Biometrika (2016), **xx**, x, *pp*. 1–16
Printed in Great Britain

Supplementary material for ‘Generalized R-squared for detecting dependence’

BY X. WANG
Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A
xufeiwang@fas.harvard.edu

B. JIANG
Two Sigma Investments, Limited Partnership, New York, New York 10013, U.S.A
bojiang83@gmail.com

AND J. S. LIU
Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A
jliu@stat.harvard.edu

SUMMARY

This document contains Supplementary Material on the following topics: (1) software implementation; (2) relationship between G-squared and segmented regression; (3) equitability study; (4) more simulations; (5) proof of the consistency of $G^2_{\rm m}$ and $G^2_{\rm t}$ for estimating the G-squared; (6) proof of the equivalence between $G^2_{\rm m}$ and R^2 in the bivariate normal case.

1. SOFTWARE IMPLEMENTATION

We provide R implementation to estimate $G^2_{\rm m}$ and $G^2_{\rm t}$ discussed in the main paper. The R package is available at `http://www.people.fas.harvard.edu/~junliu/Gs`. We studied the computing time for different methods with sample sizes $n = 50,\ 100,\ 225$ and 500. For each n we simulated 1,000 observations and recorded the computing time for every method; the average time is shown in Fig 1. The computing time for $G^2_{\rm t}$ was twice as much as the computing time for $G^2_{\rm m}$ due to the normalizing constant. This time can be further reduced by tabulating the normalizing constant for pairs of (n, λ_0). $G^2_{\rm m}$ and $G^2_{\rm t}$ were more time efficient compared with distance correlation, the method of Heller et al. (2016), and MIC$_e$.

2. SEGMENTED REGRESSION

The R-squared for segmented regression with predictor X and response Y is

$$R^2 = 1 - \frac{\sum_{h=1}^{K} n_h \widehat{\sigma}^2_h}{n\widehat{\nu}^2},$$

where $\widehat{\nu}^2$ is the sample variance of Y, n_h and $\widehat{\sigma}^2_h$ are sample size and residual variance of Y after regressing on X in segment h $(h = 1, \ldots, K)$. R^2 can be viewed as an estimator of

$$R^2_{Y|X} = 1 - E\{\mathrm{var}(Y \mid X)\}/\mathrm{var}(Y);$$

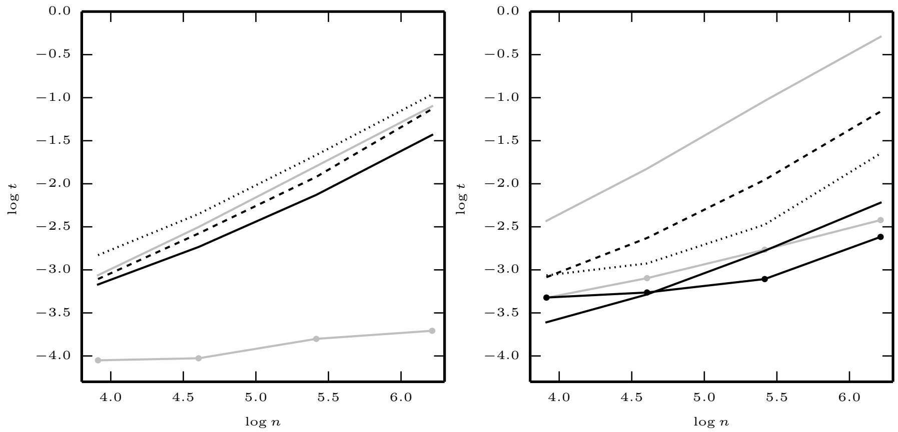

Fig. 1. The left figure shows the average computing time of $G^2_{\rm m}$ (black solid), $G^2_{\rm t}$ (grey solid), Pearson correlation (grey markers), distance correlation (black dashes) and the method of Heller et al. (2016) (black dots) for 1,000 simulations with sample sizes n =50, 100, 225 and 500; the right figure shows the average computing time of mutual information (black solid), MIC$_e$ (grey solid), alternating conditional expectation (grey markers), characteristic function (black dashes), Genest's test (black dots) and Hoeffding's test (black markers). The x-axis is the logarithm of n with base 10 and the y-axis is the logarithm of the computing time in seconds with base 10.

it is zero if and only if $E(Y \mid X)$ is a constant. $G^2_{Y|X}$ equals

$$1 - \exp\left[E\{\log \mathrm{var}(Y \mid X)\} - \log \mathrm{var}(Y)\right];$$

it is zero if and only if both $E(Y \mid X)$ and $\mathrm{var}(Y \mid X)$ are constant. $G^2_{Y|X}$ equals $R^2_{Y|X}$ when $\mathrm{var}(Y \mid X)$ is a constant, but $G^2_{Y|X}$ is more general than $R^2_{Y|X}$ since it can capture heteroscedastic effects.

Given a fixed number of segments K, computing $R^2_{Y|X}$ with the optimal segmentation is more computationally intensive than computing $G^2_{\rm m}$ and $G^2_{\rm t}$, especially when K is large. When K is unknown, we can apply the same dynamic programming algorithm for $G^2_{\rm m}$ or $G^2_{\rm t}$ and fit a penalized version of the segmented regression to avoid over-fitting. If we also require that the fitted curve be continuous, no exact numerical solution is available; we could potentially design a Markov chain Monte Carlo algorithm under a Bayesian framework.

3. EQUITABILITY

Reshef et al. (arXiv:1505.02212) gave two equivalent definitions for 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, we have $\Psi \equiv G^2_{Y|X}$. For a perfectly equitable statistic, its sampling distribution should be almost identical for different relationships with the same Ψ. But the existence of such a statistic for any well-defined large class of functional relationships remains unclear.

We repeated the equitability study by Reshef et al. (2011). Figure 2 shows the 95% confidence bands for $G^2_{\rm m}$ and $G^2_{\rm t}$, compared with alternating conditional expectation, Pearson correlation, distance correlation, and MIC$_e$ for $X \sim N(0,1)$ and

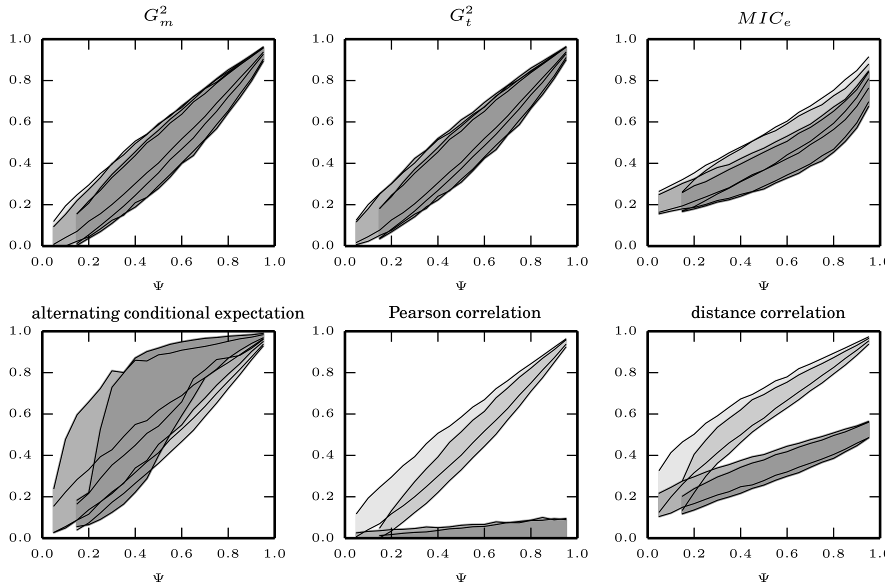

Fig. 2. The plots from the top left to the bottom right are the 95% confidence bands of Φ for the 6 indicated methods. We chose $n = 225$ and performed 1,000 replications for each relationship and each value of Ψ for Example 1–4. The shadow is the lightest for Example 1 and darkest for Example 4. Ψ is a monotone function of the signal-to-noise ratio when the error variance is constant. The y-axis shows the values of the corresponding statistic, each estimating its own population mean, which may or may not be Ψ.

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

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

Example 3. $Y = X^2/\sqrt{2} + \epsilon\sigma$ and $\epsilon \sim N(0, 1)$;

Example 4. $Y = X^2/\sqrt{2} + \epsilon\sigma$ and $\epsilon \sim N(0, e^{-|X|})$.

We chose different values of Ψ with $n = 225$ and conducted 1,000 replications for each case. The plots show that G_{m}^2 and G_{t}^2 increased along with Ψ for all relationships, as expected, and that the confidence bands obtained under different functional relationships had a similar size and location for the same Ψ. The confidence bands were also comparably narrow. The MIC$_e$ displayed good equitability, though slightly worse than G_{m}^2 and G_{t}^2, while the other three statistics did poorly for non-monotone relationships. The alternating conditional expectation tended to have a wider confidence band for Example 3 and 4 than the other methods, while the Pearson correlation and distance correlation had non-overlapping confidence intervals for different relationships when Ψ is moderately large. In other words, the Pearson correlation and distance correlation can yield drastically different values for two relationships with the same Ψ. This phenomenon was as expected, since it is known that these two statistics do not perform well for non-monotone relationships.

