ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Cao D, Chen Y, Chen J, Zhang H, Yuan Z. 2021 An improved algorithm for the maximal information coefficient and its application. *R. Soc. Open Sci.* **8**: 201424.
https://doi.org/10.1098/rsos.201424

Received: 10 August 2020
Accepted: 18 January 2021

Subject Category:
Mathematics

Subject Areas:
pattern recognition/bioinformatics

Keywords:
maximal information coefficient, χ^2-test, statistical power, equitability, K-means clustering

Authors for correspondence:
Yuan Chen
e-mail: chenyuan0510@126.com
Zheming Yuan
e-mail: zhmyuan@sina.com

THE ROYAL SOCIETY
PUBLISHING

An improved algorithm for the maximal information coefficient and its application

Dan Cao[1,2], Yuan Chen[1], Jin Chen[1], Hongyan Zhang[1] and Zheming Yuan[1]

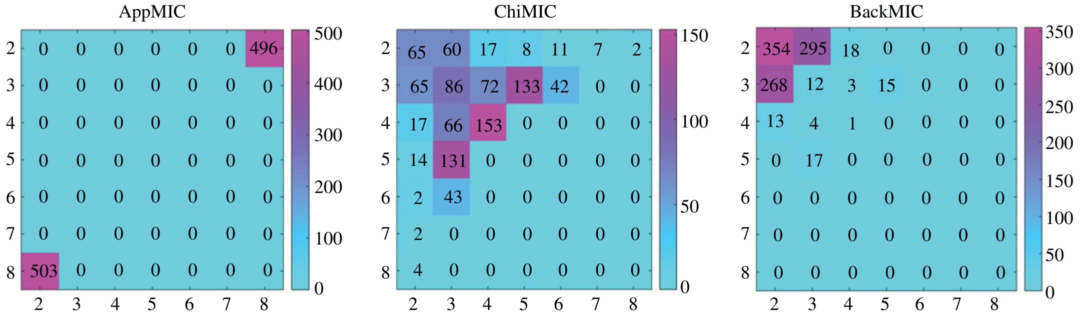
[1]Hunan Engineering and Technology Research Centre for Agricultural Big Data Analysis and Decision-making, Hunan Agricultural University, Changsha 410000, People's Republic of China
[2]Orient Science and Technology College of Hunan Agricultural University, Changsha 410000, Hunan, People's Republic of China

ZY, 0000-0002-7444-8743

The maximal information coefficient (MIC) captures both linear and nonlinear correlations between variable pairs. In this paper, we proposed the BackMIC algorithm for MIC estimation. The BackMIC algorithm adds a searching back process on the equipartitioned axis to obtain a better grid partition than the original implementation algorithm ApproxMaxMI. And similar to the ChiMIC algorithm, it terminates the grid search process by the χ^2-test instead of the maximum number of bins B(n, α). Results on simulated data show that the BackMIC algorithm maintains the generality of MIC, and gives more reasonable grid partition and MIC values for independent and dependent variable pairs under comparable running times. Moreover, it is robust under different α in B(n, α). MIC calculated by the BackMIC algorithm reveals an improvement in statistical power and equitability. We applied (1-MIC) as the distance measurement in the K-means algorithm to perform a clustering of the cancer/normal samples. The results on four cancer datasets demonstrated that the MIC values calculated by the BackMIC algorithm can obtain better clustering results, indicating the correlations between samples measured by the BackMIC algorithm were more credible than those measured by other algorithms.

1. Introduction

Correlation analysis has important applications in data mining, such as disease diagnosis [1,2], public management [3,4] and financial market analysis [5,6]. Since the maximal information coefficient (MIC) [7] was proposed as a type of nonlinear correlation measurement, it has been widely studied and applied because of its generality and approximate equitability [8–11]. In MIC, if there is a



AppMIC

	2	3	4	5	6	7	8
2	0	0	0	0	0	0	496
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	503	0	0	0	0	0	0

ChiMIC

	2	3	4	5	6	7	8
2	65	60	17	8	11	7	2
3	65	86	72	133	42	0	0
4	17	66	153	0	0	0	0
5	14	131	0	0	0	0	0
6	2	43	0	0	0	0	0
7	2	0	0	0	0	0	0
8	4	0	0	0	0	0	0

BackMIC

	2	3	4	5	6	7	8
2	354	295	18	0	0	0	0
3	268	12	3	15	0	0	0
4	13	4	1	0	0	0	0
5	0	17	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0

Figure 1. Grid frequency distribution of the three algorithms for independent variable pairs. *y*-axis and *x*-axis are the number of bins. Data size $n = 100$, 1000 replicates.

correlation between a paired variable, a grid can be drawn on the scatterplot of these two variables to encapsulate this correlation. MIC is not affected by the distribution and correlation type of the variable pairs [7]

Given a paired variable (X, Y), $X \in R^n$, $Y \in R^n$, MIC of X and Y is defined as

$$\mathrm{MIC}(X, Y) = \max_{n_x \times n_y \leq B(n,\alpha)} \left\{ \frac{\max_G (I_G(X, Y))}{\log_2 \min(n_x, n_y)} \right\},$$

where n_x and n_y are the number of bins on the *x*-axis and *y*-axis, respectively. G represents a $n_x \times n_y$ grid on (X, Y), $I_G(X, Y)$ denotes the mutual information under the grid G. $\mathrm{B}(n, \alpha)$ is a function of data size n and is equal to n^α $(0 < \alpha < 1)$, which limits the maximum number of bins. $\log_2 \min(n_x, n_y)$ is a normalization term to ensure MIC in the range of 0 to 1. MIC converges to 0 as data size $n \to \infty$ when X and Y are statistically independent; the MIC increases as the correlation between X and Y strengthens.

The calculation of MIC is computationally intensive. To estimate MIC, we need to search for the 'optimal grid' on the scatterplot of X and Y to separate the 'blank space' and 'data points' by using the least number of bins possible. With the same number of bins, the grid improves as the MIC increases. Several algorithms can be used for the approximate estimation of MIC. For ApproxMaxMI (AppMIC, downloaded from http://www.exploredata.net/Downloads/MINE-Application), proposed by Reshef *et al.* [7], one axis (the *y*-axis or the *x*-axis) is equipartitioned (an *equipartition* is a partition into either rows or columns such that each row/column contains the same number of data points), while the other axis is partitioned by dynamic programming algorithm under the limitation of $\mathrm{B}(n, \alpha)$ (α is often set to 0.55 or 0.6; in this paper, α is 0.6). However, AppMIC algorithm features two problems. First, the equipartition of one axis is neither sufficient nor necessary for MIC estimation [12]. Second, the MIC values are usually obtained when the number of bins $n_x \times n_y$ reaches $\mathrm{B}(n, \alpha)$, so the generality of MIC is closely related to $\mathrm{B}(n, \alpha)$. If a low $\mathrm{B}(n, \alpha)$ is set, then the MIC can only capture simple correlation patterns; by contrast, a high $\mathrm{B}(n, \alpha)$ will cause a non-zero score even for independent variables [7]. To solve this problem, Chen *et al.* [13] proposed the ChiMIC algorithm (downloaded from https://github.com/chenyuan0510/Chi-MIC), in which one axis is equipartitioned, and the partition of other axis is terminated by the χ^2-test. To avoid the number of bins on the equipartitioned axis always reaching $\mathrm{B}(n, \alpha)/2$, the ChiMIC algorithm replaces $\log_2 \min(n_x, n_y)$ in the MIC definition with $\log_2 n_{\mathrm{equ}}$, where n_{equ} is the number of bins on the equipartitioned axis and $\log_2 n_{\mathrm{equ}} \geq \log_2 \min(n_x, n_y)$. However, whether the normalization term can be modified remains unclear, and the equipartition restriction remains unresolved in the ChiMIC algorithm.

