Generalized R-squared for detecting dependence

Citation

Wang, X., B. Jiang, and J. S. Liu. 2017. "Generalized R-squared for detecting dependence." Biometrika 104, no. 1: 129-139. doi: 10.1093/biomet/asw071

Published Version

10.1093/biomet/asw071

Permanent link

Terms of Use

Share Your Story

Biometrika (2016), **xx**, x, *pp*. 1–14
Printed in Great Britain

Generalized R-squared for detecting dependence

BY X. WANG
Department of Statistics, Harvard University, Cambridge, Massachusetts, U.S.A
xufeiwang@fas.harvard.edu

B. JIANG
Two Sigma Investments, Limited Partnership, New York, New York, U.S.A
bojiang83@gmail.com

AND J. S. LIU
Department of Statistics, Harvard University, Cambridge, Massachusetts, U.S.A
jliu@stat.harvard.edu

SUMMARY

Detecting dependence between two random variables is a fundamental problem. Although the Pearson correlation is effective for capturing linear dependency, it can be entirely powerless for detecting nonlinear and/or heteroscedastic patterns. We introduce a new measure, G-squared, to test whether two univariate random variables are independent and to measure the strength of their relationship. The G-squared is almost identical to the square of the Pearson correlation coefficient, R-squared, for linear relationships with constant error variance, and has the intuitive meaning of the piecewise R-squared between the variables. It is particularly effective in handling nonlinearity and heteroscedastic errors. We propose two estimators of G-squared and show their consistency. Simulations demonstrate that G-squared estimates are among the most powerful test statistics compared with several state-of-the-art methods.

Some key words: Bayes factor; Coefficient of determination; Hypothesis test; Likelihood ratio.

1. INTRODUCTION

The Pearson correlation coefficient is widely used to detect and measure the dependence of two random quantities. The square of its least-squares estimate, popularly known as R-squared, is often used to quantify how linearly related two random variables are. However, the shortcomings of the R-squared as a measure of the strength of dependence are also significant, as discussed recently by Reshef et al. (2011), which has inspired the development of many new methods for detecting dependence.

The Spearman correlation calculates the Pearson correlation coefficient between rank statistics. Although more robust than the Pearson correlation, this method still cannot capture non-monotone relationships. The alternating conditional expectation method was introduced by Breiman & Friedman (1985) to approximate the maximal correlation between X and Y, i.e., to find the optimal transformations of the data, $f(X)$ and $g(Y)$, so that their correlation is maximized. The implementation of the method has its limitations because it is unfeasible to search over all possible transformations. Estimating mutual information is another popular approach due to the fact that the mutual information is zero if and only if X and Y are independent. Furthermore, Kraskov et al. (2004) proposed an efficient method by estimating the entropy of X, Y and (X, Y) separately. The method was claimed to be numerically exact for independent cases and to also work for high dimensional variables. An energy distance-based method (Szèkely et al., 2007; Szèkely & Rizzo, 2009) and a kernel-based method (Gretton et al., 2005, 2012) appeared separately in statistics and machine learning literature to solve the two-sample test problem and have corresponding usage in independence tests. The two methods were recently shown to be equivalent (Sejdinovic et al., 2013). Methods based on empirical cumulative distribution functions (Hoeffding, 1948), empirical copula (Genest & Rémillard, 2004) and empirical characteristic functions (Kankainen & Ushakov, 1998; Huskova & Meintanis, 2008) have also been proposed for detecting dependence.

Another set of approaches is based on discretizations of the random variables. Known as grid-based methods, they are primarily designed to test independence between univariate random variables. Reshef et al. (2011) introduced a new statistic, the maximum information coefficient, which focuses on the generality and equitability of a dependence statistic. Y. Reshef and coauthors (arXiv:1505.02213) proposed two new estimators for this quantity, which are empirically more powerful and easier to compute. Heller et al. (2016) proposed a grid based method, which utilizes the χ^2 statistic to test independence and is a distribution-free test.

To measure how accurately an independence test can reflect the strength of dependence between two random variables, Reshef et al. (2011) introduced the idea of equitability, which was more carefully defined and examined in (Y. Reshef and coauthors, arXiv:1505.02212). Equitability requires that the same value of the statistic implies the same amount of dependence, regardless of the type of relationship. Whether there exists a statistic that can achieve exact equitability is still subject to debate. However, given a collection of functional relationships with varying noise levels, we can compare the empirical equitability of different statistics through simulation studies.

Intuitively, if there is a functional relationship between two random variables X and Y, it is natural to estimate their relationship using a nonparametric technique and use the fraction of reduction in the sum of squares as a measure of the strength of the relationship. In this way, one can both detect dependence and provide an equitable statistic. In contrast, it is more challenging for other types of dependence measures, such as energy-based or entropy-based methods, to be equitable. Doksum et al. (1994) and Blyth (1994) discussed the correlation curve to measure the strength of the relationship. However, a direct use of nonparametric curve estimation may rely too heavily on the smoothness assumption of the relationship; it also cannot deal with heteroscedastic noises.

The G^2 proposed in this paper is derived from a regularized likelihood ratio test for piecewise linear relationships and can be viewed as an integration of continuous and discrete methods. The G-squared statistic is a function of both the conditional mean and conditional variance of one variable given the other. It is thus capable of detecting general functional relationships with heteroscedastic error variances. An estimate of G^2 can be derived via the same likelihood ratio approach as the R^2 when the true underlying relationship is linear. Thus, it is reasonable that G^2 is almost identical to the R^2 for linear relationships. Efficient estimates of G^2 can be computed quickly by a dynamic programming method, whereas Reshef et al. (2011) and Heller et al. (2016) have to consider grids on two variables simultaneously and hence require longer computational time, as shown by our simulation studies. We will also show that, in terms of both power and equitability, G^2 is among the best statistics for independence testing in consideration of a wide range of functional relationships.

