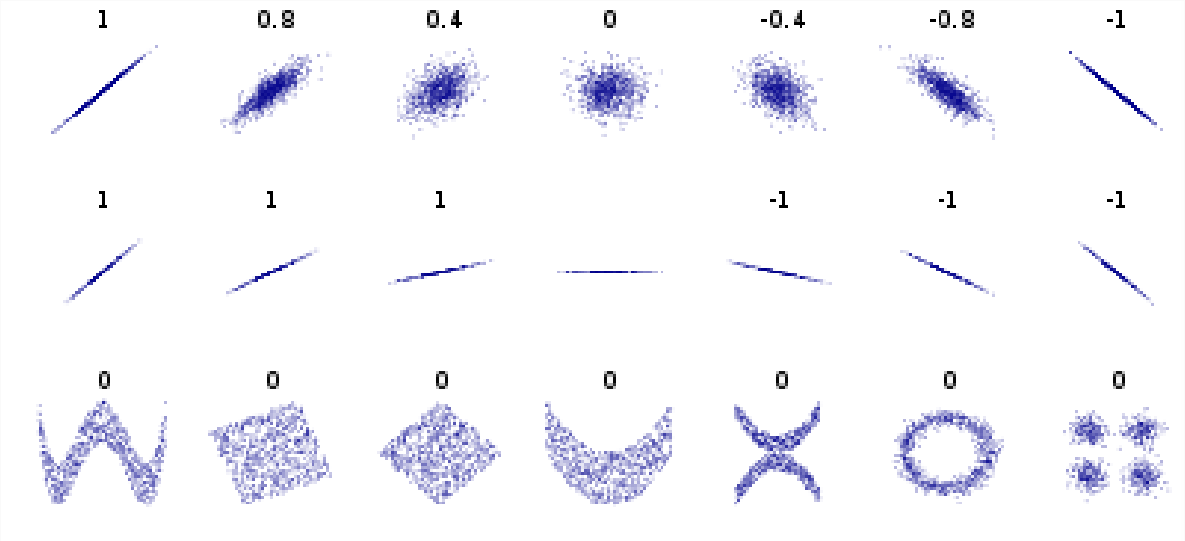
QUANTIFYING DEPENDENCE

Ella Kaye Kaspar Märtens Paul Vanetti Andi Wang

26th February 2016

Introduction

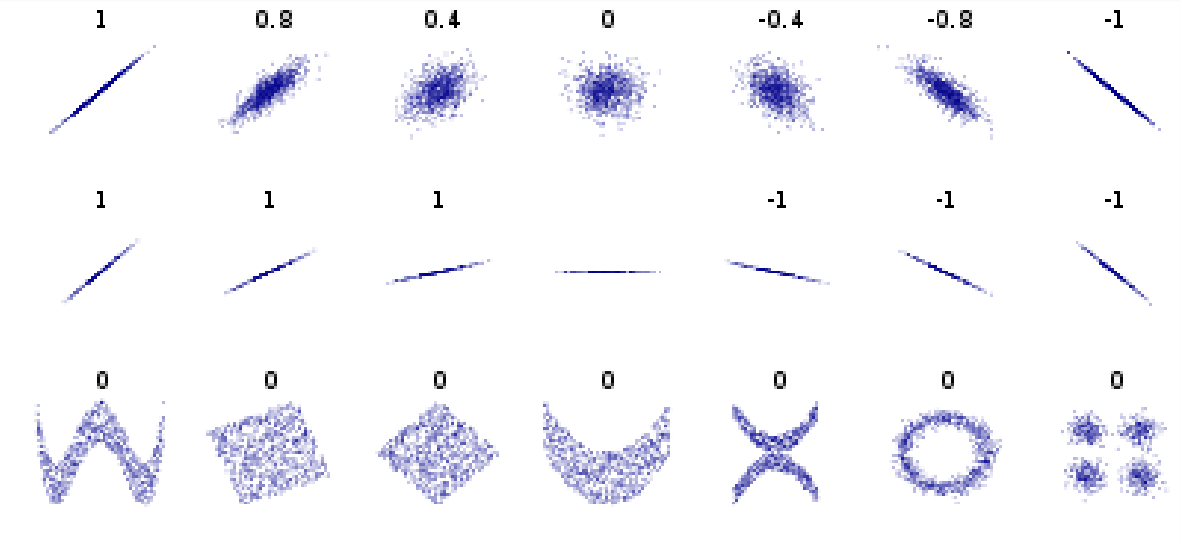
Goal: be able to quantify the level of dependence between two random variables.



https://en.wikipedia.org/wiki/Correlation_and_dependence

Introduction

Goal: be able to quantify the level of dependence between two random variables.
e.g. Pearson correlation coefficient, R^2.

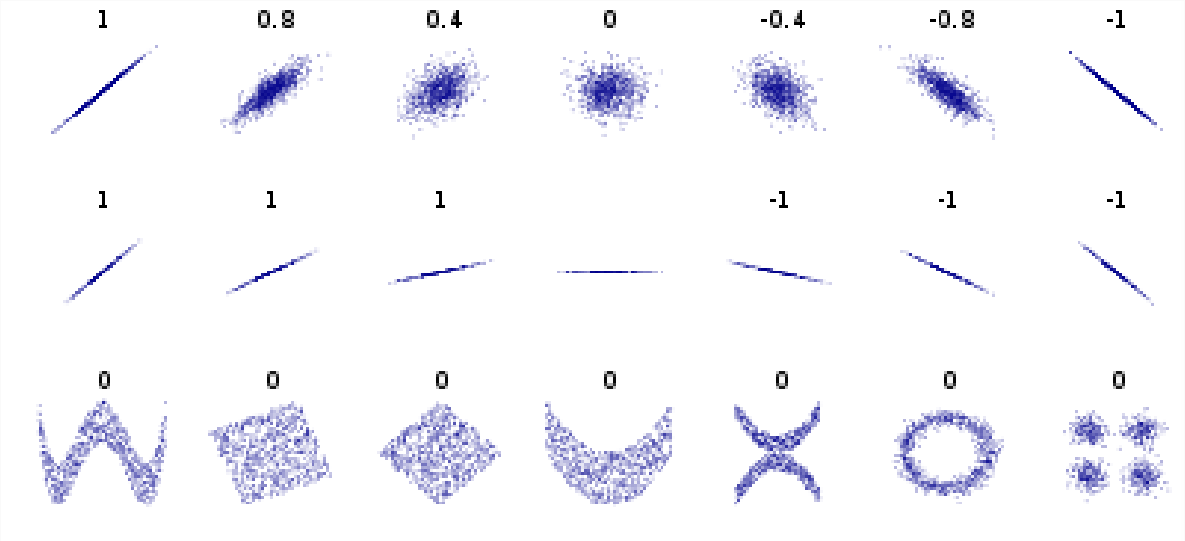


https://en.wikipedia.org/wiki/Correlation_and_dependence

Introduction

Goal: be able to quantify the level of dependence between two random variables.
e.g. Pearson correlation coefficient, R^2.

But this only captures *linear* dependence.



https://en.wikipedia.org/wiki/Correlation_and_dependence

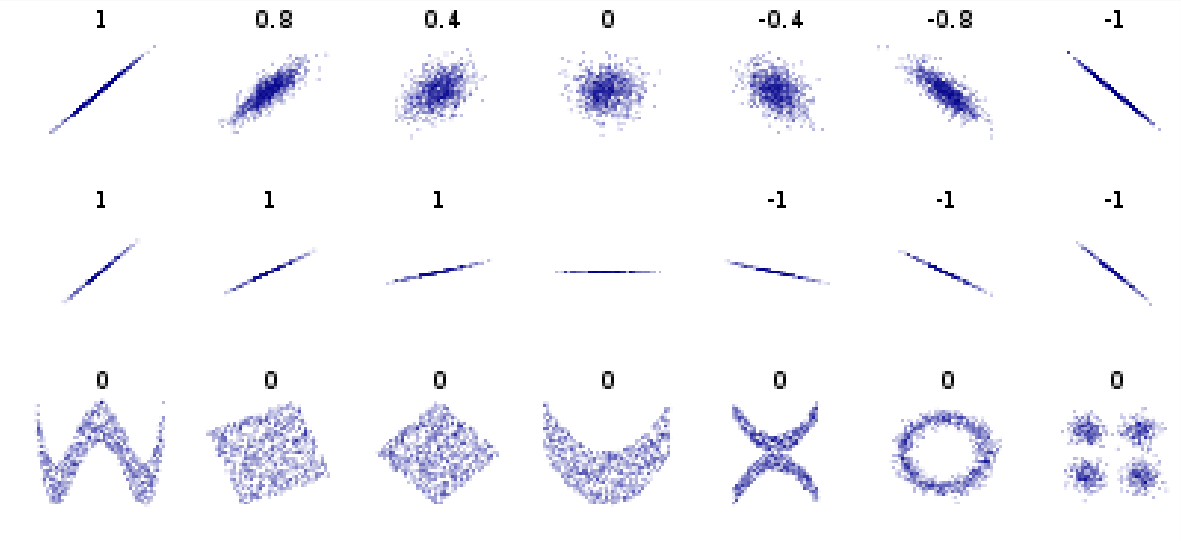
Introduction

Goal: be able to quantify the level of dependence between two random variables.

e.g. Pearson correlation coefficient, R^2.

But this only captures *linear* dependence.

Want an *equitable* statistic; no preference for a specific relationship.



https://en.wikipedia.org/wiki/Correlation_and_dependence

Mutual Information

$X, Y \sim p_{X,Y}$. Define *mutual information*

Mutual Information

$X, Y \sim p_{X,Y}$. Define *mutual information*

$$I[X; Y] = \mathbb{E}\left[\log_2 \frac{p_{X,Y}(X, Y)}{p_X(X)p_Y(Y)}\right].$$

Mutual Information

$X, Y \sim p_{X,Y}$. Define *mutual information*

$$I[X; Y] = \mathbb{E}\left[\log_2 \frac{p_{X,Y}(X, Y)}{p_X(X)p_Y(Y)}\right].$$

Measures how much information one random variable contains about the other.

Mutual Information

$X, Y \sim p_{X,Y}$. Define *mutual information*

$$I[X; Y] = \mathbb{E}\left[\log_2 \frac{p_{X,Y}(X, Y)}{p_X(X)p_Y(Y)}\right].$$

Measures how much information one random variable contains about the other.

Given a data sample, $I[X, Y]$ must be estimated.

Maximal Information Coefficient

Data set of real values $\{(x_i, y_i)\}_{i=1}^{n}$.

Maximal Information Coefficient

Data set of real values $\{(x_i, y_i)\}_{i=1}^{n}$.

Consider grids up to a specified resolution. For each grid, calculate an estimate of the mutual information based on that grid.

Maximal Information Coefficient

Data set of real values $\{(x_i, y_i)\}_{i=1}^{n}$.

Consider grids up to a specified resolution. For each grid, calculate an estimate of the mutual information based on that grid.

Adjust these values to allow comparison across grid sizes.

