Exploring World Bank country classifications: an (un)supervised analysis?

Krystian Filippo Maria Bua[1], Damiano Di Francesco[1], and Tancredi Salamone[1]

[1] *Sant'Anna School of Advanced Studies*

Abstract

Each year the World Bank classifies countries into different groups according to gross national income per-capita. Since the 2008, enormous changes have taken place. In particular, the World economy has been shaken by the biggest financial crises since the 1930s, with a global re-assessing of economic relations and with increasing concerns of environmental degradation and climate change. This turmoil suggest that a review of the income classification provided by World Bank may be needed. In this work we propose methods that possibly capture the multidimensional nature of development by extending the landscape of indicators to be used for the classification. In particular, we first ask whether is it fair to use only GNI per-capita to classify very heterogeneous economies. Then, we provide different models to exploit the multitude of indicators provided by the World Bank to predict the income per-capita of the selected countries. Finally, we provide suggestions to build upon and extend our findings.

1 Introduction

Each year the World Bank classifies 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, this indicator is to a first extent 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 important aspects. This concern is shared also by Fantom and Serajuddin (2016) who claim that some users of the classification have lamented a number of methodological issues, all converging towards the necessity to include alternative measures incorporating poverty and distributional concerns more explicitly. Building on this point, in this short report we propose methods that possibly capture the multidimensional nature of development by extending the landscape of indicators to be used for the classification.

In particular our research questions can be summarized as follows: 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 World Bank?

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

The remainder of this report is organized as follow. In Section 2 we briefly describe the data set. Then, Section 3 is devoted to the unsupervised setting. In Section 4 we provide the battery of prediction models. Finally, Section 5 concludes with future research avenues.

2 The Dataset

The World Development Indicators (WDI) is a collection of international, high-quality, cross-country comparable statistics about global development and poverty. As pointed out by the World Bank, the full database is compiled from officially-recognized sources and includes national, regional, and global estimates. The database contains 1,400 time series indicators for 217 economies and more than 40 country groups, with data for many indicators going back more than 50 years.

We select 165 countries ranging from highly developed economies to emerging ones and for each country we pick 70 indicators capturing a variety of dimensions: 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 by subtracting the mean μ and by dividing by their standard deviation σ. Even if the data are provided as a panel of time series, we focus our analysis on a cross-section dimension by picking the year 2015. The reason for this choice is that in 2015 there is a greater availability of data. Still, 1.5% of our observations are missing. As we will see in the unsupervised analysis section, we decided to impute missing values to have a complete dataset.

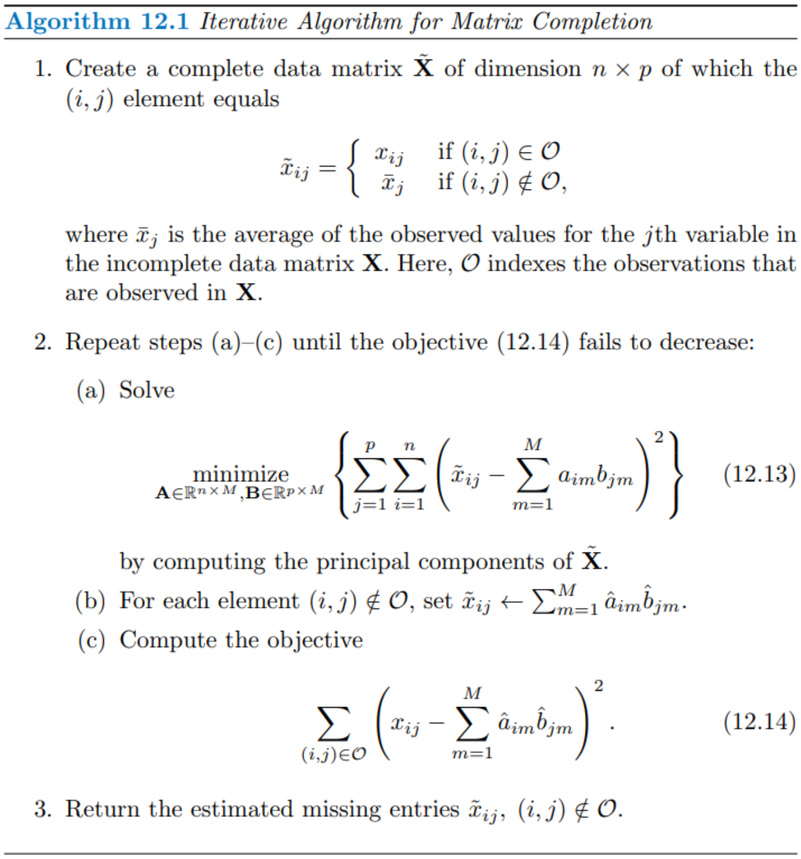
[1] World Bank Data Team (2021). New country classifications by income level: 2021–2022. See also Fantom and Serajuddin (2016).

