# STAT401 Multivariate Statistical Analysis
# Midterm Report (PCA)

2015100059, 통계학과, 윤창원

## Ⅰ. Introduction to Principal Component Analysis

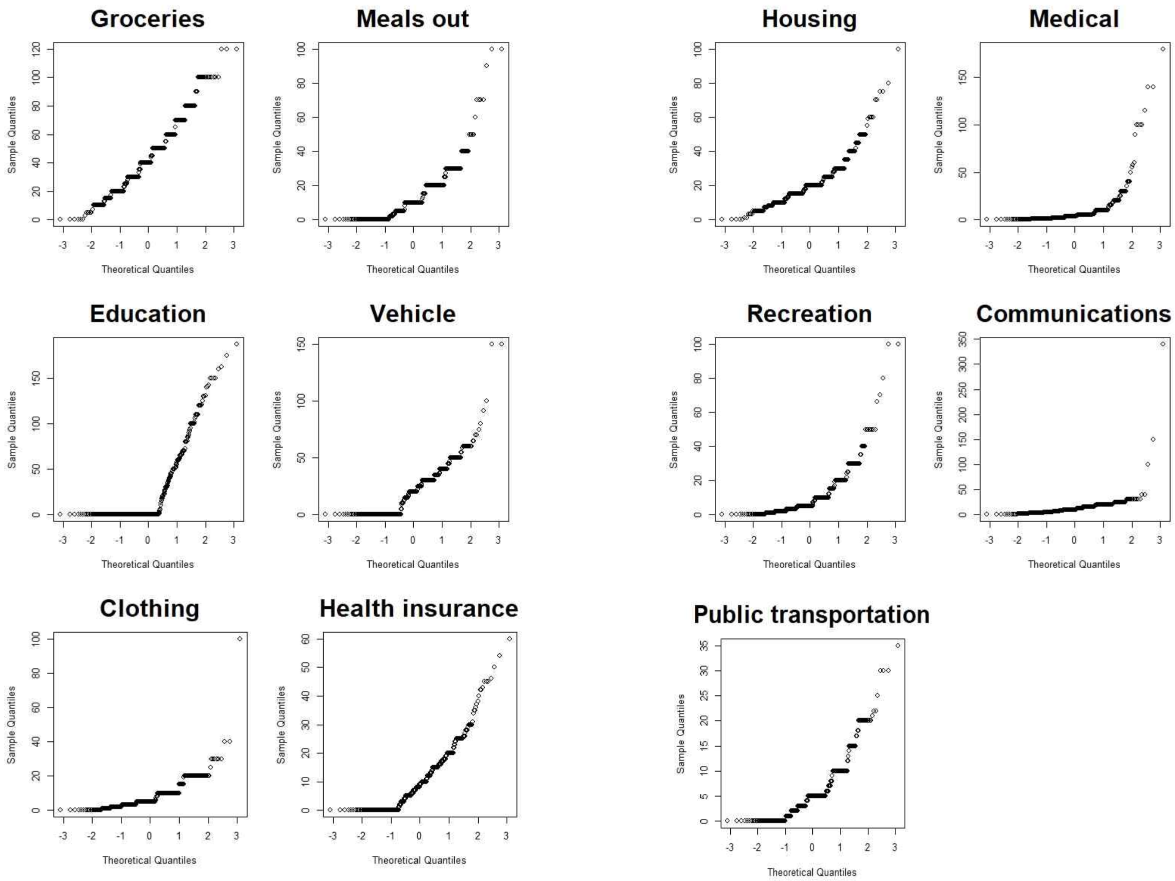
Multivariate data refers to data composed of multiple variables. When you handle those multivariate data, you often get overwhelmed by the number of dimensions it has. The purpose of Principal Component Analysis is to process the data and reproduce new variables that can explain most of the variance of the original data with fewer dimensions. Principal Components are linear combinations of variables from original data and number of Principal Components is equal to the number of variables denoted as 'p'.