Maximal Information Coefficient

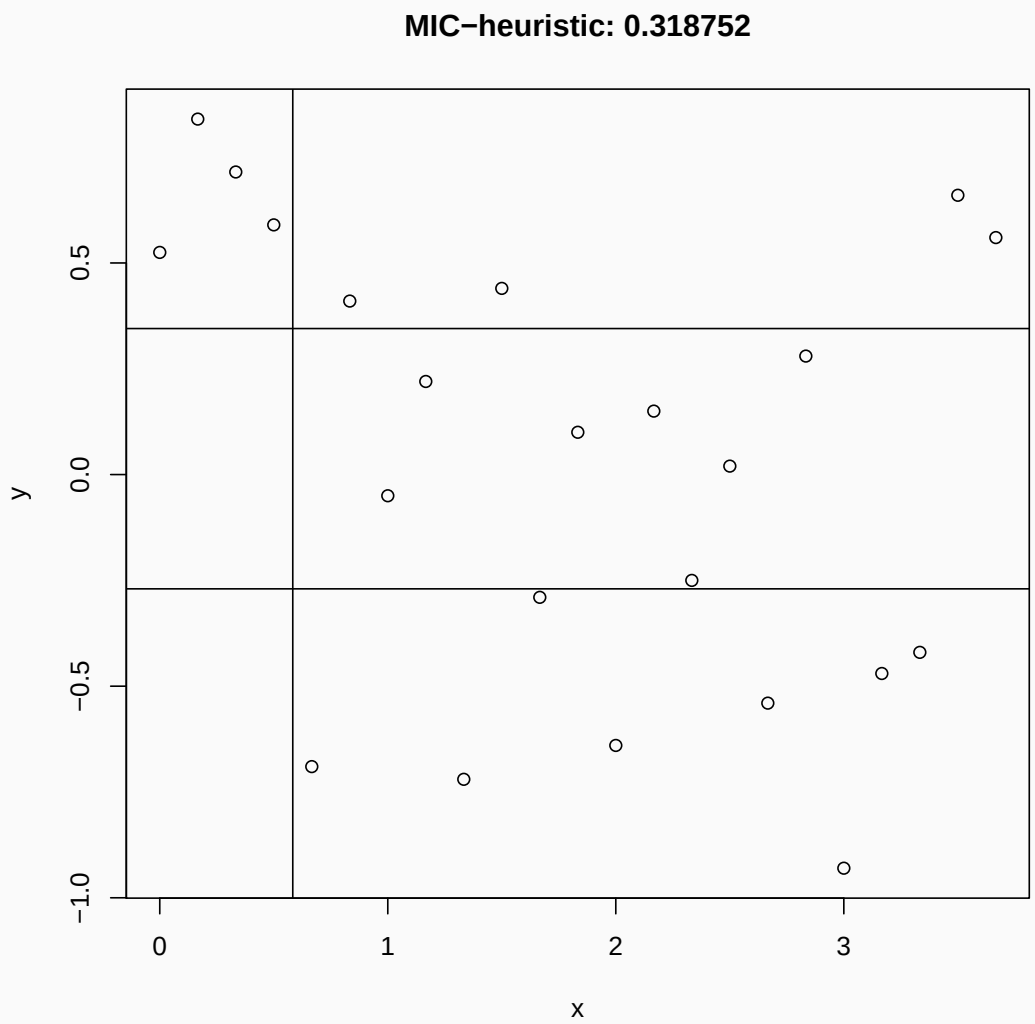
Data set of real values $\{(x_i, y_i)\}_{i=1}^{n}$.

Consider grids up to a specified resolution. For each grid, calculate an estimate of the mutual information based on that grid.

Adjust these values to allow comparison across grid sizes.

Take the maximum.

Maximal Information Coefficient



WHO data set

357 social, economic, health and political indicators from the World Health Organisation, for 202 countries.

WHO data set

357 social, economic, health and political indicators from the World Health Organisation, for 202 countries.

63546 pairs to consider

WHO data set

357 social, economic, health and political indicators from the World Health Organisation, for 202 countries.

63546 pairs to consider

around 23000 missing values (32%)

Plot from Reshef et al.

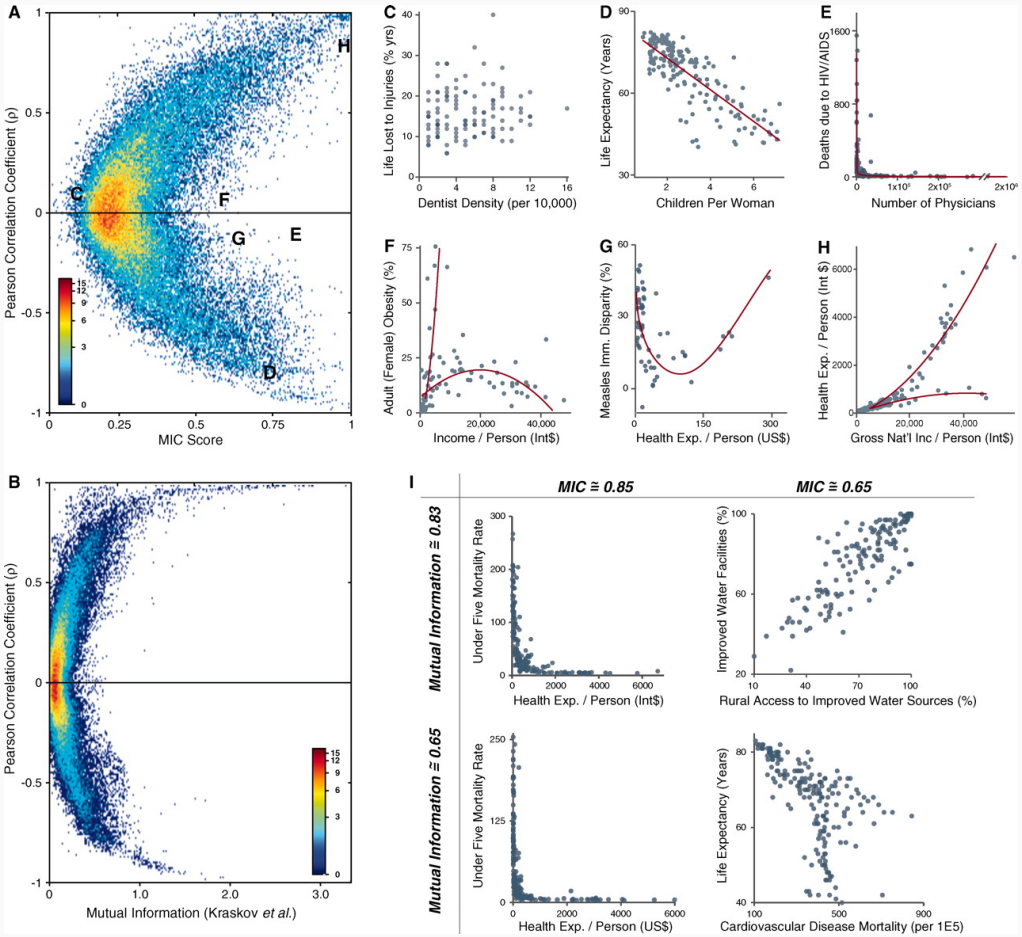
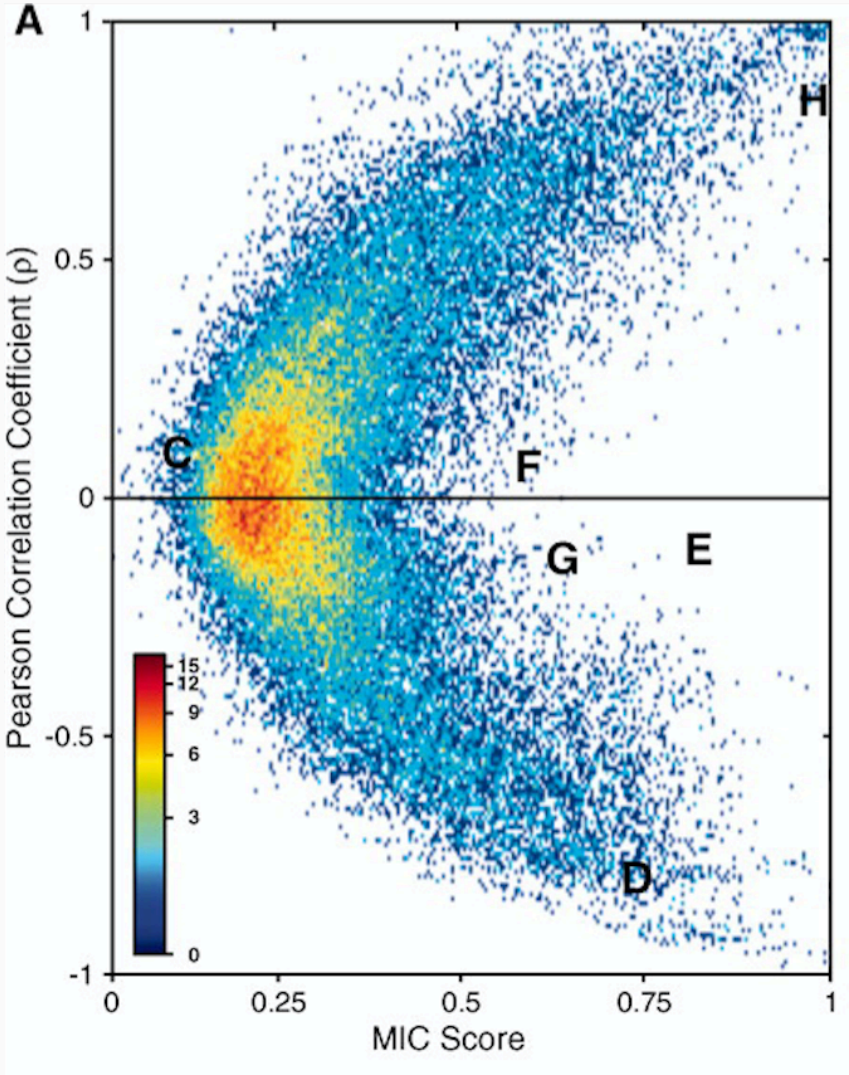


Figure: Figure 4 from [1]

MIC vs. Correlation



MIC vs. Correlation

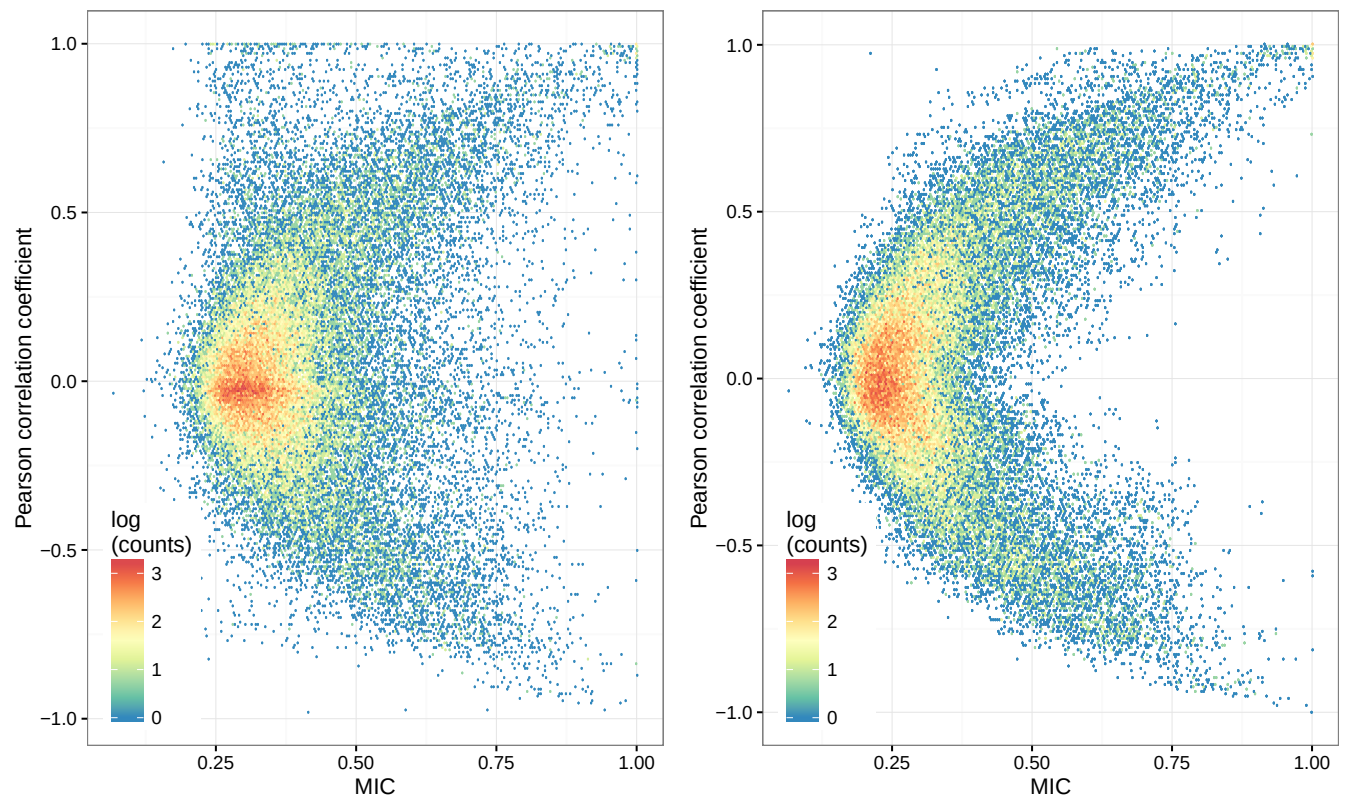
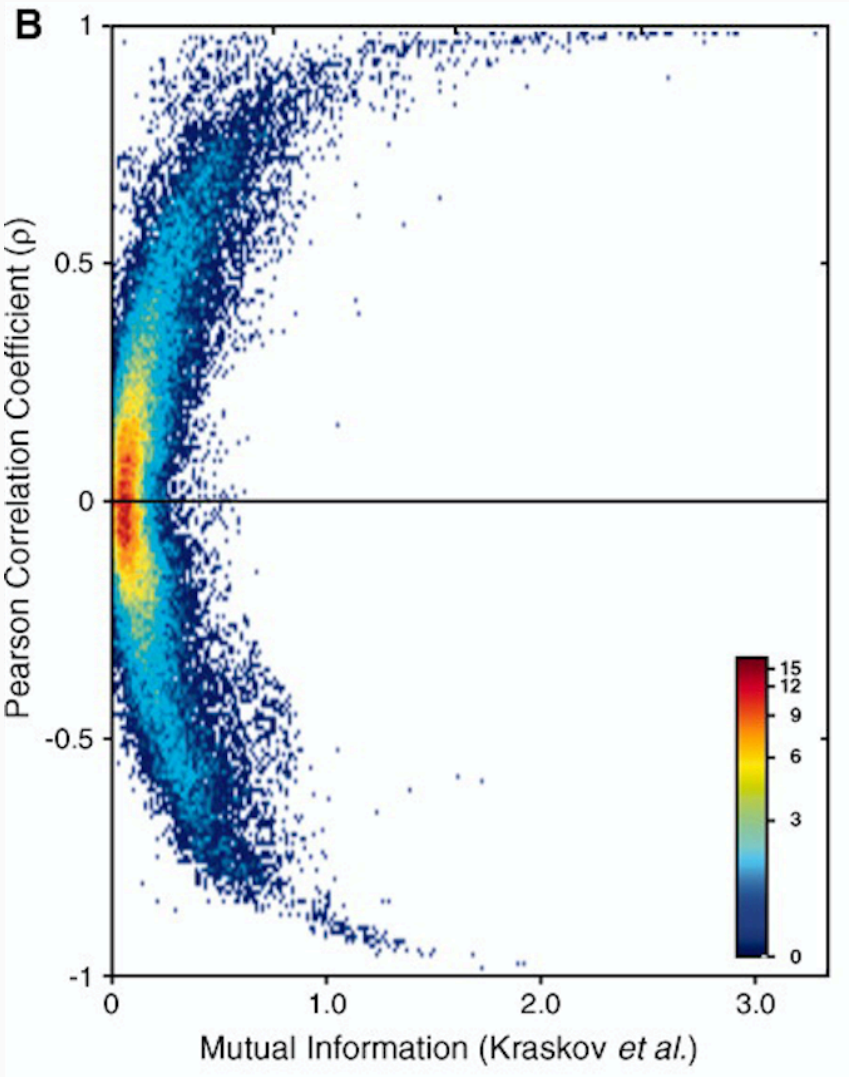


