Biometrika (2017), **104**, 1, *pp.* 129–139
Printed in Great Britain
doi: 10.1093/biomet/asw071

Generalized R-squared for detecting dependence

BY X. WANG

Department of Statistics, Harvard University, One Oxford Street, Cambridge, Massachusetts 02138, U.S.A
xufeiwang@fas.harvard.edu

B. JIANG

Two Sigma Investments, Limited Partnership, 100 Avenue of the Americas, 16th Floor, New York, New York 10013, U.S.A
bojiang83@gmail.com

AND J. S. LIU

Department of Statistics, Harvard University, One Oxford Street, Cambridge, Massachusetts 02138, U.S.A
jliu@stat.harvard.edu

SUMMARY

Detecting dependence between two random variables is a fundamental problem. Although the Pearson correlation coefficient is effective for capturing linear dependence, 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 statistic 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 estimators 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 between 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 statistic 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 nonmonotone 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 optimal transformations of the data, $f(X)$ and $g(Y)$, such that their correlation is maximized. The implementation of this method has limitations, because it is infeasible to search through 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. Kraskov et al. (2004) proposed a method that involves estimating the entropy of X, Y and (X, Y) separately. The method was claimed to be numerically exact for independent cases, and effective 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) for solving the two-sample test problem appeared separately in the statistics and machine learning literatures, 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 discretization of the random variables. Known as grid-based methods, they are primarily designed to test for independence between univariate random variables. Reshef et al. (2011) introduced the maximum information coefficient, which focuses on the generality and equitability of a dependence statistic; two more powerful estimators for this quantity were suggested by Reshef et al. (arXiv:1505.02213). Equitability requires that the same value of the statistic imply the same amount of dependence regardless of the type of the underlying relationship, but it is not a well-defined mathematical concept. We show in the Supplementary Material that the equitability of G-squared is superior to all other independence testing statistics for a wide range of functional relationships. Heller et al. (2016) proposed a grid-based method which utilizes the χ^2 statistic to test independence and is a distribution-free test. Blyth (1994) and Doksum et al. (1994) discussed using 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 of the relationship; furthermore, it cannot deal with heteroscedastic noise.

The G^2 statistic 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. It is a function of both the conditional mean and the conditional variance of one variable given the other, so it is 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 R^2 when the true underlying relationship is linear. Thus, it is reasonable that G^2 is almost identical to R^2 for linear relationships. Efficient estimates of G^2 can be computed quickly using a dynamic programming method, whereas the methods of Reshef et al. (2011) and Heller et al. (2016) consider grids on two variables simultaneously and hence require longer computational times. We will also show that, in terms of power, G^2 is one of the best statistics for independence testing when considering a wide range of functional relationships.

2. MEASURING DEPENDENCE WITH G-SQUARED

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

The R-squared statistic 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.,

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

where $L_0(\hat{\theta}_0)$ and $L(\hat{\theta})$ 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 $Y = f(X) + \epsilon\sigma_X$, with $\epsilon \sim N(0, 1)$ 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

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

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 consecutive 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

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

where $\hat{\nu}^2$ is the overall sample variance of Y and $\hat{\sigma}_h^2$ 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 the Pearson correlation coefficient, we perform the same transformation on $\textsc{lr}$. 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 $n \to \infty$,

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

When $K = 1$, the relationship degenerates to a simple linear relation 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 R-squared:

$$G^2_{Y|X} = 1 - \exp\big[E\{\log \operatorname{var}(Y \mid X)\} - \log \operatorname{var}(Y)\big],$$

in which we require that $\operatorname{var}(Y) < \infty$. Evidently, $G^2_{Y|X}$ lies between 0 and 1, and is equal to zero if and only if both $E(Y \mid X)$ and $\operatorname{var}(Y \mid X)$ are constant. The definition of $G^2_{Y|X}$ is closely related to the R-squared defined by segmented regression (Oosterbaan & Ritzema, 2006), discussed in the Supplementary Material. We symmetrize $G^2_{Y|X}$ to arrive at the following quantity as the definition of the G-squared statistic:

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

provided $\operatorname{var}(X) + \operatorname{var}(Y) < \infty$. Thus, $G^2 = 0$ if and only if $E(X \mid Y)$, $E(Y \mid X)$, $\operatorname{var}(Y \mid X)$ and $\operatorname{var}(X \mid Y)$ 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 interchanging X and Y. When $Y = f(X) + \epsilon\sigma_X$ with $\epsilon \sim N(0,1)$ 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 that 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, i.e., $S(x_i) = h$ 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 loglikelihood ratios both over unknown parameters and over all possible slicing schemes, we restrict the minimum size of each slice to $m_S \geqslant m$ and maximize the loglikelihood ratio with a penalty on the number of slices. For simplicity, let $m = \lceil n^{1/2} \rceil$. Thus, we focus on the penalized loglikelihood 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 incurred for 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 that 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

