An Empirical Analysis of Growth Regimes: Appendix on Principal Component Analysis

Andrea Mario Lavezzi* Matteo Marsili†

October 27, 2010

1 Introduction

In this Appendix we present the results of the application of Principal Component Analysis (PCA) to the database studied in Lavezzi and Marsili (2010), in order to clarify the relationship between the GM algorithm and this standard procedure for the analysis of multidimensional data, and to compare the results from their application.

PCA is a statistical method for the analysis of multidimensional data composed by correlated variables (define N the number of data, and D their dimension). By applying PCA to the original dataset, a new set of data composed by uncorrelated variables, called *Principal Components* (PC), is obtained (see, e. g., Jolliffe (2002), Ch. I). The principal aim of PCA is *to reduce* the dimensionality of the data, so that they can be expressed by $k < D$ principal components, while at the same time maximizing the amount of variation in the data that the PCs are able to explain. Solving this problem amounts to finding the eigenvalues and eigenvectors associated to the variance/covariance matrix of the variables in the original dataset (Jolliffe (2002), p. 5).

Each eigenvalue of the variance/covariance matrix of the original database corresponds to the variance of a PC. The PCs, therefore, can be ordered on the basis of the value of the eigenvalues (Jolliffe (2002), p. 5). Althought D principal components can in principle be obtained, only a number $k < D$ will be retained on the basis of the variance of each PC. Each new observation will be therefore composed by k principal components, whose values (*scores*)

*Dipartimento di Studi su Politica, Diritto e Società, Università di Palermo. Email: lavezzi@unipa.it.

†Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, Trieste. Email: marsili@ictp.trieste.it.

are calculated through linear combinations of the original variables, with weigths (*loadings*) given by the eigenvector associated to each PC.

Before presenting the results, let us compare the GM algorithm and PCA. Consider our dataset, composed by N objects (countries) of dimension D, the number of *features*.

- Both algorithms utilize information on correlations/covariances. GM, however, exploits the correlation existing *among the objects*, and is based on the corresponding $N \times N$ correlation matrix. PCA, on the contrary, is based on the or covariances *among the features*, and exploits the information contained in the corresponding $D \times D$ covariance matrix.

- While the GM method, being a clustering algorithm, *directly* unconvers information on similarity across the objects, PCA can *indirectly* provide information on similarities across the objects and can represent a: "prelude to ... cluster analysis" (Jolliffe (2002), p. 71). Specifically, in the reduction of objects' dimensionality, the "new" low-dimensional observations may display a cluster structure not visible in the original D-dimensional space. The method of Desdoigts (1999), *exploratory projection pursuit* (EPP), is a generalization of PCA. In particular, EPP reduces the dimensionality of the objects following a criterion to select an "optimal" structure of the data, but the identification of clusters of objects remains partially based on visual inspection.

In the following, however, we show that there is a correspondence between PCA and GM in the study of the relationships among the features (see Section 4.3.2 of Lavezzi and Marsili (2010)). In particular, in PCA information on *loadings* highlights the relative weight of individual features in generating the "new" coordinates. If two features have similar loadings in the PCs, this indicates that there exists a relatively high level of correlation among them, and therefore they provide a relatively similar contribution in the computation of the new coordinates. This may suggest a possibility to identify *clusters of features* from the examination of the loadings. We show that this criterium provides results similar to those presented in Section 4.3.2 of Lavezzi and Marsili (2010).

In the next section we present the results from the application of PCA to our data. Given the illustrative role of these notes, we limit our analysis to the set of variables representing the main focus of Lavezzi and Marsili (2010), i. e. those in Ex 1.

2 Application of PCA

We perform PCA with the positive and negative values of the features, to be consistent with Section 4.3.2 of Lavezzi and Marsili (2010).[1] We retain the PCs associated to an eigenvalue, λ_i, larger than the average value of the eigenvalues, $\bar{\lambda}$.[2] We will show that this choice allows to take into account approximately $80\% - 85\%$ of total variance, and is therefore consisent with the criterium based on the (relevant) amount of variance that should be explained by the PCs.

