Overview

- Causal graph refresher
- The problem
- The PC algorithm
  - Description
  - The *pcalg* package and its approach
  - An alternative approach
- Categorical variables
  - Independence and conditional independence
  - log-linear models
- My implementation
- Example

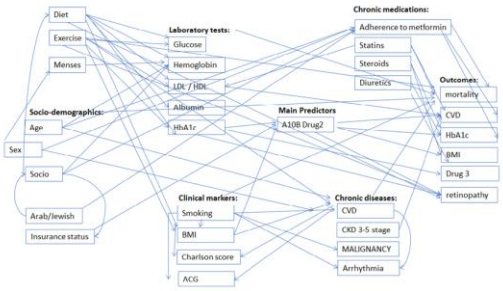
2

Causal Graphs and the PC Algorithm

Yossi Levy
28.5.2018

1

A more realistic example

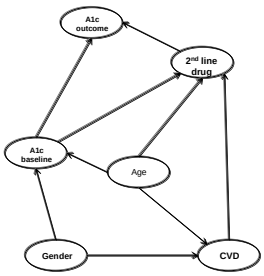


4

Causal Graph

A causal graph is a graphical model used to encode assumptions about the data-generating process

Toy example



3

The problem

- Does the model fit the data?
- Do we really need all the vertices?
- Do we really need all the edges?


```
> head(testdata, 12)
   ID alc_outcome age_at_index_date second_line_drug AlC_last_before_index gender cvd_before_index
1 138    6.233333                66             Alpha                   7.1   Male              No
2 143    6.700000                63     Sufonylureas                    7.4   Male             Yes
3 165    9.300000                65             DPP4                    8.7 Female             Yes
4 277    7.050000                64     Combinations                    7.1 Female              No
5 386    7.000000                65             DPP4                    6.9   Male              No
6 387    7.200000                62             DPP4                    7.3 Female              No
7 530    5.550000                81     Sufonylureas                    5.7   Male             Yes
8 535    6.650000                63             DPP4                    8.1 Female             Yes
9 642    8.400000                60            Other                    9.4   Male             Yes
10 718   7.050000                80     Combinations                    8.3   Male              No
11 719   7.750000                76     Sufonylureas                    7.5 Female              No
12 770   6.866667                64     Combinations                    7.1 Female              No
```

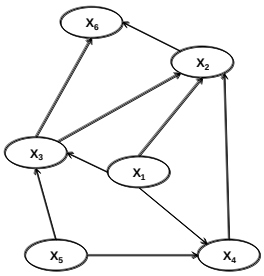

6

Graph formalism

$G=(V,E)$

$V=\{X1, X2, X3, X4, X5, X6\}$

$E=\{ (X1 , X2),$
$(X1 , X3),$
$(X1 , X4),$
$(X5 , X4),$
$(X5 , X3),$
$(X3 , X6),$
$(X3 , X2),$
$(X4 , X2),$
$(X2 , X6) \}$



5

The algorithm

- For all edges (X,Y):
  If $X \perp\!\!\!\perp Y$ - then remove the edge
- For all remaining edges (X,Y) and for all vertices Z:
  If $X \perp\!\!\!\perp Y \mid Z$ then remove the edge
- For all remaining edges (X,Y) and for all pairs of vertices (Z_1, Z_2) :
  If $X \perp\!\!\!\perp Y \mid (Z_1, ; Z_2)$ then remove the edge

And so on…

8

8

The PC algorithm basic idea

- if X and Y are independent, then there is no edge going from X to Y (or the other way)
- Extension: : if X and Y are conditionally independent given Z, then there is no edge going from X to Y (or the other way)
- And so on….

7

7

How to test for independence

The *pcalg* approach:

- Two categorical variables : chi-square test – sounds ok
- Two quantitative variables: assume X and Y have a bivariate normal distribution, then test for r(X,Y)=0
- A categorical variable X and a quantitative variable Y: ?

- It is unclear how they test for conditional independence

10

10

The pcalg package

- There is a *pcalg* package for R
- We were unable to install it due to security issues
- The manual is 167 page long
- It does wonderful things, but…
- I don't like the approach of independence testing

9

9

Different approach

Instead of significance tests, use association measures

- Two categorical variables: Crammer's V, etc.
- Two quantitative variables: Pearson's correlation coefficient
- A quantitative variable and a categorical variable: ICC

- If the values of the association measure is close enough to zero we declare no association
- What is "close enough to zero"?

