Data Analysis and Statistical Modeling

1st Project

Data analysis through statistical methods

Group 9 :

Pedro Matias 90162

Aurel Vitelariu 96728

André Mendes 98926

Duarte Flôr 98931

Salvador Caliço 98967

Profª Isabel Rodrigues

Instituto Superior Técnico

Universidade de Lisboa

Portugal, 2744-016 Porto Salvo

LETI 2023/24

Index:

Introduction

In this project for the course in Data Analysis and Statistical Modelling, we aim to analyze a data set through the use of common statistical methods. The data set concerns household budget data in Italy in 1973, obtained via survey. The variables, all numerical, are the following:

1. 'wfood': The share of total expenditure spent on food.
2. 'whouse': The share of total expenditure spent on housing and fuels.
3. 'wmisc': The share of total expenditure spent on miscellaneous items.
5. 'phouse': The price of housing and fuels.
6. 'pmisc': The price of miscellaneous items.
7. 'totexp': The total expenditure of the household.
8. 'year': The year of the observation.
9. 'income': The household's income.
10. 'size': The number of people in the household.
11. 'pct': The cell weight, which could be used in survey data to adjust the contribution of the observation to analysis, reflecting the structure of the overall population.

The miscellaneous variables consist of goods and services such as clothing, health care, transportation and communications.

1. Summary Statistics

wfood	whouse	wmisc	phouse	pmisc
Min. :0.9432	Min. :0.5451	Min. : 0.1920	Min. :0.1513	Min. :0.1564
1st Qu.:1.5842	1st Qu.:0.7851	1st Qu.: 0.8569	1st Qu.:0.1554	1st Qu.:0.1674
Median :1.8666	Median :0.9681	Median : 1.2496	Median :0.1564	Median :0.1700
Mean :2.1201	Mean :1.4329	Mean : 1.9856	Mean :0.1567	Mean :0.1707
3rd Qu.:2.2137	3rd Qu.:1.5898	3rd Qu.: 2.1236	3rd Qu.:0.1577	3rd Qu.:0.1735
Max. :6.9402	Max. :5.9897	Max. :11.3116	Max. :0.1620	Max. :0.1887

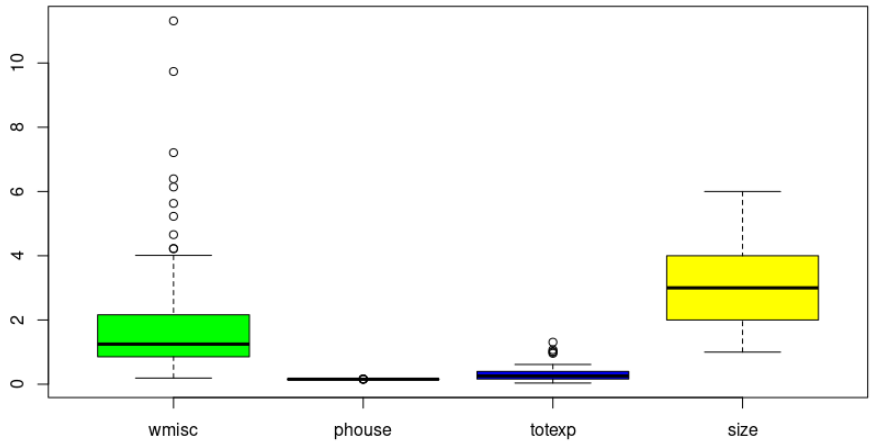
totexp	income	size	pct
Min. :0.03785	Min. : 1.000	Min. :1.000	Min. : 0.300
1st Qu.:0.16209	1st Qu.: 6.000	1st Qu.:2.000	1st Qu.: 2.625
Median :0.26135	Median :10.000	Median :3.000	Median : 5.800
Mean :0.31936	Mean : 9.884	Mean :3.128	Mean : 5.798
3rd Qu.:0.39894	3rd Qu.:14.000	3rd Qu.:4.000	3rd Qu.: 7.750
Max. :1.30951	Max. :18.000	Max. :6.000	Max. :23.400

