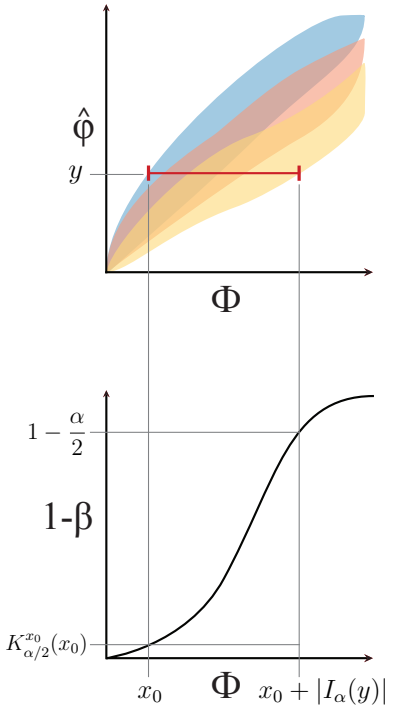


Figure A.1: The relationship between equitability and power, as in Proposition A.1.3. The top plot is the same as the one in Figure 2.1a, with the indicated interval denoting the $\mathcal{Q}$-confidence interval $I_\alpha(y)$. The bottom plot is a plot of the power function $K^{x_0}_{\alpha/2}(x)$, with the y-axis indicating statistical power. (Notice that because the null and alternative hypotheses are composite, $K^{x_0}_{\alpha/2}(x_0)$ need not equal $\alpha/2$; in general it may be lower.)

Proof. The equivalence of the two statements follows from Lemma A.1.1, which states that $y = \max R_\alpha(x_0)$ 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 $K^{x_0}_{\alpha/2}$ for $x_0 = \min I_\alpha(y)$.

Let U be the uncertain set of $K^{x_0}_{\alpha/2}$. 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 $K^{x_0}_{\alpha/2}(x_0) \leq \alpha/2 < 1 - \alpha/2$, which means that U is non-empty, 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 $\boldsymbol{\mathcal{Q}}$-confidence interval, we can find x arbitrarily close to $\max I_\alpha(y)$ from below such that $y \in R_\alpha(x)$. 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

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

i.e.,

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

But since as we already noted $y = \max R_\alpha(x_0)$, Lemma A.1.2 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, $K^{x_0}_{\alpha/2}(x) < 1 - \alpha/2$, meaning that $x \in U$.

It remains only to show that $\sup U \leq \max I_\alpha(y)$. To do so, we note that $y \notin R_\alpha(x)$ for all $x > \max I_\alpha(y)$. This implies that either $y > \max R_\alpha(x)$ or $y < \min R_\alpha(x)$. However, since $y \in R_\alpha(x_0)$ and $\max R_\alpha(\cdot)$ is an increasing function, no $x > x_0$ can have $y > \max R_\alpha(x)$. Thus the only option remaining is that $y < \min R_\alpha(x)$. 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

$$\mathbf{P}\left(\hat{\varphi}(Z) < y\right) < \frac{\alpha}{2}$$

i.e.,

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

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

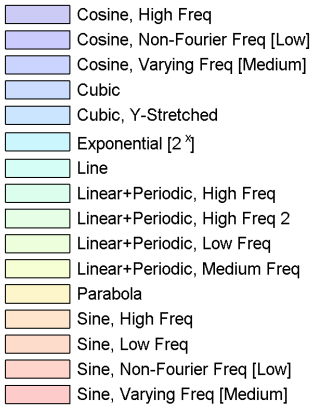
A.2 Details of empirical analyses

A.2.1 Example quantification of equitability in Figure 2.2

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 $\boldsymbol{\mathcal{Q}}$-acceptance region at each noise level. From the $\boldsymbol{\mathcal{Q}}$-acceptance regions we can then construct $\boldsymbol{\mathcal{Q}}$-confidence intervals, and the equitability of $\hat{\rho}$ is the reciprocal of the length of the largest of those intervals.

A.2.2 Functions analysed in Figures 2.2 and 3.3

Below is the legend showing which function types correspond to the colors in each of Figures 2.2 and 3.3. The functions used are the same as the ones in the equitability analyses of Reshef et al.[126].



The legend for Figures 2.2 and 3.3.

A.2.3 Parameters used in Figure 3.3

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. [126]. 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 3.3.

B

Supplementary information for Chapter 3

B.1 Proof of Theorem 3.3.1

This section is devoted to proving Theorem 3.3.1, restated below.

Theorem. *Let* $f : m^\infty \to \mathbb{R}$ *be uniformly continuous, and assume that* $f \circ r_i \to f$

pointwise. Then for every random variable (X, Y)*, we have*

$$\left(f \circ r_{B(n)}\right)\left(\widehat{M}(D_n)\right) \to f(M(X,Y))$$

in probability where D_n *is a sample of size* n *from the distribution of* (X, Y)*, provided* $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ *for some* $\varepsilon > 0$*.*

We prove the theorem by a sequence of lemmas that build on each other to bound the bias of $I^*(D, k, \ell)$. The general strategy is to capture the dependencies between different k-by-ℓ grids G by considering a "master grid" Γ that contains many more than $k\ell$ cells. Given this master grid, we first bound the difference between $I(D|_G)$ and $I((X,Y)|_G)$ only for sub-grids G of Γ. The bound is in terms of the difference between $D|_\Gamma$ and $(X,Y)|_\Gamma$. We then show that this bound can be extended without too much loss to all k-by-ℓ grids. This gives what we seek, because then the difference between $I(D|_G)$ and $I((X,Y)|_G)$ is uniformly bounded for all grids G in terms of the same random variable: $D|_\Gamma$. Once this is done, standard arguments give the consistency we seek.

In our argument we occasionally require technical facts about entropy and mutual information that are self-contained and unrelated to the central ideas. These lemmas are consolidated in Section B.12.

We begin by using one of these technical lemmas to prove a bound on the difference between $I(D|_G)$ and $I((X,Y)|_G)$ that is uniform over all grids G that are sub-grids of a much denser grid Γ. The common structure imposed by Γ will allow us to capture the

dependence between the quantities $|I(D|_G) - I((X,Y)|_G)|$ for different grids G.

Lemma B.1.1. *Let $\Pi = (\Pi_X, \Pi_Y)$ and $\Psi = (\Psi_X, \Psi_Y)$ be random variables distributed over the cells of a grid Γ, and let $(\pi_{i,j})$ and $(\psi_{i,j})$ be their respective distributions. Define*

$$\varepsilon_{i,j} = \frac{\psi_{i,j} - \pi_{i,j}}{\pi_{i,j}}.$$

Let G be a sub-grid of Γ with B cells. Then for every fixed $0 < a < 1$ we have

$$|I(\Psi|_G) - I(\Pi|_G)| \leq O\left((\log B) \sum_{i,j} |\varepsilon_{i,j}|\right)$$

when $|\varepsilon_{i,j}| \leq 1 - a$ for all i and j.

Proof. Let $P = \Pi|_G$ and $Q = \Psi|_G$ be the random variables induced by Π and Ψ respectively on the cells of G. Using the fact that $I(X,Y) = H(X) + H(Y) - H(X,Y)$, we write

$$|I(Q) - I(P)| \leq |H(Q_X) - H(P_X)| + |H(Q_Y) - H(P_Y)| + |H(Q) - H(P)|$$

where Q_X and P_X denote the marginal distributions on the columns of G and Q_Y and P_Y denote the marginal distributions on the rows. We can bound each of the terms on the right-hand side of the equation above using a Taylor expansion argument given in

Lemma B.12.1, whose proof is found in Section B.12. Doing so gives

$$|I(Q) - I(P)| \leq (\ln B) \left(\sum_i O\left(|\varepsilon_{i,*}|\right) + \sum_j O\left(|\varepsilon_{*,j}|\right) + \sum_{i,j} O\left(|\varepsilon_{i,j}|\right) \right)$$

where

$$\varepsilon_{i,*} = \frac{\sum_j (\psi_{i,j} - \pi_{i,j})}{\sum_j \pi_{i,j}}$$

and $\varepsilon_{*,j}$ is defined analogously.

To obtain the result, we observe that

$$|\varepsilon_{i,*}| = \left| \frac{\sum_j \pi_{i,j} \varepsilon_{i,j}}{\sum_j \pi_{i,j}} \right| \leq \frac{\sum_j \pi_{i,j} \left|\varepsilon_{i,j}\right|}{\sum_j \pi_{i,j}} \leq \sum_j |\varepsilon_{i,j}|$$

since $\pi_{i,j} / \sum_j \pi_{i,j} \leq 1$, and the analogous bound holds for $|\varepsilon_{*,j}|$. □

We now extend Lemma B.1.1 to all grids with B cells rather than just those that are sub-grids of the master grid Γ. The proof of this lemma relies on an information-theoretic result proven in Section B.2 that bounds the difference in mutual information between two distributions that can be obtained from each other by moving a small amount of probability mass.

Lemma B.1.2. *Let $\Pi = (\Pi_X, \Pi_Y)$ and $\Psi = (\Psi_X, \Psi_Y)$ be random variables, and let Γ be a grid. Define $\varepsilon_{i,j}$ on $\Pi|_\Gamma$ and $\Psi|_\Gamma$ as in Lemma B.1.1. Let G be any grid with B cells, and let δ (resp. d) represent the total probability mass of $\Pi|_\Gamma$ (resp. $\Psi|_\Gamma$) falling*

in cells of Γ *that are not contained in individual cells of* G*. We have that*

$$|I(\Psi|_G) - I(\Pi|_G)| \leq O\left(\left(\sum_{i,j} |\varepsilon_{i,j}| + \delta + d\right) \log B + \delta \log(1/\delta) + d \log(1/d)\right)$$

provided that the $|\varepsilon_{i,j}|$ *are bounded away from 1 and that* $d, \delta \leq 1/2$*.*

Proof. In the proof below, we use the convention that for any two grids G and G' and any random variable Z, the expression $\Delta^Z(G, G')$ denotes $|I(Z|_G) - I(Z|_{G'})|$.

Consider the grid G' obtained by replacing every horizontal or vertical line in G that is not in Γ with a closest line in Γ. The grid G' is clearly a sub-grid of Γ. Moreover, $\Pi|_{G'}$ (resp. $\Psi|_{G'}$) can be obtained from $\Pi|_G$ (resp. $\Pi|_G$) by moving at most δ (resp. d) probability mass. This can be shown to imply that

$$\Delta^{\Pi}(G, G') \leq O\left(\delta \log(1/\delta) + \delta \log B\right) \quad \text{and} \quad \Delta^{\Psi}(G', G) \leq O\left(d \log(1/d) + d \log B\right).$$

The proof of this information-theoretic fact is self-contained and so we defer it to Proposition B.2.2 in Section B.2, as it is more central to the arguments presented there.

With $\Delta^{\Phi}(G, G')$ and $\Delta^{\Psi}(G', G)$ bounded in terms of δ and d, we can bound $|I(\Psi|_G) - I(\Phi|_G)|$ using the triangle inequality by comparing it with

$$\Delta^{\Pi}(G, G') + \left|I\left(\Pi|_{G'}\right) - I\left(\Psi|_{G'}\right)\right| + \Delta^{\Psi}(G', G)$$

and bounding the middle term using Lemma B.1.1, since $G' \subset \Gamma$. □

We now use the fact that the variables $\varepsilon_{i,j}$ defined in Lemma B.1.1 are small with high probability to give a concrete bound on the bias of $I(D|_G)$ that is uniform over all k-by-ℓ grids G and that holds with high probability. It is useful at this point to recall that, given a distribution (X, Y), an *equipartition* of (X, Y) is a grid G such that all the rows of $(X, Y)|_G$ have the same probability mass, and all the columns do as well.

Lemma B.1.3. *Let D_n be a sample of size n from the distribution of a pair (X, Y) of jointly distributed random variables. For any $\alpha \geq 0$, any $\varepsilon > 0$, and any integers $k, \ell > 1$, we have that for all n*

$$|I(D_n|_G) - I((X, Y)|_G)| \leq O\left(\frac{\log(k\ell)}{C(n)^\alpha} + \frac{\log(k\ell n)}{n^{\varepsilon/4}}\right)$$

for every k-by-ℓ grid G with probability at least $1 - C(n)e^{-\Omega(n/C(n)^{1+2\alpha})}$, where $C(n) = k\ell n^{\varepsilon/2}$.

Proof. Fix n, and let Γ be an equipartition of (X, Y) into $kn^{\varepsilon/4}$ rows and $\ell n^{\varepsilon/4}$ columns. $C(n)$ is now the number of cells in Γ. Lemma B.1.2, with $\Pi = (X, Y)$ and $\Psi = D$, shows that $|I(D|_G) - I((X, Y)|_G)|$ is at most

$$O\left(\left(\sum_{i,j} |\varepsilon_{i,j}| + \delta + d\right) \log(k\ell) + \delta \log(1/\delta) + d \log(1/d)\right)$$

provided the $\varepsilon_{i,j}$ have absolute value bounded away from 1, and provided that $d, \delta \leq 1/2$.

The remainder of the proof proceeds as follows. We first show that the $\varepsilon_{i,j}$ are small

with high probability. This will both show that the lemma's requirement on the $\varepsilon_{i,j}$ holds and allow us to bound the sum in the inequality above. We will then use our bound on the $\varepsilon_{i,j}$ to bound d in terms of δ. Finally, we will bound δ using the fact that the number of rows and columns in Γ increases with n. This will give us that $d, \delta \leq 1/2$ and allow us to bound the rest of the terms in the expression above.

Bounding the $\varepsilon_{i,j}$: We bound the $\varepsilon_{i,j}$ using a multiplicative Chernoff bound. Let $\pi_{i,j}$ and $\psi_{i,j}$ represent the probability mass functions of $(X, Y)|_\Gamma$ and $D|_\Gamma$ respectively. We write

$$\begin{aligned} \mathbf{P}\left(|\varepsilon_{i,j}| \geq \delta\right) &= \mathbf{P}\left(\pi_{i,j}(1-\delta) \leq \psi_{i,j} \leq \pi_{i,j}(1+\delta)\right) \\ &\leq e^{-\Omega(n\pi_{i,j}\delta^2)} \end{aligned}$$

since $\psi_{i,j}$ is a sum of n i.i.d. Bernoulli random variables and $\mathbf{E}\left(\psi_{i,j}\right) = n\pi_{i,j}$. (See, e.g., Mitzenmacher & Upfal[94].) Setting $\delta = \sqrt{\pi_{i,j}}/C(n)^{1/2+\alpha}$ yields

$$\mathbf{P}\left(|\varepsilon_{i,j}| \geq \frac{\sqrt{\pi_{i,j}}}{C(n)^{1/2+\alpha}}\right) \leq e^{-\Omega(n/C(n)^{1+2\alpha})}.$$

A union bound over the pairs (i, j) then gives that, with the desired probability, the above bound on $|\varepsilon_{i,j}|$ holds for all i, j.

Bounding $\sum |\varepsilon_{i,j}|$*:* The bound on the $\varepsilon_{i,j}$ implies that

$$\begin{aligned}\sum_i |\varepsilon_{i,j}| &\leq \frac{1}{C(n)^{1/2+\alpha}} \sum_{i,j} \sqrt{\pi_{i,j}} \\ &\leq \frac{1}{C(n)^{1/2+\alpha}} \sqrt{C(n)} \\ &\leq \frac{1}{C(n)^{\alpha}}\end{aligned}$$

where the second line follows from the fact that the function $\sum \sqrt{\pi_{i,j}}$ is symmetric and concave and therefore, when restricted to the hyperplane $\sum \pi_{i,j} = 1$, must achieve its maximum when $\pi_{i,j} = 1/C(n)$ for all i, j.

Bounding d *in terms of* δ*:* We use our bound on the $\varepsilon_{i,j}$ to bound d. We do so by observing that it implies

$$\psi_{i,j} \leq \pi_{i,j}\left(1 + \frac{\sqrt{\pi_{i,j}}}{C(n)^{1/2+\alpha}}\right) = \pi_{i,j} + \frac{\pi_{i,j}^{3/2}}{C(n)^{1/2+\alpha}} \leq \pi_{i,j} + \frac{\pi_{i,j}}{C(n)^{1/2+\alpha}} \leq 2\pi_{i,j}$$

since $\pi_{i,j} \leq 1$ and $C(n) \geq 1$.

The connection to d comes from the fact that for any column j of Γ, this means that

$$\psi_{*,j} = \sum_i \psi_{i,j} \leq 2 \sum_i \pi_{i,j} = 2\pi_{*,j}.$$

This also applies to the sums across rows. Since d is a sum of terms of the form $\psi_{*,j}$ and $\psi_{i,*}$ for j in some index set J and i in an index set I, and δ is a sum of terms of

the form $\pi_{*,j}$ and $\pi_{i,*}$ with the same index sets, we therefore get that $d \leq 2\delta$.

Bounding δ and obtaining the result: To bound δ, we observe that because G has at most $\ell - 1$ vertical lines and $k - 1$ horizontal lines, we have

$$\delta \leq \frac{\ell}{\ell n^{\varepsilon/4}} + \frac{k}{k n^{\varepsilon/4}} \leq \frac{2}{n^{\varepsilon/4}}.$$

This bound on δ allows us to bound the terms involving d and δ by

$$\delta + d \leq O\left(\frac{1}{n^{\varepsilon/4}}\right), \quad \delta \log\left(\frac{1}{\delta}\right) + d \log\left(\frac{1}{d}\right) \leq O\left(\frac{\log n}{n^{\varepsilon/4}}\right).$$

Combining all of the bounds gives the desired result. □

Our final lemma shows that as long as $B(n)$ doesn't grow too fast, the bound from the previous lemma yields a uniform bound on the entire sample characteristic matrix. This is done by specifying an error threshold for which Lemma B.1.3 yields a bound that holds with high probability, and then invoking a union bound.

Lemma B.1.4. *Let D_n be a sample of size n from the distribution of a pair (X, Y) of jointly distributed random variables. For every $B(n) = O\left(n^{1-\varepsilon}\right)$, there exists an $a > 0$ such that for sufficiently large n,*

$$\left|\widehat{M}(D_n)_{k,\ell} - M(X,Y)_{k,\ell}\right| \leq O\left(\frac{1}{n^a}\right)$$

holds for all $k\ell \leq B(n)$ with probability $P(n) = 1 - o(1)$, where $\widehat{M}(D_n)_{k,\ell}$ is the k,ℓ-th entry of the sample characteristic matrix and $M(X,Y)_{k,\ell}$ is the k,ℓ-th entry of the population characteristic matrix of (X,Y).

Proof. Fix k,ℓ, and any α satisfying $0 < \alpha < \varepsilon/(4-2\varepsilon)$. Lemma B.1.3 implies that with high probability the difference $|\widehat{M}(D_n)_{k,\ell} - M_{k,\ell}|$ is at most

$$\begin{aligned} O\left(\frac{\log(k\ell)}{C(n)^\alpha} + \frac{\log(k\ell n)}{n^{\varepsilon/4}}\right) &\leq O\left(\frac{\log n}{C(n)^\alpha} + \frac{\log n}{n^{\varepsilon/4}}\right) \\ &\leq O\left(\frac{\log n}{n^{\alpha\varepsilon/2}} + \frac{\log n}{n^{\varepsilon/4}}\right) \end{aligned}$$

where the first inequality comes from $k\ell \leq B(n)$ and second is because $C(n) = k\ell n^{\varepsilon/2} \geq n^{\varepsilon/2}$. This bound is at most $O\left(1/n^a\right)$ for every $a < \min\{\alpha\varepsilon/2, \varepsilon/4\}$, as desired. It remains only to show that the bound holds with high probability across all $k\ell \leq B(n)$.

Lemma B.1.3 states that the probability our bound holds for one fixed pair (k,ℓ) is at least

$$1 - C(n)e^{-\Omega(n/C(n)^{1+2\alpha})} \geq 1 - O\left(n\right)e^{-\Omega(n^u)}$$

for some positive u. This is because $C(n) \leq B(n)n^{\varepsilon/2} \leq O\left(n^{1-\varepsilon/2}\right)$ for large n, and so our choice of α ensures that $C(n)^{1+2\alpha} = O\left(n^{1-u}\right)$ for some $u > 0$.

We can then perform a union bound over all pairs $k\ell \leq B(n)$: since the number of such pairs can be bounded by a polynomial in n, we have that the desired condition is satisfied for all $k\ell \leq B(n)$ with probability approaching 1. □

We are now ready to prove the main result.

Theorem. *Let $f : m^\infty \to \mathbb{R}$ be uniformly continuous, and assume that $f \circ r_i \to f$ pointwise. Then for every random variable (X, Y), we have*

$$\left(f \circ r_{B(n)}\right)\left(\widehat{M}(D_n)\right) \to f(M(X,Y))$$

in probability where D_n is a sample of size n from the distribution of (X, Y), provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.

Proof. Let N denote $B(n)$, let $M_N = r_N(M)$, and let $\widehat{M}_N(D_n) = r_N(\widehat{M}(D_n))$. We begin by writing

$$\begin{aligned}\left|f\left(\widehat{M}_N(D_n)\right) - f(M)\right| &\leq \left|f\left(\widehat{M}_N(D_n)\right) - f\left(M_N\right)\right| + \left|f\left(M_N\right) - f(M)\right| \\ &= \left|f\left(\widehat{M}_N(D_n)\right) - f\left(M_N\right)\right| + \left|\left(f \circ r_N\right)(M) - f(M)\right|\end{aligned}$$

and observing that as $n \to \infty$, the second term vanishes by the pointwise convergence of $f \circ r_i$ and the fact that $B(n) > \omega(1)$. It therefore suffices to show that the first term converges to zero in probability. Since f is uniformly continuous, we can establish this via a simple adaptation of the continuous mapping theorem, which says that if the sequence of random variables $R_n \to R$ in probability, and g is continuous, then $g(R_n) \to g(R)$ in probability. We replace R with a second sequence, and replace continuity with uniform continuity.

Let $\|\cdot\|$ denote the supremum norm on m^∞, and fix any $z > 0$. Then, for any $\delta > 0$, define

$$C_\delta = \left\{ A \in m^\infty : \exists A' \in m^\infty \text{ s.t. } \|A - A'\| < \delta, \left| f(A) - fA') \right| > z \right\}.$$

This is the set of matrices $A \in m^\infty$ for which it is possible to find, within a δ-neighborhood of A, a second matrix that f maps to more than z away from $f(A)$. Because f is uniformly continuous, there exists a δ^* sufficiently small so that $C_{\delta^*} = \emptyset$.

Suppose that $|f(\widehat{M}_N(D_n)) - f(M_N)| > z$. This means that either $\|\widehat{M}_N(D_n) - M_N\| > \delta^*$, or $M_N \in C_{\delta^*}$. The latter option is impossible since $C_{\delta^*} = \emptyset$, and Lemma B.1.4 tells us that $\mathbf{P}\left(\|\widehat{M}_N(D_n) - M_N\| > \delta^*\right) \to 0$ as n grows. We therefore have that

$$\left| f\left(\widehat{M}_N(D_n)\right) - f(M_N) \right| \to 0$$

in probability, as desired. □

B.2 Proof of Theorem 3.3.2

In this section we prove Theorem 3.3.2, reproduced below.

Theorem. *Let $\boldsymbol{\mathcal{P}}(\mathbb{R}^2)$ denote the space of random variables supported on $\mathbb{R}^2$ equipped with the metric of statistical distance. The map from $\boldsymbol{\mathcal{P}}(\mathbb{R}^2)$ to m^∞ defined by $(X, Y) \mapsto M(X, Y)$ is uniformly continuous.*

The proposition below begins our argument with the simple observation that the family of maps consisting of applying any finite grid to some $(X, Y) \in \boldsymbol{\mathcal{P}}(\mathbb{R}^2)$ is uniformly equicontinuous. The reason this holds is that $(X, Y)|_G$ is a deterministic function of (X, Y), and deterministic functions cannot increase statistical distance.

Proposition B.2.1. *Let $\mathbb{G}$ be the set of all finite grids. The family $\{(X, Y) \mapsto (X, Y)|_G : G \in \mathbb{G}\}$ is uniformly equicontinuous on $\boldsymbol{\mathcal{P}}(\mathbb{R}^2)$.*

Proof. To establish uniform equicontinuity, we need to show that, given some $(X, Y) \in \boldsymbol{\mathcal{P}}(\mathbb{R}^2)$ and some $\varepsilon > 0$, we can choose δ to satisfy the continuity condition in a way that does not depend on G or on (X, Y). But because deterministic functions cannot increase statistical distance, we have that if $(X, Y), (X', Y') \in \boldsymbol{\mathcal{P}}$ are at most ε apart then

$$\Delta\left((X, Y)|_G, (X', Y')|_G\right) \leq \Delta\left((X, Y), (X', Y')\right) = \varepsilon$$

where Δ denotes statistical distance. Choosing $\delta = \varepsilon$ therefore gives the result. □

At this point it is tempting to try to use continuity properties of discrete mutual information to obtain uniform continuity of the characteristic matrix. And indeed, this strategy does yield that each *individual* entry of the characteristic matrix is a uniformly continuous function. However, to obtain continuity of the entire (infinite) characteristic matrix we need to make a statement about all grid resolutions simultaneously. This is not straightforward because mutual information is only uniformly continuous for a

fixed grid resolution, and the family $\{(X,Y) \mapsto I((X,Y)|_G) : G \in \mathbb{G}\}$ is in fact not even equicontinuous.

The normalization in the definition of MIC_* is what allows us to establish the uniform continuity of the characteristic matrix despite this problem. To see why, suppose we have a distribution over a k-by-ℓ grid and we are allowed to move at most δ away in statistical distance for some small δ. The largest change in discrete mutual information that this can cause indeed increases as we increase k and ℓ. However, it turns out that we can bound the extent of this "non-uniformity": the proposition below shows that as we move away from a distribution, the discrete mutual information can change only proportionally to the amount of mass we move, with the proportionality constant bounded by $\log \min\{k, \ell\}$. Because $\log \min\{k, \ell\}$ is the quantity by which we regularize the entries of the characteristic matrix, this is exactly enough to make the normalized matrix continuous. This proposition is the technical heart of our continuity result. And as we show in Corollary 3.3.9 when we demonstrate the non-continuity of the non-normalized characteristic matrix mutual information, our bound is tight.

Proposition B.2.2. *Let $I_{k,\ell} : \mathcal{P}(\{1,\ldots,k\}\times\{1,\ldots,\ell\}) \to \mathbb{R}$ denote the discrete mutual information function on k-by-ℓ grids. For $0 < \delta \leq 1/4$, the maximal change in $I_{k,\ell}$ over any subset of $\mathcal{P}(\{1,\ldots,k\} \times \{1,\ldots,\ell\})$ of diameter δ (in statistical distance) is*

$$O\left(\delta \log\left(\frac{1}{\delta}\right) + \delta \log \min\{k, \ell\}\right).$$

Proof. Without loss of generality, assume $k \leq \ell$, so that $\log \min\{k, \ell\} = \log k$. Let (X, Y) and (X', Y') be two random variables distributed over $\{1, \ldots, k\} \times \{1, \ldots, \ell\}$ that are at most δ apart in statistical distance. Using $I(X, Y) = H(Y) - H(Y|X)$, we can express the difference between the mutual information of these two pairs of random variables as

$$\left|I(X, Y) - I(X', Y')\right| \leq \left|H(Y) - H(Y')\right| + \left|H(Y|X) - H(Y'|X')\right|.$$

We now use Lemma B.12.5, which relates movement of probability mass to changes in entropy and is proven in Section B.12, to separately bound each of the terms on the right hand side. Straightforward application of the lemma to $|H(Y) - H(Y')|$ shows that it is at most $2H_b(2\delta) + 3\delta \log k$, where $H_b(\cdot)$ is the binary entropy function. Since $H_b(x) \leq O(x \log(1/x))$ for x small, this is $O(\delta \log(1/\delta) + \delta \log k)$.

Bounding the term with the conditional entropies is more involved. Let $p_x =$

$\mathbf{P}(X = x)$, and let $p'_x = \mathbf{P}(X' = x)$. We have

$$\begin{aligned}
\left|H(Y|X) - H(Y'|X')\right| &= \sum_x \left|p_x H(Y|X=x) - p'_x H(Y'|X'=x)\right| \\
&\leq \sum_x \Big(p_x \left|H(Y|X=x) - H(Y'|X'=x)\right| + \qquad \text{(B.1)} \\
&\qquad\qquad \left|p'_x - p_x\right| H(Y'|X'=x)\Big) \\
&= \sum_x p_x \left|H(Y|X=x) - H(Y'|X'=x)\right| + \sum_x \left|p'_x - p_x\right| \log k \\
&\leq \sum_x p_x \left|H(Y|X=x) - H(Y'|X'=x)\right| + \delta \log k \qquad \text{(B.2)}
\end{aligned}$$

where the last line is because $\sum_x |p_x - p'_x| \leq \delta$ and $H(Y'|X'=x) \leq \log k$.

Now let δ_{x+} be the magnitude of all the probability mass entering any cell in column x, let δ_{x-} be the magnitude of all the probability mass leaving any cell in column x, and let $\delta_x = \delta_{x+} + \delta_{x-}$. Using this notation, we can again apply Lemma B.12.5 to obtain

$$\begin{aligned}
\sum_x p_x \left|H(Y|X=x) - H(Y'|X'=x)\right| &\leq \sum_x p_x \left(2H_b\left(\frac{2\delta_x}{p_x}\right) + 3\frac{\delta_x}{p_x}\log k\right) \\
&= 2\sum_x p_x H_b\left(\frac{2\delta_x}{p_x}\right) + 3\sum_x \delta_x \log k \\
&\leq 2\sum_x p_x H_b\left(\frac{2\delta_x}{p_x}\right) + 3\delta \log k \\
&\leq 2H_b(2\delta) + 3\delta \log k
\end{aligned}$$

where the last line is by application of Lemma B.12.2 from the appendix, which bounds weighted sums of binary entropies.

Combining this with Line (B.2) gives that

$$\left|H(Y|X) - H(Y'|X')\right| \leq 2H_b(2\delta) + 4\delta \log k$$

which, together with the bound on $|H(Y) - H(Y')|$ and the fact that $H_b(X) \leq O(x \log(1/x))$ for x small, gives the result. □

Having bounded the extent to which variation in mutual information depends on grid resolution, we are now ready to show the uniform continuity of the characteristic matrix.

Theorem *Let $\boldsymbol{\mathcal{P}}(\mathbb{R}^2)$ denote the space of random variables supported on $\mathbb{R}^2$ equipped with the metric of statistical distance. The map from $\boldsymbol{\mathcal{P}}(\mathbb{R}^2)$ to m^∞ defined by $(X, Y) \mapsto M(X, Y)$ is uniformly continuous.*

Proof. We complete the proof in three steps. First, we show that a certain family of functions F is uniformly equicontinuous. Second, we use this to show that a different family F' consisting of functions of the form $\sup_{g \in A} g$ with $A \subset F$ is uniformly equicontinuous. Finally, we argue that since the entries of $M(X, Y)$ consist of the functions in F', this is sufficient to establish the result.

Define

$$F = \left\{ (X, Y) \mapsto \frac{I_{k,\ell}((X, Y)|_G)}{\log \min\{k, \ell\}} : k, \ell \in \mathbb{Z}_{>1}, G \in G(k, \ell) \right\}.$$

F is uniformly equicontinuous by the following argument. Given some $\varepsilon > 0$, we know (Proposition B.2.1) that for any (X', Y') in an ε-ball around (X, Y), $(X', Y')|_G$ will remain within ε of $(X, Y)|_G$ for any G. Proposition B.2.2 then tells us that if ε is sufficiently small then the distance between $I_{k,\ell}((X', Y')|_G)$ and $I_{k,\ell}((X, Y)|_G)$ will be at most

$$O\left(\varepsilon \log(1/\varepsilon) + \varepsilon \log \min\{k, \ell\}\right).$$

After the normalization, this becomes at most $O(\varepsilon(\log(1/\varepsilon) + 1))$, which goes to zero (uniformly with respect to (X, Y)) as ε approaches zero, as desired.

Next, define

$$F' = \{(X, Y) \mapsto M(X, Y)_{k,\ell} : k, \ell \in \mathbb{Z}_{>1}\}.$$

Each map in F' is of the form $\sup_{g \in A} g$ for some $A \subset F$. Therefore, for a given $\varepsilon > 0$, whatever δ establishes the uniform equicontinuity for F can be used to establish continuity of all the functions in F'. (To see this: $\sup_{g \in A} g$ can't increase by more than ε if no g increases by more than ε, and $\sup_{g \in A} g$ is also lower bounded by any of the g's, so it can't decrease by more than ε either.) Since we can use the same δ for all of the maps in F', they therefore form a uniformly equicontinuous family.

Finally, the δ provided by the uniform equicontinuity of F' also ensures that $M(X', Y')$ is within ε of $M(X, Y)$ in the supremum norm, thus giving the uniform continuity of $(X, Y) \mapsto M(X, Y)$. □

B.3 Proof of Proposition 3.3.8

Theorem *For some function $N(k,\ell)$, let M^N be the characteristic matrix with normalization N, i.e.,*

$$M^N(X,Y) = \frac{I^*((X,Y),k,\ell)}{N(k,\ell)}.$$

If $N(k,\ell) = o(\log\min\{k,\ell\})$ along some infinite path in $\mathbb{N}\times\mathbb{N}$, then M^N and $\sup M^N$ are not continuous as functions of $\mathcal{P}([0,1]\times[0,1]) \subset \boldsymbol{\mathcal{P}}(\mathbb{R}^2)$.

Proof. Consider a random variable Z uniformly distributed on $[0,1/2]^2$. Because Z exhibits statistical independence, $I^*(Z,k,\ell)$ is zero for all k,ℓ. Now define Z_ε to be uniformly distributed on $[0,1/2]^2$ with probability $1-\varepsilon$ and uniformly distributed on the line from $(1/2,1/2)$ to $(1,1)$ with probability ε.

We lower-bound $I^*(Z_\varepsilon,k,\ell)$. Without loss of generality suppose that $k \le \ell$, and consider a grid that places all of $[0,1/2]^2$ into one cell and uniformly partitions the set $[1/2,1]^2$ into $k-1$ rows and $k-1$ columns. By considering just the rows/columns in the set $[1/2,1]^2$ we see that this grid gives a mutual information of at least $\varepsilon\log(k-1)$. Thus, we have that for all k,ℓ,

$$I^*(Z_\varepsilon,k,\ell) \ge \varepsilon\log\min\{k-1,\ell-1\}.$$

This implies that the limit of $M^N(Z_\varepsilon)$ along P is ∞, and so the distance between

$M^N(Z)$ and $M^N(Z_\varepsilon)$ in the supremum norm is infinite. □

B.4 Proof of Theorem 3.3.4

Theorem. *Let M be a population characteristic matrix. Then $M_{k,\uparrow}$ equals*

$$\max_{P \in P(k)} \frac{I(X, Y|_P)}{\log k}$$

where $P(k)$ denotes the set of all partitions of size at most k.

Proof. Define

$$M^*_{k,\uparrow} = \max_{P \in P(k)} \frac{I(X, Y|_P)}{\log k}.$$

We wish to show that $M^*_{k,\uparrow}$ is in fact equal to $M_{k,\uparrow}$. To show that $M_{k,\uparrow} \leq M^*_{k,\uparrow}$, we observe that for every k-by-ℓ grid $G = (P, Q)$, where P is a partition into rows and Q is a partition into columns, the data processing inequality gives $I((X, Y)|_G) \leq I(X, Y|_P)$. Thus $M_{k,\ell} \leq M^*_{k,\uparrow}$ for $\ell \geq k$, implying that

$$M_{k,\uparrow} = \lim_{\ell \to \infty} M_{k,\ell} \leq M^*_{k,\uparrow}.$$

It remains to show that $M^*_{k,\uparrow} \leq M_{k,\uparrow}$. To do this, we let P be any partition into k rows, and we define Q_ℓ to be an equipartition into ℓ columns. We let

$$M^*_{k,\ell,P} = \frac{I(X|_{Q_\ell}, Y|_P)}{\log k}.$$

Since $M^*_{k,\ell,P} \leq M_{k,\ell}$ when $\ell \geq k$, we have that for all P

$$\frac{I(X,Y|_P)}{\log k} = \lim_{\ell\to\infty} M^*_{k,\ell,P} \leq \lim_{\ell\to\infty} M_{k,\ell} = M_{k,\uparrow}$$

which gives that

$$M^*_{k,\uparrow} = \sup_P \frac{I(X,Y|_P)}{\log k} \leq M_{k,\uparrow}$$

as desired. □

B.5 Proof of Theorem 3.3.5

Theorem. *Given a random variable (X,Y), $M_{k,\uparrow}$ (resp. $M_{\uparrow,\ell}$) is computable to within an additive error of $O(k\varepsilon\log(1/(k\varepsilon))) + E$ (resp. $O(\ell\varepsilon\log(1/(\ell\varepsilon))) + E$) in time $O(kT(E)/\varepsilon)$ (resp. $O(\ell T(E)/\varepsilon)$), where $T(E)$ is the time required to numerically compute the mutual information of a continuous distribution to within an additive error of E.*

Proof. Without loss of generality we prove the claim only for $M_{k,\uparrow}$. Given $0 < \varepsilon < 1$, we would like a partition into rows P of size at most k such that $I(X,Y|_P)$ is maximized. We would like to use OptimizeXAxis for this purpose, but while our search problem is continuous, OptimizeXAxis can only perform a discrete search over sub-partitions of some master partition Π. We therefore set Π to be an equipartition into $1/\varepsilon$ rows and show that this gets us close enough to achieve the desired result.

With Π as described, the OptimizeXAxis provides in time $O(kT(E)/\varepsilon)$ a partition P_0 into at most k rows such that $I(X, Y|_{P_0})$ is maximized, subject to $P_0 \subset \Pi$, to within an additive error of E. To prove the claim then, we must show that the loss we incur by restricting to sub-partitions of Π costs us at most $O(k\varepsilon \log(1/(k\varepsilon)))$. In other words, we must show that

$$I(X, Y|_P) - I(X, Y|_{P_0}) \leq O(k\varepsilon)$$

where P is an optimal partition into rows. Note that we have omitted the absolute value above, since by the optimality of P, $I(X, Y|_P) \geq I(X, Y|_{P_0})$ always.