12

12

Problems

Fundamental problem

- If we do not reject the null hypothesis it does not imply that the null hypothesis is true

Even when ignoring the fundamental problem:

- When sample size is large almost all p-values will appear as significant
- Therefore we will not remove any edges, so why bother?
- Remedy: set α to be small – but how small is small enough?
- What about multiplicity issues?

11

11

Categorical variables example

- We cannot reject the independence hypotheses when conditioning on hospital

Clinical trial results: Hospital 1

	Failure	Success	Total
Treatment A	17 (21.8%)	61 (78.2%)	78
Treatment B	160 (23.3%)	527 (76.7%)	687
Total	177 (23.1%)	588 (76.9%)	765

Pearson's Chi-squared test with Yates' continuity correction
X-squared = 0.024024, df = 1, p-value = 0.8768

Clinical trial results: Hospital 2

	Failure	Success	Total
Treatment A	97 (45.3%)	117 (54.7%)	214
Treatment B	119 (49.4%)	122 (50.6%)	241
Total	216 (47.5%)	239 (52.5%)	455

Pearson's Chi-squared test with Yates' continuity correction
X-squared = 0.59218, df = 1, p-value = 0.4416

14

14

Categorical variables example

Clinical trial results

	Failure	Success	Total
Treatment A	114 (39%)	178 (61%)	292
Treatment B	279 (30.1%)	649 (69.9%)	928
Total	393 (32.2%)	827 (67.8%)	1220

Pearson's Chi-squared test with Yates' continuity correction
X-squared = 7.7901, df = 1, p-value = 0.005253

- The independence hypothesis is rejected

* Be aware that the large sample size has an impact on the p-value.

If the sample size is cut by a half and the percentages remain the same, then the result will not be statistically significant.

It is possible that the study is over-powered. Another possibility is that the effect size is larger than expected, but it is not likely.

13

13

Model formulation

$$\hat{m}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

$$\Rightarrow \log \hat{m}_{ij} = -\log n + \log n_{i\cdot} + \log n_{\cdot j}$$

$$\Rightarrow \text{Model}:$$

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \quad i = 1,2 \; j = 1,2$$

$$u_{1(1)} + u_{1(2)} = 0, \quad u_{2(1)} + u_{2(2)} = 0$$

16

16

Log linear models

- An approach to model association between categorical variables
- Expected cell counts are modeled as log(m_{ij})=….

		X_2		
		1	2	
X_1	1	n_{11}	n_{12}	$n_{1\cdot}$
	2	n_{21}	n_{12}	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	n

- Expected cell counts under independence assumption:

$$m_{11} = n \cdot P(X_1 = 1) \cdot P(X_2 = 1)$$
$$\cong n \cdot \frac{n_{1\cdot}}{n} \cdot \frac{n_{\cdot 1}}{n} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{n}$$

15

15

3 way table

Hospital 1

	Failure	Success	Total
Treatment A	17	61	78
Treatment B	160	527	687
Total	177	588	765

Hospital 2

	Failure	Success	Total
Treatment A	97	117	214
Treatment B	119	122	241
Total	216	239	455

18

18

General model for 2x2 table

Saturated model:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \quad i = 1,2 \; j = 1,2$$
$$u_{1(1)} + u_{1(2)} = 0, \quad u_{2(1)} + u_{2(2)} = 0$$
$$u_{12(1j)} + u_{12(2j)} = 0 \text{ for } j = 1,2$$
$$u_{12(i1)} + u_{12(i2)} = 0 \text{ for } i = 1,2$$

Therefore: the hypothesis of independence between X_1 and X_2 is equivalent to

H_0: $u_{12(11)}=0$

17

17

Conditional independence model

- [12][13]: Conditional Independence of X_2 and X_3 given X_1

$$\log m_{ujk} = u + \\ + u_{1(i)} + u_{2(j)} + u_{3(k)} + \\ + u_{12(ij)} + u_{13(ik)}$$

20

20

Models for 3 way tables

- [123]: Saturated model

$$\log m_{ijk} = u + \\ + u_{1(i)} + u_{2(j)} + u_{3(k)} + \\ + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + \\ + u_{123(ijk)}$$

- [1][2][3]: Independence model

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

19

19

Back to conditional independence model

- [12][13]: Conditional Independence of X_2 and X_3 given X_1

$$\log m_{ujk} = u + \\ + u_{1(i)} + u_{2(j)} + u_{3(k)} + \\ + u_{12(ij)} + u_{13(ik)}$$