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

Table 1. *Functional relationships for equitability study*

Relation Name	Function
line	x
quadratic	$(x-1/2)^2$
cubic	$4(2.4x-1.3)^3+(2.4x-1.3)^2-4(2.4x-1.3)$
exponential (10^x)	10^{10x}
exponential (2^x)	2^{2x}
L-shaped	$(x/99)I_{x\leq 0/99}+1I_{x>0.99}$
lopsided L-shaped	$200xI_{x\leq 0.005}+(-198x+19.9)I_{0.005<x\leq 0.01}+(-x/99+1/99)I_{x>0.1}$
spike	$20I_{x\leq 0.05}+(-18x+1.9)I_{0.05<x\leq 0.1}+(-x/9+1/9)I_{x>0.1}$
sigmoid	$\{50(x-0.5)+0.5\}I_{0.49<x\leq 0.51}+1I_{x>0.51}$
linear + high freq periodic	$0.1\sin\{10.6(2x-1)\}+1.1(2x-1)$
linear + high freq periodic 2	$0.2\sin\{10.6(2x-1)\}+1.1(2x-1)$
linear + low freq periodic	$0.2\sin\{4(2x-1)\}+1.1(2x-1)$
linear + medium freq periodic	$\sin(10\pi x)+x$
high freq sine	$\sin(8\pi x)$
non-Fourier freq sine	$\sin(9\pi x)$
very high freq sine	$\sin(16\pi x)$
varying freq sine	$\sin\{6\pi x(1+x)\}$
high freq cosine	$\cos(14\pi x)$
non-Fourier freq cosine	$\cos(7\pi x)$
varying freq cosine	$\sin\{5\pi x(1+x)\}$

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. We use a heat map to demonstrate the average power of a test statistic with different pairs of (x_0, x_1) $(0 < x_0 < x_1 < 1)$ in Fig. 3. Each dot in the plot represents the average power of a test statistic over a class of functional relationships; the darker the color, the higher the power. We simulated (X, Y) with the following model

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

The twenty chosen functional relationships, which were inspired by the functional relationships in (Reshef et al., arXiv:1505.02214), are shown in Table 1. We carried out the testing for $(x_0, x_1) = (i/50, j/50)$ $(i < j = 1, \ldots, 49)$. We set $n = 225$ and conducted 1,000 replications for each relationship and each pair of (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^2_{m}, G^2_{t} and MIC_e had the best equitability, followed by alternating conditional expectation and TIC_e. The average powers for G^2_{m}, G^2_{t} and MIC_e over the entire range of (x_0, x_1) $(0 < x_0 < x_1 < 1)$ were all 0.6, although G^2_{m} and G^2_{t} were slightly better for larger x_0's. Besides, using our empirical Bayes method to select λ_0, the equitability of G^2_{m} and G^2_{t} can be further improved. In comparison, all the remaining methods were not as equitable.

4. SIMULATIONS

4·1. *Consistency of G^2_{m} and G^2_{t}*

For a general relationship, the true value of G^2 is nontrivial to compute. However, we can calculate $G^2_{Y|X}$ for some special examples and evaluate the sum of squared errors of the estimators.

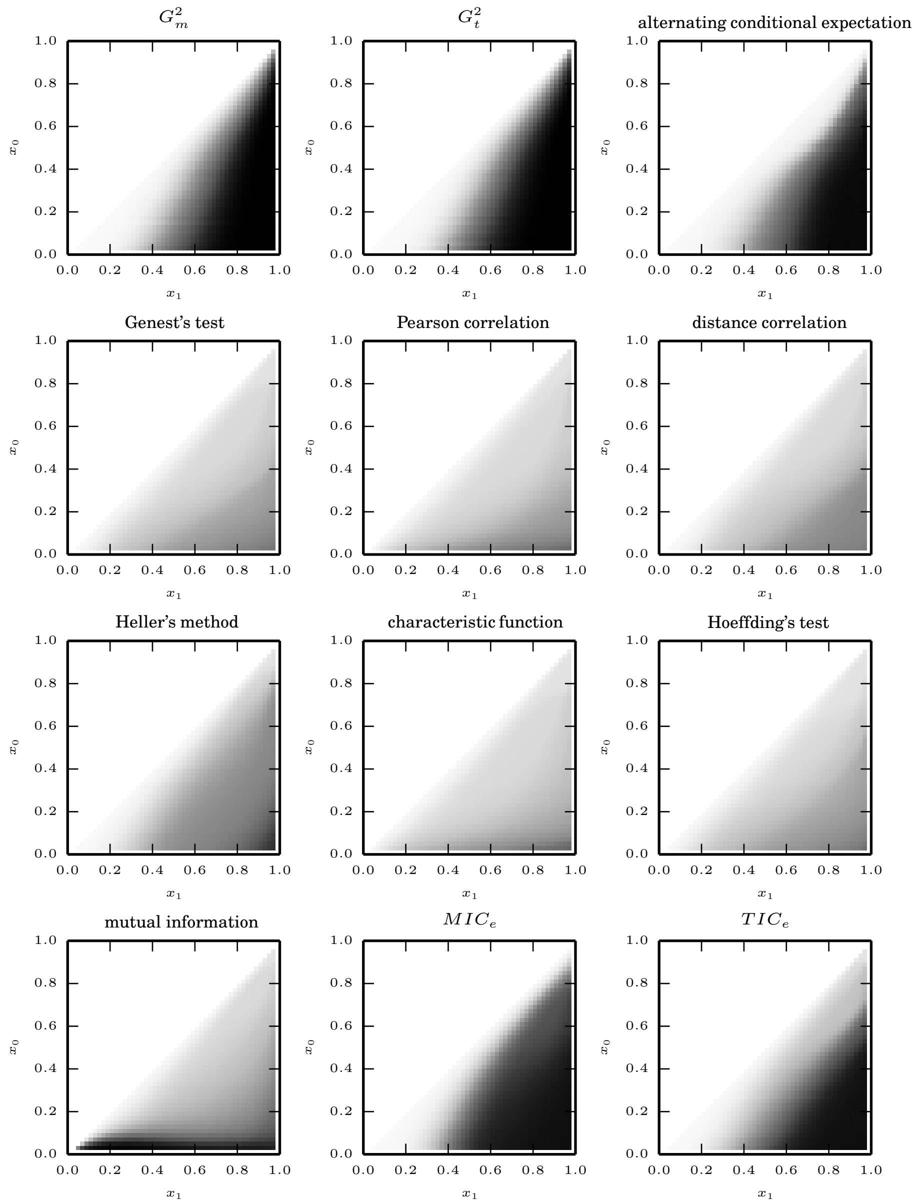

Fig. 3. Heat maps for the equitability of different methods. Each gray dot corresponding to (x_1, x_0) $(0 < x_0 < x_1 < 1)$ represents the power of the method for testing $\mathcal{H}_0 : \Psi = x_0$ against $\mathcal{H}_1 : \Psi = x_1$, averaging over a class of functions. The darker a dot, the higher the average power. We chose sample size $n = 225$ and performed 1,000 replications for each relationship and pair of (x_0, x_1) $(0 < x_0 < x_1 < 1)$.

The introduction of the working model provides a simple and intuitive derivation of $G^2_{Y|X}$. With $X \sim U(0, 1)$, we consider Example 1–4 and

Example 5. $Y = X + \epsilon\sigma$ and $\epsilon \sim \sqrt{3}U(-1, 1)$;

Example 6. $Y = X + \epsilon\sigma$ and $\epsilon \sim \sqrt{3}e^{-|X|}U(-1, 1)$;

Table 2. *Sum of squared errors for $G^2_{\rm m}$ and $G^2_{\rm t}$ with increasing n*

	$G^2_{\rm m}$				$G^2_{\rm t}$			
n	ex. 1	ex. 2	ex. 3	ex. 4	ex. 1	ex. 2	ex. 3	ex. 4
100	5.11	4.56	19.27	16.45	4.99	3.56	13.15	11.53
225	2.37	2.56	9.30	7.55	2.39	1.88	6.41	5.37
400	1.35	1.42	5.17	4.16	1.35	1.05	3.67	3.04
	$G^2_{\rm m}$				$G^2_{\rm t}$			
n	ex. 5	ex. 6	ex. 7	ex. 8	ex. 5	ex. 6	ex. 7	ex. 8
100	4.87	4.10	20.29	17.29	5.56	3.12	13.45	11.73
225	2.29	2.43	9.05	8.98	2.76	1.77	6.13	6.42
400	1.49	1.49	5.38	4.82	1.93	1.08	3.78	3.46

Example 7. $Y = X^2/\sqrt{2} + \epsilon\sigma$ and $\epsilon \sim \sqrt{3}U(-1, 1)$;

Example 8. $Y = X^2/\sqrt{2} + \epsilon\sigma$ and $\epsilon \sim \sqrt{3}e^{-|X|}U(-1, 1)$.

For Example 1, 3, 5, and 7, $G^2_{Y|X}$ is $(1+\sigma^2)^{-1}$; for Example 2, 4, 6 and 8, $G^2_{Y|X}$ is $(1 + 0.07\sigma^2)(1 + 0.52\sigma^2)^{-1}$. We chose $\sigma = 1$ and simulated $1,000$ replications for each model and sample size and used $\lambda_0 = 3$ for $G^2_{\rm m}$ and $G^2_{\rm t}$. Table 2 shows the sum of squared errors of $G^2_{\rm m}(Y \mid X, \lambda_0)$ and $G^2_{\rm t}(Y \mid X, \lambda_0)$ for the different models as n varies. We found that the sum of squared errors decreased roughly in the order of n^{-1} for both estimators and that $G^2_{\rm t}$ appeared slightly more accurate. The sum of squared errors were similar when the function relationships were the same, regardless of the error type. This confirmed that the estimation accuracies of $G^2_{\rm m}$ and $G^2_{\rm t}$ are not sensitive to the Gaussian assumption.

4·2. *More simulations for power analysis*

Table 3 lists twenty functional relationships for power analysis. For all relationships, we normalize them so that $\mathrm{var}\{f(X)\} = 1$ with $X \sim U(0, 1)$. As an intuitive presentation, Figure 4 shows the twenty simulated relationships with $G^2_{Y|X} = 0.8$. The power analysis results with six methods for the first eight relationships were in the main paper. Figure 5 presents the power for the eight relationships with the remaining six methods. The power analysis of the remaining twelve relationships with the entire twelve methods are in Figures 6–8. Figures 7 and 8 have the same legend as Figure 6. We found $G^2_{\rm m}$ and $G^2_{\rm t}$ were among the most powerful test statistics and $G^2_{\rm t}$ showed a higher power than $G^2_{\rm m}$ in most examples.

4·3. *Influence of sample size*

We ran simulations with the same setup with $n = 50,\ 100,\ 225$ and 500. Figure 9 shows the average power of $G^2_{\rm m}$, $G^2_{\rm t}$, the Pearson correlation, the distance correlation, the method of Heller et al. (2016) and TIC$_e$ against different sample sizes. We found that $G^2_{\rm m}$ and $G^2_{\rm t}$ were among the most powerful methods when n is larger than 100. When the sample size is small, the powers of $G^2_{\rm m}$ and $G^2_{\rm t}$ were slightly lower than that of Heller et al. (2016) in some cases but were still among the most powerful methods. Power analysis for more relationships are in Fig. 10–12.