We prove the desired bound by showing that there exists some $P' \subset \Pi$ such that the mutual information of $(X, Y|_{P'})$ is $O(k\varepsilon \log(1/(k\varepsilon)))$-close to that achieved with $(X, Y|_P)$. Since $P' \subset \Pi$ gives us that $I(X, Y|_{P_0}) \geq I(X, Y|_{P'})$, we may then conclude that $I(X, Y|_P) - I(X, Y|_{P_0})$ is at most $O(k\varepsilon \log(1/(k\varepsilon)))$.

We construct P' by simply replacing every horizontal line in P with a horizontal line in Π closest to it. Since there are at most $k-1$ horizontal lines in P, and each such line is contained in a row of Π containing $1/\varepsilon$ probability mass, performing this operation moves at most $(k-1)\varepsilon$ probability mass. In other words, the statistical distance between $(X, Y|_{P'})$ and $(X, Y|_P)$ is at most $(k-1)\varepsilon \leq k\varepsilon$. Thus, for sufficiently small ε, Proposition B.2.2, proven in Section B.2, can be used to show that

$$|I(X, Y|_{P'}) - I(X, Y|_P)| \leq O\left(k\varepsilon \log\left(\frac{1}{k\varepsilon}\right) + k\varepsilon \log\left(\frac{1}{\varepsilon}\right)\right)$$

which yields the desired result. □

Remark B.5.1. *We do not explore here the details of the numerical integration associated with the above theorem, since the error introduced by the numerical integration is independent of the algorithm being proposed. However, standard numerical integration methods can be used to make this error arbitrarily small with an understood complexity tradeoff (see, e.g., Stoer & Bulirsch[147]).*

B.6 Proof of Theorem 3.4.1

Theorem. *Let (X, Y) be jointly distributed random variables. Then $\partial[M] = \partial M$.*

Proof. Without loss of generality, we show that $[M]_{k,\uparrow} = M_{k,\uparrow}$. Fix any partition into rows P. If Q_ℓ is an equipartition into ℓ columns then

$$\lim_{\ell \to \infty} I(X|_{Q_\ell}, Y|_P) = I(X, Y|_P),$$

because the continuous mutual information equals the limit of the discrete mutual information with increasingly fine partitions. (See, e.g., Chapter 8 of Cover & Thomas[23] for a proof of this.) This means that, letting $P(k)$ denote the set of all partitions of size at most k, we have

$$[M]_{k,\uparrow} = \max_{P \in P(k)} \frac{I(X, Y|_P)}{\log k} = M_{k,\uparrow}$$

where the second equality follows from Proposition 3.3.4. □

B.7 Consistency of MIC_e in estimating MIC_*

The consistency of MIC_e for estimating MIC_* can be established using the same technical lemmas that we used to show that $\text{MIC} \to \text{MIC}_*$. Specifically, we can use Lemma B.1.3, which bounds the difference, for all k-by-ℓ grids G, between the sample quantity $I(D_n|_G)$ and the population quantity $I((X,Y)|_G)$ with high probability, where D_n is a sample of size n from (X,Y). That lemma yields the following fact about the sample equicharacteristic matrix, whose proof is similar to that of Lemma B.1.4.

Lemma B.7.1. *Let D_n be a sample of size n from the distribution of a pair (X,Y) of jointly distributed random variables. For every $B(n) = O\left(n^{1-\varepsilon}\right)$, there exists an $a > 0$ such that for sufficiently large n,*

$$\left|\widehat{[M]}(D_n)_{k,\ell} - [M](X,Y)_{k,\ell}\right| \leq O\left(\frac{1}{n^a}\right)$$

holds for all $k\ell \leq B(n)$ with probability $P(n) = 1 - o(1)$, where $\widehat{[M]}(D_n)_{k,\ell}$ is the k,ℓ-th entry of the sample equicharacteristic matrix and $[M](X,Y)_{k,\ell}$ is the k,ℓ-th entry of the population equicharacteristic matrix of (X,Y).

In the case of MIC, we proceeded to apply abstract continuity considerations to obtain our consistency theorem (Theorem 3.3.1) from a result analogous to the above lemma. A similar argument shows us that, in the case of the equicharacteristic matrix as well, we can estimate a large class of functions of the matrix in the same way. This is stated

formally in the theorem below. As before, we let m^∞ be the space of infinite matrices equipped with the supremum norm, and given a matrix A the projection r_i zeros out all the entries $A_{k,\ell}$ for which $k\ell > i$.

Theorem. *Let $f : m^\infty \to \mathbb{R}$ be uniformly continuous, and assume that $f \circ r_i \to f$ pointwise. Then for every random variable (X, Y), we have*

$$\left(f \circ r_{B(n)}\right)\left(\widehat{[M]}(D_n)\right) \to f([M](X, Y))$$

in probability where D_n is a sample of size n from the distribution of (X, Y), provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.

B.8 The EquicharClump algorithm

In Theorem 3.4.4, we sketched an algorithm called EquicharClump for approximating the sample equicharacteristic matrix that is more efficient than the naive computation. In this section, we describe the algorithm in detail, bound its runtime, and show that it indeed yields a consistent estimator of MIC_* from finite samples as well as a consistent independence test when used to compute the total information coefficient. We then present some empirical results characterizing the sensitivity of the algorithm to its speed-versus-optimality parameter c.

The results in this section can be summarized as follows: let (X, Y) be a pair of jointly distributed random variables, and let D_n be a sample of size n from the distribution of

(X, Y). For every $c \geq 1$, there exists a matrix $\{\widehat{M}\}^c(D_n)$ such that

1. There exists an algorithm EquicharClump for computing $r_B(\{\widehat{M}\}^c(D_n))$ in time $O(n + B^{5/2})$, which equals $O(n + n^{5(1-\varepsilon)/2})$ when $B(n) = O(n^{1-\varepsilon})$.

2. The function
$$\widetilde{\mathrm{MIC}}_{e,B}(\cdot) = \max_{k\ell \leq B(n)} \{\widehat{M}\}^c(\cdot)_{k,\ell}$$
is a consistent estimator of MIC_* provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.

3. The function
$$\widetilde{\mathrm{TIC}}_{e,B}(\cdot) = \sum_{k\ell \leq B(n)} \{\widehat{M}\}^c(\cdot)_{k,\ell}$$
yields a consistent right-tailed test of independence provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $\varepsilon > 0$

We will prove these results in order.

B.8.1 Algorithm description and analysis of runtime

We begin by describing the algorithm and bounding its runtime simultaneously. As in the proof of Theorem 3.4.3, we bound the runtime required to approximately compute only the k, ℓ-th entries of $\{\widehat{M}\}^c(D_n)$ satisfying $k \leq \ell, k\ell \leq B$. To do this, we analyze two portions of $\{\widehat{M}\}^c(D_n)$ separately: we first consider the case $\ell \geq \sqrt{B}$, in which we must compute the entries corresponding to all the pairs $\{(2, \ell), \ldots, (B/\ell, \ell)\}$. We then consider $\ell < \sqrt{B}$, in which case we need only compute the entries $\{(2, \ell), \ldots, (\ell, \ell)\}$ since the additional pairs would all have $k > \ell$.

For the case of $\ell \geq \sqrt{B}$, as in the previous theorem we can simultaneously compute using OptimizeXAxis the entries corresponding to all the pairs $\{(2, \ell), \ldots, (B/\ell, \ell)\}$

in time $O(|\Pi|^2(B/\ell)\ell) = O(|\Pi|^2 B)$, which equals $O(c^2B^3/\ell^2)$ when we set Π to be an equipartition of size cB/ℓ. Doing this for $\ell = \sqrt{B}, \ldots, B/2$ gives a contribution of the following order to the runtime.

$$O(c^2B^3) \sum_{\ell=\sqrt{B}}^{B/2} \frac{1}{\ell^2} = O\left(c^2B^3\right) O\left(\frac{1}{\sqrt{B}}\right)$$
$$= O(c^2B^{5/2})$$

For the case of $\ell < \sqrt{B}$, we can simultaneously compute using OptimizeXAxis the entries corresponding to all the pairs $\{(2,\ell), \ldots, (\ell,\ell)\}$ in time $O(|\Pi|^2\ell^2)$ which equals $O(c^2\ell^4) \leq O(c^2B^2)$ when we set Π to be an equipartition of size $c\ell$. Summing over the $O(\sqrt{B})$ possible values of ℓ with $\ell < \sqrt{B}$ gives an upper bound of $O(c^2B^{5/2})$.

B.8.2 Consistency

Let (X, Y) be a pair of jointly distributed random variables. For a sample D_n of size n from the distribution of (X, Y) and a speed-versus-optimality parameter $c \geq 1$, let $\{\widehat{M}\}^c(D_n)$ denote the matrix computed by EquicharClump. (Notice the use of curly braces to differentiate this from the sample equicharacteristic matrix $[\widehat{M}]$.) We show here that $\max_{k\ell \leq B(n)}\{\widehat{M}\}^c(D_n)_{k,\ell}$ is a consistent estimator of $\mathrm{MIC}_*(X, Y)$, and correspondingly that $\sum_{k\ell \leq B(n)}\{\widehat{M}\}^c(D_n)_{k,\ell}$ yields a consistent independence test.

The key to both consistency results is that, though in calculating the k, ℓ-th entry of

$\{\widehat{M}\}^c(D_n)$ the algorithm only searches for optimal partitions that are sub-partitions of some equipartition, the size of the equipartition used always grows as n, k, and ℓ grow large. Therefore, in the limit this additional restriction does not hinder the optimization. We present this argument by introducing a population object called the *clumped equicharacteristic matrix*. We observe that this matrix is the limit of the EquicharClump procedure as sample size grows, and then show that the supremum and partial sums of this matrix have the necessary properties.

Definition B.8.1. Let (X, Y) be jointly distributed random variables and fix some $c \geq 1$. Let

$$I^{\{c*\}}((X,Y),k,\ell) = \max_G I((X,Y)|_G)$$

where the maximum is over k-by-ℓ grids whose larger partition is an equipartition and whose smaller partition must be contained in an equipartition of size $c \cdot \max\{k, \ell\}$. The *clumped equicharacteristic matrix* of (X, Y), denoted by $\{M\}^c(X, Y)$, is defined by

$$\{M\}^c(X,Y)_{k,\ell} = \frac{I^{\{c*\}}((X,Y),k,\ell)}{\log\min\{k,\ell\}}$$

Notice that curly braces differentiate the quantities $I^{\{c*\}}$ and $\{M\}^c$ defined above from the corresponding equicharacteristic matrix quantities $I^{[*]}$ and $[M]$.

The following two results, which we state without proof, characterize the convergence of the output of EquicharClump to the clumped equicharacteristic matrix. These

lemmas can be shown using Lemma B.1.3, which simultaneously bounds the difference, for all k-by-ℓ grids G, between the sample quantity $I(D_n|_G)$ and the population quantity $I((X,Y)|_G)$ with high probability over the sample D_n of size n from (X,Y).

Lemma B.8.2. *Let D_n be a sample of size n from the distribution of a pair (X,Y) of jointly distributed random variables. For every $B(n) = O\left(n^{1-\varepsilon}\right)$, there exists an $a > 0$ such that for sufficiently large n,*

$$\left|\{\widehat{M}\}^c(D_n)_{k,\ell} - \{M\}^c(X,Y)_{k,\ell}\right| \leq O\left(\frac{1}{n^a}\right)$$

holds for all $k, \ell \leq \sqrt{B(n)}$ with probability $P(n) = 1 - o(1)$, where $\{\widehat{M}\}^c(D_n)$ denotes the matrix computed by the EquicharClump *algorithm with parameter c on the sample D_n.*

Notice that the error bound provided by the above lemma holds not for $k\ell \leq B(n)$ as in the analogous Lemma B.1.4 and Lemma B.7.1, but rather for the smaller region defined by $k, \ell \leq \sqrt{B(n)}$. However, though we do not have uniform convergence outside the region $k, \ell \leq \sqrt{B(n)}$, we do nevertheless have pointwise convergence there, as stated below.

Lemma B.8.3. *Fix $k, \ell \geq 2$. Let D_n be a sample of size n from the distribution of a pair (X,Y) of jointly distributed random variables. For every $B(n) > \omega(1)$, we have*

that

$$\{\widehat{M}\}^c(D_n)_{k,\ell} \to \{M\}^c(X,Y)_{k,\ell}$$

in probability as n grows, where $\{\widehat{M}\}^c(D_n)$ *denotes the matrix computed by the* EQUICHAR-CLUMP *algorithm with parameter c on the sample* D_n*.*

CONSISTENCY FOR ESTIMATING MIC_*

The consistency of $\{\widehat{M}\}^c(D_n)$ for estimating MIC_* follows from the following property of the clumped equicharacteristic matrix $\{M\}^c$, for which we state a proof sketch.

Proposition B.8.4. *Let* (X,Y) *be a pair of jointly distributed random variables. Then we have* $\sup\{M\}^c(X,Y) = MIC_*(X,Y)$*.*

Proof. (Sketch) Let $\{M\}^c = \{M\}^c(X,Y)$, and let $M = M(X,Y)$ be the characteristic matrix. Fix k, and consider the limit $\{M\}^c_{k,\ell}$ as ℓ grows. The grid chosen for the k,ℓ-th entry when $\ell > k$ will contain an equipartition P_ℓ of size ℓ on the x-axis, and a partition Q_ℓ of size k on the y-axis that is optimal subject to the restriction that Q_ℓ be contained in an equipartition of size $c\ell$. As ℓ grows large, the equipartition P_ℓ on the first axis will become finer and finer until in the limit $X|_{P_\ell} \to X$. And the partition Q_ℓ will be chosen from a finer and finer equipartition, so that in the limit it approaches an unconditionally optimal partition Q of size k. The convergence of Q_ℓ to the optimal partition Q of size

k can be shown to be uniform using Proposition B.2.2. This implies that

$$\{M\}^c_{k,\uparrow} = \lim_{\ell\to\infty} \{M\}^c_{k,\ell} = \max_{P\in P(k)} \frac{I(X,Y|_P)}{\log k}$$

where $P(k)$ denotes the set of all partitions of size at most k. Therefore, the boundary $\partial\{M\}^c$ of $\{M\}^c$ equals the boundary ∂M of M. Since $\mathrm{MIC}_*(X,Y) = \sup \partial M$ (Theorem 3.3.3), this implies that

$$\sup\{M\}^c \geq \sup \partial\{M\}^c = \sup \partial M = \mathrm{MIC}_*(X,Y).$$

On the other hand, $\{M\}^c \leq M$ element-wise since the optimization for the k,ℓ-th entry of $\{M\}^c$ is performed over a subset of the grids searched for the k,ℓ-th entry of M. This means that $\sup\{M\}^c \leq \sup M = \mathrm{MIC}_*(X,Y)$. □

This fact, together with the pointwise convergence of $\{\widehat{M}\}^c(D_n)$ to $\{M\}^c$, suffices to establish the consistency we seek via standard continuity arguments, which we give in the abstract lemma below. The lemma applies to a double-indexed sequence indexed by i and j; in our argument, the index i corresponds to position in the equicharacteristic matrix, and the index j corresponds to sample size. The sequence A corresponds to the output of the EquicharClump algorithm, the sequence a corresponds to the clumped equicharacteristic matrix, and the sequence B corresponds to the sample equicharacteristic matrix.

Lemma B.8.5. *Let $\{A_{ij}\}_{i,j=1}^{\infty}$ and $\{B_{ij}\}_{i,j=1}^{\infty}$ be sequences of random variables, and let $\{a_i\}_{i=1}^{\infty}$ be a non-stochastic sequence. Assume that the following conditions hold.*

1. *$A_{ij} \leq B_{ij}$ almost surely*
2. *For every i, $A_{ij} \to a_i$ in probability*
3. *$B'_j = \max_{i \leq j} B_{ij}$ satisfies $B'_j \to \sup\{a_i\}$ in probability*

Then $A'_j = \max_{i \leq j} A_{ij}$ converges in probability to $\sup\{a_i\}$ as well.

Proof. Let $a = \sup\{a_i\}$. We give the proof for the case that $a < \infty$. However, it is easily adapted to the infinite case. We must show that for every $\varepsilon > 0$ and every $0 < p \leq 1$, there exists some N such that $\mathbf{P}(|A'_j - a| < \varepsilon) > p$ for all $j \geq N$. By the definition of a, we know that there exists some k such that $|a_k - a| < \varepsilon/2$. Also, by the convergence of A_{kj} to a_k, there exists some m such that $\mathbf{P}(|A_{kj} - a_k| < \varepsilon/2) > 1 - p$ for all $j \geq m$. Thus, with probability at least $1 - p$, we have

$$\begin{aligned} |A_{kj} - a| &\leq |A_{kj} - a_k| + |a_k - a| \\ &\leq \varepsilon \end{aligned}$$

for all $j \geq m$.

Next, we observe that since $A'_j \geq A_{kj}$ for $j \geq k$, the above inequality implies that for $j \geq \max\{m, k\}$ we have $\mathbf{P}(A'_j > a - \varepsilon) > 1 - p$. It remains only to show that A'_j doesn't get too large, but this follows from the fact that $A'_j \leq B'_j$ and $B'_j \to a$ in probability. Specifically, we are guaranteed some $N \geq \max\{m, k\}$ such that $\mathbf{P}(B'_j < a + \varepsilon) > 1 - p$

for $j \geq N$. Since $B'_j < a + \varepsilon$ implies $A'_j < a + \varepsilon$, we have that $\mathbf{P}(|A'_j - a| < \varepsilon) > 1 - p$ for $j \geq N$, as desired. $\square$

Proposition B.8.6. *The function*

$$\widetilde{MIC}_{e,B}(\cdot) = \max_{k\ell \leq B(n)} \{\widehat{M}\}^c(\cdot)_{k,\ell}$$

is a consistent estimator of MIC$_*$ provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $\varepsilon > 0$, where $\{\widehat{M}\}^c(\cdot)$ is the output of the the EquicharClump *algorithm.*

Proof. Let (X, Y) be a pair of jointly distributed random variables, and let D_n be a sample of size n from the distribution of (X, Y). Let $\{(k_i, \ell_i)\}_{i=1}^{\infty} \subset \mathbb{Z}^+ \times \mathbb{Z}^+$ be a sequence of coordinates with the property that for every number B there exists an index $q(B)$ such that

$$\{(k_i, \ell_i) : i \leq q(B)\} = \{(k, \ell) : k\ell \leq B\}.$$

We define $B_{ij} = \widehat{[M]}(D_j)_{k_i,\ell_i}$, i.e., B_{ij} is the k_i, ℓ_i-th entry of the sample characteristic matrix evaluated on a sample of size j. We analogously define $A_{ij} = \{\widehat{M}\}^c(D_j)_{k_i,\ell_i}$, and we define $a_i = \{M\}^c(X, Y)_{k_i,\ell_i}$. We observe that by Proposition B.8.4, $\sup a_i = \sup\{M\}^c(X, Y) = \text{MIC}_*$.

It is straightforward to see that $A_{ij} \leq B_{ij}$. Additionally, Lemma B.8.3 shows that $A_{ij} \to a_i$ in probability, and Corollary 3.4.6, which states that MIC$_e$ is a consistent

estimator of MIC_*, shows that $B'_j = \max_{i \leq j} B_{ij} \to \mathrm{MIC}_*(X,Y)$. In the notation of the lemma, it therefore follows that $A'_j = \max_{i \leq j} A_{ij}$ converges in probability to $\mathrm{MIC}_*(X,Y)$ as well. But this means that the sub-sequence

$$A'_{q(B(n))} = \max_{i \leq q(B(n))} \{\widehat{M}\}^c(D_{q(B(n))})_{k_i,\ell_i} = \max_{k\ell \leq B(n)} \{\widehat{M}\}^c(D_{q(B(n))})_{k,\ell}$$

converges in probability to $\mathrm{MIC}_*(X,Y)$, which implies the result since the sequence A'_j is monotone. □

Consistency for total information coefficient

Similarly to the consistency argument for MIC_*, we begin by exhibiting the relevant property of the population clumped equicharacteristic matrix.

Proposition B.8.7. *Let (X,Y) be a pair of jointly distributed random variables. If X and Y are statistically independent, then $\{M\}^c(X,Y) \equiv 0$. If not, then there exists some $a > 0$ and some integer $\ell_0 \geq 2$ such that*

$$\{M\}^c(X,Y)_{k,\ell} \geq \frac{a}{\log \min\{k,\ell\}}$$

either for all $k \geq \ell \geq \ell_0$, or for all $\ell \geq k \geq \ell_0$.

Proof. (Sketch) Let $\{M\}^c = \{\widehat{M}\}^c(X,Y)$. Under independence, every entry of $\{M\}^c$ is zero since $I((X,Y)|_G) = 0$ for any grid G. For the case of dependence, the argument

is identical to that given in the proof of Proposition 3.5.3. Specifically, it can be shown that there exists some index ℓ_0, taken without loss of generality to be a column index, and some $r > 0$ such that all but finitely many of the entries in the ℓ_0-column are at least r. It can then be shown that for large k, the entries $(k, \ell_0), (k, \ell_0 + 1), \ldots, (k, k)$ have non-decreasing values of $I^{[c*]}$. This establishes the claim for $a = r \log \ell_0$. □

We now show that the above result, together with the uniform convergence of $\{\widehat{M}\}^c(D_n)$ to $\{M\}^c(X, Y)$, implies the consistency we seek.

Proposition B.8.8. *The function*

$$\widetilde{TIC}_{e,B}(\cdot) = \sum_{k\ell \leq B(n)} \{\widehat{M}\}^c(\cdot)_{k,\ell}$$

yields a consistent right-tailed test of independence provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ *for some* $\varepsilon > 0$*, where* $\{\widehat{M}\}^c(\cdot)$ *is the output of the the* EquicharClump *algorithm.*

Proof. Let (X, Y) a pair of jointly distributed random variables, and let D_n be a sample of size n from the distribution of (X, Y). It suffices to show consistency for any deterministic monotonic function of the statistic in question. We therefore choose to analyze $\widetilde{\mathrm{TIC}}_{e,B}(D_n) \log(B(n))/B(n)$.

For the null hypothesis in which X and Y are independent, we observe that since $\{\widehat{M}\}^c(D_n) \leq \widehat{[M]}(D_n)$ element-wise, $0 \leq \widetilde{\mathrm{TIC}}_{e,B}(D_n) \leq \mathrm{TIC}_{e,B}(D_n)$ as well. Moreover, the argument given in Section B.11, which shows that $\mathrm{TIC}_{e,B}(D_n)/B(n)$ con-

verges to 0 in probability under the null hypothesis, can be adapted to show that $\mathrm{TIC}_{e,B}(D_n)\log(B(n))/B(n) \to 0$ as well. Thus, $\widetilde{\mathrm{TIC}}_{e,B}(D_n)\log(B(n))/B(n)$ converges to zero in probability, as required.

For the case that X and Y are dependent, the proof is analogous to the argument given in Section B.11 for TIC_e. The only difference is that Lemma B.8.2, which guarantees the uniform convergence of $\{\widehat{M}\}^c(D_n)$ to $\{M\}^c(X,Y)$, applies only to the k,ℓ-th entries for which $k,\ell \leq \sqrt{B(n)}$, rather than the entries over which we are summing, which are those for which $k\ell \leq B(n)$. However, since we require only a lower bound on $\widetilde{\mathrm{TIC}}_{e,B}(D_n)$, we may neglect these entries because

$$\widetilde{\mathrm{TIC}}_{e,B}(D_n) = \sum_{k\ell \leq B(n)} \{\widehat{M}\}^c(D_n)_{k,\ell} \geq \sum_{k,\ell \leq \sqrt{B(n)}} \{\widehat{M}\}^c(D_n)_{k,\ell}.$$

It can then be shown, following the argument from Section B.11, that there exists some $a > 0$ depending only on B such that, with probability $1 - o(1)$,

$$\frac{\log B(n)}{B(n)}\left(\sum_{k,\ell \leq \sqrt{B(n)}} \{\widehat{M}\}^c(X,Y)_{k,\ell} - \widetilde{\mathrm{TIC}}_{e,B}(D_n)\right) \leq O\left(\frac{\#_n \log B(n)}{B(n)n^a}\right) = O\left(\frac{\log B(n)}{n^a}\right)$$

where $\#_n = B(n)$ represents the number of pairs (k,ℓ) such that $k,\ell \leq \sqrt{B(n)}$. To obtain the result, we note that this means that

$$\frac{\log B(n)}{B(n)}\widetilde{\mathrm{TIC}}_{e,B}(D_n) \geq \frac{\log B(n)}{B(n)} \sum_{k,\ell \leq \sqrt{B(n)}} \{\widehat{M}\}^c(X,Y)_{k,\ell} - O\left(\frac{\log B(n)}{n^a}\right)$$

and then invoke Proposition B.8.7, which implies that for large n

$$\sum_{k,\ell \leq \sqrt{B(n)}} \{M\}^c(X,Y) \geq \Omega\left(\frac{B(n)}{\log B(n)}\right).$$

□

B.8.3 Empirical characterization of the performance of EquicharClump

The EquicharClump algorithm has a parameter c that controls the fineness of the equipartition whose sub-partitions are searched over by the algorithm. To gain an empirical understanding of the effect of c on performance, we computed MIC_e on the set of relationships described in Section 3.4.4 using EquicharClump with different values of c. For each relationship, we compared the average MIC_e across all 500 independent samples from that relationship with different values of c. We performed this analysis at sample sizes of $n = 250$ (Figures B.1 and B.2), $n = 500$ (Figures B.3 and B.3), and $5,000$ (Figures B.5 and B.6).

We summarize our findings as follows.

- At low ($n = 250$) and medium ($n = 500$) sample sizes, using $c = 1$ introduces a downward bias for more complex relationships when $B(n) = n^{0.6}$ is used but not when $B(n) = n^{0.8}$ is used. This makes sense since the low sample size and low setting of $B(n)$ mean that the algorithm is searching over grids with relatively few cells, and so setting $c = 1$ hinders its ability to find good grids in this limited search space. This bias is almost entirely alleviated by setting $c \geq 2$.

- At high sample size ($n = 5,000$), this effect is still observable but much reduced. This makes sense since when n is large, $B(n)$ is large as well, and so the number of cells allowed in the grids being searched over is already large regardless of the exponent α used in $B(n) = n^{\alpha}$. Thus, there is less need for the robustness provided by searching for an optimal grid.

B.9 Sample equitability and power analyses

Figure B.7 contains a representative equitability analysis from Reshef et al.[125]. Figure B.8 contains power curves from Reshef et al.[125] for a large set of leading methods.

B.10 Equitability analysis of randomly chosen functions with additional noise model

Figure B.9 contains a version of the main text Figure 3.4, but where noise has been added only to the dependent variable in each functional relationship, rather to both the independent and dependent variables.

B.11 Consistency of independence testing based on TIC_e

Here we prove Propositions 3.5.3 and 3.5.4 and then use those propositions to prove Theorem 3.5.1, which shows that TIC_e can be used for independence testing.

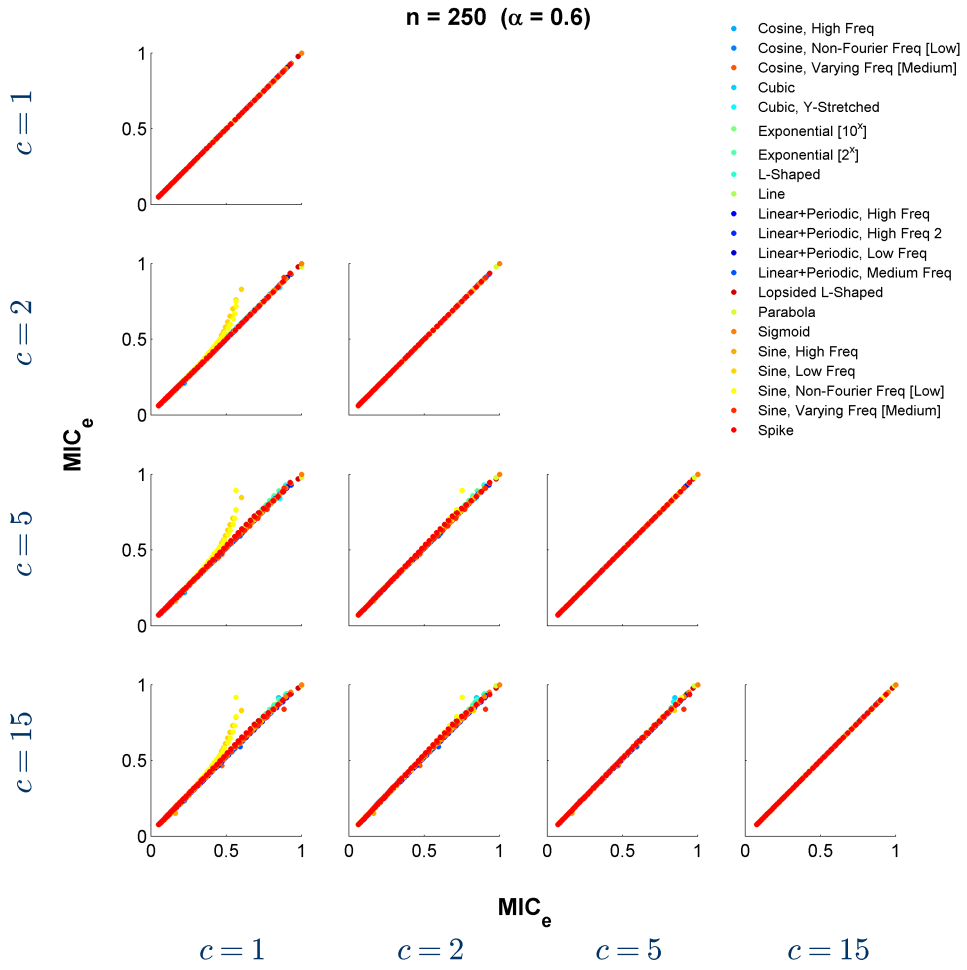


Figure B.1: The effect of the parameter c on the performance of EquicharClump, at $n = 250$ with $\alpha = 0.6$. See Section B.8.3 for details.

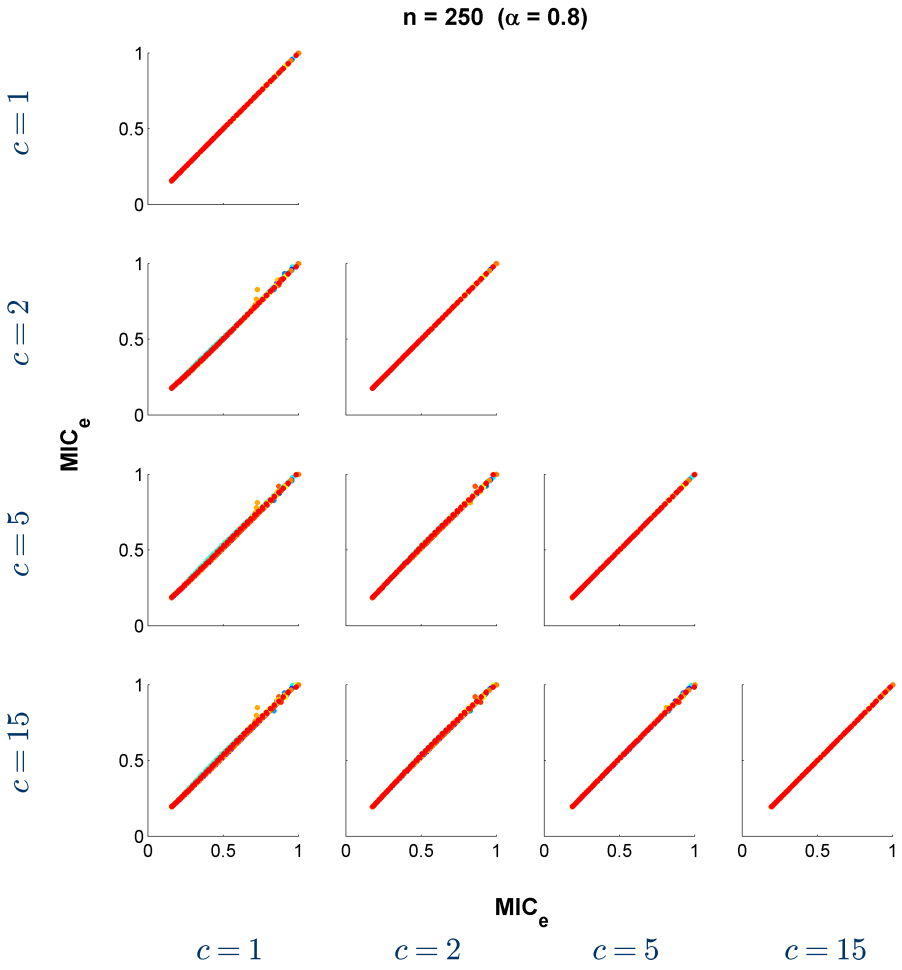


Figure B.2: The effect of the parameter c on the performance of EquicharClump, at $n = 250$ with $\alpha = 0.8$. See Section B.8.3 for details.

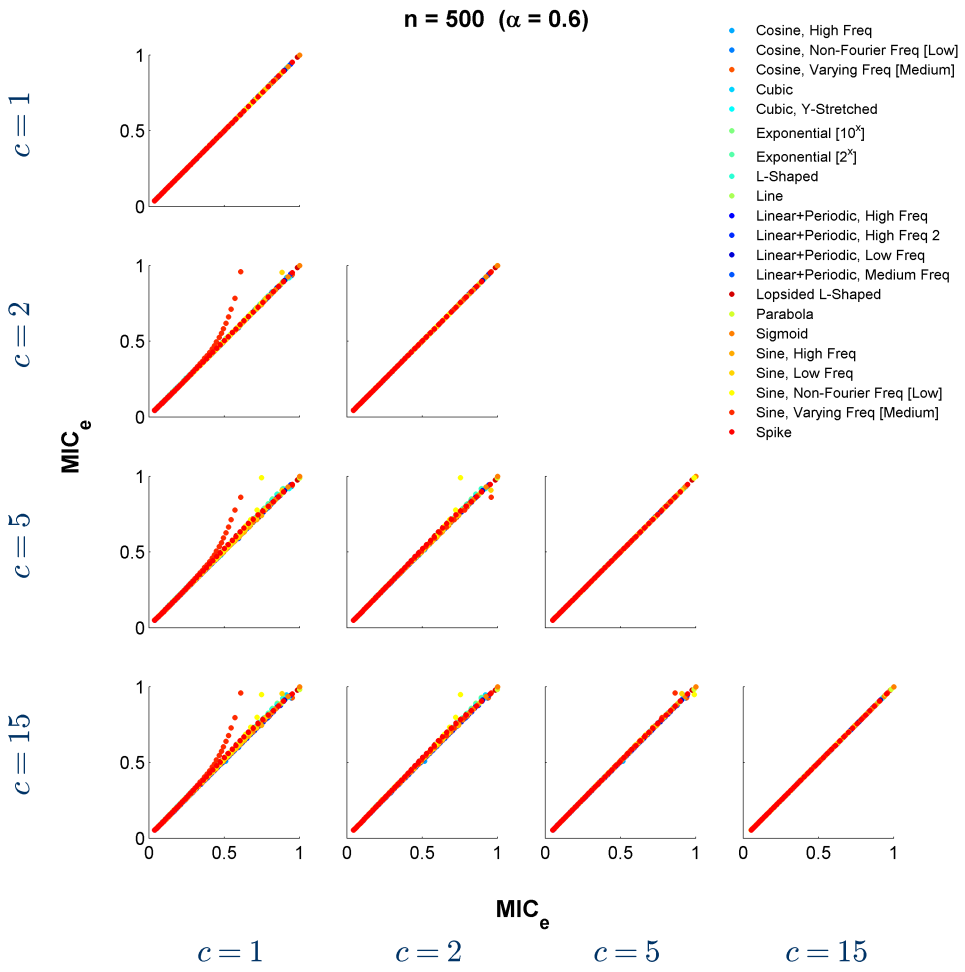


Figure B.3: The effect of the parameter c on the performance of EquicharClump, at $n = 500$ with $\alpha = 0.6$. See Section B.8.3 for details.

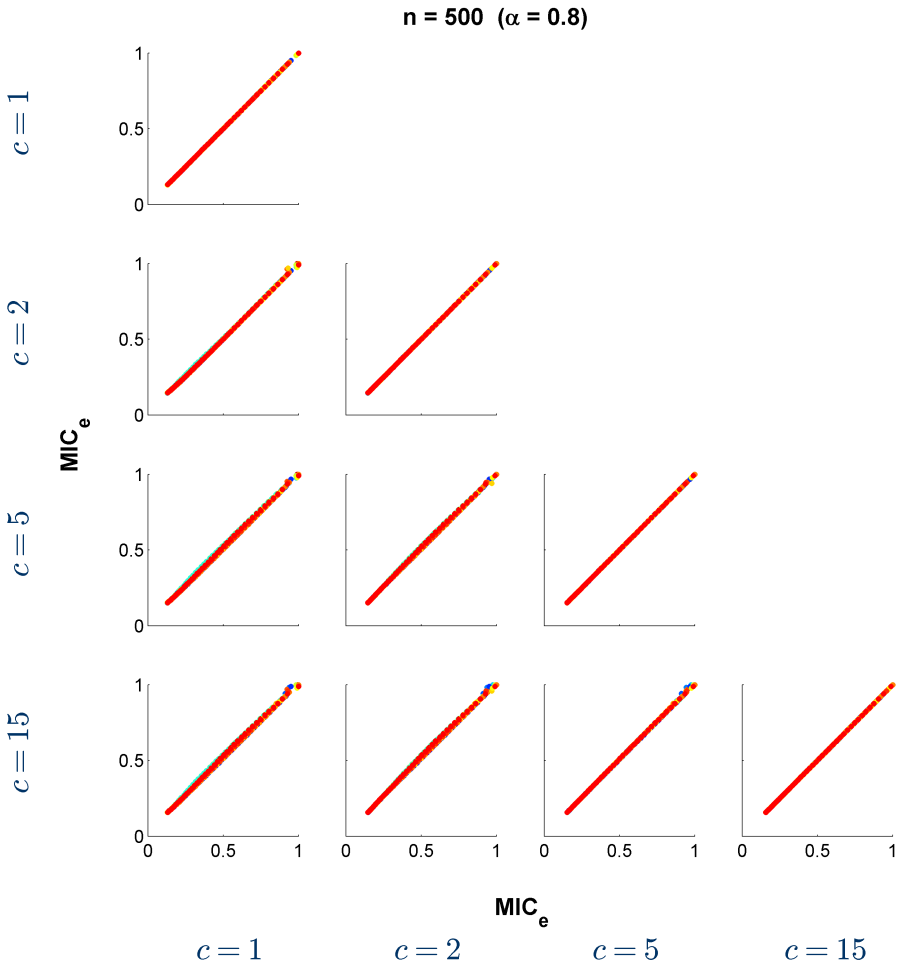


Figure B.4: The effect of the parameter c on the performance of EquicharClump, at $n = 500$ with $\alpha = 0.8$. See Section B.8.3 for details.