How to test for conditional independence?
1. Estimate the parameters of the model
2. Calculate chi-square test for goodness of fit

Recall that
1. if the sample size is large we will get small p-values
2. If we fail to reject conditional independence hypothesis it does not imply the hypothesis is true

Therefore we will use an association coefficient: Cramer's V

22

22

Two other possible models

- [1][23]: X_1 is independence of {X_2 , X_3}

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}$$

- [12][13][23]: No third order interactions

No clear interpretation

$$\log m_{ijk} = u + \\ + u_{1(i)} + u_{2(j)} + u_{3(k)} + \\ + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$

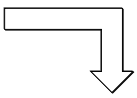
21

21

What about quantitative variables?

- Binning

	ID	age_at_index_date	alc_outcome
1	138	66	6.233333
2	143	63	6.700000
3	165	65	9.300000
4	277	64	7.050000
5	386	65	7.000000
6	387	62	7.200000
7	530	81	5.550000
8	535	63	6.650000
9	642	60	8.400000
10	718	80	7.050000
11	719	76	7.750000
12	770	64	6.866667
13	1122	68	6.475000
14	1450	71	7.200000
15	1588	80	6.966667



	age_at_index_date				
alc_outcome	(24.9,39]	(39,53]	(53,67]	(67,81]	(81,95.1]
(2.98,6.17]	424	1771	3879	2800	589
(6.17,9.34]	1887	12291	26389	14813	2474
(9.34,12.5]	457	1807	1859	536	88
(12.5,15.7]	50	147	132	36	7
(15.7,18.9]	3	7	9	4	1

24

24

Cramer's V

$$V = \sqrt{\frac{G^2/n}{min(\dim(T)) - 1}}$$

- T = the contingency table
- n = number of observations
- G^2 = log-likelihood chi-square

- V ranges from 0 to 1
- V=0 implies no association
- V=1 implies full association
- 0<V<1 : open to interpretation

23

23

Number of bins may affect V (2)

```
> t0
                    alc_outcome
age_at_index_date (2.98,6.17] (6.17,9.34] (9.34,12.5] (12.5,15.7] (15.7,18.9]
       (24.9,39]          424        1887         457          50           3
       (39,53]           1771       12291        1807         147           7
       (53,67]           3879       26389        1859         132           9
       (67,81]           2800       14813         536          36           4
       (81,95.1]          589        2474          88           7           1
> cond.ind.v(t0)
2 iterations: deviation 3.637979e-12
[1] 0.07868904
> t1
                    alc_outcome
age_at_index_date (2.98,6.17] (6.17,9.34] (9.34,12.5] (12.5,15.7] (15.7,18.9]
       (24.9,42.5]          643        3272         704          74           5
       (42.5,60]           3414       24146        2710         205           9
       (60,77.5]           4246       25168        1158          78           9
       (77.5,95.1]         1160        5268         175          15           1
> cond.ind.v(t1)
2 iterations: deviation 3.637979e-12
[1] 0.08716309
> t2
                    alc_outcome
age_at_index_date (2.98,6.17] (6.17,9.34] (9.34,12.5] (12.5,15.7] (15.7,18.9]
       (24.9,48.3]         1282        7839        1408         125           8
       (48.3,71.7]         5758       38274        2954         216          12
       (71.7,95.1]         2423       11741         385          31           4
> cond.ind.v(t2)
2 iterations: deviation 7.275958e-12
[1] 0.09617058
>
```

A "smarter" binning function is needed

26

26

Number of bins may affect V (1)

```
> t0
                    age_at_index_date
alc_outcome       (24.9,39] (39,53] (53,67] (67,81] (81,95.1]
  (2.98,6.17]           424    1771    3879    2800        589
  (6.17,9.34]          1887   12291   26389   14813       2474
  (9.34,12.5]           457    1807    1859     536         88
  (12.5,15.7]            50     147     132      36          7
  (15.7,18.9]             3       7       9       4          1
> cond.ind.v(t0)
2 iterations: deviation 3.637979e-12
[1] 0.07868904
> t1
                    age_at_index_date
alc_outcome       (24.9,39] (39,53] (53,67] (67,81] (81,95.1]
  (2.98,6.17]           424    1771    3879    2800        589
  (6.17,9.34]          1887   12291   26389   14813       2474
  (9.34,12.5]           457    1807    1859     536         88
  (12.5,18.9]            53     154     141      40          8
> cond.ind.v(t1)
2 iterations: deviation 3.637979e-12
[1] 0.09080929
> t2
                    age_at_index_date
alc_outcome       (24.9,39] (39,53] (53,67] (67,81] (81,95.1]
  (2.98,6.17]           424    1771    3879    2800        589
  (6.17,9.34]          1887   12291   26389   14813       2474
  (9.34,18.9]           510    1961    2000     576         96
> cond.ind.v(t2)
2 iterations: deviation 1.818989e-12
[1] 0.1110151
>
```