Figure: Replicating Figure 4A from [1]

I vs. Correlation



I vs. Correlation

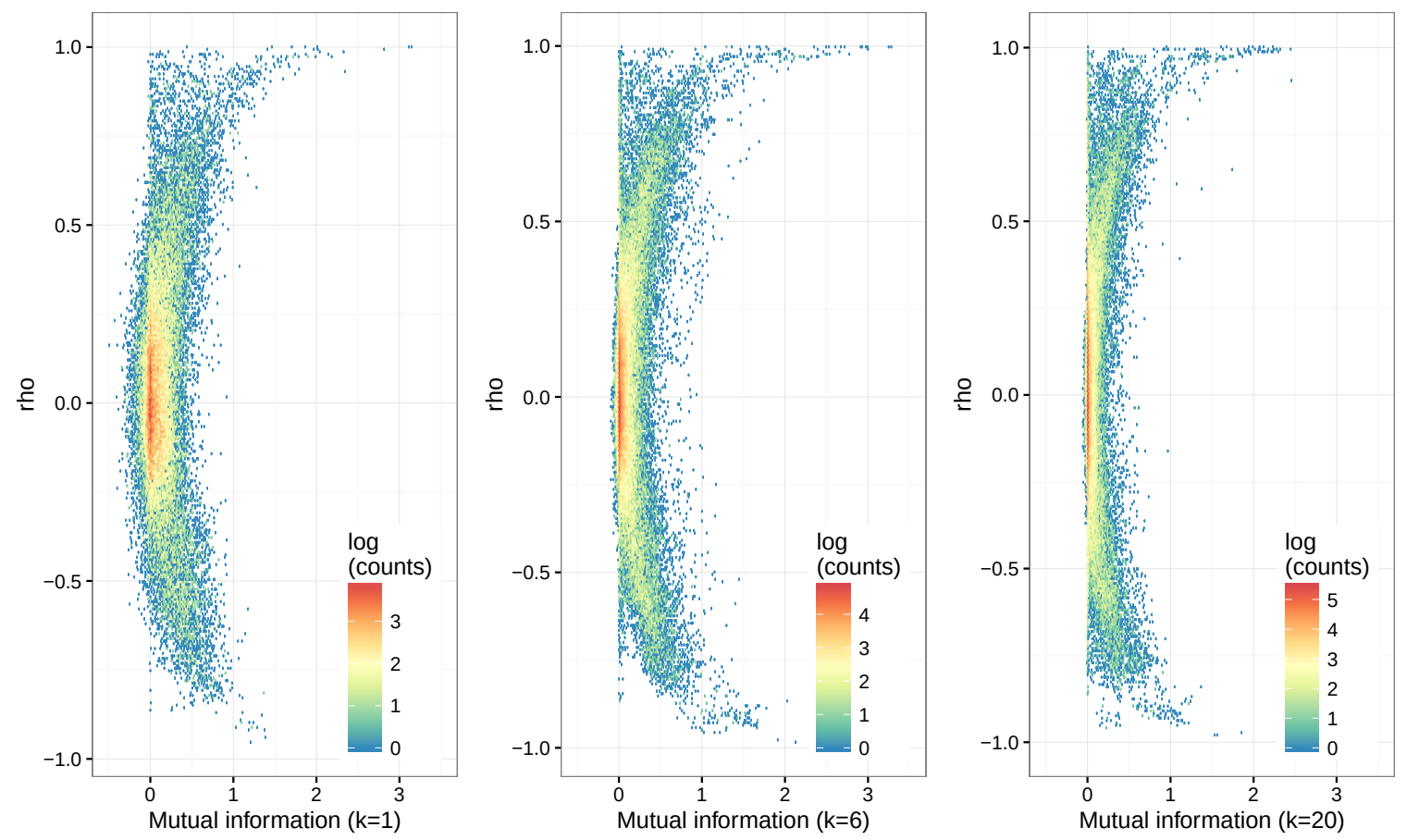
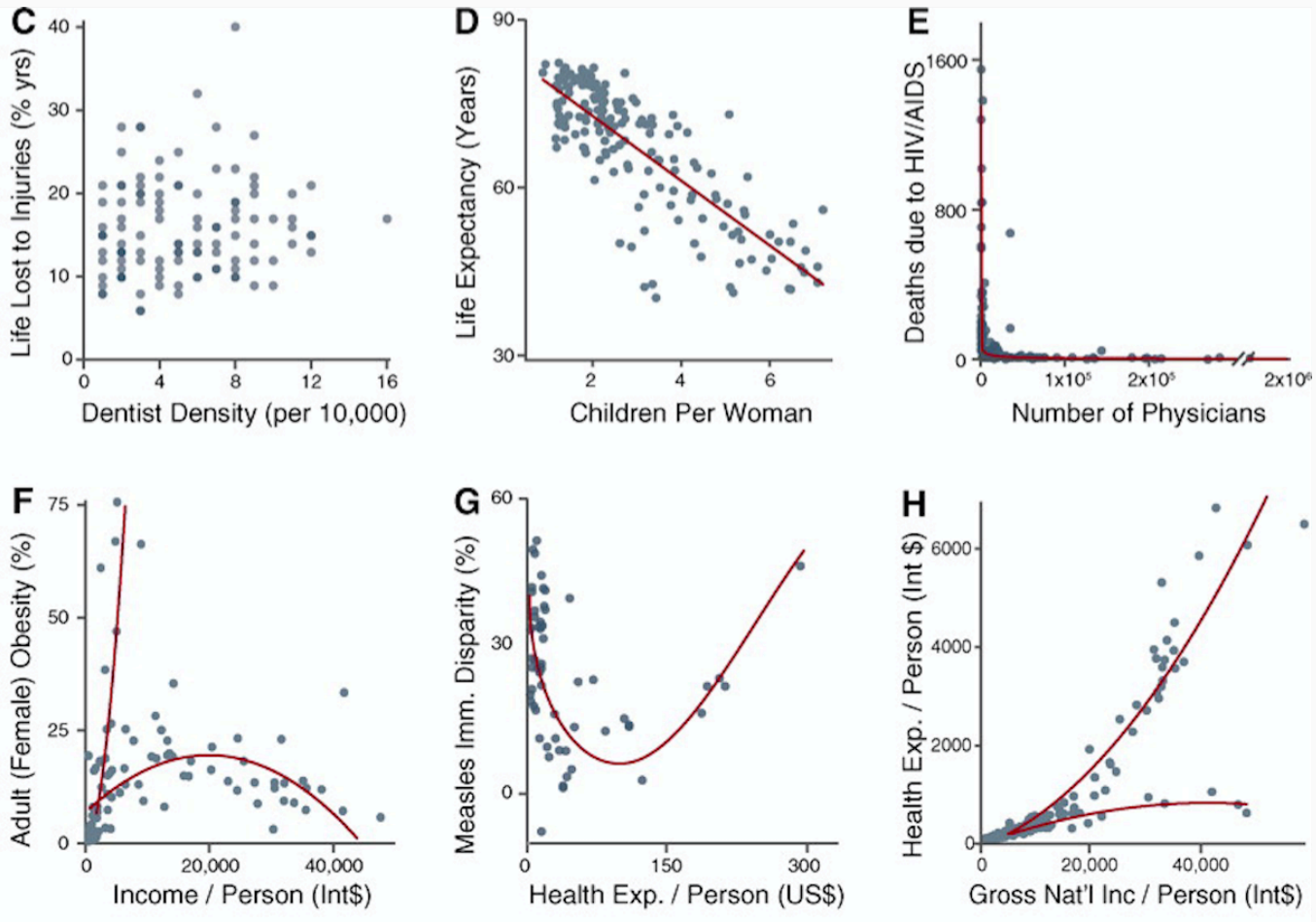


Figure: Replicating Figure 4B from [1]

Data Visualisation Choices



Yeast gene expression data

Yeast gene expression data (by Spellman et al):

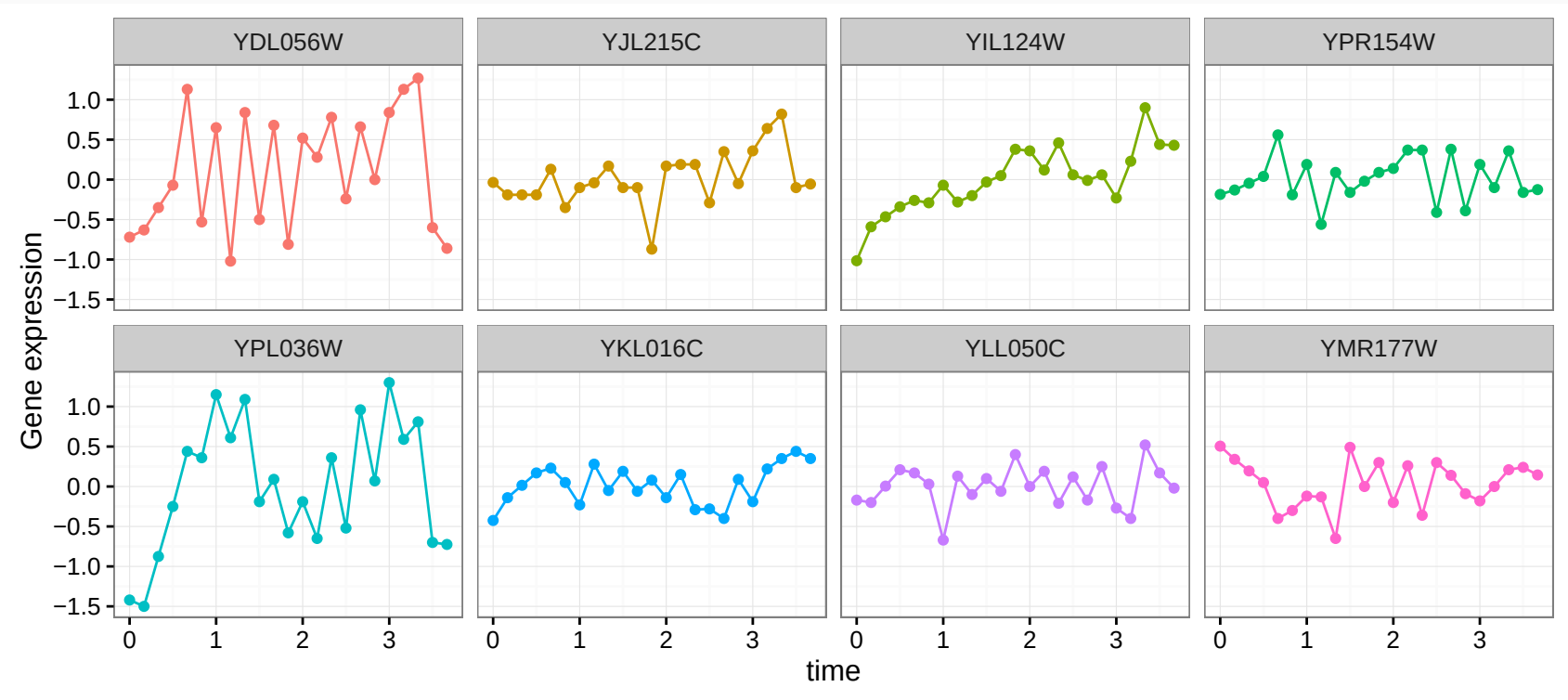
- expression for 4381 genes
- measured at 23 time points