2. MEASURING DEPENDENCE WITH G-SQUARED

2·1. *Defining the G^2 as a generalization of the R^2*

The R-squared measures how well the data fit a linear regression model. Given $Y = \mu + \beta X + e$ with $e \sim N(0, \sigma^2)$, the standard estimate of R-squared can be derived from a likelihood ratio test statistic for testing $\mathcal{H}_0 : \beta = 0$ against $\mathcal{H}_1 : \beta \neq 0$, i.e.,

$$R^2 = 1 - \left(\frac{L(\widehat{\theta})}{L_0(\widehat{\theta}_0)} \right)^{-2/n},$$

and $L_0(\widehat{\theta}_0)$ and $L(\widehat{\theta})$ are the maximized likelihoods under $\mathcal{H}_0$ and $\mathcal{H}_1$.

Throughout the paper, we let X and Y be univariate continuous random variables. As a working model, we assume that the relationship between X and Y can be characterized as $Y = f(X) + \epsilon\sigma_X$, $\epsilon \sim N(0,1)$ and $\sigma_X > 0$. If X and Y are independent, then $f(X) \equiv \mu$ and $\sigma_X^2 \equiv \sigma^2$. Now, let us look at the piecewise linear relationship

$$f(X) = \mu_h + \beta_h X, \quad \sigma_X^2 = \sigma_h^2, \quad c_{h-1} < X \leq c_h,$$

where c_h $(h = 0, \ldots, K)$ are called the breakpoints. While this working model allows for heteroscedasticity, it requires constant variance within each segment between two adjacent breakpoints. Testing whether X and Y are independent is equivalent to testing whether $\mu_h = \mu$ and $\sigma_h^2 = \sigma^2$. Given c_h $(h = 0, \ldots, K)$, the likelihood ratio test statistic can be written as

$$\textsc{lr} = \exp\left(\frac{n}{2} \log \widehat{\nu}^2 - \sum_{h=1}^{K} \frac{n_h}{2} \log \widehat{\sigma}_h^2 \right),$$

where $\widehat{\nu}^2$ is the overall sample variance of Y and $\widehat{\sigma}_h^2$ is the residual variance after regressing Y on X for $X \in (c_{h-1}, c_h]$. Because R^2 is a transformation of the likelihood ratio and converges to the square of Pearson correlation coefficient, we perform the same transformation on LR. The resulting test statistic converges to a quantity related to the conditional mean and the conditional variance of Y on X. It is easy to show that, as $n \to \infty$,

$$1 - (\textsc{lr})^{-2/n} \to 1 - \exp\left[E\{\log \mathrm{var}(Y \mid X)\} - \log \mathrm{var}(Y) \right]. \tag{1}$$

When $h = 1$, the relationship degenerates to a simple linear relationship and $1 - (\textsc{lr})^{-2/n}$ is exactly R^2.

More generally, because a piecewise linear function can approximate any almost-everywhere continuous function, we can employ the same hypothesis testing framework as above to derive (1) for any such approximation. Thus, for any pair of random variables (X, Y), the following concept is a natural generalization of the R-squared:

$$G_{Y|X}^2 = 1 - \exp\left[E\{\log \mathrm{var}(Y \mid X)\} - \log \mathrm{var}(Y) \right],$$

in which we require that $\mathrm{var}(Y) < \infty$. Evidently, $G_{Y|X}^2$ lies between zero and one, and is equal to zero if and only if both $E(Y \mid X)$ and $\mathrm{var}(Y \mid X)$ are constant. The definition of $G_{Y|X}^2$ is closely related to the R-squared defined by segmented regression (Oosterbaan & Ritzema, 2006) discussed in the Supplementary Material. We symmetrize $G_{Y|X}^2$ to arrive at the following quantity as the definition of the G-squared:

$$G^2 = \max(G_{Y|X}^2,\, G_{X|Y}^2),$$

provided $\mathrm{var}(X) + \mathrm{var}(Y) < \infty$. Thus, $G^2 = 0$ if and only if $E(X \mid Y)$, $E(Y \mid X)$, $\mathrm{var}(Y \mid X)$ and $\mathrm{var}(X \mid Y)$ are all constant, which is not equivalent to independence of X and Y. In practice, however, dependent cases with $G^2 = 0$ are rare.

2·2. *Estimation of G^2*

Without loss of generality, we focus on the estimation of $G^2_{Y|X}$; $G^2_{X|Y}$ can be estimated in the same way by flipping X and Y. When $Y = f(X) + \epsilon\sigma_X$ and $\epsilon \sim N(0,1)$ for an almost-everywhere continuous function $f(\cdot)$, we can use a piecewise linear function to approximate $f(X)$ and estimate G^2. However, in practice the number and locations of the breakpoints are unknown. We propose two estimators of $G^2_{Y|X}$, the first aiming to find the maximum penalized likelihood ratio among all possible piecewise linear approximations, and the second focusing on a Bayesian average of all approximations.

Suppose we have n sorted independent observations, (x_i, y_i) $(i = 1, \ldots, n)$, such that $x_1 < \cdots < x_n$. For the set of breakpoints, we only need to consider $c_h = x_i$. Each interval $s_h = (c_{h-1}, c_h]$ is called a slice of the observations, so that c_h $(h = 0, \ldots, K)$ divide the range of X into K non-overlapping slices. Let n_h denote the number of observations in slice h, and let $S(X)$ denote a slicing scheme of X, that is, $S(x_i) = h$ if $x_i \in s_h$, which is abbreviated as S whenever the meaning is clear. Let $|S|$ be the number of slices in S and let m_S denote the minimum size of all the slices.

To avoid overfitting when maximizing log-likelihood ratios over both unknown parameters and all possible slicing schemes, we restrict the minimum size of each slice as $m_S \geq \lceil n^{1/2} \rceil$ and maximize the log-likelihood ratio with a penalty on the number of slices. For simplicity, let $m = \lceil n^{1/2} \rceil$. Thus, we focus on the following penalized log-likelihood ratio

$$nD(Y \mid S, \lambda_0) = 2 \log \mathrm{LR}_S - \lambda_0(|S| - 1) \log n, \tag{2}$$

where LR_S is the likelihood ratio for S and $\lambda_0 \log n > 0$ is the penalty for incurring one additional slice. From a Bayesian perspective, this is equivalent to assigning the prior distribution for the number of slices to be proportional to $n^{-\lambda_0(|S|-1)/2}$. Suppose each observation x_i $(i = 1, \ldots, n-1)$ has probability $p_n = n^{-\lambda_0/2}/(1 + n^{-\lambda_0/2})$ of being the breakpoint independently. Then the probability of a slicing scheme S is

$$p_n^{|S|-1}(1 - p_n)^{n-|S|} \propto \left(\frac{p_n}{1 - p_n}\right)^{|S|-1} = n^{-\lambda_0(|S|-1)/2}.$$

When $\lambda_0 = 3$, the statistic $-nD(Y \mid S, \lambda_0)$ is equivalent to the Bayesian information criterion (Schwarz, 1978) up to a constant.

