A DATA-DRIVEN APPROACH TO BUILD FINANCIAL INDEX

Alessandro Bitetto and Paola Cerchiello

Fintech laboratory, Department of economics and management, University of Pavia, Pavia, Italy

Workshop on "Big Data & Machine Learning Applications for Central Banks", Bank of Italy

Rome, October 21-22, 2019

Motivation

Macroeconomic variables are often used to assess financial stability for countries.

To this aim synthetic indexes are typically created based on expert-judgement assumptions (e.g. weighted average).

However, all indexes can be questionable and can lead to endless debate on which one should be used as a robust financial indicator.

Here we present a data-driven statistical approach to build financial index based on intrinsic information of data.

We analyze a set of Financial Soundness Indicator (FSI) provided by International Monetary Fund ranging from 2007 to 2017 and for most of worldwide countries, including both strong and developing economies.

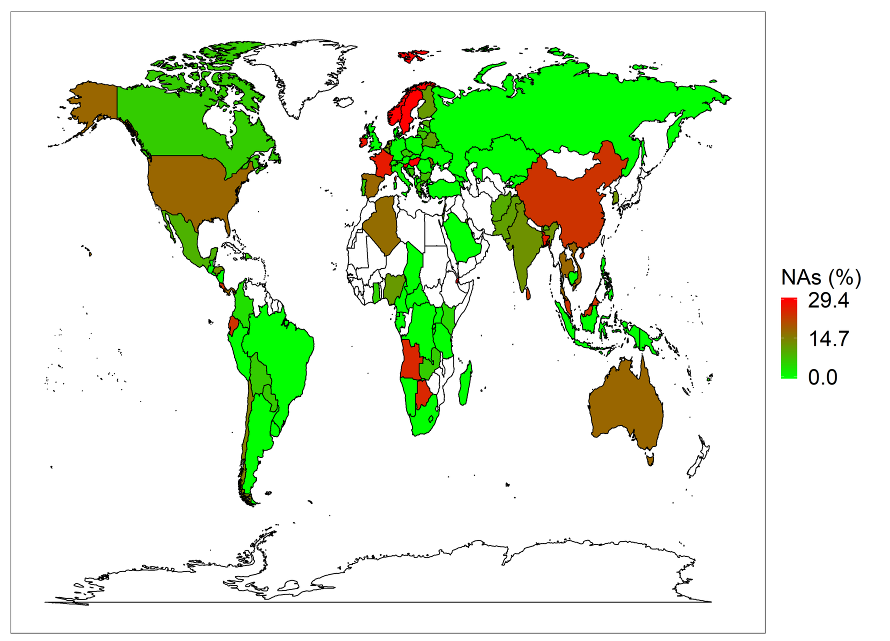
We assess data quality and recovered some missing data, experimenting with different techniques.

We test two methodologies to build the index: a **PCA-based approach** used to create a low dimensional (1 to 2 way) indicator, whereas a **network-based approach** used to estimate weights to average the FSI.

Dataset

Data used consists of:

- Financial Soundness Indicators (FSI) provided by International Monetary Fund (IMF) ranging from 2007 to 2017 and for most of worldwide countries, including both strong and developing economies for a total of 119 countries and 17 FSI (Interest margin to gross income, Return on assets, Non performing loans net of capital provisions, etc)
- 6 Hofstede Indicators (Individualism, Masculinity, etc) for each country, fixed for all years
- 2 Geographical Indicator (Latitude and Longitude)
- Final dataset has $n = 119$ countries with $p = 25$ variables for $T = 7$ years



Missing values have been recovered by comparing different methodologies:

- NIPALS
- Matrix Completion with Low Rank SVD
- Bayesian Tensor Factorization

Matrix Completion performed best

Time series of each country have been differenciated to ensure stationarity

Methodology

As the data have 3 dimensions, *Country*, *Variables* and *Time*, two complementary techniques have been used:

(a) Principal Component Analysis (PCA) to model country/variables interaction, for each year. PCA aims to create one or more index variables from a larger set of measured variables, where each index is a linear combination of the Y original variables. The model is an equation $C = w_1 Y_1 + \ldots + w_4 Y_4$

(b) Factor Analysis (FA) to model country/time interaction, for all variables. FA models the measurement of latent variables, seen through the relationships they cause in a set of Y variables. The model is a set of equations $Y_i = b_i F_1 + u_i, i = 1, \ldots, 4$

The following PCA techniques have been tested for each year:

- **PCA**
- **Robust PCA**: decompose M by solving

$$\text{minimize } \|L\|_* + \lambda\|S\|_1 \quad \text{subject to } L + S = M$$

  where $\|L\|_*$ is the nuclear norm
- **Robust Sparse PCA**: minimize

$$f(A,B) = \frac{1}{2}\|X - XBA^T - S\|_F^2 + \psi(B) + \gamma\|S\|_1$$

  where B is the sparse loading matrix, A is orthonormal, ψ is a regulizer (i.e. LASSO or Elastic Net) and S captures outliers

Robust PCA performed best with an average (over years) Explained Variance of $46 \pm 3\%$ for the first 2 PC

Due to small depth of each FSI time series the following FA approach has been used:

- Fit a Dynamic Factor Model

$$\begin{cases} \mathbf{F}_t^i = \mathbf{A}^i \mathbf{F}_{t-1}^i + \mathcal{N}(0, \mathbf{Q}^i) \\ \mathbf{y}_t^i = \mathbf{C}^i \mathbf{F}_t^i + \mathcal{N}(0, \mathbf{R}^i) \end{cases}$$

  for each of n country, obtaining *factor matrices* F^i, *factor interactions* A^i and *factor loadings* C^i, $i = 1, \ldots, n$
- Fit a Vector Auto Regressive (VAR) model in order to get $\hat{A}$ lag-1 matrix that incorporates cross-countries interaction of A^i
- Use Kalman Filter to get smoothed factors $\widehat{F^i}$ using $\hat{A}$ and $\hat{C} = diag(C^i)$ in order to get latent factors that incorporates cross-countries interactions