3 Unsupervised analysis

3.1 Dealing with missing values

Often datasets have missing values, which may be a problem. The existence of missing data may significantly affect our exercise, possibly biasing the results of the supervised and unsupervised analysis. There are numerous techniques we may apply for imputing missing data. For instance, we could remove the rows that contain missing observations and perform our data analysis on the complete rows. Alternatively, we could replace all the missing values with the mean of the observations using the non-missing entries in the original dataset. Although these imputation techniques are frequently used in the literature, they suffer from numerous issues. As a consequence, in this work we adopt a different data imputation method. In particular, we use iterative PCA as as presented in Tibshirani et al. (2021). The advantage of this technique is that it imputes missing values using a greater information set with respect to simply taking the mean of the observations, thanks to the exploitation of principal component analysis.

The algorithm works as follow



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

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

       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

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

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

Figure 1: Iterative PCA algorithm for matrix completion taken from Tibshirani et al. (2021, p. 512)

3.2 Principal Component Analysis

Principal component analysis (PCA) is a mathematical tool that is used to reduce the dimensionality of large data sets. It accomplishes this reduction by identifying new variables, called principal components, which are linear combinations of the original variables and along which the variation in the data space is maximal. Using mathematical projection, principal components define a subspace that retains most of the variation in the original high−dimensional sample. By using a few number of components, each dataset may be represented by relatively few numbers instead of by values for thousands of variables.

PCA is often employed as a pre-processing step in a supervised setting, such as principal component regression (PCR). PCA is also very sensitive to outliers. Indeed, outliers may drive the PCA components and they arbitrarily skew the solution from the desired solution. So as to avoid misleading interpretations, different robust exercises have been proposed in the literature.

In our work, we proceed as follows. First of all, we determine the number of principal components. Then, we try to interpret them in terms of the original variables. In closing, identification of outliers and a robustness analysis are performed.

To get a view of the dimensionality of the data, we begin by looking at the proportion of the variance present within each principal component. This information may be collected in a scree plot of the variances with respect to the dimensions, which is given in Figure 2.

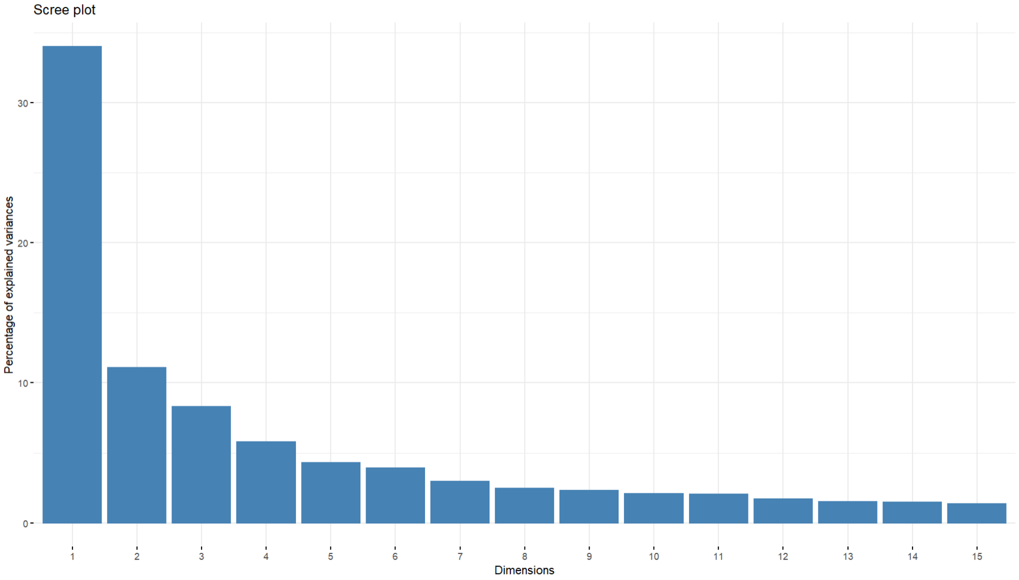


Figure 2: Scree plot

Note that although the first component has more variance than later components, the first dimension retains only 34% of the original variance and roughly 11 components are needed to retain 80% of the original variability. On the other hand, it is interesting to observe that 11 components are enough to retain all the original variance − a much smaller number than the original 70 variables, without losing much information. Looking at the cumulative percentage of variance explained (cumulative PVE) in Figure 3 with an acceptable threshold at 80%, we show that exactly 11 components are enough to represent our original data set.