Looking at the phouse and pmisc values we can see that all the measures of location are very close together. This indicates a very low standard deviation in these variables. On the other hand, there is a big interquartile range in the income variable, along with a large range (difference between maximum and minimum values). This means that both the robust (IQR) and non robust (R) measures of spread are high, which indicates a large standard deviation since a large standard deviation is caused by values more distant from the mean. In the income variable, the proximity of the median to the mean and the fact that both are about in the middle between the 1st and 3rd quartile suggests highly symmetrical data, with very little skewness. Contrast this with wmisc, which had a mean much greater than the median, and close to the 3rd quartile, along with a maximum value that is almost 10 times greater than the median, indicating that the distribution is asymmetrical and right skewed, with some outliers in the higher values.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
wfood	1	86	2.12	0.95	1.87	1.97	0.45	0.94	6.94	6.00	2.33	7.35	0.10
whouse	2	86	1.43	1.10	0.97	1.19	0.35	0.55	5.99	5.44	2.36	5.68	0.12
wmisc	3	86	1.99	1.97	1.25	1.59	0.71	0.19	11.31	11.12	2.52	7.20	0.21
phouse	4	86	0.16	0.00	0.16	0.16	0.00	0.15	0.16	0.01	0.69	1.11	0.00
pmisc	5	86	0.17	0.01	0.17	0.17	0.00	0.16	0.19	0.03	0.51	1.66	0.00
totexp	6	86	0.32	0.24	0.26	0.28	0.18	0.04	1.31	1.27	1.76	3.54	0.03
income	7	86	9.88	5.01	10.00	9.90	5.93	1.00	18.00	17.00	-0.02	-1.22	0.54
size	8	86	3.13	1.71	3.00	3.04	1.48	1.00	6.00	5.00	0.44	-0.98	0.18
pct	9	86	5.80	4.12	5.80	5.39	3.71	0.30	23.40	23.10	1.44	3.77	0.44

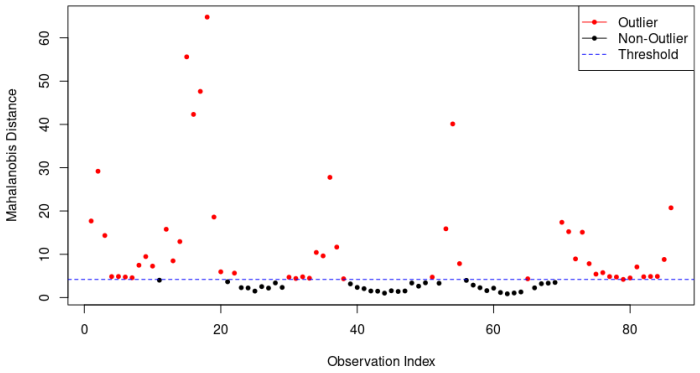
As we can now confirm, the standard deviation of phouse and pmisc are both close to 0, as we expected from the close values of their measures of location. The standard deviation of the income is the highest among all the analyzed variables, as we predicted earlier. Its skewness is -0.02 which is close to 0, thus confirming the hypothesis of a symmetrical curve. The skewness of wmisc is positive and far from 0, also confirming the right skewness of this variable's distribution. The fact that the trimmed mean of wmisc is significantly smaller than the regular mean confirms the presence of outliers of higher value than the 3rd quartile. It also demonstrates the robustness of trimmed means in comparison to normal means, since the normal mean is highly affected by the outliers.

Another manner of confirming the presence of outliers in wmisc is the fact that the MAD is significantly smaller than the standard deviation. Both are measures of dispersion, but the MAD is more robust to outliers since its formula uses the median instead of the mean, which is used in the standard deviation. A MAD much lower than the standard deviation happens when there are outliers since outliers cause a high increase in the standard deviation but have no such effect on the MAD.