Here's how to do it. We extract eigenvalues and eigenvectors from covariance matrix of the original data. Assume Σ(cov matrix) to be positive definite and make eigenvectors orthonormal. Then, for each eigenvalue and eigenvector pairs ($\lambda_i$,$a_i$) with decreasing order of eigenvalue, i-th Principal Component $y_i$ is defined as $y_i = a_ix$ where x = ($x_1$, $x_2$,...,$x_p$) from original data. Now, by Maximization of Quadratic Forms Theorem, $y_i$ is a linear combination with maximum variance possible while var($y_i$) = $\lambda_i$ and cov($y_i$,$y_j$) = 0 (i≠j). And also, it is easy to show that $\sum_{i=1}^{p} var(x_i) = \sum_{i=1}^{p} var(y_i)$ through spectral decomposition of Σ. Hence, through this calculation we can reproduce the data with same amount of variance while each Principal Components having maximum variance possible. This means that first few Principal Components can explain most of the variance of the original data which is the exact purpose of PCA.

## Ⅱ. Data Exploration

The data to be analyzed is from Korean Labor & Income Panel Study and have 16 columns with 500 observations. Each observation is a household. Columns include household ID, household head's Gender, household head's Age, monthly Income, 11 categories of monthly expenses and monthly savings. Income, expenses and savings are in unit of KRW 10,000 and there are 3 missing values in savings. Among them, this report will conduct PCA on 11 categories of monthly expenses.

Before conducting Principal Component Analysis, it is useful to check whether each columns of data follows normal distribution and the whole data follows multivariate normal distribution. For this purpose, qqplot was used.



<Figure 1>

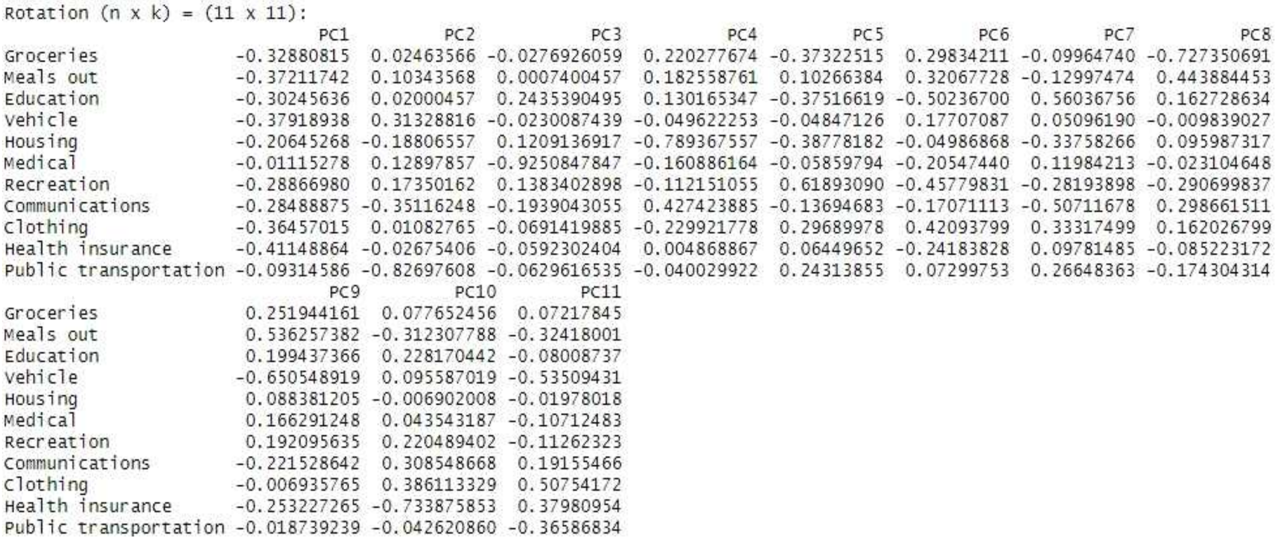
If each columns follows normal distribution, the qqplot will show a 45 degree straight line. However, except the 'Groceries' column, it is clear that other columns deviate from the line hence not following normal distribution, For 'Groceries' column, Shapiro-Wilk normality test was conducted for further investigation. However, it showed p-value of 4.907e-10 which means we should reject the null hypothesis 'The column follows normal distribution'. In conclusion, Every column of the data doesn't follow normal distribution. Since each marginal distributions of multivariate normal distribution should be normal, we can also conclude that the data does not follow multivariate normal distribution.

## Ⅲ. Conducting PCA

When conducting PCA, a researcher can choose whether to use Covariance matrix or Correlation matrix to calculate eigenvalues and eigenvectors from. The main difference between them is whether each columns share the same unit or scale. If they don't, columns with large variance merely due to it's large unit will exert too much influence when we use covariance matrix. Hence they have unnecessarily large power constructing Principal Components. However, if we use correlation matrix, each columns now has equal variance of 1 giving equal weight to each columns.

