Statistics of Big Data, Fall 23/24

Homework 3

Due date: 14 March 2024 **before class**

1. **Detecting signal in noise**

   In this exercise we seek to identify some signal hidden in high dimensional noise. The file *covtrain.csv* contains a matrix X of $n = 1000$ observations of dimension $p = 500$. Data were generated from the model Boaz Nadler used in his talk:

   $$x \sim \mathcal{N}(0_p, \sum_{j=1}^{K} \lambda_j v_j v_j^T + I)$$

   with $K << p$ dimension of signal. Note that this also assumes that $v_j \perp v_l$ for $j \neq l$, and that we used $\sigma^2 = 1$ for simplicity.

   Our task is to investigate the eigen decomposition of $X^T X/n$ (or PCA of X) to try and find K, the directions, and relate it to the theory and results presented by Boaz.

   (a) Plot the empirical distribution of the eigenvalues of $X^T X/n$ and compare it to the null distribution under the Marchenko-Pasteur law. What do you conclude about the likely number of identifiable non-null signals in this data?

   (b) Compare the top eigenvalues to the magnitude $(1 + \sqrt{p/n})^2$ expected if signal is below the "phase transition" threshold. Are your conclusions similar?

   (c) Now project the matrix X on the 10 top eigenvectors/PCs $\hat{v}_j$ (by multiplying each row by $\hat{v}_j$), and calculate the norms of these vectors. How are they related to the corresponding eigenvalues? Explain it algebraically.

   (d) Next read another independent matrix drawn from the same distribution in *covtest.csv.* Perform the same 10 projections for this matrix and calculate the norms. Explain the results in light of your findings in the previous items.

   (e) (* Extra credit) Next, can we infer on the nature of the vectors v_j?
   Hint: The structure is relatively simple.
   You can use any graphical, intuitive or other method to try and figure it out, but to get credit you then need to find a way to justify your guess in a relevant measurable way.

   Some code hints for this problem are in the file *pca.r.*

2. **Statistical network modeling**

   (a) For the $p1$ model, we described a model with the following potential predictors (in addition to overall rate θ): popularity (indegree) β_i per node, friendliness/expansiveness (outdegree) α_i per node, and overall reciprocity (mutuality) ρ . We said that the likelihood is complex because of the normalization factor λ_{ij} and therefore the usual solution is to use pseudo-(log)likelihood:

$$PL(\theta,\alpha,\beta,\rho\ ;\ Y) = \sum_{i\neq j} \log\left\{Pr(Y_{ij}|Y_{ji}\ ;\ \theta,\alpha,\beta,\rho)\right\}.$$

As we showed maximizing PL is a logistic regression, and the function pstar in R also offers statistical inference as if it was truly a logistic regression, although PL is not a proper log-likelihood.

Assume now we choose to fit a model without mutuality, that is we assume $\rho = 0$, explain accurately and rigorously with specific formulas why we can now solve the proper maximum likelihood problem as a logistic regression. Explain what this means for the inference (Wald statistics, AIC, etc.) that the pstar function offers in this case.

(b) Simulate a 100-node Erdos-Renyi graph with ($G(N,p)$ or $G(N,E)$) with 10% of nodes (for example, rate parameter $p = 0.1$) and plot the resulting distribution of node ranks. Relate this distribution to a known distribution and explain the relation.

(c) Similarly simulate 100-node Preferential attachment graph, with an initial 10 nodes with $G(10,0.1)$ connectivity, then each new node connects to 10 existing nodes, with rate proportional to their current connections (for example, you can randomly draw 10 existing edges, and connect the new node to one of their ends randomly). Plot the distribution of node ranks and compare it to the one of Erdos-Renyi, and explain the difference.

(d) (* Extra credit) Investigate the E-coli network (data(ecoli) in R), and offer various insights, based on sound visual or mathematical arguments, for example:

- Viewed as an undirected graph, does it resemble a small-world preferential attachment network or an Erdos-Renyi type graph?
- How do indegrees and outdegrees of nodes related to each other, what does it tell us about the nature of the network? Which would be more important variables in a $p1$ model?
- If you embed the entire network or subnetworks (to make it computationally feasible) using ergmm, what do you learn about the structure of the network?
- etc.