Growth and Gender Equality: a multivariate approach

Ferrando Alberto - 150022854

April 13, 2016

Abstract

The following paper describes a study conducted with exploratory multivariate techniques on a Gender Equality and Economic dataset. Its main aim is to explore potential relationships among the variables and extrapolate useful research questions. From such analysis, it emerges that the main variables which really differentiates countries are not average income levels or GDP but gender equality in secondary education and the level of infant mortality. This interesting view could be subsequently researched using inferential methods. Secondly, a potential research hypothesis is discussed that identifies those previous variables as drivers (and not consequences) of good economic performance. Finally, some interesting relationships among variables are discussed.

1 Introduction and modelling framework

The common macroeconomic view usually differentiate among countries on the basis of GDP or other income level indexes. However, such distinction on the pure basis of economic performance may actually hide a subtler (and more intriguing) story. In this paper we try to lay the foundations for a study that could potentially change this common economic distinction by trying to find possible alternative variables that better reflect the development levels of countries. Other authors have already approached this aim, see for example (Daly & Cobb, 1989) where the Index of Sustainable Economic Welfare (ISEW) is developed. A nice summary of the debate on the issue is provided by (Schepelmann, Goossens, & Makipaa, 2009). In it the authors suggest to go beyond GDP because it "does not properly account for social and environmental costs and benefits".

Nevertheless, their researches and views are still linked on constructing a new measure from scratch. In this paper on the contrary, a more empirical view is adopted: real data is used in order to determine what measure(s) really differentiate countries among themselves; in other words: what is the real underlying characteristic(s) that can enable the researchers to really categorize a country?

We are going to develop some hypothesized answers to that (and more) question(s) using exploratory multivariate techniques. In particular *Principal Component Analysis* (PCA) and *K-Means Clustering*.

1.1 The Data

The dataset used has dimensions 144x34 and was generated by the World Bank DataBank. It comprises several (all numeric) indicators measuring different aspects of an economy *e.g.* GDP, dependency ratio, percentage of MPs who are women, etc. All the measurement refer to the year 2013. In other words, this dataset is not a time series but a single snapshot in time. Similarly, many countries were included: for a complete list of variables and observations, see the appendix. The original dataset included a significant number of missing values: thanks to Kim-Yen Nguyen, the missing values have been filled-in with the value for the next available year. Nevertheless, even with this now complete dataset, some variables still presented missing values. To overcome this problem, and avoid loosing important observations, the dataset was treated through *Fully Conditional Specification* (FCS) (Van Buuren, 2007). Such method is based on the idea of constructing an imputation model for every Y_j (*i.e.* every variable) by specifying a conditional density: $P(Y_j|X, Y_{-j}, R, \theta_j)$ where:

- X is a complete set of covariates. Obliviously, to initialize the algorithm, the missing X are firstly simply guessed.
- Y_{-j} is the set of variables in Y expect Y_j.
- R the set of binary response indicators for each variable k: $R = (R_1, ..., R_k)$. If $R_j = 1$, Y_j is missing.
- θ_j is a set of parameters for Y_j.

FCS is done by iterating over all conditionally specified imputation models, which can differ for every Y_j. In such an algorithm, one iteration consist of a passage through all Y. In theory, this process should be repeated m times to decrease uncertainty. However, the author did not find any method to merge together the results of PCA/K-means in the literature: hence for the purpose of this study, $m = 1$. However, given that the missing data was only 4.39%, the gain from the saved observations clearly surpasses the harm of (possible) increased uncertainty. Moreover, we have to stress that a variable (GDP) has been transformed by taking the log_{10} in order to reduce its range in line with the other variables.

1.2 Multivariate techniques

To explore the dataset and generate hypothesis we used two main multivariate techniques: PCA and K-means clustering. PCA (Wold, Esbensen, & Geladi, 1987) extracts the most

important information by an orthogonal transformation of the data matrix. In mathematical terms, the eigenvectors α_k and eigenvalues $lambda_i$ of var-covariance matrix Σ are estimated. They represent, respectively, the variances of the projections of the data on the new axes and the contributions of each variable on them. Such technique allows the user to identify the dominant trend in a data matrix. However, we have to bear in mind that the results depend significantly on the scaling used. In this study, both the covariance and correlation matrix have been used.