In this study, we proposed an improved approximation algorithm called BackMIC for MIC estimation. This algorithm adds a searching back process on the equipartitioned axis to remove the restriction of equipartition and control the search process based on the χ^2-test for both the *y*- and *x*-axes. Results on simulated and real data demonstrated that the BackMIC algorithm exhibits better performance in measuring the correlations between independent and dependent variable pairs compared with the AppMIC and ChiMIC algorithms.

2. Results and discussion

2.1. Comparison of grids and estimated MICs for independent variable pairs

The expected grid for independent variable pairs is 2 × 2 [13]. Figure 1 shows the grid frequency distribution obtained by the AppMIC, ChiMIC and BackMIC algorithms for computing the MIC values of independent variable pairs at 1000 repetitions. When data size $n = 100$ and $\mathrm{B}(n, \alpha) = 16$,

Table 1. MICs estimated by different algorithms for various data sizes. The MIC values are represented as the average value ± the standard deviation of 1000 repetitions.

data size	AppMIC	ChiMIC	BackMIC
250	0.1813 ± 0.0234	0.0390 ± 0.0284	0.0556 ± 0.0180
500	0.1545 ± 0.0152	0.0188 ± 0.0165	0.0342 ± 0.0104
1000	0.1051 ± 0.0103	0.0119 ± 0.0092	0.0204 ± 0.0065
2000	0.0730 ± 0.0070	0.0066 ± 0.0053	0.0119 ± 0.0032
4000	0.0490 ± 0.0040	0.0041 ± 0.0035	0.0069 ± 0.0018
10 000	0.0383 ± 0.0024	0.0020 ± 0.0013	0.0032 ± 0.0008
20 000	0.0221 ± 0.0011	0.0010 ± 0.0009	0.0018 ± 0.0005

Table 2. Thirteen functional correlations.

functions	X	$f(X)$
line	[0, 1]	X
parabolic	[−0.5, 0.5]	$4X^2$
cubic	[−1.3, 1.1]	$4X^3 + X^2 - 4X$
exponential	[0, 10]	2^X
non-Fourier freq [low]	[0, 1]	$\cos(4\pi X)$
non-Fourier freq [medium]	[0, 1]	$\cos(8\pi X)$
non-Fourier freq [high]	[0, 1]	$\cos(12\pi X)$
linear + period freq [low]	[0, 1]	$(1/5)\sin(4(2X-1)) + (11/10)(2X-1)$
linear + period freq [medium]	[0, 1]	$\sin(10\pi X) + X$
linear + period freq [high]	[0, 1]	$\frac{1}{10}\sin(10.6(2X-1)) + \frac{11}{10}(2X-1)$
varying freq [low]	[0, 1]	$\cos(4\pi X\ (1+X))$
varying freq [medium]	[0, 1]	$\cos(8\pi X\ (1+X))$
varying freq [high]	[0, 1]	$\cos(12\pi X\ (1+X))$

almost all the grids of the AppMIC algorithm were concentrated in 2 × 8 and 8 × 2. Most grids of the ChiMIC algorithm were concentrated in $a \times b$, where $a \leq 5$ and $b \leq 5$. By contrast, almost all the grids of the BackMIC algorithm were concentrated in 2 × 2, 2 × 3 and 3 × 2; the most frequent grid was 2 × 2. The grids of the BackMIC algorithm were the closest to the expected one.

The MIC of independent variable pairs should converge to 0 as data size $n \rightarrow \infty$ [7]. Table 1 shows the MIC values estimated by the three algorithms (i.e. AppMIC, ChiMIC and BackMIC) for different data sizes. BackMIC was similar to ChiMIC without expanding the normalization term, which was closer to zero than AppMIC under the same data size.

2.2. Comparison of grids and estimated MICs for dependent variable pairs

The MIC of variable pairs with noiseless functional correlations should be 1 [7]. We used the BackMIC algorithm to calculate the MIC values of 13 pairs of noiseless functional correlations (table 2). All MIC values were 1, indicating that the BackMIC algorithm maintained the generality of MIC.

For the 'I'-type dataset, the expected grid and MIC were 3 × 3 and 0.2561, respectively (figure 2*a*). The grids of the AppMIC, ChiMIC and BackMIC algorithms were 14 × 3, 3 × 3 and 3 × 3, respectively, and the estimated MICs were 0.2457, 0.1808 and 0.2561, respectively (figure 2*b–d*). Only the grid and estimated value of the BackMIC algorithm were in line with expectations. For the chequerboard dataset, the expected grid and MIC were 5 × 5 and 0.3835, respectively (figure 2*e*). The grids of the AppMIC, ChiMIC and BackMIC algorithms were 9 × 5, 5 × 6 and 5 × 6, respectively, and the estimated