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

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

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

Hence, the overall G-squared can be estimated as

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

By definition, $G^2_{\mathrm{m}}(\lambda_0)$ lies between 0 and 1, and $G^2_{\mathrm{m}}(\lambda_0) = R^2$ when the optimal slicing schemes for both directions have only one slice. Later, we will show that when X and Y follow a bivariate normal distribution, $G^2_{\mathrm{m}}(\lambda_0) = R^2$ 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 modelling assumption. On the other hand, however, the Bayesian formalism may guide us to a desirable integration strategy for the slicing scheme. We therefore 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

$$\mathrm{BF} = \sum_{S:\, m_s \geqslant m} \omega_S \times \frac{p_S(y_1, \ldots, y_n)}{p_0(y_1, \ldots, y_n)}, \tag{3}$$

where m_S is the minimum size of all the slices of S. 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

$$p(y_1, \ldots, y_n) \approx \mathrm{L} \exp\{-k(\log n - \log 2\pi)/2\}, \tag{4}$$

where L is the maximized likelihood. In our set-up, the number of parameters k for the null model is 2, and for an alternative model with a slicing scheme S it is $3|S|$. Inserting expression (4) into both the numerator and the denominator of (3), we obtain

$$\mathrm{BF} \approx \sum_{S:\, m_s \geqslant m} \omega_S \, \mathrm{LR}_S \exp\{-(3|S| - 2)(\log n - \log 2\pi)/2\}. \tag{5}$$

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

$$\mathrm{BF}(\lambda_0) = \left[\sum_{S:\, m_S \geqslant m} n^{-\{\lambda_0(|S|-1)\}/2} \right]^{-1} \sum_{S:\, m_S \geqslant m} \left(\frac{2\pi}{n} \right)^{(3|S|-2)/2} \exp\left\{ \frac{n}{2} D(Y \mid S, \lambda_0) \right\}. \tag{6}$$

As we will discuss in § 2·5, $\mathrm{BF}(\lambda_0)$ can serve as a marginal likelihood function for λ_0 and be used to find an optimal λ_0 suitable for a particular dataset. This quantity also looks like an average version of G^2_{m}, but with an additional penalty. Since $\mathrm{BF}(\lambda_0)$ can take values below 1, its transformation $1 - \mathrm{BF}(\lambda_0)^{-2/n}$, as in the case where we derived R^2 via the likelihood ratio test, can take negative values, especially when X and Y are independent. 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:

$$\mathrm{BF}^*(\lambda_0) = \left[\sum_{S:\, m_S \geqslant m} n^{-\{\lambda_0(|S|-1)\}/2} \right]^{-1} \sum_{S:\, m_S \geqslant m} \exp\left\{ \frac{n}{2} D(Y \mid S, \lambda_0) \right\}.$$

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

$$G_{\mathrm{t}}^2(Y \mid X, \lambda_0) = 1 - \mathrm{BF}^*(\lambda_0)^{-2/n},$$

which we call the total G-squared, and define

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

We show later that $G_{\mathrm{m}}^2(\lambda_0)$ and $G_{\mathrm{t}}^2(\lambda_0)$ are both consistent estimators of G^2.

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

In order to show that $G_{\mathrm{m}}^2(\lambda_0)$ and $G_{\mathrm{t}}^2(\lambda_0)$ converge to G^2 as the sample size goes to infinity, we introduce the notation $\mu_X(y) = E(X \mid Y = y)$, $\mu_Y(x) = E(Y \mid X = x)$, $\nu_X^2(y) = \mathrm{var}(X \mid Y = y)$ and $\nu_Y^2(x) = \mathrm{var}(Y \mid X = x)$, and assume 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

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

almost surely.

With these preparations, we can state our main results.

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

$$G_{\mathrm{m}}^2(Y \mid X, \lambda_0) \to G_{Y|X}^2, \quad G_{\mathrm{t}}^2(Y \mid X, \lambda_0) \to G_{Y|X}^2$$

almost surely as $n \to \infty$. Thus, $G_{\mathrm{m}}^2(\lambda_0)$ and $G_{\mathrm{t}}^2(\lambda_0)$ are consistent estimators of G^2.