Yeast gene expression data

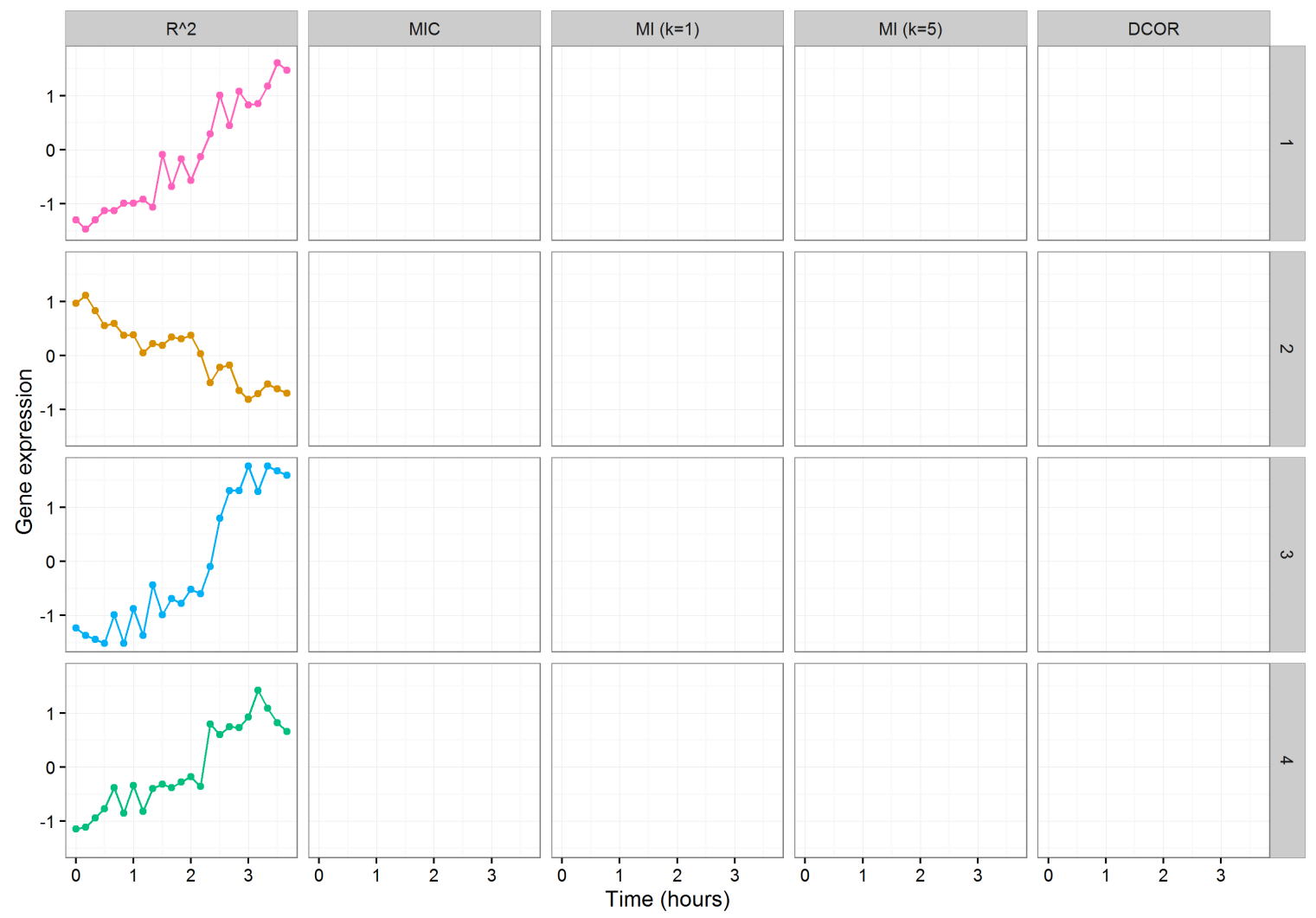
Yeast gene expression data (by Spellman et al):