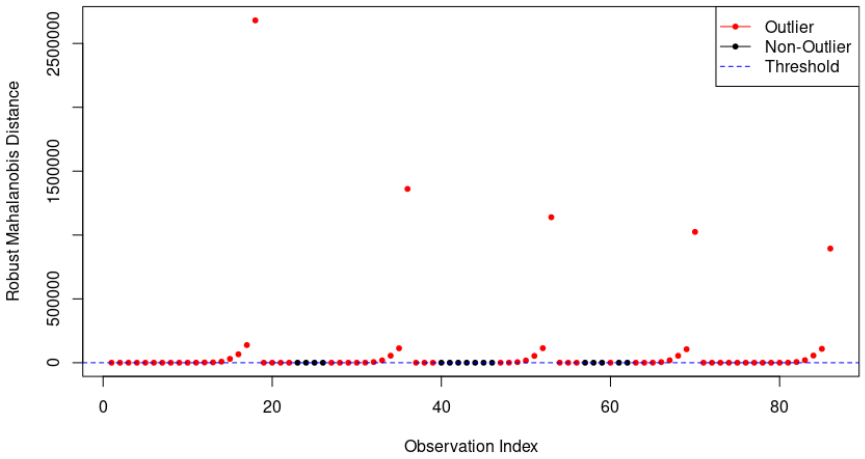
As we can see from the boxplot, and just as was predicted, wmisc has a lot of outliers that are higher in value than the 1.5*IQR+3rd quartile. On the other hand, size has no outliers, which makes sense given that the mean and the trimmed mean are in close proximity to one another, and so are the MAD and standard deviation. Another aspect we can observe in the boxplots are the differences in scale between phouse and totexp with a smaller scale, and wmisc and size with a bigger one. This leads to higher values of standard deviation in the variables with the bigger scale, and might indicate a necessity for some standardization in the data for certain analysis, such as using standardized PCs instead of non standardized ones.

So far we have detected outliers with regard to a single variable, but since we are analyzing multivariate data there are outliers that can only be detected when looking at all the variables at the same time. One effective method to achieve this is the Mahalanobis distance.



The threshold was set as the square root of the 0.975 quantile of the chi-squared distribution with as many degrees of freedom as variables analyzed. As we can see clearly, the data has many outliers, and some have a Mahalanobis distance more than 6 times bigger than others. This suggests that they are extreme outliers.

The Mahalanobis distance can be made more robust to outliers if it is combined with the Mininum Covariance Determinant (MCD) estimator, which is used to calculate robust estimates of the central value and covariance of the data.



The threshold in this graph was set to be the same value. In this plot we can see that certain outliers have a robust Mahalanobis distance many orders of magnitude greater than anything we saw in the previous graph. Those are the extreme outliers also identified earlier. An explanation for what happened is that these extreme outliers were significantly influencing the calculation of the sample covariance matrix and sample mean used in the traditional Mahalanobis distance, and the Mahalanobis distance was then underestimating the degree to which the outliers were distant from the central tendency of the data. The MCD estimates are less affected by those outliers, thus demonstrating the real degree to which they are far away from the central tendency of the data.

	wfood	whouse	wmisc	phouse	pmisc	totexp	income	size	pct
wfood	1.000	0.396	0.215	0.261	0.343	-0.274	-0.253	0.042	-0.514
whouse	0.396	1.000	0.837	0.297	0.027	0.489	0.388	-0.459	-0.626
wmisc	0.215	0.837	1.000	0.072	-0.174	0.563	0.601	-0.399	-0.604
phouse	0.261	0.297	0.072	1.000	0.059	0.291	-0.023	-0.056	-0.222
pmisc	0.343	0.027	-0.174	0.059	1.000	-0.539	-0.720	-0.203	0.169
totexp	-0.274	0.489	0.563	0.291	-0.539	1.000	0.870	0.022	-0.422
income	-0.253	0.388	0.601	-0.023	-0.720	0.870	1.000	0.072	-0.482
size	0.042	-0.459	-0.399	-0.056	-0.203	0.022	0.072	1.000	0.056
pct	-0.514	-0.626	-0.604	-0.222	0.169	-0.422	-0.482	0.056	1.000