A proof of the theorem and numerical studies of the estimators' consistency are provided in the Supplementary Material. It is expected that $G_{\mathrm{m}}^2(\lambda_0)$ should converge to G^2 because of the way it is constructed. It is surprising that $G_{\mathrm{t}}^2(\lambda_0)$ 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 $G_{\mathrm{m}}^2(\lambda_0)$ and $G_{\mathrm{t}}^2(\lambda_0)$ for testing whether X and Y are independent. The null distributions of the two statistics depend on the marginal distributions of X and Y, and can be generated empirically using permutation. One can also perform a quantile-based transformation on X and Y so that their marginal distributions become 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 statistic when λ_0 is large enough.

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

$$\mathrm{pr}\{G^2_{\mathrm{m}}(\lambda_0) = R^2\} > 1 - 3n^{-\lambda_0/3+5}.$$

So, for $\lambda_0 > 18$ and $n \to \infty$, we have $G^2_{\mathrm{m}}(\lambda_0) = R^2$ 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 $G^2_{\mathrm{m}}(\lambda_0)$ to be very close to R^2 in the bivariate normal setting.

2·4. *Dynamic programming algorithm for computing G^2_{m} and G^2_{t}*

The brute force calculation of either G^2_{m} or G^2_{t} 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 only $O(n^2)$. The algorithms for computing $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ and $G^2_{\mathrm{t}}(Y \mid X, \lambda_0)$ are roughly the same except for one operation, namely 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 x values sorted 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 $\alpha = \exp(\lambda)$. 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 $K_i = \{1\} \cup \{k : k = m+1, \ldots, i-m+1\}$, $l_{k:i} = -(i-k)\log(\hat{\sigma}^2_{k:i})/2$ and $L_{k:i} = \exp\{l_{k:i}\}$, with $\hat{\sigma}^2_{k:i}$ being the residual variance of regressing y on x for observations (x_j, y_j) $(j = k, \ldots, i)$.

Step 3. The final result is

$$G^2_{\mathrm{m}} = 1 - \exp\{M_n - \lambda\}, \quad G^2_{\mathrm{t}} = 1 - (T_n/B_n)^{-2/n}.$$

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 $G^2_{\mathrm{m}}(X \mid Y, \lambda_0)$ and $G^2_{\mathrm{t}}(X \mid Y, \lambda_0)$ similarly to get $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$. 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 take $\lambda_0 = 3$ for G^2_{m} and G^2_{t} 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 therefore propose a data-driven strategy for choosing λ_0 adaptively. The quantity $\mathrm{BF}(\lambda_0)$ in (6) can be viewed as an approximation to $\mathrm{pr}(y_1, \ldots, y_n \mid \lambda_0)$ up to a normalizing constant. Hence we can use the

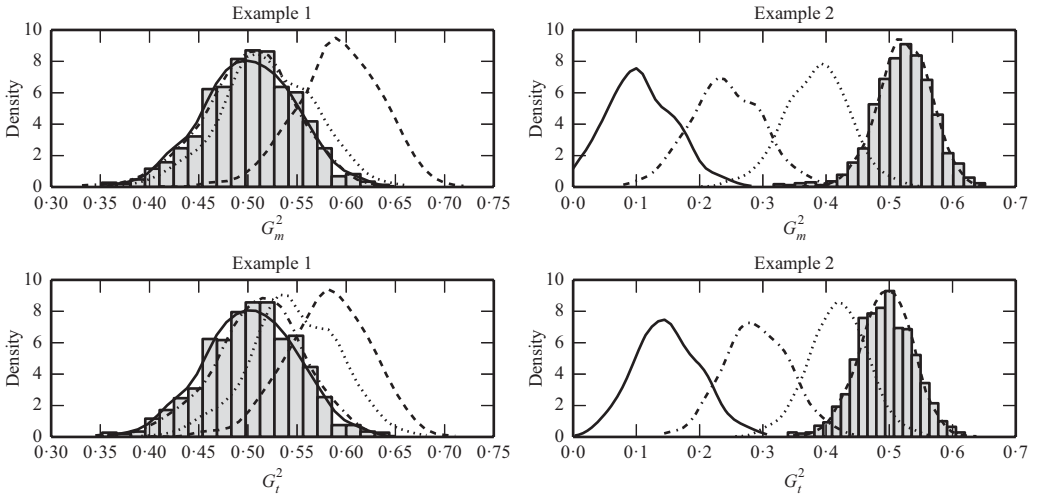

Fig. 1. Sampling distributions of G^2_{m} and G^2_{t} under the two models described in § 2·5 with $G^2_{Y|X} = 0{\cdot}5$ for $\lambda_0 = 0{\cdot}5$ (dashed), 1·5 (dotted), 2·5 (dot-dash) and 3·5 (solid). The density function in each case is estimated by the histogram. The sampling distributions of G^2_{m} and G^2_{t} with the empirical Bayes selection of λ_0 are shaded grey and overlaid on top of the other density functions.