K-means clustering on the contrary, allows the user to identify k groupings of data, which are similar in some respect. Of course, the definition of similarity depends strictly on the measure used. In this case, Euclidean distance is the choice given that all the variables are continuous; Nevertheless, it is very sensitive to outliers. A potential alternative would have been *hierarchical clustering*: however, given that the dataset does not present any implicit hierarchical structure this method was not considered. For an in depth review on the topic see (Hartigan & Wong, 1979).[1]

1.2.1 Assumptions

Multivariate methods used in this study are based on three basic assumptions:

- Multivariate normality
- Absence of Outliers
- Linearity

Moreover, they are also quite sensible to deviations from symmetry. Before conducting the analysis, some checks are done: the dataset used, unfortunately, presents significant violations of those assumptions. For example, numerous outliers were identified when plotting the robust Mahalanobis distances against χ^2 quantiles; Checking linearity is difficult due to de large number of variables which makes the use of a scatterplot matrix impossible. Nevertheless, by controlling a sample of variables, non-linearities were obvious: see for example Figure 1, left plot. Also multivariate Normality is violated, as is depicted by the QQ-plot in Figure 1, right plot. In addition, also symmetry is violated: many different variables had an highly skewed distribution. Plots are omitted for sake of simplicity.

[1] In truth, PCA and k-means are just special cases of one another: PCs are the continuous generalization of the discrete cluster indicators in k-means. For a proof of this: (Ding & He, 2004).

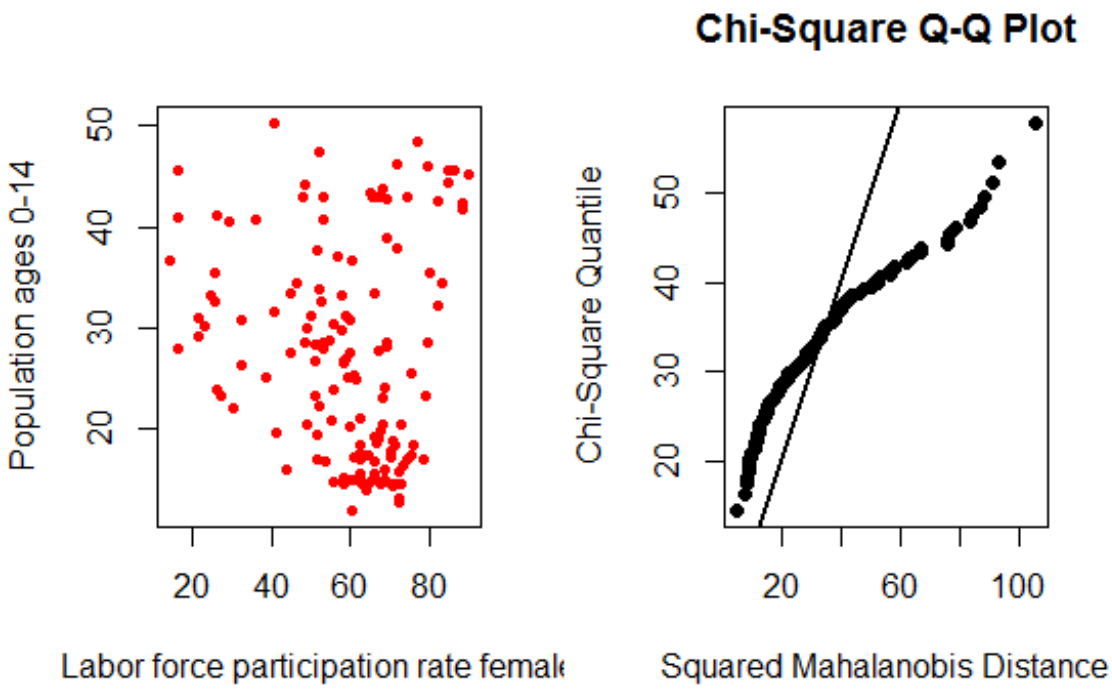

Figure 1: (Left) Scatterplot of Labor force participation rate (female) and Population ages 0-14: no linear relationship is visible. (Right) QQ-plots of empirical quantiles and inverse ^{2} distribution.

Therefore, the reader has to bear in mind such violations when interpreting the results, which can obliviously be significantly distorted. Nevertheless, the aim of this study is not conducting inference but generating hypothesis: hence, no much harm could be done by such violations.

2 Results of the analysis

By applying PCA we are able to extract the components (*i.e.* a new set of axis) that account for as much variability as possible. In addition, each variable has a value for this axis (called *loading*): thanks to this we can identify which variable has most influence on the position of the component.