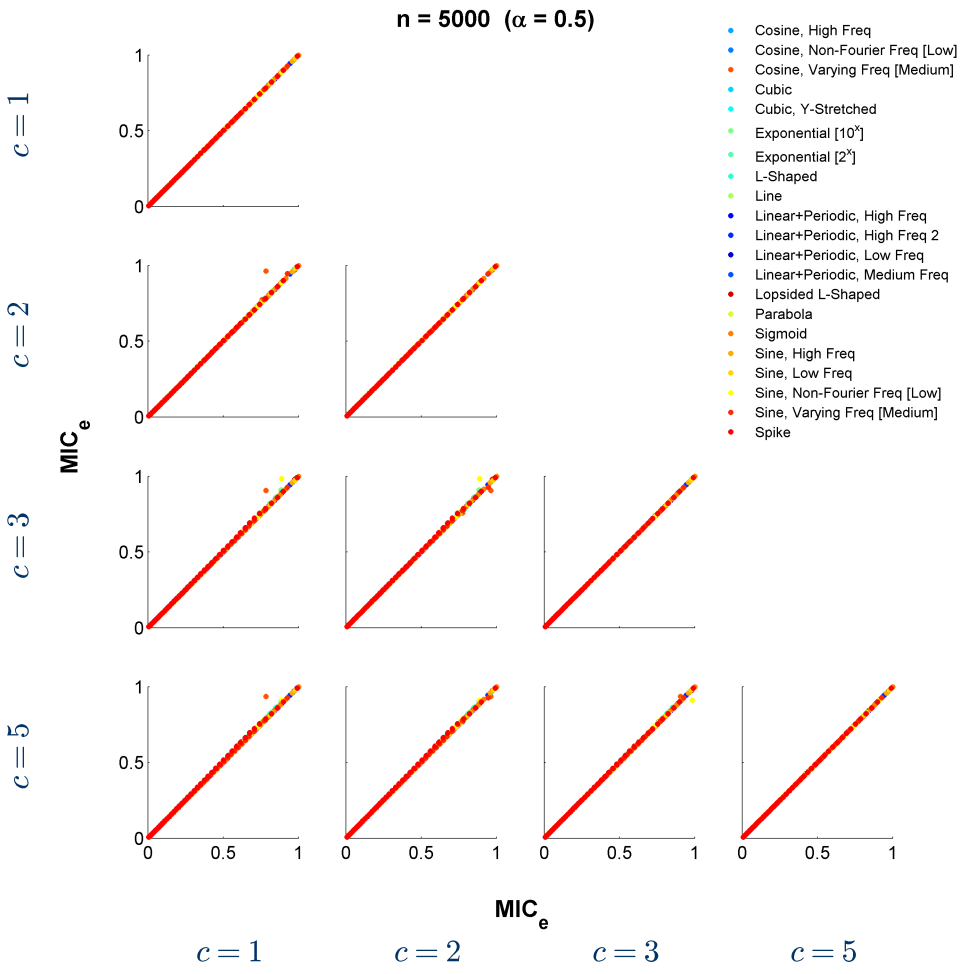


Figure B.5: The effect of the parameter c on the performance of EquicharClump, at $n = 5,000$ with $\alpha = 0.5$. See Section B.8.3 for details.

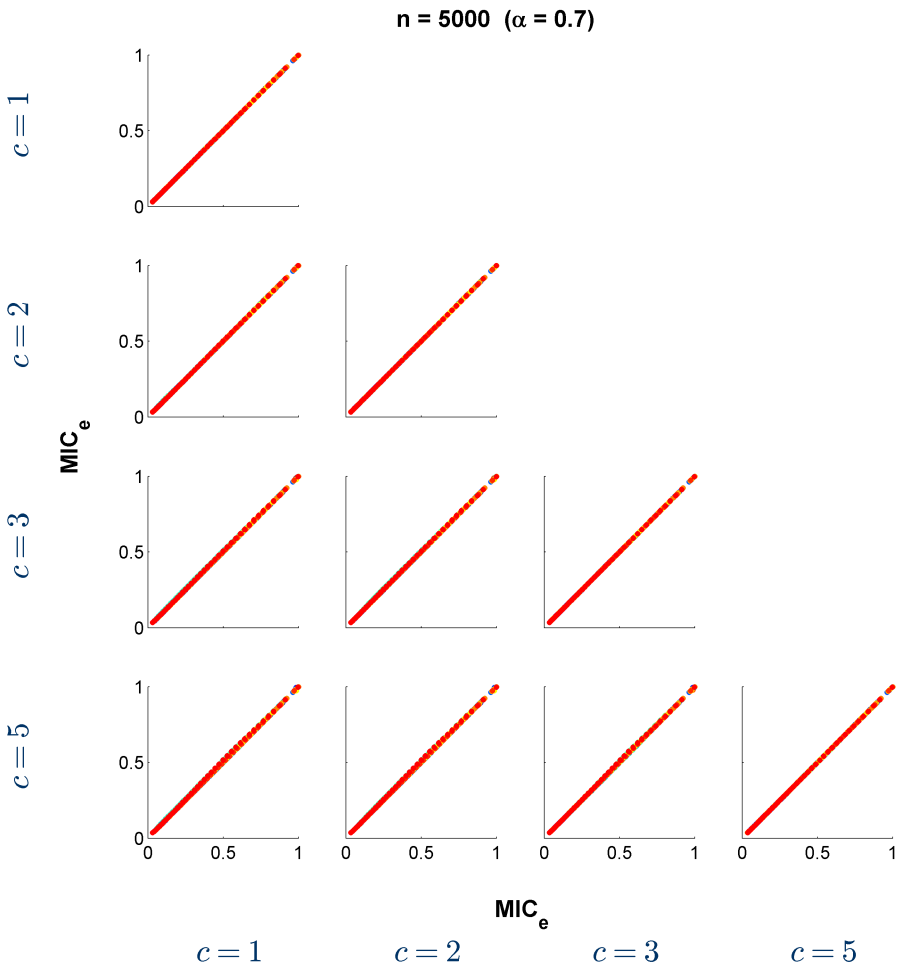


Figure B.6: The effect of the parameter c on the performance of EquicharClump, at $n = 5,000$ with $\alpha = 0.7$. See Section B.8.3 for details.

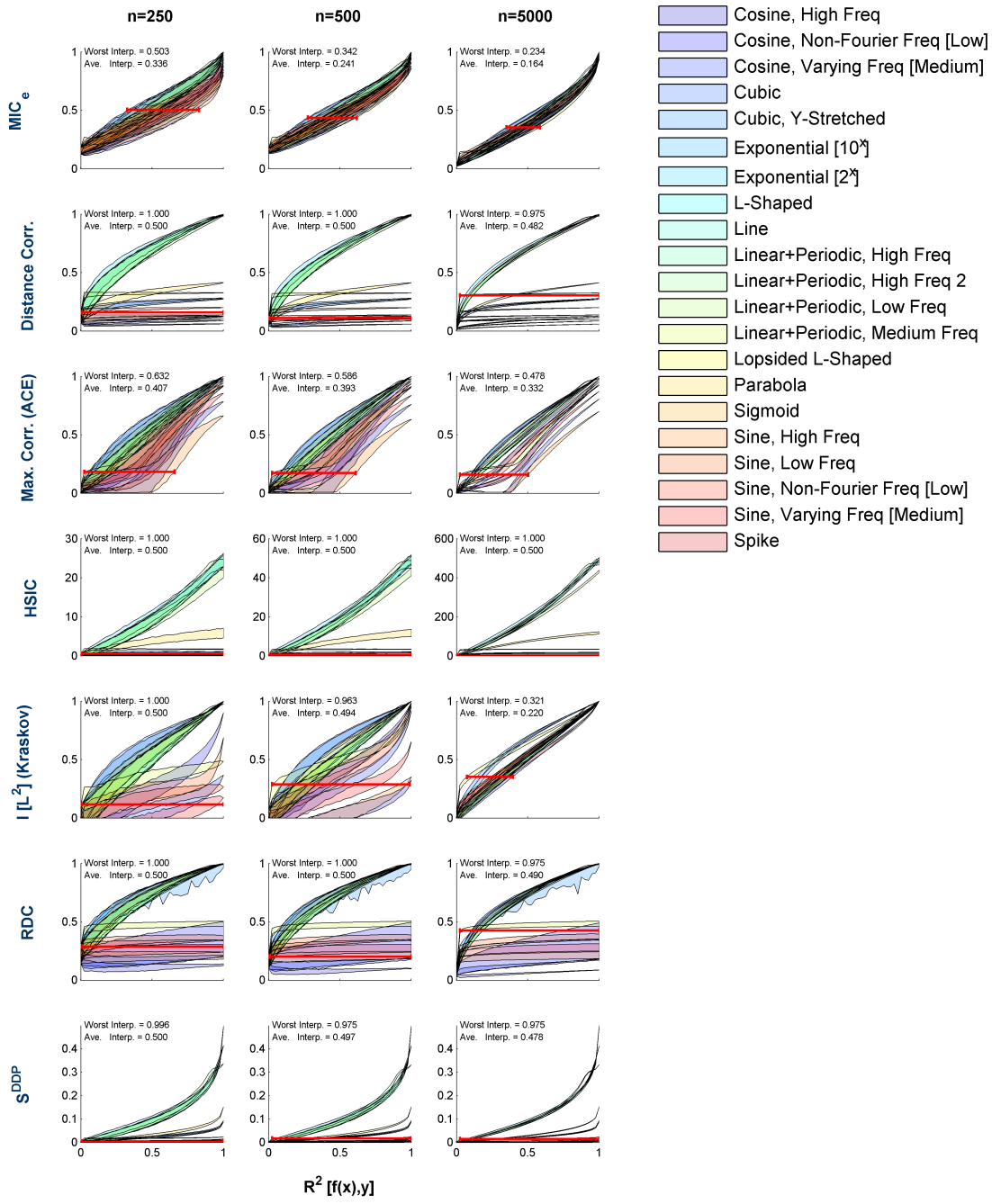


Figure B.7: The equitability of measures of dependence on a set of noisy functional relationships, reproduced from Reshef et al.[125]. *[Narrower is more equitable.]* The plots were constructed as in Figure 3.3. Mutual information, estimated using the Kraskov estimator, is represented using the squared Linfoot correlation.

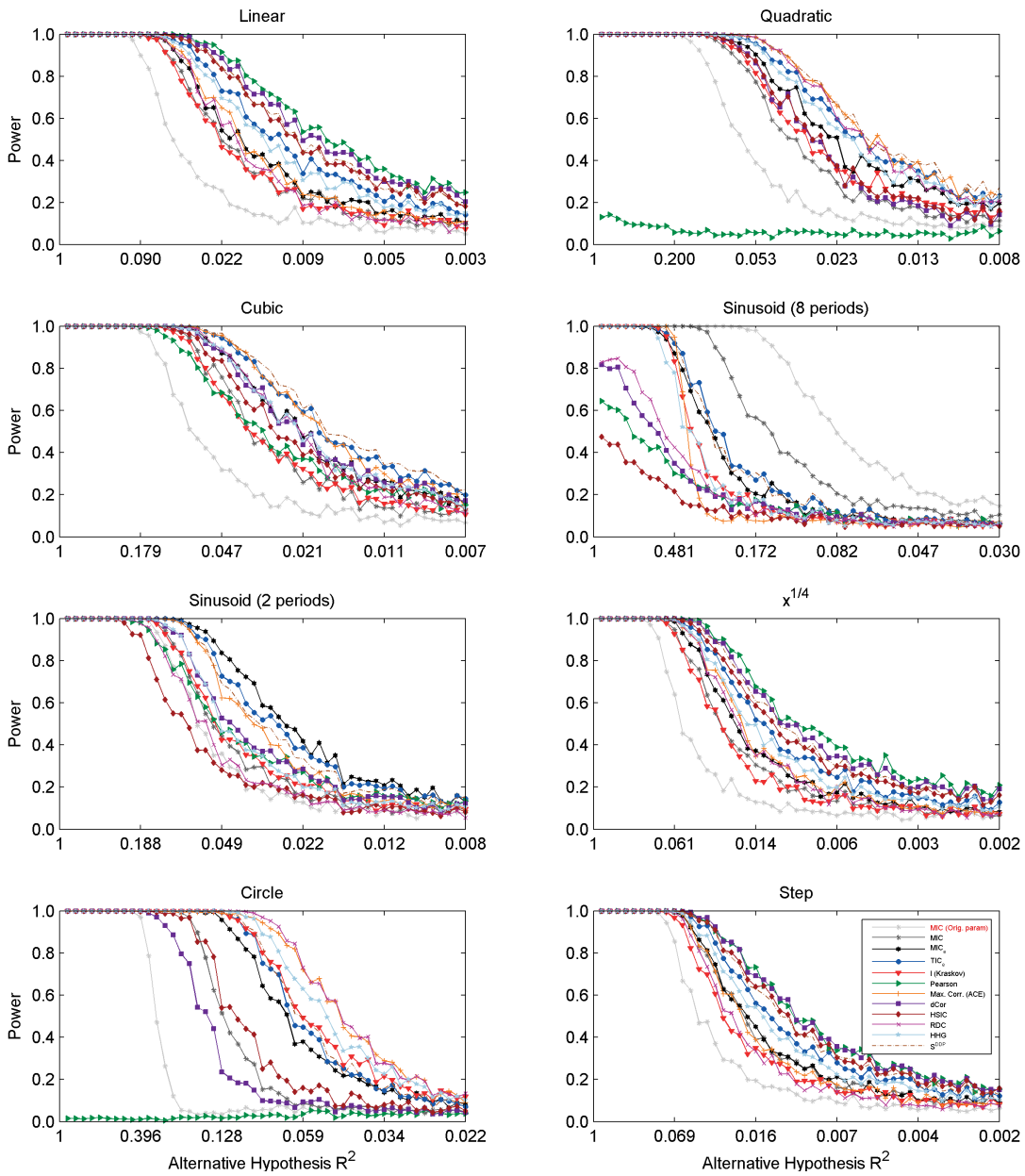


Figure B.8: Power of independence testing using several leading measures of dependence, on the relationships chosen by Simon & Tibshirani[144], at 50 noise levels with linearly increasing magnitude for each relationship and $n = 500$. To enable comparison of power regimes across relationships, the x-axis of each plot lists R^2 rather than noise magnitude.

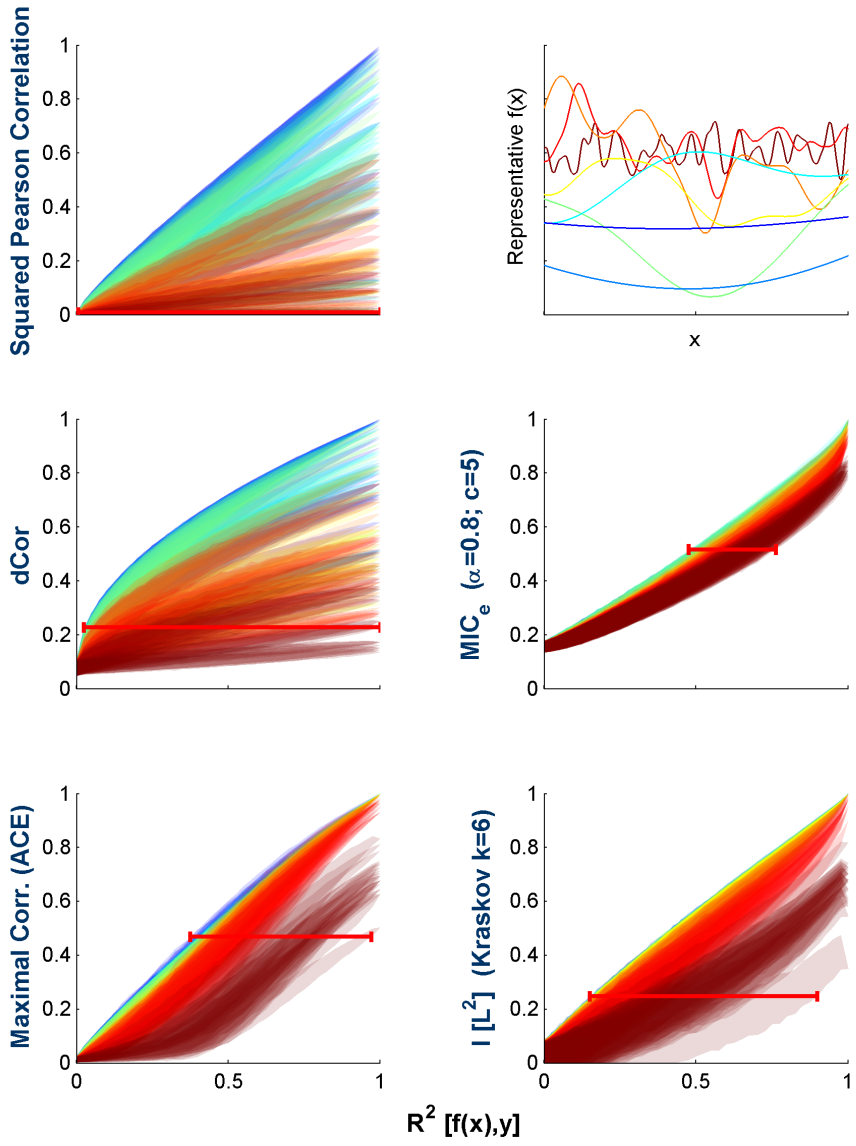


Figure B.9: Equitability of methods examined on functions randomly drawn from a Gaussian process distribution, using a different noise model. This figure is identical to Figure 3.4, but with noise added only to the dependent variable in each relationship. Each method is assessed as in Figure 3.4, with a red interval indicating the widest range of R^2 values corresponding to any one value of the statistic; the narrower the red interval, the higher the equitability. Sample relationships for each Gaussian process bandwidth are shown in the top right with matching colors.

B.11.1 Proof of Proposition 3.5.3

Proposition *Let (X, Y) be a pair of jointly distributed random variables. If X and Y are statistically independent, then $M(X, Y) \equiv [M](X, Y) \equiv 0$. If not, then there exists some $a > 0$ and some integer $\ell_0 \geq 2$ such that*

$$M(X, Y)_{k,\ell}, [M](X, Y)_{k,\ell} \geq \frac{a}{\log \min\{k, \ell\}}$$

either for all $k \geq \ell \geq \ell_0$, or for all $\ell \geq k \geq \ell_0$.

Proof. We give the proof only for $[M] = [M](X, Y)$, with the understanding that all parts of the argument are either identical or similar for $M(X, Y)$. When X and Y are independent, then for any grid G, $(X, Y)|_G$ exhibits independence as well. Therefore $I((X, Y)|_G) = 0$ for all grids G, and so every entry of $[M]$, being a supremum over such quantities, is 0.

For the case that X and Y are dependent, our strategy is to first find, without loss of generality, a column of $[M]$ almost all of whose values are bounded away from zero, and then argue that this suffices.

The dependence of X and Y implies that $\mathrm{MIC}_*(X, Y) > 0$. By Corollary 3.4.3, which states that $\sup \partial[M] = \mathrm{MIC}_*(X, Y)$, we therefore know that there is at least one non-zero element of the boundary of $[M]$, as defined in Definition 3.3.12. Without loss of

generality, suppose that this element is $[M]_{\uparrow,\ell_0} = \lim_{k\to\infty}[M]_{k,\ell_0}$. The fact that this limit is strictly positive implies that there exists some $k_0 \geq \ell_0$ and some $r > 0$ such that $[M]_{k,\ell_0} \geq r$ for all $k \geq k_0$. That is, all but finitely many of the entries in the ℓ_0-th column of $[M]$ are at least r.

We now show that the existence of such a column suffices to prove the claim. Fix some $k > k_0$ and note that this implies that $k > \ell_0$. We argue that for all ℓ in $\{\ell_0, \ldots, k\}$, the desired condition holds. Since $k > \ell_0$, the term $I^{[*]}((X,Y),k,\ell_0)$ in the definition of $[M]_{k,\ell_0}$ is a maximization over grids that have an equipartition of size k on one axis and an optimal partition of size ℓ_0 on the other. Since we allow empty rows/columns in the maximization, substituting any ℓ satisfying $\ell_0 \leq \ell \leq k$ therefore does not constrain the maximization in any way and so it cannot decrease $I^{[*]}$. In other words, for ℓ satisfying $\ell_0 \leq \ell \leq k$, we have

$$I^{[*]}((X,Y),k,\ell) \geq I^{[*]}((X,Y),k,\ell_0).$$

Since $k \geq \ell, \ell_0$, the normalizations in the definition of $[M]_{k,\ell}$ and $[M]_{k,\ell_0}$ are $\log \ell$ and $\log \ell_0$ respectively. Therefore, we have that

$$[M]_{k,\ell} \geq [M]_{k,\ell_0} \frac{\log \ell_0}{\log \ell} \geq \frac{r \log \ell_0}{\log \ell}$$

where the last inequality is because $k > k_0$. Setting $a = r \log \ell_0$ then gives the result. $\square$

B.11.2 Proof of Proposition 3.5.4

Proposition *Let (X, Y) be a pair of jointly distributed random variables. If X and Y are statistically independent, then $S_B(M(X,Y)) = S_B([M](X,Y)) = 0$ for all $B > 0$. If not, then $S_B(M(X,Y))$ and $S_B([M](X,Y))$ are both $\Omega(B \log\log B)$.*

Proof. We give the argument for $M = M(X,Y)$ only, but the argument holds as stated for $[M](X,Y)$ as well.

The result follows from the guarantee given by the Proposition 3.5.3 above. In the case of independence, the proposition tells us that $M \equiv 0$, which immediately gives that $S_B(M) = 0$ for all $B > 0$. For the case of dependence, the proposition implies that there is some $a > 0$ and some integer $\ell_0 \geq 2$ such that, without loss of generality, $M_{k,\ell} \geq a/\log \ell$ for all $k \geq \ell \geq \ell_0$. We convert this into a lower bound on $S_B(M)$.

The key is to write the sum one column at a time, counting how many entries in each column both satisfy $k \geq \ell \geq \ell_0$ and $k\ell \leq B$. For any ℓ satisfying $\ell_0 \leq \ell \leq \sqrt{B}$, the entries $(\ell, \ell), \ldots, (B/\ell, \ell)$ meet this criterion, and there are $B/\ell_0 - (\ell_0 - 1)$ of them. Moreover, since the guarantee of Proposition 3.5.3 tells us that all of these entries are

at least $a/\log \ell$, we can lower-bound $S_B(M)$ as follows.

$$\begin{aligned} S_B(A) &\geq \sum_{\ell=\ell_0}^{\sqrt{B}} \frac{a}{\log \ell} \left(\frac{B}{\ell} - (\ell - 1) \right) \\ &= aB \sum_{\ell=\ell_0}^{\sqrt{B}} \frac{1}{\ell \log \ell} - a \sum_{\ell=\ell_0}^{\sqrt{B}} \frac{\ell - 1}{\log \ell} \\ &= a \left(B \sum_{\ell=\ell_0}^{\sqrt{B}} \frac{1}{\ell \log \ell} - O(B) \right) \\ &= \Omega(B \log \log B) \end{aligned}$$

where the second-to-last equality is because $(\ell - 1)/\log \ell \leq \ell$, and the last equality is because $\sum_{i=i_0}^{n} 1/(i \log i)$ grows like $\log \log n$. □

B.11.3 Proof of Theorem 3.5.1

Theorem *The statistics TIC_B and $TIC_{e,B}$ yield consistent right-tailed tests of independence, provided $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $\varepsilon > 0$.*

Proof. We give the proof for TIC only; however, the argument holds as stated for TIC_e as well.

Let (X, Y) be jointly distributed random variables, and let D_n be a sample of size n from the distribution of (X, Y). Let $M = M(X, Y)$ be the characteristic matrix of (X, Y) and let $\widehat{M}(D_n)$ be the sample characteristic matrix. It suffices to establish the result for a deterministic monotonic function of $\text{TIC}_B(D_n)$. We therefore show

convergence of $\mathrm{TIC}_B(D_n)/B(n)$ to zero under the null hypothesis of independence and to ∞ under any alternative. Our general strategy for doing so is to translate known bounds on our error at estimating entries of M into bounds on the difference between $\mathrm{TIC}_B(D_n)/B(n) = S_{B(n)}(\widehat{M}(D_n))/B(n)$ and $S_B(M)/B(n)$. We then obtain the result by invoking Proposition 3.5.4, which implies that $S_B(M)/B(n)$ is zero under the null hypothesis but grows without bound under the alternative.

We know from Lemma B.1.4 (Lemma B.7.1 for the equicharacteristic matrix) that there exists some $a > 0$ depending only on B such that

$$\left|\widehat{M}(D_n)_{k,\ell} - M_{k,\ell}\right| \leq O\left(\frac{1}{n^a}\right)$$

for all $k\ell \leq B(n)$ with probability $1 - o(1)$. This means that with probability $1 - o(1)$ we have

$$\frac{1}{B(n)}\left|\mathrm{TIC}_B(D_n) - S_{B(n)}(M)\right| \leq O\left(\frac{\#_n}{B(n)n^a}\right)$$

where $\#_n$ is the number of pairs (k, ℓ) such that $k\ell \leq B(n)$. It can be shown by taking the integral of B/x with respect to x that $\#_n = O(B(n)\log B(n))$. Therefore, the error in the above bound is at most $O(\log B(n)/n^a) = O(1/\mathrm{poly}(n))$ for our choice of $B(n)$.

We now use Proposition 3.5.4 to show that this bound gives the desired result. Under the null hypothesis of independence, the proposition says that $S_{B(n)}(M) = 0$ always, and so since B is a growing function the bound implies that $\mathrm{TIC}_B(D_n)/B(n) \to 0$ in

probability. Under the alternative hypothesis in which (X, Y) exhibit a dependence, the proposition implies that $S_{B(n)}(M)/B(n) > \omega(1)$. Since B is a growing function of n, this means that for any $r > 0$, the probability that $S_{B(n)}(M)/B(n) > r$ goes to 1 as n grows. In other words, $\mathrm{TIC}_B(D_n)/B(n) \to \infty$ in probability. □

B.12 Information-theoretic lemmas

Lemma B.12.1. *Let Π and Ψ be random variables distributed over a discrete set of states Γ, and let (π_i) and (ψ_i) be their respective distributions. Let $P = f(\Pi)$ and $Q = f(\Psi)$ for some function f whose image is of size B. Define*

$$\varepsilon_i = \frac{\psi_i - \pi_i}{\pi_i}.$$

Then for every $0 < a < 1$ there exists some $A > 0$ such that

$$|H(Q) - H(P)| \leq (\log B)\, A \sum_i |\varepsilon_i|$$

when $|\varepsilon_i| \leq 1 - a$ for all i.

Proof. We prove the claim with entropy measured in nats. A rescaling then gives the general result.

Let (p_i) and (q_i) be the distributions of P and Q respectively, and define

$$e_i = \frac{q_i - p_i}{p_i}$$

analogously to ε_i. Before proceeding, we observe that

$$e_i = \sum_{j \in f^{-1}(i)} \frac{\pi_j}{p_i} \varepsilon_j.$$

We now proceed with the argument. We have from Roulston [136] that

$$\begin{aligned} |H(Q) - H(P)| &\leq \left| \sum_i \left(e_i p_i (1 + \ln p_i) + \frac{1}{2} e_i^2 p_i + O\left(e_i^3\right) \right) \right| && \text{(B.3)} \\ &\leq \left| \sum_i e_i p_i \right| + \left| \sum_i e_i p_i \ln p_i \right| + \frac{1}{2} \left| \sum_i e_i^2 p_i \right| + \left| \sum_i O\left(e_i^3\right) \right| && \text{(B.4)} \\ &= \left| \sum_i e_i p_i \ln p_i \right| + \frac{1}{2} \sum_i e_i^2 p_i + \left| \sum_i O\left(e_i^3\right) \right| && \text{(B.5)} \end{aligned}$$

where the final equality is because $\sum_i e_i p_i = \sum_i q_i - \sum_i p_i = 0$. We proceed by bounding each of the terms in Equation B.5 separately.

To bound the first term, we write

$$\left| \sum_i e_i p_i \ln p_i \right| \leq - \sum_i |e_i| p_i \ln p_i.$$

We then note that $-\sum_i p_i \ln p_i \leq \ln B$, and since each of the summands has the same

sign this means that $-p_i \ln p_i \leq \ln B$. We also observe that

$$|e_i| \leq \left| \sum_{j \in f^{-1}(i)} \frac{\pi_j}{p_i} \varepsilon_j \right| \leq \sum_j \frac{\pi_j}{p_i} |\varepsilon_j| \leq \sum_j |\varepsilon_j|$$

since $\pi_j / p_i \leq 1$. Together, these two facts give

$$\begin{aligned} -\sum_i |e_i| p_i \ln p_i &\leq (\ln B) \sum_i |e_i| \\ &\leq (\ln B) \sum_i |\varepsilon_i| \end{aligned}$$

The second inequality is because each e_i is a weighted average of a set of ε_i and each ε_i enters into the expression of exactly one e_i.

To bound the second term, we use the fact that $p_i \leq 1$ for all i, and so

$$\sum_i e_i^2 p_i \leq \sum_i e_i^2.$$

We then write

$$\begin{aligned} \sum_i e_i^2 &= \sum_i \left(\sum_{j \in f^{-1}(i)} \frac{\pi_j}{p_i} \varepsilon_j \right)^2 \\ &\leq \sum_i \sum_{j \in f^{-1}(i)} \frac{\pi_j}{p_i} \varepsilon_j^2 \\ &\leq \sum_j \varepsilon_j^2 \\ &= \sum_j O\left(|\varepsilon_j|\right) \end{aligned}$$

where the second line is a consequence of the convexity of $f(x) = x^2$ and the third line is because the sets $f^{-1}(i)$ partition Γ.

To bound the third term, we write

$$\left| \sum_i O\left(e_i^3\right) \right| \leq \sum_i O\left(|e_i|^3\right)$$

and then proceed as we did with the second term, using the fact that $f(x) = x^3$ is convex for $x \geq 0$. This gives

$$\sum_i O\left(|e_i|^3\right) \leq \sum_i O\left(|\varepsilon_i|^3\right) = \sum_i O\left(|\varepsilon_i|\right)$$

completing the proof. □

Lemma B.12.2. *Let $\{w_i\} \subset [0,1]$ be a set of size n with $\sum_i w_i \leq 1$, and let $\{u_i\}$ be a set of n non-negative numbers satisfying $\sum_i u_i = a$ and $u_i \leq w_i$. Then*

$$\sum_{i=1}^{n} w_i H_b\left(\frac{u_i}{w_i}\right) \leq H_b\left(a\right)$$

where H_b is the binary entropy function.

Proof. Consider the random variable X taking values in $\{0, \ldots, n\}$ that equals zero with probability $1 - \sum_i w_i$ and equals i with probability w_i for $0 < i \leq n$. Define the random

variable Y taking values in $\{0,1\}$ by

$$\mathbf{P}\left(Y=0|X=i\right)=\begin{cases}0 & i=0\\ u_i/w_i & 0<i\leq n\end{cases}.$$

The function we wish to bound equals $H(Y|X)\leq H(Y)$. We therefore observe that

$$\sum_{i=1}^{n} w_i H_b\left(\frac{u_i}{w_i}\right)\leq H(Y).$$

The result follows from the observation that

$$\mathbf{P}\left(Y=0\right)=\sum_i \mathbf{P}\left(X=i\right)\frac{u_i}{w_i}=\sum_i u_i\leq a.$$

□

Lemma B.12.3. *Let X be a random variable distributed over k states, with $\mathbf{P}\left(X=x\right)=p_x$. Let $\alpha_x\geq 0$ be such that $\sum\alpha_x=\delta$, and define the random variable X' by $\mathbf{P}\left(X'=x\right)=(p_x+\alpha_x)/(1+\delta)$. We have*

$$\left|H(X')-H(X)\right|\leq H_b(\delta)+\delta\log k$$

where H_b is the binary entropy function.

Proof. Define a new random variable Z by

$$\mathbf{P}\left(Z=0|X'=x\right)=\frac{p_x}{p_x+\alpha_x},\quad \mathbf{P}\left(Z=1|X'=x\right)=\frac{\alpha_x}{p_x+\alpha_x}.$$

We will use the fact that $H(X'|Z=0)=H(X)$ to obtain the required bound.

To upper bound $H(X')-H(X)$, we write

$$\begin{aligned}
H(X')-H(X) &\leq H(X',Z)-H(X)\\
&= H(Z)+\mathbf{P}\left(Z=0\right)H(X'|Z=0)+\mathbf{P}\left(Z=1\right)H(X'|Z=1)-H(X)\\
&\leq H_b(\delta)+(1-\delta)H(X)+\delta H(X'|Z=1)-H(X)\\
&= H_b(\delta)-\delta H(X)+\delta\log k\\
&\leq H_b(\delta)+\delta\log k
\end{aligned}$$

where in the fourth line we have used that $H(X'|Z=1)\leq\log k$.

To upper bound $H(X)-H(X')$, we write

$$\begin{aligned}
H(X')+H(Z) &\geq H(X',Z)\\
&\geq \mathbf{P}\left(Z=0\right)H(X'|Z=0)\\
&= (1-\delta)H(X)
\end{aligned}$$

which yields

$$H(X') \geq (1-\delta)H(X) - H_b(\delta)$$

since $H(Z) = H_b(\delta)$. Thus, we have

$$H(X) - H(X') \leq \delta H(X) + H_b(\delta) \leq \delta \log k + H_b(\delta).$$

□

Lemma B.12.4. *Let X be a random variable distributed over k states, with $\mathbf{P}(X = x) = p_x$. Let $\alpha_x \leq 0$ be such that $\sum |\alpha_x| = \delta$, and define the random variable X' by $\mathbf{P}(X' = x) = (p_x + \alpha_x)/(1-\delta)$. We have*

$$\left|H(X') - H(X)\right| \leq H_b\left(\frac{\delta}{1-\delta}\right) + \frac{\delta}{1-\delta}\log k$$

where H_b is the binary entropy function. In particular, when $\delta \leq 1/3$ we have

$$\left|H(X') - H(X)\right| \leq H_b(2\delta) + 2\delta \log k.$$

Proof. We observe that we can get from X' to X by adding $\delta/(1-\delta)$ probability mass and rescaling. The previous lemma then gives the result. □

Lemma B.12.5. *Let X be a random variable distributed over k states, with $\mathbf{P}(X = x) = p_x$. Let α_x be such that $\sum |\alpha_x| = \delta$, and define the random variable X' by $\mathbf{P}(X' = x) =$*

$(p_x + \alpha_x)/(1 - \sum \alpha_x)$. *That is, X' is the result of changing the probability of state x by α_x and then re-normalizing to obtain a valid distribution. If $\delta \leq 1/4$, we have*

$$\left|H(X') - H(X)\right| \leq 2H_b(2\delta) + 3\delta \log k$$

where H_b is the binary entropy function.

Proof. Let δ_+ be the total magnitude of all the positive α_x, and let δ_- be the total magnitude of all the negative α_x. We first add all the mass we're going to add, and apply the first of the previous two lemmas. Then we remove all the mass we are going to remove, and apply the second of the two previous lemmas. This yields a bound of

$$\begin{aligned}
& H_b(\delta_+) + \delta_+ \log k + H_b\left(2\frac{\delta_-}{1+\delta_+}\right) + 2\frac{\delta_-}{1+\delta_+}\log k \\
\leq\ & H_b(\delta_+) + \delta_+ \log k + H_b(2\delta_-) + 2\delta_- \log k \\
\leq\ & H_b(2\delta) + \delta \log k + H_b(2\delta) + 2\delta \log k \\
\leq\ & 2H_b(2\delta) + 3\delta \log k
\end{aligned}$$

where the first inequality is because $1 + \delta_+ \leq 1 + \delta < 2$ and $2\delta_- \leq 2\delta \leq 1/2$, and the second inequality is because $\delta_+ \leq \delta < 2\delta \leq 1/2$. □

C

Supplementary information for Chapter 4

C.1 Online empirical supplement

A full set of simulation analyses of power, equitability, and runtime, including sensitivity analyses and additional samples sizes and models, are available for download at http://www.exploredata.net/ftp/aoas_empirical_supplement.zip.

C.2 Supplementary methods

C.2.1 Equitability analyses

We describe here the details of an example analysis of the equitability with respect to R^2 of the squared sample correlation $\hat{\rho}^2$ on a specific set $\boldsymbol{\mathcal{Q}}$, since the equitability analyses of the other methods in this paper all follow the same pattern.

First, we fix the set of standard relationships $\boldsymbol{\mathcal{Q}}$ to be a set of noisy functional relationships with the appropriate type of marginals and noise. In this example, noise might be added along the Y-axis only, and so the relationships in $\boldsymbol{\mathcal{Q}}$ take the form $(X + \varepsilon, f(X) + \varepsilon'_\sigma)$ with $\varepsilon = 0$, $\varepsilon'_\sigma \sim \mathcal{N}(0, \sigma^2)$, f ranging over the set F of functions indicated in Table C.1, and X chosen to be the marginal distribution of the uniform distribution over the graph of f. Next, to analyze the equitability with respect to R^2 of $\hat{\rho}^2$, we generate, for 41 different noise levels σ and for every function $f \in F$, 500 samples from the relationship $Z = (X, f(X) + \varepsilon'_\sigma)$ with a sample size of $n = 500$. Using these, we estimate the 5th and 95th percentiles of the sampling distribution of $\hat{\rho}^2$ on Z. These allow us to estimate the level-0.05 $\boldsymbol{\mathcal{Q}}$-acceptance region at the value of R^2 corresponding to each noise level. The $\boldsymbol{\mathcal{Q}}$-acceptance regions then enable us to construct $\boldsymbol{\mathcal{Q}}$-confidence intervals, and our estimate of the equitability is then the reciprocal of the length of the longest $\boldsymbol{\mathcal{Q}}$-confidence interval.