maximum likelihood principle to choose the λ_0 that maximizes $\mathrm{BF}(\lambda_0)$. We then use the chosen λ_0 to compute G^2_{m} and G^2_{t} as estimators of G^2. In practice, we evaluate $\mathrm{BF}(\lambda_0)$ for a finite set of λ_0 values, such as $\{0{\cdot}5,\ 1,\ 1{\cdot}5,\ 2,\ 2{\cdot}5,\ 3,\ 3{\cdot}5,\ 4\}$, and pick the λ_0 value that maximizes $\mathrm{BF}(\lambda_0)$; $\mathrm{BF}(\lambda_0)$ can be computed efficiently via a dynamic programming algorithm similar to that described in § 2·4. As an illustration, we consider the sampling distributions of $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$ with $\lambda_0 = 0{\cdot}5,\ 1{\cdot}5,\ 2{\cdot}5$ and $3{\cdot}5$ for the following two scenarios:

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

Example 2. $X \sim N(0,1)$ and $Y = \sin(4\pi x)/0{\cdot}7 + \sigma\epsilon$ with $\epsilon \sim N(0,1)$.

We simulated $n = 225$ data points. For each model, we set $\sigma = 1$ so that $G^2_{Y|X} = 0{\cdot}5$ and performed 1000 replications. Figure 1 shows histograms of $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$ with different λ_0 values. The results demonstrate that for relationships which can be approximated well by a linear function, a larger λ_0 is preferred because it penalizes the number of slices more heavily, so that the resulting sampling distributions are less biased. On the other hand, for complicated relationships such as trigonometric functions, a smaller λ_0 is preferable because it allows more slices, which can help to 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 dataset from both examples and resulting in the most accurate estimates of G^2. Additional simulation studies and discussion of the consistency of the data-driven strategy can be found in the Supplementary Material.

3. POWER ANALYSIS

Next, we compare the power of different independence testing methods for various 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 Reshef et al. (arXiv:1505.02214) and a 2012 online note by N. Simon and R. J. Tibshirani.

For different functional relationships $f(X)$ and different noise levels σ^2, we let

$$X \sim \text{Un}(0,1), \quad Y = f(X) + \epsilon\sigma, \quad \epsilon \sim N(0,1),$$

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

Figure 2 shows power comparisons for eight functional relationships. We set the sample size to $n = 225$ and performed 1000 replications for each relationship and each $G^2_{Y|X}$ value. For the sake of clarity, here we plot only Pearson correlation, distance correlation, the method of Heller et al. (2016), TIC_e, G^2_{m} and G^2_{t}. For any method with tuning parameters, we chose the parameter values 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 can be seen that G^2_{m} and G^2_{t} performed robustly, and were always among the most powerful methods, with G^2_{t} being slightly more powerful than G^2_{m} in nearly all the examples. They outperformed the 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 power 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 for $n = 50$, 100 and 400, and found that G^2_{m} and G^2_{t} still had high power regardless of n, although their advantages were much less obvious when n was small. More details can be found in the Supplementary Material.