Optimal number of factors has been set to 2 with Y-reconstruction error validation

Index Validation

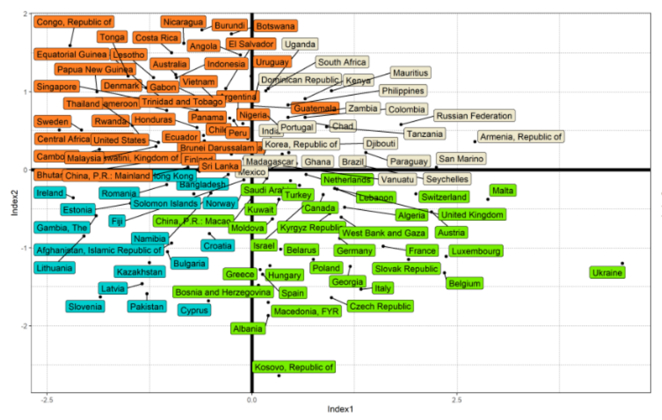
Both methodologies produce continuous value for the 2 components of the index In order to get a binary index the following procedure has been followed:

- set a threshold and get the binary index, i.e. 0 or 1
- perform a regression task where target is a economic variable (such as *GDP* or *Non Performing Loans*) and regressors are the 2 binary using different partitioning algorithm, such as *Random Forest* and *Gradient Boosting Machine*
- evaluate prediction accuracy and outliers for different threshold

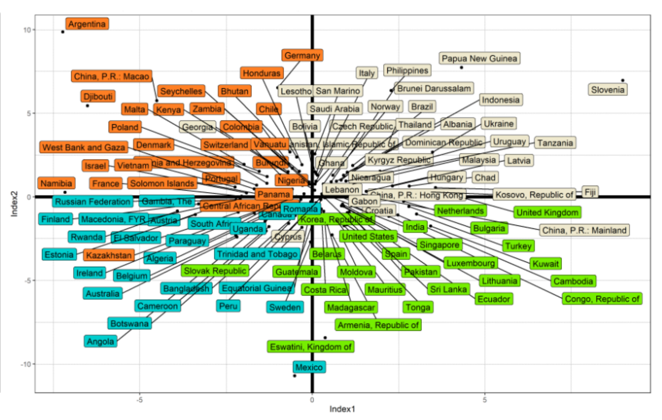
Robust threshold has been set to 0 for both indices

Results and future work

Results from both methodology can be visualized by scatter plot and clusterized by their binarized index value:

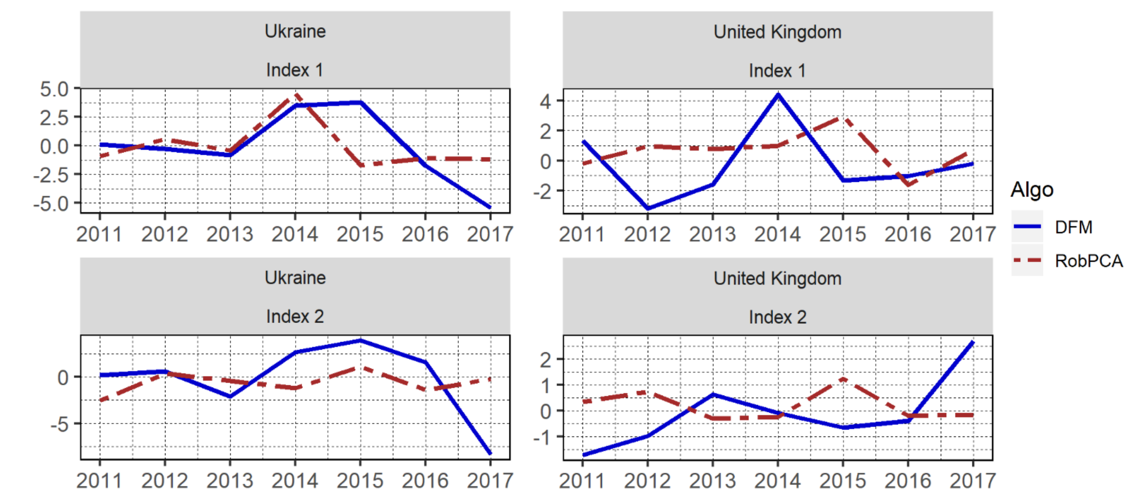


Robust PCA index for 2014



DFM index for 2014

Temporal evolution of both index can highlight analogies or differences:



Index evolution over years for Ukraine and United Kingdom

Index must be compared with other economical indices and find meaningful economical explanation

Index predictive power must be futher validated in regression/classification tasks

Additional methodologies involving *Network Theory* should be tested for comparison:

- *Factorial Graphical Model* for a time-independent estimation
- *Time Series Chain Graphical Model* for time-dependent estimation

and centrality measures could be used as FSI weights

References

1. Hastie T., Mazumder R., Lee J. D., Zadeh R., (2015) Matrix completion and low-rank SVD via fast alternating least squares
2. Khan S. A., Ammad-ud-din M., tensorBF: an R package for Bayesian tensor factorization
3. Candes E. J., Li X., Ma Y., Wright J., (2009) Robust Principal Component Analysis?
4. Erichson N. B., Zheng P., Manohar K., Brunton S. L., Kutz J. N., Aravkin A. Y., (2018) Sparse Principal Component Analysis via Variable Projection
5. Holmes E. E., WardE. J., Scheuerell M. D., (2018) Analysis of multivariate time-series using the MARSS package