2.1 All countries

Applying PCA to the whole sample we find seven PCs (ordered as PC1, PC2, ..., on the basis of their variance), that explain approximately 88% of total variance. Table 1 contains the results.[3]

[1] We have, therefore, $N = 61$ objects of dimension $D = 26$.

[2] See Jolliffe (2002), p. 115.

[3] We report only the first D elements of the eigenvectors associated to each PCs. The remaining D elements are simply the negative values of those reported. In fact, in our case, if we consider the features and their negative values as our set of $D \times 2$ variables, computation of the PCs requires to solve the following problem:

$$\lambda \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{\Omega} & -\mathbf{\Omega} \\ -\mathbf{\Omega} & \mathbf{\Omega} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \tag{1}$$

where $\mathbf{\Omega}$ is the $(D \times D)$ variance-covariance matrix of the features, $-\mathbf{\Omega}$ is the $(D \times D)$ variance-covariance matrix of the features and their negative values, and the vectors of loadings of the PCs are the $(2D \times 1)$ eigenvectors, represented in Equation (1) as vectors with two components of dimension D, i. e. $[\mathbf{x}\ \mathbf{y}]^T$. It is easy to show that $\mathbf{y} = -\mathbf{x}$.

Feature	PC1	PC2	PC3	PC4	PC5	PC6	PC7
FreeS	0.27	0.04	-0.01	-0.12	-0.01	-0.15	0.25
PE60	0.27	0.15	0.03	-0.04	-0.17	0.08	-0.12
SE60	0.26	0.19	-0.01	-0.06	-0.12	0.06	0.1
-GGap60	0.24	0.15	-0.09	-0.22	0.01	-0.16	0.03
NoElMa	0.21	-0.28	0.01	-0.02	0.15	0.09	0.18
Struct	0.2	0.09	0.17	0.06	-0.05	0.43	-0.24
ElMach	0.2	-0.27	0.08	0.11	0.19	-0.21	0.04
Transp	0.19	-0.24	-0.02	-0.1	0.1	-0.16	-0.51
-LF6085	0.19	0.07	-0.31	0.04	0.28	0.18	0.11
G6085	0.14	-0.18	0.13	0.47	-0.12	0.09	0.13
-GovC	0.1	0.28	0.14	0.3	-0.09	-0.37	-0.07
Trade	0.05	-0.29	-0.1	-0.16	-0.52	-0.02	0.08
-MinDep	0	0.01	-0.56	0.26	-0.11	-0.03	-0.13
$\lambda_i/\bar{\lambda}$	9.86	4.05	2.34	2.2	1.69	1.64	1.01
Prop. var	0.38	0.16	0.09	0.08	0.07	0.06	0.04
Cum. var.	0.38	0.54	0.63	0.71	0.78	0.84	0.88

Table 1: PCA, Ex 1, whole sample

In Table 1 we ordered the features on the basis of their loadings in PC1. We notice that: the features having the four largest loadings in PC1 which, alone, explains about 38% of the variance, are those found in Cluster All.I;[4] three of them (PE60, SE60 and -GGap60) have also similar loadings in PC2;[5] the subsequent four elements of PC1 are the four investment components, three of which are in Cluster All.2. Trade and GovC, and MinDep, have very small loadings in PC1 but have relatively high loadings, respectively, in PC2 and PC3.

The interpretation of this result is that a high portion of variability in the data is first of all provided by human capital, institutions and initial conditions. After accounting for this, investment becomes important. In a further step, a contrast between trade openness and government consumption emerges, indicating that: "after [the variation captured by the first two PCs] has been accounted for, the main source of variation is between [countries] with large [GovC] measurements relative to [Trade], and [countries] with the converse relationship." (Jolliffe (2002), p. 97). The variation based on natural resource abundance becomes important only in PC3. Hence, some correspondence with the clusters of features presented in Table 9 of Lavezzi and Marsili (2010) emerges from this analysis.