This is the sample correlation matrix of the data. Correlation matrices entries can vary from -1, indicating that the 2 variables are related exactly linearly via a function of negative slope, to 1, in which the slope is positive. A 0 correlation indicates no linear association between the data, and then the closer we are to 1 or -1 the closer we are to a perfect linear association of positive or negative slope, respectively. We can see a high positive correlation between wmisc and income and a negative correlation between food and income . This indicates that as the income of a family rises, they tend to spend a higher share of their income on miscellaneous items and services, and a lower share on their food expenses. This is natural, since higher income families can more easily afford to pay for their food and still have enough spare money for non essential expenses, such as expensive clothing.

2. PCA

a) Original Scale Vs Standardized PCA

i) Considering the variables in original scale and the classical sample covariance estimate

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	5.852	3.3449	2.01619	1.06602	0.52455	0.44129	0.09303	0.003133
Proportion of Variance	0.670	0.2189	0.07953	0.02223	0.00538	0.00381	0.00017	0.000000
Cumulative Proportion	0.670	0.8889	0.96840	0.99064	0.99602	0.99983	1.00000	1.000000

Rotation (n x k) = (10 x 10):

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
wfood	8.841211e-03	0.2247595681	-5.292400e-02	3.370465e-01	-6.294434e-01	6.546675e-01	-9.062840e-02	6.567126e-04
whouse	1.120935e-01	0.1281884502	2.812497e-01	2.798690e-01	-5.640639e-01	-6.898985e-01	1.394490e-01	-7.907288e-04
wmisc	2.496289e-01	0.1023187197	5.158382e-01	6.417128e-01	4.809459e-01	1.315034e-01	-2.700070e-02	-8.449225e-04
phouse	2.696112e-05	0.0001488031	4.326958e-06	3.329495e-05	-1.087949e-03	-9.676312e-04	-1.187095e-02	-1.099709e-01
pmisc	-5.123386e-04	0.0007774599	5.148334e-04	8.140204e-04	3.818402e-04	-1.673469e-03	-4.275520e-03	9.939282e-01
totexp	3.390530e-02	-0.0238491573	6.692404e-03	4.176383e-03	-3.559646e-02	-1.644460e-01	-9.847582e-01	-3.109370e-03
year	0.000000e+00	0.0000000000	4.336809e-19	0.000000e+00	5.551115e-17	-5.551115e-17	-3.330669e-16	-3.400058e-16
income	7.899824e-01	-0.5785721416	-4.034437e-02	-9.041812e-02	-1.470484e-01	9.404625e-02	3.015391e-02	1.244323e-03
size	-5.219430e-03	-0.1079481904	-7.694423e-01	5.829939e-01	1.286743e-01	-1.976209e-01	2.802547e-02	-2.866030e-04
pct	-5.475336e-01	-0.7586921600	2.414425e-01	2.195501e-01	-1.219639e-01	5.667878e-02	-2.962307e-03	1.335554e-04

Retaining the first two principal components is suggested, following a principal components analysis which considers variables in their original scale and uses the classical sample covariance estimate.

In this scenario, the matrix of loadings comprises two columns, each representing a principal component, and ten rows, corresponding to each of the ten initial variables.

Analyzing the loadings of each variable on the principal components (PCs) helps in understanding their contributions to the PCs. As an example, in PC1, 'income' has the highest loading, followed by 'pct', 'whouse', and 'wmisc'. This indicates that variables such as household income, cell weight (utilized in survey data to modify the observation's impact on the analysis, mirroring the general population's structure), the proportion of total expenditure on housing and fuels, and the price of miscellaneous items are most strongly associated with PC1. Conversely, variables like 'phouse', 'pmisc', 'wfood', 'size', and 'totexp', which represent the price of housing and fuels, price of miscellaneous items, proportion of total expenditure on food, household size, and total household expenditure, respectively, show a lesser association with PC1.