4·4. *Simulation for the empirical Bayes selection of λ_0*

We examined the distributions of $G^2_{\rm m}(\lambda_0)$ and $G^2_{\rm t}(\lambda_0)$ with $\lambda_0 = 0.5,\ 1.5,\ 2.5$ and 3.5 for $X \sim N(0, 1)$ and

Example 9. $Y = X + \sigma\epsilon$ and $\epsilon \sim N(0, 1)$.

Example 10. $Y = \sin(4\pi x)/0.7 + \sigma\epsilon$ and $\epsilon \sim N(0, 1)$.

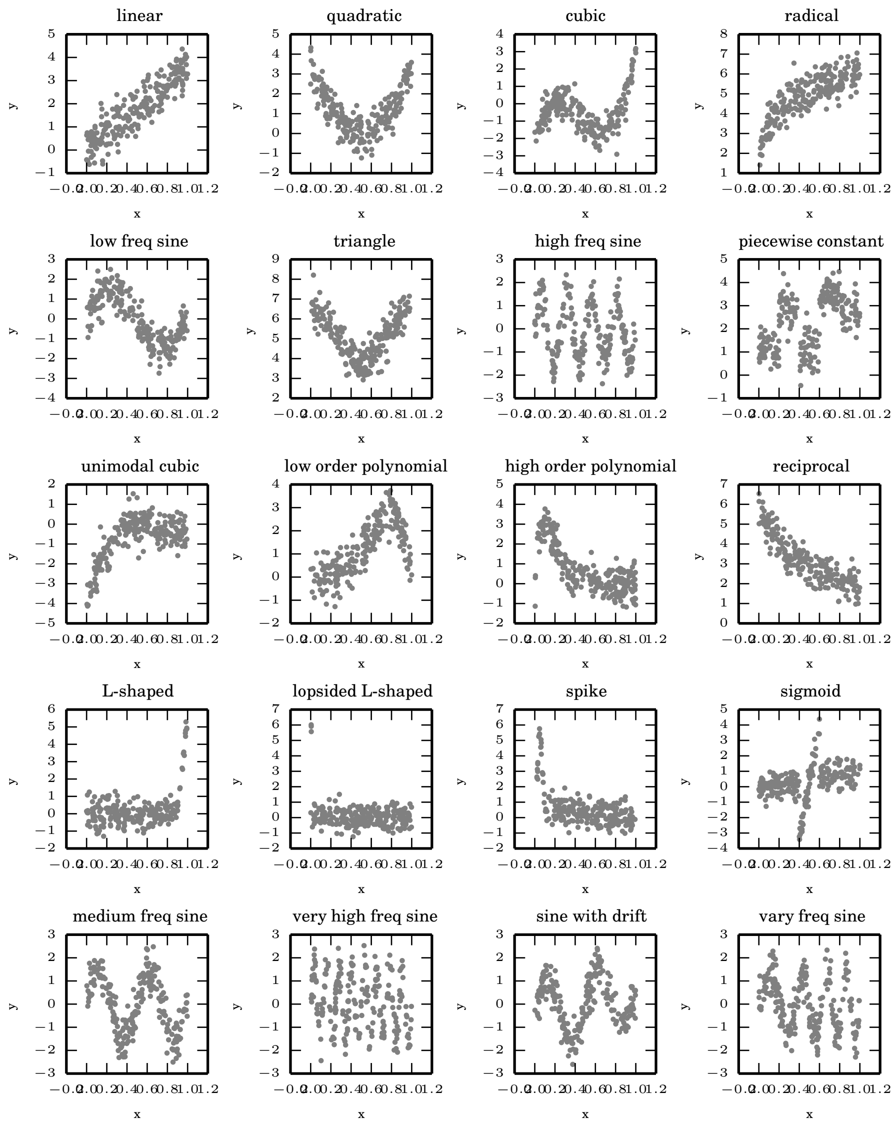

Fig. 4. Scatter plots for the twenty functional relationships in Table 3 with $n = 225$. We chose $\sigma = 0.5$ for each relationship and $G^2_{Y|X} = 0.8$.

Similar to Section 2.5 of the main paper, we chose $n = 225$ and computed G^2_{m} and G^2_{t} with data-driven λ_0. For each model we performed 1,000 replications and chose $\sigma = 9.95$ so that $G^2_{Y|X} = 0.01$. Figure 13 presents the same analysis as Figure 1 of the main paper but here X and Y were almost independent. A larger λ_0 was preferable for both models; this is because a small λ_0 tended to use more slices than necessary. The data-driven λ_0 still gave the most accurate estimates of the $G^2_{Y|X}$. Consistency of the data-driven estimators is proven in Section 5·2.

Table 3. *Functional relationships for power analysis*

Relation Name	Function
linear	x
quadratic	$(x-1/2)^2$
cubic	$32(x-1/3)^3-12(x-1/3)^2-3(x-1/3)$
radical	$x^{0.25}$
low freq sine	$\sin(2\pi x)$
triangle	$(1-x)I_{x<0.5}+xI_{x\geq 0.5}$
high freq sine	$\sin(8\pi x)$
piecewise constant	$0.287I_{x\leq 0.2}+0.796I_{0.2<x\leq 0.4}+0.290I_{0.4<x\leq 0.6}$ $+0.924I_{0.6<x\leq 0.8}+0.717I_{x>0.8}$
unimodal cubic	$32(x-2/3)^3-12(x-2/3)^2-3(x-2/3)$
low order polynomial	$x^4(1-x)$
high order polynomial	$x(1-x)^9$
reciprocal	$1/(x+0.5)$
L-shaped	$(x/90)I_{x\leq 0.9}+(90x-81)I_{x>0.9}$
lopsided L-shaped	$200xI_{x\leq 0.005}+(-198x+19.9)I_{0.005<x\leq 0.01}+(-x/99+1/99)I_{x>0.1}$
spike	$20xI_{x\leq 0.05}+(-18x+1.9)I_{0.05<x\leq 0.1}+(-x/9+1/9)I_{x>0.1}$
sigmoid	$\{50(x-0.5)+0.5\}I_{0.4<x\leq 0.6}+I_{x>0.6}$
medium freq sine	$\sin(4\pi x)$
very high freq sine	$\sin(16\pi x)$
sine with drift	$\sin\{2\pi(2x-1)\}+(2x-1)/2$
vary freq sine	$\sin\{4\pi x(1+x)\}$

5. PROOFS

5·1. *Proof of Theorem 1 - consistency*

The following lemma is needed for the main theorem.

LEMMA 1. *Suppose X and Y are univariate continuous random variables with $|X|,\ |Y|<B$ and $\mathrm{var}(Y)>b^{-2}$. Given n observations (x_i,y_i) $(i=1,\dots,n)$ and let $\widehat{\sigma}^2$ be the residual variance after regressing Y on X. Then,*

$$\mathrm{pr}\left[\left|\widehat{\sigma}^2-\left\{\mathrm{var}(Y)-\frac{\mathrm{cov}^2(X,Y)}{\mathrm{var}(X)}\right\}\right|>\epsilon\right]\leq 10e^{-C(B,b)n\epsilon^2}$$

with $C(B,b)=(288b^2B^4)^{-1}\min\{1,\ (4b^2B^2)^{-1}\}$ and $\epsilon>0$ small enough.

Proof of Lemma 1. Without loss of generality, we assume $E(X)=E(Y)=0$, $\mathrm{var}(X)=\mathrm{var}(Y)=1$ and $E(XY)=\rho$. By definition

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}y_i^2-\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)^2-\frac{\left\{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-(\frac{1}{n}\sum_{i=1}^{n}x_i)(\frac{1}{n}\sum_{i=1}^{n}y_i)\right\}^2}{\frac{1}{n}\sum_{i=1}^{n}x_i^2-(\frac{1}{n}\sum_{i=1}^{n}x_i)^2}.$$

Then $x_i^2,\, y_i^2\in[0,\ B^2]$, $x_iy_i\in[-B^2,\ B^2]$. According to Hoeffding's inequality,

$$\mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n}x_i\right|>\epsilon/6\right),\quad \mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n}y_i\right|>\epsilon/6\right),\quad \mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n}x_i^2-1\right|>\epsilon/6\right),$$
$$\mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n}y_i^2-1\right|>\epsilon/6\right),\quad \mathrm{pr}\left(\left|\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\rho\right|>\epsilon/6\right)\leq 2\exp\{-c(B)n\epsilon^2\}$$

with $c(B) = (72B^2)^{-1} \min(1,\ B^{-2})$. If $\epsilon < 1$ and

$$\left|\frac{1}{n}\sum_{i=1}^n x_i\right|, \quad \left|\frac{1}{n}\sum_{i=1}^n y_i\right|, \quad \left|\frac{1}{n}\sum_{i=1}^n x_i^2 - 1\right|, \quad \left|\frac{1}{n}\sum_{i=1}^n y_i^2 - 1\right|, \quad \left|\frac{1}{n}\sum_{i=1}^n x_i y_i - \rho\right| \le \epsilon/6,$$

we have

$$\begin{aligned}\left|\widehat{\sigma}^2 - 1 + \rho^2\right| &\le \left|1 - \frac{1}{n}\sum_{i=1}^n y_i^2\right| + \left|\frac{1}{n}\sum_{i=1}^n y_i\right|^2 + \frac{|\frac{1}{n}\sum_{i=1}^n x_i^2 - (\frac{1}{n}\sum_{i=1}^n x_i)^2 - 1|\rho^2}{|\frac{1}{n}\sum_{i=1}^n x_i^2 - (\frac{1}{n}\sum_{i=1}^n x_i)^2|} \\ &\quad \frac{|\left\{\frac{1}{n}\sum_{i=1}^n x_i y_i - (\frac{1}{n}\sum_{i=1}^n x_i)(\frac{1}{n}\sum_{i=1}^n y_i)\right\}^2 - \rho^2|}{|\frac{1}{n}\sum_{i=1}^n x_i^2 - (\frac{1}{n}\sum_{i=1}^n x_i)^2|} \\ &\le \frac{4(\epsilon/6 + \epsilon^2/36)}{1 - \epsilon/6 - \epsilon^2/36} < \epsilon.\end{aligned}$$

So $\mathrm{pr}\left(\left|\widehat{\sigma}^2 - 1 - \rho^2\right| > \epsilon\right) \le 10\exp\{-c(B)n\epsilon^2\}$. For general cases, define

$$X' = \frac{X - E(X)}{\mathrm{sd}(X)}, \quad Y' = \frac{Y - E(Y)}{\mathrm{sd}(Y)}.$$