expression for 4381 genes

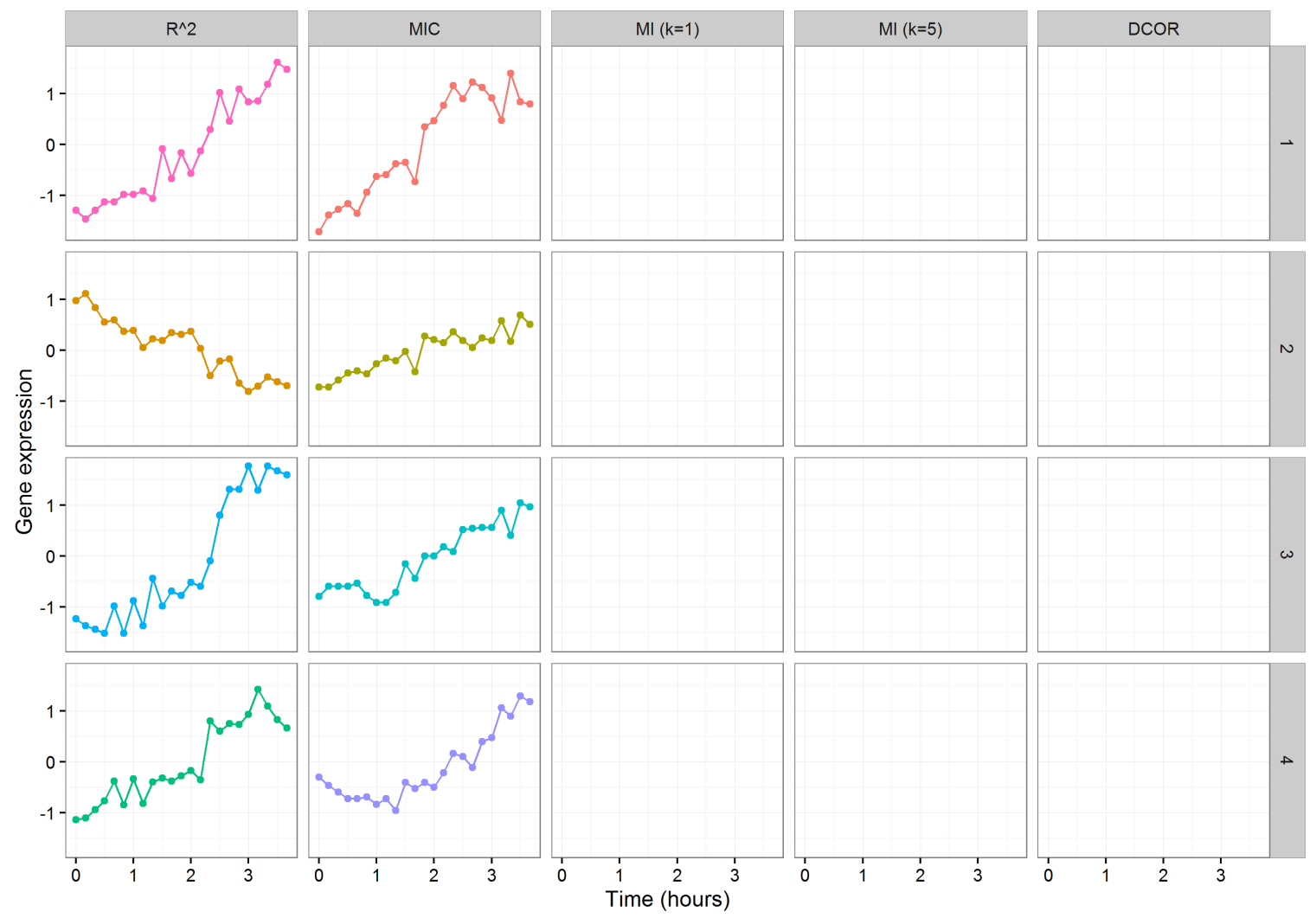
measured at 23 time points



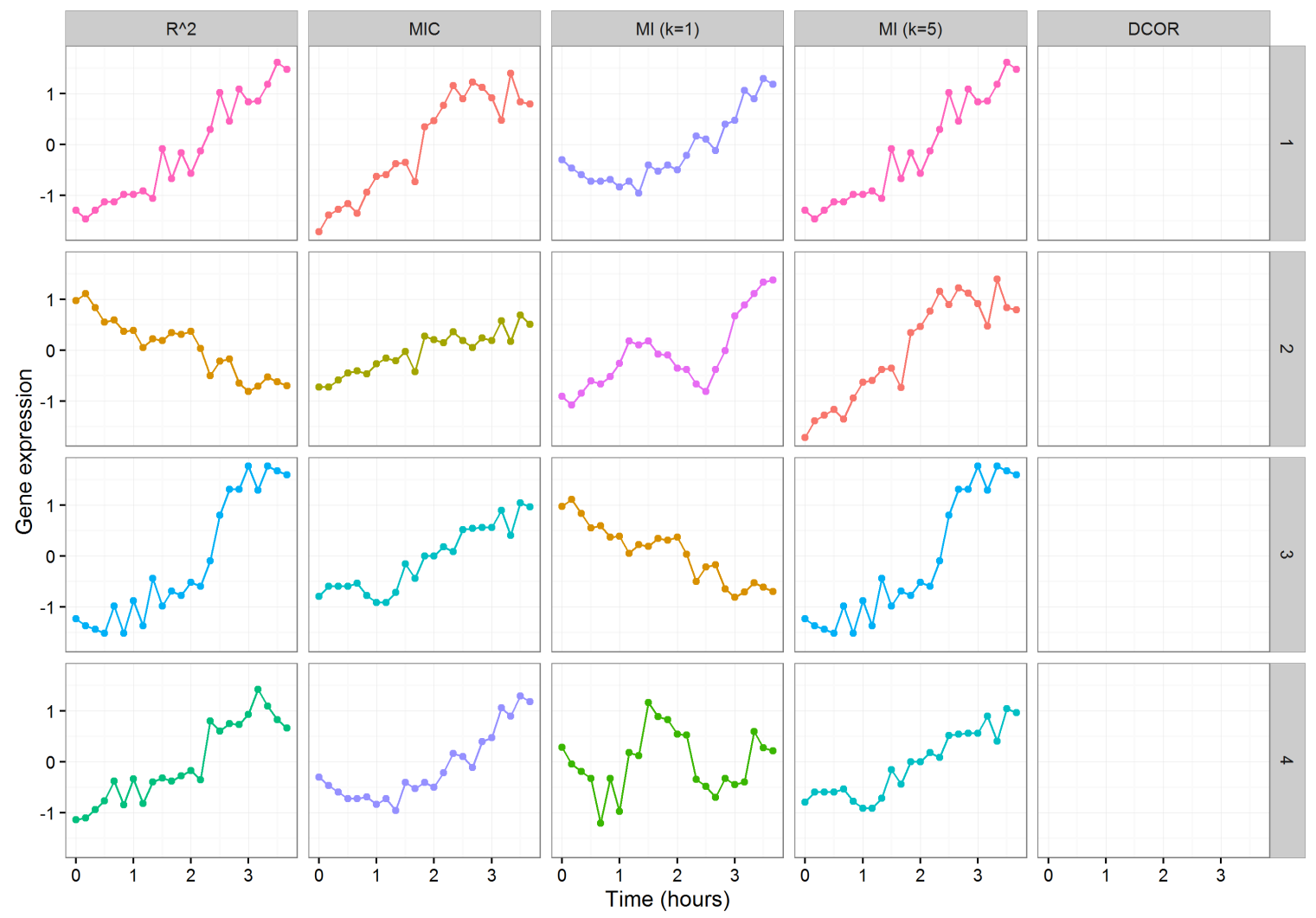
Comparing methods: TOP 4 dependencies



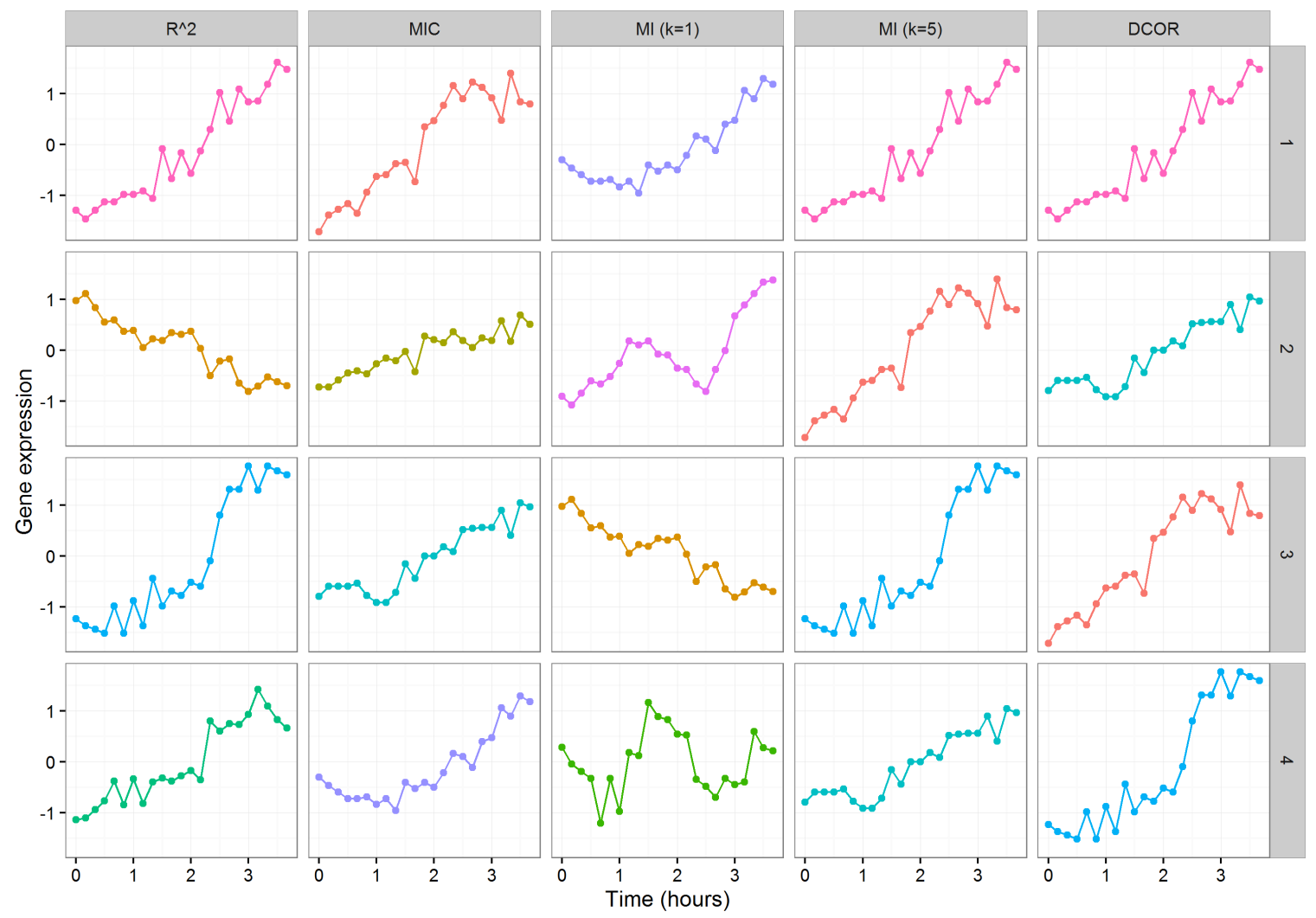
Comparing methods: TOP 4 dependencies



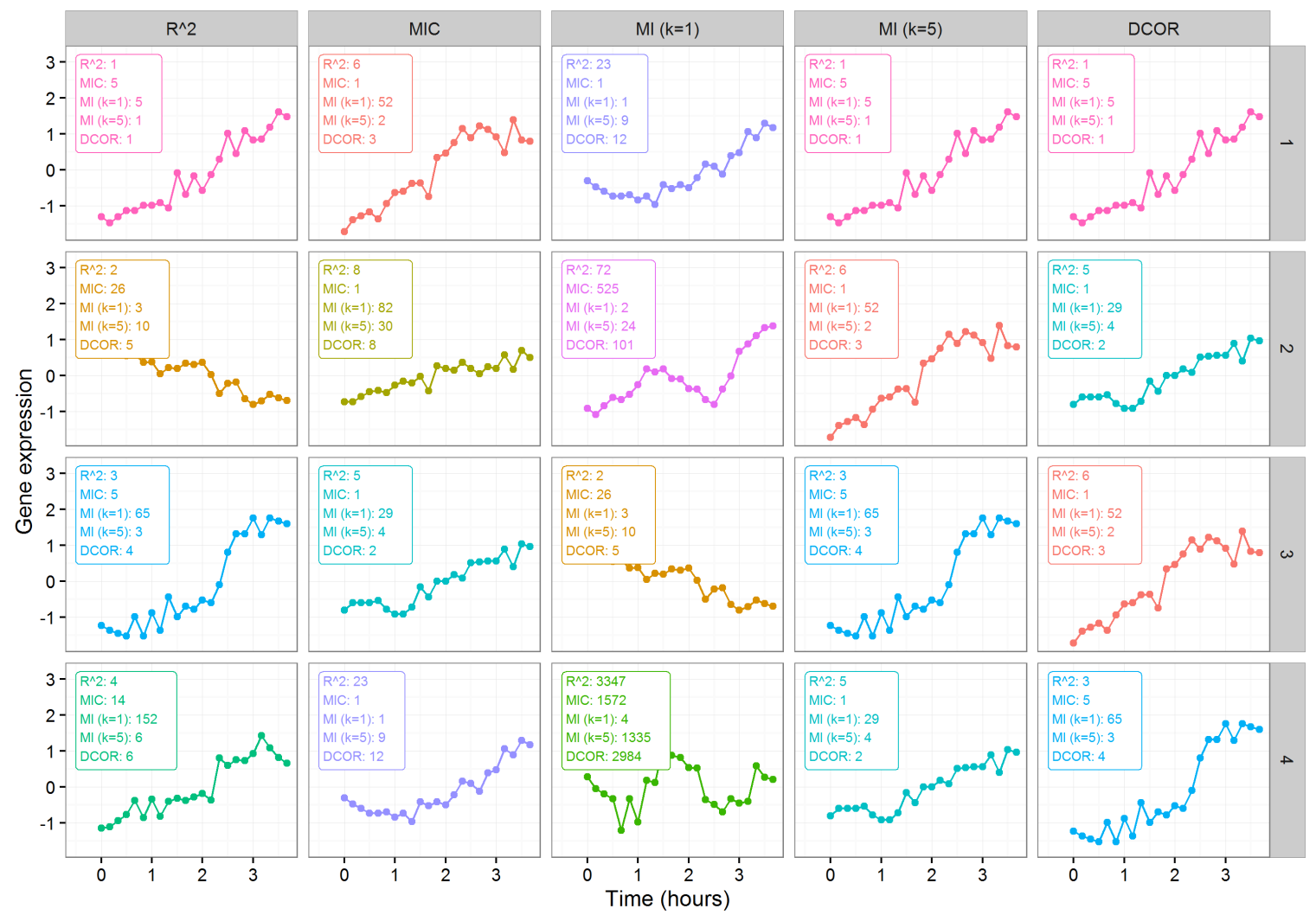
Comparing methods: TOP 4 dependencies



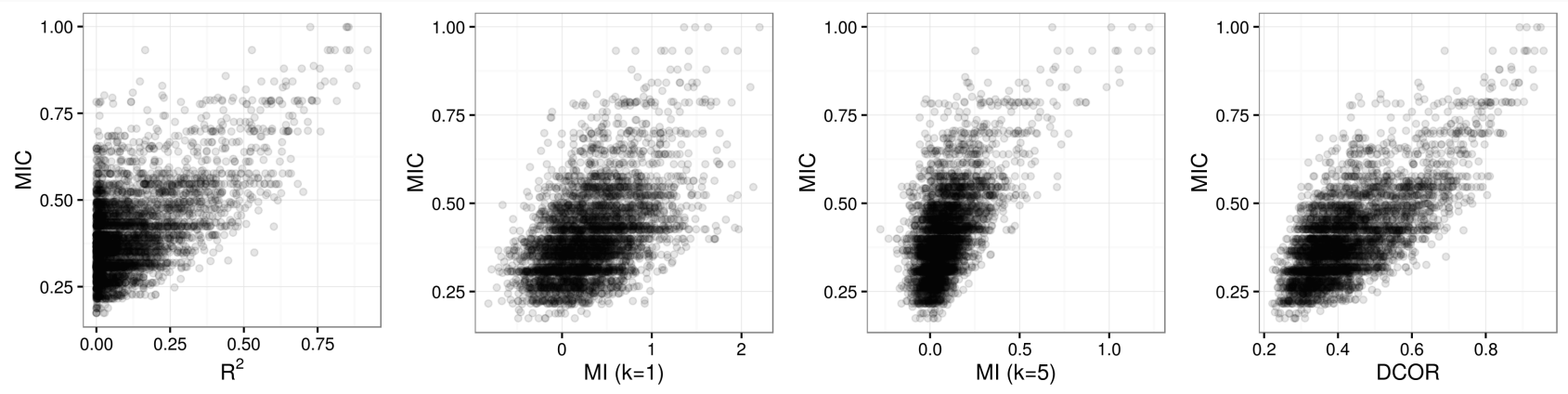
Comparing methods: TOP 4 dependencies



Comparing methods: TOP 4 dependencies



Comparing methods



Testing Statistical Power

We would like our dependency measures to be able to detect true dependencies.

Testing Statistical Power

We would like our dependency measures to be able to detect true dependencies.

Generated data according to a true relationship (e.g. linear, parabolic, sinusoidal).

Testing Statistical Power

We would like our dependency measures to be able to detect true dependencies.

Generated data according to a true relationship (e.g. linear, parabolic, sinusoidal).

Add noise, and calculate the dependency statistics.

Testing Statistical Power

We would like our dependency measures to be able to detect true dependencies.

Generated data according to a true relationship (e.g. linear, parabolic, sinusoidal).

Add noise, and calculate the dependency statistics.

Permute the y-values to obtain 'independent' null data and calculate the statistics again.

Testing Statistical Power

We would like our dependency measures to be able to detect true dependencies.

Generated data according to a true relationship (e.g. linear, parabolic, sinusoidal).

Add noise, and calculate the dependency statistics.

Permute the y-values to obtain 'independent' null data and calculate the statistics again.

Repeat this many times. Estimate the power by the proportion of values calculated from noisy data which are above the 95th-percentile of values from null data.

