

Statistical Learning Presentation

Exploring World Bank country classifications

Krystian Filippo Maria Bua

Damiano Di Francesco

Tancredi Salamone

PhD Program - Sant'Anna School of Advanced Studies

Motivation I

- Each year the World Bank classifies the world's economies into **four income groups** — high, upper-middle, lower-middle and low - by considering Gross National Income (GNI) per capita (current US\$)¹
- Of course, the indicator is a good proxy for the overall economic development level of a country. Nonetheless, the picture can be more complex: countries within the same income group may still vary a lot in different aspects
- Our '*research questions*': Is it fair to use only GNI per capita to classify very heterogeneous economies? If the income of a country turns out to be the most important welfare indicator, can we predict its values using a bunch of other socio-economic indicators provided by the WB?

¹World Bank Data Team (2021). New country classifications by income level: 2021–2022. See also Fantom and Serajuddin (2016).

Motivation II

- Starting from a group of various indicators selected from the *World Bank database* we aim to
 - apply **factor analysis** (PCA) to see what dimension these indicators could represent, followed by **cluster analysis** to attempt to re-classify these economies
 - employ a mix of dimension reduction, regularization (shrinkage) as well as subset selection methods to build a battery of prediction models for world economies' GDP per capita (PPP)
 - The idea is to select the 'best' one in terms of test Mean squared error performance, estimated using a k -fold cross-validation approach

Our dataset

- The *World Development Indicators* (*WDI*) is a collection of comparable statistics about global development.
- We extract data for **165** countries ranging from highly developed economies to emerging ones.
- **70** indicators are selected from a variety of topics: Economy & Growth, Education, Environment, Gender, Health, Social Development, Trade, Social Protection, Labor and Urban Development.
- In order to avoid 'scale issues', all the indicators are scaled using their mean μ and standard deviation σ .
- We choose **2015** as the cross-section because is the year with less missing data. Still, **1.5%** of our observations are missing.

Dealing with missing values

Algorithm 12.1 *Iterative Algorithm for Matrix Completion*

1. Create a complete data matrix $\tilde{\mathbf{X}}$ of dimension $n \times p$ of which the (i, j) element equals

$$\tilde{x}_{ij} = \begin{cases} x_{ij} & \text{if } (i, j) \in \mathcal{O} \\ \bar{x}_j & \text{if } (i, j) \notin \mathcal{O}, \end{cases}$$

where \bar{x}_j is the average of the observed values for the j th variable in the incomplete data matrix \mathbf{X} . Here, \mathcal{O} indexes the observations that are observed in \mathbf{X} .

2. Repeat steps (a)–(c) until the objective (12.14) fails to decrease:
 - (a) Solve

$$\underset{\mathbf{A} \in \mathbb{R}^{n \times M}, \mathbf{B} \in \mathbb{R}^{p \times M}}{\text{minimize}} \left\{ \sum_{j=1}^p \sum_{i=1}^n \left(\tilde{x}_{ij} - \sum_{m=1}^M a_{im} b_{jm} \right)^2 \right\} \quad (12.13)$$

by computing the principal components of $\tilde{\mathbf{X}}$.

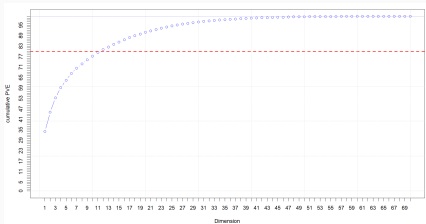
- (b) For each element $(i, j) \notin \mathcal{O}$, set $\tilde{x}_{ij} \leftarrow \sum_{m=1}^M \hat{a}_{im} \hat{b}_{jm}$.
- (c) Compute the objective

$$\sum_{(i,j) \in \mathcal{O}} \left(x_{ij} - \sum_{m=1}^M \hat{a}_{im} \hat{b}_{jm} \right)^2. \quad (12.14)$$

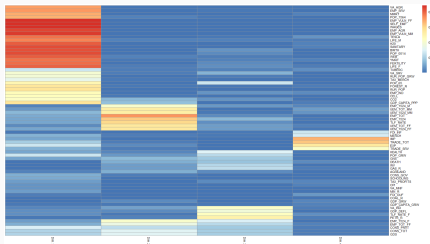
3. Return the estimated missing entries \tilde{x}_{ij} , $(i, j) \notin \mathcal{O}$.
-