The PCA executed on the covariance matrix returns interesting and clear results. First of all, the first three components are selected through the aid of a screeplot (Figure 2) and their loadings inspected. Kaiser's criterion would suggest to choose 9: however, such number of components would pose a significant obstacle to interpretation. It is therefore not followed.

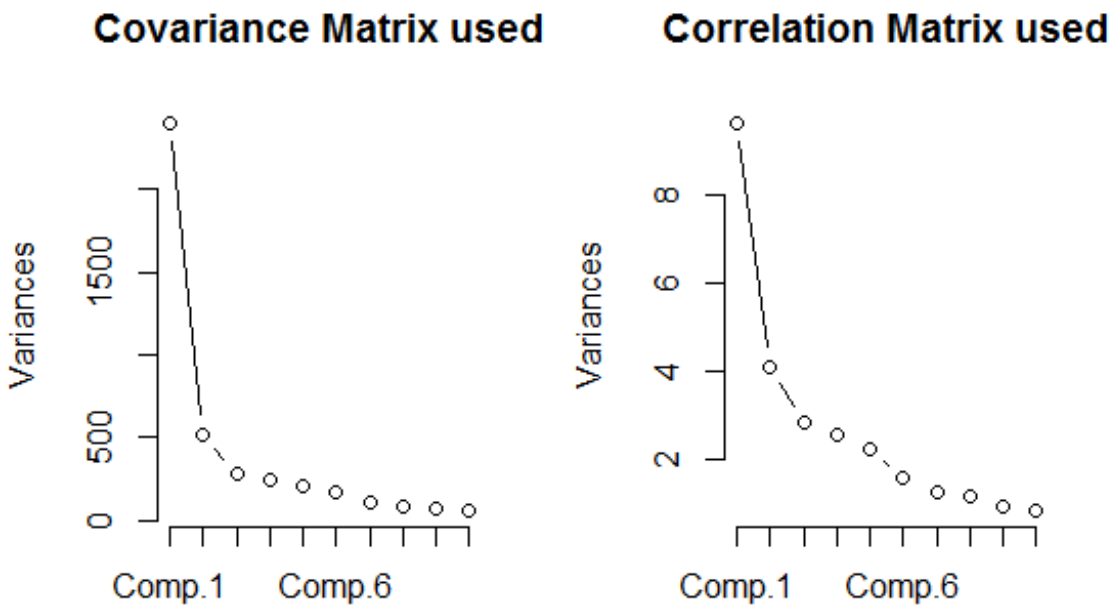

Figure 2: Screeplots showing the variance of PCs depending if the: (Left) Covariance Matrix or (Right) Correlation Matrix was used.

It turns out that, for the first PC, the most important variable are the secondary education enrolment ratio for females, the same measure for males and the level of infant mortality. In addition, infant mortality has opposite sign than the other two. We could therefore speculate a sort of *bipolar component*: countries that have high infant mortality present low secondary education enrolment for both sexes. Other variables that seem to have an impact (even if less marked) are: dependency ratio, proportion of population aged 0-14 and life expectancy at birth (female/male). On the second component, the importance shifts even more on female-specific measures. In particular: female labour force participation rate, firms with active female ownership and female primary education enrolment. The third component presents a quite similar picture: see 1 in the appendix for a review of all loadings. The interesting detail is that purely economic indicators do not appear in the components until PC9 with GDP growth and even then the loading is quite small in absolute value.

From such figures a clear conclusion emerges: the real difference between countries seem to lie in how they compare in terms of gender equality, education and infant mortality[2]

[2]not a proxy for healthcare spending, since public health expenditure did not appear important on the first components.

and not in respect of income or industrial production (as mainstream economic research suggest).

Additionally, a possible aid to interpretation is adding the names of countries to the PCs scatterplot. The result can be seen in Figure 3 where PC1 is plotted against PC2 and PC3[3].