Then $E(X') = E(Y') = 0$, $\mathrm{var}(X') = \mathrm{var}(Y') = 1$ and $|X'|, |Y'| < 2bB$. Thus,

$$\begin{aligned}&\mathrm{pr}\left[\left|\widehat{\sigma}^2 - \left\{\mathrm{var}(Y) - \frac{\mathrm{cov}^2(X,Y)}{\mathrm{var}(X)}\right\}\right| > \epsilon\right] \\ &= \mathrm{pr}\left[\left|\widehat{\sigma}'^2 - \{1 - \mathrm{cov}^2(X',Y')\}\right| > \frac{\epsilon}{\mathrm{var}(Y)}\right] \\ &\le 10\exp\{-\frac{c(2bB)}{\mathrm{var}(Y)^2}n\epsilon^2\} = 10\exp\{-C(B,b)n\epsilon^2\}\end{aligned}$$

with $C(B,b) = (288b^2B^4)^{-1}\min\{1,\ (4b^2B^2)^{-1}\}$. □

Proof of Theorem 1. We only need to prove that $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ and $G^2_{\mathrm{t}}(Y \mid X, \lambda_0)$ are consistent estimators of $G^2_{Y|X}$. If so, by switching X and Y, we must have that $G^2_{\mathrm{m}}(X \mid Y, \lambda_0)$ and $G^2_{\mathrm{t}}(X \mid Y, \lambda_0)$ are consistent estimators of $G^2_{X|Y}$ which guarantees the consistency of $G^2_{\mathrm{m}}(\lambda_0)$ and $G^2_{\mathrm{t}}(\lambda_0)$.

We first introduce some notations that will appear later. Suppose $|X|,\ |Y| < B$. Condition 1 shows that $\nu_X(y) > b^{-2}$ almost surely. Let $m = \lceil n^{1/2} \rceil$ be the minimum size of slices, and let $s \in S$ denote a slice and p_s be the probability that an observation falls in s. Let E_s, var_s, and cov_s denote the mean, variance and covariance conditional on slice s. Finally, define

$$\sigma_s^2 = \mathrm{var}_s(Y) - \frac{\mathrm{cov}_s^2(X,Y)}{\mathrm{var}_s(X)}.$$

Then by definition

$$\sigma_s^2 \ge \mathrm{var}_s(Y) - \mathrm{var}_s\{E(Y \mid X)\} = E_s\{\mathrm{var}(Y \mid X)\} \ge \exp[E_s\{\log\mathrm{var}(Y \mid X)\}] \ge b^{-2}.$$

For observations (x_i, y_i) $(i = 1, \ldots, n)$, let $\hat{\nu}^2$ be the estimated variance of Y and $\widehat{\sigma}_s^2$ be the residual variance after regressing Y on X in slice s. Besides, we use the following inequality

$$1 - x^{-1} < \log x < x - 1, \quad x > 0$$

throughout the proof.

Now we prove that $G^2_{\mathrm{m}}(Y \mid X, \lambda_0)$ is a consistent estimator for $G^2_{Y|X}$. Define

$$d_{Y|X} = \log \mathrm{var}(Y) - E\left\{\log \mathrm{var}(Y \mid X)\right\},$$

so $G^2_{Y|X} = 1 - \exp(-d_{Y|X})$. Because

$$G^2_{\mathrm{m}}(Y \mid X) = 1 - \exp\{-\max_{S\colon m_S \geq m} D(Y \mid S, \lambda_0)\},$$

we only need to show the consistency of

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

We prove this in two steps:

Step 1: We show that there exists $\eta(n) > 0$ and $\eta(n) \to 0$ as $n \to \infty$, such that

$$\mathrm{pr}\left\{\limsup_{n\to\infty} D(Y \mid X, \lambda_0) < d_{Y|X} + \eta(n)\right\} = 1,$$

which means that $D(Y \mid X, \lambda_0)$ is almost surely smaller than $d_{Y|X}$. Because for any slicing scheme S, $\log \mathrm{var}(Y) - \sum_{s\in S} p_s \log \sigma_s^2 \leq d_{Y|X}$, it is enough to show that there is $\eta(n)$ such that

$$\mathrm{pr}\left\{\limsup_{n\to\infty} D(Y \mid S, \lambda_0) - \log \mathrm{var}(Y) + \sum_{s\in S} p_s \log(\sigma_s^2) < \eta(n)\right\} = 1.$$

Let $\delta(n) = \log(n) n^{-1/4}$. By definition of $D(Y \mid S, \lambda_0)$, we have

$$\begin{aligned} & D(Y \mid S, \lambda_0) - \log \mathrm{var}(Y) + \sum_{s\in S} p_s \log(\sigma_s^2) \\ \leq & \left\{\log \hat{\nu}^2 - \log \mathrm{var}(Y)\right\} + \sum_{s\in S}\left(p_s - \frac{n_s}{n}\right)\log \sigma_s^2 + \sum_{s\in S}\frac{n_s}{n}\left(\log \sigma_s^2 - \log \widehat{\sigma}_s^2\right). \end{aligned}$$

First, we consider $\log \hat{\nu}^2 - \log \mathrm{var}(Y)$. By Hoeffding's inequality, for $0 < \epsilon < 2$,

$$\begin{aligned} & \mathrm{pr}\left\{|\hat{\nu}^2 - \mathrm{var}(Y)| > \epsilon\right\} \\ \leq & \mathrm{pr}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\{y_i - E(Y)\}^2 - \mathrm{var}(Y)\right| > \epsilon/2\right] + \mathrm{pr}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} y_i - E(Y)\right| > \epsilon/2\right\} \\ \leq & 4\exp\left[-n\epsilon^2 \min\{1,\ (4B^2)^{-1}\}(8B^2)^{-1}\right], \end{aligned}$$

we have

$$\begin{aligned} & \mathrm{pr}\left\{\log \hat{\nu}^2 - \log \mathrm{var}(Y) > \delta(n)\right\} \\ \leq & \mathrm{pr}\left\{\hat{\nu}^2 - \mathrm{var}(Y) > \mathrm{var}(Y)\delta(n)\right\} \leq 4n^{-C_1 n^{1/2}\log n} \end{aligned} \tag{1}$$

with $C_1 = \min\{1,\ (4B^2)^{-1}\}(8b^4B^2)^{-1}$.

Second, we consider $\sum_{s\in S}(p_s - n_s/n)\log \sigma_s^2$. Let us define a new random variable Z and $Z = \log \sigma_s^2$ if X is in slice s. Let z_i $(i = 1, \ldots n)$ be n independent observations of Z, then,

$$E(Z) = \sum_{s\in S} p_s \log \sigma_s^2, \quad \frac{1}{n}\sum_{i=1}^{n} z_i = \sum_{s\in S}\frac{n_s}{n}\log \sigma_s^2.$$

By Hoeffding's inequality and the fact that $\sigma_s^2 \in [b^{-2},\ B^2]$,

$$\text{pr}\left\{\left|\sum_{s\in S}(p_s - \frac{n_s}{n})\log\sigma_s^2\right| > \delta(n)\right\} \leq 2n^{-C_2 n^{1/2}\log n} \tag{2}$$

with $C_2 = \min(1/|\log B|,\ 1/|\log b|)^2/2$.

Third, we focus on the difference between $\log\widehat{\sigma}_s^2$ and $\log\sigma_s^2$. Consider a slicing scheme Q_n of n^4 slices such that an observation falls in each slice equally. Given n observations, the probability for any of the n^4 slices containing more than one observations is smaller than

$$n^4\left\{1 - \left(1+n^{-3}\right)\left(1-n^{-4}\right)^n\right\} \leq n^{-2}.$$

Then event

$$E_{1,n} = \{\text{each slice of } Q_n \text{ has at most one observation}\}$$

satisfies $\text{pr}\,(\liminf_{n\to\infty} E_{1,n}) = 1$. Thus, we only need to consider slicing schemes that are more refined than Q_n, denoted as $S \preceq Q_n$. Define the set of slices as

$$\Xi = \{s \mid \text{there exists } S \preceq Q_n \text{ such that } s \in S\}.$$

The set Ξ contains at most $n^4(n^4+1)/2 = O(n^8)$ slices. Each slice $s \in \Xi$ contains at least m observations. By Lemma 1, if $\delta(n) < 0.5b^{-2}$,

$$\begin{aligned}
&\text{pr}\left\{\log\sigma_s^2 - \log\widehat{\sigma}_s^2 > \delta(n)\right\} \qquad (3)\\
&\leq P\{\sigma_s^2/\widehat{\sigma}_s^2 - 1 > \delta(n)\}\\
&\leq \text{pr}\left\{|\widehat{\sigma}_s^2 - \sigma_s^2| > \delta(n)\right\} + P\left\{|\widehat{\sigma}_s^2 - \sigma_s^2| > \delta(n)\widehat{\sigma}_s^2,\ |\widehat{\sigma}_s^2 - \sigma_s^2| \leq \delta(n)\right\}\\
&\leq 20n^{-C_3\log(n)}.
\end{aligned}$$

with $C_3 = C(B,b)\min\{1,\ (4b^4)^{-1}\}$. Let $\eta(n) = 3\delta(n)$ and event

$$E_{2,n} = \{\max_{S\preceq Q_n} D(Y \mid S, \lambda_0) < d_{Y|X} + \eta(n)\}.$$

Combine the results of (1)–(3), we have $\text{pr}\,(\liminf_{n\to\infty} E_{1,n}\cap E_{2,n}) = 1$, which means that $G_{\text{m}}^2(Y \mid X, \lambda_0)$ is almost surely smaller than $G_{Y|X}^2$.