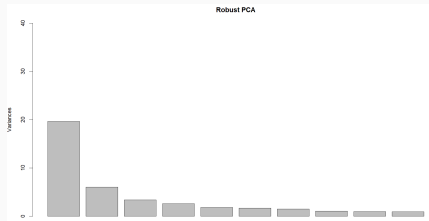
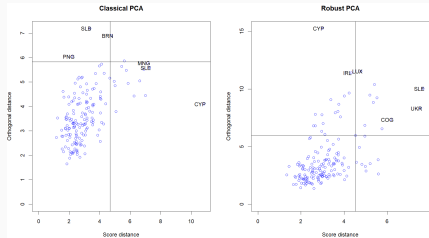
Missing values
imputation using
iterative PCA
cfr. James et al.
(2013)

- With PCA, we are looking for a subspace that grasp as much as possible of the variation in the original dataset.
- We will also employ PCA as a pre-processing step in our supervised analysis (e.g. PC regression).
- PCA is very sensitive to *outliers*. Indeed, outliers may drive the PCA components. So as to avoid misleading interpretations, a robust exercise is performed.
- *In our exercise*, we proceed as follow. First of all, we determine the number of PCs. Then, we try to interpret them in terms of the original variables. Finally, identification of outliers and robustness analysis.



- Number of components: **11**
- Factor's meaning (PVE > 5%):
 - **Factor 1** – Quality of life: access to essential services (*tesla, cell, web*), sanitation and vulnerability (*sanitary, H2O, mortality rates*).
 - **Factor 2** – Employment situation: *young, female and male*.
 - **Factor 3** – Industrial development: *value added industry, rents and GDP*.
 - **Factor 4** – Economic openness: *trade total, imports, FDI and exports*.





- Orthogonal and score distances are used to find outliers and extreme objects.
- Outliers detected by PCA: 2.
- Outliers detected by RPCA: 10
- The PVE by each PC changes in Robust PCA.
- *Main Issues* of PCA:
 - The complexity of the PC model may be assessed using different (and arbitrary) rules.
 - Distances are dependent on the number of PCs.

Clustering I: HC and K-means

- Here, we are going to focus on **hard partitioning** algorithms.
- Looking at the *HC* dendrogram, we detect **two** clusters as two branches that occur roughly at about the same horizontal distance. ▶ Dendrogram
- These two clusters divide economies into: *high & upper-middle income* and *lower-middle & low income*.
- Still, some 'sub-clusters' seem to emerge in all cases, e.g. small but rich countries, oil-rich Gulf monarchies.
- K-means algorithm provides similar results. ▶ K-clustering and silhouette

Clustering II: HCPC, CoV and Fuzzy k-means

- Partitioning through **HCPC** and **CoV** techniques. In particular:
 - **HCPC**: *PCA* on the original data \rightarrow we perform *HC* on the **11** PCs.
 - **CoV**: *HC* on the **70** variables \rightarrow *PCA* on the **4** cluster of variables detected \rightarrow for each cluster, we extract only the PCs with $PVE \geq 80\%$ \rightarrow *HC* on the **17** PCs.
- **Soft clustering** (Fuzzy *k - means*) in which we allow each data point to be part to more than one cluster.

Clustering overview

Table 1: Clustering results for the battery of models.

Model	Hartigan Index	AS	Gap Statistics
AHC	3	2	19
K – means	3	2	3
HCPC	3	2	17
CoV_{AHC}	3	2	18
CoV_{k-means}	3	2	7
Fuzzy k – means	-	2 ²	-

²Fuzzy silhouette index is reported with clustering index value equal to 0.462. Other indexes have been explored: AS, partition coefficient, partition entropy and modified partition coefficient. 2 clusters are confirmed.

