Homework 3: the Axis of History?

36-467, Fall 2020

Due at 6 pm on Thursday, 24 September 2020

The data set `http://www.stat.cmu.edu/~cshalizi/dst/20/data/soccomp.irep1.csv` contains values for nine different measures of social complexity of a range of historic and pre-historic polities around the world, compiled by a large team of historians, archaeologists and anthropologists. Each row records measurements (or estimates) for the following variables:

- `NGA`, the "natural geographical area" where the polity was located;
- `PolID`, an abbreviated name for the polity;
- `Time`, the year (in terms of the common era) for the data point;
- `PolPop`, the total population of the polity;
- `PolTerr`, the territory controlled by the polity;
- `CapPop`, the population of the polity's capital;
- `levels`, a composite of the number of hierarchical levels for government[1], religious figures, settlements, etc.;
- `government`, a composite measure[2] of the extent to which the polity had professional soldiers, judges, a legal code, a permanent bureaucracy, etc.;
- `infrastructure`, a composite measure of the extent to which the polity built and maintained various forms of infrastructure (irrigation, canals, bridges, roads, etc.);
- `writing`, a composite measure of the sophistication of the polity's system of writing or other ways of permanently storing information;
- `texts`, a composite measure of the extent to which the polity composed and propagated literature about a range of advanced subjects (calenders, religious rituals, history, fiction, etc.);

[1] A modern American example would be that each town or city has a government, under a county government, under a state government, under the federal government.

[2] More precisely, the people who collected the data had a list of 11 features of government, and coded them as present or absent for each polity; the `government` variable is the average. Details will be available in the solutions.

- `money`, a composite measure of the extent to which the polity used various forms of money.

1. *Initial explorations*

   (a) (3) What is the earliest date for a data point? What is the most recent data point? What it the median date?

   (b) (3) Calculate a table of summary statistics for the nine complexity measures.

   (c) (5) The summary statistics for the population variables `PolPop` and `CapPop` should look strange. These variables have been transformed from their original values[3]. Explain how we can be sure that these numbers are not the actual populations.

   (d) (2) Make a guess about what the transformation was, and check your guess. (We will use the transformed values, so you do not need to get this right to complete the rest of the homework.)

   (e) (5) Find the all the covariances between the complexity measures, and display it as a table. (Don't show too many decimal places.) *Hint:* What does `var(x)` do when `x` is an $n \times p$ matrix?

   (f) (2) Create a similar table of the correlations. *Hint:* What does `cor(x)` do?

2. *Principal components of historical complexity* Perform a principal components analysis of the complexity measures.

   (a) (5) Explain why, in this case, it makes sense to scale the variables going in to the PCA to all have variance 1.

   (b) (5) Make a plot of the R^2 we would get from using the $1, 2, \ldots 9$ principal components. The R^2 for using all 9 should be exactly 1; why? How many PCs do we need to capture 75% of the variance? To capture 90%?

   (c) (5) Display the first three principal component vectors. (You can decide whether this is better done as a table, or as one or more figures.)

   (d) (5) Describe, in words, what kind of polity would get a high (very positive) score on PC1, and what kind of polity would get a low (very negative) score.

   (e) (5) Describe, in words, the kinds of polities that would get high and low scores on PC2.

3. *PC1 over time.* All these sub-problems will be easier if you first add the scores on PC1 to the data frame as a new column.

[3] This is not mentioned in the paper describing the data.

(a) (5) Make a single plot, containing all the data points, showing time on the horizontal axis and score on PC1 on the vertical axis. If there are any patterns, describe them; if not, say that.

(b) (10) Make *separate* plots of PC1 over time for the following areas: Cahokia; Kachi Plain; Latium; Middle Yellow River Valley; Niger Inland Delta; Susiana. For each of these six plots, describe its shape in words.

(c) (2) What are the modern countries corresponding to these six areas? (A single area may correspond to more than one modern country.) — *Hint:* You'll want an atlas or an encyclopedia.

(d) (5) Describe, in words, the common pattern to the shape of the six plots, or explain why there isn't a common pattern.

(e) (5) Describe, in words, what this pattern says about how the polities changed over time. (If you *convincingly* argued in the previous problem that there is no common pattern, these are free points.)

4. *PC2*

(a) (2) What is the correlation between scores on PC1 and PC2?

(b) (5) What should the correlation be theoretically? Is the difference between the theory and the calculated value a cause for concern?

(c) (5) Make a plot of scores on PC1 against scores on PC2. Describe any patterns.

(d) (5) Based on the previous plot, are the two PCs statistically independent of each other? Is this a problem?

5. (1) How much time did you spend on this problem set?

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, without dangling or useless commands. All parts of all problems are answered with coherent sentences, and raw computer code or output are only shown when explicitly asked for.

Extra Credit: Multiple Imputation, Multiple Analyses

Because information about many of the variables is missing for many polities, missing values have been filled in by means of a technique called **multiple imputation**. The details are not important for our purposes[4], but what matters is that the filled-in or imputed values were somewhat random. The data set we have been working with is just one of the twenty that were created. You can find the full data at `http://www.stat.cmu.edu/~cshalizi/dst/20/data/soccomp.csv`, which has all the same columns as the one you've worked with, plus an extra column, `irep`, which records the imputation run (or "replication"). The data we have used so far is just the `irep==1` rows.

1. (1) Re-do the PCA, using only the `irep==19` rows. Report the first 3 PC vectors, in the same format you used in Problem 2c. How much, if at all, would your interpretations differ from those you gave in Problems 2d and 2e?

2. (2) For each imputation replication, re-run the PCA you did in Problem 2. Do not report all 20 PCAs. Instead, report the means of the vectors for the first 3 PCs, and give the standard deviation for each coordinate of each PC. Again, use the same table or figure format you used in Problem 2c. *Hints:* Doing one PCA of all the data at once isn't the right thing here. (Why not?) Instead, wrap what you did for Problem 1 in a `for` loop. (Alternatively, if you know about "split-apply-combine", use that.)

3. (1) Based on the last problem, do you need to change any of your interpretations of the principal component vectors from Problems 2d and 2e?

4. (1) Now that you have done 20 PCAs, you have twenty different scores on PC1 for each data point. Re-do the plots from Problem 3b to show the mean complexity over time, $\pm$ 2 standard deviations. *Warning:* This is likely to be tricky/annoying to code.

[4] In brief: when variable A was missing but variables B, C, and D were present, a model was fit to all of the data points which had complete records for variables A, B, C, D. This was then used to predict the missing value of A, which was perturbed by noise, whose distribution was also estimated from the complete records.