x(t) = ((-67/37 sin(39/25-12 t)-112/51 sin(25/16-10 t)-61/14 sin(69/44-6 t)+1164/17 sin(t+11/7)+611/131 sin(2 t+179/38)+103/41 sin(3 t+179/38)+189/20 sin(4 t+41/26)+355/41 sin(5 t+41/26)+121/26 sin(7 t+30/19)+6/29 sin(8 t+75/16)+11/20 sin(9 t+49/31)+43/33 sin(11 t+41/26)+2690/13) theta(67 pi-t) theta(t-63 pi)+(-89/79 sin(58/41-12 t)-41/16 sin(19/13-11 t)-10/13 sin(101/67-10 t)-151/33 sin(65/43-7 t)-120/31 sin(58/39-6 t)-31/46 sin(93/65-5 t)-113/33 sin(239/159-2 t)-1517/25 sin(25/16-t)+33/8 sin(3 t+43/27)+17/7 sin(4 t+103/22)+23/21 sin(8 t+38/23)+57/23 sin(9 t+68/41)-725/121) theta(63 pi-t) theta(t-59 pi)+(-73/38 sin(134/89-10 t)-181/40 sin(91/58-9 t)-77/54 sin(42/31-4 t)+1417/33 sin(t+61/39)+328/19 sin(2 t+63/40)+727/51 sin(3 t+113/24)+2971/27 sin(5 t+85/54)+11/9 sin(6 t+21/17)+70/9 sin(7 t+35/22)+62/23 sin(8 t+241/52)+73/26 sin(11 t+11/7)+37/17 sin(12 t+73/45)-31973/87) theta(59 pi-t) theta(t-55 pi)+(-4429/46 sin(69/44-t)+1722/29 sin(2 t+63/40)+276/37 sin(3 t+49/31)+379/18 sin(4 t+30/19)+40/19 sin(5 t+147/92)+281/29 sin(6 t+49/31)+23/25 sin(7 t+47/29)+161/31 sin(8 t+19/12)+11/49 sin(9 t+51/29)+98/31 sin(10 t+74/47)+3/5 sin(11 t+164/35)+17/22 sin(12 t+37/24)-3180/29) theta(55 pi-t) theta(t-51 pi)+(-23/4 sin(58/39-17 t)-118/23 sin(26/17-16 t)-48/29 sin(190/127-15 t)-86/39 sin(49/32-14 t)-421/31 sin(87/56-9 t)-1201/50 sin(48/31-6 t)-417/5 sin(53/34-4 t)-18496/349 sin(53/34-3 t)-2629/31 sin(25/16-2 t)+7245/88 sin(t+69/44)+13676/291 sin(5 t+41/26)+191/11 sin(7 t+49/31)+1550/141 sin(8 t+83/52)+90/17 sin(10 t+136/29)+190/49 sin(11 t+29/18)+14/19 sin(12 t+50/31)+331/24 sin(13 t+8/5)-9905/27) theta(51 pi-t) theta(t-47 pi)+(-926/161 sin(1/58-11 t)-137/19 sin(1/167-9 t)-1241/98 sin(1/38-5 t)-257/12 sin(1/62-3 t)+797/13 sin(t+1/339)+364/11 sin(2 t+81/26)+641/40 sin(4 t+22/7)+491/46 sin(6 t+103/33)+251/28 sin(7 t+1/88)+47/6 sin(8 t+47/15)+158/25 sin(10 t+117/37)+75/14 sin(12 t+85/27)-22535/33) theta(47 pi-t) theta(t-43 pi)+(-11/17 sin(22/25-12 t)-44/31 sin(62/41-11 t)-35/19 sin(87/56-8 t)-58/33 sin(98/65-5 t)+1948/41 sin(t+5/31)+682/27 sin(2 t+155/43)+247/15 sin(3 t+33/25)+156/31 sin(4 t+16/11)+75/26 sin(6 t+78/25)+339/97 sin(7 t+101/36)+3/28 sin(9 t+119/43)+21/32 sin(10 t+1381/345)-16712/25) theta(43 pi-t) theta(t-39 pi)+(-5/12 sin(20/41-6 t)-23/26 sin(1/35-5 t)-38/31 sin(58/39-4 t)+587/24 sin(t+49/22)+425/63 sin(2 t+281/78)+50/29 sin(3 t+91/31)+4/7 sin(7 t+29/37)+8/37 sin(8 t+87/31)+10/51 sin(9 t+47/28)+3/7 sin(10 t+63/17)+1/17 sin(11 t+29/12)+5/29 sin(12 t+67/15)-11925/17) theta(39 pi-t) theta(t-35 pi)+(-9/25 sin(41/36-5 t)-39/38 sin(4/3-4 t)-27/13 sin(1/48-3 t)-79/33 sin(77/60-2 t)+818/29 sin(t+19/15)+3/8 sin(6 t+194/47)+7/23 sin(7 t+71/23)-9803/20) theta(35 pi-t) theta(t-31 pi)+(-397/22 sin(2/15-t)+62/21 sin(2 t+2/11)+58/27 sin(3 t+37/9)+11/17 sin(4 t+65/38)+9/14 sin(5 t+89/31)+114/229 sin(6 t+27/31)+3/35 sin(7 t+185/42)-34395/43) theta(31 pi-t) theta(t-27 pi)+(-93/70 sin(1/20-7 t)-97/28 sin(86/115-3 t)+27721/616 sin(t+5/43)+221/44 sin(2 t+5/51)+96/35 sin(4 t+10/3)+14/15 sin(5 t+77/50)+41/42 sin(6 t+11/5)-33646/41) theta(27 pi-t) theta(t-23 pi)+(-34/25 sin(25/29-6 t)-59/27 sin(31/26-4 t)-23/4 sin(38/33-2 t)+2071/35 sin(t+24/13)+101/42 sin(3 t+22/27)+30/17 sin(5 t+57/40)+13/18 sin(7 t+27/10)-35021/76) theta(23 pi-t) theta(t-19 pi)+(-19/33 sin(25/18-4 t)-47/30 sin(54/37-3 t)+1195/14 sin(t+21/26)+12/13 sin(2 t+93/34)+18/13 sin(5 t+43/27)-7251/20) theta(19 pi-t) theta(t-15 pi)+(-247/45 sin(53/40-10 t)-107/24 sin(19/51-8 t)+3579/23 sin(t+237/71)+7181/68 sin(2 t+77/51)+1574/61 sin(3 t+80/29)+...

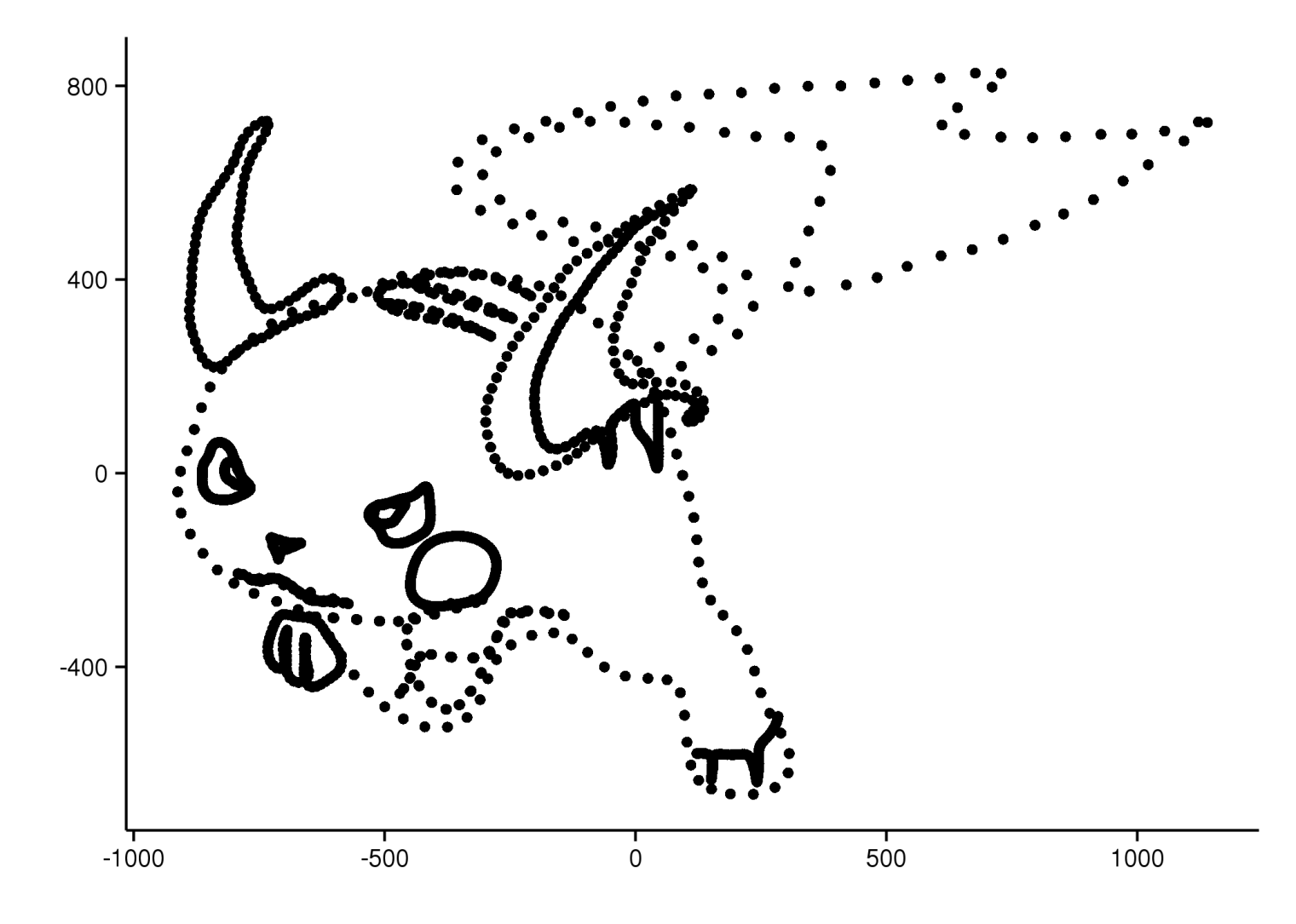
...+466/23 sin(4 t+72/19)+97/39 sin(5 t+128/43)+201/29 sin(6 t+11/5)+329/31 sin(7 t+127/31)+48/37 sin(9 t+99/31)+74/27 sin(11 t+23/28)+53/20 sin(12 t+143/31)-771/16) theta(15 pi-t) theta(t-11 pi)+(-1/13 sin(2/3-16 t)-2/39 sin(43/46-14 t)-26/17 sin(9/22-9 t)-219/58 sin(41/33-7 t)+4135/46 sin(t+115/28)+2992/43 sin(2 t+80/43)+943/42 sin(3 t+159/38)+174/35 sin(4 t+301/150)+671/84 sin(5 t+137/34)+67/19 sin(6 t+13/25)+51/16 sin(8 t+3/11)+49/16 sin(10 t+46/47)+21/22 sin(11 t+108/31)+75/43 sin(12 t+31/38)+16/15 sin(13 t+36/13)+34/31 sin(15 t+37/14)-37942/49) theta(11 pi-t) theta(t-7 pi)+(-25/49 sin(33/32-14 t)-67/5 sin(13/33-4 t)-1711/46 sin(31/77-3 t)+2714/5 sin(t+5/8)+1370/23 sin(2 t+56/19)+179/9 sin(5 t+57/17)+81/11 sin(6 t+118/41)+244/23 sin(7 t+83/41)+1069/91 sin(8 t+66/35)+233/37 sin(9 t+115/43)+106/31 sin(10 t+12/17)+63/19 sin(11 t+5/8)+145/22 sin(12 t+4/13)+87/37 sin(13 t+6/13)+73/33 sin(15 t+62/21)+77/26 sin(16 t+25/12)+49/29 sin(17 t+49/15)-17486/67) theta(7 pi-t) theta(t-3 pi)+(-35/29 sin(19/26-27 t)-102/41 sin(12/25-26 t)-255/47 sin(41/36-18 t)-1242/73 sin(13/23-12 t)-1113/32 sin(15/17-6 t)-2824/31 sin(49/32-3 t)+9066/19 sin(t+33/20)+3402/13 sin(2 t+52/27)+981/11 sin(4 t+32/19)+1561/17 sin(5 t+37/25)+46 sin(7 t+137/60)+1062/35 sin(8 t+159/40)+135/7 sin(9 t+178/59)+127/25 sin(10 t+169/36)+109/8 sin(11 t+63/47)+359/29 sin(13 t+88/41)+103/15 sin(14 t+92/29)+367/54 sin(15 t+63/22)+181/23 sin(16 t+98/29)+74/35 sin(17 t+427/95)+65/49 sin(19 t+19/24)+4/11 sin(20 t+25/29)+25/18 sin(21 t+33/41)+180/43 sin(22 t+37/11)+63/23 sin(23 t+56/17)+38/23 sin(24 t+202/47)+44/37 sin(25 t+52/15)+47/25 sin(28 t+73/16)+7/6 sin(29 t+90/31)+853/3) theta(3 pi-t) theta(t+pi)) theta(sqrt(sgn(sin(t/2))))