Treating the slicing scheme as a nuisance parameter, we can maximize over all allowable slicing schemes to obtain that

$$D(Y \mid X, \lambda_0) = \max_{m_S \geq m} D(Y \mid S, \lambda_0).$$

Our first estimator of $G^2_{Y|X}$, which we call G^2_m with m representing the maximum likelihood ratio, can be defined as

$$G^2_m(Y \mid X, \lambda_0) = 1 - \exp\{-D(Y \mid X, \lambda_0)\}.$$

Thus, the overall G-squared can be estimated as

$$G^2_m(\lambda_0) = \max\{G^2_m(Y \mid X, \lambda_0),\ G^2_m(X \mid Y, \lambda_0)\}.$$

By definition, $G^2_m(\lambda_0)$ lies between 0 and 1 and $G^2_m(\lambda_0) = R^2$ when the optimal slicing schemes for both directions have only one slice. Later, we will show that when X and Y are a bivariate normal, $G^2_m(\lambda_0) = R^2$ almost surely for large λ_0.

Another attractive way to estimate G^2 is to integrate out the nuisance slicing scheme parameter. A full Bayesian approach would require us to compute the Bayes factor (Kass & Raftery, 1995), which may be undesirable since we do not wish to impose too strong a modeling assumption. On the other hand, however, the Bayesian formalism may guide us to a desirable integration strategy for the slicing scheme. We thus put the problem into a Bayes framework and compute the Bayes factor for comparing the null and alternative models. The null model is only one model while the alternative is any piecewise linear model, possibly with countably infinite pieces. Let $p_0(y_1, \ldots, y_n)$ be the marginal probability of the data under the null. Let ω_S be the prior probability for slicing scheme S and let $p_S(y_1, \ldots, y_n)$ denote the marginal probability of the data under S. The Bayes factor can be written as

$$\textsc{bf} = \sum_{m_s \geq m} \omega_S \times p_S(y_1, \ldots, y_n)/p_0(y_1, \ldots, y_n). \tag{3}$$

The marginal probabilities are not easy to compute even with proper priors. Schwarz (1978) states that if the data distribution is in the exponential family and the parameter is of dimension k, the marginal probability of the data can be approximated as

$$p(y_1, \ldots, y_n) \approx \textsc{l} \exp\left\{-k(\log n - \log 2\pi)/2\right\}, \tag{4}$$

where $\textsc{l}$ is the maximized likelihood. In our setup, the number of parameters k for the null model is two, and for an alternative model with a slicing scheme S is $3|S|$. Plugging expression (4) into both the numerator and the denominator of (3), we obtain

$$\textsc{bf} \approx \sum_{S\colon m_s \geq m} \omega_S \textsc{lr}_S \exp\left\{-(3|S| - 2)(\log n - \log 2\pi)/2\right\}. \tag{5}$$

If we take $\omega_S \propto n^{-\lambda_0(|S|-1)/2}$ $(\lambda_0 > 0)$, which corresponds to the penalty term in (2) and is involved in defining G^2_m, the approximated Bayes factor can be restated as

$$\textsc{bf}(\lambda_0) = \left\{\sum_{S\colon m_S \geq m} n^{-\frac{\lambda_0(|S|-1)}{2}}\right\}^{-1} \sum_{S\colon m_S \geq m} \left(\frac{2\pi}{n}\right)^{\frac{3|S|-2}{2}} \exp\left\{\frac{n}{2} D(Y \mid S, \lambda_0)\right\}. \tag{6}$$

As we will discuss in Section 2·5, $\textsc{bf}(\lambda_0)$ can serve as a marginal likelihood function for λ_0 and can be used to find an optimal λ_0 suitable for a particular data set. This quantity also looks like an average version of G^2_m, but with an additional penalty. Since $\textsc{bf}(\lambda_0)$ can take values below 1, its transformation $1 - \textsc{bf}(\lambda_0)^{-2/n}$, as in the case where we derived the R^2 via the likelihood ratio test, can take negative values, especially when X and Y are independent, and it is therefore not an ideal estimator of G^2.

By removing the model size penalty term in (5), we obtain a modified version, which is simply a weighted average of the likelihood ratios and is guaranteed to be greater than or equal to 1:

$$\textsc{bf}^*(\lambda_0) = \left\{\sum_{S\colon m_S \geq m} n^{-\frac{\lambda_0(|S|-1)}{2}}\right\}^{-1} \sum_{S: m_S \geq m} \exp\left\{\frac{n}{2} D(Y \mid S, \lambda_0)\right\}.$$

We can thus define a quantity similar to our likelihood formulation of R-squared,

$$G^2_t(Y \mid X, \lambda_0) = 1 - \textsc{bf}^*(\lambda_0)^{-2/n},$$

which we call the total G-squared, and define

$$G_t^2(\lambda_0) = \max\{G_t^2(Y \mid X, \lambda_0),\ G_t^2(X \mid Y, \lambda_0)\}.$$

We show later that $G_m^2(\lambda_0)$ and $G_t^2(\lambda_0)$ are both consistent estimators of G^2.

2·3. *Theoretical properties of the G^2 estimators*

In order to show that $G_m^2(\lambda_0)$ and $G_t^2(\lambda_0)$ converge to G^2 as the sample size goes to infinity, we introduce the notations: $\mu_X(y) = E(X \mid Y = y)$, $\mu_Y(x) = E(Y \mid X = x)$, $\nu_X^2(y) = \mathrm{var}(X \mid Y = y)$ and $\nu_Y^2(x) = \mathrm{var}(Y \mid X = x)$ as well as the following regularity conditions:

Condition 1. The random variables X and Y are bounded continuously with finite variances such that $\nu_Y^2(x), \nu_X^2(y) > b^{-2} > 0$ almost everywhere for some constant b.