[4]In Table 9 of Lavezzi and Marsili (2010) we reported the cluster including GGap60, -FreeS, -PE60 and -SE60. See Footonote 46 of Lavezzi and Marsili (2010) for an explanation.

[5]In Desdoigts (1999), p. 312, the first three elements of the first projection vector which identifies the separation of OECD from non-OECD countries, are: SE60, GGap60 and PE60.

A visual representation provides a convenient way to grasp further insights from Table 1. Figure 1 contains a *biplot* which displays each feature in Table 1 through vectors (arrows), whose coordinates are the loadings of the first two principal components, PC1 and PC2. Also, it displays the observations relative to each country in two dimensions, where the first (second) coordinate is a linear combination of the original data with weights given by PC1 (PC2).[6]

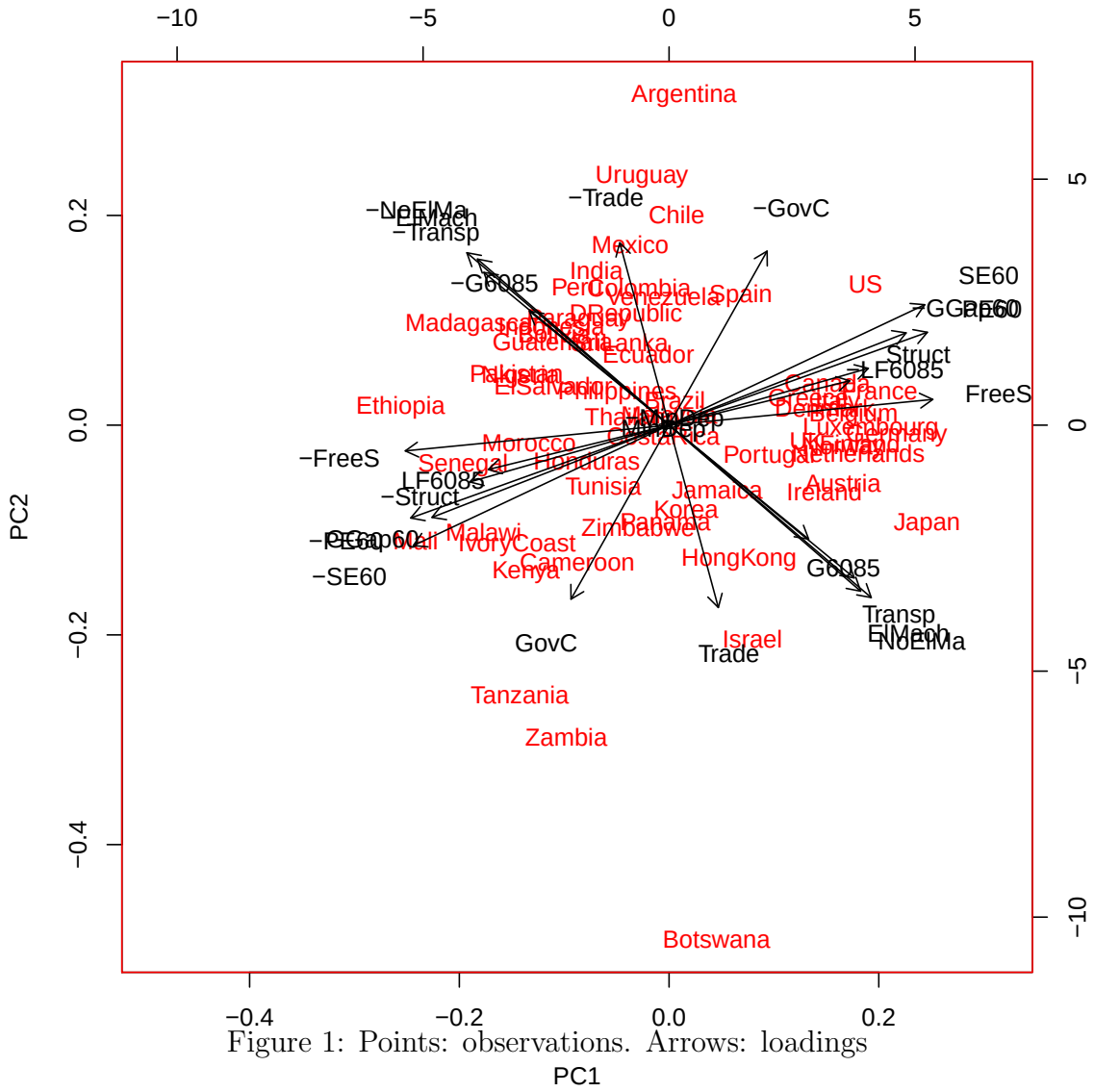

Figure 1: Points: observations. Arrows: loadings

Figure 1 shows that, clearly, PC1 explains a large part of the variance in the data, while PC2 a lower part, which would be much smaller if Argentina and Botswana were excluded (the latter is an outlier in Lavezzi and Marsili (2010)). The clustering of the countries is far

[6] Jolliffe (2002), pp. 90-103, provides a discussion of biplots. In the present analysis we consider the case of $\alpha = 1$, which ensures that the biplot contains the PC scores and the PC loadings.

from clear although, moving along PC1 from the left, we find countries approximately ordered according by their development level, from Sub-Saharan to OECD countries.