C.2.2 Power analyses

For the power analyses, we determined, for each of the eight relationship types chosen by Simon and Tibshirani* 100 noise levels evenly distributed over the range of noise levels yielding $R^2 = 1.0$ (no noise) and $R^2 = 10^{-3}$ (substantial noise). For each noise level, we then drew $1,000$ independent samples, each of size $n = 500$, from the corresponding distribution representing our alternative hypothesis. We also drew $1,000$ independent samples from a corresponding null hypothesis chosen to have the same marginals. Power was then estimated at a significance level of 0.05. This resulted in a power estimate for every 4-tuple of method, relationship type, noise level, and parameter setting.

Quantifying power via integration under power curves

As mentioned in the main text, one of the methods we used to aggregate power across noise levels in a way that was comparable across relationship types was by integrating under the power curve generated for the method in question on the relationship type in question. There are a few details worth mentioning about this procedure. First, we integrated with respect to absolute noise magnitude and not with respect to R^2. When integration is performed with respect to R^2, performance on cleaner relationships is emphasized, typically leading to improved performance of MIC_e and TIC_e relative to

* Note that one of the relationship types chosen by Simon and Tibshirani was a circle. We heuristically defined the R^2 of a noisy circle to be the average of the R^2 values, computed separately, of the top and bottom halves.

the other statistics examined.

When integration is done with respect to noise magnitude instead of R^2 threshold, the upper limit of integration needs to be defined. To ensure this upper limit was consistent across relationship types, we defined it by choosing a universal R^2 threshold, call it ε, and computing, for each relationship type, the noise magnitude necessary to reach an R^2 of ε. We used values of $\varepsilon \in \{10^{-3}, 10^{-2.5}, 10^{-2}\}$, and our results were robust to this variation of ε (see Empirical Supplement 2A).

Figures 4.4 and 4.5a were then created using, for each method, the parameter setting with the best performance by this aggregated metric, averaged over relationship types. (See Figure C.6 for parameter sweeps.)

Quantifying power via R^2 threshold

Figure 4.5b was created by selecting for each method the parameter setting that minimized the average over relationship types of the minimal R^2 necessary to achieve 50% power on each relationship type. (See Figure C.7 for parameter sweeps.) Our results were robust to the power threshold used (see Empirical Supplement 2B).

C.2.3 Power analyses on randomly chosen functions

To assess whether the parameters chosen for each method in the two power analyses described in Section C.2.2 yielded performance that generalized well to new relationship types, we created a set of 160 functions randomly drawn from Gaussian process

distributions with radial basis function kernels. We used the eight kernel bandwidths in the set $\{0.01, 0.025, 0.05, 0.1, 0.2, 0.25, 0.5, 1\}$, and batched together the 20 random functions of each bandwidth as a "relationship type". This yielded the power curves shown in Figure C.1.

We then repeated the two analyses described in Section C.2.2 on this set of power curves using the parameters that performed best on the fixed set of relationships in those analyses. This yielded similar results, with the top methods remaining S^{DDP} and TIC_e for each of the two respective ways of quantifying power; with dCor remaining among the top three methods when power is quantified using area under the power curve and generally not otherwise; and with MIC_e performing quite well, comparably to and usually better than MIC. For a version of Figure 4.5 but with the randomly chosen functions instead of the fixed relationship set, see Figure C.2; for the full set of results with power assessed using varying noise and power thresholds, see Empirical Supplements 2A and 2B.

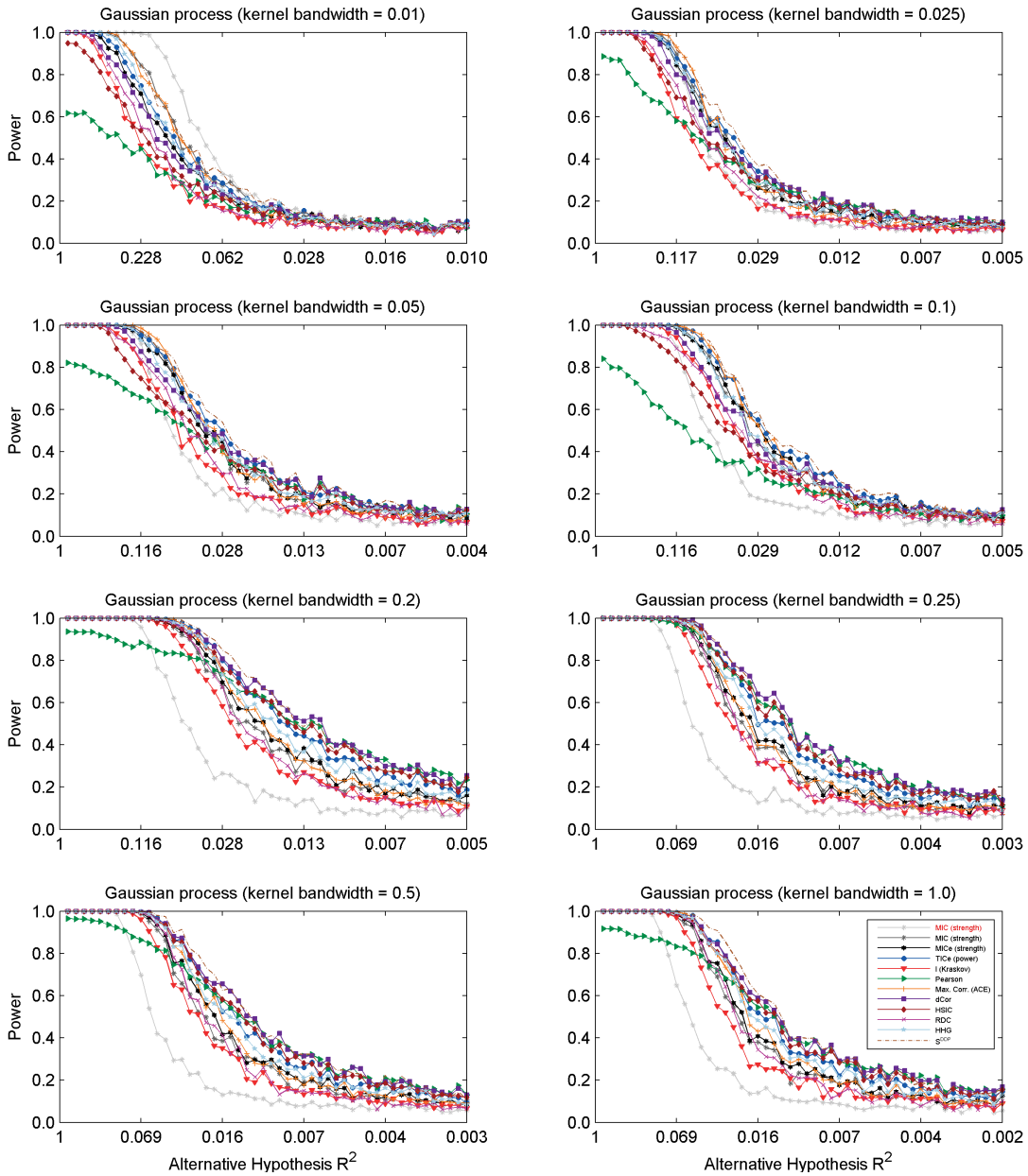
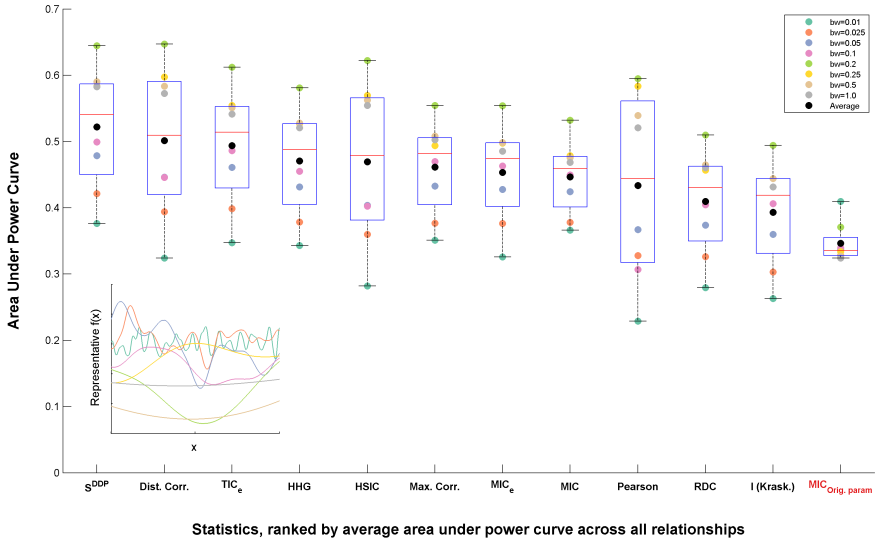


Figure C.1: Power of independence testing using the measures of dependence examined, on the 160 randomly chosen relationships described in Section C.2.3, at 50 noise levels with linearly increasing magnitude for each relationship and $n = 500$. To enable comparison of power regimes across relationships, the x-axis of each plot lists R^2 rather than noise magnitude. For each statistic that has a parameter, the value used was the same as in Figure 4.4.

(a)



(b)

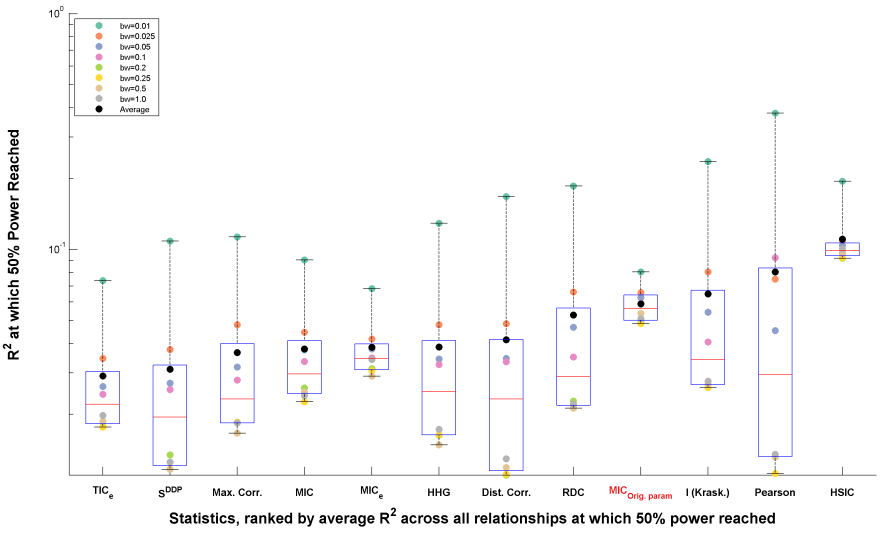


Figure C.2: Measures of dependence ranked by the power of their corresponding independence tests on 160 randomly chosen noisy functional relationships. For each measure of dependence and each relationship type, power was quantified using (a) the area under the power curve *[higher is more powerful]*, or (b) the minimal R^2 at which at least 50% power is achieved *[lower is more powerful]*. The collection of these scores across the eight different Gaussian process kernel bandwidths is then plotted for each method along with quartiles. The inset plot in (a) shows example relationships with each bandwidth. For each statistic that has a parameter, the value used was the same as in Figure 4.5a and Figure 4.5b respectively. The sample size is $n = 500$.

C.2.4 Data generation

Definition of sampling/noise models used

Each of the twelve sampling/noise models $\boldsymbol{\mathcal{Q}}$ consists of an independent-variable marginal distribution from the set

$$\left\{ \begin{array}{ll} \text{points sampled evenly along the graph of } f(X) & [E, f(X)] \\ \text{points sampled evenly along the } X \text{ range} & [E, X] \\ \text{points sampled uniformly along the graph of } f(X) & [U, f(X)] \\ \text{points sampled uniformly along the } X \text{ range} & [U, X] \end{array} \right\}$$

and a noise distribution from the set

$$\left\{ \begin{array}{ll} \text{Gaussian noise added along Y-axis} & [\mathcal{N}_y] \\ \text{Gaussian noise added along both axes} & [\mathcal{N}_x, \mathcal{N}_y] \\ \text{Gaussian noise added along X-axis} & [\mathcal{N}_x] \end{array} \right\}.$$

In this appendix, we refer to these noise models using abbreviations of the form, e.g., $[E, f(X), \mathcal{N}_y]$, which would correspond to a model in which the independent variable is sampled evenly along the curve described by $f(X)$ and Gaussian noise is added only to the dependent coordinate. Appendix C.2.4 contains definitions of the functions used.

Definitions of functions used

Tables C.1 and C.2 contain the definitions of the functions used to assess the equitability and statistical power against independence, respectively, of measures of dependence throughout this paper. The functions used for all analyses of power against independence (Table C.2) are taken from Simon & Tibshirani[144].

#	Function Name	Definition	
1	Cosine, High Freq	$y = \cos(14\pi x)$	$x \in [0, 1]$
2	Cosine, Non-Fourier Freq [Low]	$y = \cos(7\pi x)$	$x \in [0, 1]$
3	Cosine, Varying Freq [Medium]	$y = \sin(5\pi x(1 + x))$	$x \in [0, 1]$
4	Cubic	$y = 4x^3 + x^2 - 4x$	$x \in [-1.3, 1.1]$
5	Cubic, Y-stretched	$y = 41(4x^3 + x^2 - 4x)$	$x \in [-1.3, 1.1]$
6	Exponential [10^x]	$y = 10^x$	$x \in [0, 10]$
7	Exponential [2^x]	$y = 2^x$	$x \in [0, 10]$
8	L-shaped	$y = \begin{cases} x/99 & \text{if } x \leq \frac{99}{100} \\ 1 & \text{if } x > \frac{99}{100} \end{cases}$	$x \in [0, 1]$
9	Line	$y = x$	$x \in [0, 1]$
10	Linear+Periodic, High Freq	$y = \frac{1}{10}\sin(10.6(2x - 1)) + \frac{11}{10}(2x - 1)$	$x \in [0, 1]$
11	Linear+Periodic, High Freq 2	$y = \frac{1}{5}\sin(10.6(2x - 1)) + \frac{11}{10}(2x - 1)$	$x \in [0, 1]$
12	Linear+Periodic, Low Freq	$y = \frac{1}{5}\sin(4(2x - 1)) + \frac{11}{10}(2x - 1)$	$x \in [0, 1]$
13	Linear+Periodic, Medium Freq	$y = \sin(10\pi x) + x$	$x \in [0, 1]$
14	Lopsided L-shaped	$y = \begin{cases} 200x & \text{if } x < \frac{1}{200} \\ -198x + \frac{199}{100} & \text{if } \frac{1}{200} \leq x < \frac{1}{100} \\ -\frac{x}{99} + \frac{1}{99} & \text{if } x \geq \frac{1}{100} \end{cases}$	$x \in [0, 1]$
15	Parabola	$y = 4x^2$	$x \in [-\frac{1}{2}, \frac{1}{2}]$
16	Sigmoid	$y = \begin{cases} 0 & \text{if } x \leq \frac{49}{100} \\ 50(x - \frac{1}{2}) + \frac{1}{2} & \text{if } \frac{49}{100} \leq x \leq \frac{51}{100} \\ 1 & \text{if } x > \frac{51}{100} \end{cases}$	$x \in [0, 1]$
17	Sine, High Freq	$y = \sin(16\pi x)$	$x \in [0, 1]$
18	Sine, Low Freq	$y = \sin(8\pi x)$	$x \in [0, 1]$
19	Sine, Non-Fourier Freq [Low]	$y = \sin(9\pi x)$	$x \in [0, 1]$
20	Sine, Varying Freq [Medium]	$y = \sin(6\pi x(1 + x))$	$x \in [0, 1]$
21	Spike	$y = \begin{cases} 20 & \text{if } x < \frac{1}{20} \\ -18x + \frac{19}{10} & \text{if } \frac{1}{20} \leq x < \frac{1}{10} \\ -\frac{x}{9} + \frac{1}{9} & \text{if } x \geq \frac{1}{10} \end{cases}$	$x \in [0, 1]$

Table C.1: Definitions of the functions used to analyze equitability. Under noise/sampling models containing X noise or independent-variable marginal distributions other than $[E, f(X)]$ or $[U, f(X)]$, functions 6, 8, 14, 16, and 21 were excluded due to poor performance across all methods tested. This is presumably due to the fact that a) horizontally perturbing points in a very steep portion of a function drastically changes the distribution in question, and b) sampling uniformly along the x-axis under-samples a large part of the graph of a function if that graph contains very steep portions.

Function Name	**Definition**	
Line	$y = x$	$x \in [0, 1]$
Quadratic	$y = 4x^2$	$x \in [-\frac{1}{2}, \frac{1}{2}]$
Cubic	$y = 128(x - \frac{1}{3})^3 - 48(x - \frac{1}{3})^3 - 12(x - \frac{1}{3})$	$x \in [0, 1]$
Sinusoid (8 periods)	$y = \sin(16\pi x)$	$x \in [0, 1]$
Sinusoid (2 periods)	$y = \sin(4\pi x)$	$x \in [0, 1]$
$x^{1/4}$	$y = x^{1/4}$	$x \in [0, 1]$
Circle	$y = \pm\sqrt{1 - (2x - 1)^2}$	$x \in [0, 1]$
Step	$y = \begin{cases} 0 & \text{if } x \leq \frac{1}{2} \\ 1 & \text{if } x > \frac{1}{2} \end{cases}$	$x \in [0, 1]$

Table C.2: Definitions of the functions from Simon & Tibshirani[144] used to analyze statistical power against independence.

C.3 Additional equitability results

Sample Size	Model: $p_x(X)$	Model: Noise (Gaussian)	Maximal Corr. (ACE)	Pearson	dCor	HSIC	I [L^2] (Kraskov, k=1)	I [L^2] (Kraskov, k=6)	RDC	HHG	S^{DDP}	TIC_e	MIC_e	MIC
$n = 250$	Even Along $f(X)$	Y-Noise	0.58	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.88	0.39	0.45
	Even Along X	Y-Noise	0.52	1.00	1.00	1.00	0.99	0.92	1.00	0.98	0.99	0.87	0.35	0.42
	Even Along $f(X)$	XY-Noise	0.63	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.50	0.55
	Even Along X	XY-Noise	0.66	1.00	1.00	1.00	1.00	0.88	1.00	1.00	0.99	0.98	0.63	0.72
	Even Along $f(X)$	X-Noise	0.64	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.94	0.50	0.55
	Even Along X	X-Noise	0.66	1.00	1.00	1.00	1.00	0.90	1.00	1.00	0.99	0.98	0.63	0.68
	Uniform Along $f(X)$	Y-Noise	0.56	1.00	1.00	1.00	0.65	1.00	1.00	0.98	0.98	0.83	0.43	0.45
	Uniform Along X	Y-Noise	0.51	1.00	1.00	1.00	0.59	0.87	1.00	0.98	0.98	0.83	0.42	0.42
	Uniform Along $f(X)$	XY-Noise	0.61	1.00	1.00	1.00	0.69	1.00	1.00	0.98	0.98	0.94	0.50	0.53
	Uniform Along X	XY-Noise	0.68	1.00	1.00	1.00	0.76	0.98	1.00	0.98	0.98	0.98	0.65	0.71
	Uniform Along $f(X)$	X-Noise	0.61	1.00	1.00	1.00	0.70	1.00	1.00	0.98	0.98	0.94	0.50	0.54
	Uniform Along X	X-Noise	0.68	1.00	1.00	1.00	0.81	0.98	1.00	0.98	0.98	0.98	0.64	0.72
		Worst Case	**0.68**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.98**	**0.65**	**0.72**
$n = 500$	Even Along $f(X)$	Y-Noise	0.53	1.00	1.00	1.00	0.45	0.75	1.00	0.98	0.98	0.70	0.24	0.27
	Even Along X	Y-Noise	0.48	1.00	1.00	1.00	0.43	0.48	1.00	0.97	0.98	0.70	0.21	0.24
	Even Along $f(X)$	XY-Noise	0.59	1.00	1.00	1.00	0.97	0.96	1.00	0.98	0.98	0.77	0.34	0.38
	Even Along X	XY-Noise	0.63	1.00	1.00	1.00	0.95	0.84	1.00	0.98	0.98	0.95	0.48	0.51
	Even Along $f(X)$	X-Noise	0.58	1.00	1.00	1.00	0.97	0.95	1.00	0.98	0.98	0.80	0.35	0.38
	Even Along X	X-Noise	0.63	1.00	1.00	1.00	0.96	0.87	1.00	0.98	0.98	0.95	0.48	0.51
	Uniform Along $f(X)$	Y-Noise	0.52	1.00	1.00	1.00	0.45	0.69	1.00	0.96	0.98	0.68	0.25	0.28
	Uniform Along X	Y-Noise	0.48	1.00	1.00	1.00	0.43	0.48	1.00	0.95	0.98	0.68	0.22	0.25
	Uniform Along $f(X)$	XY-Noise	0.59	1.00	1.00	1.00	0.57	0.83	1.00	0.98	0.98	0.74	0.35	0.38
	Uniform Along X	XY-Noise	0.65	1.00	1.00	1.00	0.67	0.75	1.00	0.98	0.98	0.93	0.51	0.56
	Uniform Along $f(X)$	X-Noise	0.59	1.00	1.00	1.00	0.59	0.84	1.00	0.98	0.98	0.75	0.35	0.39
	Uniform Along X	X-Noise	0.65	1.00	1.00	1.00	0.73	0.80	1.00	0.98	0.98	0.94	0.51	0.56
		Worst Case	**0.65**	**1.00**	**1.00**	**1.00**	**0.97**	**0.96**	**1.00**	**0.98**	**0.98**	**0.95**	**0.51**	**0.56**
$n = 5000$	Even Along $f(X)$	Y-Noise	0.44	1.00	0.98	1.00	0.18	0.08	1.00	--	0.96	0.44	0.15	0.16
	Even Along X	Y-Noise	0.40	1.00	0.98	1.00	0.18	0.07	1.00	--	0.96	0.43	0.12	0.15
	Even Along $f(X)$	XY-Noise	0.48	1.00	0.98	1.00	0.41	0.32	1.00	--	0.98	0.50	0.23	0.24
	Even Along X	XY-Noise	0.53	1.00	0.98	1.00	0.56	0.49	1.00	--	0.98	0.69	0.36	0.41
	Even Along $f(X)$	X-Noise	0.48	1.00	0.98	1.00	0.77	0.37	1.00	--	0.98	0.51	0.24	0.24
	Even Along X	X-Noise	0.53	1.00	0.98	1.00	0.83	0.58	1.00	--	0.98	0.69	0.37	0.40
	Uniform Along $f(X)$	Y-Noise	0.44	1.00	0.98	1.00	0.17	0.09	1.00	--	0.97	0.44	0.16	0.16
	Uniform Along X	Y-Noise	0.41	1.00	0.98	1.00	0.18	0.07	1.00	--	0.97	0.43	0.13	0.15
	Uniform Along $f(X)$	XY-Noise	0.49	1.00	0.98	1.00	0.37	0.30	1.00	--	0.98	0.50	0.24	0.25
	Uniform Along X	XY-Noise	0.54	1.00	0.98	1.00	0.53	0.47	1.00	--	0.98	0.69	0.39	0.43
	Uniform Along $f(X)$	X-Noise	0.49	1.00	0.98	1.00	0.41	0.34	1.00	--	0.98	0.51	0.25	0.25
	Uniform Along X	X-Noise	0.54	1.00	0.98	1.00	0.61	0.55	1.00	--	0.98	0.70	0.39	0.44
		Worst Case	**0.54**	**1.00**	**0.98**	**1.00**	**0.83**	**0.58**	**1.00**	**--**	**0.98**	**0.70**	**0.39**	**0.44**

Table C.3: A summary of the worst-case equitability of measures of dependence for a variety of noise models, independent-variable marginal distributions, and sample sizes. *[Smaller values correspond to better equitability.]* Each number is a worst-case $\mathcal{Q}$-confidence interval length for a given statistic in a given setting. Table cells are colored proportionally (red = interval of length 0; white = interval of length 1). Figures analogous to Figures 4.1 and C.3 for all the settings presented in this table are included in Empirical Supplement 1A. For statistics whose performance was dependent on parameter settings, at each sample size results are presented for parameter settings that maximize worst-case equitability across all twelve of the noise/marginal distributions tested at that sample size. Results are not presented for HHG for $n = 5,000$ as it was prohibitively computationally expensive to analyze at this sample size.

Sample Size	Model: $p_x(X)$	Model: Noise (Gaussian)	Maximal Corr. (ACE)	Pearson	dCor	HSIC	I [L²] (Kraskov, k=1)	I [L²] (Kraskov, k=6)	RDC	HHG	S^{DDP}	TIC_e	MIC_e	MIC
$n = 250$	Even Along $f(X)$	Y-Noise	0.38	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.45	0.28	0.32
	Even Along X	Y-Noise	0.31	0.50	0.50	0.50	0.50	0.49	0.50	0.50	0.50	0.45	0.26	0.30
	Even Along $f(X)$	XY-Noise	0.41	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.49	0.34	0.37
	Even Along X	XY-Noise	0.41	0.50	0.50	0.50	0.50	0.48	0.50	0.50	0.50	0.49	0.40	0.43
	Even Along $f(X)$	X-Noise	0.41	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.49	0.34	0.37
	Even Along X	X-Noise	0.41	0.50	0.50	0.50	0.50	0.49	0.50	0.50	0.50	0.49	0.40	0.43
	Uniform Along $f(X)$	Y-Noise	0.37	0.50	0.50	0.50	0.36	0.50	0.50	0.50	0.50	0.45	0.29	0.32
	Uniform Along X	Y-Noise	0.32	0.50	0.50	0.50	0.33	0.48	0.50	0.50	0.50	0.44	0.28	0.30
	Uniform Along $f(X)$	XY-Noise	0.40	0.50	0.50	0.50	0.40	0.50	0.50	0.50	0.50	0.48	0.34	0.37
	Uniform Along X	XY-Noise	0.42	0.50	0.50	0.50	0.43	0.49	0.50	0.50	0.50	0.48	0.41	0.43
	Uniform Along $f(X)$	X-Noise	0.41	0.50	0.50	0.50	0.41	0.50	0.50	0.50	0.50	0.48	0.35	0.37
	Uniform Along X	X-Noise	0.42	0.50	0.50	0.50	0.44	0.49	0.50	0.50	0.50	0.48	0.42	0.43
		Average Case	**0.39**	**0.50**	**0.50**	**0.50**	**0.45**	**0.49**	**0.50**	**0.50**	**0.50**	**0.47**	**0.34**	**0.37**
$n = 500$	Even Along $f(X)$	Y-Noise	0.36	0.50	0.50	0.50	0.26	0.46	0.50	0.50	0.49	0.38	0.19	0.21
	Even Along X	Y-Noise	0.29	0.50	0.50	0.50	0.24	0.35	0.50	0.49	0.50	0.37	0.17	0.19
	Even Along $f(X)$	XY-Noise	0.39	0.50	0.50	0.50	0.49	0.49	0.50	0.50	0.50	0.45	0.24	0.27
	Even Along X	XY-Noise	0.40	0.50	0.50	0.50	0.49	0.47	0.50	0.49	0.50	0.48	0.30	0.32
	Even Along $f(X)$	X-Noise	0.40	0.50	0.50	0.50	0.49	0.49	0.50	0.50	0.50	0.45	0.24	0.27
	Even Along X	X-Noise	0.40	0.50	0.50	0.50	0.49	0.48	0.50	0.49	0.50	0.48	0.30	0.33
	Uniform Along $f(X)$	Y-Noise	0.36	0.50	0.50	0.50	0.26	0.44	0.50	0.49	0.49	0.38	0.19	0.21
	Uniform Along X	Y-Noise	0.30	0.50	0.50	0.50	0.24	0.35	0.50	0.49	0.49	0.37	0.18	0.19
	Uniform Along $f(X)$	XY-Noise	0.39	0.50	0.50	0.50	0.33	0.47	0.50	0.49	0.49	0.44	0.25	0.27
	Uniform Along X	XY-Noise	0.41	0.50	0.50	0.50	0.38	0.44	0.50	0.49	0.49	0.48	0.31	0.33
	Uniform Along $f(X)$	X-Noise	0.40	0.50	0.50	0.50	0.33	0.47	0.50	0.49	0.49	0.44	0.25	0.27
	Uniform Along X	X-Noise	0.41	0.50	0.50	0.50	0.39	0.46	0.50	0.49	0.49	0.48	0.31	0.33
		Average Case	**0.38**	**0.50**	**0.50**	**0.50**	**0.37**	**0.45**	**0.50**	**0.49**	**0.49**	**0.43**	**0.24**	**0.26**
$n = 5000$	Even Along $f(X)$	Y-Noise	0.30	0.50	0.48	0.50	0.11	0.07	0.491	--	0.478	0.254	0.106	0.11
	Even Along X	Y-Noise	0.23	0.50	0.48	0.50	0.10	0.06	0.491	--	0.478	0.247	0.0937	0.10
	Even Along $f(X)$	XY-Noise	0.33	0.50	0.48	0.50	0.25	0.22	0.49	--	0.478	0.344	0.164	0.17
	Even Along X	XY-Noise	0.34	0.50	0.48	0.50	0.32	0.29	0.488	--	0.478	0.424	0.228	0.24
	Even Along $f(X)$	X-Noise	0.33	0.50	0.48	0.50	0.42	0.24	0.491	--	0.479	0.352	0.166	0.17
	Even Along X	X-Noise	0.35	0.50	0.48	0.50	0.43	0.32	0.488	--	0.478	0.427	0.228	0.24
	Uniform Along $f(X)$	Y-Noise	0.31	0.50	0.48	0.50	0.11	0.07	0.497	--	0.478	0.255	0.109	0.11
	Uniform Along X	Y-Noise	0.23	0.50	0.48	0.50	0.10	0.06	0.497	--	0.478	0.247	0.0962	0.11
	Uniform Along $f(X)$	XY-Noise	0.34	0.50	0.48	0.50	0.23	0.20	0.498	--	0.478	0.345	0.168	0.17
	Uniform Along X	XY-Noise	0.35	0.50	0.48	0.50	0.30	0.28	0.497	--	0.477	0.425	0.236	0.25
	Uniform Along $f(X)$	X-Noise	0.34	0.50	0.48	0.50	0.24	0.21	0.498	--	0.478	0.352	0.171	0.18
	Uniform Along X	X-Noise	0.35	0.50	0.48	0.50	0.33	0.30	0.495	--	0.477	0.427	0.237	0.25
		Average Case	**0.32**	**0.50**	**0.48**	**0.50**	**0.25**	**0.19**	**0.49**	**--**	**0.48**	**0.34**	**0.17**	**0.17**

Table C.4: A summary of the average-case equitability of measures of dependence for a variety of noise models, independent-variable marginal distributions, and sample sizes. *[Smaller values correspond to better equitability.]* Each number is an average $\mathcal{Q}$-confidence interval length for a given statistic in a given setting. Table cells are colored proportionally (red = interval of length 0; white = interval of length 1). Figures analogous to Figures 4.1 and C.3 for all the settings presented in this table are included in Empirical Supplement 1A. For statistics whose performance was dependent on parameter settings, we used the same parameter settings as in Table C.3. Results are not presented for HHG for $n = 5,000$ as it was prohibitively computationally expensive to analyze at this sample size.

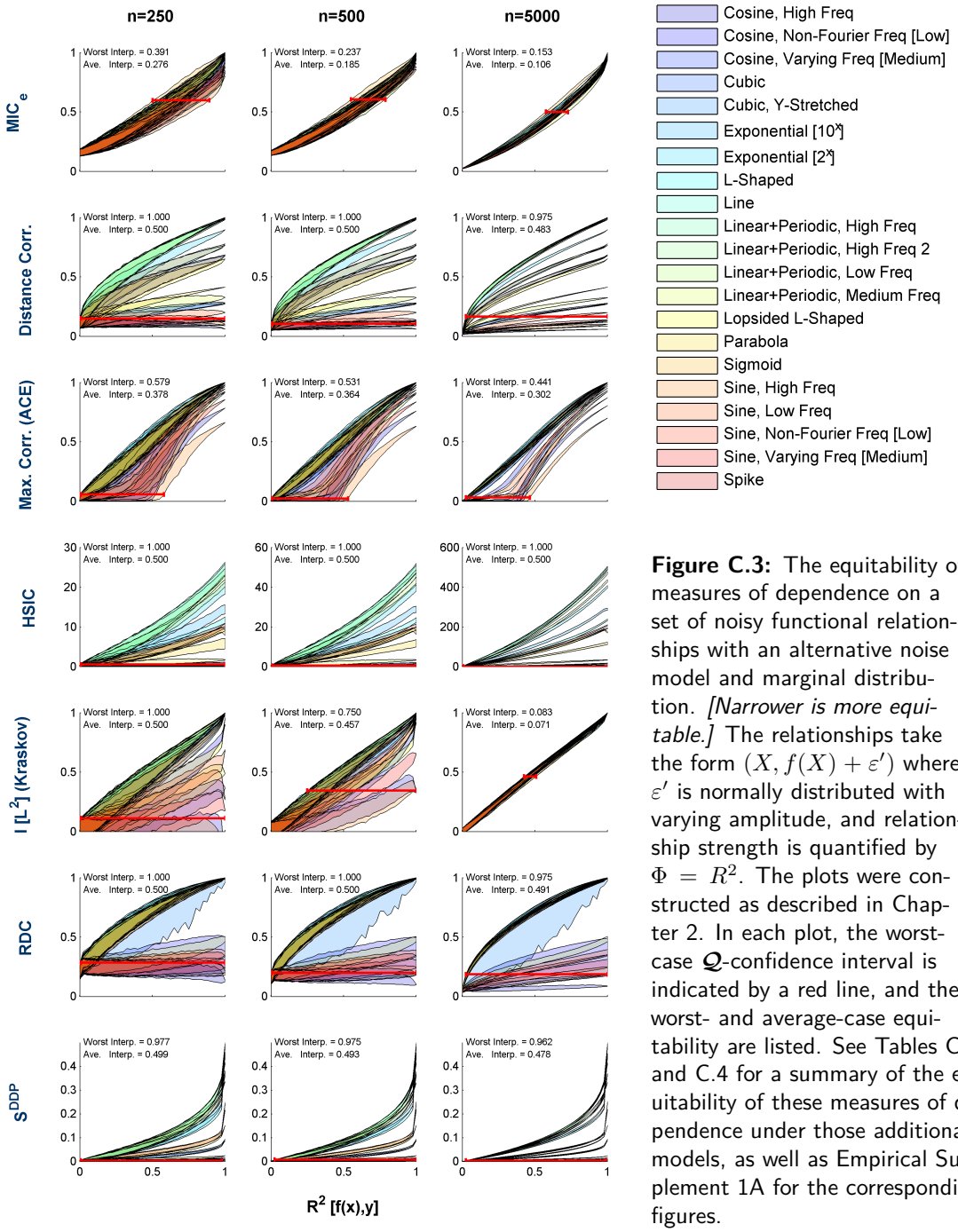


Figure C.3: The equitability of measures of dependence on a set of noisy functional relationships with an alternative noise model and marginal distribution. *[Narrower is more equitable.]* The relationships take the form $(X, f(X) + \varepsilon')$ where ε' is normally distributed with varying amplitude, and relationship strength is quantified by $\Phi = R^2$. The plots were constructed as described in Chapter 2. In each plot, the worst-case $\mathcal{Q}$-confidence interval is indicated by a red line, and the worst- and average-case equitability are listed. See Tables C.3 and C.4 for a summary of the equitability of these measures of dependence under those additional models, as well as Empirical Supplement 1A for the corresponding figures.

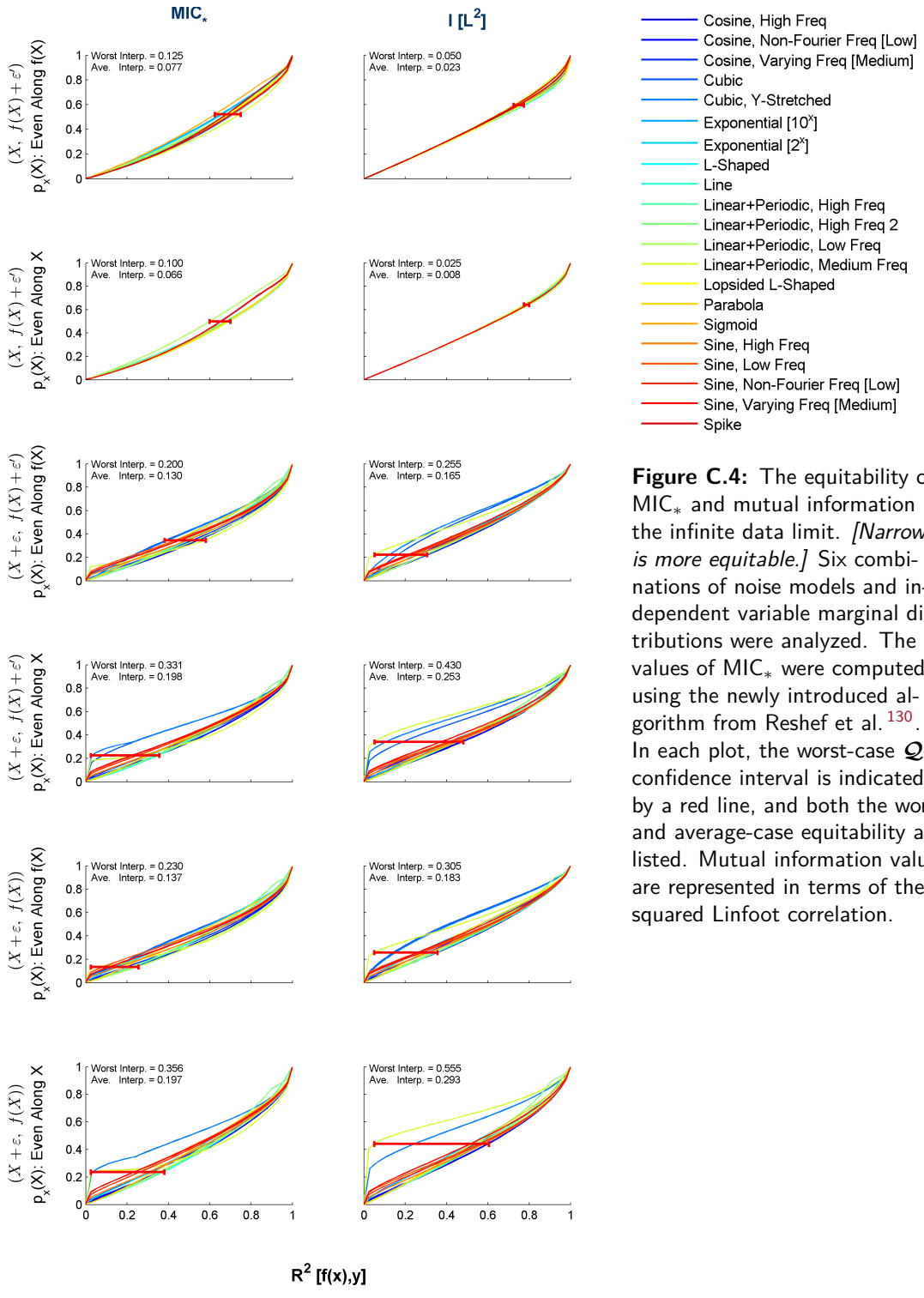


Figure C.4: The equitability of MIC_* and mutual information in the infinite data limit. *[Narrower is more equitable.]* Six combinations of noise models and independent variable marginal distributions were analyzed. The values of MIC_* were computed using the newly introduced algorithm from Reshef et al.[130]. In each plot, the worst-case $\mathcal{Q}$-confidence interval is indicated by a red line, and both the worst- and average-case equitability are listed. Mutual information values are represented in terms of the squared Linfoot correlation.