y(t) = ((-122/39 sin(36/23-10 t)-227/16 sin(47/30-5 t)+171/8 sin(t+11/7)+887/37 sin(2 t+63/40)+794/35 sin(3 t+52/33)+425/19 sin(4 t+52/33)+25/24 sin(6 t+52/33)+34/31 sin(7 t+35/22)+39/17 sin(8 t+51/32)+115/17 sin(9 t+49/31)+33/46 sin(11 t+35/22)+45/58 sin(12 t+27/17)-27039/46) theta(67 pi-t) theta(t-63 pi)+(-97/19 sin(59/39-11 t)-2/21 sin(33/82-10 t)-674/23 sin(36/23-3 t)-677/39 sin(66/43-2 t)-347/73 sin(304/203-t)+219/5 sin(4 t+51/32)+283/39 sin(5 t+39/23)+326/37 sin(6 t+47/29)+104/27 sin(7 t+43/26)+41/23 sin(8 t+31/19)+103/49 sin(9 t+43/26)+152/91 sin(12 t+35/23)+4799/64) theta(63 pi-t) theta(t-59 pi)+(-99/28 sin(95/61-12 t)-53/17 sin(25/16-7 t)-638/21 sin(58/37-5 t)+1635/46 sin(t+85/54)+72/23 sin(2 t+25/16)+181/22 sin(3 t+64/41)+367/19 sin(4 t+63/40)+226/25 sin(6 t+202/43)+8/7 sin(8 t+38/25)+80/23 sin(9 t+113/24)+240/49 sin(10 t+174/37)+23/33 sin(11 t+5/3)+5633/16) theta(59 pi-t) theta(t-55 pi)+(-457/45 sin(80/51-3 t)-8497/36 sin(69/44-t)+11/18 sin(2 t+117/25)+137/15 sin(4 t+30/19)+79/18 sin(5 t+49/31)+249/37 sin(6 t+41/26)+155/36 sin(7 t+30/19)+155/34 sin(8 t+65/41)+44/15 sin(9 t+165/103)+73/31 sin(10 t+46/29)+47/38 sin(11 t+13/8)+37/34 sin(12 t+51/32)+6758/23) theta(55 pi-t) theta(t-51 pi)+(-873/238 sin(106/71-14 t)-108/35 sin(118/79-13 t)-179/24 sin(38/25-12 t)-242/21 sin(49/32-11 t)-13/25 sin(43/36-10 t)-225/52 sin(59/38-8 t)-1016/47 sin(67/43-7 t)-1823/30 sin(39/25-4 t)-1150/29 sin(64/41-3 t)-454/13 sin(36/23-t)+1079/59 sin(2 t+36/23)+1417/58 sin(5 t+63/40)+313/14 sin(6 t+41/26)+336/73 sin(9 t+45/29)+10/9 sin(15 t+11/7)+111/43 sin(16 t+59/36)+34/29 sin(17 t+56/33)-9919/29) theta(51 pi-t) theta(t-47 pi)+(-37/19 sin(5/31-10 t)+790/31 sin(t+64/21)+81/16 sin(2 t+14/39)+169/31 sin(3 t+249/71)+649/130 sin(4 t+7/48)+122/31 sin(5 t+218/67)+43/15 sin(6 t+3/41)+105/37 sin(7 t+173/55)+172/69 sin(8 t+5/36)+55/29 sin(9 t+137/43)+62/37 sin(11 t+81/25)+51/37 sin(12 t+2/23)-16525/69) theta(47 pi-t) theta(t-43 pi)+(-28/17 sin(19/27-7 t)-785/22 sin(17/47-3 t)+2329/65 sin(t+58/13)+565/22 sin(2 t+59/31)+691/31 sin(4 t+23/70)+53/35 sin(5 t+39/25)+99/29 sin(6 t+93/20)+23/18 sin(8 t+15/23)+23/28 sin(9 t+95/26)+11/29 sin(10 t+73/20)+92/35 sin(11 t+191/53)+14/47 sin(12 t+138/37)-44969/122) theta(43 pi-t) theta(t-39 pi)+(-3/17 sin(1/38-12 t)-6/5 sin(19/22-4 t)-83/46 sin(47/59-3 t)+439/29 sin(t+119/33)+300/41 sin(2 t+47/29)+38/27 sin(5 t+296/71)+21/29 sin(6 t+53/28)+5/32 sin(7 t+16/23)+7/19 sin(8 t+45/91)+3/13 sin(9 t+107/25)+9/41 sin(10 t+17/9)+2/35 sin(11 t+69/32)-4835/32) theta(39 pi-t) theta(t-35 pi)+(-9/14 sin(1/7-7 t)-37/36 sin(41/47-5 t)-157/59 sin(19/47-3 t)+1717/74 sin(t+1/15)+161/50 sin(2 t+78/35)+25/19 sin(4 t+47/16)+23/24 sin(6 t+71/24)-5095/63) theta(35 pi-t) theta(t-31 pi)+(-9/49 sin(38/29-5 t)+375/16 sin(t+163/39)+13/14 sin(2 t+17/21)+18/19 sin(3 t+67/26)+22/39 sin(4 t+49/30)+13/30 sin(6 t+16/9)+10/41 sin(7 t+114/31)-29/36) theta(31 pi-t) theta(t-27 pi)+(-83/28 sin(29/20-3 t)+1369/25 sin(t+107/23)+131/22 sin(2 t+24/23)+47/21 sin(4 t+61/48)+19/18 sin(5 t+146/35)+43/42 sin(6 t+23/20)+29/37 sin(7 t+123/31)-347/81) theta(27 pi-t) theta(t-23 pi)+(-201/100 sin(68/91-4 t)+1526/29 sin(t+135/43)+149/18 sin(2 t+97/38)+101/19 sin(3 t+137/30)+14/13 sin(5 t+32/31)+85/64 sin(6 t+31/27)+24/23 sin(7 t+50/17)-1227/14) theta(23 pi-t) theta(t-19 pi)+(-68/69 sin(5/12-5 t)-6417/85 sin(20/33-t)+311/74 sin(2 t+94/33)+107/35 sin(3 t+124/29)+27/43 sin(4 t+46/11)-5948/29) theta(19 pi-t) theta(t-15 pi)+(-72/35 sin(8/21-11 t)-113/31 sin(14/19-8 t)-92/15 sin(4/7-5 t)+7212/31 sin(t+59/31)+...

...+1707/20 sin(2 t+42/25)+875/33 sin(3 t+138/31)+945/43 sin(4 t+177/62)+68/33 sin(6 t+131/54)+278/29 sin(7 t+123/32)+47/23 sin(9 t+29/35)+15/4 sin(10 t+147/38)+35/23 sin(12 t+21/29)+12683/49) theta(15 pi-t) theta(t-11 pi)+(-52/31 sin(32/29-10 t)-89/33 sin(27/28-8 t)-128/13 sin(24/35-6 t)+5680/29 sin(t+32/15)+1160/17 sin(2 t+39/20)+877/34 sin(3 t+128/29)+139/20 sin(4 t+62/23)+1319/120 sin(5 t+380/117)+166/31 sin(7 t+61/18)+73/38 sin(9 t+24/95)+45/31 sin(11 t+152/71)+56/25 sin(12 t+6/23)+31/24 sin(13 t+106/39)+3/14 sin(14 t+430/123)+21/22 sin(15 t+56/33)+53/80 sin(16 t+142/33)+11153/25) theta(11 pi-t) theta(t-7 pi)+(-13548/29 sin(101/72-t)+2803/21 sin(2 t+67/28)+1713/29 sin(3 t+50/19)+1951/49 sin(4 t+123/44)+569/18 sin(5 t+139/49)+255/29 sin(6 t+164/37)+269/21 sin(7 t+4/43)+505/39 sin(8 t+19/52)+302/41 sin(9 t+35/34)+34/11 sin(10 t+26/21)+373/71 sin(11 t+32/27)+106/33 sin(12 t+74/49)+41/18 sin(13 t+163/109)+9/46 sin(14 t+270/61)+106/31 sin(15 t+13/9)+65/37 sin(16 t+47/36)+62/25 sin(17 t+43/31)-2707/19) theta(7 pi-t) theta(t-3 pi)+(-106/39 sin(17/60-26 t)-27/26 sin(5/24-20 t)-269/36 sin(7/50-11 t)-308/17 sin(9/20-6 t)+6803/48 sin(t+67/30)+4066/25 sin(2 t+183/43)+2096/27 sin(3 t+145/67)+498/17 sin(4 t+43/31)+1181/36 sin(5 t+32/27)+327/14 sin(7 t+47/20)+293/18 sin(8 t+40/9)+1190/69 sin(9 t+81/29)+181/26 sin(10 t+68/33)+263/81 sin(12 t+41/44)+8/7 sin(13 t+116/31)+43/17 sin(14 t+105/43)+145/39 sin(15 t+154/51)+2623/874 sin(16 t+14/5)+64/23 sin(17 t+135/29)+27/22 sin(18 t+47/19)+15/19 sin(19 t+37/13)+23/19 sin(21 t+59/18)+17/29 sin(22 t+39/10)+68/31 sin(23 t+122/33)+61/71 sin(24 t+61/81)+9/7 sin(25 t+126/31)+81/37 sin(27 t+117/31)+27/67 sin(28 t+106/27)+17/16 sin(29 t+93/22)+27524/47) theta(3 pi-t) theta(t+pi)) theta(sqrt(sgn(sin(t/2))))

for $t \in [0, 68\pi]$, and where theta denotes the indicator function of $[0, \infty)$.

Testing Statistical Power



Testing Statistical Power

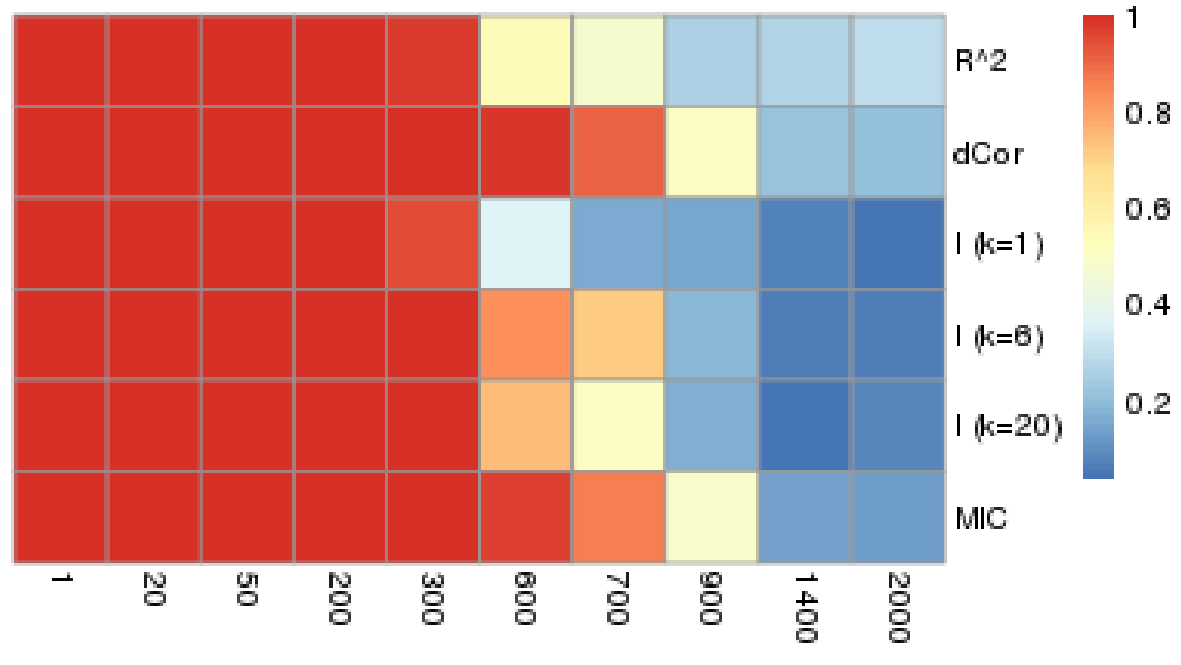


Figure: Heatmap of the statistical power (indicated by color) of various methods (y-axis) for the Raichu-like curve, for different noise levels (the standard deviation of the Gaussian noise added to the y_{R}-components on the x-axis)

Computing MIC

MIC is defined as a maximum over all grids of less than a certain size:

$$\max_{|R||C|<B(n)} MI(R, C)$$

Computing MIC

MIC is defined as a maximum over all grids of less than a certain size:

$$\max_{|R||C|<B(n)} MI(R, C)$$

Number of possible grids is

$$\binom{N-1}{|R|-1} \times \binom{N-1}{|C|-1}$$

Computing MIC

MIC is defined as a maximum over all grids of less than a certain size:

$$\max_{|R||C|<B(n)} MI(R, C)$$

Number of possible grids is

$$\binom{N-1}{|R|-1} \times \binom{N-1}{|C|-1}$$

Too many possibilities to check!