Condition 2. The functions $\mu_Y(x)$, $\mu_X(y)$, $\nu_Y^2(x)$ and $\nu_X^2(y)$ have continuous derivatives almost everywhere.

Condition 3. There exists a constant $C > 0$ such that

$$\max\{\left|\mu_X'(y)\right|,\ \left|\nu_X'(y)\right|\} \le C\nu_X(y), \quad \max\{\left|\mu_Y'(x)\right|,\ \left|\nu_Y'(x)\right|\} \le C\nu_Y(x)$$

almost surely.

With these preparations, we can state our main results.

THEOREM 1. *Under Conditions 1-3, for all $\lambda_0 > 0$,*

$$G_m^2(Y \mid X, \lambda_0) \to G_{Y|X}^2, \quad G_t^2(Y \mid X, \lambda_0) \to G_{Y|X}^2$$

almost surely as $n \to \infty$. Thus, $G_m^2(\lambda_0)$ and $G_t^2(\lambda_0)$ are consistent estimators of G^2.

A proof of the theorem and numerical studies of the consistency are in the Supplementary Material. It is expected that $G_m^2(\lambda_0)$ should converge to G^2 just because of its construction. It is surprising that $G_t^2(\lambda_0)$ also converges to G^2. The result, which links G^2 estimation with the likelihood ratio and Bayesian formalism, suggests that most of the information up to the second moment has been fully utilized in the two test statistics. The theorem thus supports the use of $G_m^2(\lambda_0)$ and $G_t^2(\lambda_0)$ for testing whether X and Y are independent. The null distributions of the two statistics depend on the marginal distributions of X and Y, which can be generated empirically using permutation. One can also do a quantile-based transformation on X and Y such that their marginal distributions are standard normal; however, the G^2 based on the transformed data tends to lose some power.

When X and Y are bivariate normal, the G-squared statistic is almost the same as the R-squared when λ_0 is large enough.

THEOREM 2. *If X and Y follow bivariate normal distribution, then for n large enough*

$$\mathrm{pr}\left\{G_m^2(\lambda_0) = R^2\right\} > 1 - 3n^{-\lambda_0/3+5}.$$

So for $\lambda_0 > 18$ and $n \to \infty$, we have $G_m^2(\lambda_0) = R^2$ almost surely .

The lower bound on λ_0 is not tight and can be relaxed in practice. Empirically, we have observed that $\lambda_0 = 3$ is large enough for $G_m^2(\lambda_0)$ to be very close to R^2 in the bivariate normal setting.

2·4. *Dynamic programming algorithm for computing G_m^2 and G_t^2*

The brute force calculation of either G_m^2 or G_t^2 has a computational complexity of $O(2^n)$ and is prohibitive in practice. Fortunately, we have found a dynamic programming scheme for computing both quantities with a time complexity of $O(n^2)$. The algorithms for computing $G_m^2(Y \mid X, \lambda_0)$ and $G_t^2(Y \mid X, \lambda_0)$ are roughly the same except for one operation, i.e., maximization versus summation, and can be summarized by the following steps:

Step 1 (*Data preparation*). Arrange the observed pairs (x_i, y_i) $(i = 1, \ldots, n)$ according to the sorted xs from low to high. Then normalize y_i $(i = 1, \ldots, n)$ such that $\sum_{i=1}^n y_i = 0$ and $\sum_{i=1}^n y_i^2 = 1$.

Step 2 (*Main algorithm*). Define $m = \lceil n^{1/2} \rceil$ as the smallest slice size, $\lambda = -\lambda_0 \log(n)/2$ and $\alpha = e^{\lambda}$. Initialize three sequences: $(M_i,\ B_i,\ T_i)$ $(i = 1, \ldots, n)$ with $M_1 = 0$ and $B_1 = T_1 = 1$. For $i = m, \ldots, n$, recursively fill in entries of the tables with

$$M_i = \max_{k \in K_i} \left(\lambda + M_k + l_{k:i}\right), \quad B_i = \sum_{k \in K_i} \alpha B_k, \quad T_i = \sum_{k \in K_i} \alpha T_k L_{k:i},$$

where $K_i = \{1\} \cup \{k : k = m+1, \ldots, i-m+1\}$, $l_{k:i} = -(i-k)\log(\widehat{\sigma}_{k:i}^2)/2$ and $L_{k:i} = \exp\{l_{k:i}\}$, with $\widehat{\sigma}_{k:i}^2$ as the residual variance of regressing y on x for observations (x_j, y_j) $(j = k, \ldots, i)$.

Step 3. The final result is

$$G_m^2 = 1 - \exp\left\{M_n - \lambda\right\}, \quad G_t^2 = 1 - (T_n/B_n)^{-2/n}.$$

Here, M_i $(i = m, \ldots, n)$ stores the partial maximized likelihood ratio up to the ordered observation (x_k, y_k) $(k = 1, \ldots, i)$, B_i $(i = m, \ldots, n)$ stores the partial normalizing constant, and T_i $(i = m, \ldots, n)$ stores the partial sum of the likelihood ratios. When n is extremely large, we can speed up the algorithm by considering fewer slice schemes. For example, we can divide X into chunks of size m by rank and consider only slicing schemes between the chunks. For this method, the computational complexity is $O(n)$. We can compute $G_m^2(X \mid Y, \lambda_0)$ and $G_t^2(X \mid Y, \lambda_0)$ similarly to get $G_m^2(\lambda_0)$ and $G_t^2(\lambda_0)$. Empirically, the algorithm is faster than many other powerful methods as shown in the Supplementary Material.