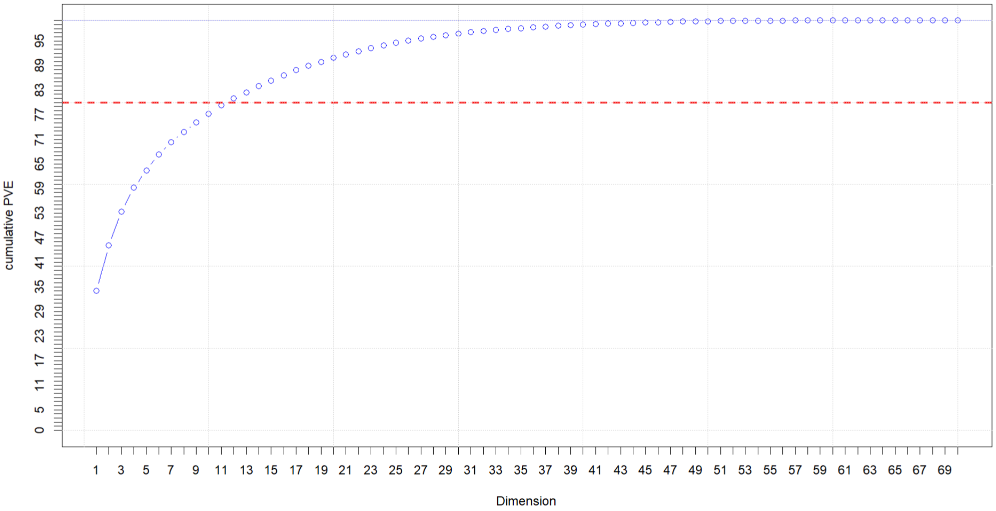


Figure 3: Cumulative PVE and optimal dimensions

Let's take a look of the heat-map of factor loadings, which is basically the correlation coefficient for the variable and component. In short, it shows the amount of variance explained by the indicator on that particular factor. For a matter of space, in Figure 4 we propose only the 4 principal components with a PVE $> 5\%$. Investigating which variables have high correlation with each dimension, we can try to grasp each component's meaning.

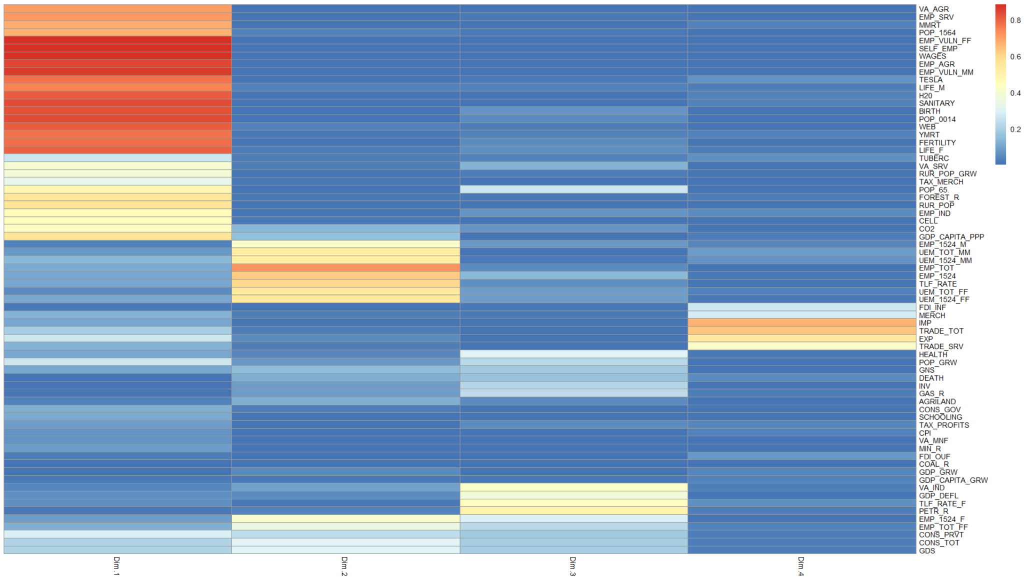


Figure 4: Heat-map of factor loadings

- *Factor 1* – **Quality of life**: this dimension is strongly driven by indicators related to access to essential utilities (*tesla, cell, web*); sanitation and vulnerability (*sanitary, H20, mortality rates*); and health situation (*fertility, life expectancy at birth*).

- *Factor 2* – **Employment situation**: the labour market aspect of the economy is bunched up by this component. In particular, a strong correlation is detected with indicators as *young, female and male employment/unemployment.*

- *Factor 3* – **Industrial development**: a few indicators about *value added industry, rents and GDP* characterize this dimension.

- *Factor 4* – **Economic openness**: this dimension is represented by trade flows, foreign direct investment inflows/outflows and financial capital inflows (*trade total*, *trade services*, *imports*, *FDI* and *exports*).

However, a few indicators have very low correlation with all 4 factors, such as *agriland*, *tax profits*, *population growth* and *investment*. Such low correlation may make sense as these indicators are more like the end-products of many elements in the economy. Thus, they cannot be grouped into the 4 factors above.

The last part of this section is devoted to robust our analysis. As we well-know, standard PCA uses the classical sample covariance matrix for computing the n-th dimensions, but this technique is very sensitive to outliers. In general, an outlier is an observation which does not obey to the pattern of the majority of the data. In order to deal with these outliers, several robust techniques have been proposed in the literature. In this work, we focus on the ROBPCA algorithm[2]. The advantage of using this method is that is suitable both for symmetric distributions and skewed data.

In Figure 5, we present the results. Three types of outliers can be distinguished, namely good leverage points (small orthogonal distance and large score distance), orthogonal outliers (large orthogonal distance and small score distance) and bad leverage points (large orthogonal distance and projection on the PCA space far from the regular data). Notably, from the two outlier-maps Robust PCA is able to detect more outliers than Classical PCA. Moreover, more observations are detected as bad leverage points such as Congo, Ukraine and Luxembourg among others.