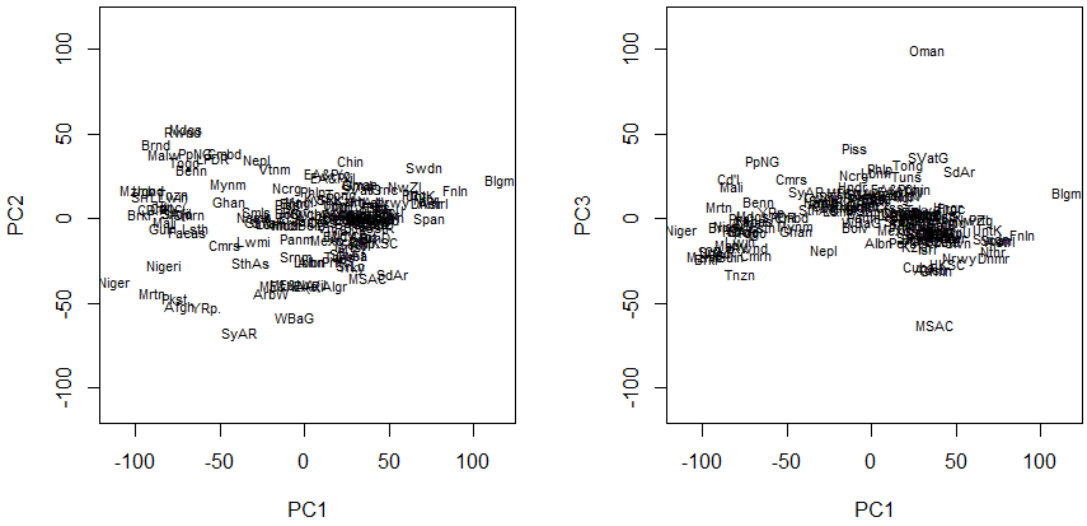

Figure 3: Scatterplots of PC1 vs PC2 and PC3

From Figure 3 we can distinguish a clear pattern: first of all, PC1 is the main driver of diversity. Secondly, European countries seem to be mainly positioned at high levels of PC1. In fact, those nations present low levels of infant mortality and high levels of secondary education for both sexes. Conversely, the left corners mainly present the group of African countries, which perform poorly on those levels. From such findings, another possible hypothesis would be that economic growth is in truth conditional on equality, education and good healthcare and not vice-versa.

Nonetheless, if we plot the first two principal components and we assign to the bubbles the GDP growth of each country, a contrasting pattern appears as depicted in Figure 4. It is possible to see that countries that score lower on PC1 (hence have high infant mortality and low gender equality) have actually higher economic performances. Hence, the hint we get from this plot is actually the opposite than before. However, to confirm this aspect, a time-series analysis would be required: maybe their gender equalities and infant mortality are decreasing over time, boosting economic growth (a possibility that is in line with our hypothesis).

[3]PC2 vs. PC3 is not shown because it did not present any interesting pattern

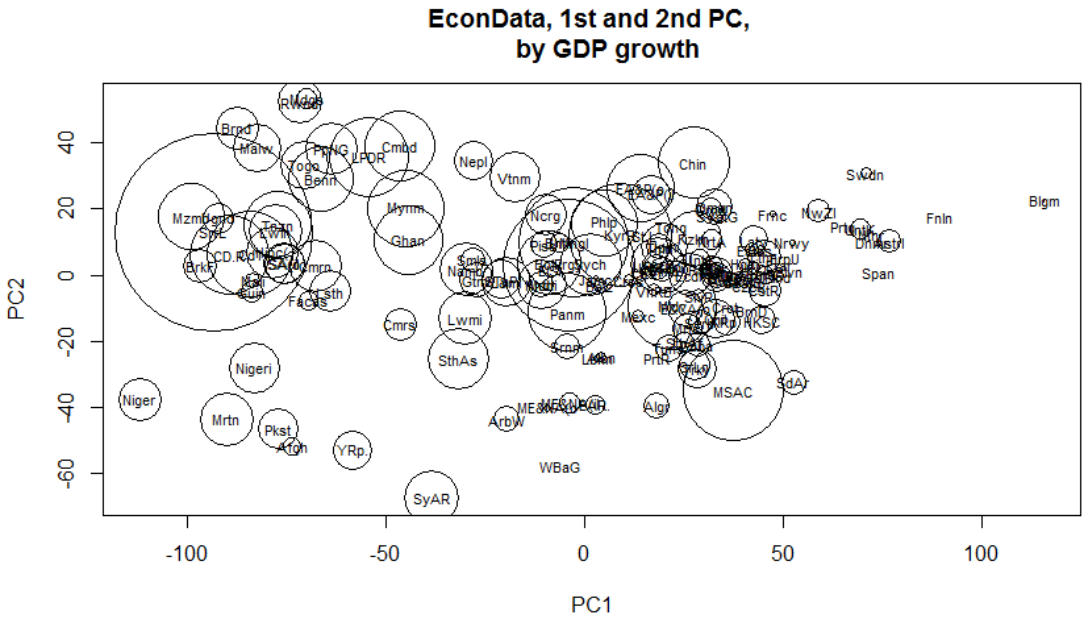

Figure 4: Bubbleplot of PC1 vs PC2. The area of the bubble is proportional to a country GDP growth

PCA can also be run on the correlation matrix. In this case the loadings structure is less clear and for this reason the covariance method is preferred. Nevertheless, a similar pattern of important variables could be identified: the most important loadings still identify gender and demographic indicators first, rather than purely economic indices.