2·5. *An empirical Bayes strategy for selecting* λ_0

Although the choice of the penalty parameter λ_0 is not critical for the general use of G^2, we typically use $\lambda_0 = 3$ for G_m^2 and G_t^2 because $D(Y \mid X, 3)$ is equivalent to the Bayesian information criterion. Fine-tuning λ_0 can improve the estimation of G^2. We thus propose a data-driven strategy for choosing λ_0 adaptively. $\textsc{bf}(\lambda_0)$ in (6) can be viewed as an approximation to $\text{pr}(y_1, \ldots, y_n \mid \lambda_0)$ up to a normalizing constant. We thus can use the maximum likelihood principle to choose the λ_0 that maximizes $\textsc{bf}(\lambda_0)$. We then use the chosen λ_0 to compute G_m^2 and G_t^2 as estimators of G^2. In practice, we evaluate $\textsc{bf}(\lambda_0)$ for a set of discrete λ_0 values, e.g., $\{0.5,\ 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 3.5,\ 4\}$, and pick the one that maximizes $\textsc{bf}(\lambda_0)$. $\textsc{bf}(\lambda_0)$ can be computed efficiently via a dynamic programming algorithm similar to that described in Section 2·4. As an illustration, we consider the sampling distributions of $G_m^2(\lambda_0)$ and $G_t^2(\lambda_0)$ with $\lambda_0 = 0.5,\ 1.5,\ 2.5$ and 3.5 for the following two examples

Example 1. $X \sim N(0,1)$, $Y = X + \sigma\epsilon$ and $\epsilon \sim N(0,1)$.

Example 2. $X \sim N(0,1)$, $Y = \sin(4\pi x)/0.7 + \sigma\epsilon$ and $\epsilon \sim N(0,1)$.

We simulated $n = 225$ data points. For each model, we set $\sigma = 1$ so that $G_{Y|X}^2 = 0.5$ and performed 1,000 replications. Figure 1 shows histograms of $G_m^2(\lambda_0)$ and $G_t^2(\lambda_0)$ with different λ_0 values. The results demonstrate that, for relationships that can be approximated well by a linear

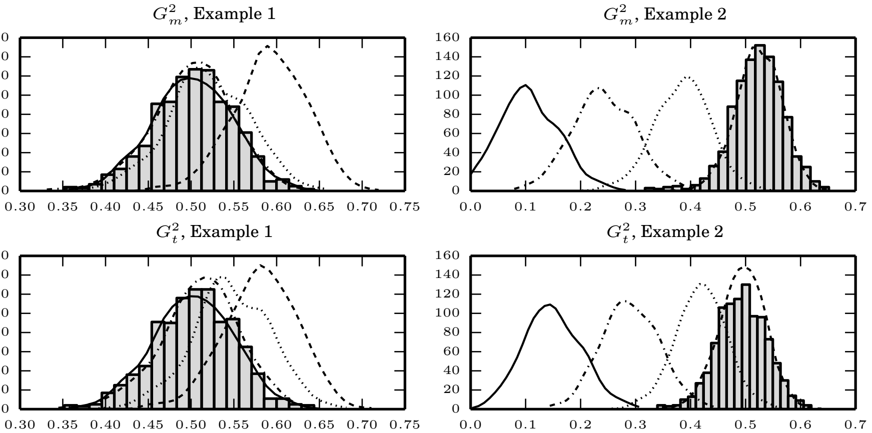

Fig. 1: Sampling distributions of G^2_m and G^2_t under the two models in Section 2·5 with $G^2_{Y|X} = 0.5$ for λ_0 = 0.5 (dashes), 1.5 (dots), 2.5 (dot-dash) and 3.5 (solid). The density function in each case is estimated by the histogram. The sampling distributions of G^2_m and G^2_t with the empirical Bayes selection of λ_0 are in gray shadow and overlaid on top of other density functions.

function, a larger λ_0 is preferred because it penalizes the number of slices more heavily and the resulting sampling distributions are less biased. On the other hand, for complicated relationships such as the trigonometric function, a smaller λ_0 is preferable because it allows more slices, which can help capture fluctuations in the functional relationship. The figure also shows that the empirical Bayes selection of λ_0 worked very well, leading to a proper choice of λ_0 for each simulated data set from both examples and resulting in the most accurate estimates of G^2. Additional simulation studies and consistency of the data-driven strategy are in the Supplementary Material.

3. SIMULATION STUDIES

3·1. *Power analysis*

Now we compare the power of different independence testing methods for various types of relationships. Here, we again fixed $\lambda_0 = 3$ for both G^2_m and G^2_t. Other methods we tested include the alternating conditional expectation (Breiman & Friedman, 1985), Genest's test (Genest & Rémillard, 2004), Pearson correlation, distance correlation (Szèkely et al., 2007), the method of Heller et al. (2016), the characteristic function method (Kankainen & Ushakov, 1998), Hoeffding's test (Hoeffding, 1948), the mutual information method (Kraskov et al., 2004) and two methods, MIC_e and TIC_e, based on the maximum information criterion (Reshef et al., 2011). We follow the procedure for computing the powers of different methods as described in previous studies of (D. Reshef and coauthors, arXiv:1505.02214) and a 2012 online note by N. Simon and R. Tibshirani.

For different functional relationships $f(X)$ and different values of noise levels σ^2, we simulated (X, Y) with the following model:

$$X \sim U(0,1), \quad Y = f(X) + \epsilon\sigma, \quad \epsilon \sim N(0,1).$$

where $\mathrm{var}\{f(X)\} = 1$. Thus $G^2_{Y|X} = (1+\sigma^2)^{-1}$ is a monotone function of the signal-to-noise ratio and it is of interest for us to observe how the performances of different methods deteriorate as the signal strength weakens for various functional relationships. We used permutations to generate the null distribution and to set the rejection region for all testing methods in all examples.

Figure 2 shows the power comparisons for eight functional relationships. We set the sample size $n = 225$ and performed 1,000 replications for each relationship and $G^2_{Y|X}$ value. We only plot Pearson correlation, distance correlation, method by Heller et al. (2016), TIC$_e$, G^2_m and G^2_t for a clear presentation. More simulations are in the Supplementary Material. For any method with tuning parameters, we chose the ones that resulted in the highest average power over all the examples. Due to computational concerns, we chose $K = 3$ for the method of Heller et al. (2016). It is seen that G^2_m and G^2_t performed robustly, always being among the most powerful methods, with G^2_t slightly more powerful than G^2_m in almost all examples. They outperformed other methods in cases such as the high frequency sine, triangle and piecewise constant functions, where piecewise linear approximation is more appropriate than other approaches. For monotonic examples such as linear and radical relationships, G^2_m and G^2_t had slightly lower powers than Pearson correlation, distance correlation and the method of Heller et al. (2016), but were still highly competitive.