Step 2: Next, we show that there exists $\eta'(n) > 0$ and $\eta'(n) \to 0$ as $n \to \infty$, such that

$$\text{pr}\left\{\liminf_{n\to\infty} D(Y \mid X, \lambda_0) > d_{Y|X} - \eta'(n)\right\} = 1,$$

which means that $D(Y \mid X, \lambda_0)$ is almost surely larger than $d_{Y|X}$. We just need to prove that for any sample size n, there exists a slicing scheme T_n such that

$$\text{pr}\left(\liminf_{n\to\infty} E_{3,n}\cap E_{4,n}\right) = 1,$$

where

$$E_{3,n} = \{\text{each slice of } T_n \text{ contains at least } m \text{ samples}\}$$

and

$$E_{4,n} = \{D(Y \mid T_n, \lambda_0) > d_{Y|X} - \eta'(n)\}.$$

Consider a slicing scheme T_n of $\lfloor n^{1/4}\rfloor$ slices such that an observation falls in one slice equally. Then, we further divide each slice into $\lfloor n^{1/2}\rfloor$ bins such that an observation falls in each bin equally. Given n observations, the probability that each bin contains at least one observation is greater than

$$1-\lfloor n^{1/4}\rfloor\lfloor n^{1/2}\rfloor(1-n^{-3/4})^n > 1-\lfloor n^{1/4}\rfloor\lfloor n^{1/2}\rfloor e^{-n^{1/4}},$$

so each slice of T_n contains at least m observations. Then, $\mathrm{pr}\,(\liminf_{n\to\infty} E_{3,n}) = 1$. Define

$$\Delta_n(T_n) = \log \mathrm{var}(Y) - \sum_{s\in T_n} p_s \log \mathrm{var}_s(Y).$$

We first consider the difference between $D(Y \mid T_n, \lambda_0) - \Delta_n(T_n)$:

$$\begin{aligned}
&D(Y \mid T_n, \lambda_0) - \Delta_n(T_n)\\
&\geq \left\{\log \hat{\nu}^2 - \log \mathrm{var}(Y)\right\} + \sum_{s\in T_n}\left(p_s - \frac{n_s}{n}\right)\log \mathrm{var}_s(Y) + \sum_{s\in T_n}\frac{n_s}{n}\{\log \mathrm{var}_s(Y) - \log \widehat{\sigma}_s^2\}\\
&\quad -\lambda_0 n^{-3/4}\log n.
\end{aligned}$$

Similar as (1), if $\delta(n) < 0.5b^{-2}$,

$$\begin{aligned}
&\mathrm{pr}\left\{\log \hat{\nu}^2 - \log \mathrm{var}(Y) < -\delta(n)\right\} \qquad (4)\\
&\leq \mathrm{pr}\left\{1 - \mathrm{var}(Y)/\hat{\nu}^2 < -\delta(n)\right\}\\
&\leq \mathrm{pr}\left\{|\hat{\nu}^2 - \mathrm{var}(Y)| > \delta(n)\right\} + P\left\{|\hat{\nu}^2 - \mathrm{var}(Y)| > \delta(n)\hat{\nu}^2,\ |\hat{\nu}^2 - \mathrm{var}(Y)| \leq \delta(n)\right\}\\
&\leq 4n^{-C_4 n^{1/2}\log n}
\end{aligned}$$

with $C_4 = (8B^2)^{-1}\min\{1,\ (4B^2)^{-1}\}\min\{1, (4b^4)^{-1}\}$. Similar as (2), we have

$$\mathrm{pr}\left\{\left|\sum_{s\in S}(p_s - \frac{n_s}{n})\log \mathrm{var}_s(Y)\right| > \delta(n)\right\} \leq 2n^{-C_2 n^{1/2}\log n}. \qquad (5)$$

Besides, $\mathrm{var}_s(Y) \geq \sigma_s^2$ and

$$\begin{aligned}
&\mathrm{pr}\left\{\log \mathrm{var}_s(Y) - \log \widehat{\sigma}_s^2 < -\delta(n)\right\} \qquad (6)\\
&\leq \mathrm{pr}\left\{\log \sigma_s^2 - \log \widehat{\sigma}_s^2 < -\delta(n)\right\}\\
&\leq \mathrm{pr}\left\{1 - \widehat{\sigma}_s^2/\sigma_s^2 < -\delta(n)\right\}\\
&\leq \mathrm{pr}\left\{|\widehat{\sigma}_s^2 - \sigma_s^2| \geq b^{-2}\delta(n)\right\} \leq 10n^{-C(B,b)b^{-4}\log(n)}.
\end{aligned}$$

Now, define $\delta_1(n) = 3\delta(n) + \lambda_0\log(n)n^{-3/4}$ and event

$$E_{5,n} = \{D(Y \mid T_n, \lambda_0) > \Delta_n(T_n) - \delta_1(n)\}.$$

By (4)–(6), $\mathrm{pr}\,(\liminf_{n\to\infty} E_{3,n}\cap E_{5,n}) = 1$.

The only problem left is how to control the difference between $\Delta_n(T_n)$ and $d_{Y|X}$, which is

$$\Delta_n(T_n) - d_{Y|X} = \sum_{s\in T_n} p_s\left\{\frac{1}{p_s}\int_s \log \nu_Y^2(x) f_X(x)dx - \log \mathrm{var}_s(Y)\right\}.$$

Denote the probability density function of X as $f_X(x)$. For one slice s, because X is a continuous random variable, set

$$\frac{1}{p_s}\int_s \mu_Y(x) f_X(x)dx = \mu_Y(x_s'), \quad \frac{1}{p_s}\int_s \log \nu_Y^2(x) f_X(x)dx = \log \nu_Y^2(x_s''),$$

where x'_s and x''_s lie in the slice almost surely. Then

$$
\begin{aligned}
&\log \nu_Y^2(x''_s) - \log \mathrm{var}_s(Y) \\
&= \log \nu_Y^2(x''_s) - \log \left[\frac{1}{p_s} \int_s \nu_Y^2(x) f_X(x) dx + \frac{1}{p_s} \int_s \left\{ \mu_Y(x) - \mu_Y(x'_s) \right\}^2 f_X(x) dx \right] \\
&= \log \nu_Y^2(x''_s) - \log \left[\nu_Y^2(x''_s) + \frac{1}{p_s} \int_s \int_{x''_s}^{x} 2\nu_Y(z)\nu'_Y(z) dz f_X(x) dx \right. \\
&\qquad \left. + \frac{1}{p_s} \int_s \left\{ \int_{x'_s}^{x} \mu'_Y(z) dz \right\}^2 f_X(x) dx \right] \\
&\geq \log \nu_Y^2(x''_s) - \log \left[\nu_Y^2(x''_s) + \int_s 2\nu_Y(x) |\nu'_Y(x)| dx + \left\{ \int_s |\mu'_Y(x)| dx \right\}^2 \right].
\end{aligned}
$$

According to Condition 3, we have

$$
\begin{aligned}
&\log \nu_Y^2(x''_s) - \log \mathrm{var}_s(Y) \\
&\geq \log \nu_Y^2(x'') - \log \left\{ \nu_Y^2(x'') + 2C \int_s \nu_Y^2(x) dx + C^2 \int_s 1 dx \int_s \nu_Y^2(x) dx \right\} \\
&\geq - \frac{\int_s \nu_Y^2(x) dx \left(2C + C^2 \int_s 1 dx \right)}{\nu_Y^2(x'')} \\
&\geq -2b^2 B^2 C (1 + BC) \int_s 1 dx.
\end{aligned}
$$

Then, we can conclude

$$
\begin{aligned}
\Delta_n(T_n) - d_{Y|X} &\geq -2 p_s b^2 B^2 C (1 + BC) \sum_{s \in T_n} \int_s 1 dx \\
&\geq -4 \lfloor n^{1/4} \rfloor^{-1} (1 + BC) C b^2 B^3 = -\delta_2(n).
\end{aligned}
$$

Therefore, let $\eta'(n) = \delta_1(n) + \delta_2(n)$, we have $\mathrm{pr}\left(\liminf_{n\to\infty} E_{3,n} \cap E_{4,n} \right) = 1$, which means $G_{\mathrm{m}}^2(Y \mid X, \lambda_0)$ is almost surely larger than $G_{Y|X}^2$. By Steps 1 and 2, we can conclude that $G_{\mathrm{m}}^2(Y \mid X, \lambda_0)$ is a consistent estimator of $G_{Y|X}^2$.

To prove the consistency of $G_{\mathrm{t}}^2(Y \mid X, \lambda)$, we introduce a new quantity $Z(\lambda_0) = \sum_{S:\, m_S \geq m} n^{-\lambda_0(|S|-1)/2}$; $Z(\lambda_0)$ is bounded by 1 and $(1 + n^{-\lambda_0/2})^n$. By definition of $G_{\mathrm{m}}^2(Y \mid X, \lambda_0)$ and $G_{\mathrm{t}}^2(Y \mid X, \lambda_0)$, we have

$$
\begin{aligned}
\left\{ 1 - G_{\mathrm{t}}^2(Y \mid X, \lambda_0) \right\}^{-n/2} &= Z(\lambda_0)^{-1} \sum_{S:\, m_S \geq m} \exp\{ \frac{n}{2} D(Y \mid S, \lambda_0) \} \\
&\geq Z(\lambda_0)^{-1} \exp\{ \frac{n}{2} D(Y \mid X, \lambda_0) \}, \\
\left\{ 1 - G_{\mathrm{t}}^2(Y \mid X, \lambda_0) \right\}^{-n/2} &\leq Z(\lambda_0)^{-1} \sum_{S:\, m_S \geq m} \exp\{ \frac{n}{2} D(Y \mid S, \frac{\lambda_0}{2}) - \frac{\lambda_0}{4} (|S| - 1) \log(n) \} \\
&\leq Z(\lambda_0)^{-1} Z(\frac{\lambda_0}{2}) \exp\{ \frac{n}{2} D(Y \mid X, \frac{\lambda_0}{2}) \}.
\end{aligned}
$$

By the consistency of $D(Y \mid X, \lambda_0)$ and $D(Y \mid X, \lambda_0/2)$, we prove that $G_{\mathrm{t}}^2(Y \mid X, \lambda_0)$ is an consistent estimator of $G_{Y|X}^2$. □