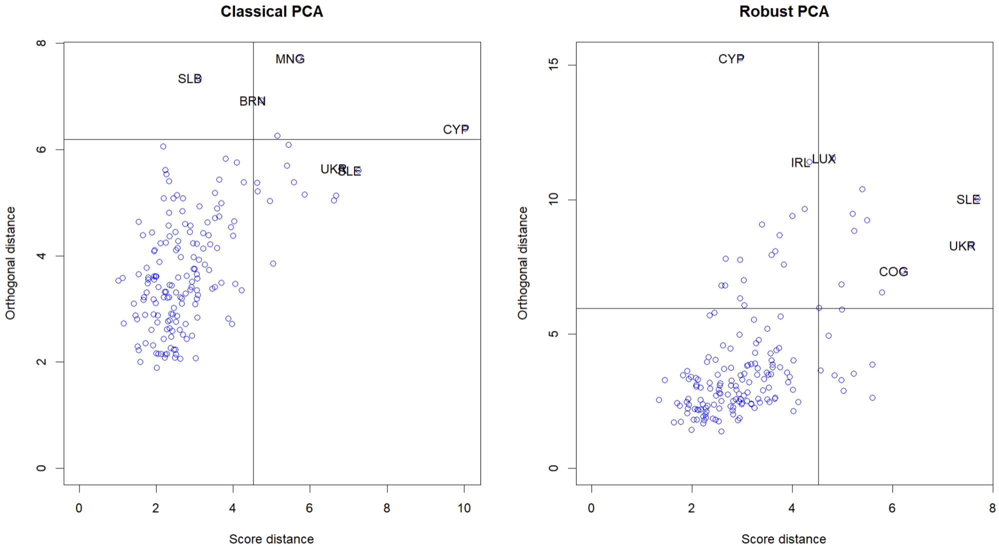


Figure 5: Classical/Robust PCA - Outlier maps

[2] For more about the description of ROBPCA algorithm and outlier sensitivity of PCA see Hubert et al. (2009).

Finally, from the scree plots in Figure 6, although for the first component the variance is still quite high (19.5%), we observe as the total variance of each principal component tend to be lower in the Robust PCA case than Classical PCA.

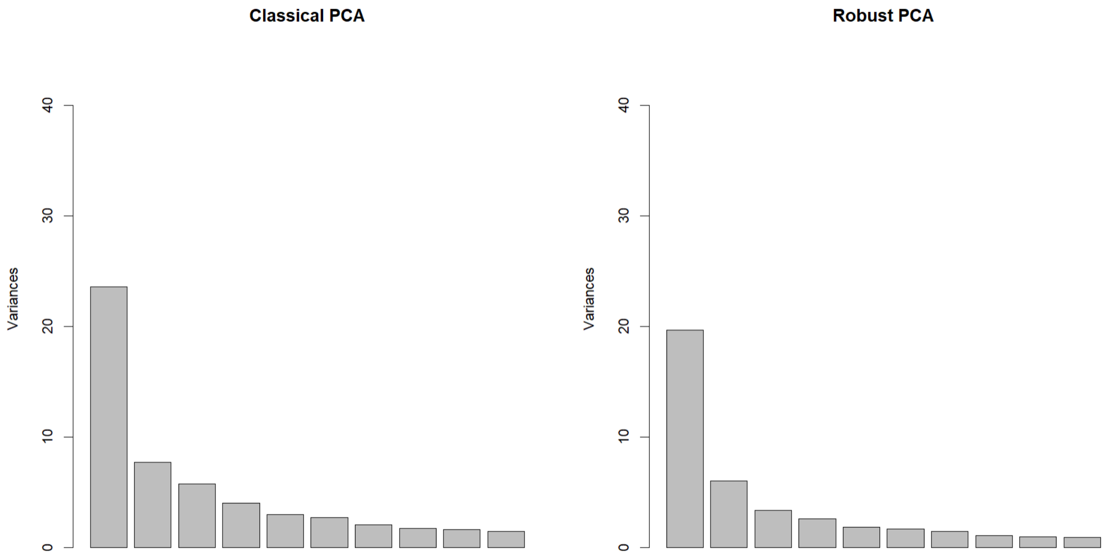


Figure 6: Classical/Robust PCA - PCs variances

3.3 Cluster analysis

Cluster analysis is one of the main techniques for analyzing multivariate data set in an unsupervised setting. The objective is to partition countries into distinct groups with similar characteristics. In what follows, we will first of all presents the most commonly used clustering approaches, i.e. *hierarchical clustering* and *K-means clustering*. Then, some refinements are explored.

3.3.1 Hard partitioning

The first technique we employ is the hierarchical clustering (AHC henceforth), a bottom-up approach which allows to progressively aggregate observations into clusters by considering a dissimilarity matrix. Note that such technique requires to specify both a measure of dissimilarity and a linkage function, that is a function which allows to define the dissimilarity between two groups of observations. As it is standard, we consider the Euclidean distance and use Ward's (1963) linkage function.[3] The resulting dendogram is shown in Figure 7. From a first visual inspection, it seems that two clusters, or possibly three, can be identified. The cluster on the left contains all *high & upper-middle income* countries while the one on the right includes all *lower-middle & low income* countries. Moreover, note that some other 'sub-clusters' consistently emerge, whatever linkage function one uses. Among others, the group of 'oil exporters' (yellow) and that of small, open and advanced economies (light-blue)– e.g. Luxembourg and Singapore – can be easily spotted. To refine this result, we perform K-means clustering, which, unlike the previous technique, requires the user to set the

[3] Note that we also performed the analysis using other linkage functions. The resulting dendograms are, as one would have expected, different, but the main conclusions do not change.