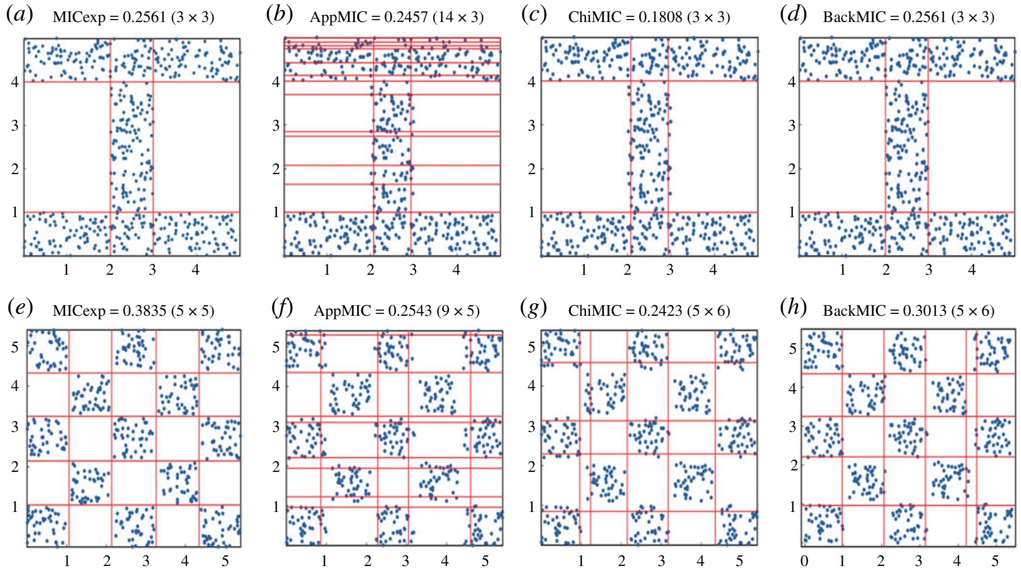


Figure 2. Comparison of the grids and estimated MICs for the 'I'-type and chequerboard datasets. Data size $n = 520$; each square contains 40 data points.

MICs were 0.2543, 0.2423 and 0.3013, respectively (figure 2*f*–*h*). The grid and estimated MIC of the BackMIC algorithm were the closest to expectations relative to those of the other algorithms.

The actual MIC values of noisy linear, parabolic and sinusoidal correlations were difficult to determine, as shown in figure 3. We compared the three algorithms under the criterion that 'with the same number of bins, the larger the MIC values, the better the grid'. The AppMIC algorithm was excluded from the comparison because of its excessive bins (all reached 2 × 21 for data size $n = 500$ and $B(n, \alpha) = 42$; figure 3*a*, *d* and *g*). A comparison of the MIC values of the three functional correlations obtained by the ChiMIC and BackMIC algorithms revealed that the latter always achieved higher MIC values with the same or fewer number of bins than the former. Thus, the grid and estimated MICs obtained by the BackMIC algorithm were more reasonable than those by the ChiMIC algorithm; no axis was equipartitioned.

2.3. Comparison of robustness

The correlation strength of a given variable pair should be certain. However, as shown in figure 4, AppMICs varied with the change of α in $B(n, \alpha)$ for noisy linear, parabolic and sinusoidal correlations, because more bins generally result in larger MIC values. BackMICs remained almost constant because the χ^2-test was used instead of $B(n, \alpha)$ to terminate grid optimization. Therefore, the BackMIC algorithm was more robust in measuring the correlation between variables than the AppMIC algorithm. The variation in ChiMICs was not obvious either; however, ChiMICs were always lower than BackMICs due to the equipartition restriction and harsh normalization term in the ChiMIC algorithm.

2.4. Comparison of statistical power

Statistical power refers to the probability of correctly accepting the alternative hypothesis in the hypothesis test [14]. As the statistical power increases, the probability of making type II error decreases [15,16].

For the null hypothesis of statistical independence, for each dataset, statistical power is computed on the dependent variable pairs as well as on independent variable pairs; the statistical power of each statistic is defined as the fraction of dependent variable pairs yielding a statistic value greater than 95% (significance level is 0.05) of the values yielded by the independent variable pairs [13]. The statistical power of AppMIC, ChiMIC and BackMIC for the above five functional correlations (figures 2 and 3) at different noise amplitudes is shown in figure 5. The power of BackMIC was significantly higher than those of AppMIC and ChiMIC.

2.5. Comparison of equitability

If a statistic assigns similar scores to equally noisy correlations of different types, then the statistic has the property of equitability [17]. Equitability allows us to specify a threshold correlation strength below