Observing the arrows, a pattern for the clustering of features seems to emerge. In particular, some salient aspects of the cluster structure in Table 9 of Lavezzi and Marsili (2010) appear in Figure 1 through the similar position of the vectors, determined by the loadings in PC1 and PC2. In particular: FreeS, SE60, PE60 and -GGap60, which form Cluster All.I in Lavezzi and Marsili (2010) have similar values in the first two PCs, and therefore appear very close in Figure 1.[7] Also, G6085, Transp, ElMach and NoElMach, which form Cluster All.II, have very similar loadings. The vectors representing other features, which do not appear in any significant cluster in Table 9 of Lavezzi and Marsili (2010), appear quite dispersed in Figure 1.[8]

In the next sections, we perform PCA on the four clusters identified in Ex1, following the same order of Lavezzi and Marsili (2010), that is: 4 - 2 - 3 - 1.

2.2 Cluster 4

Results for countries in Cluster 4, mostly Sub-Saharan countries, are very different from those obtained for the whole sample. Table 2 and Figure 2 contain the results.

[7] In Desdoigts (1999), p. 313 SE60, PE60 and GGap60 are the most relevant elements along the first direction of projection.

[8] This type of analysis of the features is absent in Desdoigts (1999) although, as remarked, he utilizes a method close in spirit to PCA.

Feature	PC1	PC2	PC3	PC4
-G6085	0.4	0.15	-0.22	0.3
MinDep	0.4	-0.28	-0.05	-0.37
-Transp	0.28	0.32	0.02	0.07
-PE60	0.21	0.04	0.22	-0.01
GovC	0.19	-0.33	0.01	0.22
-Struct	0.09	0.09	0.22	-0.05
-SE60	0.08	-0.13	0.19	-0.14
-ElMach	0.04	0.18	-0.02	-0.16
-Trade	0.04	0.22	0.3	0.11
-FreeS	0.04	-0.02	0.29	-0.04
-NoElMa	0.02	0.26	-0.26	-0.38
-LF6085	0.02	0.1	0.28	-0.13
GGap60	0	-0.02	0.07	0.02
$\lambda_i/\bar{\lambda}$	4.53	3.79	1.57	1.19
Prop. var	0.35	0.29	0.12	0.09
Cum. var.	0.35	0.65	0.77	0.86