25

25

R code

```
# run pc algorithm
# I use the value of tol=0.08 as an arbitrary cut off for the sake
# of the example. In real life we will probably need to test the sensitivity
# of the run to the tol value

G= dag2G(g)
tol=0.08
p=length(G[[2]])
for (nz in 0:(p-2)){
  nedge=length(G[[2]])
  print(paste("step ", as.character(nz), ": ", as.character(nedge), " edges", sep="" ))
  sink(file="temp")
  v= test.graph(G, nz, testdata)
  sink()
  if (max(v)==1){
    print (" No more edges that can be rmoved")
    break
  }
  G=modify.graph(G,v, tol)
  nedge=nedge-length(G[[2]])
  print(paste("step ", as.character(nz), ": ", as.character(nedge), " edges removed", sep="" ))
}


newg = G2dag(G)
plot(graphLayout(newg))
```

28

28

Our implementation

- Bin quantitative variables into k bins, that is, reduce the problem into categorical data analysis problem
- Fit log-linear model for (conditional) independence to the data, and calculate the log-likelihood chi-square statistic
- Use Cramer's V to determine association instead of significance testing

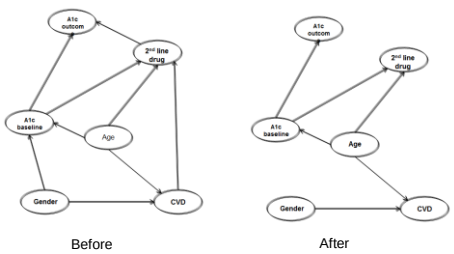
Questions

- How many bins?
- What will be the threshold for association?

27

27

Output: modified graph



Before After

30

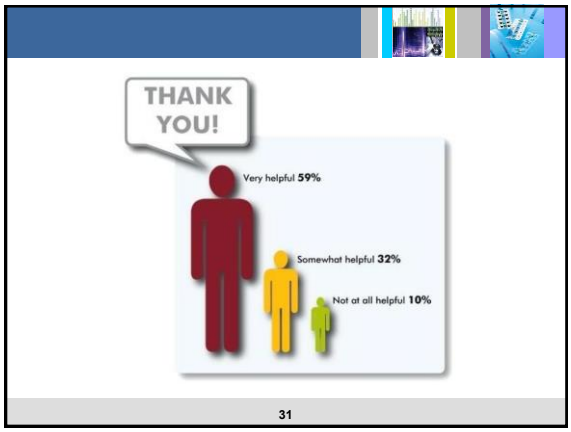
30

R run

```
> print(paste("Initinal graph has", as.character(length(G[[1]])), "vertices and", as.character(length(G[[
2]])),"edges"))
[1] "Initinal graph has 6 vertices and 9 edges"
> for (nz in 0:(p-2)){
+   nedge=length(G[[2]])
+   print(paste("step ", as.character(nz), ": ", as.character(nedge), " edges", sep="" ))
+   sink(file="temp")
+   v= test.graph(G, nz, testdata)
+   sink()
+   if (max(v)==1){
+     print (" No more edges that can be rmoved")
+     break
+   }
+   G=modify.graph(G,v, tol)
+   nedge=nedge-length(G[[2]])
+   print(paste("step ", as.character(nz), ": ", as.character(nedge), " edges removed", sep="" ))
+ }
[1] "step 0: 9 edges"
[1] "step 0: 2 edges removed"
[1] "step 1: 7 edges"
[1] "step 1: 1 edges removed"
[1] "step 2: 6 edges"
[1] "step 2: 0 edges removed"
[1] "step 3: 6 edges"
[1] "step 3: 0 edges removed"
[1] "step 4: 6 edges"
[1] " No more edges that can be rmoved"
> print(paste("Final graph has", as.character(length(G[[1]])), "vertices and", as.character(length(G[[2]]
)),"edges"))
[1] "Final graph has 6 vertices and 6 edges"
>
```

29

29


THANK YOU!
Very helpful 59%
Somewhat helpful 32%
Not at all helpful 10%
31

31