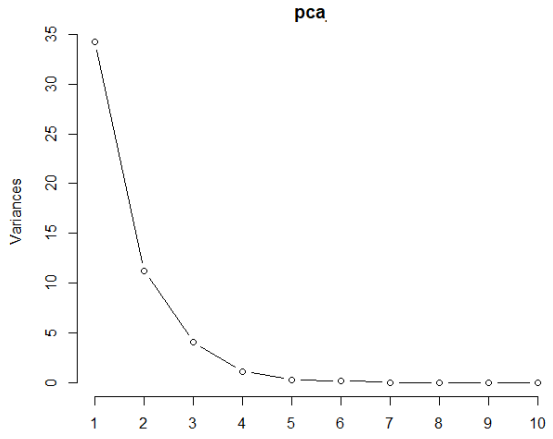
By comparing a variable's loadings across various principal components (PCs), we can understand its influence on each PC. Take 'income', for instance: its loading is higher for PC1 (0.789) than for PC2 (-0.578), indicating that 'income' has a more significant relationship with PC1 than with PC2.

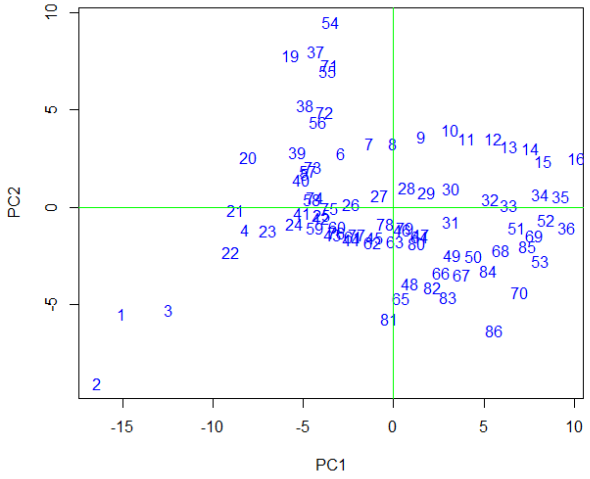
For PC2, 'income' has the highest loading, succeeded by 'pct' and 'wfood', indicating that variables like household income, cell weight, and the share of total food expenditure are most closely related to PC2. Conversely, variables such as 'phouse', 'pmisc', 'totexp', 'wmisc', 'size', and 'whouse', which exhibit the lowest loadings for PC2, demonstrate a less significant connection with this principal component.

Given that PC1 and PC2 heavily factor in these two elements (income, pct), it can be inferred that PC1 examines the impact of these substances on specific economic or demographic trends, while PC2 likely delves into their influence on a different aspect of the data, perhaps related to spending patterns or resource allocation.

The matrix of loadings is instrumental in revealing the connections between the original variables and the principal components (PCs), offering a deeper understanding of the data's inherent structure. The predominant share of the data's variance is accounted for by the first two PCs. Specifically, PC1, with a standard deviation of 5.852, accounts for 67% of the variance, and PC2, having a standard deviation of 3.3449, contributes to 21.89% of the variance. Collectively, these two PCs elucidate 88.89% of the total variance in the data.

The scree plot, a visual depiction of the variance explained by each principal component (PC), shows a marked decline in the curve after the second PC. This indicates that the initial two PCs account for most of the data's variation, with the subsequent PCs contributing minimally to additional variance. Consequently, keeping just the first two PCs is a prudent decision. It enables capturing the bulk of the data's variance while also simplifying its dimensionality. Moreover, these first two PCs probably represent the most significant patterns and trends within the data, enhancing their interpretability.