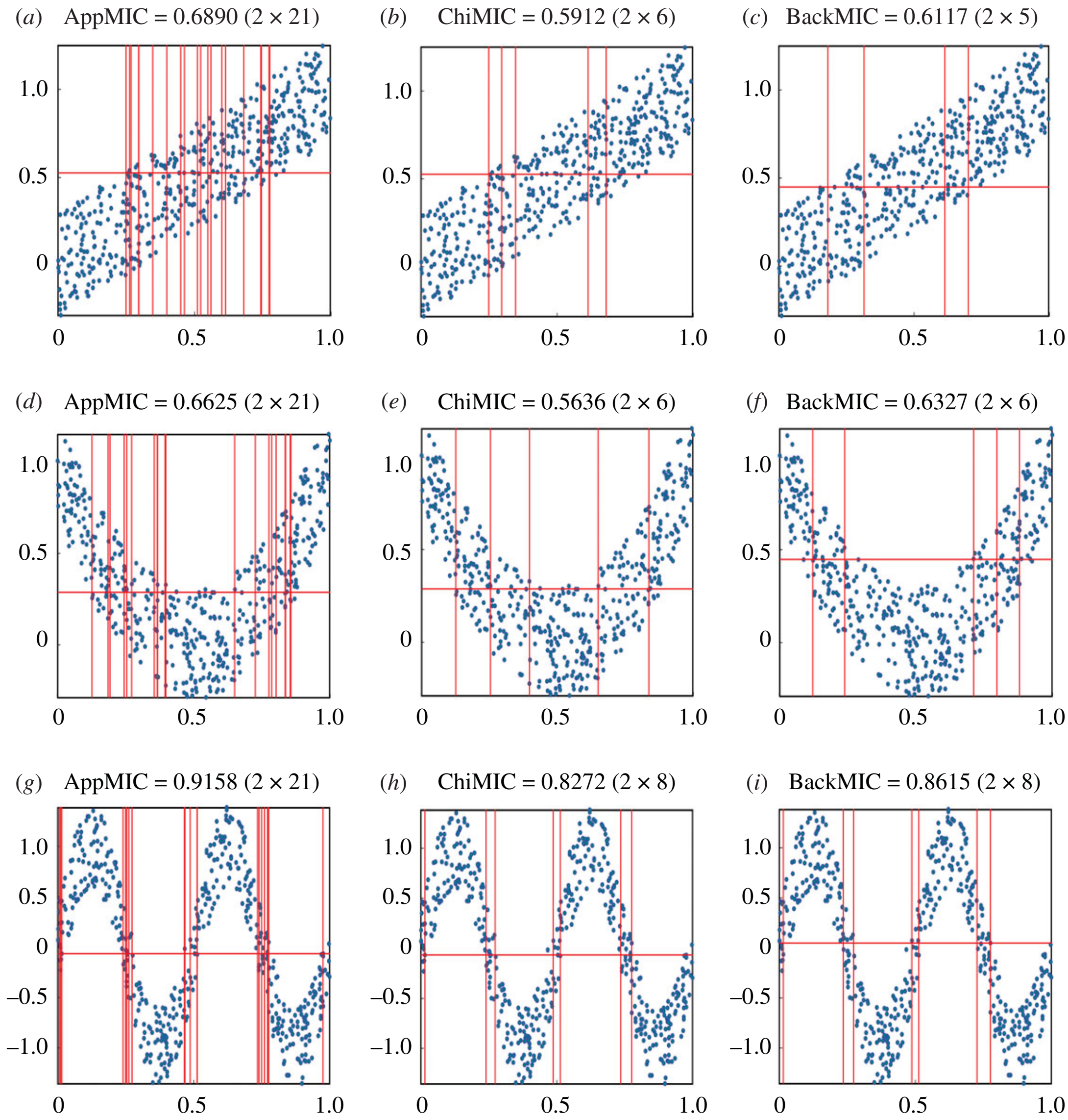


Figure 3. Comparison of the grids and estimated MICs for linear (*a*–*c*), parabolic (*d*–*f*) and sinusoidal (*g*–*i*) correlations. The data points are simulated from $Y = f(X) + \zeta$, where $f(X) = X$, $4(X - 0.5)^2$ and $\sin(4\pi X)$, respectively. X and ζ are random variables drawn uniformly from (0, 1) and (−0.3, 0.3), respectively; data size $n = 500$.

which we are uninterested and search for correlations whose strength is greater than the threshold [18]. Perfect equitability does not exist [19]. We tested the approximate equitability of AppMIC, ChiMIC and BackMIC on the basis of 13 functional correlations listed in table 2. For each correlation, we analyse the equitability by generating a noiseless data sequence with a data size $n = 500$ and 301 data series with noise ε added to $f(X)$, where ε is a uniformly distributed random variable from $-b$ to b, and b denotes the noise levels selected from [0, 3] with step size of 0.01. The results confirmed that the approximate equitability of BackMIC was better than those of AppMIC and ChiMIC (figure 6).

2.6. Comparison of computational cost

We compared the computational time of the AppMIC, ChiMIC and BackMIC algorithms by using different data sizes for two types of variable pairs (the independent variable pairs and the parabolic functional variable pairs with noise level of 0.5). The results in table 3 showed that the ChiMIC algorithm ran faster than the AppMIC algorithm because the former used the χ^2-test to terminate grid optimization earlier than the latter. The running time of the BackMIC algorithm was almost twice as much as that of the ChiMIC algorithm, because the BackMIC algorithm added a searching back process for an optimal partition on the originally equipartitioned axis compared with the ChiMIC algorithm. Compared with the AppMIC algorithm, the BackMIC algorithm was slower when the data size was small. As the data

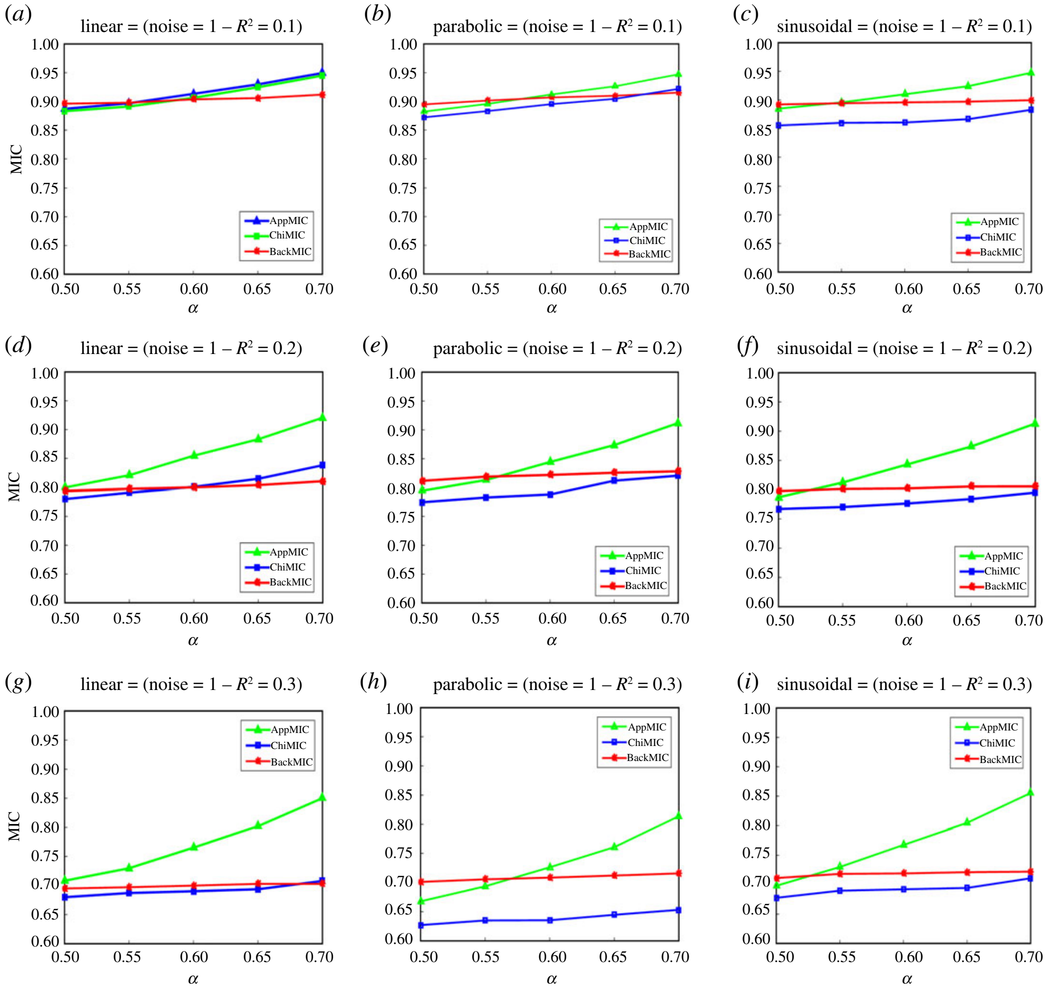


Figure 4. Comparison of AppMIC, ChiMIC and BackMIC for linear, parabolic and sinusoidal correlations at different α in B(n, α). R^2 in ‘noise’ is the squared Pearson correlation coefficient of $f(X)$ and Y, where $f(X)$ is the same as in figure 3 and $Y = f(X) + \zeta$. For linear and parabolic correlations, ζ is drawn uniformly at noise levels 0.2, 0.35 and 0.5, respectively; for sinusoidal correlation, ζ is drawn uniformly at noise levels 0.6, 0.9 and 1.2, respectively. The MIC values were the average of 500 repetitions with data size $n = 500$.

size increased, the BackMIC algorithm was able to catch up with the AppMIC algorithm. In terms of independent variable pairs, the BackMIC algorithm ran faster than the AppMIC algorithm.