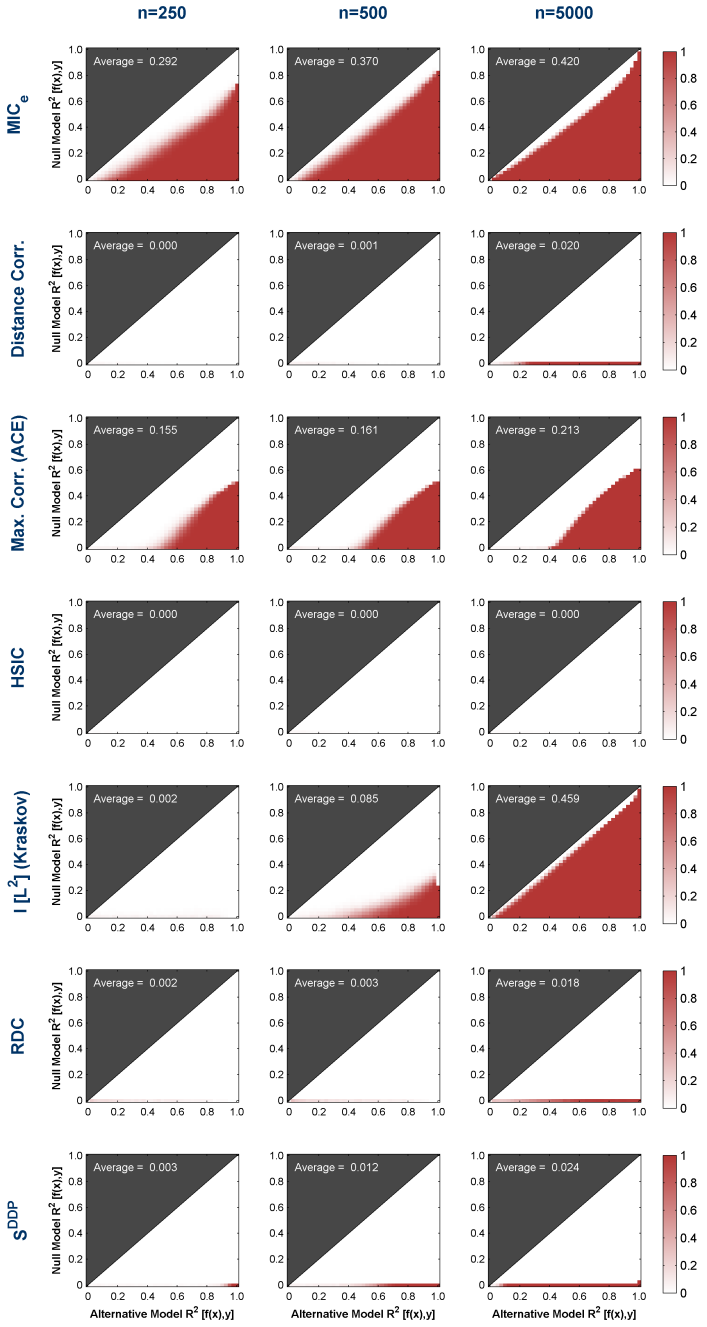


Figure C.5: The equitability with respect to $\Phi = R^2$ of measures of dependence on the noisy functional relationships analyzed in Figure C.3, visualized in terms of power. *[Redder is more equitable.]* Plots were generated as in Figure 4.2. Mutual information, estimated using the Kraskov estimator, is represented using the squared Linfoot correlation. For every parametrized statistic whose parameter meaningfully affects equitability, results are presented at each sample size using parameter settings that maximize equitability across all twelve of the noise/marginal distributions tested at that sample size. See Tables C.3 and C.4 for a summary of the equitability of these measures of dependence under those additional models, as well as Empirical Supplement 1B for the corresponding figures.

C.4 Results of parameter sweeps for power against independence

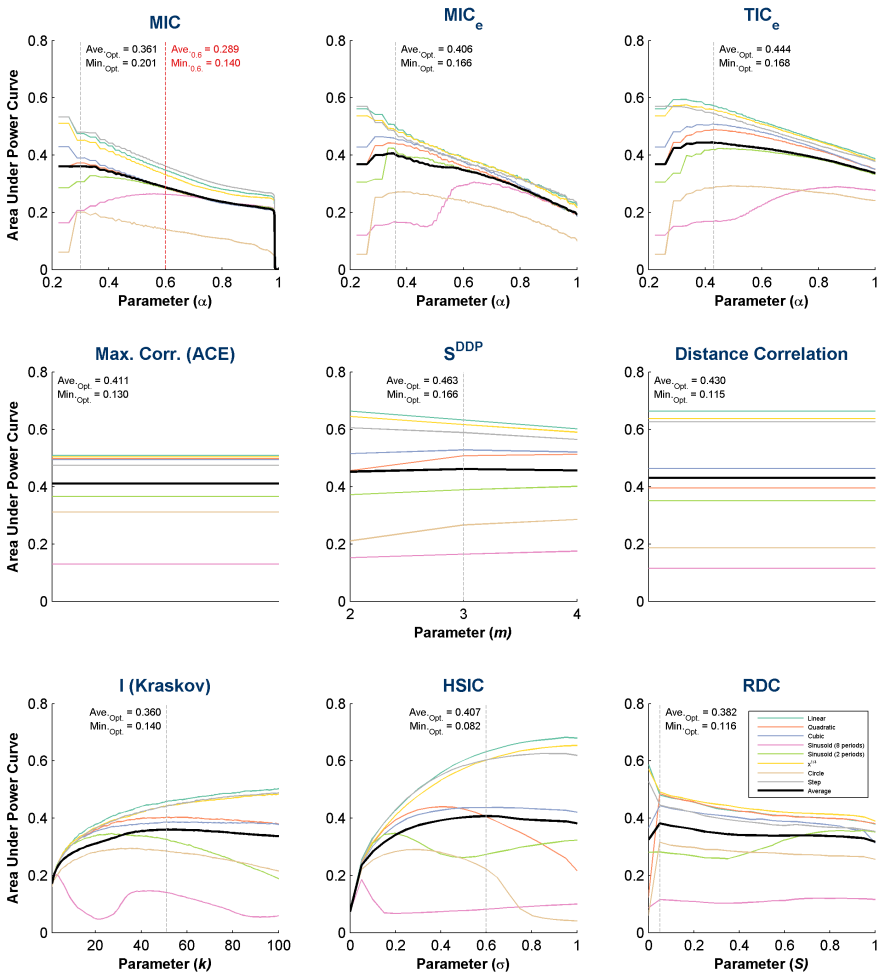


Figure C.6: Power against independence as a function of the parameter of each measure of dependence. *[Higher is more powerful.]* For each measure of dependence, we computed power curves over a range of parameters using the relationships from Simon & Tibshirani [144]. All power curves were computed as functions of the R^2 of the noisy relationship comprising the alternative hypothesis, and the area under each power curve was computed. We show for each statistic the area under the power curve for each relationship type as a function of that statistic's parameter. The black line represents the average area under the power curves across all relationship types, and the vertical dotted line represents the optimal parameter setting. Average and worst-case performance across relationship types are listed for the optimal parameter setting of each statistic. For the MIC statistic from Reshef et al. [121], the red line represents the default parameter setting used by Simon and Tibshirani.

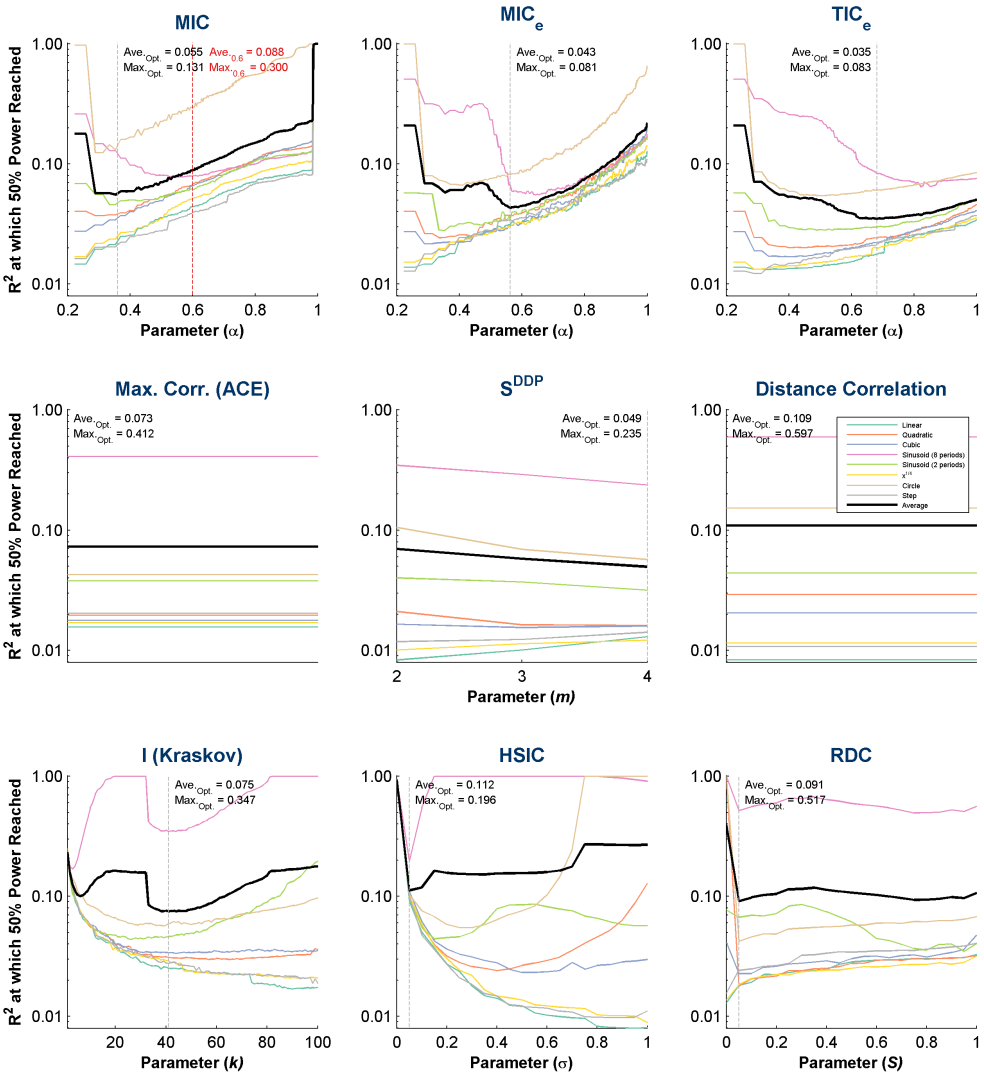


Figure C.7: Power against independence as a function of the parameter of each measure of dependence, with overall power quantified differently than in Figure C.6. *[Lower is more powerful.]* As in Figure C.6, we compute power curves for a range of parameters of each measure of dependence using the relationships from Simon & Tibshirani[144]. Here, in order to aggregate the power of a given test across relationship types, the power curve of each test was computed as a function of the R^2 of the noisy relationship being tested, and the R^2 at which 50% power is achieved for each relationship type was determined. This number is graphed for each relationship type and statistic as a function of that statistic's parameter. The black line represents the average R^2 at which 50% power is achieved across all relationships tested, and the vertical dotted line represents the optimal parameter setting. Average and worst-case performance across relationship types are listed for the optimal parameter setting of each statistic. For the MIC statistic from Reshef et al.[121], the red line represents the default parameter setting, which was used by Simon and Tibshirani.

C.5 The equitability-runtime trade-off

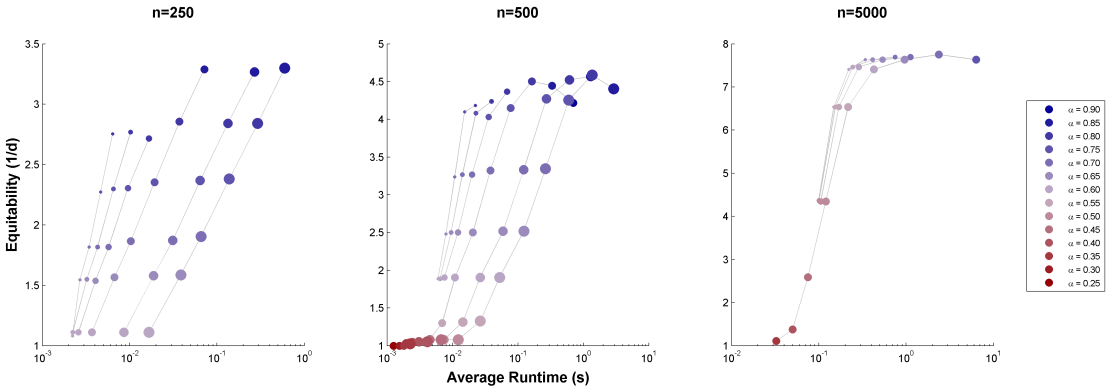


Figure C.8: The relationship between equitability and runtime of MIC_e. Sample sizes are $n = 250$ (left), 500 (middle), and $5,000$ (right). Each plot shows, as α varies, the worst-case equitability of MIC_e with the given value of α on the model used in Figure 4.6 graphed against the runtime of MIC_e with the same value of α. The multiple series in every plot correspond to different values of c, with marker size indicating the size of c. The values of c used are 1, 2, 3, 5, 10, and 15. ($c = 10$ and $c = 15$ are ommitted from the analysis for $n = 5,000$.) As α increases, we generally see a rise in equitability but also in runtime.

C.6 Parameter values used in analyses

Parameter sweeps were performed for all methods in evaluating their equitability and statistical power against independence.

C.6.1 Parameter values used in equitability analyses

For each method, results are presented for the parameter values tested that maximized worst-case equitability across all models $\mathcal{Q}$ examined, at each sample size (see Table C.5). Results for all parameter values tested, including for some methods not included in the

figures here due to space constraints, can be found in the empirical supplement.

In the case of RDC and HSIC the parameter values tested did not have a strong effect on equitability, so we present performance for the default / rule of thumb parameter values. That is, the random sampling parameters, (S_x, S_y), of RDC and the RBF kernel bandwidth parameters, (σ_x, σ_y), used for HSIC were set independently for each of the two samples being tested to the Euclidean distance empirical median (values of $\{0-, 25-, 50-, 75-, 100-\}$%-ile pairwise distances were also tested for these parameters). For RDC, the number of random features was set to $k = 10$. For the Kraskov mutual information estimator, $k = 1$, $k = 6$, $k = 10$, and $k = 20$ were tested. In the case of S^{DDP}, values of $m > 3$ were prohibitively computationally expensive to run for this analysis. For MIC_e, at $n = 250$, 500, and $5,000$, the ranges of α tested were $\{0.60, 0.65, ..., 0.80, 0.85\}$, $\{0.25, 0.30, ..., 0.85, 0.90\}$, and $\{0.35, 0.40, ..., 0.70, 0.75\}$, respectively.

Sample size	**MIC_e**		**TIC_e**		S^{DDP}	**I (Kraskov)**	**RDC**		**HSIC**
	α	c	α	c	m	k	S_x, S_y	k	σ_x, σ_y
250	0.75	15	0.80	3	2	6	Median pair. dist.	10	Median pair. dist.
500	0.80	5	0.80	3	2	6	Median pair. dist.	10	Median pair. dist.
$5,000$	0.65	3	0.70	3	2	6	Median pair. dist.	10	Median pair. dist.

Table C.5: Parameters used in the equitability analyses.

C.6.2 Parameter values used in statistical power analyses

Tables C.6 and C.7 summarize the optimal parameters identified for tests for independence based on the methods examined, using area under the power curves and a 50%

power threshold, respectively, as the optimization criterion. The parameters in Table C.6 were used to generate the power curves in Figures 4.4 and C.1. The parameter ranges tested for each statistic can be observed from Figures C.6 and C.7.

Sample size	MIC_e		TIC_e		MIC		S^{DDP}	I (Kraskov)	RDC		HSIC
	α	c	α	c	α	c	m	k	S_x, S_y	k	σ_x, σ_y
100	0.48	5	0.50	5	0.40	5	3	13	5%-ile pair. dist.	10	45%-ile pair. dist.
500	0.35	5	0.38	5	0.30	5	3	50	5%-ile pair. dist.	10	60%-ile pair. dist.

Table C.6: Best parameters for testing for independence, identified by maximizing the average area under the power curves generated by a given test for the set of relationships examined.

Sample size	MIC_e		TIC_e		MIC		S^{DDP}	I (Kraskov)	RDC		HSIC
	α	c	α	c	α	c	m	k	S_x, S_y	k	σ_x, σ_y
100	0.74	5	0.96	5	0.48	5	5	12	5%-ile pair. dist.	10	30%-ile pair. dist.
500	0.56	5	0.68	5	0.36	5	4	41	5%-ile pair. dist.	10	5%-ile pair. dist.

Table C.7: Best parameters for testing for independence, identified by minimizing the average across relationship types of the minimal R^2 for which the power of a given test remained above 50%.

C.6.3 Parameter values used in runtime analyses

For methods whose runtime did not strongly depend on parameter settings, default parameter values were used. That is, the Kraskov mutual information estimator was run using $k = 6$, and the random sampling parameters, (S_x, S_y), of RDC and the RBF kernel bandwidth parameters, (σ_x, σ_y), used for HSIC were set independently for each of the two samples being tested to the Euclidean distance empirical median. In the case of RDC, the number of random features was set to $k = 10$, as in the runtime analysis in Lopez-Paz et al.[86]. Since the runtime of $\text{MIC}_e/\text{TIC}_e$ depends on parameter

choice, results are presented for three different sets of parameter settings, determined by assessing performance in our simulations and shown in Table C.8.

Sample size	Power		Fast equitability		Equitability	
	α	c	α	c	α	c
50	0.54	5	0.75	3	0.85	5
100	0.48	5	0.70	2	0.80	5
500	0.36	5	0.65	1	0.80	5
$1,000$	0.32	5	0.60	1	0.75	4
$5,000$	0.26	5	0.50	1	0.65	1
$10,000$	0.24	5	0.45	1	0.60	1

Table C.8: Parameters used in the runtime analysis of MIC_e presented in Table 4.1.

For MIC_e, the three sample-size-dependent parameter settings optimize for maximal power against independence, 80% of optimal equitability (fast equitability), and 99% of optimal equitability. For sample sizes for which results were not available, parameter values were estimated via interpolation/extrapolation using a power curve. As pointed out in Appendix C.8, these parameter settings depend on the set of relationships being examined, and, for example, for relationship suites with less complex relationships than the ones examined in the analyses here, lower values of α would perform well and be more computationally efficient.

The only other method whose runtime was affected by its parameter was S^{DDP}. At the sample size regimes tested in the equitability and power analyses, only three parameter settings led to practical runtimes for S^{DDP} ($m = 2, 3$, and 4). The runtime performance for all three of those parameter settings of S^{DDP} is presented.

In this analysis, the Kraskov mutual information estimator was run using a pre-compiled C binary, MIC was computed approximately using the APPROX-MIC algorithm[121] in Java, and MIC_e was run in Java. The other statistics were run using their respective R functions/packages.

Remark C.6.1. *In the runtime analysis, dCor was run with the standard R package, which is $O(n^2)$; as of this writing there is a new, faster estimator of the same population quantity that is computable in time $O(n \log n)$*[64].

C.6.4 Parameter values used in analysis of the WHO data set

For independence testing

For analyzing power against independence, we set the parameters of each method in a sample size-dependent manner by interpolating between the optimal parameters for $n = 100$ and $n = 500$ determined by our simulations and listed in Table C.6. This resulted in the following settings for each parametrized method.

$\text{MIC}_e/\text{TIC}_e$ Since in this procedure TIC_e is the statistic used for independence testing, we optimized the parameter of TIC_e for this purpose. TIC_e had an optimal $\alpha = 0.5$ and $\alpha = 0.38$ for $n = 100$ and $n = 500$, respectively, which corresponds to a maximal grid resolution of 10 cells in both cases. Therefore, we set α to be $\log_n 10$.

Mutual information For the Kraskov mutual information estimator, the optimal k was 13 for $n = 100$ and 50 for $n = 500$, both of which are approximately $0.1n$. We therefore used $k = 0.1n$ in our analysis.

HSIC The optimal σ_x and σ_y were the 45-th percentile pairwise distance at $n = 100$ and the 60-th percentile pairwise distance at $n = 500$. We therefore used the median pairwise distance given that our sample size range was closer to $n = 100$.

RDC and S^{DDP} Both of these methods had the same optimal parameter for $n = 100$ and $n = 500$: $m = 3$ for S^{DDP} and 5-th percentile pairwise distance for RDC.

For measuring relationship strength

Mutual information estimation has two different parameter regimes for equitability and independence testing respectively. We therefore also ran the Kraskov mutual information estimator with the equitability-optimized parameter ($k = 6$) from the $n = 250$ column of Table C.5. Wherever we quantify relationship strength in the WHO analysis using mutual information, we do so using the equitability-optimized parameter; wherever we quantify power against independence, we do so using the power-optimized parameter listed above.

For MIC_e/TIC_e, we supplied MIC_e with an equitability-optimized parameter $\alpha = 0.75$ from the $n = 250$ column of Table C.5, since the value of MIC_e is the one used to quantify

relationship strength once a null of independence has been rejected by TIC_e. For speed, we ran MIC_e with $c = 5$ rather than $c = 15$.

All the other statistics evaluated either did not show non-trivial equitability at any parameter setting or were not parametrized.

C.7 Highlighted results from analysis of WHO data set

For performance of $\text{MIC}_e/\text{TIC}_e$ on previously highlighted relationships from Reshef et al.[121], see Table C.9. For plots of three highlighted relationships ranked highly by $\text{MIC}_e/\text{TIC}_e$ but not by other methods, see Figure C.9.

X	Y	MIC_e	TIC_e p-value
Number of physicians	Deaths due to HIV/AIDS	0.55529	$\leq 5.96 \times 10^{-7}$
Per cap. gvmnt expend. on health	Measles immunization disparity	0.58093	$< 9.53 \times 10^{-6}$
Children per woman	Life expectancy at birth	0.60372	$\leq 5.96 \times 10^{-7}$
Adult female obesity (%)	Income per person	0.51699	$\leq 5.96 \times 10^{-7}$
Gross national income per capita	Health expenditure per person	0.85086	$\leq 5.96 \times 10^{-7}$

Table C.9: Performance of $\text{MIC}_e/\text{TIC}_e$ on relationships previously identified by MIC in Reshef et al.[121] in the WHO data set.

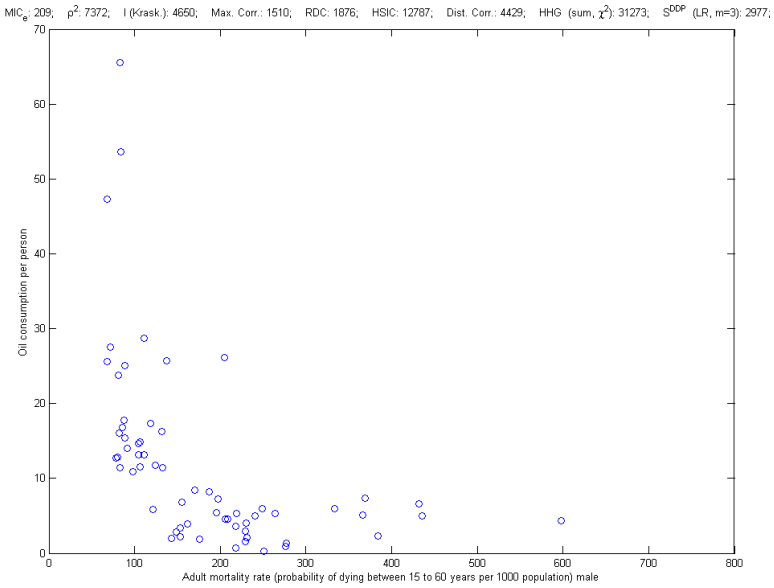
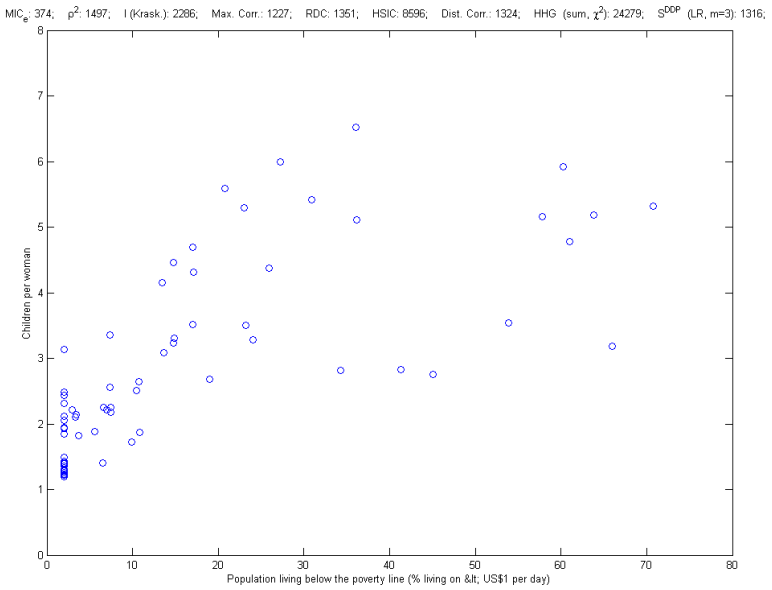


Figure C.9: Highlighted relationships found in the WHO data set using $\text{MIC}_e/\text{TIC}_e$. Each plot lists along the top the ranks given by each method to the relationship in question. (Continued on the following page.)



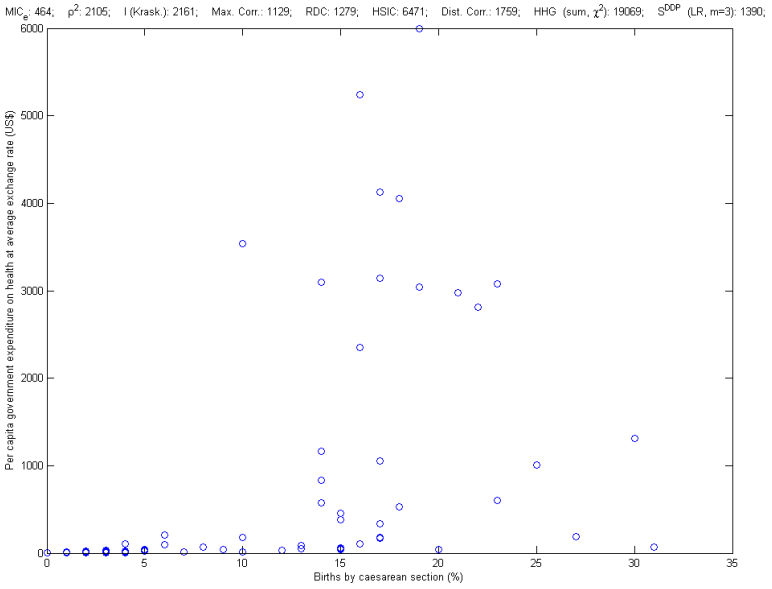


Figure C.9: Highlighted relationships found in the WHO data set using $\text{MIC}_e/\text{TIC}_e$ (continued from previous page).

C.8 Choosing parameters for MIC_e/TIC_e: a practical guide

Here we give guidelines for setting parameters for MIC_e/TIC_e. The two parameters required by these statistics are the parameter α discussed above, which governs the maximal grid resolution $B(n)$ of the estimator according to $B(n) = n^{\alpha}$, and c, an optional parameter that controls a speed-versus-optimality trade-off in the algorithm. We discuss each of these in turn.

C.8.1 Choosing α

There are two main considerations involved in choosing α. The first, which is suggested by the power-equitability tradeoff shown in the main text, is how much we care about power against independence relative to equitability. The second consideration is whether we expect to see complex relationships in our data. These considerations can can be reframed in terms of hypothesis testing as follows:

1. *Is our null hypothesis statistical independence or presence of a weak dependence?* When using MIC_e (or, more likely, TIC_e) to generate tests for statistical dependence one should use a lower value of α, while if one is interested in equitability, a larger α is required.

2. *What is our most complex alternative hypothesis?* Since α places an upper bound on the resolution of grids that can be explored by the estimators, it restricts the complexity of structure that can be detected. Thus, as the relationship class of interest grows to include more complex structure relative to sample size, the value of α should be increased.

Balancing the two considerations For the specific values of α that maximized power against independence of TIC_e and equitability of MIC_e, respectively, in our analyses, see Appendix C.6. The tables generally show that a) when optimizing for statistical power against independence in the sample-size regimes analyzed here, one should use an α that leads to $B(n) = n^\alpha$ being approximately between 4 (for less complex alternative hypotheses) and 12 (for more complex alternative hypotheses)[†], and b) when optimizing for equitability, one should use an α approximately between 0.5 (when n is larger) and 0.75 (when n is smaller).

Equitability and computational efficiency For large n, the parameters suggested above for equitability are likely needlessly computationally expensive. This is because as n grows, the maximal allowed grid resolution of the statistic $B(n) = n^\alpha$ will outstrip the complexity of most alternative hypotheses that we are liable to encounter in practice.

For example, at $n = 5,000$, $B(n) = 70$ provides good equitability on the set of functions and noise models tested in this paper. If this level of equitability is acceptable to us, we may set $\alpha = \log_n 70$ for $n \geq 5,000$, which means that $B(n) = 70$ always. Given that the runtime of the search procedure in MIC_e is $O(n^{5\alpha/2})$, which is $O(n)$

[†] Of course, for even more complex alternative hypotheses, a larger $B(n)$ will lead to better performance, provided the sample size allows for detection of the level of complexity in question. In particular, we suspect that for MIC_e and TIC_e, $B(n) > \omega(1)$ is necessary for consistency against all alternatives of the resulting independence test. For MIC and TIC, even just $B(n) = 4$ yields a statistic that is consistent against all alternatives. (See, e.g., Lemma 6.7 in the supplemental online materials of Reshef et al.[121].)

for $\alpha = 0.4$, a less extreme version of this strategy that maintains consistency and gives asymptotically linear runtime is to allow α to decrease for large n until $\alpha = 0.4$ is reached, and then to keep it at 0.4. In the example above, this happens around $n = 40,000$. And indeed, the equitability of MIC_e at this sample size with $\alpha = 0.4$ appears quite good (see Empirical Supplement 1F).

For more on how to balance runtime and equitability, see: Figure C.8, which graphs equitability on our set of functional relationships against runtime as α, n, and c are varied; Table C.8, which suggests values of α at several sample sizes that yield 80% of the best observed equitability for MIC_e at each sample size; and the discussion in the next section, where we examine the runtime of MIC_e compared to other statistics.

C.8.2 Choosing c

The parameter c determines the coarseness of the discretization of the grid search in the algorithm that computes MIC_e, with larger values of c corresponding to finer discretization[130]. Characterizing the effect of c on the bias and variance of MIC_e is an avenue of future work. However, using $c = 5$ seems to provide good performance in most settings, and in more computationally constrained settings setting even $c = 1$ appears to result in only moderate performance loss[130]. For an illustration of the effect of c on equitability, see Figure C.8, which graphs equitability on our set of functional relationships against runtime as α, n, and c are varied.

C.9 Example of a noisy functional relationship that leads to poor equitability

Here we prove the claim that under some noise models a step function will lead to poor equitability for all three of the methods that showed non-trivial equitability in our empirical simulations: MIC_e, mutual information, and maximal correlation. We first define the family of relationships in question.

Definition C.9.1. Let X be uniformly distributed over $[0, 1]$ and define

$$f(x) = \begin{cases} -1 & x \in [0, 1/2] \\ \\ 1 & x \in (1/2, 1] \end{cases}.$$

For $\varepsilon \geq 0$, define $S_\varepsilon = (X, f(X) + U_\varepsilon)$, where U_ε is uniformly distributed over $[-\varepsilon, \varepsilon]$ and is independent of X.

We now prove that S_ε will receive high scores from all three methods, over a wide range of R^2 values. In the proposition below, MIC_*, I, and C denote the population value of MIC_e, mutual information in bits, and the population maximal correlation respectively.

Proposition C.9.2. *For $\varepsilon < 1$, the relationship S_ε satisfies $MIC_*(S_\varepsilon) = 1$, $I(S_\varepsilon) = 1$, and $C(S_\varepsilon) = 1$.*

Proof. The key observation is that when $\varepsilon < 1$, the distribution of S_ε has the property that the support of its y-values when $x \in [0, 1/2]$ is disjoint from the support of its y-values when $x \in (1/2, 1]$. This yields the result, via different arguments, for all three methods.

$\text{MIC}_*(S_\varepsilon) \geq 1$ because when $\varepsilon < 1$, we can construct a two-by-two grid whose mutual information is 1 (by equipartitioning both axes into two parts). Since $\text{MIC}_* \leq 1$ always, this means that $\text{MIC}_*(S_\varepsilon) = 1$.

To show that $I(S_\varepsilon) = 1$, we write $I(X;Y) = H(X) - H(X|Y)$ where H denotes differential entropy. For S_ε, $H(X) = \log_2(1-0) = \log_2(1) = 0$. And since $X|Y$ is always uniformly distributed over an interval of length $1/2$, we have $H(X|Y) = \log_2(1/2) = -1$, from which the result follows.

For C, we notice that when $\varepsilon < 1$ then the function

$$g(y) = \begin{cases} -1 & y \in [-1-\varepsilon, -1+\varepsilon] \\ 1 & y \in [1-\varepsilon, 1+\varepsilon] \\ 0 & \text{otherwise} \end{cases}$$

is well defined. We also notice that $(f(X), g(f(X) + U_\varepsilon))$ is uniformly distributed over $\{(-1,-1),(1,1)\}$. Since the correlation of this distribution is 1, and since $C(X,Y) \geq C(u(X), v(Y))$ for all measurable functions u, v, it follows that $C(S_\varepsilon) \geq 1$. Finally, since $C(X,Y) \leq 1$ always, we have that $C(S_\varepsilon) = 1$. □

Together with the observation that $R^2(S_\varepsilon) = 1/(1+\varepsilon^2)$, this yields the following corollary. Recall that an equitability of 1 represents worst possible performance and an equitability of ∞ represents perfect equitability.

Corollary C.9.3. *The equitability with respect to R^2 of MIC_*, mutual information, and maximal correlation on the set $\mathcal{Q} = \{S_\varepsilon : 0 \leq \varepsilon \leq \infty\}$ is at most 2.*

D

Supplementary information for Chapter 5

D.1 Model and estimands

D.1.1 The model

Let M be the length of the genome. Given a genotype vector $x \in \mathbb{R}^M$ of an individual sampled randomly from some population distribution and a vector $\beta \in \mathbb{R}^M$ of causal

SNP effects, we model the phenotype y with a standard linear model:

$$y|\beta, x \sim \mathcal{N}(x^T\beta, \sigma_e^2). \tag{D.1}$$

We assume that the genotypes are standardized in the population, i.e., that $E(x_m) = 0$ and $E(x_m^2) = 1$ for all SNPs m. We assume the same of the phenotype: $E(y) = 0$ and $E(y^2) = 1$. Because our GWAS sample will be very large, these assumptions are for expositional convenience only.

The last ingredient of our model is the connection between β and the signed functional annotation of interest $v \in \mathbb{R}^M$. To get this, we assume that β is sampled from a distribution satisfying

$$E(\beta|v) = \mu v, \quad \text{cov}(\beta|v) = \sigma^2 I \tag{D.2}$$

where μ and σ are scalars.

D.1.2 The estimands

The first estimand we might be interested in is μ, which would tell us the expected change in the per-normalized-genotype effect β_m of SNP m for every unit increase of v_m. However, this estimand depends on the units of v: if we multiply v by a constant c, then μ is decreased by a factor of c. We therefore introduce a second estimand, the *functional correlation* r_f, which is defined as the genetic correlation between y and the

100%-heritable phenotype $x^T v$, i.e.,

$$r_f := \text{corr}(x^T\beta, x^T v). \tag{D.3}$$

Under our model,

$$\begin{aligned}\text{cov}(x^T\beta, x^T v) &= E(\beta^T x x^T v) &\text{(D.4)}\\ &= E(\beta)^T E(xx^T)v &\text{(D.5)}\\ &= \mu v^T R v &\text{(D.6)}\end{aligned}$$

where $R = E(xx^T) \in \mathbb{R}^{M\times M}$ is the (signed) population LD matrix of the genotypes, and v is fixed and known. Since

$$\text{var}(x^T v) = E(v^T x x^T v) = v^T R v, \tag{D.7}$$

we obtain

$$r_f = \frac{\text{cov}(x^T\beta, x^T v)}{\sqrt{\text{var}(x^T\beta)\text{var}(x^T v)}} = \mu\sqrt{\frac{v^T R v}{h_g^2}}. \tag{D.8}$$

where $h_g^2 = \text{var}(x^T\beta)$ is the SNP-heritability of the phenotype. Note that r_f can also be derived under a model in which v is also modeled as random and jointly distributed with β, in which case r_f is equal to a standard random-effects genetic correlation[15]. The choice to model v as fixed here arises from the fact that, since it is a complicated

biological object, we wish to make as few assumptions as possible about its structure.