The results of k-means clustering on the dataset actually supports our previous view. Indeed, if $k = 2$, the clusters are easily spotted for component 1 against 2 and 3. As before, we identify a red group (which is mainly composed by poor countries, mostly African) and a blue group (developped countries) separated by different scores on PC1. On the contrary, the two clusters are not distinguishable with respect to the scores of PC2 against PC3. The scatterplot matrix in Figure 5 illustrates the point[4].

Finally, visualizing variables relationships with the added information from clustering can reveal interesting (and puzzling) patterns. For example, the most counter-intuitive of them is the female enrolment in secondary education plotted against proportion of female MPs (Member of Parliament) in Figure 6: confusingly, no relationship between the two is present. It seems therefore that high levels of education and political representation are not related for women. Another possible hypothesis then emerges: could it be possible that increasing women education does not automatically increase their political and civil powers?

[4]In this scatterplot, the signs of components are reversed with respect to previous plots. This is due to the R function used in analysis

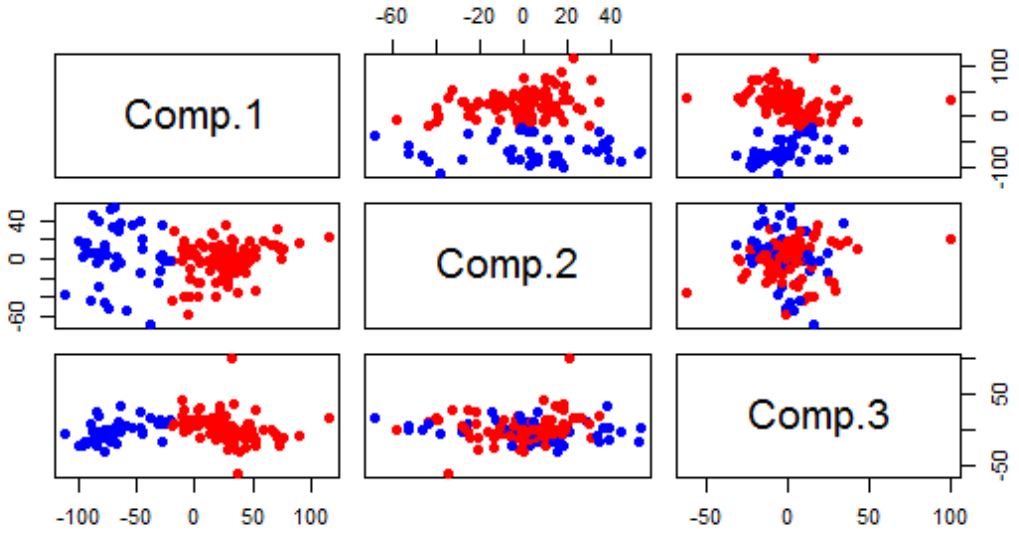

Figure 5: Scatterplot matrix of PCs. Two groups are identified.