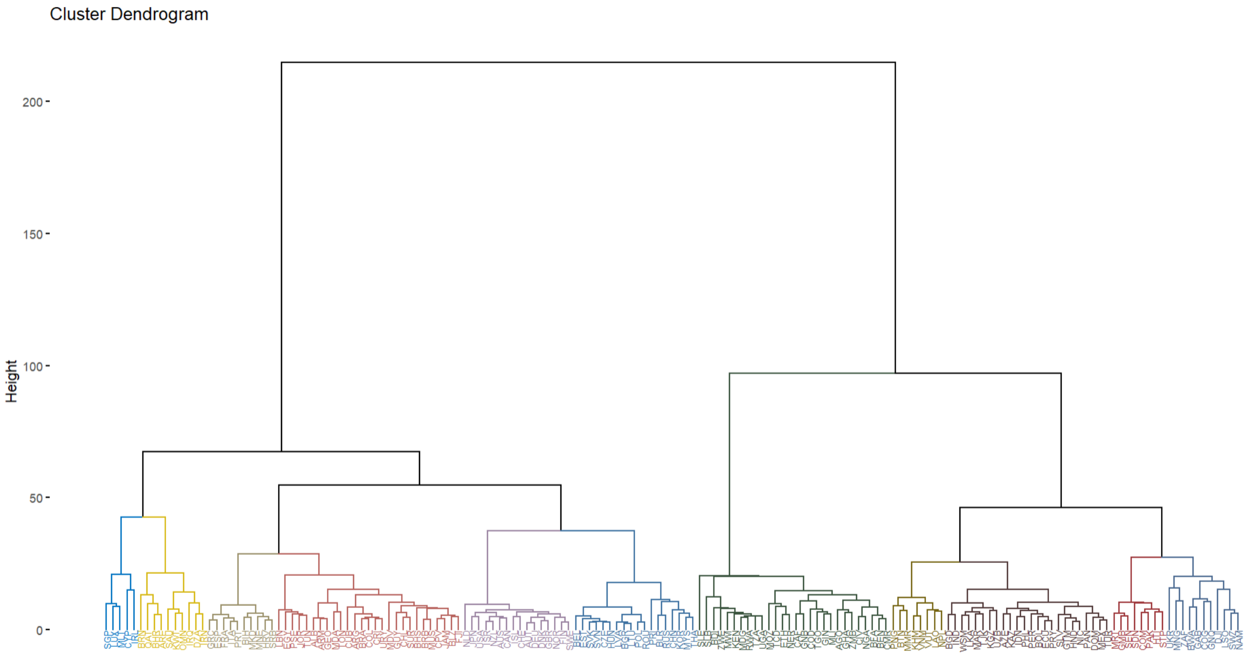


Figure 7: Hierarchical clustering dendogram

desired number of clusters. The algorithm will then choose the best clustering, where *best* means the one which minimizes the *within-cluster variation*. In Figure 8 is shown the result of the k-means with $k = 2$, k being the specified number of clusters. Note that in this particular plot, the two clusters are plotted against the first two principal components. In order to choose the best number of clusters, we performed a battery of tests. In most cases, the *best* number was either 2 or 3, even though some tests provided way different results. In Table 1, we show three out of the many tests we run – the first two rows are referred to the AHC and the k-means, respectively. It can be seen that while the Hartigan index and the average silhouette point to 2-3 clusters, the gap statistics (Tibshirani et al., 2001) suggests 19 clusters in the case of AHC.

3.3.2 Other methods

Unsatisfied with the conclusion that the best number of clusters is either two or three, we go through different avenues to see if this result is robust. First of all, we explore the possibility that the high number of features might make the clustering somewhat unstable. To this end, we use different approaches to reduce the dimensionality of the problem and then perform the methods employed before to this *reduced* data-set.

The first method is to use PCA to shrink the number of features. In particular, we perform AHC on the first 11 principal components, which explain up to 80% of total variability. Although the number of features is considerably smaller than before, the result does not sensibly change. Indeed, as it is reported in Table 1 (HCPC, third row), the detected number of clusters is still either two or three. Note however that also in this case the gap statistics conveys a totally divergent result.

A second approach that we considered is that of reducing the dimensionality by firstly grouping all the variables into clusters; then, on each identified cluster, we perform PCA to extract synthetic measures which