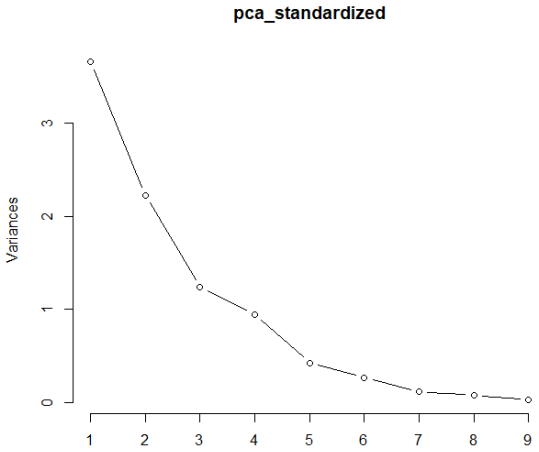
ii) Considering the standardized variables

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8
Standard deviation     1.911 1.4912 1.1136 0.9747 0.65512 0.51838 0.34243 0.28709
Proportion of Variance 0.406 0.2471 0.1378 0.1056 0.04769 0.02986 0.01303 0.00916
Cumulative Proportion  0.406 0.6530 0.7908 0.8963 0.94404 0.97390 0.98693 0.99608

Rotation (n x k) = (9 x 9):
              PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8
wfood   0.07224769  0.5232317 -0.43247509  0.27065009  0.19912335 -0.32651204 -0.07044712  0.5351650 -
whouse  0.42211622  0.3198689  0.15153863 -0.01782561 -0.16995172 -0.28146791 -0.62784711 -0.3556345
wmisc   0.45341470  0.1466416  0.23041003  0.18229196 -0.17122020 -0.34344830  0.68260711 -0.2137879 -
phouse  0.14024710  0.2131578 -0.33828545 -0.83808156  0.14736934 -0.03836403  0.23827715 -0.0722351
pmisc  -0.23253904  0.4886306  0.12623409 -0.05578643 -0.70523250  0.33061564  0.11489475  0.2080142
totexp  0.43097220 -0.2720412 -0.03795584 -0.27426638 -0.35183077  0.10628594 -0.23130008  0.3057148 -
income  0.42870966 -0.3543488 -0.02476320  0.10528164 -0.06841822  0.05880053  0.09392370  0.4763345
size   -0.11823181 -0.2756725 -0.71671861  0.17680096 -0.46563805 -0.20938869  0.02812433 -0.3147969
pct    -0.39212479 -0.2065134  0.30133210 -0.26484474 -0.19440833 -0.72583238 -0.04246844  0.2710651
```

After performing a principal components analysis and standardizing the variables, our analysis suggests retaining the first four principal components. These components together account for a significant portion of the data's variance: specifically, PC1 explains 40.6%, PC2 24.71%, PC3 13.78%, and PC4 10.56%, summing up to 89.63% in total. This decision is in line with the common practice of retaining the principal components that collectively explain roughly 80%-90% of the variance, justifying our choice to keep the first four.

The scree plot, which graphically represents the variance explained by each principal component (PC), indicates a notable decline in the curve beyond the fourth PC. Therefore, choosing to keep the first three PCs is advantageous, as it enables us to retain most of the data's variance while simultaneously simplifying its dimensionality.



b) Reflecting on the results from both PCA analyses:

Standardized Data PCA
- 40.6% of the variance is accounted for by PC1.
- PC2 accounts for another 24.71%.
- Together, PC1 and PC2 explain a total of 65.3% of the variance.

Original Scale PCA- 67% of the variance is explained by PC1.
- PC2 explains an additional 21.89%.
- In sum, PC1 and PC2 on this scale account for 88.9% of the variance.

With the original scale PCA accounting for a greater portion of variance with its first two principal components compared to the standardized PCA, there's an indication towards favoring the original scale PCA when variables are on a similar scale and unit. Conversely, for variables with significant scale or unit differences, the standardized PCA could be more fitting, despite its lower variance proportion, as it equalizes the scale differences by standardizing the variables.