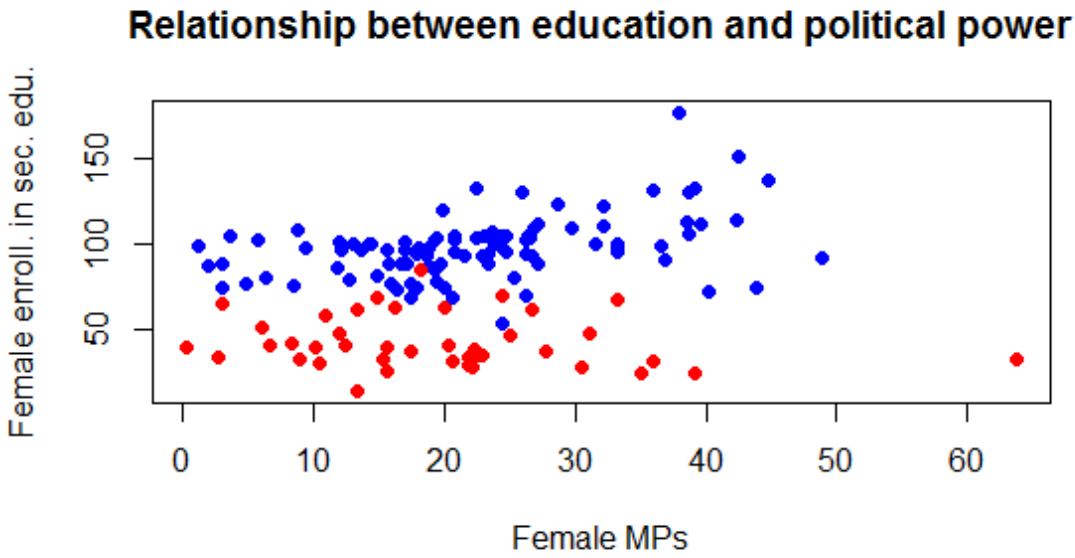

Figure 6: Political representation and high education do not seem to be related

3 Conclusion

In this study, two multivariate methods were applied, namely PCA and K-means clustering on an Economic dataset. Unfortunately, all their assumptions were greatly violated. Nevertheless, given the exploratory nature of the analysis, the results were still interpretable. First of all, GDP did not seem to be a useful variable to differentiate between countries: on the contrary, gender equality in secondary education and infant mortality were the ones who varied the most, and hence the most important to categorize nations. For this reason, the mainstream economic view was challenged given that it poses disproportionate weight on income levels. Secondly, a causal relationship between those measures and economic growth was hypothesised, even if two different visualisation perspectives provided opposite evidences. Thirdly, clustering methods were applied which confirmed our interpretation of the PCs by identifying two main groups of countries, depending on their score with respect to gender equality and infant mortality. Finally, the clustering results were also used to explore the relationship between original variables from the dataset. One plot in particular revealed that no link between female secondary education and women political power exists. The author underlines again that this study as exploratory aims only, hence it does not constitute enough evidence to support the generated hypothesis. For such goal, deeper inferential methods would be required.

References

Daly, H. E., & Cobb, J. (1989). For the common good beacon press. *Boston*.

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on machine learning* (p. 29).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Schepelmann, P., Goossens, Y., & Makipaa, A. (2009). *Towards sustainable development: Alternatives to gdp for measuring progress* (No. 42). Wuppertal Spezial, Wuppertal Institut für Klima, Umwelt und Energie.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, *16*(3), 219–242.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52.

Appendices

A PCs Scores

Variable	PC1	PC2	PC3
Age dependency	-0.290		-0.111
Labor force part. female		0.650	-0.472
Labor force female		0.299	-0.258
Life expectancy female	0.158		
Life expectancy male	0.133		
Population ages 65+			-0.100
Population ages 0-14	-0.194		
Unemployment female		-0.164	0.104
Mortality rate infant	-0.411		
Enrolment primary edu female		0.317	
Enrolment primary edu male		0.249	0.108
Enrolment secon edu female	0.602		
Enrolment secon edu male	0.527		
Women MPs			-0.269
Firms with female top mgmt		0.237	0.396
Firms with female ownership		0.452	0.630

Table 1: PCs loadings for PCA on the covariance matrix. The empty slots refer to values close to 0

B Full list of variables and Countries

- Age Dependency Ratio (%): ratio of dependent population (age ¡15 or ¿64) to working-age population (age 15-64).
- GDP Growth (%): Gross Domestic Product growth is the annual percentage variation in all of an economys income (wages, interest, profits, and rents or expenditures) consumption, investment, government purchases and net exports.
- GDP PPP (current international $): GDP in terms of purchase power parity (PPP): number of units of a countrys currency required to buy the same amount of goods and services in the domestic market as U.S dollar would buy in the United States.
- Public Health Expenditure (%): proportion in terms of GDP
- Inflation Consumer Price (%): annual percentage change in the cost of acquiring a determined basket of goods and services.