4. Discussion

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

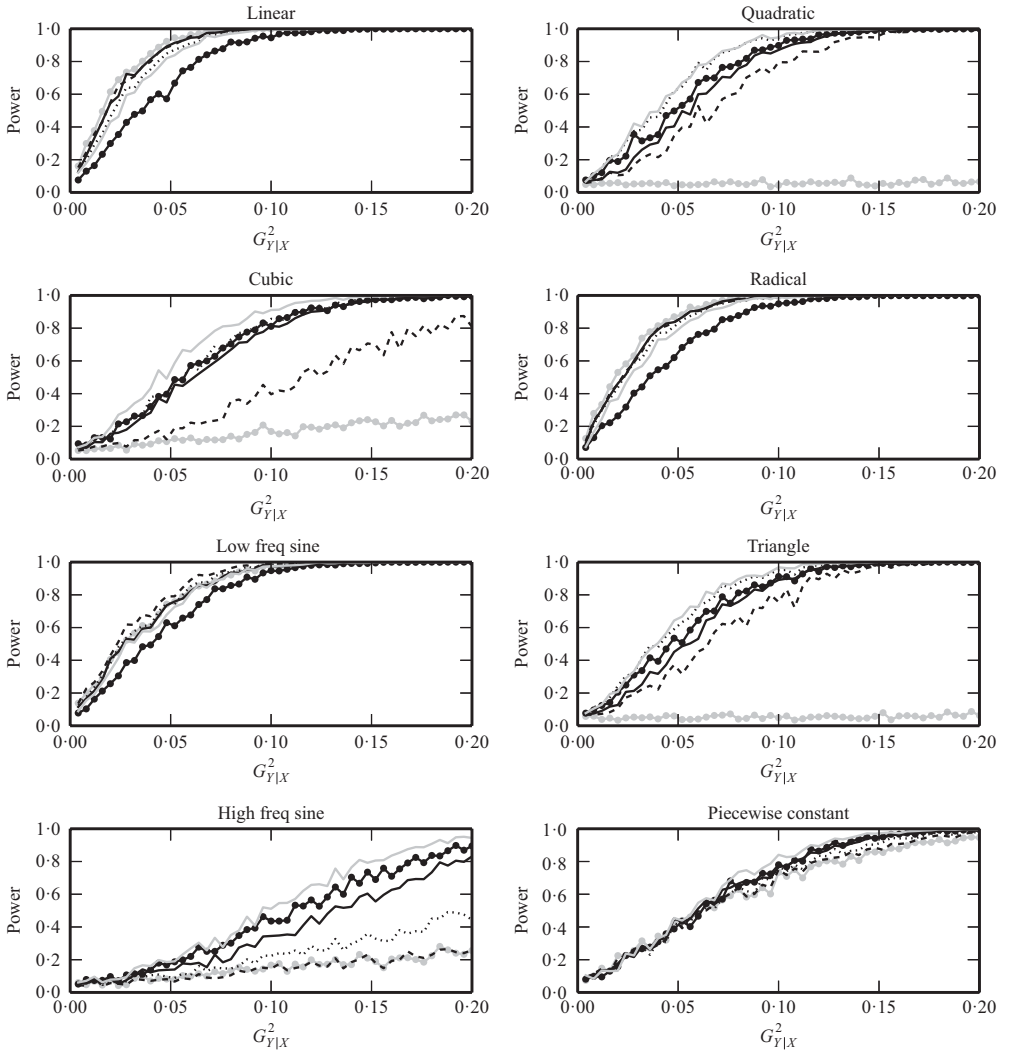

Fig. 2. The powers of G^2_{m} (black solid), G^2_{t} (grey solid), Pearson correlation (grey circles), distance correlation (black dashed), the method of Heller et al. (2016) (black dotted) and TIC_e (black circles) for testing independence between X and Y when the underlying true functional relationships are linear, quadratic, cubic, radical, low-frequency sine, triangle, high-frequency sine, and piecewise constant. The horizontal axis represents $G^2_{Y|X}$, a monotone function of the signal-to-noise ratio, and the vertical axis is the power. We chose $n = 225$ and performed 1000 replications for each relationship and each $G^2_{Y|X}$ value.

ACKNOWLEDGEMENT

We are grateful to the two referees for helpful comments and suggestions. This research was supported in part by 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

Supplementary material available at *Biometrika* online includes proofs of the theorems, software implementation details, discussions on segmented regression, a study of equitability, and more simulation results.

References

Blyth, S. (1994). Local divergence and association. *Biometrika* **91**, 579–84.

Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Am. Statist. Assoc.* **80**, 580–98.

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**, 571–82.

Genest, C. & Rémillard, B. (2004). Test of independence and randomness based on the empirical copula process. *Test* **13**, 335–69.

Gretton, A., Gousquet, O., Smola, A. & Schlkopf, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. *Algor. Learn. 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**, 723–73.

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. Statist.* **19**, 546–57.

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**, 225–33.

Kankainen, A. & Ushakov, N. G. (1998). A consistent modification of a test for independence based on the empirical characteristic function. *J. Math. Sci.* **89**, 1486–94.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.

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*. Wageningen, Netherlands: International Institute for Land Reclamation and Improvement, pp. 217–20.

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**, 1518–24.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. & Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41**, 2263–91.

Széekely, G. J. & Rizzo, M. L. (2009). Brownian distance correlation. *Ann. Appl. Statist.* **12**, 1236–65.

Szèkely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **12**, 2769–94.

[*Received on* 8 *February* 2016. *Editorial decision on* 26 *November* 2016]