The decision regarding the PCA method to employ should consider both the scale of the original variables and how interpretable the principal components are in each method.Given that the variables are not uniform in scale, using standardized PCA is more appropriate due to its ability to handle varying scales more effectively. This approach is preferable over the original scale PCA, which would be more suitable if the variables were relatively uniform in scale and required less variance normalization.

3. Classical and Robust PCA (With 5 Outliers)

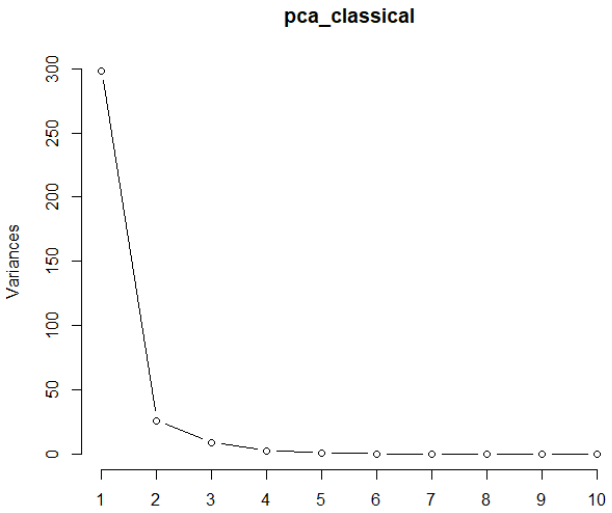
To evaluate the potential effects and outcomes of outliers in both classical and robust PCA, the primary subset of data underwent modification. This change involved substituting the first five observations with their values scaled down to 0.01 of the original figures.

a) Classical PCA

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     17.2701 5.08165 3.00047 1.59833 1.00124 0.47912 0.42621 0.09042
Proportion of Variance  0.8849 0.07661 0.02671 0.00758 0.00297 0.00068 0.00054 0.00002
Cumulative Proportion   0.8849 0.96149 0.98820 0.99578 0.99876 0.99944 0.99998 1.00000

Rotation (n x k) = (10 x 10):
              PC1           PC2          PC3           PC4           PC5          PC6           PC7           PC8
wfood  0.028398422  0.0314229104 -0.239779395 -0.2244748142 -0.2169070667  0.33247727  0.8489090447 -0.109930215
whouse 0.020452407 -0.1210100353 -0.233768220  0.1443701529 -0.2638937273  0.81676210 -0.3944185307  0.130923792
wmisc  0.030283110 -0.2813205812 -0.288702625  0.3575486761 -0.7287469292 -0.42069278 -0.0024397012 -0.026272003
phouse 0.002112679  0.0002282894 -0.000414970 -0.0001705208  0.0002568380  0.00111084 -0.0007281301 -0.011968930
pmisc  0.002288054  0.0008608948 -0.001083966  0.0002743623 -0.0006991902  0.00011163 -0.0019220165 -0.003907094
totexp 0.004983221 -0.0392467958  0.013779610 -0.0001365647 -0.0009218166  0.08095615 -0.1542044911 -0.983743318
year   0.984217337  0.1155434926 -0.107216679  0.0027573139  0.0613965067 -0.03673173 -0.0364833986  0.001548665
income 0.150423088 -0.8690679539  0.427442954 -0.0262010827  0.1060668214  0.09014063  0.1362839550  0.027370718
size   0.044663324  0.0677042339  0.278764714 -0.7847018297 -0.4913171919 -0.04999981 -0.2342323308  0.034598732
pct    0.067215665  0.3612111959  0.729660996  0.4295067834 -0.3095571821  0.16384883  0.1588367573 -0.015034189
```

With five variables in the dataset replaced by outliers, we conducted a principal components analysis on the original scale variables, applying the standard sample covariance estimate. This analysis led us to the conclusion that only the first principal component should be retained.