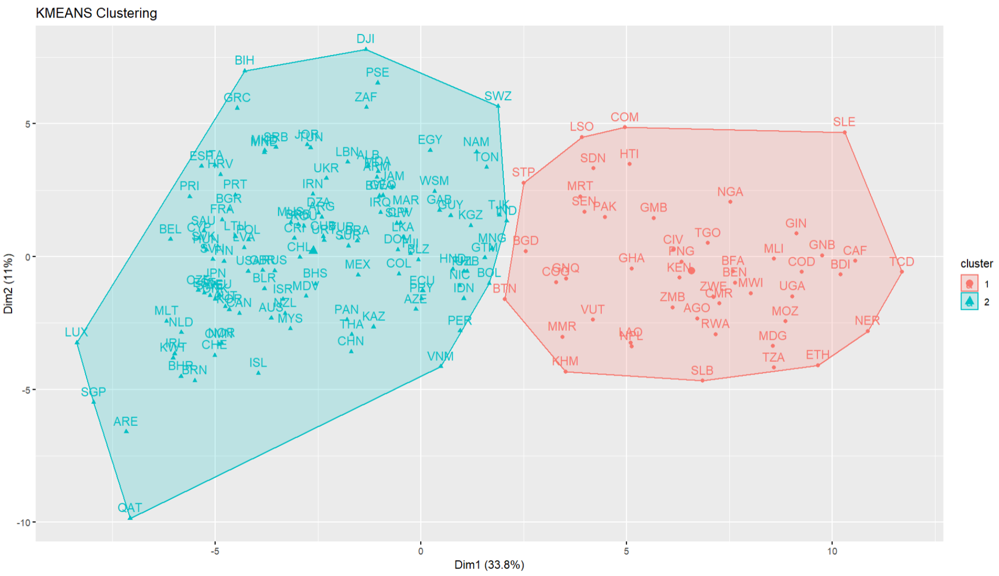


Figure 8: K-means clustering

are then used to cluster the countries. Before commenting the results, it might be interesting to spend few words on the problem of clusterizing variables. The basic idea is to arrange the variables into homogeneous clusters so that variables in each cluster are strongly related to each other. In this particular exercise, we performed the hierarchical clustering algorithm which aims at finding a partition $\mathcal{P}_K$ which maximizes the homogeneity function $\mathcal{H}$ defined as

$$\mathcal{H}(\mathcal{P}_K) = \sum_{k=1}^{K} H(C_k) = \sum_{k=1}^{K} \sum_{x_j \in C_k} r^2_{x_j, y_k} + \sum_{k=1}^{K} \sum_{z_j \in C_k} \eta^2_{y_k | z_j}$$

where y_k is a synthetic variable of the cluster k (i.e. the first principal component), C_k the cluster k, x_j the j-th column of the set of quantitative variables, z_j the j-th column of the set of qualitative variables, r^2 the squared Pearson correlation and η^2 the correlation ratio. Note that r^2 is used to measure the strength of the link between y_k and the quantitative variables in C_k – irrespective of the sign of the relation –, while η^2 is used to measure the link between y_k and the quantitative variables in C_k. For all details, please refer to Chavent et al. (2011). Based on the resulting dendogram (Figure 9), we choose 4 clusters.

On each of the identified clusters, we perform a PCA analysis and extract a number of dimensions which allows to explain at least 80% of the variability, for a total of 17 principal components. On these PCs, we carry out the AHC and k-means algorithm. Again, as shown in Table 1, the results reached before are retained.

Finally, recognizing that countries are very much heterogeneous in a number of aspects and that hard-partitioning may not be appropriate in this case, we explore a soft-partitioning method, namely the *fuzzy k-means*. Contrary to the standard k-means, by applying the fuzzy k-means each data point is not uniquely assigned to one cluster, but it is allowed to have a *fuzzy* memberships to all clusters. Indeed, the result of this algorithm is a NxK (N number of observations and K number of clusters) *membership matrix*, in which each entry gives a *degree* of belonging of country n to cluster k. Note that this algorithm requires the

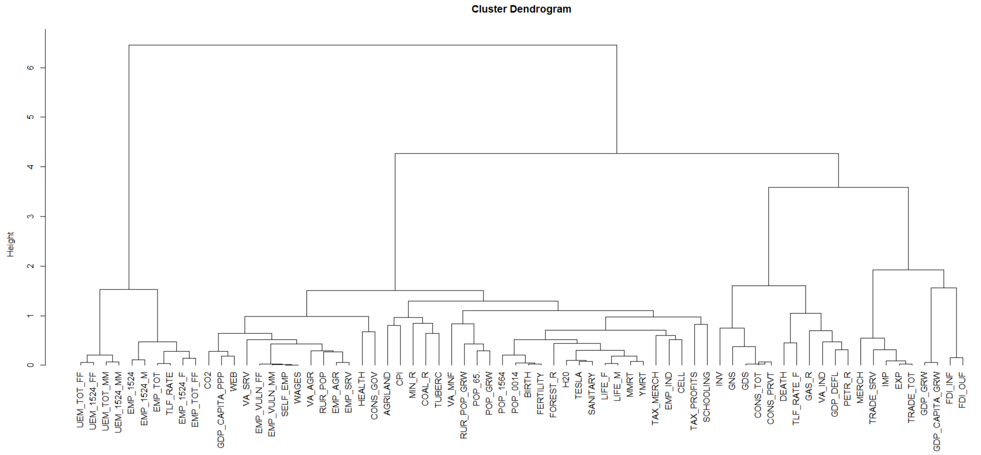


Figure 9: Cluster of Variables

user to set an extra parameter m, the *fuzzifier*, which critically influences the final result. Being this just an exploratory work, we used the standard value $m = 2$. Note however that this is not always the correct choice, as Schwämmle and Jensen (2010) show. Be that as it may, the best number of clusters detected according to the fuzzy silhouette index – a slight modification of the standard silhouette index – is still two.

We thus can conclude that, although some 'sub-clusters' do emerge, only two clusters consistently emerge: one composed by *poor* countries and one by *rich* countries.

Table 1: Clustering results for the battery of models.