In addition to μ and r_f, we might wish to know how much total phenotypic variance is explained by the signed contribution of v to β. This parameter, h_v^2, is defined by

$$h_v^2 := \mathrm{var}(\mu x^T v) = \mu^2 v^T R v. \tag{D.9}$$

This is equal to the prediction r^2 that we would obtain if we tried to predict y from $x^T v$. If we scale h_v^2 by the total heritability of y, we obtain the proportion of heritability explained by the signed contribution of v, i.e.,

$$\frac{h_v^2}{h_g^2} = \frac{\mu^2 v^T R V}{h_g^2} = r_f^2. \tag{D.10}$$

We remark that for annotations with small support, r_f and its associated quantities should generally expected to be small in magnitude. To see this, define $h_{|v|}^2$ to be the prediction r^2 that we would obtain if we predicted y from an optimal predictor that was constrained to be zero outside the support of v. By construction we have $h_v^2 \leq h_{|v|}^2$, but since $h_{|v|}^2$ is the total phenotypic variance explained by SNPs in the support of v, this implies that $r_f^2 \leq h_v^2/h_g^2 \leq h_{|v|}^2/h_g^2$ is at most the proportion of heritability explained by the SNPs in the support of v.

D.2 Derivations and description of method

D.2.1 Main derivation

Now suppose that N individuals $x_1, \ldots, x_N$ have been sampled i.i.d. from the population with corresponding phenotypes $y_1, \ldots y_N$, and that we are given the vector of marginal correlations between each SNP and the trait, i.e., we are given

$$\hat{\alpha} := \frac{1}{N} \sum_{n=1}^{N} x_n y_n \in \mathbb{R}^M. \tag{D.11}$$

It is easily shown that $E(\hat{\alpha}|\beta) = R\beta$ (see Proposition D.7.2), from which it follows that

$$E(\hat{\alpha}|v) = E(E(\hat{\alpha}|\beta, v)|v) \tag{D.12}$$

$$= E(R\beta|v) \tag{D.13}$$

$$= \mu R v. \tag{D.14}$$

This means that naive regression of $\hat{\alpha}$ on the *signed LD profile* Rv of v is an unbiased estimator of μ. However, ordinary least-squares is optimal only when the observations have i.i.d. noise. In this regression, each SNP provides one observation $(\hat{\alpha}_m, (Rv)_m)$, but under our model the covariance of $\hat{\alpha}_m$ and $\hat{\alpha}_{m'}$ given Rv is non-zero. Therefore, if we can model this covariance structure properly, we should be able to use generalized

least-squares to reduce variance and increase power. In Theorem D.7.4, we show that indeed

$$\text{cov}(\hat{\alpha}|v) \approx \sigma^2 R^2 + \frac{R}{N} =: \Omega. \tag{D.15}$$

The default version of signed LD profile regression estimates Ω from the reference panel and the chi-squared statistics of the GWAS in question and then performs generalized least-squares using a pseudo-inverse of Ω to de-couple correlated errors among SNPs. It can be shown that if a) all causal SNPs are typed, b) sample size is infinite, and c) R is invertible, this method is equivalent to estimating β via $R^{-1}\hat{\alpha}$ and then regressing this estimate on v to obtain μ, which is the optimal approach in that setting. Note that because we generate P-values for hypothesis testing empirically (see below), we are guaranteed that our generalized least-squares scheme will remain well-calibrated even if our estimate of the matrix Ω is inaccurate due to, e.g., mis-match between the reference panel and the study population.

The point estimate arising from the regression described above is an estimate $\hat{\mu}$ of μ. To obtain an estimate of r_f, we plug into Equation D.3, estimating h_g^2 using the "aggregate estimator" of heritability[40] given by

$$\hat{h_g^2} := \frac{|\hat{\alpha}|_2^2 - \frac{M}{N}}{\frac{1}{M_{5,50}} \sum_m \widehat{\ell_m}} \tag{D.16}$$

where $|\hat{\alpha}|_2$ is the ℓ_2-norm of $\hat{\alpha}$, $\widehat{\ell_m}$ is a reference-panel-based estimate of the LD-

score $\ell_m := \sum_{m'} R^2_{mm'}$ of SNP m, and $M_{5,50}$ is the number of causal SNPs with MAF between 5% and 50%. Equation D.3 also has a $v^T Rv$ term; for convenience we approximate this term by $v^T v$; our simulations show that we do not suffer from this approximation, and it is empirically quite accurate for our annotations (data not shown).

To estimate h^2_v/h^2_g, we use the jackknife to estimate the sampling variance $\widehat{\tau^2}$ of the statistic $\widehat{r_f}$, and then report $\widehat{r_f}^2 - \widehat{\tau^2}$. Though this is an exactly unbiased estimate of h^2_v only if $\widehat{r_f}$ is normally distributed and the jackknife provides an accurate estimate of the sampling variance of μ, our simulations show that it is very close to unbiased in practice. Note that while we use a jackknife estimate of the variance of $\widehat{r_f}$ to estimate r^2_f, this is not how we compute P-values for null hypothesis testing; for details of null hypothesis testing, see below.

To estimate h^2_v, we simply multiply our estimate of $r^2_f = h^2_v/h^2_g$ by our estimate of h^2_g.

D.2.2 Untyped SNPs

Typically, our set of potentially causal SNPs is much larger than the set of SNPs for which we have GWAS summary statistics. Signed LD profile works well in such scenarios: it simply uses only the entries of Rv corresponding to typed SNPs in the regression. Because drastically different sets of typed SNPs require estimation of Ω anew, we estimate Ω assuming that all non-MHC HapMap3 SNPs are typed, and then restrict the summary statistics for each trait analyzed to non-MHC HapMap3 SNPs only.

D.2.3 Null hypothesis testing

To test the null hypothesis $H_0 : \mu = 0$ (or, equivalently, $H_0 : r_f = 0$), we split the genome into approximately 300 blocks of approximately the same size with the block boundaries constrained to fall on estimated recombination hotspots[10]. We then define the null distribution of our statistic as the distribution arising from independently multiplying v by an independent random sign for each block. We perform this empirical sign-flipping many times to obtain an approximation of the null distribution and corresponding P-values. Our use of sign-flipping ensures that any true positives found by our method are the result of genuine first-moment effects; if in contrast we estimated standard errors using least-squares theory or a re-sampling method such as the jackknife or bootstrap, our method might inappropriately reject the null hypothesis only because the variance of β is higher in parts of the genome where Rv is large in magnitude. This would make our method susceptible to confounding due to unsigned enrichments, as might arise from the co-localization of TF binding sites with enriched regulatory elements such as enhancer regions. Additionally, the fact that we flip the signs of SNPs in each block together ensures that our null distribution preserves any potential relationship of our annotation to the LD structure of the genome. In choosing how many blocks to use for this procedure, we took into account that i) the fewer blocks we use the fewer assumptions we make about LD structure and the faster we can compute P-values, and ii) the more blocks we use the higher the precision of the P-values that we can obtain.

Our choice to use 300 blocks is a compromise between these two considerations.

D.2.4 Controlling for covariates and the signed background model

Given a signed covariate $u \in \mathbb{R}^M$, we can perform inference on the signed effect of v conditional on u. This is done by first regressing Ru out of $\hat{\alpha}$ and out of Rv using the generalized least-squares method outlined above, and then proceeding as usual with the residuals of $\hat{\alpha}$ and Rv. This can be done simultaneously for multiple covariates u.

Unless stated otherwise, all analyses in this paper are done controlling for a "signed background model" consisting of 5 annotations $u^1, \ldots, u^5$, defined by

$$u_m^i = \mathbf{1}\left\{\text{MAF}_m \text{ is in } i\text{-th quintile}\right\} \sqrt{2\text{MAF}_m(1-\text{MAF}_m)^{1+\alpha_s}} \tag{D.17}$$

where MAF_m is the minor allele frequency of SNP m and α_s is a parameter describing the MAF-dependence of the signed effect of minor alleles on phenotype. Based on the literature on MAF-dependence of the unsigned effects $\text{var}(\beta_m)$, we set $\alpha_s = -0.3$[139].

D.3 Computational considerations

We model the LD matrix R as being block-diagonal, with the block endpoints defined by recombination hotspots[10]. This allows both more statistically efficient estimation of the true Rv as well as more efficient computation.

For estimating Ω, we use the above block-diagonal decomposition, together with a

truncated singular value decomposition applied in each block. Specifically, we store enough singular vectors to capture 95% of the spectrum of each LD block. This is a pre-processing step that need only be carried out once per reference panel, and the relevant outputs of this step for the 1000G Phase 3 Europeans can be downloaded from our website.

D.4 Additional interpretation of results

We list complementary evidence for the associations in Table 5.1 that are not discussed in the main text.

POL2/TBP and Crohns (+) We found several associations between binding of RNA polymerase II (POL2) and Crohn's disease, matched by an association with the same direction of effect for TATA binding protein (TBP), which forms part of the transcription pre-initiation complex. These associations are consistent with the theory that, due to the time-scale required for many immune responses, those responses tend to involve up-regulation rather than down-regulation of gene expression since the former can be done more quickly than the latter. This can either happen via rapid, de novo recruitment of RNA polymerase,[27] or via release of a "poised" state wherein POL2 is present at the gene promoter but remains paused until an additional signal quickly triggers transcription[100,170]. This theory predicts that we would see positive associations for many POL2 annotations provided they contain enough genes to have a substantial

number of immune genes, and that the association should be stronger for annotations that are biased toward POL2 binding specifically near immune genes. Indeed, the most significant association we detected was for a POL2 annotation in the lymphoblastoid cell line GM18951, and using the ChIP-seq enrichment tool GREAT[92], we found that the ChIP-seq peaks used to create this annotation showed enrichment in several gene ontology biological processes corresponding to immune response programs such as "antigen binding" (3.2x enrichment, binomial $p = 1.57 \times 10^{-33}$), "cellular response to type I interferon" (4.4x enrichment, binomial $p = 1.76 \times 10^{-84}$), and "interferon-gamma-mediated signalling pathway" (2.8x enrichment, $p = 5.0 \times 10^{-46}$). Thus, these results may provide evidence that, due to the bias of the immune system toward up-regulation, variants that increase the amount of bound polymerase at immune response genes impart an increased risk of Crohn's disease.

SP1 AND ANOREXIA (-) There is evidence that increased SP1 activity is protective for several psychiatric conditions[20,45]. Separately, SP1 has been shown to be regulated by insulin levels[107,175] and in turn to modulate expression in the hypothalamus of POMC, an important regulator of appetite[180,172]. It has also been shown to have a key role in the induction of leptin following insulin-stimulated glucose metabolism in adipocytes[96]. However, since SP1 has many binding partners, it could also be that this association is driven by the binding of one those partners.

FOS AND HDL (+) In mice, liver-specific overexpression of the *FOS* gene leads to increased intrahepatic cholesterol and modulation of genes in metabolic pathways connected to cholesterol and fatty acid biosynthesis[5]. FOS has also been shown to be up-regulated when HeLa cells are grown in a sterol-depleted medium designed to activate cellular sterol homeostatic machinery[7], and the AP-1 complex that it forms has been shown to be down-regulated by high-cholesterol diet in model organisms[68]. A different mechanism is suggested by the fact that in humans, a mutation in the *FOS* gene is associated with congenital generalized lipodystrophy, a phenotype characterized by absence of adipocytes[77].

CTCF AND ECZEMA (+) Cell-type-specific deletion of CTCF in Langerhans cells (skin-resident dendritic cells) leads to a reduced pool of systemic dendritic cells and a much reduced pool of Langerhans cells[73]. More generally, CTCF has been shown to alter immune processes[117,159,106,102], and a genetic variant associated with risk of atopy has been shown to alter CTCF binding[138]. Another possible mechanism for our result is suggested by the fact that the cohesin complex, which frequently binds to DNA in the same locations as CTCF, is necessary for maintenance of epidermal progenitor cells[103]. A third possibility is that, given the negative relationship between CTCF binding and methylation, together with the genome-wide depletion of methylation observed in atopy[74], this association could be explained by SNPs that alter CpG islands in ways that affect methylation and therefore CTCF binding.

ELF1 AND CROHN'S (+) ELF1 is a hematopoietic and immune regulator[46] that lies in a genome-wide significant Crohn's disease locus in a GWAS of a Japanese population[171]. Alterations of ELF1 expression are associated with autoimmunity[46]. Additionally, its binding sites are enriched for autoimmune risk SNPs[38] and are overrepresented among promoters of genes that are differentially expressed in inflammatory bowel disease intestinal tissue relative to control tissue[112].

E2F1 AND CROHN'S (+) E2F1 has roles in immunity, and E2f1-deficient mice challenged with lipopolysaccharide exhibit an attenuated inflammatory response[167]. Additionally, chronic colonic inflammation is associated with release of E2F1 inhibition and activation of E2F1 target genes[177]. Finally, activity of RB, an upstream regulator of the E2F1 pathway, is a highly sensitive and specific test for distinguishing Crohn's disease from ulcerative colitis in some cases, with RB activity being elevated in Crohn's disease[145].

ETS1 AND CROHN'S (+) ETS1 is known to regulate genes involved in immunity[46] and the *ETS1* gene lies in GWAS loci for multiple autoimmune traits[55,173]. Absence of ETS1 leads to T lymphocytes that respond more weakly to stimuli and are fewer in number[101,37], as well as to abnormal B lymphocyte differentiation[11]. Additionally, adoptive transfer of Th cells with no functional copies of the *Ets-1* gene fails to induce colitis in severe combined immunodeficient (SCID) mice, whereas Th cells with func-

tional copies of the gene do induce colitis[50]. Finally, viable *Ets-1*-deficient mice are protected from inflammatory bowel disease relative to wild-type mice[97].

D.5 Supplementary Tables

Table D.1: Summary information about ChIP-seq annotations used in analyses. v denotes annotation, M denotes the total number of SNPs in the reference panel, $|v|_0$ denotes the number of SNPs with non-zero values of v, and $|v|_2$ denotes the 2-norm of v.

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
HAIB	SKNSHRA	CTCF	0.880098	18646	0.19	13.20
BROAD	NHA	CTCF	0.869841	27912	0.28	12.68
HAIB	A549	CTCFSC5916	0.866840	21517	0.22	12.73
UW	NB4	CTCF	0.866150	25419	0.25	13.23
UW	HRE	CTCF	0.864149	28846	0.29	13.64
HAIB	A549	CTCFSC5916	0.863801	21011	0.21	13.41
UTA	HUVEC	CTCF	0.861944	21000	0.21	14.18
BROAD	HUVEC	CTCF	0.859699	29576	0.30	12.68
UW	HFF	CTCF	0.859124	25034	0.25	11.61
UW	RPTEC	CTCF	0.858547	44995	0.45	17.53
BROAD	HMEC	CTCF	0.858372	27488	0.27	12.58
UW	HASP	CTCF	0.858100	29663	0.30	14.75
UW	GM12878	CTCF	0.858056	25981	0.26	13.11
UW	A549	CTCF	0.857446	35097	0.35	15.54
UW	HFFMYC	CTCF	0.857241	38004	0.38	14.93
UTA	GM12878	CTCF	0.856204	24907	0.25	15.67
UW	GM06990	CTCF	0.855834	33120	0.33	14.51
UW	HMF	CTCF	0.854815	35825	0.36	16.13
UW	HCFAA	CTCF	0.854650	26214	0.26	13.36
UW	GM12874	CTCF	0.854489	24822	0.25	12.73
UW	HEK293	CTCF	0.854351	31140	0.31	15.48
UTA	HEPG2	CTCF	0.853428	17547	0.18	13.62
UW	MCF7	CTCF	0.852776	40427	0.40	17.06
UW	NHEK	CTCF	0.852312	31784	0.32	13.27
HAIB	H1HESC	CTCFSC5916	0.852040	30644	0.31	18.33
UW	HVMF	CTCF	0.851735	33859	0.34	14.79
UW	GM12875	CTCF	0.851254	26436	0.26	13.21
UW	HCT116	CTCF	0.851195	36485	0.36	15.57
UW	GM12865	CTCF	0.850843	29599	0.30	14.14

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
HAIB	HEPG2	CTCFSC5916	0.850684	29285	0.29	17.25
UW	HRPE	CTCF	0.850296	33503	0.34	16.27
BROAD	H1HESC	CTCF	0.849116	47350	0.47	20.96
UW	GM12872	CTCF	0.847288	34212	0.34	15.09
SYDH	H1HESC	RAD21	0.846410	35780	0.36	17.12
UW	BE2C	CTCF	0.846211	41476	0.41	15.80
UW	HPF	CTCF	0.845889	29441	0.29	14.13
UW	NHLF	CTCF	0.845237	24971	0.25	11.64
BROAD	NHDFAD	CTCF	0.844702	33708	0.34	14.84
UW	SAEC	CTCF	0.843178	27722	0.28	13.59
BROAD	HSMMT	CTCF	0.843109	39253	0.39	14.10
BROAD	GM12878	CTCF	0.842508	39752	0.40	14.28
BROAD	NHLF	CTCF	0.842394	30215	0.30	12.99
UW	HELAS3	CTCF	0.842036	24028	0.24	11.95
UW	GM12864	CTCF	0.841830	33480	0.33	14.86
UW	SKNSHRA	CTCF	0.841702	26551	0.27	13.96
UW	HCM	CTCF	0.839966	42907	0.43	15.57
UTA	GLIOBLA	CTCF	0.839859	37388	0.37	18.58
UTA	K562	CTCF	0.838050	27610	0.28	16.98
UW	HUVEC	CTCF	0.837666	23780	0.24	12.51
UW	K562	CTCF	0.835751	30678	0.31	14.23
UW	GM12873	CTCF	0.834805	36107	0.36	15.83
UW	HMEC	CTCF	0.834803	36092	0.36	14.96
BROAD	HEPG2	CTCF	0.834631	36924	0.37	14.72
BROAD	HSMM	CTCF	0.833446	34415	0.34	15.13
UW	HEPG2	CTCF	0.831350	31010	0.31	15.52
UW	HPAF	CTCF	0.830419	40688	0.41	16.57
UW	AG09309	CTCF	0.830321	31862	0.32	13.56
BROAD	HELAS3	CTCF	0.828969	49347	0.49	15.31
UW	BJ	CTCF	0.828852	32555	0.33	13.39
BROAD	NHEK	CTCF	0.828230	37413	0.37	14.19
UW	HEE	CTCF	0.828217	33823	0.34	13.55
UW	HAC	CTCF	0.828210	36662	0.37	13.83
UTA	HELAS3	CTCF	0.828109	25915	0.26	16.07
UW	AG04450	CTCF	0.827331	32761	0.33	13.88

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
UTA	PROGFIB	CTCF	0.826811	22840	0.23	14.38
HAIB	ECC1	CTCFC	0.826438	15251	0.15	8.81
BROAD	DND41	CTCF	0.824320	38541	0.39	13.81
HAIB	H1HESC	RAD21	0.823698	47411	0.47	22.20
SYDH	IMR90	CTCFB	0.820777	26982	0.27	13.99
UW	AG09319	CTCF	0.820556	33669	0.34	14.46
UW	HBMEC	CTCF	0.819613	41152	0.41	16.62
UW	WI38	CTCF	0.819609	25725	0.26	10.62
UTA	H1HESC	CTCF	0.818739	22472	0.22	15.80
UTA	A549	CTCF	0.817553	32700	0.33	17.81
UW	AG10803	CTCF	0.817006	29517	0.30	13.69
BROAD	OSTEOBL	CTCF	0.816996	53644	0.54	16.04
UW	HCPE	CTCF	0.816798	42276	0.42	16.83
SYDH	GM12878	CTCFSC15914C20	0.815991	30691	0.31	15.49
UTA	MCF7	CTCF	0.815467	49073	0.49	22.63
BROAD	K562	CTCF	0.815351	52427	0.52	15.60
UW	WERIRB1	CTCF	0.815231	30972	0.31	15.58
UTA	MCF7	CTCF	0.814259	37438	0.37	18.94
UW	AOAF	CTCF	0.810198	25402	0.25	12.89
UW	CACO2	CTCF	0.808883	28146	0.28	12.68
UW	AG04449	CTCF	0.808085	24368	0.24	14.42
SYDH	K562	CTCFB	0.807922	34266	0.34	15.56
HAIB	HEPG2	RAD21	0.806753	31414	0.31	14.66
UW	NHDFNEO	CTCF	0.805912	34150	0.34	13.07
UTA	FIBROBL	CTCF	0.802580	24917	0.25	14.54
HAIB	K562	CTCFC	0.800330	29034	0.29	14.24
SYDH	HEPG2	RAD21	0.795326	24061	0.24	10.74
SYDH	GM12878	RAD21	0.793772	22165	0.22	9.93
UTA	GM19240	CTCF	0.787095	24254	0.24	14.44
UTA	GM19238	CTCF	0.784621	28109	0.28	15.19
UTA	NHEK	CTCF	0.782123	28029	0.28	15.70
HAIB	T47D	CTCFSC5916	0.780735	20119	0.20	9.44
UTA	GM12891	CTCF	0.776692	23165	0.23	13.77
SYDH	GM12878	SMC3AB9263	0.775055	22604	0.23	9.36
HAIB	GM12878	RAD21	0.773313	19232	0.19	10.90

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
UTA	MCF7	CTCF	0.771586	32289	0.32	17.54
SYDH	IMR90	RAD21	0.771096	21035	0.21	10.62
UTA	GM19239	CTCF	0.770649	21921	0.22	12.29
UTA	GM12892	CTCF	0.764533	27003	0.27	14.40
SYDH	K562	SMC3AB9263	0.764408	17833	0.18	8.29
HAIB	K562	RAD21	0.762473	17349	0.17	10.54
UW	HL60	CTCF	0.760612	11834	0.12	6.43
SYDH	HEPG2	MAFKAB50322	0.756003	36764	0.37	16.31
SYDH	HEK293	POL2	0.750713	11423	0.11	2.57
HAIB	SKNSHRA	RAD21	0.748781	34221	0.34	14.81
UTA	MCF7	CTCF	0.744677	33804	0.34	16.07
UTA	A549	POL2	0.743474	13317	0.13	2.99
UTA	MCF7	CTCF	0.737779	31703	0.32	15.80
SYDH	HELAS3	RAD21	0.732822	23726	0.24	9.90
UTA	GLIOBLA	POL2	0.730622	12444	0.12	2.89
SYDH	A549	RAD21	0.726374	15727	0.16	8.17
SYDH	GM10847	POL2	0.725536	11162	0.11	2.82
SYDH	K562	RAD21	0.719791	11216	0.11	5.92
UTA	HUVEC	POL2	0.710965	9848	0.10	2.62
SYDH	GM18526	POL2	0.704244	15927	0.16	3.59
SYDH	HELAS3	SMC3AB9263	0.703877	25410	0.25	9.28
SYDH	MCF10AES	CFOS	0.695666	52371	0.52	14.00
SYDH	GM15510	POL2	0.692228	18641	0.19	3.92
SYDH	GM12878	ZNF143166181AP	0.691695	16121	0.16	6.52
SYDH	MCF10AES	CFOS	0.689921	41778	0.42	11.91
SYDH	HEPG2	SMC3AB9263	0.683574	21539	0.22	8.17
SYDH	MCF10AES	CFOS	0.678308	49334	0.49	12.33
SYDH	MCF10AES	CFOS	0.672546	37719	0.38	10.03
SYDH	H1HESC	ZNF143	0.665846	25229	0.25	8.50
SYDH	GM18951	POL2	0.662339	23305	0.23	4.19
SYDH	K562	NFYB	0.661296	9570	0.10	3.91
HAIB	GM12878	GABP	0.660956	5625	0.06	2.43
HAIB	ECC1	POL2	0.657365	19849	0.20	3.32
UTA	MCF7	POL2	0.652882	18193	0.18	3.05
HAIB	HEPG2	TAF1	0.650101	16181	0.16	2.94

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
SYDH	K562	IRF1	0.649426	12976	0.13	3.16
SYDH	K562	POL2	0.647737	16308	0.16	3.35
SYDH	GM12892	POL2	0.645338	23295	0.23	4.12
SYDH	HEPG2	MAFKSC477	0.643218	24770	0.25	9.07
UTA	MCF7	POL2	0.642949	15229	0.15	2.94
SYDH	NB4	POL2	0.641432	16158	0.16	3.31
SYDH	K562	POL2	0.640277	15063	0.15	2.99
SYDH	K562	POL2	0.635903	17161	0.17	3.22
SYDH	K562	ZNF143	0.634772	23343	0.23	7.50
SYDH	HEPG2	MAFFM8194	0.634067	25009	0.25	8.93
HAIB	GM12878	ELF1SC631	0.631869	20946	0.21	5.37
HAIB	H1HESC	TAF1	0.627966	21837	0.22	3.08
HAIB	HEPG2	GABP	0.627412	9290	0.09	3.07
SYDH	HEPG2	CEBPB	0.625633	34970	0.35	14.15
SYDH	K562	POL2	0.624054	15843	0.16	3.14
SYDH	IMR90	MAFK	0.620883	25154	0.25	8.57
SYDH	GM18505	POL2	0.618220	24625	0.25	3.97
UTA	HELAS3	POL2	0.617348	19384	0.19	3.25
UTA	PROGFIB	POL2	0.617226	14761	0.15	2.91
SYDH	GM19099	POL2	0.606235	22799	0.23	4.01
SYDH	GM19193	POL2	0.604915	24050	0.24	3.91
SYDH	K562	POL2	0.602457	15110	0.15	2.90
HAIB	SKNSH	TAF1	0.601160	11185	0.11	2.76
SYDH	HCT116	POL2	0.598756	17455	0.17	2.72
SYDH	PBDE	POL2	0.596470	22492	0.22	3.29
HAIB	K562	TAF1	0.594640	13400	0.13	3.11
UTA	MCF7	POL2	0.587761	14677	0.15	2.73
SYDH	MCF10AES	POL2	0.581721	22034	0.22	3.45
BROAD	K562	PLU1	0.578953	19126	0.19	2.78
SYDH	IMR90	CEBPB	0.577892	44228	0.44	14.66
HAIB	A549	CREB1SC240	0.576054	13155	0.13	3.07
UTA	K562	POL2	0.575441	19966	0.20	3.30
HAIB	GM12878	PU1	0.574256	27757	0.28	9.34
SYDH	GM12878	POL2	0.573648	23803	0.24	3.93
UTA	GM12878	POL2	0.572056	17552	0.18	3.00

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
HAIB	GM12878	NRSF	0.568899	5888	0.06	3.82
BROAD	K562	PHF8A301772A	0.566331	27457	0.27	2.88
SYDH	RAJI	POL2	0.564973	21621	0.22	3.36
SYDH	HEPG2	POL2	0.563102	18212	0.18	2.71
HAIB	K562	YY1	0.558414	10704	0.11	2.79
HAIB	A549	POL2	0.555363	31308	0.31	3.68
HAIB	A549	POL2	0.553825	29976	0.30	3.58
HAIB	GM12878	YY1SC281	0.553334	26103	0.26	5.34
SYDH	GM12878	POL2	0.552473	11117	0.11	2.41
HAIB	GM12891	PU1	0.551608	28912	0.29	9.97
HAIB	GM12878	TAF1	0.551273	12105	0.12	2.98
SYDH	A549	CEBPB	0.551046	26389	0.26	9.72
SYDH	HUVEC	CFOS	0.550936	42775	0.43	7.57
HAIB	A549	TAF1	0.550319	11038	0.11	2.08
HAIB	GM12892	POL2	0.548292	23439	0.23	3.42
HAIB	HELAS3	TAF1	0.547530	14406	0.14	2.81
HAIB	HEPG2	POL24H8	0.547414	18782	0.19	3.01
SYDH	HEPG2	JUND	0.545643	23439	0.23	5.68
SYDH	HELAS3	HAE2F1	0.544870	9314	0.09	1.47
SYDH	HELAS3	POL2	0.543185	29222	0.29	3.11
HAIB	GM12892	TAF1	0.542027	8249	0.08	2.23
SYDH	K562	MAZAB85725	0.541193	33691	0.34	6.34
SYDH	MCF10AES	POL2	0.541022	25900	0.26	3.53
SYDH	H1HESC	MAFK	0.540650	8262	0.08	2.09
HAIB	A549	ETS1	0.539878	6635	0.07	2.60
SYDH	GM12891	POL2	0.538971	24040	0.24	3.79
HAIB	K562	GABP	0.535852	12143	0.12	3.59
HAIB	K562	E2F6	0.535787	20429	0.20	2.89
HAIB	HEPG2	YY1SC281	0.535256	17564	0.18	3.27
HAIB	HCT116	POL24H8	0.534399	29439	0.29	4.18
SYDH	HELAS3	ELK4	0.533836	6984	0.07	2.00
HAIB	U87	NRSF	0.533645	10740	0.11	3.53
SYDH	H1HESC	TBP	0.533586	17933	0.18	3.13
SYDH	GM12878	ELK112771	0.532557	5585	0.06	1.90
UTA	H1HESC	POL2	0.528904	15666	0.16	2.28

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
HAIB	HEPG2	POL2	0.527603	26528	0.27	3.51
HAIB	GM12878	PMLSC71910	0.523565	21007	0.21	3.16
HAIB	HEPG2	NRSF	0.522989	11697	0.12	3.82
HAIB	K562	ELF1SC631	0.521651	20676	0.21	5.35
SYDH	GM12878	NFYB	0.521437	14633	0.15	3.58
HAIB	GM12891	TAF1	0.520083	10825	0.11	2.70
HAIB	HUVEC	POL2	0.519612	24168	0.24	3.11
HAIB	A549	ELF1	0.516848	8792	0.09	2.24
HAIB	PFSK1	FOXP2	0.514938	15908	0.16	2.79
SYDH	MCF10AES	E2F4	0.514526	12559	0.13	2.58
SYDH	HELAS3	NFYA	0.513807	5483	0.05	1.98
SYDH	K562	HMGN3	0.513410	18241	0.18	2.26
SYDH	HELAS3	NFYB	0.512540	6653	0.07	2.22
SYDH	HUVEC	CJUN	0.510520	20080	0.20	4.26
HAIB	HUVEC	POL24H8	0.509722	35149	0.35	4.72
HAIB	HEPG2	ELF1SC631	0.509441	13489	0.13	3.73
SYDH	K562	MAFKAB50322	0.508412	13001	0.13	3.37
HAIB	GM12891	POL2	0.505543	17852	0.18	2.78
SYDH	H1HESC	USF2	0.503572	5202	0.05	2.27
HAIB	H1HESC	GABP	0.501419	5292	0.05	1.53
SYDH	K562	E2F4	0.500739	7900	0.08	1.74
SYDH	K562	MAFF	0.499311	17035	0.17	4.41
SYDH	IMR90	POL2	0.499139	21099	0.21	2.57
HAIB	H1HESC	USF1	0.498243	16631	0.17	6.39
HAIB	K562	MAX	0.494249	42934	0.43	5.98
SYDH	HELAS3	POL2S2	0.492278	14434	0.14	2.32
HAIB	H1HESC	NRSF	0.491469	8454	0.08	5.74
SYDH	HELAS3	MAZAB85725	0.489070	16019	0.16	2.24
HAIB	HELAS3	NRSF	0.488734	6360	0.06	4.97
HAIB	GM12891	YY1SC281	0.487772	11490	0.11	2.73
HAIB	HEPG2	SIN3AK20	0.487522	17653	0.18	2.53
HAIB	HELAS3	POL2	0.487393	28715	0.29	3.64
HAIB	K562	POL2	0.486825	36854	0.37	3.37
SYDH	HEPG2	MAX	0.486481	11059	0.11	1.92
HAIB	GM12878	SP1	0.486260	15317	0.15	3.48

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
SYDH	HEPG2	POL2	0.484689	20477	0.20	2.83
HAIB	GM12892	POL24H8	0.483645	20500	0.21	2.59
HAIB	K562	ETS1	0.483398	10444	0.10	2.37
SYDH	GM12878	MAZAB85725	0.483322	22411	0.22	3.16
SYDH	HELAS3	CJUN	0.478779	16492	0.16	2.98
SYDH	K562	CFOS	0.478299	5481	0.05	2.17
SYDH	HEPG2	MXI1	0.477728	21106	0.21	3.26
HAIB	H1HESC	POL2	0.476246	26239	0.26	2.59
SYDH	K562	CEBPB	0.474134	28505	0.29	9.12
HAIB	U87	POL24H8	0.473137	23582	0.24	3.29
SYDH	K562	MAX	0.471849	29516	0.30	4.86
HAIB	A549	GABP	0.471447	13855	0.14	3.02
SYDH	HELAS3	CHD2	0.471053	19320	0.19	3.33
SYDH	K562	E2F6	0.470723	16483	0.16	2.33
HAIB	GM12878	EGR1	0.468941	10841	0.11	2.08
SYDH	HUVEC	MAX	0.466519	6425	0.06	1.93
HAIB	GM12878	RUNX3SC101553	0.466113	56840	0.57	8.61
HAIB	GM12878	USF1	0.465793	7272	0.07	2.57
HAIB	K562	USF1	0.464692	12871	0.13	4.61
BROAD	K562	RBBP5A300109A	0.463994	20083	0.20	1.84
SYDH	K562	TBP	0.463143	17767	0.18	3.22
HAIB	K562	SIN3AK20	0.463116	8897	0.09	1.77
SYDH	K562	CMYC	0.462873	32161	0.32	5.06
SYDH	A549	MAX	0.461439	9266	0.09	1.72
SYDH	HELAS3	MAX	0.458337	29171	0.29	4.12
HAIB	HEPG2	USF1	0.457588	12887	0.13	3.90
SYDH	K562	CCNT2	0.456697	21697	0.22	2.94
SYDH	GM12878	MXI1	0.456679	19923	0.20	2.77
HAIB	GM12892	YY1	0.456003	12740	0.13	2.83
HAIB	GM12891	POL24H8	0.455418	17929	0.18	2.50
SYDH	HELAS3	CEBPB	0.450802	39105	0.39	7.92
SYDH	NB4	MAX	0.449059	28193	0.28	4.72
SYDH	HEPG2	TBP	0.448004	13778	0.14	2.88
HAIB	HCT116	YY1SC281	0.447206	9601	0.10	2.36
UTA	MCF7	CMYC	0.446932	17429	0.17	2.52

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
SYDH	K562	CMYC	0.446684	26346	0.26	3.95
HAIB	SKNSHRA	YY1SC281	0.445929	13128	0.13	2.71
HAIB	H1HESC	YY1SC281	0.445242	15591	0.16	2.65
SYDH	HELAS3	JUND	0.444612	22640	0.23	4.23
SYDH	HEPG2	MAZAB85725	0.444409	12934	0.13	1.88
UTA	MCF7	CMYC	0.443654	24235	0.24	3.51
HAIB	A549	USF1	0.441291	7881	0.08	2.59
SYDH	HEPG2	CJUN	0.440671	8890	0.09	1.91
HAIB	SKNSHRA	USF1SC8983	0.439829	12682	0.13	3.64
SYDH	GM12878	MAX	0.439437	14531	0.15	2.21
HAIB	K562	POL24H8	0.438629	19971	0.20	3.52
HAIB	PFSK1	NRSF	0.435981	9928	0.10	4.63
SYDH	H1HESC	SIN3ANB6001263	0.433869	26283	0.26	2.93
UTA	HEPG2	POL2	0.432243	21612	0.22	2.23
HAIB	A549	FOSL2	0.430795	23494	0.24	3.95
HAIB	SKNSH	POL24H8	0.427949	22879	0.23	3.35
SYDH	HUVEC	POL2	0.427119	11883	0.12	1.94
HAIB	K562	YY1	0.426097	19380	0.19	3.54
UCHICAGO	K562	EFOS	0.425453	6855	0.07	1.91
SYDH	H1HESC	CHD2	0.424343	6252	0.06	1.25
SYDH	MCF7	HAE2F1	0.423359	27514	0.28	2.20
HAIB	K562	SP1	0.422803	6215	0.06	1.58
SYDH	K562	JUND	0.420900	30409	0.30	5.93
SYDH	HELAS3	ZNF143	0.420784	5406	0.05	2.13
HAIB	A549	YY1C	0.420411	11293	0.11	2.20
SYDH	GM12878	POL2S2	0.420026	12996	0.13	1.84
HAIB	GM12878	POL2	0.419133	48007	0.48	3.33
HAIB	PFSK1	TAF1	0.415078	6236	0.06	1.35
HAIB	K562	PU1	0.411073	15386	0.15	4.70
SYDH	GM12878	CHD2AB68301	0.410210	16016	0.16	2.63
SYDH	NB4	CMYC	0.406744	23774	0.24	3.73
HAIB	H1HESC	TAF7SC101167	0.406696	10442	0.10	1.54
SYDH	H1HESC	CEBPB	0.405410	11800	0.12	3.73
SYDH	MCF10AES	STAT3	0.404351	33486	0.33	5.08
HAIB	GM12878	POL24H8	0.402366	31663	0.32	2.85

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
HAIB	SKNSH	NRSF	0.401931	7233	0.07	3.71
HAIB	K562	ZBTB7ASC34508	0.399912	19683	0.20	2.16
HAIB	K562	EGR1	0.399163	24881	0.25	3.28
SYDH	MCF10AES	STAT3	0.398512	29538	0.30	4.81
SYDH	K562	CHD2AB68301	0.398431	7834	0.08	2.01
HAIB	SKNMC	POL24H8	0.393543	21485	0.21	2.96
HAIB	H1HESC	POL24H8	0.391510	19419	0.19	1.99
HAIB	K562	CTCFLSC98982	0.391258	5891	0.06	2.85
SYDH	MCF10AES	STAT3	0.388008	31591	0.32	4.98
HAIB	A549	USF1	0.387810	6778	0.07	1.84
HAIB	HEPG2	FOXA1SC6553	0.386906	33656	0.34	5.34
SYDH	MCF10AES	STAT3	0.385338	25848	0.26	4.56
HAIB	SKNSH	NRSF	0.385146	14169	0.14	3.45
SYDH	GM12891	NFKB	0.383466	29206	0.29	4.56
HAIB	H1HESC	SP1	0.380258	12393	0.12	2.05
SYDH	MCF10AES	CMYC	0.379656	27000	0.27	4.33
SYDH	HEPG2	CEBPB	0.379397	11572	0.12	4.10
HAIB	K562	NRSF	0.379106	9598	0.10	4.30
SYDH	GM12878	USF2	0.377835	6661	0.07	2.16
SYDH	HELAS3	TBP	0.376722	17555	0.18	3.06
UTA	K562	CMYC	0.372061	5833	0.06	1.68
HAIB	K562	ATF3	0.371010	10360	0.10	2.78
SYDH	HELAS3	MXI1AF4185	0.368398	12174	0.12	1.83
HAIB	HEPG2	FOSL2	0.367104	16407	0.16	3.44
SYDH	K562	CMYC	0.366773	21209	0.21	3.20
SYDH	HELAS3	MAFK	0.366364	9993	0.10	1.82
SYDH	HELAS3	P300SC584SC584	0.364830	18694	0.19	2.54
HAIB	HEPG2	SP1	0.364172	21711	0.22	3.58
HAIB	K562	PMLSC71910	0.362038	18655	0.19	2.75
HAIB	K562	FOSL1SC183	0.359258	6436	0.06	2.20
HAIB	GM12878	BCL11A	0.358333	12360	0.12	2.80
SYDH	GM12878	SIN3ANB6001263	0.356799	13694	0.14	1.61
SYDH	K562	CJUN	0.354626	5656	0.06	1.98
SYDH	GM12878	TBP	0.353883	15238	0.15	2.78
HAIB	HEPG2	FOXA1SC101058	0.353734	29596	0.30	4.83

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
HAIB	HEPG2	CEBPBSC150	0.348724	9795	0.10	3.67
HAIB	A549	NRSF	0.348252	12999	0.13	3.65
HAIB	GM12878	BATF	0.347600	18755	0.19	3.78
HAIB	A549	USF1	0.347257	8140	0.08	2.22
BROAD	H1HESC	RBBP5A300109A	0.343881	25833	0.26	1.35
HAIB	GM12892	PAX5C20	0.343844	8182	0.08	1.34
BROAD	K562	POL2B	0.341811	15495	0.15	1.86
HAIB	GM12878	NFICSC81335	0.341187	33737	0.34	3.76
SYDH	HELAS3	RFX5200401194	0.341053	15994	0.16	2.36
HAIB	GM12878	IRF4SC6059	0.340861	14517	0.15	2.83
HAIB	GM12878	POU2F2	0.336826	18566	0.19	2.97
HAIB	HEPG2	FOXA2SC6554	0.336085	27428	0.27	4.48
HAIB	SKNSH	SIN3AK20	0.336066	13855	0.14	1.95
HAIB	GM12878	ATF2SC81188	0.335843	26054	0.26	3.55
SYDH	HELAS3	USF2	0.329562	8429	0.08	1.85
SYDH	HELAS3	E2F1	0.328842	5081	0.05	0.74
SYDH	MCF10AES	CMYC	0.327448	19677	0.20	2.88
HAIB	HEPG2	HNF4ASC8987	0.325563	13192	0.13	3.15
SYDH	K562	UBTFSAB1404509	0.325086	14930	0.15	1.59
UCHICAGO	K562	EJUND	0.323401	26489	0.26	3.49
UTA	GM12878	CMYC	0.322020	5627	0.06	0.63
BROAD	K562	SAP3039731	0.320382	11693	0.12	1.16
SYDH	K562	CMYC	0.318111	11312	0.11	2.06
HAIB	H1HESC	EGR1	0.317297	7071	0.07	0.68
HAIB	K562	CEBPBSC150	0.311232	18052	0.18	3.71
HAIB	H1HESC	SIN3AK20	0.310984	7354	0.07	1.48
SYDH	GM15510	NFKB	0.309530	13887	0.14	2.14
BROAD	K562	HDAC1SC6298	0.308889	15009	0.15	1.08
SYDH	GM19099	NFKB	0.308646	6705	0.07	1.71
HAIB	GM12878	FOXM1SC502	0.307947	26561	0.27	2.91
HAIB	PANC1	POL24H8	0.306956	11954	0.12	1.43
HAIB	HEPG2	HNF4GSC6558	0.305644	14815	0.15	2.92
HAIB	HEPG2	JUND	0.305335	14409	0.14	2.61
SYDH	K562	TAL1SC12984	0.304212	18090	0.18	4.50
HAIB	HEPG2	CEBPDSC636	0.303716	8698	0.09	1.82

Continued on next page

Table D.1 (Continued)

Lab	Cell line	Experiment	BASSET AUPRC	$\|v\|_0$	$\|v\|_0/M$ (%)	$\|v\|_2$
SYDH	K562	CORESTSC30189	0.303011	28293	0.28	3.98
SYDH	K562	BHLHE40NB100	0.301552	19955	0.20	2.77
HAIB	GM12878	EBF1SC137065	0.301285	24230	0.24	3.43

This table can be downloaded as an Excel file from
https://www.biorxiv.org/content/early/2017/10/17/204685.figures-only.