For our data, even if each columns shares the same unit each categories of expenses has different scale. For example, expenses for education would normally be much higher than expenses for public transportation for most of the households. Hence, the report will use correlation matrix instead of the covariance matrix.

The result of PCA using correlation matrix is as follows.



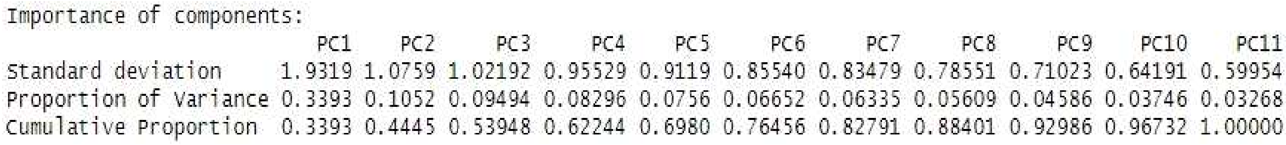
```
Rotation (n x k) = (11 x 11):
                              PC1         PC2           PC3          PC4         PC5         PC6         PC7          PC8
Groceries             -0.32880815  0.02463566 -0.0276926059  0.220277674 -0.37322515  0.29834211 -0.09964740 -0.727350691
Meals out             -0.37211742  0.10343568  0.0007400457  0.182558761  0.10266384  0.32067728 -0.12997474  0.443884453
Education             -0.30245636  0.02000457  0.2435390495  0.130165347 -0.37516619 -0.50236700  0.56036756  0.162728634
Vehicle               -0.37918938  0.31328816 -0.0230087439 -0.049622253 -0.04847126  0.17707087  0.05096190 -0.009839027
Housing               -0.20645268 -0.18806557  0.1209136917 -0.789367557 -0.38778182 -0.04986868 -0.33758266  0.095987317
Medical               -0.01115278  0.12897857 -0.9250847847 -0.160886164 -0.05859794 -0.20547440  0.11984213 -0.023104648
Recreation            -0.28866980  0.17350162  0.1383402898 -0.112151055  0.61893090 -0.45779831 -0.28193898 -0.290699837
Communications        -0.28488875 -0.35116248 -0.1939043055  0.427423885 -0.13694683 -0.17071113 -0.50711678  0.298661511
Clothing              -0.36457015  0.01082765 -0.0691419885 -0.229921778  0.29689978  0.42093799  0.33317499  0.162026799
Health insurance      -0.41148864 -0.02675406 -0.0592302404  0.004868867  0.06449652 -0.24183828  0.09781485 -0.085223172
Public transportation -0.09314586 -0.82697608 -0.0629616535 -0.040029922  0.24313855  0.07299753  0.26648363 -0.174304314
                               PC9         PC10        PC11
Groceries              0.251944161  0.077652456  0.07217845
Meals out              0.536257382 -0.312307788 -0.32418001
Education              0.199437366  0.228170442 -0.08008737
Vehicle               -0.650548919  0.095587019 -0.53509431
Housing                0.088381205 -0.006902008 -0.01978018
Medical                0.166291248  0.043543187 -0.10712483
Recreation             0.192095635  0.220489402 -0.11262323
Communications        -0.221528642  0.308548668  0.19155466
Clothing              -0.006935765  0.386113329  0.50754172
Health insurance      -0.253227265 -0.733875853  0.37980954
Public transportation -0.018739239 -0.042620860 -0.36586834
```

<Figure 2>

For choosing how many Principal Components to use, normally 3 methods are used.

1) Choosing enough Principal Components to take up over 70% to 90% percent of total variance

For our data, cumulative proportion of variance is as follows.



```
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     1.9319 1.0759 1.02192 0.95529 0.9119 0.85540 0.83479 0.78551 0.71023 0.64191 0.59954
Proportion of Variance 0.3393 0.1052 0.09494 0.08296 0.0756 0.06652 0.06335 0.05609 0.04586 0.03746 0.03268
Cumulative Proportion  0.3393 0.4445 0.53948 0.62244 0.6980 0.76456 0.82791 0.88401 0.92986 0.96732 1.00000
```