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

4 Supervised analysis

Our starting point was to re-classify economies according to a bunch of socio-economic indicators in order to lay out a potentially richer classification than the one based solely on income. Especially from the cluster analysis, it turned out that the income classification is not a so bad representation of the differences among economies. While we inevitably loose some details on the multi-faceted nature of development, the income-based classification provided by the World Bank tends to be very similar to the one we proposed so far. Now, if income is *prima facie* a good proxy for the overall welfare of a country, we can ask whether we can predict its value taking as predictors all the other indicators we picked from the WDI.

Our strategy in this section is then to build a battery of prediction models comprising dimension reduction techniques, standard shrinkage methods and post-Lasso estimation procedures. The best model will be the one displaying the lowest test mean squared error. First we present each prediction model. Then, we explain in detail the cross validation approach adopted. Finally, we comment the results on the MSE.

4.1 Dimension reduction

Our first model is a *principal component regression* (PCR). Instead of using the original predictors in a regression setting, we extract the principal components and use them as predictors in a linear regression fit using least squares

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

where GDP_i is the income per capita in PPP of the country i and z_{im} are the $M < p$ linear combinations of our original p predictors X_j:

$$z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

As argued in Tibshirani et al. (2021), if the components are chosen wisely, then such dimension reduction framework can outperform the simple OLS.

Note that we perform two 'types' of PCR. First, a standard PCR using the first 17 principal components.[4] Then, exploiting the clustering of variables in the unsupervised setting, we also run a PCR in which the regressors are the first principal component of each cluster of variable we obtained. This choice is motivated by the fact that the first principal component is a good synthetic measure of the clusters of variables. On this, one further point is worth mentioning, that is one could also take the centroids of each cluster as the synthetic measure to be considered and run an OLS accordingly.

4.2 Regularization

We also perform standard Ridge and Lasso regressions. We do not spend much words on this part since it is quite standard. The most important estimation procedure is the Post-lasso method presented in the following section. For an extensive treatment on shrinkage methods see Tibshirani et al. (2021).

4.3 Post-model selection estimation

Most importantly, we use Lasso as a *feature selection* procedure. In particular we run a OLS post-Lasso, i.e. we apply ordinary least squares to the model selected by first-step penalized estimators. The rationale for this type of exercise is that post-model selection estimators may have smaller bias with respect to traditional

[4] We run different PCRs with different number of components. It turned out that the best-performing PCR was the one with 17 principal components.

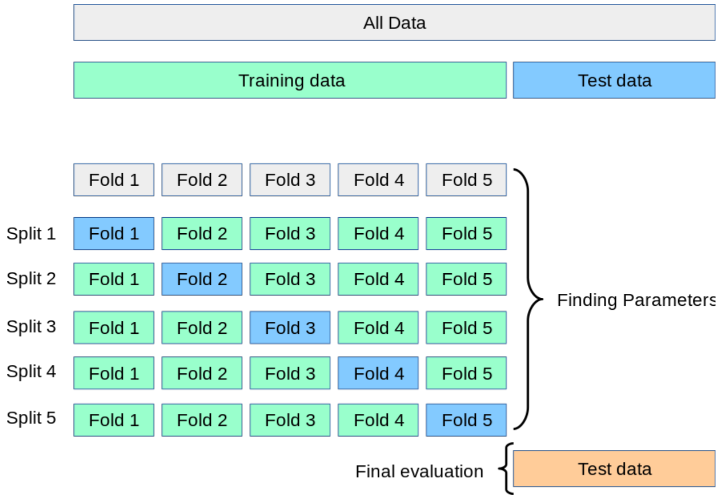


Figure 10: k-fold cross validation procedure.

regression techniques and may outperform Lasso, in the sense of a strictly faster rate of convergence (Belloni and Chernozhukov, 2013).

First we run a simple OLS with the regressors selected by Lasso. Then, more rigorously, we apply a best-subset selection strategy on this set regressors, that is we fit a separate least squares regression for each possible combination of the p predictors that survive Lasso. The 'final' model will chosen according to different criteria, i.e. Adjusted-R^2, Mallow's C_p, BIC and cross-validation.

4.4 Results

Finally, we evaluate our prediction models using a k-fold cross validation approach. A sketchy representation of our methodology is provided in Figure 10. In particular, we split our dataset into a training dataset comprising 75% of observations and a test dataset covering the remaining 25%. Then, we estimate each model in the training set using a k-fold approach (see Tibshirani et al., 2001, for details). We set $k = 10$. The final evaluation of each model is done by calculating the test mean squared error (MSE), i.e. the difference between the predicted values of the estimated models and the actual data of the test dataset.

The results are provided in Table 2. As you can see, the worst model is the PCR of the clustering of variable but also Lasso and Ridge do not perform relatively well. Interestingly enough, post-Lasso estimators outperform standard shrinkage techniques. For instance, if we run a simple OLS on the eight regressors that survived to Lasso we get a test MSE of 0.2071. However, one might argue that the eight-regressors OLS is not the best model. This is why we decided to perform also a best-subset selection on these eight regressors. We see from Table 2 that the best model retains four regressors out of eight, that are CO2 emissions, private consumption, access to internet, and trade in service. This choice has been done on the basis of different criteria as shown in Fig. 10. Even if the technical minimum of some criteria is six, a clear 'elbow' is visible already at four. To wrap up, on the basis of our results, the best prediction model for the GDP per capita (PPP) of the selected economies is thus the OLS post-Lasso with four regressors.