Table 2: PCA, Ex 1, Cluster 4

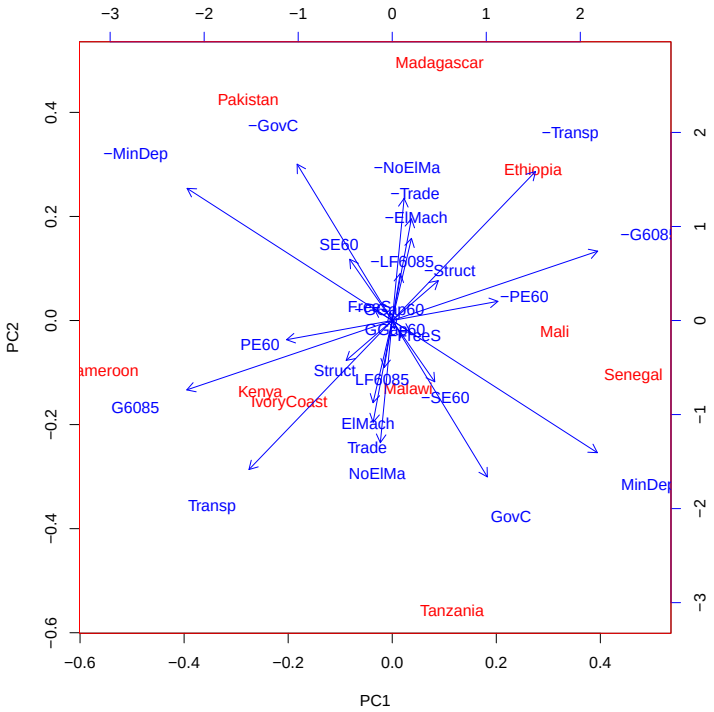

Figure 2: Points: obs. Arrows: loadings

We find four PCs, explaining about 86% of total variance. Although patterns are somewhat less visible than those for all countries in Section 2.1, the correspondence with the cluster structure of the features in Lavezzi and Marsili (2010) is found in the following aspects: the proximity of the vectors representing -FreeS and GGap60 (although they are not clearly visible in Figure 2), i. e. the most significant members of Cluster 4.I; the closeness of ElMach and NoElMa (members of Cluster 4.II) which, however, in the present setting also appear related to LF6085 and Trade; the similarity of the loadings of Transp, Trade and G6085, i. e. three out of four members of Cluster 4.III.[9]

2.3 Cluster 2

Table 3 and Figure 3 present the results for Cluster 2, including Latin American and African countries, plus the Philippines.

[9]Along PC1, it is possible to identify a partition of countries in Clusters 5 and 3 of Lavezzi and Marsili (2010), with the former below zero, and the latter above.

Transp	0.46	-0.13	0.1	0.14
GovC	0.31	-0.01	-0.01	-0.46
NoElMa	0.2	-0.11	-0.03	0.4
ElMach	0.17	-0.01	0.03	0.19
MinDep	0.17	-0.01	-0.45	-0.08
-SE60	0.15	0.35	0.22	0.02
Trade	0.15	-0.12	0	-0.19
-G6085	0.14	0.05	0.3	-0.05
Struct	0.11	-0.08	-0.12	0.01
GGap60	0.08	0.17	-0.09	0.02
-FreeS	0.08	0.37	-0.23	0.15
-PE60	0.03	0.38	0.11	-0.07
-LF6085	0.03	0.12	-0.24	-0.03
$\lambda_i/\bar{\lambda}$	5.45	2.55	1.56	1.05
Prop. var	0.43	0.2	0.12	0.08
Cum. var.	0.43	0.63	0.75	0.83