Computing MIC

What they do: MIC-Heuristic

Optimization can be done efficiently in *one axis only*

Computing MIC

What they do: MIC-Heuristic

- Optimization can be done efficiently in *one axis only*
- So partition x evenly and optimize over y (then vice-versa)

Computing MIC

What they do: MIC-Heuristic

- Optimization can be done efficiently in *one axis only*
- So partition x evenly and optimize over y (then vice-versa)

Restricts search to grids with one axis partitioned evenly

Computing MIC

What they do: MIC-Heuristic

- Optimization can be done efficiently in *one axis only*
- So partition x evenly and optimize over y (then vice-versa)

Restricts search to grids with one axis partitioned evenly

Strict subset of all grids

Computing MIC

What they do: MIC-Heuristic

- Optimization can be done efficiently in *one axis only*
- So partition x evenly and optimize over y (then vice-versa)

Restricts search to grids with one axis partitioned evenly

Strict subset of all grids

MIC-Heuristic $\leq$ MIC-Exact

Computing MIC

Spellman yeast data - $N = 23$ time series for each gene

Computing MIC

Spellman yeast data - $N = 23$ time series for each gene

$B(23) = 23^{0.6} = 6.56$

Computing MIC

Spellman yeast data - $N = 23$ time series for each gene

$B(23) = 23^{0.6} = 6.56$

MIC-Exact feasible in this setting

Can compare MIC-Exact and MIC-Heuristic

Computing MIC

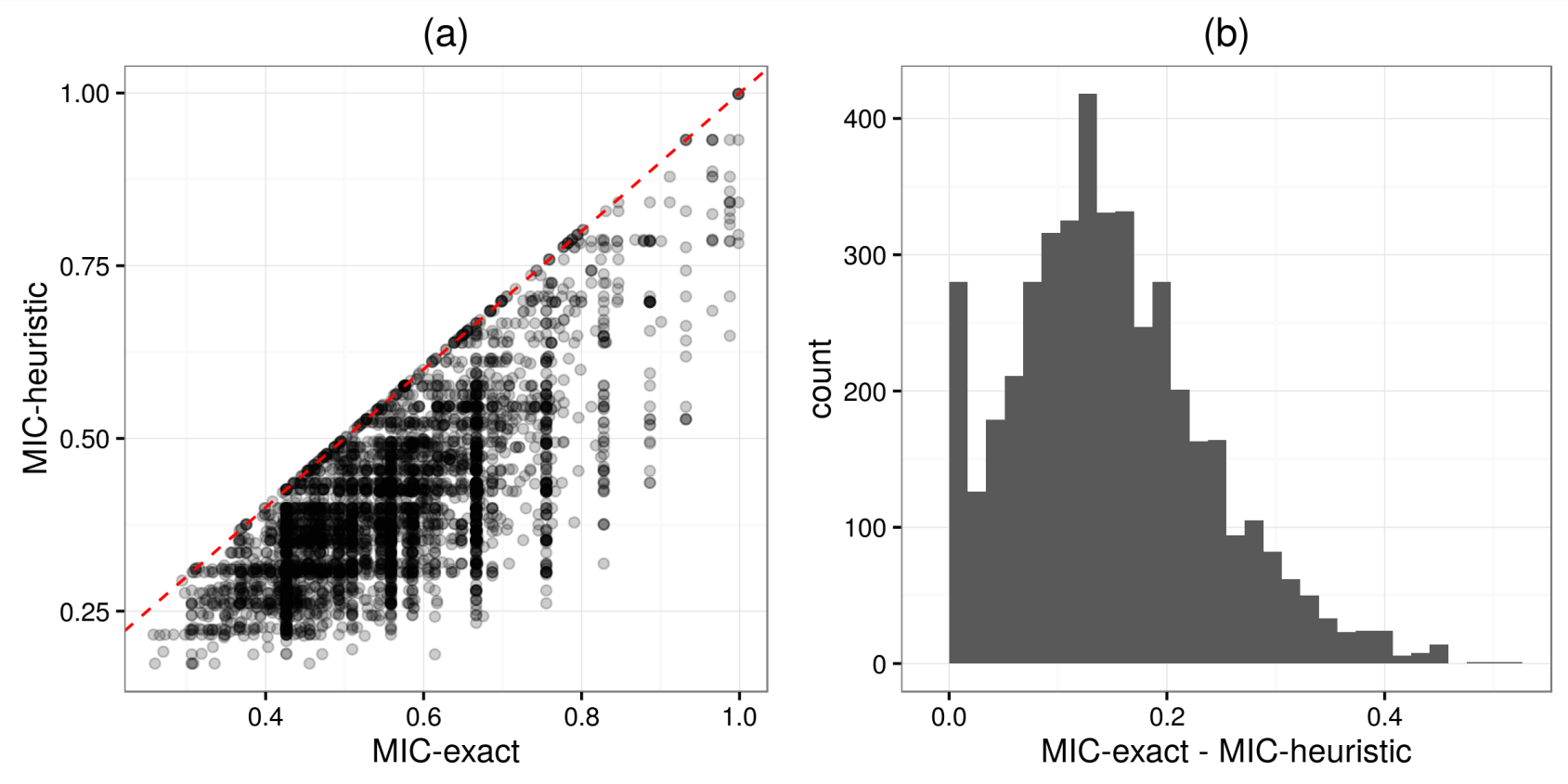
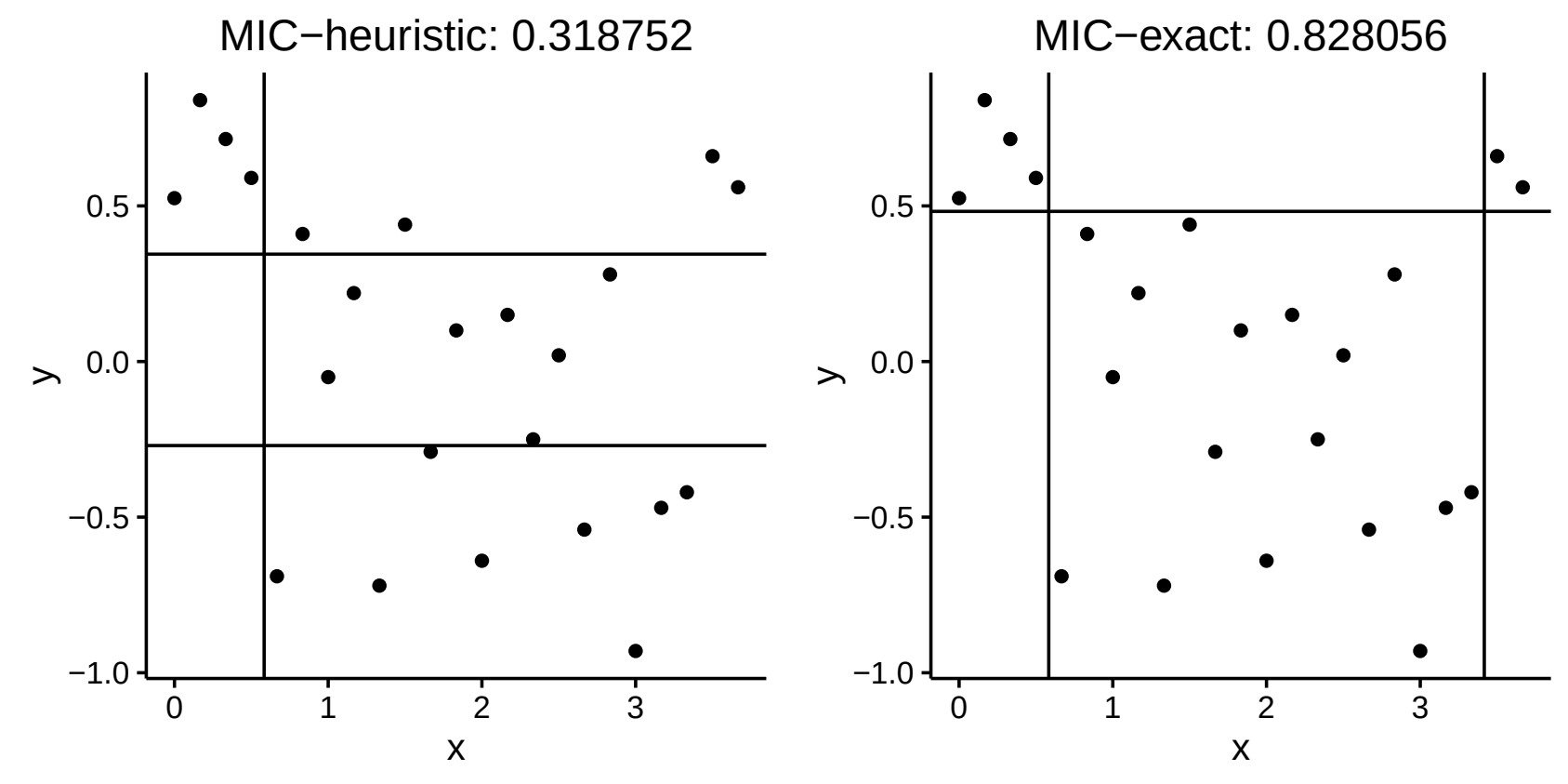
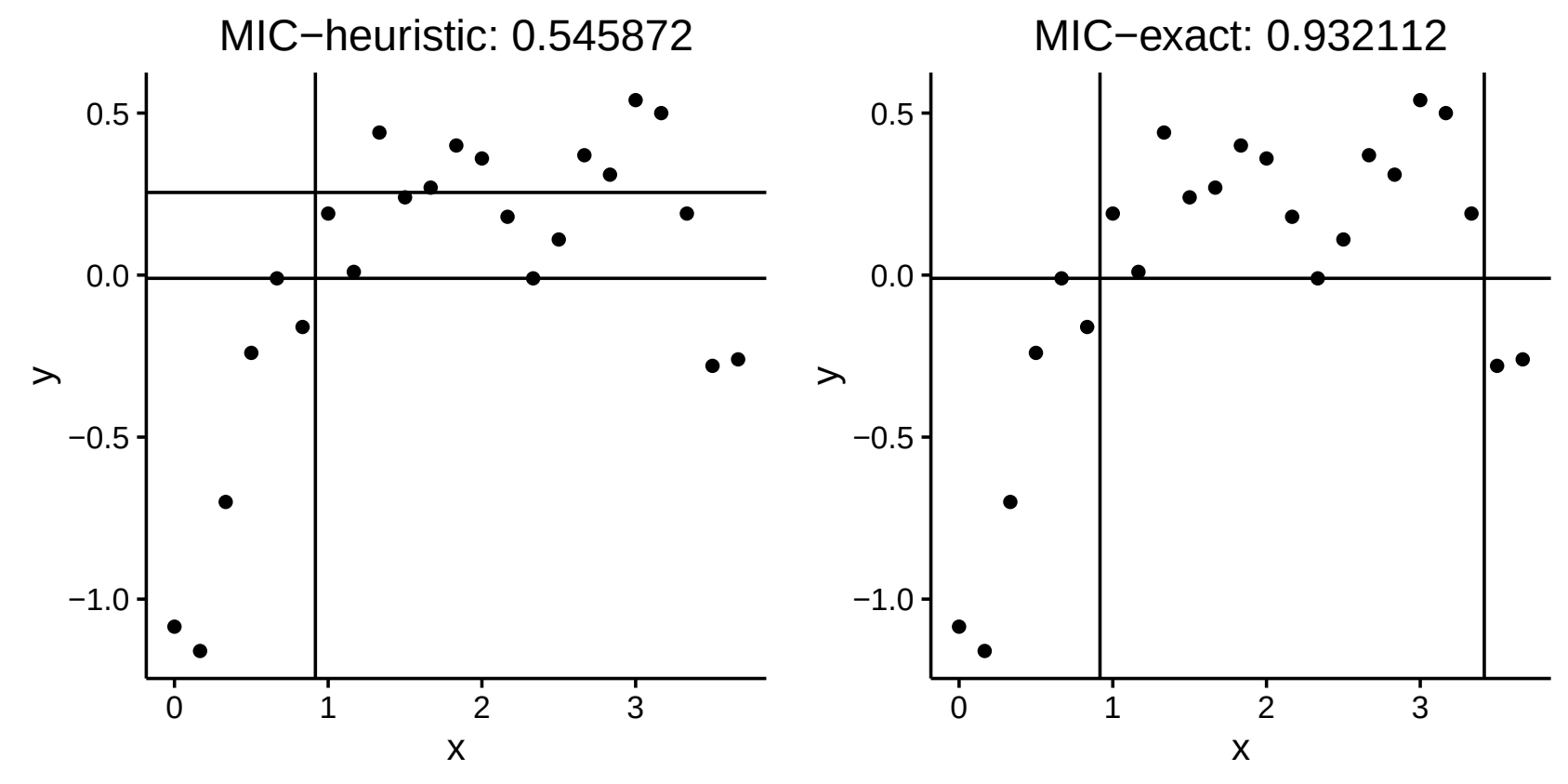


Figure: Discrepancies for Spellman yeast data

Computing MIC



Computing MIC



Computing MIC

What are the effects of this heuristic? As N increases?

Computing MIC

What are the effects of this heuristic? As N increases?

Equitability: lower scores for data not matching equipartition requirement

Computing MIC

What are the effects of this heuristic? As N increases?

Equitability: lower scores for data not matching equipartition requirement

Power: lower scores means higher power for those relationships which are detected

Computing MIC

What are the effects of this heuristic? As N increases?

Equitability: lower scores for data not matching equipartition requirement

Power: lower scores means higher power for those relationships which are detected

Better algorithms possible?

Bibliography

[1] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.