Post-Lasso estimation. The best model retains 4 regressors out of 7: CO2 emissions, private consumption, access to internet, and trade in service.

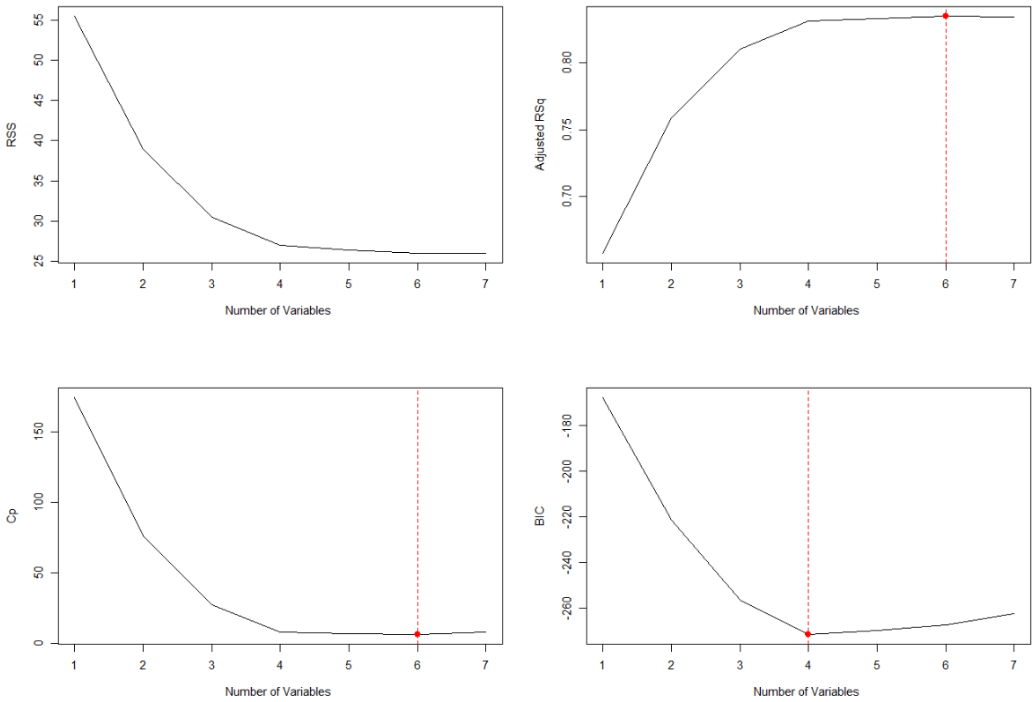


Figure 11

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
$\mathbf{PCR_{var}}$	4	0.3394

5 Conclusions and future plans

In this work, we revisit the World bank countries classification using a statistical learning approach. Selecting a richness of socio-economic indicators from the WDI for 165 economies, we first apply PCA and cluster analysis to provide an alternative - possibly richer - classification with respect the income-based one provided by the World Bank. In the second part of the report, we focus more on a supervised setting. In particular, we build a battery of prediction models for the per-capita GDP of the selected countries and we find that Post-Lasso estimators actually outperform more traditional regression techniques in terms of test MSE, as argued by Belloni and Chernozhukov (2013).

Needless to say, the analysis of this work is rather tentative and partial. One could extend the exercise in several dimensions. First, on the unsupervised setting, we could use other soft-partition clustering techniques as for instance the Gaussian Mixture. Moreover, given the interesting findings of the robust-PCA, it would be interesting to perform HCPC using the robust-PCA instead of the normal one.

Arguably, the most interesting extensions can be done in the supervised setup. As we said at the beginning, we focus our analysis on 2015. However, it is natural to carry out the exercise exploiting the time series dimensions of the World Bank database. First, it would be interesting to perform a 'comparative statics' of the type pre- vs post-financial crisis. In other words, one could do the same exercise selecting the same countries (or a subset of them) in 2006 and, say, 2010 - or in general post-2008/9 - and study their differences. Second, one could adopt a fully dynamic approach and try to predict GDP per-capita using high-dimensional time series techniques such as Dynamic Factor Models. The main difficulty here is that we would have a panel of time series with a lot observations. Thus, we may need to resort to some form of dynamic dimension reduction techniques before moving to forecasting.

6 Bibliography

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Chavent, M., Kuentz, V., Liquet, B., and Saracco, L. (2011). Clustofvar: An r package for the clustering of variables. *arXiv preprint arXiv:1112.0295*.

Fantom, N. J. and Serajuddin, U. (2016). The world bank's classification of countries by income. *World Bank Policy Research Working Paper*, (7528).

Hubert, M., Rousseeuw, P., and Verdonck, T. (2009). Robust pca for skewed data and its outlier map. *Computational Statistics & Data Analysis*, 53(6):2264–2274.

Schwämmle, V. and Jensen, O. N. (2010). A simple and fast method to determine the parameters for fuzzy c–means cluster analysis. *Bioinformatics*, 26(22):2841–2848.

Tibshirani, R., Hastie, T., Witten, D., and James, G. (2021). *An introduction to statistical learning: With applications in R*. Springer.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.