This first principal component, PC1, with a standard deviation of 17.27, accounts for a significant 88.49% of the data's variance. This high percentage falls within the usual practice of preserving principal components that collectively explain roughly 80%-90% of the variance, hence justifying our decision to retain only the first component.

Utilizing the classical PCA method in this modified dataset, our results varied notably from previous analyses, prompting us to opt for only the first principal component. The introduction of outliers significantly skewed the results of the classical PCA. These outliers concentrated a major part of the variance within PC1, thereby impacting the overall variance distribution and influencing our choice to focus solely on the first principal component for our analysis. This outcome highlights the significant effect outliers can have on statistical methods like PCA.

b) Robust PCA based on the MCD estimate

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7      PC8
Standard deviation     6.1089 3.6045 1.90938 1.19835 0.57274 0.50949 0.10802 0.003728
Proportion of Variance 0.6665 0.2321 0.06511 0.02565 0.00586 0.00464 0.00021 0.000000
Cumulative Proportion  0.6665 0.8985 0.96365 0.98930 0.99516 0.99979 1.00000 1.000000
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
wfood	2.924272e-02	-2.344854e-01	2.247047e-01	2.207899e-01	3.379763e-01	8.477066e-01	1.099298e-01	8.446892e-04
whouse	-1.212241e-01	-2.322530e-01	-1.442184e-01	2.658596e-01	8.146340e-01	-3.988894e-01	-1.309239e-01	-7.901885e-04
wmisc	-2.813263e-01	-2.885656e-01	-3.573581e-01	7.290911e-01	-4.214444e-01	-6.635926e-04	2.627236e-02	-8.326086e-04
phouse	-1.846071e-05	-1.841258e-04	1.765087e-04	-1.231720e-04	1.186622e-03	-6.559496e-04	1.197226e-02	-1.032525e-01
pmisc	5.931258e-04	-8.182846e-04	-2.673007e-04	8.501886e-04	1.911684e-04	-1.833669e-03	3.910883e-03	9.946492e-01
totexp	-3.964293e-02	1.349258e-02	1.306095e-04	8.086341e-04	8.008149e-02	-1.546583e-01	9.837445e-01	-2.854594e-03
year	-1.368674e-16	-1.923698e-16	2.117322e-16	1.168868e-15	3.223668e-15	3.169629e-15	3.824559e-15	4.073727e-12
income	-8.832921e-01	4.246374e-01	2.611729e-02	-1.062098e-01	9.106179e-02	1.359128e-01	-2.737093e-02	1.264102e-03
size	6.014425e-02	2.836062e-01	7.847846e-01	4.911518e-01	-5.188575e-02	-2.344009e-01	-3.459898e-02	-3.250955e-04
pct	3.462816e-01	7.395086e-01	-4.294589e-01	3.106198e-01	1.647095e-01	1.579280e-01	1.503447e-02	2.205600e-04

After introducing outliers into the dataset by modifying five variables, we conducted a robust principal components analysis. In contrast to the classical PCA approach used previously, this robust analysis led us to a different conclusion: the retention of the first two principal components is optimal.

In this robust PCA, the first principal component (PC1), with a standard deviation of 6.1089, accounts for 66.65% of the data's variance. The second principal component (PC2), with a standard deviation of 3.604, explains an additional 23.21% of the variance. Combined, these two components capture 89.85% of the variance.

This decision to select the first two components, rather than focusing all the variance explanation on just the first component as in the classical PCA, is a direct response to the presence of outliers in our dataset. By choosing two components, we avoid the overfitting issue observed in the classical PCA, where the skewed results led to an excessive concentration of variance in the first component. The robust PCA method, in this case, provides a more balanced and nuanced understanding of the data's structure, considering the impact of the outliers. This approach underscores the importance of adapting analysis techniques to the specific characteristics of the dataset, especially when dealing with anomalies like outliers.

Bibliography

http://qed.econ.queensu.ca/jae/2000-v15.3/bollino-perali-rossi/readme.bpr.txt