Table 3: PCA, Ex 1, Cluster 2

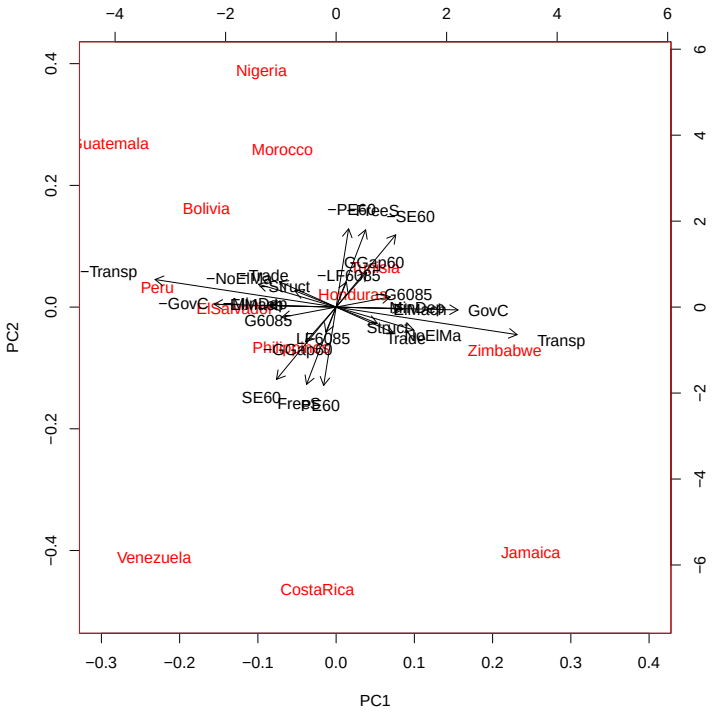

Figure 3: Points: observations. Arrows: loadings

Applying PCA to Cluster 2 we find four PCs, accounting for 83% of total variance. Notice first of all that PC1 explains much more variance than PC2. In Figure 3 the features form two groups, according to their loadings in PC1 and PC2. In particular, the first group includes most variables with the largest loadings in PC1, while the second mostly includes those with the largest loadings in PC2. The first one includes features belonging to Clusters 2.I, 2.IV and 2.V (plus G6085) (see Table 9 of Lavezzi and Marsili (2010)), while the second one contains GGap60, -FreeS, -LF6085, -SE60, and -PE60 belonging, respectively, to Clusters 2.II and 2.III.[10]

2.4 Cluster 3

Table 4 and Figure 4 contain the results for Cluster 3, which includes Latin American and Asian countries.

[10]No particular evidence appears with respect to the membership of the countries to Clusters 4, 6 and 10 in Table 3 of Lavezzi and Marsili (2010), which formed Cluster 2 in the second step of the application of the GM algorithm.

Feature	PC1	PC2	PC3	PC4
G6085	0.38	0.05	-0.15	-0.03
-Struct	0.37	-0.39	0.14	0.02
Trade	0.25	0.43	0.41	0
GGap60	0.2	-0.12	0.01	-0.24
LF6085	0.2	0.09	0.04	0.06
Transp	0.17	0.17	-0.4	0.12
NoElMa	0.16	0.12	-0.23	-0.18
-MinDep	0.09	-0.13	-0.05	0.42
ElMach	0.08	-0.07	-0.17	-0.05
-SE60	0.07	-0.07	0.18	-0.05
-GovC	0.04	-0.12	0.12	0.35
-FreeS	0.03	0.15	-0.06	0.26
-PE60	0.03	-0.17	0.03	-0.15
$\lambda_i/\bar{\lambda}$	4.4	2.86	1.57	1.13
Prop. var	0.35	0.23	0.12	0.09
Cum. var.	0.35	0.57	0.7	0.79