- Labour Force Participation Rate, Female or Male (%): proportion of female or male population aged 15-64 that is economically active. This includes people who are currently working and who are actively seeking employment.
- Labour Force, Female (%): female labour force as a percentage of the total labour force.
- Labour Force, Male (%): male labour force as a percentage of the total labour force.
- Life Expectancy at Birth, female or male (years): the lifespan estimated for females or males based on the current mortality rates at the time of birth.
- Lifetime Risk of Maternal Death (%): probability that a fifteen-year-old female will die eventually from a maternal cause assuming that current levels of fertility and mortality do not change in the future, taking into account competing causes of death.
- Population Aged 65 and above (%): in terms of total population
- Female Population (%): in percent of total population
- Employers, female (%): the female employer percentage of employment that hold the type of jobs defined as a self-employment jobs.
- Employers, total (%): the total employer percentage of employment that hold the type of jobs defined as a self-employment jobs.
- Gross capital formation (%): gross domestic investment as a percentage of GDP.
- Unemployment, total (%): share of the labor force that is without work but available for and seeking employment.
- Unemployment, female (%): percentage of female unemployment in female labor force.
- Unemployment, male (%): percentage of male unemployment in male labor force.
- Mortality rate, infant (per 1,000 live births): number of infants dying before reaching one year of age, per 1000 live births in given year.
- Gross enrolment ratio, (primary or secondary), female (%): female gross enrollment in (primary or secondary) education, regardless of age, divided by the female population of official (primary or secondary) education age. This can exceed 100
- Gross enrolment ratio, (primary or secondary), male (%): same as above for males instead of females

- Gross enrolment ratio, (primary, secondary, or tertiary), gender parity index (GPI): the ratio of female gross enrolment to male gross enrolment in (primary, secondary, or tertiary) education. If GPI is 1, it indicates parity between females and males. GPI less than 1 indicates fewer females than males enrolled.

- Proportion of seats held by women in national parliaments (%): percentage of parliamentary seats in single or lower chamber held by females. This variable is used to measure the position of female in society.

- Firms with female top manager (% of firms): the percentage of firms in the private sector which have females as the highest-ranking manager (e.g. CEO). This variable also includes owners who work as the manager of the firm.

- Firms with female participation in ownership (% of firms): the percentage of firms with female among the principal owners.

Countries: Afghanistan, Albania, Algeria, Arab World, Australia, Austria, Belarus, Belgium, Belize, Benin, Bhutan, Bolivia, Brunei, Bulgaria, Burkina Faso, Burundi, Cabo Verde, Cambodia, Cameroon, Caribbean small states, Central Europe and the Baltics, Chile, China, Colombia, Comoros, Congo Dem. Rep., Costa Rica, Cote d'Ivoire, Croatia, Cuba, Czech Republic, Denmark, East Asia & Pacific (all income levels), East Asia & Pacific (developing only), Ecuador, Egypt Arab Rep., El Salvador, Estonia, Euro area, Europe & Central Asia (all income levels), Europe & Central Asia (developing only), European Union, Finland, Fragile and conflict affected situations, France, Georgia, Germany, Ghana, Grenada, Guatemala, Guinea, Heavily indebted poor countries (HIPC), High income, High income: nonOECD, High income: OECD, Honduras, Hong Kong SAR China, Hungary, Israel, Jamaica, Kazakhstan, Korea Rep., Kyrgyz Republic, Lao PDR, Latin America & Caribbean (all income levels), Latin America & Caribbean (developing only), Latvia, Lebanon, Lesotho, Lithuania, Low & middle income, Low income, Lower middle income, Macao SAR China, Madagascar, Malawi, Mali, Mauritania, Mauritius, Mexico, Middle East & North Africa (all income levels), Middle East & North Africa (developing only), Middle income, Moldova, Mongolia, Mozambique, Myanmar, Namibia, Nepal, Netherlands, New Caledonia, New Zealand, Nicaragua, Niger, Nigeria, North America, Norway, OECD members, Oman, Pacific island small states, Pakistan, Palau, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Puerto Rico, Russian Federation, Rwanda, Sao Tome and Principe, Saudi Arabia, Serbia, Seychelles, Sierra Leone, Slovak Republic, Slovenia, Small states, South Africa, South Asia, Spain, Sri Lanka, St. Lucia, St. Vincent and the Grenadines, Sub-Saharan Africa (all income levels), Sub-Saharan Africa (developing only), Suriname, Sweden, Syrian Arab Republic, Tanzania, Thailand, Togo, Tonga, Tunisia, Turkey, Uganda, Ukraine, United Kingdom, United States, Upper middle income, Venezuela RB, Vietnam, West Bank and Gaza, Yemen Rep.