Leo's talk

Leonardo Comandini

May 3, 2017

1 An intermediate goal

Our dataset is large, but we need to refine it to perform the techniques we have been shown.
So after the first phase of exploration we want to extract a matrix without missing values.
We did several attempts, but we are presenting only one. It's a rather simple idea but in this situation it does the job.

2 Fix the best year

First of all, we decide to temporarily get rid of the time dimension to make the things a little easier.
We decided to select the year with highest number of indicators to have a wider selection among them.

3 Shrink the countries with too few indicators

Then we computed the number of indicators for each countries and we dropped the countries with the lowest number of indicators.
We have been very conservative in fact some states that are still in are not so significative, like 'St. Kitts and Nevis'.

4 Select the indicators among the fullest

After that we did the opposite: for each indicator we compute the number of countries then we ordered them starting from the fullest.
(hist with yellow bars)
So now we can carefully have a look at the best, let's say, 100 indicators and among them select manually the ones we want to base our analysis on.

5 Next steps

After the selection we need to do some further steps:

1. we need to have a look at the matrix that won't be 100% full.
2. there could be some indicators or countries that have several missing values, we should drop them.
3. at this point it still may happen that there are some missing values. The reason is probably related to technical difficulties in collecting the data.
We need to fill them with various techniques, for instance we can us the value at the previous year.
4. after all this steps we can go back on the time dependency.
After this process we have a selection of indicators and one of countries, instead of selecting values only for 2010 we are doing the same things for all the years in a given interval (for instance 2000-2010).
After some further fixing we should have a full 3d matrix, upon which we can perform PCA for each year and find some dynamics.