Table D.2: Numerical results for Figure 5.1. We list all P-values used for the simulations of a) no enrichment, b) unsigned enrichment, and c) directional effects of minor alleles, with and without the 5-MAF-bin signed background model.

This table can be downloaded as an Excel file from
https://www.biorxiv.org/content/early/2017/10/17/204685.figures-only.

Table D.3: Numerical results for Figure 5.2. We list a) estimated power, with standard errors, for both methods analyzed in Figure 5.2a, b) mean estimate of r_f, with standard error, for all values of r_f simulated, together with quantiles of the sampling distribution of our estimator.

This table can be downloaded as an Excel file from
https://www.biorxiv.org/content/early/2017/10/17/204685.figures-only.

Table D.4: List of molecular traits analyzed. We list the set of molecular traits analyzed, with number of typed SNPs for each trait.

This table can be downloaded as an Excel file from

https://www.biorxiv.org/content/early/2017/10/17/204685.figures-only.

Table D.5: Details of analysis of molecular traits. We list a) the set of 92 significant associations at per-trait FDR$< 5\%$ for the BLUEPRINT gene expression analysis, with laboratory, cell line, and TF listed for each significant annotation, along with estimated r_f, P-value, and whether the TF is known to be activating; b) the set of 22 significant associations at per-trait FDR $< 5\%$ for the NTR gene expression analysis; c) the side-by-side comparison of z-scores from the BLUEPRINT neutrophil expression analysis and the NTR analysis; d) the set of 13 significant associations at per-trait FDR $< 5\%$ for the BLUEPRINT H3K4me1 analysis; e) the set of 3 significant associations at per-trait FDR $< 5\%$ for the BLUEPRINT H3K27ac analysis; and f) the set of 40 significant associations at per-trait FDR $< 10\%$ for the BLUEPRINT methylation analysis.

This table can be downloaded as an Excel file from

https://www.biorxiv.org/content/early/2017/10/17/204685.figures-only.

Table D.6: List of diseases and complex traits analyzed. We list the set of diseases and complex traits analyzed, with sample size, number of typed SNPs, and estimated SNP-heritability for each trait.

This table can be downloaded as an Excel file from

https://www.biorxiv.org/content/early/2017/10/17/204685.figures-only.

Table D.7: Details of analysis of 46 diseases and complex traits. We list a) the set of 77 significant associations at per-trait FDR$< 5\%$ for the TF annotations, with laboratory, cell line, and transcription factor listed for each significant annotation, along with estimated r_f and P-value; b) the set of 4 significant associations at per-trait FDR $< 5\%$ for the alternate set of 382 annotations defined using the same set of SNPs with non-zero effects but with the directionality of effect determined by minor allele coding rather than predicted TF binding, for SNPs in the bottom quintile of the MAF spectrum; c) quantification of unsigned heritability explained by signed enrichments reported in (a). Specifically: because r_f^2 for an annotation can never exceed the total proportion of heritability explained by the SNPs with nonzero values of the annotation, we computed for each association the ratio of estimated r_f^2 to the proportion of SNPs with nonzero values of the annotation. We found that in some cases the signed signal was not only non-trivially different from zero but also substantial enough to imply an unsigned enrichment.

This table can be downloaded as an Excel file from

https://www.biorxiv.org/content/early/2017/10/17/204685.figures-only.

Table D.8: Numerical results for Figure 5.5. For each result in the figure, we list i) the numerical values used to make the plot of $\hat{\alpha}$ against Rv, and ii) the association summary statistics used to make the Manhattan plot. In the case of IRF1-Crohn's, we also list Z-scores for association with *IRF1* in whole blood from the NTR data set.

SNP	P(in causal set)	Causal post. prob.	Z
rs10189857	0.25	1	8.0933
rs356991	0.128176	0.512705	6.03
rs168565	0.0366951	0.14678	5.9928
rs6545816	0.154972	0.619888	5.4231
rs6545817	0.0950247	0.380099	5.3862
rs243071	0.25	1	-5.2992

Table D.9: Fine mapping of EDU signal at *BCL11A* locus. We list the six SNPs in the 95% credible set when running the CAVIAR method with the parameter $c = 4$. rs10189857 is an intronic SNP in the *BCL11A* gene. (Results with $c = 2$ and $c = 3$ were similar.)

D.6 Supplementary Figures

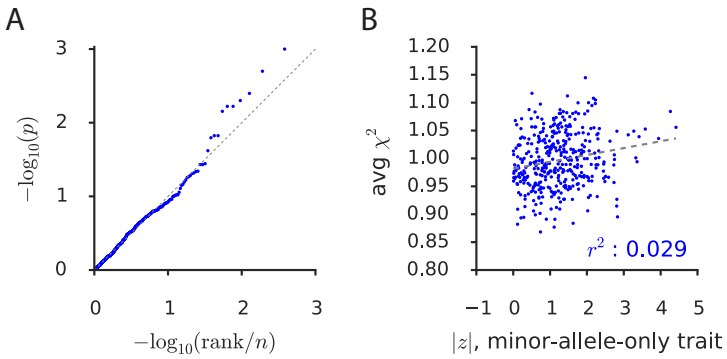


Figure D.1: Per-annotation analyses of null calibration. (a) For each annotation, we used the Simes test[143] to assess the p-value threshold at which the Benjamini-Hochberg procedure would lead to any rejections among 1000 simulated phenotypes with no unsigned enrichment or functional correlation, and we visualized the resulting set of 382 p-values using a q-q plot. These p-values appear uniformly distributed, as would be expected in the scenario of proper calibration. (b) The average χ^2 statistic across the 1000 null simulations for each annotation, plotted against the magnitude of that annotation's z-score for correlation with a 100%-heritable trait whose causal SNPs are exactly the bottom fifth of the MAF spectrum, with minor alleles always being trait-increasing. (Statistical significance of the trend is difficult to assess because many annotations are correlated, inducing a complex dependence structure among the 382 points on the plot.)

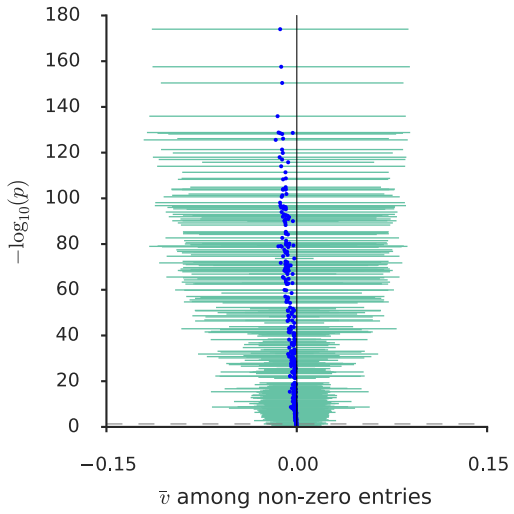


Figure D.2: Relationship of annotations to minor alleles. For each annotation, we computed the mean and standard deviation of the predicted effect of the minor allele among all SNPs with non-zero values of the annotation. We then performed a chi-squared test for the mean being non-zero and plotted $-\log_{10}(p)$ against the mean for each annotation. The green intervals show the standard deviation, in order to give a sense for the scale on which to interpret the mean-shift. The dotted gray line indicates the threshold for FDR significance. 373 of the 382 annotations exceeded this threshold.

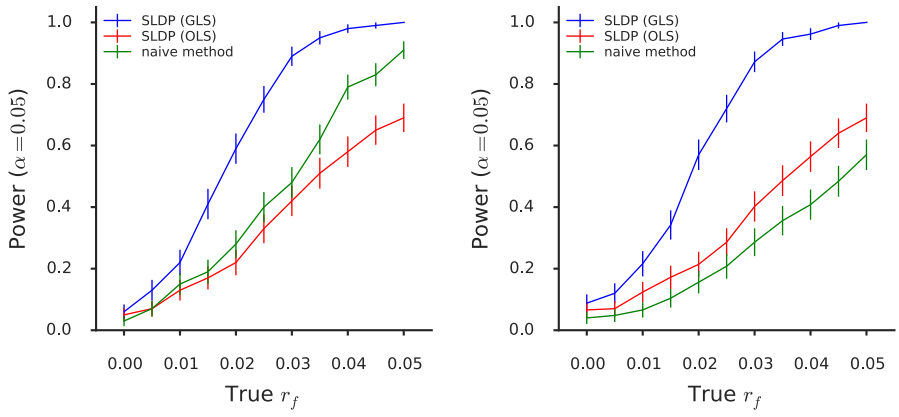


Figure D.3: Power comparison of signed LD profile regression to additional methods. Power curves comparing signed LD profile regression using generalized least-squares (GLS; i.e., weighting) to both ordinary (i.e., unweighted) regression of the GWAS summary statistics on the signed LD profile as well as to a naive method that simply regresses the GWAS summary statistics on the raw annotation. (Left) power comparison with 19.5% of causal SNPs typed, (Right) power comparison with only 9.75% of causal SNPs typed. The real phenotypes analyzed all have at most 11.9% of causal SNPs typed. SLDP regression with default weights is the most powerful method in both regimes. Additionally, the power of the naive method suffers when fewer SNPs are typed, while the power of SLDP regression is far less sensitive to this change.

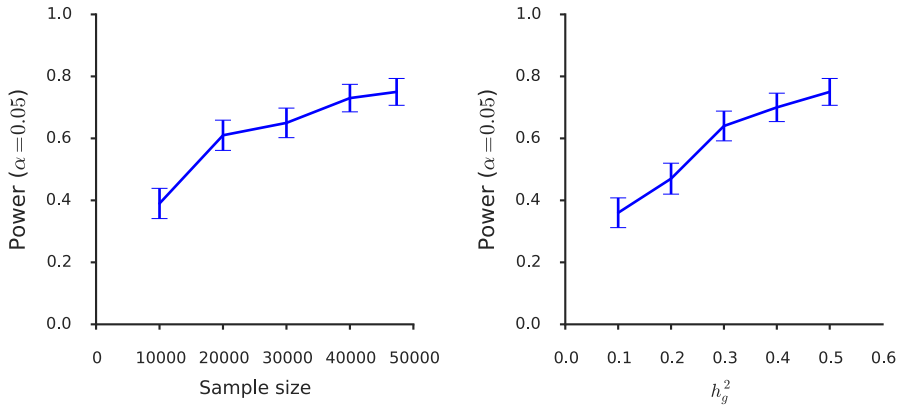


Figure D.4: Effect of sample size and heritability on power. Power of signed LD profile regression as a function of (left) sample size, and (right) overall trait heritability, when proportion of heritability explained by the signed effect is held constant. Error bars indicate standard errors of power estimates.

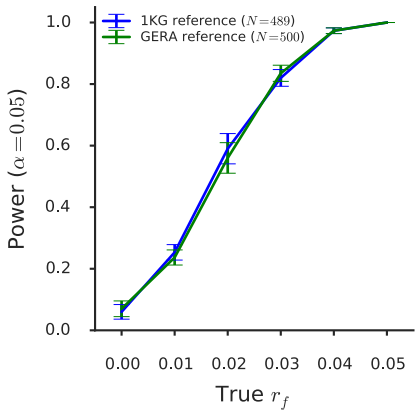


Figure D.5: Effect of reference panel on power. Power of signed LD profile regression as a function of effect size as measured by r_f, with either a 1000G reference panel or a randomly chosen in-sample reference panel of comparable size. Error bars indicate standard errors of power estimates.

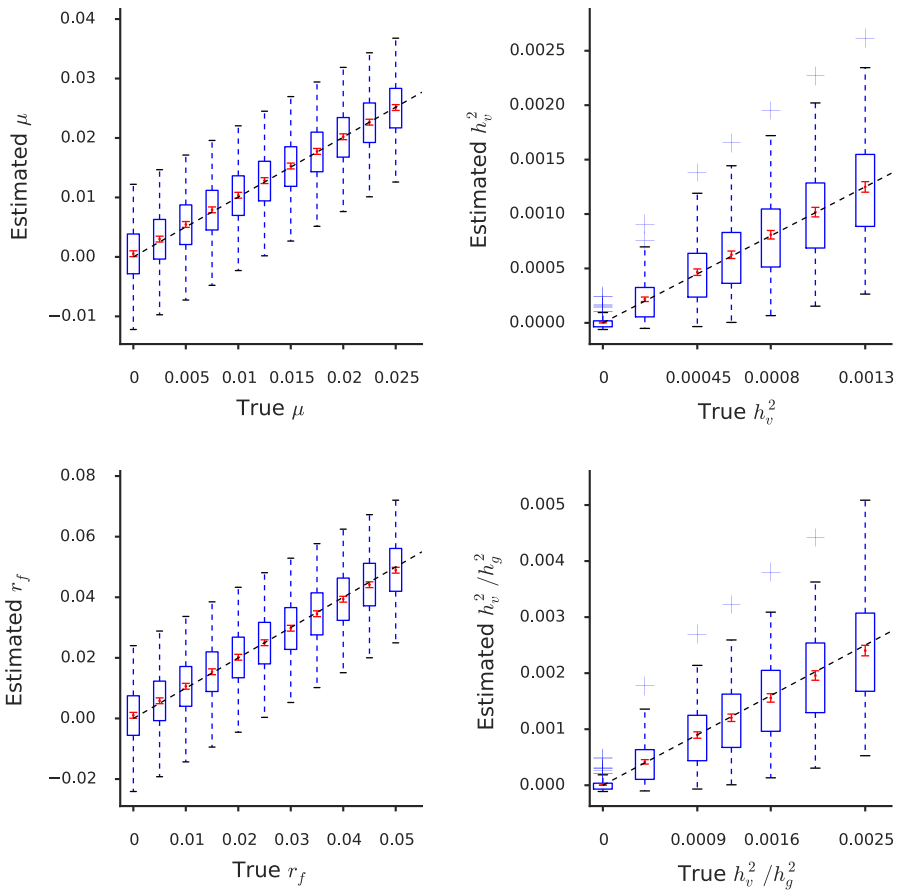


Figure D.6: Bias in estimation of additional estimands. Assessments of the bias of signed LD profile regression with an out-of-sample reference panel in estimating μ, h_v^2, r_f, and h_v^2/h_g^2. For definitions of these quantities, see Supplementary Note.

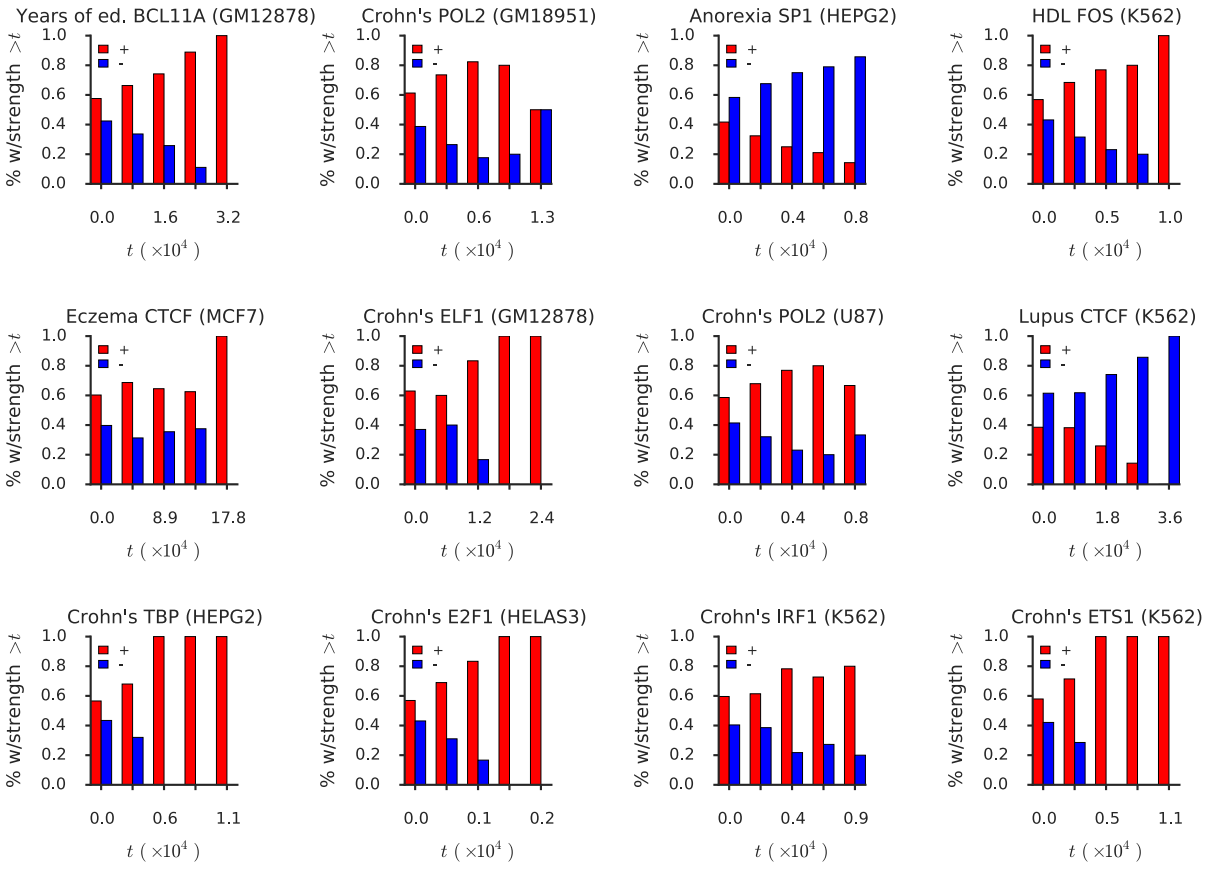


Figure D.7: Distribution of covariance between GWAS summary statistics and signed LD profile. For each of our twelve independent results, we plot, for a variety of thresholds t, the fraction of the approximately 300 independent genomic blocks with $|\text{cov}(\hat{\alpha}, Rv)| > t$ in which the covariance is positive versus negative. There is an excess of blocks in which sign of the covariance matches the genome-wide direction of effect. (We note that, as this figure illustrates, our results do not imply that the sign of the covariance matches the genome-wide direction of effect in *all* blocks.)

D.7 The distribution of GWAS summary statistics

We define the vector $\hat{\alpha}$ of marginal correlations between SNPs and trait and derive its first two moments under a variety of relevant models, building up to the signed LD profile regression model.

D.7.1 Definitions

Let M be the number of SNPs in the genome. Assume we have sampled N genotype vectors $x_1, \ldots, x_N$ i.i.d. from some population distribution, and that the phenotypes $y_1, \ldots, y_N$ of those individuals satisfy

$$y_n = x_n^T \beta + \varepsilon_n \tag{D.18}$$

where $\beta \in \mathbb{R}^M$ is the vector of true causal SNP effects on trait, and $\varepsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$ are independent of the x_n. We assume throughout this section that genotypes are standardized in the population, i.e., $E(x_{nm}) = 0$ and $E(x_{nm}^2) = 1$ for all n, m. We assume the same of the phenotype: $E(y_n) = 0$ and $E(y_n^2) = 1$ for all n. These assumptions are for expositional convenience.

Let $X \in \mathbb{R}^{N \times M}$ be the matrix whose n-th row is x_n^T, and let $Y \in \mathbb{R}^N$ be the vector

whose n-th entry is y_n. The vector

$$\hat{\alpha} = \frac{X^T Y}{N}, \tag{D.19}$$

which has as its m-th entry the in-sample marginal correlation between SNP m and the trait, is the vector of *GWAS summary statistics*.

Having defined $\hat{\alpha}$, we now proceed to derive its first two moments, initially for fixed X and fixed β, and then for fixed β only. After doing so, we will impose the distributional assumption on β used in signed LD profile regression and, by marginalizing out β according to this distribution, we will obtain the result required for this paper.

D.7.2 Derivation for fixed X and fixed β

When both X and β are fixed, the following proposition[142] gives the moments of $\hat{\alpha}$.

Proposition D.7.1. *Under the model defined above, $\hat{\alpha}$ satisfies*

$$\hat{\alpha}|X, \beta \sim \mathcal{N}\left(\hat{R}\beta, \sigma_e^2 \frac{\hat{R}}{N}\right) \tag{D.20}$$

where $\hat{R} = X^T X/N$ is the sample covariance matrix of the genotypes.

Proof. Let $\varepsilon \in \mathbb{R}^N$ be the vector whose n-th entry is ε_n. When X and β are both fixed,

it is easy to see that

$$\hat{\alpha} = \frac{1}{N} X^T Y \tag{D.21}$$

$$= \frac{1}{N} X^T (X^T \beta + \varepsilon) \tag{D.22}$$

$$= \hat{R}\beta + \frac{1}{N} X^T \varepsilon. \tag{D.23}$$

The result follows from normality of ε, together with $E(\varepsilon) = 0$, and $\text{var}(X^T\varepsilon/N) = \sigma_e^2 X^T X/N^2 = \sigma_e^2 \hat{R}/N$.

□

D.7.3 Derivation for random X and fixed β

When working with summary statistics, it is desirable to explicitly model the relationship between the unobserved individuals and the LD reference panel by assuming the individuals were drawn from a population distribution whose LD properties we are given by the reference panel. The following result states the moments of $\hat{\alpha}$ when we do so. We prove the result assuming Gaussian genotypes, but it can be shown to be robust to this assumption provided there is a lower bound on minor allele frequency relative to sample size.

Proposition D.7.2. *Under the model defined above and assuming Gaussian genotypes, $\hat{\alpha}$ satisfies*

$$\hat{\alpha}|\beta \sim \left[R\beta, \frac{1}{N}\left(R + R\beta\beta^T R\right) \right] \tag{D.24}$$

where $R = cov(x_n) \in \mathbb{R}^{M \times M}$ is the population covariance matrix of the genotypes, and the notation $[,]$ is used to specify the mean and covariance of the distribution without specifying any higher moments.

Proof. Application of the law of total expectation to the result from Proposition D.7.1 readily gives

$$E(\hat{\alpha}|\beta) = E\left(E(\hat{\alpha}|X,\beta)|\beta\right) \tag{D.25}$$

$$= E(\hat{R}\beta|\beta) \tag{D.26}$$

$$= R\beta. \tag{D.27}$$

Application of the law of total covariance yields

$$\mathrm{cov}(\hat{\alpha}|\beta) = E\left(\mathrm{cov}(\hat{\alpha}|X,\beta)|\beta\right) + \mathrm{cov}\left(E(\hat{\alpha}|X,\beta)|\beta\right) \tag{D.28}$$

$$\sigma_e^2 \frac{\hat{R}}{N} + \mathrm{cov}(\hat{R}\beta|\beta). \tag{D.29}$$

It is left then only to analyze $\mathrm{cov}(\hat{R}\beta|\beta) = E(\hat{R}\beta\beta^T\hat{R}) - R\beta\beta^T R$. To do so, we note

that

$$\begin{aligned}
\mathrm{cov}(\hat{R}\beta|\beta)_{mm'} &= \left(E(\hat{R}\beta\beta^T\hat{R}) - R\beta\beta^T R\right)_{mm'} && \text{(D.30)}\\
&= \sum_{i,j}\left(E\left(\hat{R}_{mi}\beta_i\beta_j\hat{R}_{jm'}\right) - R_{mi}\beta_i\beta_j R_{jm'}\right) && \text{(D.31)}\\
&= \sum_{i,j}\beta_i\beta_j\left(E\left(\hat{R}_{mi}\hat{R}_{m'j}\right) - R_{mi}R_{m'j}\right) && \text{(D.32)}\\
&= \frac{1}{N}\sum_{i,j}\beta_i\beta_j\left(R_{mm'}R_{ij} + R_{mj}R_{m'i}\right) && \text{(D.33)}\\
&= \frac{1}{N}R_{mm'}\sum_{i,j}\beta_i\beta_j R_{ij} + \frac{1}{N}\sum_{i,j}\beta_i\beta_j R_{mj}R_{m'i} && \text{(D.34)}\\
&= \frac{1}{N}R_{mm'}\beta^T R\beta + \frac{1}{N}\sum_{i,j}\beta_i\beta_j R_{mj}R_{m'i} && \text{(D.35)}
\end{aligned}$$

where Equation D.33 follows from the fact that for Gaussian genotypes, Isselis' theorem implies that

$$E(\hat{R}_{mi}\hat{R}_{m'j}) = R_{mi}R_{m'j} + \frac{1}{N}(R_{mm'}R_{ij} + R_{mj}R_{m'i}). \qquad \text{(D.36)}$$

The result of this argument can be summarized across all pairs of SNPs m, m' by

$$\mathrm{cov}(\hat{R}\beta|\beta) = \frac{1}{N}\left((\beta^T R\beta)R + R\beta\beta^T R\right), \qquad \text{(D.37)}$$

whereupon noticing that $\beta^T R\beta + \sigma_e^2 = \mathrm{var}(y_n) = 1$ completes the proof. □

Corollary D.7.3. *Under the model defined above, $\hat{\alpha}$ approximately satisfies*

$$\hat{\alpha}|\beta \sim \left[R\beta, \frac{R}{N}\right] \tag{D.38}$$

where $R = cov(x_n) \in \mathbb{R}^{M\times M}$ is the population covariance matrix of the genotypes.

Proof. For typical β it can be shown that $\beta\beta^T$ will be close to a scalar multiple of the identity. In this case, taking traces shows that the term $R\beta\beta^T R/N$ In Proposition D.7.2 will be negligible compared to the R/N term. □

D.7.4 Derivation for random X and random β

We now assume the full signed LD profile regression model, i.e., we fix some signed annotation $v \in \mathbb{R}^M$, and let $\beta \sim [\mu v, \sigma^2]$. Under this model, we have the following result.

Theorem D.7.4. *If $\beta \sim [\mu v, \sigma^2]$ for some $v \in \mathbb{R}^M$ and $\sigma^2 > 0$, then $\hat{\alpha}$ approximately satisfies*

$$\hat{\alpha}|v \sim \left[\mu Rv, \sigma^2 R^2 + \frac{R}{N}\right] \tag{D.39}$$

where $R = cov(x_n) \in \mathbb{R}^{M\times M}$ is the population covariance matrix of the genotypes.

Proof. The law of total expectation applied to the result of Corollary D.7.3 yields $E(\hat{\alpha}|v) = \mu Rv$

as desired. The law of total covariance yields

$$\begin{aligned}
\operatorname{cov}(\hat{\alpha}|v) &\approx E\left(\operatorname{cov}(\hat{\alpha}|\beta)|v\right) + \operatorname{cov}\left(E(\hat{\alpha}|\beta)|v\right) && \text{(D.40)}\\
&= \frac{R}{N} + \operatorname{cov}(R\beta|v) && \text{(D.41)}\\
&= \frac{R}{N} + R\operatorname{cov}(\beta|v)R && \text{(D.42)}\\
&= \frac{R}{N} + \sigma^2 R^2 && \text{(D.43)}
\end{aligned}$$

as desired. □

References

[1] 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74.

[2] Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8), 831–838.

[3] Amente, S., Lania, L., & Majello, B. (2011). Epigenetic reprogramming of Myc target genes. *American Journal of Cancer Research*, 1(3), 413–418.

[4] Arbiza, L., Gronau, I., Aksoy, B. A., Hubisz, M. J., Gulko, B., Keinan, A., & Siepel, A. (2013). Genome-wide inference of natural selection on human transcription factor binding sites. *Nature Genetics*, 45(7), 723–729.

[5] Bakiri, L., Hamacher, R., Graña, O., Guío-Carrión, A., Campos-Olivas, R., Martinez, L., Dienes, H. P., Thomsen, M. K., Hasenfuss, S. C., & Wagner, E. F. (2017). Liver carcinogenesis by FOS-dependent inflammation and cholesterol dysregulation. *Journal of Experimental Medicine*, (pp. jem.20160935).

[6] Banda, Y., Kvale, M. N., Hoffmann, T. J., Hesselson, S. E., Ranatunga, D., Tang, H., Sabatti, C., Croen, L. A., Dispensa, B. P., Henderson, M., Iribarren, C., Jorgenson, E., Kushi, L. H., Ludwig, D., Olberg, D., Quesenberry, C. P., Rowell, S., Sadler, M., Sakoda, L. C., Sciortino, S., Shen, L., Smethurst, D., Somkin, C. P., Eeden, S. K. V. D., Walter, L., Whitmer, R. A., Kwok, P.-Y., Schaefer, C., & Risch, N. (2015). Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics*, 200(4), 1285–1295.

[7] Bartz, F., Kern, L., Erz, D., Zhu, M., Gilbert, D., Meinhof, T., Wirkner, U., Erfle, H., Muckenthaler, M., Pepperkok, R., & Runz, H. (2009). Identification

of Cholesterol-Regulating Genes by Targeted RNAi Screening. *Cell Metabolism*, 10(1), 63–75.

[8] Basak, A., Hancarova, M., Ulirsch, J. C., Balci, T. B., Trkova, M., Pelisek, M., Vlckova, M., Muzikova, K., Cermak, J., Trka, J., Dyment, D. A., Orkin, S. H., Daly, M. J., Sedlacek, Z., & Sankaran, V. G. (2015). *BCL11A* deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. *The Journal of Clinical Investigation*, 125(6), 2363–2368.

[9] Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.

[10] Berisa, T. & Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2), 283–285.

[11] Bories, J. C., Willerford, D. M., Grévin, D., Davidson, L., Camus, A., Martin, P., Stéhelin, D., & Alt, F. W. (1995). Increased T-cell apoptosis and terminal B-cell differentiation induced by inactivation of the Ets-1 proto-oncogene. *Nature*, 377(6550), 635–638.

[12] Bornstein, C., Winter, D., Barnett-Itzhaki, Z., David, E., Kadri, S., Garber, M., & Amit, I. (2014). A negative feedback loop of transcription factors specifies alternative dendritic cell chromatin states. *Molecular cell*, 56(6), 749–762.

[13] Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7), 1177–1186.

[14] Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391), 580–598.

[15] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., & Neale, B. M. (2015a). An Atlas of Genetic Correlations across Human Diseases and Traits. *Nature genetics*, 47(11), 1236–1241.

[16] Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly,

M. J., Price, A. L., & Neale, B. M. (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3), 291–295.

[17] Casella, G. & Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

[18] Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4, 7.

[19] Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., Datta, A., Richardson, D., Burden, F., Mead, D., Mann, A. L., Fernandez, J. M., Rowlston, S., Wilder, S. P., Farrow, S., Shao, X., Lambourne, J. J., Redensek, A., Albers, C. A., Amstislavskiy, V., Ashford, S., Berentsen, K., Bomba, L., Bourque, G., Bujold, D., Busche, S., Caron, M., Chen, S.-H., Cheung, W., Delaneau, O., Dermitzakis, E. T., Elding, H., Colgiu, I., Bagger, F. O., Flicek, P., Habibi, E., Iotchkova, V., Janssen-Megens, E., Kim, B., Lehrach, H., Lowy, E., Mandoli, A., Matarese, F., Maurano, M. T., Morris, J. A., Pancaldi, V., Pourfarzad, F., Rehnstrom, K., Rendon, A., Risch, T., Sharifi, N., Simon, M.-M., Sultan, M., Valencia, A., Walter, K., Wang, S.-Y., Frontini, M., Antonarakis, S. E., Clarke, L., Yaspo, M.-L., Beck, S., Guigo, R., Rico, D., Martens, J. H. A., Ouwehand, W. H., Kuijpers, T. W., Paul, D. S., Stunnenberg, H. G., Stegle, O., Downes, K., Pastinen, T., & Soranzo, N. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, 167(5), 1398–1414.e24.

[20] Chen-Plotkin, A. S., Sadri-Vakili, G., Yohrling, G. J., Braveman, M. W., Benn, C. L., Glajch, K. E., DiRocco, D. P., Farrell, L. A., Krainc, D., Gines, S., MacDonald, M. E., & Cha, J.-H. J. (2006). Decreased association of the transcription factor Sp1 with genes downregulated in Huntington's disease. *Neurobiology of Disease*, 22(2), 233–241.

[21] Cirovic, B., Schönheit, J., Kowenz-Leutz, E., Ivanovska, J., Klement, C., Pronina, N., Bégay, V., & Leutz, A. (2017). C/EBP-Induced Transdifferentiation Reveals Granulocyte-Macrophage Precursor-like Plasticity of B Cells. *Stem Cell Reports*, 8(2), 346–359.

[22] Cleveland, W. S. & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403), 596–610.

[23] Cover, T. & Thomas, J. (2006). *Elements of Information Theory*. New York: John Wiley & Sons, Inc.

[24] Cowley, G. S., Weir, B. A., Vazquez, F., Tamayo, P., Scott, J. A., Rusin, S., East-Seletsky, A., Ali, L. D., Gerath, W. F., Pantel, S. E., Lizotte, P. H., Jiang, G., Hsiao, J., Tsherniak, A., Dwinell, E., Aoyama, S., Okamoto, M., Harrington, W., Gelfand, E., Green, T. M., Tomko, M. J., Gopal, S., Wong, T. C., Li, H., Howell, S., Stransky, N., Liefeld, T., Jang, D., Bistline, J., Meyers, B. H., Armstrong, S. A., Anderson, K. C., Stegmaier, K., Reich, M., Pellman, D., Boehm, J. S., Mesirov, J. P., Golub, T. R., Root, D. E., & Hahn, W. C. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific Data*, 1, sdata201435.