We also studied the performances of these methods with different sample sizes, i.e. for n =50, 100 and 400, respectively, and found that G^2_m and G^2_t still showed high power regardless of n although their advantages were much less obvious when n is small. More details can be found in the Supplementary Material.

3·2. *Equitability*

Y. Reshef and coauthors (arXiv:1505.02212) gave two equivalent theoretical definitions of the equitability of a statistic that measures dependence. Intuitively, equitable statistics can be used to gauge the degree of dependence. They used $\Psi = \mathrm{cor}^2\{Y, f(X)\}$ to define the degree of dependence when the dependence of Y on X can be described by a functional relationship. When $\mathrm{var}(Y \mid X)$ is a constant, $\Psi \equiv G^2_{Y|X}$. For a perfectly equitable statistic, its sampling distribution should be almost identical for different relationships with the same Ψ.

We repeated the equitability study by Reshef et al. (2011). In Fig. 3, we plot the 95% confidence bands of G^2_m and G^2_t, compared with alternating conditional expectation, Pearson correlation, distance correlation and MIC$_e$, for the following relationships:

Example 3. $X \sim U(0,1), Y = X + \epsilon\sigma$ and $\epsilon \sim N(0,1)$.

Example 4. $X \sim U(0,1), Y = X + \epsilon\sigma$ and $\epsilon \sim N(0, e^{-|X|})$.

Example 5. $X \sim U(0,1), Y = \frac{X^2}{\surd 2} + \epsilon\sigma$ and $\epsilon \sim N(0,1)$.

Example 6. $X \sim U(0,1), Y = \frac{X^2}{\surd 2} + \epsilon\sigma$ and $\epsilon \sim N(0, e^{-|X|})$.

We choose different Ψ values with $n = 225$ and 1,000 replications for each case. The plots show that G^2_m and G^2_t increased along with Ψ for all relationships, as they should, and that the confidence bands obtained under different functional relationships have a similar size and location for the same Ψ. The confidence bands are also comparably narrow. The MIC$_e$ displayed good performances of equitability, though slightly worse than G^2_m and G^2_t, while other three statistics did poorly for non-monotone relationships. The alternating conditional expectation tended to have a wider confidence band for Example 5 and 6 than the aforementioned three methods, while Pearson correlation and distance correlation had non-overlapping confidence intervals for

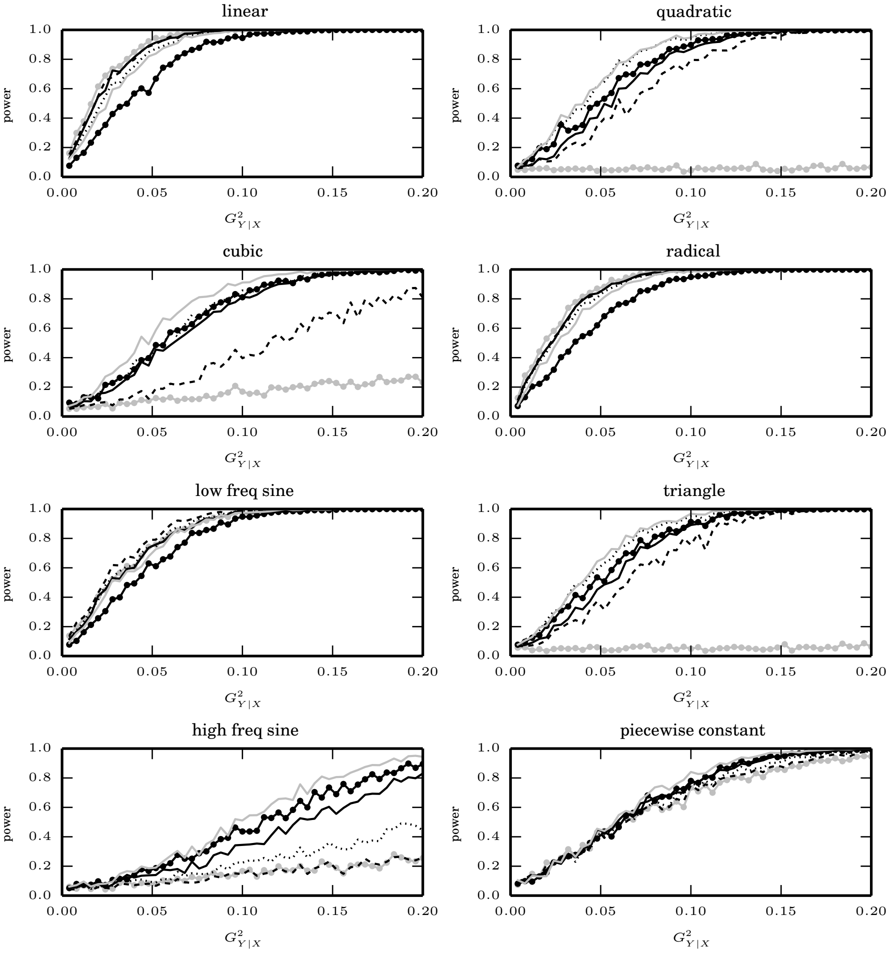

Fig. 2: The powers of G^2_m (black solid), G^2_t (grey solid), Pearson correlation (grey markers), distance correlation (black dashes), method of Heller et al. (2016) (black dots) and TIC$_e$ (black markers) for testing independence between X and Y when the underlying true functional relationships are linear, quadratic, cubic, radical, low freq sine, triangle, high freq sine and piecewise constant, respectively. The x-axis is $G^2_{Y|X}$, a monotone function of the signal-to-noise ratio, and the y-axis is the power. We chose $n = 225$ and performed 1,000 replications for each relationship and $G^2_{Y|X}$.

different relationships when Ψ is moderately large. In other words, Pearson correlation and distance correlation can yield drastically different values for two relationships with the same Ψ. This phenomenon is as expected since it is known that these two statistics do not perform well for non-monotone relationships.