5·2. *Consistency of G^2_{m} and G^2_{t} with empirical Bayes selection of λ_0*

Suppose λ^* is the optimal λ_0 that maximizes $\mathrm{BF}(\lambda_0)$ from a range $[\lambda_1,\ \lambda_2]$ with $\lambda_1 > 0$. Then $Z(\lambda_2) \le Z(\lambda^*) \le Z(\lambda_1)$ and

$$
\begin{aligned}
G^2_{\mathrm{m}}(Y \mid X, \lambda^*) &\le G^2_{\mathrm{m}}(Y \mid X, \lambda_1), \\
\left\{1 - G^2_{\mathrm{m}}(Y \mid X, \lambda^*)\right\}^{-n/2} &= \exp\{\frac{n}{2} D(Y \mid X, \lambda^*)\} \\
&\ge Z(\lambda_2)^{-1} \sum_{S:\ m_S \ge m} \exp\{\frac{n}{2} D(Y \mid S, \lambda^* + \lambda_2)\} \\
&\ge Z(\lambda_2)^{-1} \left\{1 - G^2_{\mathrm{m}}(Y \mid X, 2\lambda_2)\right\}^{-n/2}, \\
\left\{1 - G^2_{\mathrm{t}}(Y \mid X, \lambda^*)\right\}^{-n/2} &= Z(\lambda^*)^{-1} \sum_{S:\ m_S \ge m} \exp\{\frac{n}{2} D(Y \mid S, \lambda^*)\} \\
&\ge Z(\lambda_1)^{-1} \left\{1 - G^2_{\mathrm{m}}(Y \mid X, \lambda_2)\right\}^{-n/2}, \\
\left\{1 - G^2_{\mathrm{t}}(Y \mid X, \lambda^*)\right\}^{-n/2} &\le Z(\lambda^*)^{-1} \sum_{S:\ m_S \ge m} \exp\{\frac{n}{2} D(Y \mid S, \lambda_1)\} \\
&\le Z(\lambda_2)^{-1} Z(\lambda_1) \left\{1 - G^2_{\mathrm{t}}(Y \mid X, \lambda_1)\right\}^{-n/2}.
\end{aligned}
$$

By the consistency of $G^2_{\mathrm{m}}(Y \mid X, \lambda_1)$, $G^2_{\mathrm{m}}(Y \mid X, 2\lambda_2)$, $G^2_{\mathrm{m}}(Y \mid X, \lambda_2)$ and $G^2_{\mathrm{t}}(Y \mid X, \lambda_1)$, we conclude that $G^2_{\mathrm{m}}(Y \mid X, \lambda^*)$ and $G^2_{\mathrm{t}}(Y \mid X, \lambda^*)$ are consistent estimators. Then the estimators with data-driven λ_0 are consistent.

5·3. *Proof of Theorem 2 - Equivalence between G^2_{m} and R^2*

The following lemma is needed for the main theorem.

LEMMA 2. *Let* $(p_1, p_2, p_3) \sim \mathrm{Dir}(k_1, k_2, 2)$ *and*

$$\Lambda(q, p) = (k_1 - 1) \log \frac{q_1}{p_1} + (k_2 - 1) \log \frac{q_2}{p_2}.$$

Then for any $k_1,\ k_2 \ge 3$, $q_1,\ q_2 > 0,\ q_1 + q_2 = 1$ *and function* $\delta(p) > 0$,

$$\mathrm{pr}\left\{\Lambda(q, p) \ge \delta(p)\right\} \le (k_1 + k_2)^3 \int_0^1 e^{-\delta(p)} dp.$$

Proof of Lemma 2. By definition, we have

$$p_1^{k_1 - 1} p_2^{k_2 - 1} (1 - p_1 - p_2) \le q_1^{k_1 - 1} q_2^{k_2 - 1} e^{-\Lambda(q, p)},$$

so that

$$
\begin{aligned}
&\mathrm{pr}\left\{\Lambda(q, p) \ge \delta(p)\right\} \\
&= \frac{(k_1 + k_2 + 1)!}{(k_1 - 1)!(k_2 - 1)!} \int_{\Lambda(q,p) \ge \delta(p)} p_1^{k_1 - 1} p_2^{k_2 - 1} (1 - p_1 - p_2) dp_1 dp_2 \\
&\le \frac{(k_1 + k_2 + 1)!}{(k_1 - 1)!(k_2 - 1)!} q_1^{k_1 - 1} q_2^{k_2 - 1} \int_{\Lambda(q,p) \ge \delta(p)} e^{-\Lambda(q,p)} dp_1 dp_2 \\
&\le (k_1 + k_2)^3 \frac{(k_1 + k_2 - 2)!}{(k_1 - 1)!(k_2 - 1)!} q_1^{k_1 - 1} q_2^{k_2 - 1} \int_{\Lambda(q,p) \ge \delta(p)} e^{-\Lambda(q,p)} dp_1 dp_2 \\
&\le (k_1 + k_2)^3 \int_0^1 e^{-\delta(p)} dp.
\end{aligned}
$$

□

Proof of Theorem 2. If the slice scheme on X has only one slice, we have

$$D(Y \mid S, \lambda_0) = \log \hat{\nu}^2 - \log \widehat{\sigma}^2 = -\log(1 - R^2),$$

where $\widehat{\sigma}^2$ is the residual variance after regressing Y on X. Intuitively, if Y and X follow a bivariate normal, the optimal slice scheme is only one slice in each direction. Now, we show that

$$\text{pr}\left\{D(Y \mid X, \lambda_0) + \log(1 - R^2) > 0\right\} < 1.5 n^{-\lambda_0/3+5}.$$

For any slice scheme S,

$$D(Y \mid S, \lambda_0) + \log(1 - R^2) = \log \widehat{\sigma}^2 - \sum_{s \in S} \frac{n_s}{n} \log(\widehat{\sigma}_s^2) - \frac{\lambda_0}{n}(|S| - 1)\log n.$$

Without loss of generality, we assume that $\text{var}(Y) = 1$ and $x_1 < \ldots < x_n$. Suppose the connected slices each has n_i $(i = 1, \ldots |S|)$ observations. For $1 \leq j < k \leq n$, define

$$\Delta(j, k, \lambda_0) = \frac{k}{n} \log\{\widehat{\sigma}^{(k)}\}^2 - \frac{j}{n} \log\{\widehat{\sigma}^{(j)}\}^2 - \frac{k-j}{n} \log\{\widehat{\sigma}^{(k,j)}\}^2 - \frac{\lambda_0}{n} \log n.$$

Here, $\{\widehat{\sigma}^{(j)}\}^2$ is the residual variance of regressing y_i on x_i $(i = 1, \ldots, j)$, $\{\widehat{\sigma}^{(k)}\}^2$ is the residual variance of regressing y_i on x_i $(i = 1, \ldots, k)$ and $\{\widehat{\sigma}^{(k,j)}\}^2$ is the residual variance of regressing y_i on x_i $(i = j+1, \ldots, k)$. For given j, k, let

$$p_1 = \frac{j\{\widehat{\sigma}^{(j)}\}^2}{k\{\widehat{\sigma}^{(k)}\}^2}, \quad p_2 = \frac{(k-j)\{\widehat{\sigma}^{(k,j)}\}^2}{k\{\widehat{\sigma}^{(k)}\}^2}, \quad q_1 = \frac{j}{k}, \quad q_2 = 1 - q_1.$$

Then according to Cochran's theorem, we have

$$\begin{aligned}
&(p_1, p_2, 1 - p_1 - p_2) \sim \text{Dir}(j - 2, k - j - 2, 2), \\
&n\Delta(j, k, \lambda_0) = \Lambda(q, p) - \lambda_0 \log(n) + 3\log(q_1/p_1) + 3\log(q_2/p_2).
\end{aligned}$$

By Lemma 2 we have

$$\text{pr}\left\{\Lambda(q, p) > \lambda_0 \log(n)/3\right\} \leq k^3 n^{-\lambda_0/3} \leq n^{-\lambda_0/3+3}.$$

At the same time,

$$\begin{aligned}
&\text{pr}\left\{3\log(q_1/p_1) > \lambda_0 \log(n)/3\right\} \\
&= \frac{(k-3)!}{(j-3)!(k-j-1)!} \int_0^{q_1 n^{-\lambda_0/9}} p^{j-3}(1-p)^{k-j-1} dp \\
&\leq \frac{(k-3)!}{(j-3)!(k-j-1)!} \frac{1}{j-2} (q_1 n^{-\lambda_0/9})^{j-2} \\
&= (j/k)^{j-2} \frac{(k-3)!}{(j-2)!(k-j-1)!} \frac{1}{n^{\lambda_0(j-2)/9}} \leq \frac{1}{n^{(j-2)(\lambda_0/9-1)}}
\end{aligned}$$

If $n \geq 25$, we have $\text{pr}\left\{\Delta(j, k, \lambda_0) > 0\right\} \leq 3n^{-\lambda_0/3+3}$. On the other hand, for any slicing scheme with $|S| \geq 2$, $D(Y \mid S, \lambda_0) + \log(1 - R^2)$ equals

$$\sum_{h=1}^{|S|-1} \Delta\Big(\sum_{l=1}^{h} n_l, \sum_{l=1}^{h+1} n_l, \lambda_0\Big)$$

So

$$
\begin{aligned}
&\operatorname{pr}\left\{D(Y \mid X, \lambda_0) + \log(1 - R^2) > 0\right\} \\
\leq\ &\operatorname{pr}\left\{\max_{m \leq j < k \leq n-m} \Delta(j, k, \lambda_0) > 0\right\} \\
\leq\ &\sum_{m \leq j < k \leq n-m} \operatorname{pr}\left\{\Delta(j, k, \lambda_0) > 0\right\} < 1.5 n^{-\lambda_0/3+5}.
\end{aligned}
$$

Since X and Y are symmetric, the result tells us that $P\left\{G^2_{\mathrm{m}}(\lambda_0) = R^2\right\} > 1 - 3n^{-\lambda_0/3+5}$. When $\lambda_0 > 18$, we have $G^2_{\mathrm{m}}(\lambda_0) = R^2$ almost surely. □

REFERENCES

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**, 1–54.

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.