[25] Csiszár, I. (2008). Axiomatic characterizations of information measures. *Entropy*, 10(3), 261–273.

[26] Csiszár, I. & Shields, P. C. (2004). Information theory and statistics: A tutorial. *Communications and Information Theory*, 1(4), 417–528.

[27] Davari, K., Lichti, J., Gallus, C., Greulich, F., Uhlenhaut, N. H., Heinig, M., Friedel, C. C., & Glasmacher, E. (2017). Rapid Genome-wide Recruitment of RNA Polymerase II Drives Transcription, Splicing, and Translation Events during T Cell Responses. *Cell Reports*, 19(3), 643–654.

[28] Davey Smith, G. & Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1), R89–R98.

[29] Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542), 223–228.

[30] Deplancke, B., Alpern, D., & Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. *Cell*, 166(3), 538–554.

[31] Dias, C., Estruch, S. B., Graham, S. A., McRae, J., Sawiak, S. J., Hurst, J. A., Joss, S. K., Holder, S. E., Morton, J. E. V., Turner, C., Thevenon, J., Mellul, K., Sánchez-Andrade, G., Ibarra-Soria, X., Deriziotis, P., Santos, R. F., Lee, S.-C., Faivre, L., Kleefstra, T., Liu, P., Hurles, M. E., Fisher, S. E., & Logan, D. W. (2016). BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription. *The American Journal of Human Genetics*, 99(2), 253–274.

[32] Ding, A. A., Dy, J. G., Li, Y., & Chang, Y. (2017). A robust-equitable measure for feature ranking and selection. *Journal of Machine Learning Research*, 18(71), 1–46.

[33] Ding, A. A. & Li, Y. (2013). Copula correlation: An equitable dependence measure and extension of pearson's correlation. *arXiv preprint arXiv:1312.7214*.

[34] Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7), 1853–1866.e17.

[35] Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. *Nature*, 452(7186), 423–428.

[36] Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S., & Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature Biotechnology*, 34(11), 1180–1190.

[37] Eyquem, S., Chemin, K., Fasseu, M., & Bories, J.-C. (2004). The Ets-1 transcription factor is required for complete pre-T cell receptor function and allelic exclusion at the T cell receptor β locus. *Proceedings of the National Academy of Sciences of the United States of America*, 101(44), 15712–15717.

[38] Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J., Shishkin, A. A., & others (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539), 337–343.

[39] Faust, K. & Raes, J. (2012). Microbial interactions: from networks to models. *Nature reviews. Microbiology*, 10(8), 538.

[40] Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada, Y., Raychaudhuri, S., Daly, M. J., Patterson, N., Neale, B. M., & Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11), 1228–1235.

[41] Frank, D. A. (2009). Targeting transcription factors for cancer therapy. *IDrugs: the investigational drugs journal*, 12(1), 29–33.

[42] Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., & others (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics*, 42(12), 1118–1125.

[43] Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., & Engreitz, J. M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, (pp. aag2445).

[44] Funnell, A. P. W., Prontera, P., Ottaviani, V., Piccione, M., Giambona, A., Maggio, A., Ciaffoni, F., Stehling-Sun, S., Marra, M., Masiello, F., Varricchio, L., Stamatoyannopoulos, J. A., Migliaccio, A. R., & Papayannopoulou, T. (2015). 2p15-p16.1 microdeletions encompassing and proximal to BCL11A are associated with elevated HbF in addition to neurologic impairment. *Blood*, 126(1), 89–93.

[45] Fusté, M., Pinacho, R., Meléndez-Pérez, I., Villalmanzo, N., Villalta-Gil, V., Haro, J. M., & Ramos, B. (2013). Reduced expression of SP1 and SP4 transcription factors in peripheral blood mononuclear cells in first-episode psychosis. *Journal of Psychiatric Research*, 47(11), 1608–1614.

[46] Gallant, S. & Gilkeson, G. (2006). ETS transcription factors and regulation of immunity. *Archivum Immunologiae et Therapiae Experimentalis*, 54(3), 149–163.

[47] Gill, T. et al. (2002). Obesity in the pacific: Too big to ignore. *World Health Organization Regional Office for the Western Pacific, Secretariat of the Pacific Community*.

[48] Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, A. J., Mesirov, J. P., & Haining, W. N. (2016). Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity*, 44(1), 194–206.

[49] Gorfine, M., Heller, R., & Heller, Y. (2012). Comment on "Detecting novel associations in large data sets". *Unpublished*. Available at http://www.math.tau.ac.il/~ruheller/Papers/science6.pdf.

[50] Grenningloh, R., Kang, B. Y., & Ho, I.-C. (2005). Ets-1, a functional cofactor of T-bet, is essential for Th1 inflammatory responses. *Journal of Experimental Medicine*, 201(4), 615–626.

[51] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1), 723–773.

[52] Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory* (pp. 63–77).: Springer.

[53] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2007). A kernel statistical test of independence. In *Advances in neural information processing systems* (pp. 585–592).

[54] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2008). A kernel statistical test of independence.

[55] Groner, Y., Ito, Y., Liu, P., Neil, J. C., Speck, N. A., & van Wijnen, A. (2017). *RUNX Proteins in Development and Cancer*. Springer. Google-Books-ID: 1pNcDgAAQBAJ.

[56] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E.,

Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., Raitakari, O. T., Kuusisto, J., Laakso, M., Price, A. L., Pajukanta, P., & Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3), 245–252.

[57] Hansen, S. K., Baeuerle, P. A., & Blasi, F. (1994). Purification, reconstitution, and I kappa B association of the c-Rel-p65 (RelA) complex, a strong activator of transcription. *Molecular and Cellular Biology*, 14(4), 2593–2603.

[58] Heller, R., Heller, Y., & Gorfine, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2), 503–510.

[59] Heller, R., Heller, Y., Kaufman, S., Brill, B., & Gorfine, M. (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(29), 1–54.

[60] Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, (pp. 546–557).

[61] Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B., & Eskin, E. (2014). Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics*, 198(2), 497–508.

[62] Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., D'Amato, M., Deffontaine, V., Dmitrieva, J., Docampo, E., Elansary, M., Farh, K. K.-H., Franke, A., Gori, A.-S., Goyette, P., Halfvarson, J., Haritunians, T., Knight, J., Lawrance, I. C., Lees, C. W., Louis, E., Mariman, R., Meuwissen, T., Mni, M., Momozawa, Y., Parkes, M., Spain, S. L., Théâtre, E., Trynka, G., Satsangi, J., van Sommeren, S., Vermeire, S., Xavier, R. J., International Inflammatory Bowel Disease Genetics Consortium, Weersma, R. K., Duerr, R. H., Mathew, C. G., Rioux, J. D., McGovern, D. P. B., Cho, J. H., Georges, M., Daly, M. J., & Barrett, J. C. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, 547(7662), 173–178.

[63] Huff, C. D., Witherspoon, D. J., Zhang, Y., Gatenbee, C., Denson, L. A., Kugathasan, S., Hakonarson, H., Whiting, A., Davis, C. T., Wu, W., Xing, J., Watkins,

W. S., Bamshad, M. J., Bradfield, J. P., Bulayeva, K., Simonson, T. S., Jorde, L. B., & Guthery, S. L. (2012). Crohn's disease and genetic hitchhiking at IBD5. *Molecular Biology and Evolution*, 29(1), 101–111.

[64] Huo, X. & Szekely, G. J. (2014). Fast computing for distance covariance. *arXiv preprint arXiv:1410.1503*.

[65] Huo, X. & Székely, G. J. (2016). Fast computing for distance covariance. *Technometrics*, 58(4), 435–447.

[66] Jiang, B., Ye, C., & Liu, J. S. (2015). Nonparametric k-sample tests via dynamic slicing. *Journal of the American Statistical Association*, 110(510), 642–653.

[67] Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., & others (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422), 119–124.

[68] Kálmán, J., Kudchodkar, B. J., Krishnamoorthy, R., Dory, L., Lacko, A. G., & Agarwal, N. (2001). High cholesterol diet down regulates the activity of activator protein-1 but not nuclear factor-kappa B in rabbit brain. *Life Sciences*, 68(13), 1495–1503.

[69] Karczewski, K. J., Dudley, J. T., Kukurba, K. R., Chen, R., Butte, A. J., Montgomery, S. B., & Snyder, M. (2013). Systematic functional regulatory assessment of disease-associated variants. *Proceedings of the National Academy of Sciences*, 110(23), 9607–9612.

[70] Kelley, D. R. & Reshef, Y. A. (2017). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *bioRxiv*, (pp. 161851).

[71] Kelley, D. R., Reshef, Y. A., Belanger, D., McLean, C., Snoek, J., & Bileschi, M. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *bioRxiv*, (pp. 161851).

[72] Kelley, D. R., Snoek, J., & Rinn, J. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, (pp. gr.200535.115).

[73] Kim, T.-G., Kim, M., Lee, J.-J., Kim, S. H., Je, J. H., Lee, Y., Song, M.-J., Choi, Y., Chung, Y. W., Park, C. G., Cho, J. W., Lee, M.-G., Lee, Y.-S., & Kim, H.-P. (2015). CCCTC-binding factor controls the homeostatic maintenance and migration of Langerhans cells. *The Journal of Allergy and Clinical Immunology*, 136(3), 713–724.

[74] Kim, Y.-J., Park, S.-W., Kim, T.-H., Park, J.-S., Cheong, H. S., Shin, H. D., & Park, C.-S. (2013). Genome-wide methylation profiling of the bronchial mucosa of asthmatics: Relationship to atopy. *BMC Medical Genetics*, 14, 39.

[75] Kimura, T., Nakayama, K., Penninger, J., Kitagawa, M., Harada, H., Matsuyama, T., Tanaka, N., Kamijo, R., Vilček, J., Mak, T. W., & Taniguchi, T. (1994). Involvement of the IRF-1 Transcription Factor in Antiviral Responses to Interferons. *Science*, 264(5167), 1921–1924.

[76] Kinney, J. B. & Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*.

[77] Knebel, B., Kotzka, J., Lehr, S., Hartwig, S., Avci, H., Jacob, S., Nitzgen, U., Schiller, M., März, W., Hoffmann, M. M., Seemanova, E., Haas, J., & Muller-Wieland, D. (2013). A mutation in the c-Fos gene associated with congenital generalized lipodystrophy. *Orphanet Journal of Rare Diseases*, 8, 119.

[78] Konstantinopoulos, P. A. & Papavassiliou, A. G. (2011). Seeing the Future of Cancer-Associated Transcription Factor Drug Targets. *JAMA*, 305(22), 2349–2350.

[79] Kraskov, A., Stogbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69.

[80] Laiosa, C. V., Stadtfeld, M., & Graf, T. (2006). Determinants of lymphoid-myeloid lineage diversification. *Annual Review of Immunology*, 24, 705–738.

[81] Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, 47(8), 955–961.

[82] Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen,

T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G., & Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291.

[83] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)*, 27(12), 1739–1740.

[84] Linfoot, E. (1957). An informational measure of correlation. *Information and Control*, 1(1), 85–89.

[85] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., & Price, A. L. (2017). Mixed model association for biobank-scale data sets. *bioRxiv*, (pp. 194944).

[86] Lopez-Paz, D., Hennig, P., & Schölkopf, B. (2013). The randomized dependence coefficient. In *Advances in Neural Information Processing Systems* (pp. 1–9).

[87] Lopez-Paz, D., Muandet, K., Schölkopf, B., & Tolstikhin, I. (2015). Towards a learning theory of causation. In *International Conference on Machine Learning (ICML)*.

[88] Martinato, F., Cesaroni, M., Amati, B., & Guccione, E. (2008). Analysis of Myc-Induced Histone Modifications on Target Chromatin. *PLOS ONE*, 3(11), e3650.

[89] Mathelier, A., Shi, W., & Wasserman, W. W. (2015). Identification of altered cis-regulatory elements in human disease. *Trends in genetics: TIG*, 31(2), 67–76.

[90] Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., & Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (New York, N.Y.)*, 337(6099), 1190–1195.

[91] Maurano, M. T., Wang, H., John, S., Shafer, A., Canfield, T., Lee, K., & Stamatoyannopoulos, J. A. (2015). Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Reports*, 12(7), 1184–1195.

[92] McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5), 495–501.

[93] Michelson, A. M. (2002). Deciphering genetic regulatory codes: A challenge for functional genomics. *Proceedings of the National Academy of Sciences*, 99(2), 546–548.

[94] Mitzenmacher, M. & Upfal, E. (2005). *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press.

[95] Moon, Y.-I., Rajagopalan, B., & Lall, U. (1995). Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3), 2318–2321.

[96] Moreno-Aliaga, M. J., Swarbrick, M. M., Lorente-Cebrián, S., Stanhope, K. L., Havel, P. J., & Martínez, J. A. (2007). Sp1-mediated transcription is involved in the induction of leptin by insulin-stimulated glucose metabolism. *Journal of Molecular Endocrinology*, 38(5), 537–546.

[97] Mouly, E., Chemin, K., Nguyen, H. V., Chopin, M., Mesnard, L., Leite-de Moraes, M., Burlen-defranoux, O., Bandeira, A., & Bories, J.-C. (2010). The Ets-1 transcription factor controls the development and function of natural regulatory T cells. *The Journal of Experimental Medicine*, 207(10), 2113–2125.

[98] Murrell, B., Murrell, D., & Murrell, H. (2014). R2-equitability is satisfiable. *Proceedings of the National Academy of Sciences*, 111(21), E2160–E2160.

[99] Murrell, B., Murrell, D., & Murrell, H. (2016). Discovering general multidimensional associations. *PloS one*, 11(3), e0151551.

[100] Muse, G. W., Gilchrist, D. A., Nechaev, S., Shah, R., Parker, J. S., Grissom, S. F., Zeitlinger, J., & Adelman, K. (2007). RNA polymerase is poised for activation across the genome. *Nature Genetics*, 39(12), 1507–1511.

[101] Muthusamy, N., Barton, K., & Leiden, J. M. (1995). Defective activation and survival of T cells lacking the Ets-1 transcription factor. *Nature*, 377(6550), 639–642.

[102] Nikolic, T., Movita, D., Lambers, M. E., de Almeida, C. R., Biesta, P., Kreefft, K., de Bruijn, M. J., Bergen, I., Galjart, N., Boonstra, A., & Hendriks, R. (2014). The DNA-binding factor Ctcf critically controls gene expression in macrophages. *Cellular and Molecular Immunology*, 11(1), 58–70.

[103] Noutsou, M., Li, J., Ling, J., Jones, J., Wang, Y., Chen, Y., & Sen, G. L. (2017). The Cohesin Complex Is Necessary for Epidermal Progenitor Cell Function through Maintenance of Self-Renewal Genes. *Cell Reports*, 20(13), 3005–3013.

[104] Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H., & Nakatani, Y. (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell*, 87(5), 953–959.

[105] Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S. F. W., Oskarsson, S., Pickrell, J. K., Thom, K., Timshel, P., de Vlaming, R., Abdellaoui, A., Ahluwalia, T. S., Bacelis, J., Baumbach, C., Bjornsdottir, G., Brandsma, J. H., Pina Concas, M., Derringer, J., Furlotte, N. A., Galesloot, T. E., Girotto, G., Gupta, R., Hall, L. M., Harris, S. E., Hofer, E., Horikoshi, M., Huffman, J. E., Kaasik, K., Kalafati, I. P., Karlsson, R., Kong, A., Lahti, J., van der Lee, S. J., deLeeuw, C., Lind, P. A., Lindgren, K.-O., Liu, T., Mangino, M., Marten, J., Mihailov, E., Miller, M. B., van der Most, P. J., Oldmeadow, C., Payton, A., Pervjakova, N., Peyrot, W. J., Qian, Y., Raitakari, O., Rueedi, R., Salvi, E., Schmidt, B., Schraut, K. E.,

Shi, J., Smith, A. V., Poot, R. A., St Pourcain, B., Teumer, A., Thorleifsson, G., Verweij, N., Vuckovic, D., Wellmann, J., Westra, H.-J., Yang, J., Zhao, W., Zhu, Z., Alizadeh, B. Z., Amin, N., Bakshi, A., Baumeister, S. E., Biino, G., Bønnelykke, K., Boyle, P. A., Campbell, H., Cappuccio, F. P., Davies, G., De Neve, J.-E., Deloukas, P., Demuth, I., Ding, J., Eibich, P., Eisele, L., Eklund, N., Evans, D. M., Faul, J. D., Feitosa, M. F., Forstner, A. J., Gandin, I., Gunnarsson, B., Halldórsson, B. V., Harris, T. B., Heath, A. C., Hocking, L. J., Holliday, E. G., Homuth, G., Horan, M. A., Hottenga, J.-J., de Jager, P. L., Joshi, P. K., Jugessur, A., Kaakinen, M. A., Kähönen, M., Kanoni, S., Keltigangas-Järvinen, L., Kiemeney, L. A. L. M., Kolcic, I., Koskinen, S., Kraja, A. T., Kroh, M., Kutalik, Z., Latvala, A., Launer, L. J., Lebreton, M. P., Levinson, D. F., Lichtenstein, P., Lichtner, P., Liewald, D. C. M., Cohort Study, L., Loukola, A., Madden, P. A., Mägi, R., Mäki-Opas, T., Marioni, R. E., Marques-Vidal, P., Meddens, G. A., McMahon, G., Meisinger, C., Meitinger, T., Milaneschi, Y., Milani, L., Montgomery, G. W., Myhre, R., Nelson, C. P., Nyholt, D. R., Ollier, W. E. R., Palotie, A., Paternoster, L., Pedersen, N. L., Petrovic, K. E., Porteous, D. J., Räikkönen, K., Ring, S. M., Robino, A., Rostapshova, O., Rudan, I., Rustichini, A., Salomaa, V., Sanders, A. R., Sarin, A.-P., Schmidt, H., Scott, R. J., Smith, B. H., Smith, J. A., Staessen, J. A., Steinhagen-Thiessen, E., Strauch, K., Terracciano, A., Tobin, M. D., Ulivi, S., Vaccargiu, S., Quaye, L., van Rooij, F. J. A., Venturini, C., Vinkhuyzen, A. A. E., Völker, U., Völzke, H., Vonk, J. M., Vozzi, D., Waage, J., Ware, E. B., Willemsen, G., Attia, J. R., Bennett, D. A., Berger, K., Bertram, L., Bisgaard, H., Boomsma, D. I., Borecki, I. B., Bültmann, U., Chabris, C. F., Cucca, F., Cusi, D., Deary, I. J., Dedoussis, G. V., van Duijn, C. M., Eriksson, J. G., Franke, B., Franke, L., Gasparini, P., Gejman, P. V., Gieger, C., Grabe, H.-J., Gratten, J., Groenen, P. J. F., Gudnason, V., van der Harst, P., Hayward, C., Hinds, D. A., Hoffmann, W., Hyppönen, E., Iacono, W. G., Jacobsson, B., Järvelin, M.-R., Jöckel, K.-H., Kaprio, J., Kardia, S. L. R., Lehtimäki, T., Lehrer, S. F., Magnusson, P. K. E., Martin, N. G., McGue, M., Metspalu, A., Pendleton, N., Penninx, B. W. J. H., Perola, M., Pirastu, N., Pirastu, M., Polasek, O., Posthuma, D., Power, C., Province, M. A., Samani, N. J., Schlessinger, D., Schmidt, R., Sørensen, T. I. A., Spector, T. D., Stefansson, K., Thorsteinsdottir, U., Thurik, A. R., Timpson, N. J., Tiemeier, H., Tung, J. Y., Uitterlinden, A. G., Vitart, V., Vollenweider, P., Weir, D. R., Wilson, J. F., Wright, A. F.,

Conley, D. C., Krueger, R. F., Davey Smith, G., Hofman, A., Laibson, D. I., Medland, S. E., Meyer, M. N., Yang, J., Johannesson, M., Visscher, P. M., Esko, T., Koellinger, P. D., Cesarini, D., & Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533(7604), 539–542.

[106] Ouboussad, L., Kreuz, S., & Lefevre, P. F. (2013). CTCF depletion alters chromatin structure and transcription of myeloid-specific factors. *Journal of Molecular Cell Biology*, 5(5), 308–322.

[107] Pan, X., Solomon, S. S., Borromeo, D. M., Martinez-Hernandez, A., & Raghow, R. (2001). Insulin deprivation leads to deficiency of Sp1 transcription factor in H-411E hepatoma cells and in streptozotocin-induced diabetic ketoacidosis in the rat. *Endocrinology*, 142(4), 1635–1642.

[108] Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation*, 15(6), 1191–1253.

[109] Parnas, O., Jovanovic, M., Eisenhaure, T. M., Herbst, R. H., Dixit, A., Ye, C. J., Przybylski, D., Platt, R. J., Tirosh, I., Sanjana, N. E., Shalem, O., Satija, R., Raychowdhury, R., Mertins, P., Carr, S. A., Zhang, F., Hacohen, N., & Regev, A. (2015). A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell*, 162(3), 675–686.

[110] Pasaniuc, B. & Price, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, 18(2), 117–127.

[111] Paulson, K. G., Iyer, J. G., & Nghiem, P. (2011). Asymmetric lateral distribution of melanoma and merkel cell carcinoma in the united states. *Journal of the American Academy of Dermatology*, 65(1), 35–39.

[112] Peloquin, J. M., Goel, G., Kong, L., Huang, H., Haritunians, T., Sartor, R. B., Daly, M. J., Newberry, R. D., McGovern, D. P., Yajnik, V., Lira, S. A., & Xavier, R. J. (2016). Characterization of candidate genes in inflammatory bowel disease–associated risk loci. *JCI Insight*, 1(13).

[113] Phillips, J. E. & Corces, V. G. (2009). CTCF: Master Weaver of the Genome. *Cell*, 137(7), 1194–1211.

[114] Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3), 447–455.

[115] Poole, C. J. & van Riggelen, J. (2017). MYC—Master Regulator of the Cancer Epigenome and Transcriptome. *Genes*, 8(5).

[116] Price, A. L., Spencer, C. C. A., & Donnelly, P. (2015). Progress and promise in understanding the genetic basis of common diseases. *Proc. R. Soc. B*, 282(1821), 20151684.

[117] Qi, C.-F., Martensson, A., Mattioli, M., Dalla-Favera, R., Lobanenkov, V. V., & Morse, H. C. (2003). CTCF functions as a critical regulator of cell-cycle arrest and death after ligation of the B cell receptor on immature B cells. *Proceedings of the National Academy of Sciences of the United States of America*, 100(2), 633–638.

[118] Raj, P., Rai, E., Song, R., Khan, S., Wakeland, B. E., Viswanathan, K., Arana, C., Liang, C., Zhang, B., Dozmorov, I., Carr-Johnson, F., Mitrovic, M., Wiley, G. B., Kelly, J. A., Lauwerys, B. R., Olsen, N. J., Cotsapas, C., Garcia, C. K., Wise, C. A., Harley, J. B., Nath, S. K., James, J. A., Jacob, C. O., Tsao, B. P., Pasare, C., Karp, D. R., Li, Q. Z., Gaffney, P. M., & Wakeland, E. K. (2016). Regulatory polymorphisms modulate the expression of HLA class II molecules and promote autoimmunity. *eLife*, 5, e12089.

[119] Rana, T. M., Yau, E. H., Kummetha, I. R., Lichinchi, G., Tang, R., & Zhang, Y. (2017). Genome-wide CRISPR screen for essential cell growth mediators in mutant KRAS colorectal cancers. *Cancer Research*.

[120] Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, 10(3), 441–451.

[121] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062), 1518–1524.

[122] Reshef, D. N., Reshef, Y. A., Mitzenmacher, M., & Sabeti, P. C. (2013). Equitability analysis of the maximal information coefficient, with comparisons. *arXiv preprint arXiv:1301.6314*.

[123] Reshef, D. N., Reshef, Y. A., Mitzenmacher, M., & Sabeti, P. C. (2014a). Cleaning up the record on the maximal information coefficient and equitability. *Proceedings of the National Academy of Sciences*, 111(33), E3362–E3363.

[124] Reshef, D. N., Reshef, Y. A., Mitzenmacher, M., & Sabeti, P. C. (2014b). Cleaning up the record on the maximal information coefficient and equitability. *Proceedings of the National Academy of Sciences*, 111(33), E3362–E3363.

[125] Reshef, D. N., Reshef, Y. A., Sabeti, P. C., & Mitzenmacher, M. (2015a). An empirical study of leading measures of dependence. *arXiv preprint arXiv:1505.02214*.

[126] Reshef, D. N., Reshef, Y. A., Sabeti, P. C., Mitzenmacher, M., et al. (2018a). An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics*, 12(1), 123–155.

[127] Reshef, D. N., Reshef, Y. A., Sabeti, P. C., Mitzenmacher, M., et al. (2018b). An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics*, 12(1), 123–155.

[128] Reshef, D. N., Reshef, Y. A., Sabeti, P. C., Mitzenmacher, M., et al. (2018c). An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics*, 12(1), 123–155.

[129] Reshef, Y. A., Finucane, H. K., Kelley, D. R., Gusev, A., Kotliar, D., Ulirsch, J. C., Hormozdiari, F., O'Connor, L., van de Geijn, B., Loh, P.-R., Grossman, S., Bhatia, G., Gazal, S., Palamara, P. F., Pinello, L., Patterson, N., Adams, R., & Price, A. (2017). Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *bioRxiv*, (pp. 204685).

[130] Reshef, Y. A., Reshef, D. N., Finucane, H. K., Sabeti, P. C., & Mitzenmacher, M. (2016a). Measuring dependence powerfully and equitably. *Journal of Machine Learning Research*, 17(212), 1–63.

[131] Reshef, Y. A., Reshef, D. N., Finucane, H. K., Sabeti, P. C., & Mitzenmacher, M. (2016b). Measuring Dependence Powerfully and Equitably. *Journal of Machine Learning Research*, 17(212), 1–63.

[132] Reshef, Y. A., Reshef, D. N., Sabeti, P. C., & Mitzenmacher, M. (2015b). Equitability, interval estimation, and statistical power. *arXiv preprint arXiv:1505.02212*.

[133] Reshef, Y. A., Reshef, D. N., Sabeti, P. C., & Mitzenmacher, M. M. (2015c). Equitability, interval estimation, and statistical power. *arXiv:1505.02212 [cs, math, q-bio, stat]*.

[134] Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., & Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), 317–330.

[135] Romano, S., Vinh, N. X., Verspoor, K., & Bailey, J. (2017). The randomized information coefficient: assessing dependencies in noisy data. *Machine Learning*, (pp. 1–41).

[136] Roulston, M. S. (1999). Estimating the errors on measured entropy and mutual information. *Physica D: Nonlinear Phenomena*, 125(3), 285–294.

[137] Scharer, C. D., Blalock, E. L., Barwick, B. G., Haines, R. R., Wei, C., Sanz, I., & Boss, J. M. (2016). ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells. *Scientific Reports*, 6, 27030.

[138] Schmiedel, B. J., Seumois, G., Samaniego-Castruita, D., Cayford, J., Schulten, V., Chavez, L., Ay, F., Sette, A., Peters, B., & Vijayanand, P. (2016). 17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells. *Nature Communications*, 7, 13426.

[139] Schoech, A., Jordan, D., Loh, P.-R., Gazal, S., O'Connor, L., Balick, D. J., Palamara, P. F., Finucane, H., Sunyaev, S. R., & Price, A. L. (2017). Quantification of frequency-dependent genetic architectures and action of negative selection in 25 UK Biobank traits. *bioRxiv*, (pp. 188086).

[140] Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K., et al. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5), 2263–2291.

[141] Sharrocks, A. D., Brown, A. L., Ling, Y., & Yates, P. R. (1997). The ETS-domain transcription factor family. *The International Journal of Biochemistry & Cell Biology*, 29(12), 1371–1387.

[142] Shi, H., Kichaev, G., & Pasaniuc, B. (2016). Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics*, 99(1), 139–153.

[143] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3), 751–754.

[144] Simon, N. & Tibshirani, R. (2012). Comment on "Detecting novel associations in large data sets". *Unpublished*.

[145] Soletti, R. C., Rodrigues, N. A. L. V., Biasoli, D., Luiz, R. R., de Souza, H. S. P., & Borges, H. L. (2013). Immunohistochemical Analysis of Retinoblastoma and β-Catenin as an Assistant Tool in the Differential Diagnosis between Crohn's Disease and Ulcerative Colitis. *PLOS ONE*, 8(8), e70786.

[146] Speed, T. (2011). A correlation for the 21st century. *Science*, 334(6062), 1502–1503.

[147] Stoer, J. & Bulirsch, R. (1980). *Introduction to Numerical Analysis*. Springer-Verlag.

[148] Stone, C. J. (1977). Consistent nonparametric regression. *The annals of statistics*, (pp. 595–620).

[149] Storey, J. D. & Tibshirani, R. (2003a). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440–9445.

[150] Storey, J. D. & Tibshirani, R. (2003b). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440–9445.

[151] Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I., Lam, D., Liberzon, A., Toder, C., Bagul, M., Orzechowski, M., Enache, O. M., Piccioni, F., Berger, A. H., Shamji, A., Brooks, A. N., Vrcic, A., Flynn, C., Rosains, J., Takeda, D., Davison, D., Lamb, J., Ardlie, K., Hogstrom, L., Gray, N. S., Clemons, P. A., Silver, S., Wu, X., Zhao, W.-N., Read-Button, W., Wu, X., Haggarty, S. J., Ronco, L. V., Boehm, J. S., Schreiber, S. L., Doench, J. G., Bittker, J. A., Root, D. E., Wong, B., & Golub, T. R. (2017). A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. *bioRxiv*, (pp. 136168).

[152] Sugiyama, M. & Borgwardt, K. (2013). Measuring statistical dependence via the mutual information dimension. In *The International Joint Conferences on Artificial Intelligence (IJCAI)* (pp. 1692–1698).: AAAI Press.

[153] Székely, G. J. & Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4), 1236–1265.

[154] Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794.

[155] Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C.-L., Liu, E. T., Wilczynski, G. M., Plewczynski, D., Li, G., & Ruan, Y. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, 163(7), 1611–1627.

[156] Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*, 165(6), 1519–1529.

[157] The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74.

[158] The UniProt Consortium (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169.

[159] Torrano, V., Chernukhin, I., Docquier, F., D'Arcy, V., León, J., Klenova, E., & Delgado, M. D. (2005). CTCF regulates growth and erythroid differentiation of human myeloid leukemia cells. *The Journal of Biological Chemistry*, 280(30), 28152–28161.

[160] Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., & Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2), 124–130.

[161] Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., Meyers, R. M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W. F. J., Howell, S., Merkel, E., Ghandi, M., Garraway, L. A., Root, D. E., Golub, T. R., Boehm, J. S., & Hahn, W. C. (2017). Defining a Cancer Dependency Map. *Cell*, 170(3), 564–576.e16.

[162] Turk-Browne, N. B. (2013). Functional interactions as big data in the human brain. *Science*, 342(6158), 580–584.

[163] van Oevelen, C., Collombet, S., Vicent, G., Hoogenkamp, M., Lepoivre, C., Badeaux, A., Bussmann, L., Sardina, J. L., Thieffry, D., Beato, M., Shi, Y., Bonifer, C., & Graf, T. (2015). C/EBPα Activates Pre-existing and De Novo Macrophage Enhancers during Induced Pre-B Cell Transdifferentiation and Myelopoiesis. *Stem Cell Reports*, 5(2), 232–247.

[164] Verbanck, M., Chen, C.-Y., Neale, B., & Do, R. (2017). Widespread pleiotropy confounds causal relationships between complex traits and diseases inferred from Mendelian randomization. *bioRxiv*, (pp. 157552).

[165] Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., Thurman, R. E., Kaul, R., Myers, R. M., & Stamatoyannopoulos, J. A. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research*, 22(9), 1680–1688.

[166] Wang, X., Jiang, B., & Liu, J. S. (2017). Generalized r-squared for detecting dependence. *Biometrika*, 104(1), 129–139.

[167] Warg, L. A., Oakes, J. L., Burton, R., Neidermyer, A. J., Rutledge, H. R., Groshong, S., Schwartz, D. A., & Yang, I. V. (2012). The role of the E2F1 transcription factor in the innate immune response to systemic LPS. *American Journal of Physiology. Lung Cellular and Molecular Physiology*, 303(5), L391–400.

[168] Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.-H., Abdellaoui, A., Batista, S., Butler, C., Chen, G., Chen, T.-H., D'Ambrosio, D., Gallins, P., Ha, M. J., Hottenga, J. J., Huang, S., Kattenberg, M., Kochar, J., Middeldorp, C. M., Qu, A., Shabalin, A., Tischfield, J., Todd, L., Tzeng, J.-Y., van Grootheest, G., Vink, J. M., Wang, Q., Wang, W., Wang, W., Willemsen, G., Smit, J. H., de Geus, E. J., Yin, Z., Penninx, B. W. J. H., & Boomsma, D. I. (2014). Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*, 46(5), 430–437.

[169] Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., Hua, Y., Gueroussov, S., Najafabadi, H. S., Hughes, T. R., Morris, Q., Barash, Y., Krainer, A. R., Jojic, N., Scherer, S. W., Blencowe, B. J., & Frey, B. J. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), 1254806.

[170] Xu, J., Grant, G., Sabin, L. R., Gordesky-Gold, B., Yasunaga, A., Tudor, M., & Cherry, S. (2012). Transcriptional Pausing Controls A Rapid Antiviral Innate Immune Response In Drosophila. *Cell host & microbe*, 12(4), 531–543.

[171] Yamazaki, K., Umeno J, Takahashi A, Hirano A, Johnson Ta, Kumasaka N, Morizono T, Hosono N, Kawaguchi T, Takazoe M, Yamada T, Suzuki Y, Tanaka H,

Motoya S, Hosokawa M, Arimura Y, Shinomura Y, Matsui T, Matsumoto T, Iida M, Tsunoda T, Nakamura Y, Kamatani N, & Kubo M (2013). A genome-wide association study identifies 2 susceptibility Loci for Crohn's disease in a Japanese population. *Gastroenterology*, 144(4), 781–788.

[172] Yang, G., Lim, C.-Y., Li, C., Xiao, X., Radda, G. K., Li, C., Cao, X., & Han, W. (2009). FoxO1 inhibits leptin regulation of pro-opiomelanocortin promoter activity by blocking STAT3 interaction with specificity protein 1. *The Journal of Biological Chemistry*, 284(6), 3719–3727.

[173] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010a). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7), 565–569.

[174] Yang, W., Shen, N., Ye, D.-Q., Liu, Q., Zhang, Y., Qian, X.-X., Hirankarn, N., Ying, D., Pan, H.-F., Mok, C. C., Chan, T. M., Wong, R. W. S., Lee, K. W., Mok, M. Y., Wong, S. N., Leung, A. M. H., Li, X.-P., Avihingsanon, Y., Wong, C.-M., Lee, T. L., Ho, M. H. K., Lee, P. P. W., Chang, Y. K., Li, P. H., Li, R.-J., Zhang, L., Wong, W. H. S., Ng, I. O. L., Lau, C. S., Sham, P. C., Lau, Y. L., & Asian Lupus Genetics Consortium (2010b). Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. *PLoS genetics*, 6(2), e1000841.

[175] Yasui, D., Peedicayil, J., & Grayson, D. R. (2016). *Neuropsychiatric Disorders and Epigenetics*. Academic Press. Google-Books-ID: _dycBAAAQBAJ.

[176] Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, 103(481), 309–316.

[177] Ying, L., Marino, J., Hussain, S. P., Khan, M. A., You, S., Hofseth, A. B., Trivers, G. E., Dixon, D. A., Harris, C. C., & Hofseth, L. J. (2005). Chronic inflammation promotes retinoblastoma protein hyperphosphorylation and E2F1 activation. *Cancer Research*, 65(20), 9132–9136.

[178] Zeng, H., Edwards, M. D., Liu, G., & Gifford, D. K. (2016a). Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics (Oxford, England)*, 32(12), i121–i127.

[179] Zeng, H., Hashimoto, T., Kang, D. D., & Gifford, D. K. (2016b). GERV: A statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics*, 32(4), 490–496.

[180] Zhang, X., Yang, R., Jia, Y., Cai, D., Zhou, B., Qu, X., Han, H., Xu, L., Wang, L., Yao, Y., & Yang, G. (2014). Hypermethylation of Sp1 binding site suppresses hypothalamic POMC in neonates and may contribute to metabolic disorders in adults: Impact of maternal dietary CLAs. *Diabetes*, 63(5), 1475–1487.

[181] Zhao, J., Giles, B. M., Taylor, R. L., Yette, G. A., Lough, K. M., Ng, H. L., Abraham, L. J., Wu, H., Kelly, J. A., Glenn, S. B., Adler, A. J., Williams, A. H., Comeau, M. E., Ziegler, J. T., Marion, M., Alarcón-Riquelme, M. E., Networks, f. t. B., GENLES, Alarcón, G. S., Anaya, J.-M., Bae, S.-C., Kim, D., Lee, H.-S., Criswell, L. A., Freedman, B. I., Gilkeson, G. S., Guthridge, J. M., Jacob, C. O., James, J. A., Kamen, D. L., Merrill, J. T., Sivils, K. M., Niewold, T. B., Petri, M. A., Ramsey-Goldman, R., Reveille, J. D., Scofield, R. H., Stevens, A. M., Vilá, L. M., Vyse, T. J., Kaufman, K. M., Harley, J. B., Langefeld, C. D., Gaffney, P. M., Brown, E. E., Edberg, J. C., Kimberly, R. P., Ulgiati, D., Tsao, B. P., & Boackle, S. A. (2016a). Preferential association of a functional variant in complement receptor 2 with antibodies to double-stranded DNA. *Annals of the Rheumatic Diseases*, 75(1), 242–252.

[182] Zhao, M., Wang, J., Liao, W., Li, D., Li, M., Wu, H., Zhang, Y., Gershwin, M. E., & Lu, Q. (2016b). Increased 5-hydroxymethylcytosine in CD4(+) T cells in systemic lupus erythematosus. *Journal of Autoimmunity*, 69, 64–73.

[183] Zhou, J. & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10), 931–934.

[184] Zhu, X. & Stephens, M. (2017). A large-scale genome-wide enrichment analysis identifies new trait-associated genes, pathways and tissues across 31 human phenotypes. *bioRxiv*, (pp. 160770).