An alternative strategy to study equitability uses a hypothesis testing framework, i.e., to test $\mathcal{H}_0 : \Psi = x_0$ against $\mathcal{H}_1 : \Psi = x_1$ $(x_1 > x_0)$ on a broad set of functional relationships using a statistic. The more powerful a test statistic for this testing problem with all types of relationships, the better its equitability. For each aforementioned method, we performed right-tailed tests with

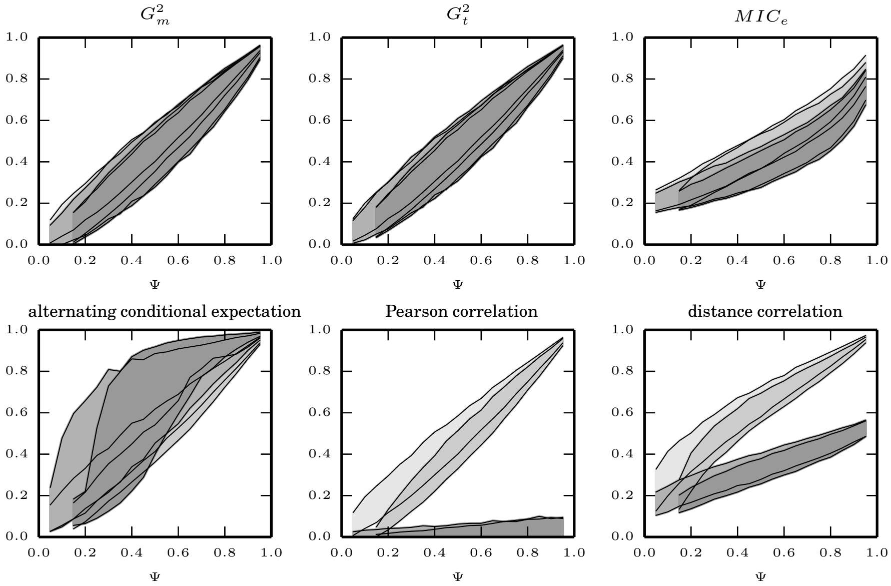

Fig. 3: The plots from the top left to the bottom right are the 95% confidence bands for G_m^2, G_t^2, $\textsc{mic}_e$, alternating conditional expectation, Pearson correlation and distance correlation, respectively. We chose $n = 225$ and performed 1,000 replications for each relationship and each Ψ value for the four examples in Section 3·2. The fill is the lightest for Example 3 and darkest for Example 6. Ψ is a monotone function of the signal-to-noise ratio when the error variance is constant, and the y-axis shows the values of the corresponding statistic, each estimating its own population mean, which may or may not be Ψ.

the type-I error fixed at $\alpha = 0.05$ and different combinations of (x_0, x_1) $(0 < x_0 < x_1 < 1)$. Given a fixed sample size, a perfectly equitable statistic should yield the same power for all kinds of relationships so that it is able to reflect the degree of dependency by a single value regardless of the type of relationship. In reality, most statistics can perform well only for a small class of relationships. In Fig. 4, we use a heat map to demonstrate the average power of a test statistic with different pairs of (x_0, x_1) $(0 < x_0 < x_1 < 1)$. Each dot in the plot represents the average power of a testing method over a class of functional relationships; the darker the color is, the higher the power. We used the same set of functional relationships as in N. Reshef and coauthors (arXiv:1505.02214) and carried out the testing for $(x_0, x_1) = (i/50, j/50)$ $(i < j = 1, \ldots, 49)$. We set the sample size as $n = 225$ and conducted 1,000 replications for each relationship and each (x_0, x_1) $(0 < x_0 < x_1 < 1)$. For any method with a tuning parameter, we chose parameters that resulted in the greatest average power. We observed that G_m^2, G_t^2 and $\textsc{mic}_e$ had the best equitability, followed by alternating conditional expectation and $\textsc{tic}_e$. The average powers for G_m^2, G_t^2 and $\textsc{mic}_e$ over the entire range of (x_0, x_1) $(0 < x_0 < x_1 < 1)$ were all 0.6, although G_m^2 and G_t^2 were slightly better for larger x_0's. Besides, with our empirical Bayes method for selecting λ_0, the equitability of G_m^2 and G_t^2 can be further improved. In comparison, all the remaining methods were not as equitable.