2.7. Comparison of AppMIC, ChiMIC and BackMIC applied in clustering for cancer classification

Given that cancer samples and normal samples have different gene expression levels [20], clustering algorithms such as K-means locate samples into different clusters on the basis of the similarity (distance) between samples; these clusters can then be used for cancer classification [21,22]. To evaluate the performance of MIC obtained by the AppMIC, ChiMIC and BackMIC algorithms, we replaced the Euclidean distance with (1-MIC) as the distance measurement between two samples in the K-means algorithm. Cancer gene expression datasets GSE37023, GSE29272 and GSE35602 were used in our work. Table 4 shows that the purity and Rand index (RI) of clustering results based on (1-BackMIC) were higher than those based on (1-AppMIC) and (1-ChiMIC), and they were even better than those based on Euclidean distance in the original K-means.

3. Datasets

3.1. Simulated data

Five dependent correlations used in figures 2 and 5 are defined in table 5. (X_0,Y_0) and (X_1,Y_1) are pairs of random variables drawn uniformly from the solid squares of a ‘I’-type pattern and 5 × 5 chequerboard

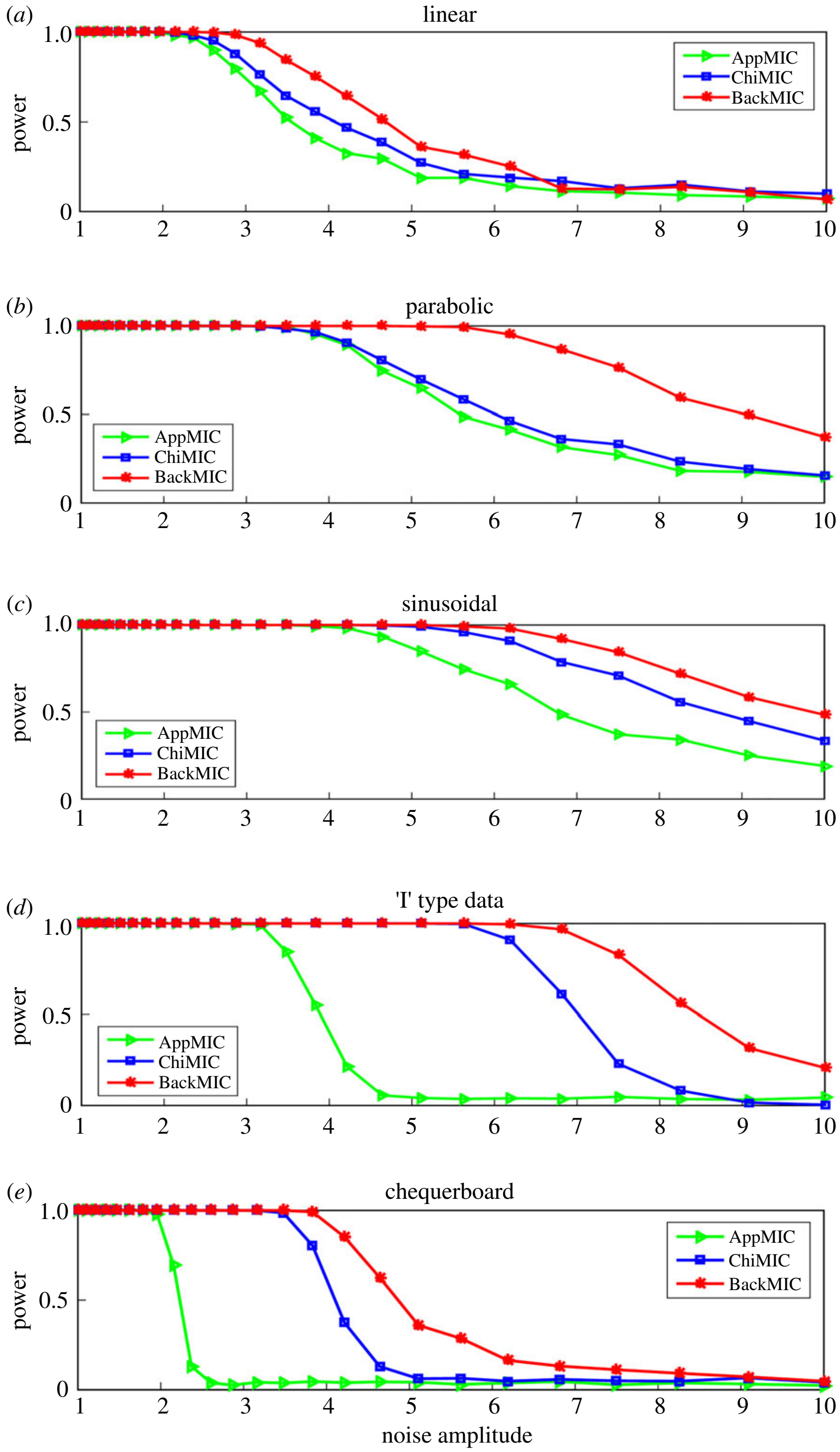


Figure 5. Statistical power of AppMIC, ChiMIC and BackMIC at different noise amplitudes. The statistical power was estimated via 500 simulations with data size $n = 500$.

[18], where each square has sides with a length of 1. In figure 2, a and η are equal to 0 in 'I'-type pattern and 5 × 5 chequerboard. In figure 5, the noise aptitude a for the statistical power calculation is 25 noise amplitudes with logarithmic distribution from 1 to 10; X, ξ and η are random variables drawn from the normal distribution N (0,1).

In figure 4, the noise added to linear, parabolic and sinusoidal functional correlations is defined as $Y = f(X) + \text{noise\_level} \times (2\text{rand}(n,1) - 1)$, where rand($n$,1) is used to generate n uniformly distributed numbers in [0, 1].

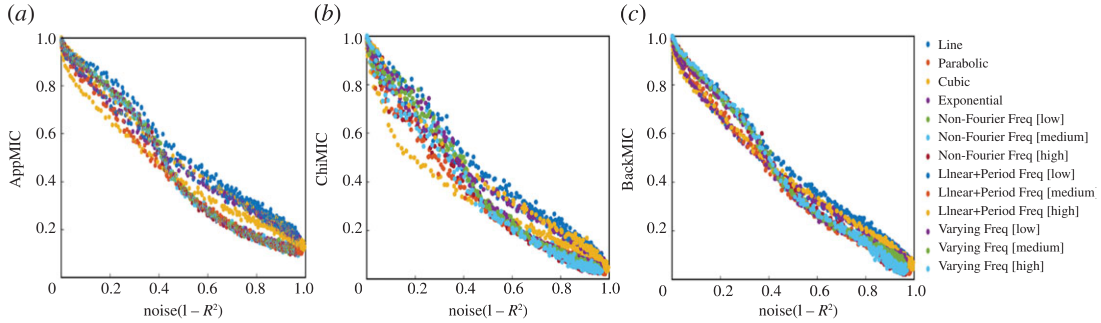


Figure 6. Equitability of AppMIC, ChiMIC and BackMIC. These values are the average of 500 repetitions with data size $n = 500$.