Table 4: PCA, Ex 1, Cluster 3

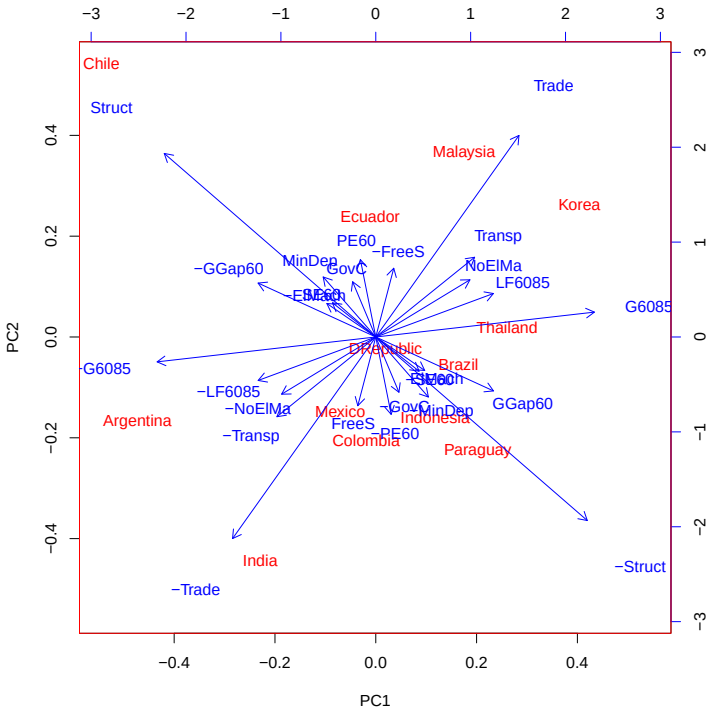

Figure 4: Points: observations. Arrows: loadings

In Cluster 3 we find five PCs, accounting for 79% of total variance. Both PC1 and PC2 explain a relatively large portion of variance. The cluster structure of Table 9 of Lavezzi and Marsili (2010) can be retraced in the relatively similar position of the vectors representing G6085, LF6085, NoElMa and Trasp (elements of Cluster 3.I), and of GGap60 and -Struct (the two most significant members of Cluster 3.II).[11]

2.5 Cluster 1

In this section we perform PCA on Cluster 1, i. e. the cluster of OECD countries. Table 5 and Figure 5 contain the results.

[11]Reading along the PC1, it is possible to partially retrace the partition of the countries in the clusters that formed Cluster 3, i. e. Clusters 7, 9, 11 and 12 of Table 3 in Lavezzi and Marsili (2010).

Feature	PC1	PC2	PC3	PC4
-GGap60	0.45	0.25	-0.08	-0.06
-G6085	0.33	0.02	0.04	-0.08
-Struct	0.23	-0.04	0.19	0.14
Trade	0.22	-0.51	-0.35	-0.07
LF6085	0.14	0.21	0.02	-0.2
-NoElMa	0.14	-0.02	0.28	0.41
FreeS	0.13	0.05	-0.17	0
GovC	0.12	-0.04	0.11	-0.03
-ElMach	0.1	-0.32	0.39	-0.15
-MinDep	0.05	-0.08	-0.18	0.42
SE60	0.04	0.11	-0.11	-0.06
Transp	0.02	0.11	-0.13	0.22
PE60	0.01	0	0.02	0.1
$\lambda_i/\bar{\lambda}$	3.36	2.71	1.67	1.18
Prop. var	0.29	0.23	0.14	0.1
Cum. var.	0.29	0.52	0.66	0.76

