Homework 11: Very PC Urban Economics

36-402, Spring 2016

Due at 11:59 pm on Thursday, 28 April 2016

This assignment requires the use of the PC algorithm, which we went over in class on 11 April and is described in detail in Chapter 28. You will need to install the `pcalg` package; for the easiest visualization of the resulting graphical models, you will also need to install the `Rgraphviz` package (which is not on CRAN but on Bioconductor). Since the latter can be somewhat tricky to install, it is acceptable to draw theoretical or estimated graphical models using other means.

We return to the data set on urban economies from homework 5.

1. A simple theory, supported by some of the original researchers on urban scaling, is that increasing population causes higher per-capita income, and also separately causes more of the city's economy to be in high-value industries, such as the four industries contained in the data set. Population, on this theory, is the common cause of all the other variables in our data set.

   (a) (4) Draw the directed graphical model. (If you believe more than one graph is compatible with the statement of the theory, explain why, and draw just one of the graphs.)

   (b) (4) In this model, is per-capita dependent on the share of the city's economy derived from finance, or independent? Are those two variables conditionally independent given population? Conditionally independent given population and the share of management in the city's economy?

   (c) (5) Check those three predictions of dependence or independence against the data.

2. A different theory is that high-value industries tend to be sited in larger cities to have access to more customers. According to this theory, then, population causes industry shares, and industry shares cause per-capita income, but there is no direct effect of population on income.

   (a) (4) Draw the directed graphical model. (If you believe more than one graph is compatible with the statement of the theory, explain why, and draw just one of the graphs.)

(b) (3) In this model, are population and per-capita income dependent? Find a set of one or more variables we could condition on which make population and per-capita income independent. (At least one such set exists, even if you think that population and per-capita income are unconditionally independent.)

(c) (2) In this model, are the shares of finance and of information technology independent, given population? Were they independent given population in the previous model?

(d) (5) Find at least one conditional independence relation, involving both population and per-capita income, which should be *different* between the two models, i.e., variables which are conditionally independent in one model but conditionally dependent in the other.

(e) (5) According to the data, is per-capita income independent of population given the shares of all four industries? Should it be, on this theory?

(f) (5) Explain at least one drawback to testing this theory with *this* particular data set.

3. Yet a third theory is that different cities acquire different industries more or less by chance (access to supplies or geographic advantages, successful early entrants to the market, good policy, dumb luck, etc.); that some industries pay much better than others; and that people move to places where the income level is high, and are pretty indifferent to everything else about the city[1].

   (a) (4) Draw the directed graphical model. (If you believe more than one graph is compatible with the statement of the theory, explain why, and draw just one of the graphs.)

   (b) (2) Find a conditional independence statement which is true of this model and the model from problem 1, or explain why there are none.

   (c) (2) Find a conditional independence statement which is true of this model and the model from problem 2, or explain why there are none.

   (d) (2) Find a conditional independence, involving both population and per capita income, which holds in this model but does not hold in either the model from problem 1 or that from problem 2.

   (e) (5) Does the data support the conditional independence from problem 3d?

4. Using the `pc` function from the `pcalg` package, with the Gaussian conditional independence test and $\alpha = 0.05$, find an estimate of the causal model. *Hint:* There are quite a few missing values.

[1] Or they care about so many distinct things, for so many distinct reasons, that they look indifferent in the aggregate.

Note: In this problem and the next, if you believe that `pc` is reporting that more than one graph is compatible with the data, repeat any estimates, explanations, etc., for every equivalent graph. (You may summarize, if some parts are the same across all or most graphs.)

   (a) (1) Draw the graphical model

   (b) (4) Explain, in words, the resulting causal structure.

   (c) (5) Linearly regress each variable which has parents (in the estimated graph) on those parents, and report the coefficients and their standard errors. Can you make sense of the signs?

   (d) (5) What, according to the estimated model, would be the effect on per-capita income of doubling the population of an average-sized city? What would be the effect on per-capita income of increasing the share of information technology in the city's economy by 10 percentage points?

   (e) (2) Explain, in words, what the conditional independence test presumes about both the marginal distribution of each variable, and about the shape of the relationship between pairs of variables.

   (f) (3) Are those presumptions plausible for this data? (Give evidence, not just opinions.)

5. Repeat the estimation from the previous problem, but replacing both population and per-capita income with their logarithms.

   (a) (1) Draw the graphical model.

   (b) (4) Explain, in words, how the graph differs from the previous one (if at all).

   (c) (5) Re-calculate the effects you estimated in problem 4d.

   (d) (3) Are the presumptions of the conditional independence test satisfied after this transformation? (Again, given evidence.)

6. (5) Explain why at least one crucial assumption of the PC algorithm is probably not met for either problem 4 or 5. Either suggest a way you could use this data to check the assumption (without implementing it), or explain why you cannot.

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.

Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.

EXTRA CREDIT (5): Write a function to implement the nonparametric conditional independence test based on the bandwidths of kernel density estimates described in the text. Make sure your function works with `pcalg`. Re-do problems 4 and 5. What changes?

EXTRA CREDIT (5): Re-do problems 4 and 5 using the `rfci` algorithm from the `pcalg` package, and carefully explain the results. (You will get no extra credit just for running the algorithm, or even just for producing pictures.)