Table 3. Running time for calculating MICs for different data sizes. The corresponding time was represented as the average value ± standard deviation over 100 repetitions on a Windows 7 64-bit operating system (RAM: 8.00 GB, CPU: 2.60 GHz).

data	independent variable pair			parabolic function (noise level = 0.5)		
size	AppMIC	ChiMIC	BackMIC	AppMIC	ChiMIC	BackMIC
250	0.0184 ± 3×10^{-3}	0.0130 ± 1×10^{-3}	0.0240 ± 3×10^{-3}	0.0226 ± 3×10^{-3}	0.0147 ± 3×10^{-3}	0.0299 ± 4×10^{-3}
500	0.0401 ± 0.01	0.0357 ± 0.01	0.0683 ± 0.02	0.0605 ± 0.01	0.0508 ± 0.01	0.1001 ± 0.01
1000	0.3172 ± 0.04	0.1026 ± 0.01	0.1774 ± 0.04	0.2225 ± 0.03	0.2114 ± 0.03	0.3856 ± 0.03
2000	1.4335 ± 0.10	0.6163 ± 0.13	1.0452 ± 0.20	0.9633 ± 0.10	0.8694 ± 0.07	1.5492 ± 0.16
4000	6.7017 ± 0.41	2.4343 ± 0.38	4.2755 ± 0.49	3.7833 ± 0.35	3.0101 ± 0.15	4.7625 ± 0.48

Table 4. Performance of different distance measurements for K-means clustering. Bold items highlight the highest prediction accuracy.

datasets	methods	K-means (Euclidean)	K-means (1-AppMIC)	K-means (1-ChiMIC)	K-means (1-BackMIC)
GSE37023^1	purity	0.7815	0.8940	0.9007	**0.9338**
	RI	0.6539	0.8061	0.8167	**0.8780**
GSE37023^2	purity	0.7385	0.8769	0.8000	**0.9077**
	RI	0.6077	0.8048	0.6750	**0.8298**
GSE29272	purity	0.5075	0.9366	0.9478	**0.9627**
	RI	0.4982	0.8807	0.9073	**0.9141**
GSE35602	purity	**1.0000**	0.9198	0.9412	**1.0000**
	RI	**1.0000**	0.8342	0.8859	**1.0000**

3.2. Real datasets

We used GSE37023 [23], GSE29272 [24] and GSE35602 [25] to verify the reliability of BackMIC, as described in table 6. In GSE37023, two datasets from platforms GPL96 and GPL97 were used respectively, denoted as GSE37023^1 and GSE37023^2. In GSE35602, only the dataset from the platform GPL6480 was used. All the datasets were obtained from the Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/). Probe IDs were converted to gene symbols according to GEO platform (GPL). If several probes were mapped to the same gene symbol, their average value was taken as the expression value of this gene. An implementation of the BackMIC algorithm can be downloaded at https://github.com/Caodan82/BackMIC.

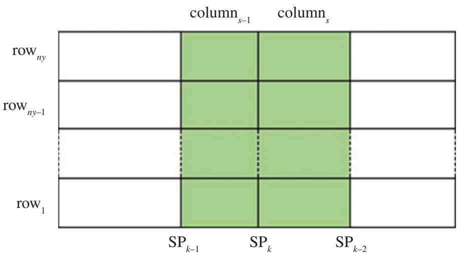


Figure 7. Instruction for the calculation of the χ^2 statistic [13].

Table 5. Five dependent datasets simulated for statistical power estimation.

dependent correlations	X	Y
'l'-type data	X_0	$Y_0 + a\eta$
chequerboard	X_1	$Y_1 + a\eta$
linear	ξ	$X + a\eta$
parabolic	ξ	$4(X - 0.5)^2 + a\eta$
sinusoidal	ξ	$\sin(4\pi X) + a\eta$

Table 6. Four cancer gene expression datasets.

GEO accession	no. genes	no. classes	description
GSE37023[1]	13 515	2	tumours (112) and non-malignant (39)
GSE37023[2]	11 349	2	tumours (29) and non-malignant (36)
GSE29272	13 515	2	cancer (134) and normal (134)
GSE35602	19 595	2	cancer (26) and normal (8)

4. Methods

4.1. χ^2-test to terminate grid optimization

The BackMIC algorithm uses the χ^2-test to terminate grid optimization, which is similar to the ChiMIC algorithm. Given an optimal segment point, if the p-value of the χ^2-test for the optimal segment point is lower than a given threshold (here, threshold = 0.01), the segment point is valid and the BackMIC algorithm continues to search for the next optimal segment point. Otherwise, the BackMIC algorithm stops its searching process.

We take the kth optimal segment point SP_k on the x-axis as an example. Suppose that the y-axis is partitioned into n_y bins, and SP_k is between SP_{k-2} and SP_{k-1}. The x-axis is partitioned into $(s-1)$th and sth columns (figure 7). The coloured $n_y \times 2$ contingency table in figure 7 is called the detection area of SP_k for the χ^2-test. The χ^2 statistic is defined as follows [13]:

$$\chi^2_{\text{SP}_k} = \sum_{j=1}^{n_y} \sum_{i=s-1}^{s} \frac{(n_{j,i} - n_{*,i} n_{j,*}/N_d)^2}{n_{*,i} n_{j,*}/N_d},$$

where $n_{j,i}$ is the number of data points in the bin of row_j and column_i, $n_{*,i}$ is the number of data points in the bins of column_i, $n_{j,*}$ is the number of data points in the bins of row_j of the detection area, and N_d is the total number of data points in the bins of the detection area. If $n_y = 2$, the χ^2 statistic needs to be corrected according to the following formula [26]:

$$\chi^2_{\text{SP}_k} = \sum_{j=1}^{n_y} \sum_{i=s-1}^{s} \frac{(|n_{j,i} - n_{*,i} n_{j,*}/N_d| - 0.5)^2}{n_{*,i} n_{j,*}/N_d}.$$

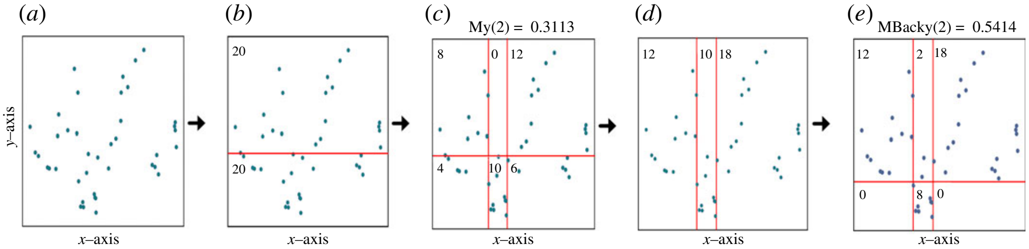


Figure 8. Simulation process of the BackMIC algorithm. (*a*) The scatterplot of a paired variable with data size $n = 40$; (*b*) equipartition the *y*-axis into 2 bins; (*c*) search for the optimal partition on the *x*-axis to obtain the largest normalized mutual information $\mathrm{M}_y(2) = 0.3113$; (*d*) fix the partition of the *x*-axis obtained in (*c*); (*e*) search back for the optimal partition on the *y*-axis to obtain the largest normalized mutual information $\mathrm{MBack}_y(2) = 0.5414$.