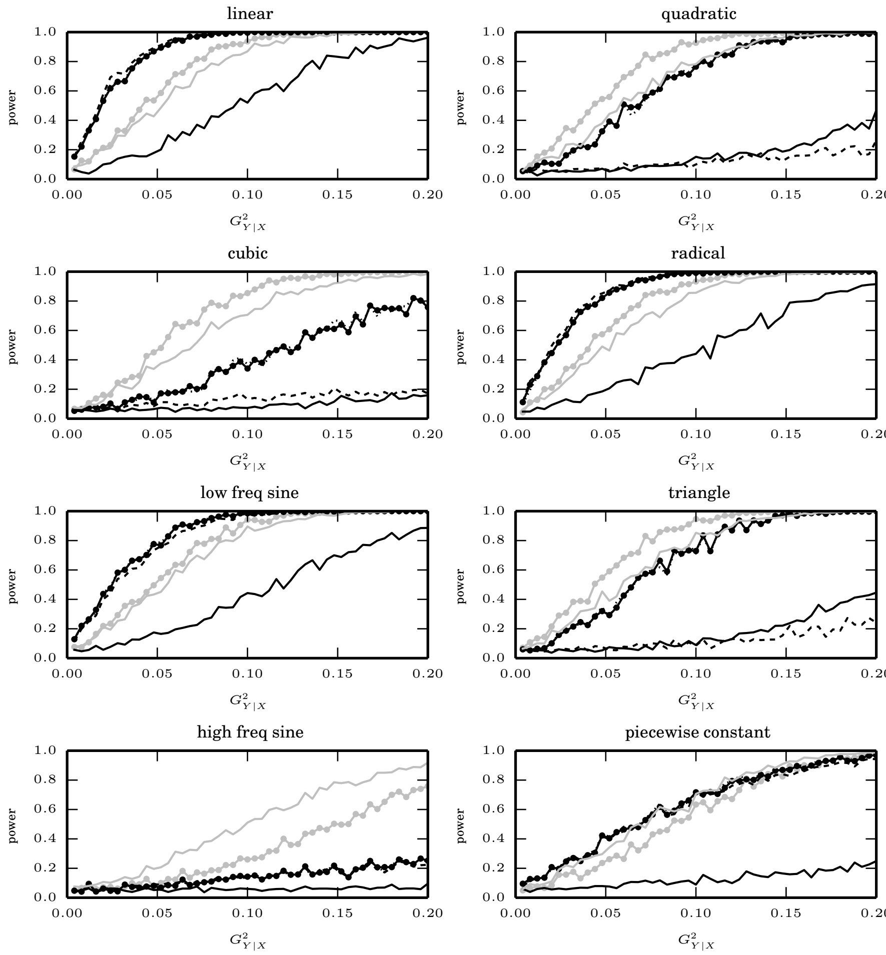

Fig. 5. The powers of mutual information (black solid), MIC$_e$ (grey solid), alternating conditional expectation (grey markers), characteristic function (black dashes), Genest's test (black dots) and Hoeffding's test (black markers) for independence test between X and Y when the function relationships are linear, quadratic, cubic, radical, low freq sine, triangle, high freq sine and piecewise constant. The x-axis is $G^2_{Y|X}$ and the y-axis is the power.

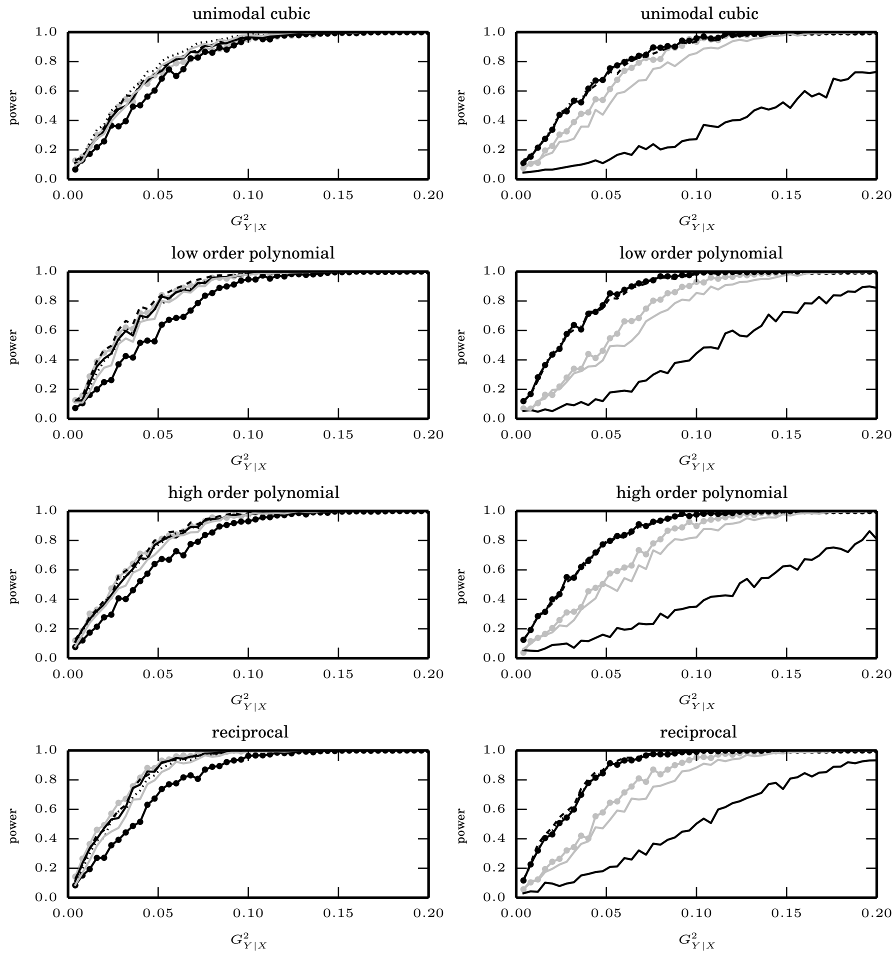

Fig. 6. The left column presents the powers of G^2_{m} (black solid), G^2_{t} (grey solid), Pearson correlation (grey markers), distance correlation (black dashes), the method of Heller et al. (2016) (black dots) and TIC_e (black markers) for independence test between X and Y when the function relationships are power functions; the right column presents the powers of mutual information (black solid), MIC_e (grey solid), alternating conditional expectation (grey markers), characteristic function (black dashes), Genest's test (black dots) and Hoeffding's test (black markers). The x-axis is $G^2_{Y|X}$ and the y-axis is the power.

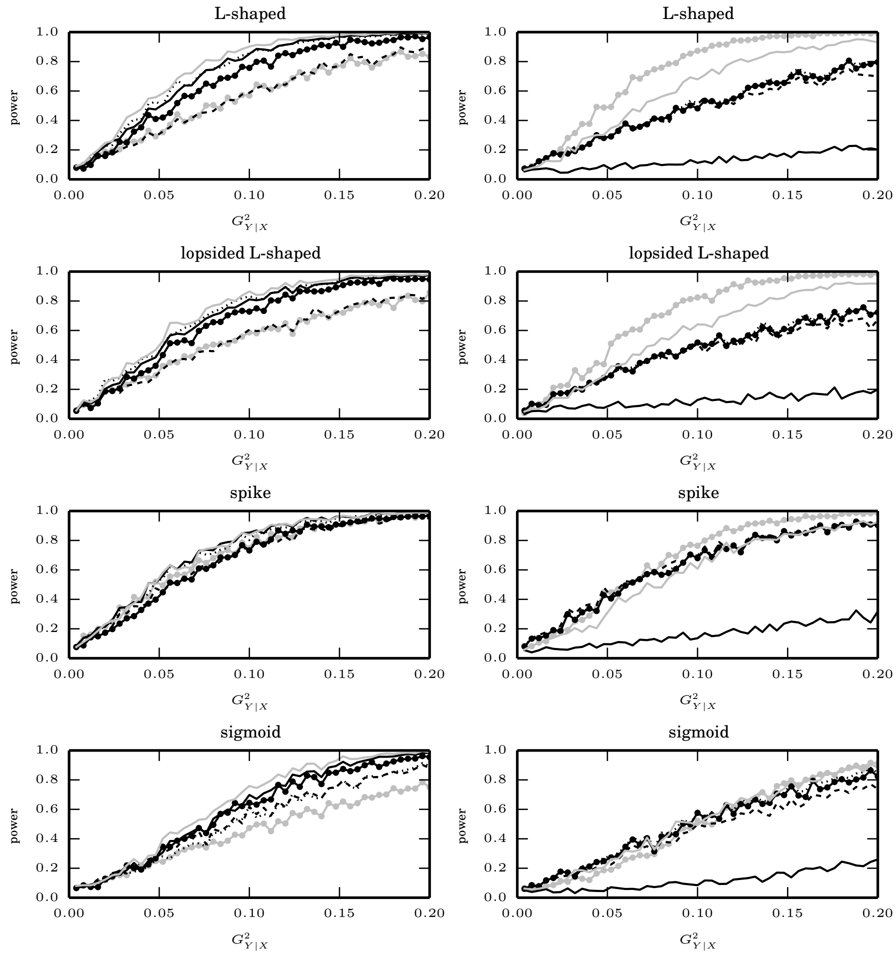

Fig. 7. The powers for independence test between X and Y when the function relationship are piecewise linear functions.

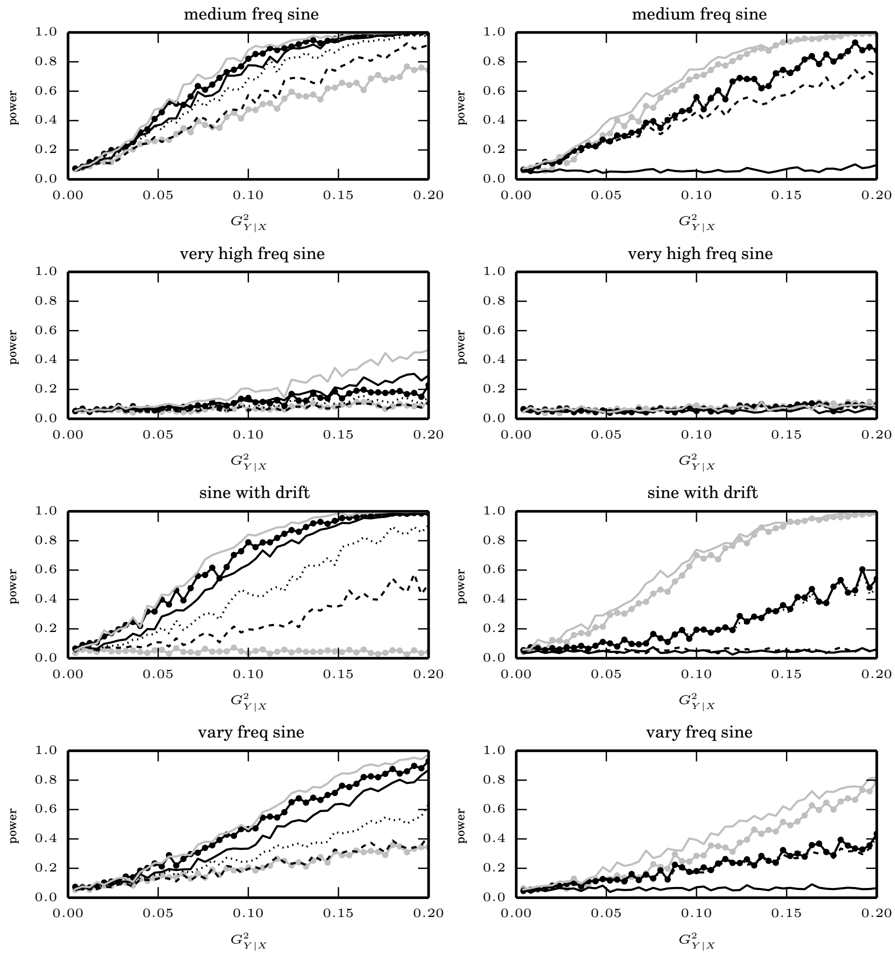

Fig. 8. The powers for independence test between X and Y when the function relationships are trigonometric functions.