Table 5: PCA, Ex 1, Cluster 1

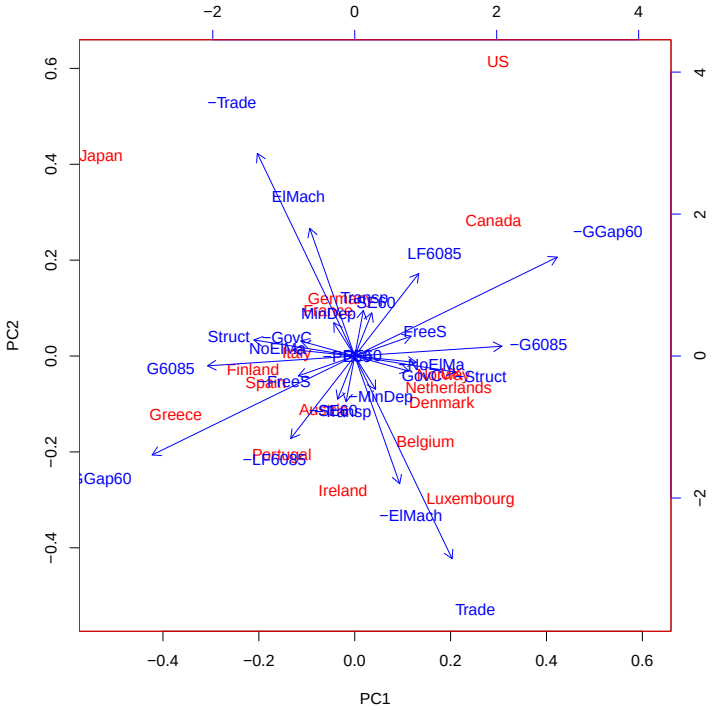

Figure 5: Points: observations. Arrows: loadings

We found 4 PCs, accounting for 76% of total variance. The strength of the relationship between the initial productivity gap and the growth rate, evidence of convergence and already emphasized in Lavezzi and Marsili (2010), is retraceable in the preminent position that these two features have in PC1 (also visible in Figure 5).[12]

2.6 Summing up

Table 2.6 summarizes the results on the number of PCs across the different samples considered.

	All countries	Cl 4	Cl 2	Cl 3	Cl 1
# PCs	7	4	4	4	4
% var.	88%	86%	83%	79%	76%
av. var	12.57	21.5	20.75	19.75	19

Table 6: Summary of results on the number of PCs. The first line contains the number of PCs in the samples considered; the second (third) line reports the amount of (average) variance explained by the PCs

[12] Along the PC1, it is possible to partially retrace the partition of the countries in the clusters that formed Cluster 1, i. e. Clusters 1, 2 and 8 of Table 3 in Lavezzi and Marsili (2010).

The number of PCs decreases when we move from the analysis of the whole sample to that of individual clusters, showing that the heterogeneity within subsamples can be accounted for by a lower number of variables than in the whole sample. The average amount of variance explained by the PCs increases from about 12% to approximately 20%. The next section draws the conclusions.

3 Concluding remarks

In these notes we applied Principal Component Analysis (PCA) to the database analyzed in Lavezzi and Marsili (2010) with the GM clustering algorithm. We clarified the differences between the two methods and showed that, nonetheless, there is consistency among the results obtained with them.

We showed that the number and the composition of the PCs is different when analyzing the whole sample, with respect to the analysis of the four clusters of countries identified in Lavezzi and Marsili (2010): the number of PCs decreases, and the role played by the different features is different across clusters and between the clusters and the whole sample. This is consistent in particular with the aspect of *clustering the features* of Lavezzi and Marsili (2010): part of the structure of the clusters of features of Lavezzi and Marsili (2010) can be retraced in the groups of features that, applying PCA, display similar loadings in in the first PCs.[13]

These results cofirm the view that the growth process is characterized by qualitatively different regimes, an aspect which can be overlooked when cross-sectional studies are carried out on large samples of economies, without paying attention to the differences in their development stages.

References

Desdoigts, A. (1999), "Patterns of Economic Development and Formation of Clubs", *Journal of Economic Growth*, 4, 305-330.

Jolliffe, I. T. (2002), *Principal Component Analysis*, Springer.

Lavezzi, A. M. and M. Marsili (2010), "An Empirical Analysis of Growth Regimes", mimeo.

[13]We found much less correspondence between the classification of the countries along the first PC identified, and the cluster structure of countries obtained by the application of the GM algorithm.