4.2. BackMIC algorithm

The BackMIC algorithm involves two phases to obtain the optimal partition on the scatterplot of a variable pair (see algorithm 1). First, based on an equipartition of n_y bins on one axis (we take the *y*-axis as example), the BackMIC algorithm locates an optimal partition of the *x*-axis through the dynamic programming algorithm to achieve the largest normalized mutual information under the restriction of the χ^2-test, which is similar to the ChiMIC algorithm [13]. Unlike the AppMIC and ChiMIC algorithms, the BackMIC algorithm fixes the partition of the *x*-axis obtained in the previous step and searches back for the optimal partition of the *y*-axis instead of equipartitioning. Therefore, the BackMIC algorithm controls the bins of both the *y*- and *x*-axes by the χ^2-test and obtains unequipartitioning for both the *y*- and *x*-axes. For $n_y = 2$, the simulation process of the BackMIC algorithm is shown in figure 8. From the results we find that, compared with equipartitioning of the *y*-axis, when the *y*-axis is unequipartitioned, larger normalized mutual information can be obtained (0.5414 versus 0.3113).

Algorithm 1 BackMIC algorithm

Input: $(X, Y) = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is a pair of ordered variables, $B=n^{0.6}$, threshold=0.01

```
for $n_y$=2 to ceil(B/2)
    EquipartitionOf$y$ into $n_y$ rows; PartionOf$x$ $\leftarrow\varphi$; BackPartionOf$y$$\leftarrow\varphi$
    for $n_x$=2 to $n$
        search for SegPointOf$x_{nx}$ maximizing $\mathrm{I_G}(X, Y)/\log_2 \min(n_x, n_y)$;
        Obtain $\chi^2_{\mathrm{SegPointOf}x_{nx}}$ and $p$-value$_{\mathrm{SegPointOf}x_{nx}}$;
        if $p$-value$_{\mathrm{SegPointOf}x_{nx}}$ > threshold then exit for end if;
        PartionOf$x$ $\leftarrow$ PartionOf$x$ $\cup${ SegPointOf$x_{nx}$ };
    end for
    $\mathrm{M}_y(n_y)\leftarrow\mathrm{I_G}(X, Y)/\log_2 \min(\mathrm{length(PartionOf}x)+1, n_y)$;
    for $s$ =2 to $n$
        Fix PartionOf$x$, search for SegPointOf$y_s$ maximizing $\mathrm{I_G}(X, Y)/\log_2 \min$
      ($s$,length(PartionOf$x$)+1);
        obtain $\chi^2_{\mathrm{SegPointOf}y_s}$ and $p$-value$_{\mathrm{SegPointOf}y_s}$;
        if $p$-value$_{\mathrm{SegPointOf}y_s}$ > threshold then exit for end if;
        BackPartionOf$y$ $\leftarrow$ BackPartionOf$y$ $\cup$ { SegPointOf$y_s$ };
    end for
    $\mathrm{MBack}_y(n_y)\leftarrow\mathrm{I_G}(X,Y)/\log_2 \min(\mathrm{length(BackPartionOf}y)+1, \mathrm{length(PartionOf}x)+1)$;
end for
Switch the axes, repeat the above steps, and obtain $\mathrm{M}_x$ and $\mathrm{MBack}_x$.
```

Output **MIC = max $\{\mathrm{M}_x, \mathrm{MBack}_x, \mathrm{M}_y, \mathrm{MBack}_y\}$**

4.3. K-means clustering algorithm

Suppose there are M samples with K classes in a dataset, $C = \{c_1, c_2, \ldots, c_K\}$ is the sample set of real classifications, and c_i is the sample set of the *i*th class (given that four gene expression datasets are of binary class, K is 2 here). The K-means clustering algorithm is performed using Matlab scripts downloaded at http://people.revoledu.com/kardi/tutorial/kMean/download.htm [27], which proceeds by randomly selecting K initial clustering centres and then assigns each sample to the nearest clustering centre [28]. For each dataset, we used the top 1000 genes with the largest variance to calculate the distance between samples. Suppose that the clustering results of K-means is $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$, where ω_i is set of the *i*th cluster.

Two commonly used evaluation criteria of clustering algorithms, namely, purity and RI, were used in this paper. Purity is the proportion of correctly clustered samples in total samples, and it can be calculated by [29]

$$\text{Purity} = \frac{1}{M}\sum_{i}\max_{j}|c_i \cap \omega_j|.$$

RI refers to the proportion of concordant sample pairs in the total number of sample pairs [30]. A is the number of sample pairs placed in the same group in both C and Ω, and B is the number of sample pairs placed in different groups in both C and Ω. RI is defined as follows [31]:

$$\text{RI} = \frac{A+B}{M(M-1)}.$$

5. Conclusion

In this paper, we introduced the BackMIC algorithm for better MIC estimation. The BackMIC algorithm added a searching back process to obtain an optimal partition for the equipartitioned axis, making it more likely to obtain the true MIC value. Meanwhile, the BackMIC algorithm used the χ^2-test to ensure that each introduced optimal segment point can significantly increase the MIC value. This effectively avoided unreasonable grid refinement and made MIC value independent of B(n, α) which improved the robustness of MIC.

The results on simulation data showed that, compared with the AppMIC and ChiMIC algorithms, the BackMIC algorithm can effectively reduce the MIC value of independent variable pairs without expanding the normalization term in MIC definition; if there is a functional correlation between variable pairs, the MIC calculated by the BackMIC algorithm is equal to 1, maintaining the generality of MIC; if there is noisy correlation between variable pairs, the BackMIC algorithm usually obtained larger MIC values with less bins; moreover, the statistical power and equitability of MIC calculated by the BackMIC algorithm are better. When applying the (1-MIC) value as the distance measurement between cancer samples and normal samples in K-means algorithm, experiments on four cancer datasets showed that the MIC values calculated by the BackMIC algorithm can obtain better clustering results. All evidence verified that the BackMIC algorithm improves MIC estimation.

Data accessibility. GSE37023 is at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37023. GSE29272 is at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29272. GSE35602 is at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35602. An implementation of the BackMIC algorithm is stored in GitHub: https://github.com/Caodan82/BackMIC, and has been archived within the Zenodo repository: https://zenodo.org/record/4280173#.X7X3z_nDpw8 (https://doi.org/10.5281/zenodo.4280173).