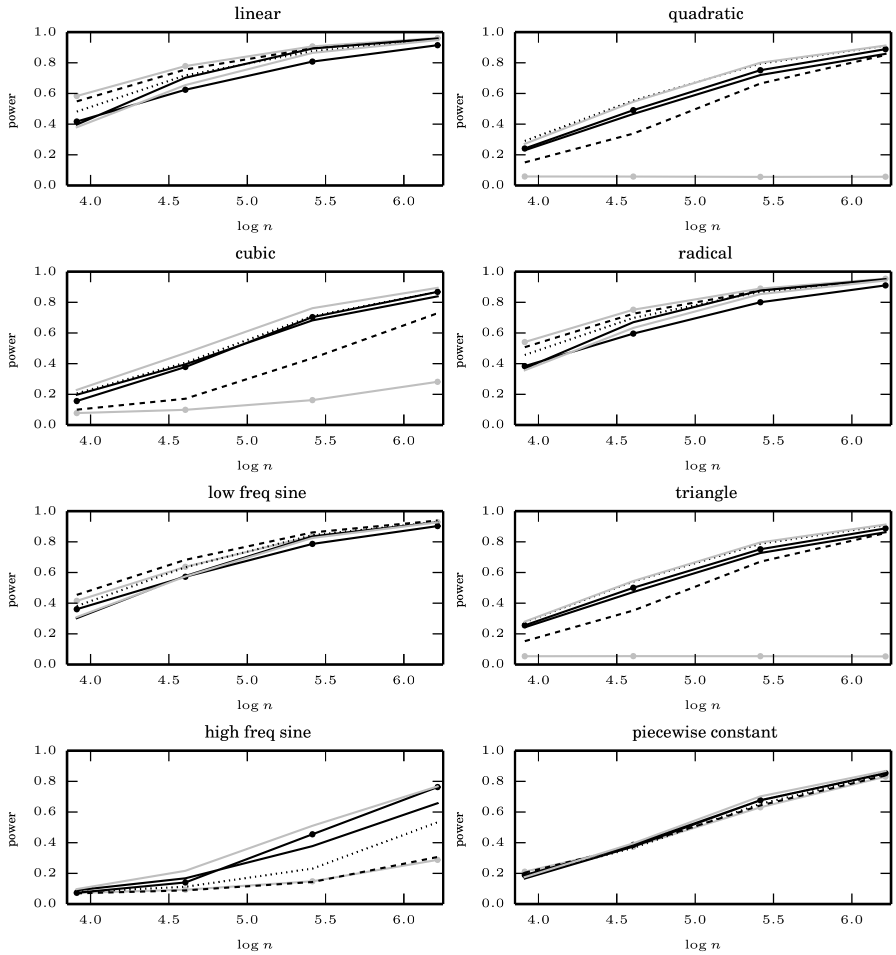

Fig. 9. The average powers of G^2_{m} (black solid), G^2_{t} (grey solid), Pearson correlation (grey markers), distance correlation (black dashes), the method of Heller et al. (2016) (black dots) and TIC_e (black markers) for testing independence between X and Y with $n = 50$, 100, 225 and 500. The underlying true functional relationships are linear, quadratic, cubic, radical, low freq sine, triangle, high freq sine and piecewise constant. The x-axis is logarithm of n with base 10 and the y-axis is the average power.

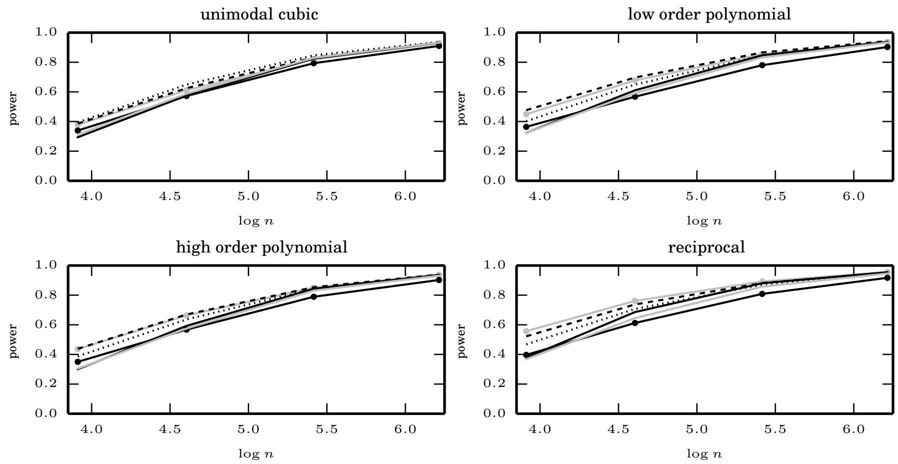

Fig. 10. The average powers for independence test between X and Y when the function relationships are power functions.

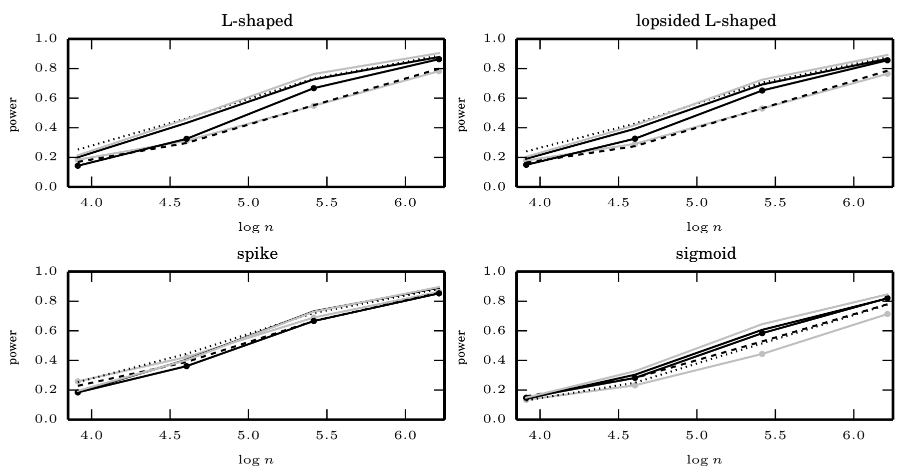

Fig. 11. The average powers for independence test between X and Y when the function relationship are piecewise linear functions.

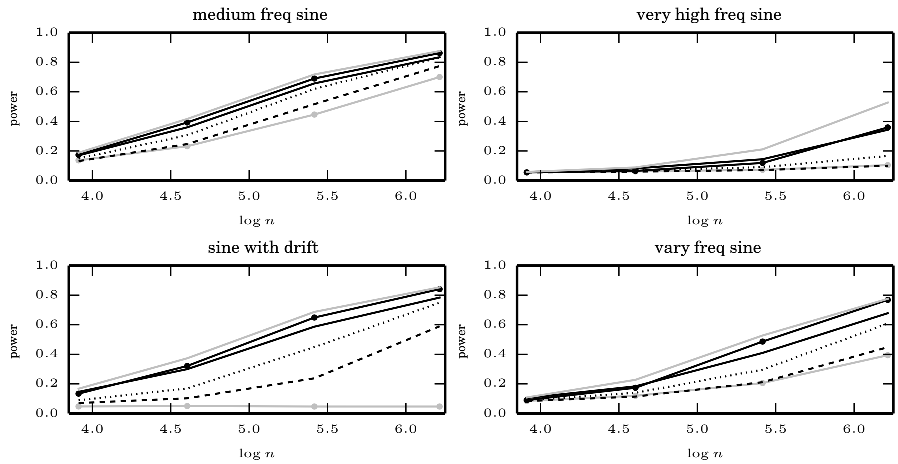

Fig. 12. The average powers for independence test between X and Y when the function relationships are trigonometric functions.

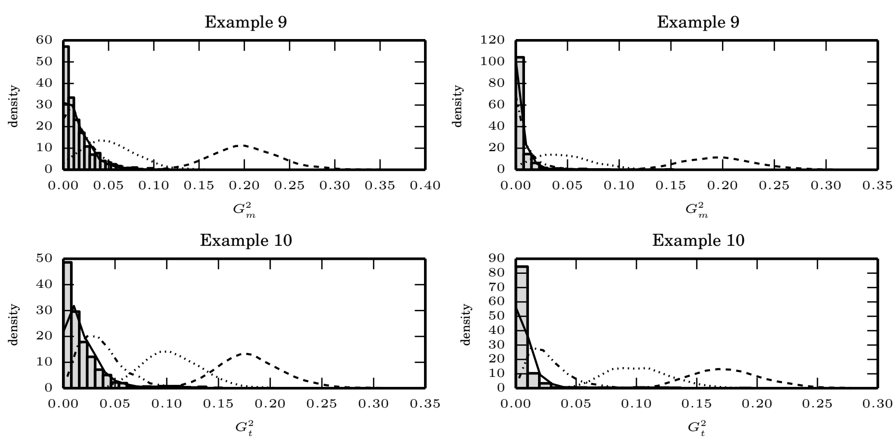

Fig. 13. Sampling distributions of G^2_{m} and G^2_{t} under the two models in Section 4·4 with $G^2_{Y|X} =$ 0.01 and λ_0 = 0.5 (dashes), 1.5 (dots), 2.5 (dot-dash) and 3.5 (solid). The density function in each case was estimated by the histogram. The sampling distributions of G^2_{m} and G^2_{t} with empirical Bayes selection of λ_0 were in gray shadow and overlaid on top of other density functions.