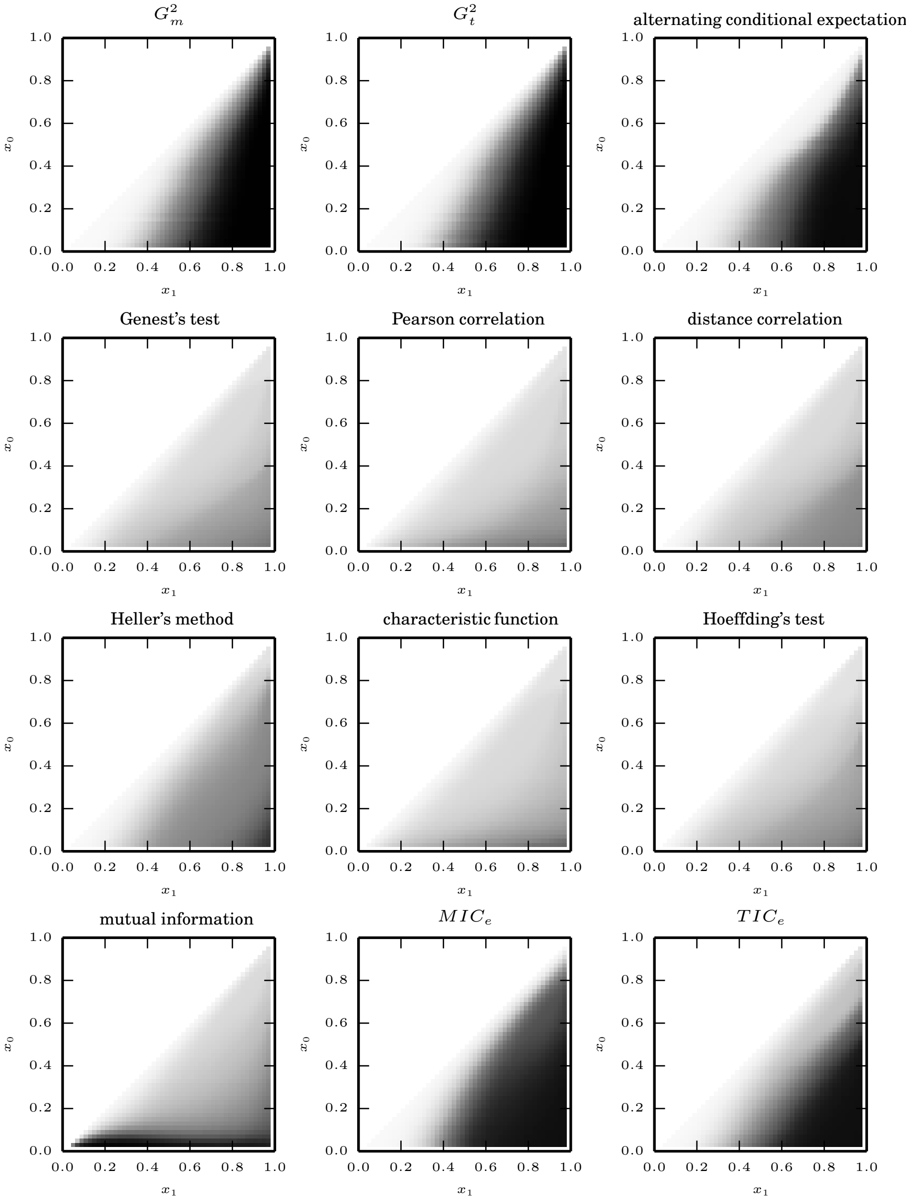

Fig. 4: Heat map plot for comparing equitability of different methods. From the top left to the bottom right: G_m^2, G_t^2, alternating conditional expectation, Genest's test, Pearson correlation, distance correlation, the method of Heller et al. (2016), characteristic function method, Hoeffding's test, mutual information, MIC_e and TIC_e. The value corresponding to (x_1, x_0) $(0 < x_0 < x_1 < 1)$ is the power of the method for testing the hypothesis: $\mathcal{H}_0 : \Psi = x_0$ against $\mathcal{H}_1 : \Psi = x_1$, averaging over a class of functions. The darker a dot, the higher the average power of the corresponding test. We chose sample size $n = 225$ and performed 1,000 replications for each relationship and (x_0, x_1) $(0 < x_0 < x_1 < 1)$.

4. Discussion

The G-squared can be viewed as a direct generalization of the R-squared. While maintaining the same interpretability as the R-squared, the G-squared is also a powerful and equitable measure of dependence for general relationships. Instead of resorting to curve-fitting methods for estimating the underlying relationship and the G-squared, we employed the more flexible piecewise linear approximations with penalty and dynamic programming algorithms. Although we only consider piecewise linear functions, one can potentially approximate a relationship between two variables with piecewise polynomials or other flexible basis functions, with perhaps additional penalty terms to control the complexity. Furthermore, it is a worthwhile effort to generalize the slicing idea for testing dependence between two multivariate random variables.

Acknowledgment

We are grateful to the two referees for helpful comments and suggestions. This research was supported in part by grants from the U.S. National Science Foundation and National Institutes of Health. We thank Ashley Wang for her proofreading of the paper. The views expressed herein are the authors alone and are not necessarily the views of Two Sigma Investments, Limited Partnership or any of its affiliates.

Supplementary material

Further material available at *Biometrika* online includes proofs of theorems, software implementation details, discussions on segmented regression and more simulation results.

References

Blyth, S. (1994). Local divergence and association. *Biometrika* **91**, 579–584.

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

Doksum, K., Blyth, S., Bradlow, E., Meng, X. & Zhao, H. (1994). Correlation curves as local measures of variance explained by regression. *J. Am. Statist. Assoc* **89**, 571–582.

Genest, C. & Rémillard, B. (2004). Test of independence and randomness based on the empirical copula process. *Test* **13**, 335–369.

Gretton, A., Gousquet, O., Smola, A., & Schlkopf, B. (2005). Measuring Statistical Dependence with Hilbert-Schmidt Norms. *Algorithmic Learning Theory* **3734**, 63–77.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schlkopf, B. & Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–773.

Heller, R., Heller, Y., Kaufman, S., Brill, B. & Gorfine, M. (2016). Consistent distribution-free K-sample and independence tests for univariate random variables. *J. Mach. Learn. Res.* **17**, 1–54.

Hoeffding, W. (1948). A non-parametric test of independence. *Ann. Stat.* **19**, 546–557.

Hušková, M. & Meintanis, S. (2008). Testing procedures based on the empirical characteristic functions I: Goodness-of-fit, testing for symmetry and independence. *Tatra Mt. Math. Publ* **39**, 225–233.

Jiang, B., Ye, C. & Liu, J. S. (2015). A consistent modification of a test for independence based on the empirical characteristic function. *J. Am. Statist. Assoc* **101**, 642–653.

Kankainen, A. & Ushakov, N. G. (1998). A consistent modification of a test for independence based on the empirical characteristic function. *J. Math. Sci.* **89**, 1486–1494.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc* **90**, 773-795.

Kraskov, A., Stogbauer, H. & Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E* **69.6**, 066138.

Oosterbaan, R. J. & Ritzema, H. P. (2006). *Drainage principles and applications*, Ritzema, H. P., 217–220 Wageningen, Netherlands: International Institute for Land Reclamation and Improvement. ISBN 90-70754-3-39.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. & Sabeti, P. S. (2011). Detecting Novel Associations in Large Data Sets. *Science* **334**, 1518–1524.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.

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

SZÈKELY, G. J., RIZZO, M. L. & BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Stat.* **12**, 2769–2794.

SZÉEKELY, G. J. & RIZZO, M. L. (2009). Brownian distance correlation. *Ann. Appl. Stat.* **12**, 1236–1265.