Authors' contributions. D.C., Y.C. and Z.Y. contributed to the design of algorithm, D.C. carried out the analysis and drafted the manuscript, J.C. and H.Z. helped draft the manuscript, all authors reviewed and gave final approval for submission.

Competing interests. We have no competing interests.

Funding. This research was supported by the National Natural Science Foundation of China (grant no. 61701177), the Scientific Research Foundation of Education Office of Hunan Province, China (grant nos. 17A096 and 18A105).

References

1. Alizadeh AA *et al.* 2020 Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511. (doi:10.1038/35000501)
2. Dharmarajan A, Velmurugan T. 2015 Lung cancer data analysis by k-means and farthest first clustering algorithms. *Indian J. Sci. Technol.* **8**, 1–8. (doi:10.17485/ijst/2015/v8i15/73329)
3. Acquisti A, Gross R. 2009 Predicting social security numbers from public data. *Proc. Natl Acad. Sci. USA* **106**, 10 975–10 980. (doi:10.1073/pnas.0904891106)
4. Panori A, Psycharis Y. 2019 Exploring the links between education and income inequality at the municipal level in Greece. *Appl. Spat. Anal. Policy* **12**, 101–126. (doi:10.1007/s12061-017-9234-6)
5. Almog A, Shmueli E. 2019 Structural entropy: monitoring correlation-based networks over time with application to financial markets. *Sci. Rep.-UK* **9**, 10 832–10 844. (doi:10.1038/s41598-019-47210-8)
6. Wadhwani SB. 1999 The US stock market and the global economic crisis. *Natl Inst. Econ. Rev.* **167**, 86–105. (doi:10.1371/journal.pone.0067721)
7. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. 2011 Detecting novel associations in large data sets. *Science* **334**, 1518–1524. (doi:10.1126/science.1205438)
8. Lyu H, Wan M, Han J, Liu R, Wang C. 2017 A filter feature selection method based on the Maximal Information Coefficient and Gram-Schmidt Orthogonalization for biomedical data

mining. *Comput. Biol. Med.* **89**, 264–274. (doi:10.1016/j.compbiomed.2017.08.021)

9. Liu HM, Rao N, Yang D, Yang L, Li Y, Ou F. 2014 A novel method for identifying SNP disease association based on maximal information coefficient. *Genet. Mol. Res.* **13**, 10 863–10 877. (doi:10.4238/2014.December.19.7)
10. Iuchi H, Sugimoto M, Tomita M. 2018 MICOP: maximal information coefficient-based oscillation prediction to detect biological rhythms in proteomics data. *BMC Bioinf.* **19**, 249. (doi:10.1186/ s12859-018-2257-4)
11. Souza RSD, Maio U, Biffi V, Ciardi B. 2018 Robust PCA and MIC statistics of baryons in early minihaloes. *Mon. Not. R. Astron. Soc.* **440**, 240–248. (doi:10.1093/mnras/stu274)
12. Zhang Y, Jia S, Huang H, Qiu J, Zhou C. 2014 A novel algorithm for the precise calculation of the maximal information coefficient. *Sci. Rep.-UK* **4**, 6662. (doi:10.1038/ srep06662)
13. Chen Y, Zeng Y, Luo F, Yuan Z. 2016 A new algorithm to optimize maximal information coefficient. *PLoS ONE* **11**, e0157567. (doi:10. 1371/journal.pone.0157567)
14. Lenth RV. 2007 Statistical power calculations. *J. Anim. Sci.* **85**, E24–EE9. (doi:10.2527/jas. 2006-449)
15. Moher D, Dulberg CS, Wells GA. 1994 Statistical power, sample size, and their reporting in randomized controlled trials. *J. Am. Med. Assoc.* **272**, 122–124. (doi:10.1097/00008506-199501000-00018)
16. Cohen J. 1992 Statistical power analysis. *Curr. Dir. Psychol. Sci.* **1**, 98–101. (doi:10.1111/1467-8721.ep10768783)
17. Reshef YA, Reshef DN, Sabeti PC, Mitzenmacher MM. 2020 Equitability, interval estimation, and statistical power. *Stat. Sci.* **35**, 202–217. (doi:10. 1214/19-sts719)
18. Kinney JB, Atwal GS. 2014 Equitability, mutual information, and the maximal information coefficient. *Proc. Natl Acad. Sci. USA* **111**, 3354–3359. (doi:10.1073/ pnas. 1309933111)
19. Reshef DN, Reshef YA, Mitzenmacher M, Sabeti PC. 2014 Cleaning up the record on the maximal information coefficient and equitability. *Proc. Natl Acad. Sci. USA* **111**, E3362–E3363. (doi:10.1073/pnas. 1408920111)
20. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW. 1997 Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272. (doi:10.1126/science.276.5316.1268)
21. Kim EY, Kim SY, Ashlock D, Nam D. 2009 MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinf.* **10**, 260. (doi:10.1186/1471-2105-10-260)
22. Dubey AK, Gupta U, Jain S. 2016 Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. *Int. J. Comput. Assist. Radiol. Surg.* **11**, 2033–2047. (doi:10.1007/s11548-016-1437-9)
23. Wu Y *et al.* 2013 Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. *Gut* **62**, 1100–1111. (doi:10.1136/ gutjnl-2011-301373)
24. Wang G *et al.* 2013 Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in China. *PLoS ONE* **8**, e63826. (doi:10.1371/journal.pone. 0063826)
25. Nishida N *et al.* 2012 Microarray analysis of colorectal cancer stromal tissue reveals upregulation of two oncogenic miRNA clusters. *Clin. Cancer Res.* **18**, 3054–3070. (doi:10.1158/ 1078-0432.CCR-11-1078)
26. Yates F. 1934 Contingency tables involving small numbers and the χ^2 test. *J. R. Stat. Soc. Suppl.* **1**, 217–235. (doi:10.2307/2983604)
27. Teknomo K. 2006 K-means clustering tutorial. *Medicine* **100**, 3.
28. Hartigan JA, Wong MA. 1979 Algorithm AS 136: a k-means clustering algorithm. *J. R. Stat. Soc. C-Appl.* **28**, 100–108. (doi:10.2307/2346830)
29. Manning CD, Raghavan P, Schütze H. 2008 *Introduction to information retrieval*, pp. 234–265. Cambridge, UK: Cambridge University Press.
30. Rand WM. 1971 Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850. (doi:10.1080/01621459. 1971.10482356)
31. Park HS, Jun CH. 2009 A simple and fast algorithm for K-medoids clustering. *Exp. Syst. Appl.* **36**, 3336–3341. (doi:10.1016/j.eswa.2008.01.039)