Homework 2: Financial Blocks

36-720, Fall 2016

Due at 11:59 pm on 21 September 2016

Notation: This problem set is about block models for n-node networks, with k blocks. *Assume that the graph is directed.* Write Z_i for the block to which node i belongs, and b_{rs} for the probability of an edge from a node in block r to one in block s; call the $k \times k$ matrix $\mathbf{b}$ the **affinity matrix**. The number of nodes in block r will be n_r, and the number of edges from r to s e_{rs} (random variable, E_{rs}). In a stochastic block model, the number of nodes in each block is also random, and the random variable is N_r.

Data set: our data set for this assignment, `http://www.stat.cmu.edu/~cshalizi/networks/16-1/hw/2/fj.txt` comes from economic history. For each country in the data set, for the years 1890, 1900 and 1910, it records whether financial newspapers in that country quoted *local* exchange rates for "bills of exchange" (roughly, checks) drawn on banks located in the other country. This is a directed network, since, e.g., Chinese newspapers might list exchange rates for British currency without British newspapers reciprocally quoting prices for Chinese currency. Many historians have, heuristically, divided the economies of the period into a "core" and "periphery", with peripheral countries connected to each other via the core. We write $A_{ijt} = 1$ if country i lists the currency of country j in year t.[1]

1. *Inference* Suppose that we have the adjacency matrix $\mathbf{A}$, that we know the block assignment vector Z, and that we wish to do inference on the affinity matrix $\mathbf{b}$.

   (a) (3) Derive the log-likelihood function for $\mathbf{b}$. Show that it takes the form of an exponential family, with natural sufficient statistics e_{rs}, and find the natural parameters.

   (b) (2) Derive the maximum likelihood estimator of $\mathbf{b}$, i.e., derive the MLE for each b_{rs}. Call this $\hat{\mathbf{b}}$.

   (c) (5) Show that $\hat{\mathbf{b}}$ is consistent and asymptotically Gaussian, and find the asymptotic variance. (You may take the central limit theorem for IID variables as given.)

[1] Citations to the data source, and its initial analyses by economic historians, will be given in the solutions.

(d) (5) Find the covariance $\text{Cov}\left[\hat{b}_{rs}, \hat{b}_{qt}\right]$. (If you can only find an asymptotic covariance, explain why.)

(e) (5) Describe the shape and location (in $\mathbb{R}^{k \times k}$) of an asymptotic confidence region for $\mathbf{b}$, with coverage level $1 - \alpha$.

2. *Transitivity* A graph is said to be **t**ransitive if two nodes which are both connected to a third node are more likely to be connected than two randomly-chosen nodes, i.e., if $\mathbb{P}(A_{ik} = 1 | A_{ij} = 1, A_{kj} = 1) > \mathbb{P}(A_{ik} = 1)$.

   (a) (2) Show that
   $$\mathbb{P}(A_{ik} = 1) = \sum_{r=1}^{k} \sum_{s=1}^{k} b_{rs} \frac{n_r n_s}{n^2} \tag{1}$$
   Hint: See notes from class on 14 September.

   (b) (3) Show that
   $$\mathbb{P}(A_{ik} = 1 | A_{ij} = 1, A_{kj} = 1) = \sum_{r,s,q} b_{rq} \mathbb{P}(Z_i = r, Z_j = s, Z_k = q | A_{ij} = 1, A_{kj} = 1) \tag{2}$$
   Hint: See notes from class on 14 September.

   (c) (5) Express $\mathbb{P}(Z_i = r, Z_j = s, Z_k = q | A_{ij} = 1, A_{kj} = 1)$ in terms of $\mathbf{b}$ and n_r, n_q, n_s. *Hint:* Bayes's rule.

   (d) (5) Suppose that the affinity matrix takes the special, simple form $b_{rr} = \rho p$, $b_{rs} = p$ for $r \neq s$. Find necessary and sufficient conditions for the model to show transitivity. These conditions will at least involve ρ, and may also involve p, k, n, and the n_r.

3. *Initial data examination* Load the data and use it to prepare three graphs, for 1890, 1900 and 1910. Note that only the columns `country_A`, `country_B`, `quote1890`, `quote1900` and `quote1910` are relevant for us.

   (a) (5) What are the densities of the three graphs? The diameters? The average pairwise geodesic distance? The number of triangles? The number of two-stars? What fraction of edges are reciprocated?

   (b) (5) For 1900, calculate the in-degree, the eigenvector centrality and the betweenness centrality of each country. Plot these three centralities against country names (alphabetically), and make scatterplots of the centralities against each other. Comment.

4. *Core and periphery*

   (a) Consider a two-block model, where one block, the core, consists of France, Germany and Great Britain, and all other countries belong to the other block, the periphery.

      i. (1) Draw the graph of the 1900 network, with nodes colored according to their block.

ii. (5) Report the maximum likelihood estimate of **b**, along with (asymptotic) 95% confidence limits, for 1890, 1900 and 1910.

iii. (4) Draw three graphs depicting this block model in the three data-collection years, with weighted edges and self-loops.

(b) Consider a three block model where the core remains the same, but the periphery is split into two blocks, one (say the "intermediates") containing Austria-Hungary, Belgium, Switzerland, Spain, Italy, the Netherlands, Russia, and the USA.

i. (1) Draw the graph of the 1900 network, with nodes colored according to their block.

ii. (5) Report the MLE in the same way you did for the two-block model.

iii. (4) Draw graphs depicting this block model, as you did for the two-block model.

5. *Two blocks or three?*

(a) (5) Find the log-likelihood of the graph for 1900 under the two- and three- block models given in the previous problem. Explain why the three block model *must* have a higher likelihood.

(b) (5) Write a function to fit the two- and three- block models to new graphs (with the same node set), and return the difference in log-likelihoods. How do you know that the code works?

Hints: You can assume the nodes are either always labeled (with the same labels), or always given in the same order. Also, if it is more convenient to work from an adjacency matrix rather than a graph object, that's OK too.

(c) (5) Write a function to generate a new graph, on the same set of nodes, using the affinity matrix you estimated from the two-block model for 1900. How do you know your code works? (It is OK to generate an adjacency matrix rather than a graph object.)

(d) (5) By repeated simulation, find the distribution of the log-likelihood difference when the graph is generated from the two-block model.

(e) (5) Find a p-value for testing the null hypothesis that the two-block model is correct, against the alternative of the three-block model.

(f) *Extra credit (5)*: Is the distribution of log-likelihood differences from your simulation χ^2? If so, with how many degrees of freedom? If not, can you explain why it isn't?

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.