Moving to Supervised Analysis

- Our aim: predicting GDP per capita (PPP) of world economies using a bunch of socio-economic indicators provided by the Wold Bank Database
- **Strategy:** run a battery of prediction models comprising
 - **Dimension reduction:** two 'types' of PC regression
 - **Regularization:** Ridge and Lasso
 - **OLS post-Lasso estimation:** simple OLS and BSS
- Models assessment using a **k-fold cross validation** approach

Dimension reduction

- Instead of using the original predictors in a regression setting, we construct the first $M = 17$ principal components and take them as predictors in a linear regression fit using least squares

$$GDP_i = \theta_0 + \sum_{m=1}^M \theta_m z_{im} + \varepsilon_i, \quad i = 1, \dots, n$$

- If the components are chosen wisely, then such dimension reduction framework can outperform OLS
- Two 'types' of PCR:
 - Standard PCR \rightarrow *orthogonal*
 - PCR using as regressors the first principal component of each cluster of variable \rightarrow *mild correlation*

Shrinkage and Post-selection

- First, we perform standard Ridge and Lasso regression
- Second, we use Lasso as a *feature selection* procedure. In particular
 - **OLS post-Lasso**: we apply ordinary least squares (OLS) to the model selected by first-step penalized estimators - in our case Lasso → post-model selection estimators may have smaller bias (Belloni and Chernozhukov 2013)
 - **BSS post-Lasso**: we fit a separate least squares regression for each possible combination of the p predictors that survive Lasso.
 - The best model retain 4 regressors out of 7: CO2 emissions, private consumption, access to internet, and trade in service

► Selection graphs




Assessing models performance

Table 2: Test MSE results for the battery of prediction models, computed with a k -fold cross validation approach (with $k = 10$) and **lambda.1se**.

Model	#var	Test MSE
Ridge	69	0.2888
Lasso	7	0.2667
Lasso + OLS	7	0.2071
Lasso + BSS	4	0.2061
PCR	17	0.2351
PCR_{var}	4	0.3394

- **Unsupervised analysis**
 - Soft partition using Gaussian mixture and HCPC Robust
- **Supervised analysis:** exploit the time series dimension of our dataset
 - pre- vs post-crisis comparison
 - dynamic approach: time series clustering, dynamic PCA
 - GDP forecasting using e.g. dynamic factor analysis

References

-  Belloni, Alexandre and Victor Chernozhukov (2013). “Least squares after model selection in high-dimensional sparse models”. In: *Bernoulli* 19.2, pp. 521–547.
-  Fantom, Neil James and Umar Serajuddin (2016). “The World Bank’s classification of countries by income”. In: *World Bank Policy Research Working Paper* 7528.
-  James, Gareth et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

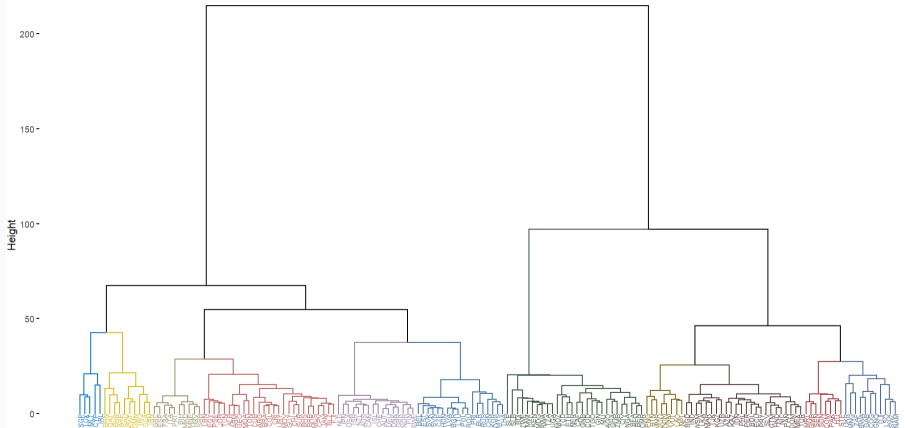


THE reference

Dendrogram

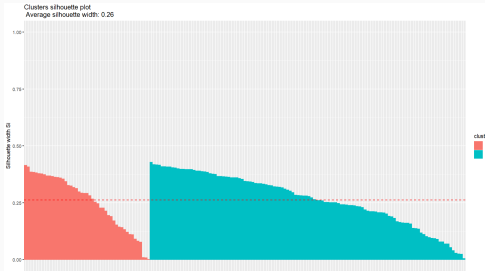
► Clustering I: HC and K-means

Cluster Dendrogram



K-clustering and silhouette

► Clustering II: HCPC, CoV and Fuzzy k-means



Selection graphs

► Shrinkage and Post-selection