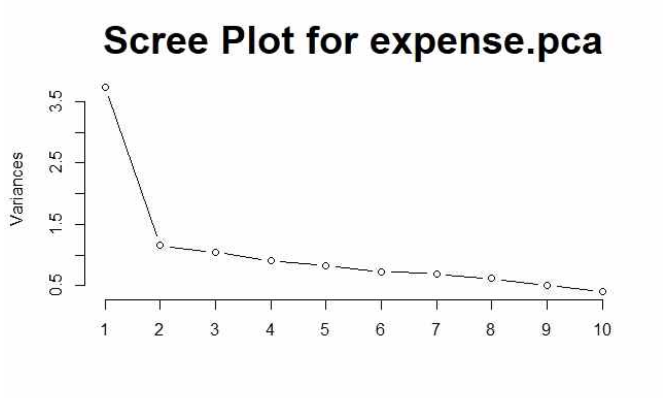
<Figure 3>

Hence we can choose up to PC5 which can explain 69.8% of the total variance.

2) Choosing Principal Components that have larger variance than the average

Since we used correlation matrix, the average of variance of Principal Components equals 1. Hence we can choose Principal Components up to those which have standard deviation larger than 1 from <Figure 3> which are PC1, PC2, and PC3.

3) Using scree plot for variance.



<Figure 4>

The elbow point of the scree plot is 2. Hence we can choose up to PC2.

Combining all 3 methods, this report chose to use up to PC3.

After choosing how many Principal Components to use, we can now interpret the meaning of each Principal Components.

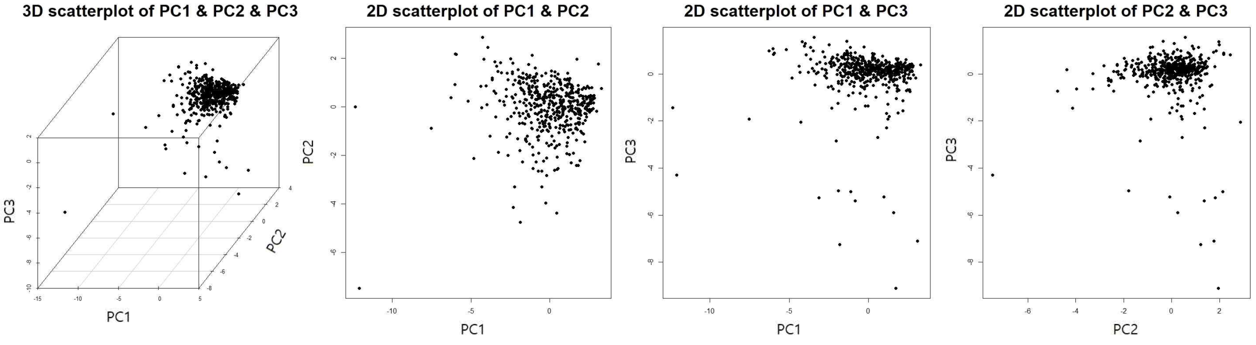
(From Figure 2)

PC1: All coefficients have the same sign with similar values. This means the first Principal Components measures the overall expenses from every category.

PC2: When we delete columns with coefficients less than 0.2, remaining coefficients are {Vehicle: 0.3, Communications: -0.35, Public Transportation: -0.82}. Since Public Transportation has the largest absolute value (almost near 1) with big difference with others, the second Principal Components can be interpreted as almost public transportation expenses itself.

PC3: Deleting columns with coefficients less than 0.2 leaves {Education: 0.24, Medical: -0.92} to remain. As PC2, PC3 also nearly depend on just one column which is Medical with absolute coefficient value nearly 1. Hence, third Principal Component is almost Medical expenses itself.

Using this 3 Principal Components we can calculate Principal Component scores which are projections of original data to the Principal Components. They are new coordinates for original data as Principal Components their new axis.



<Figure 5>

By checking the scatter plot of Principal Components, one can see that the data is uncorrelated. This is because we calculate the Principal Components to be cov($y_i$,$y_j$) = 0 ($i \neq j$). This is a very useful characteristic of PCA since many statistical methods such as linear regression demand each variables to be uncorrelated. Correlation coefficients between Principal Components are as follow.

| Cor(PC1, PC2) | Cor(PC1, PC3) | Cor(PC2, PC3) |
|---|---|---|
| 9.333965e-15 (near 0) | -2.722083e-16 (near 0) | -6.251198e-17 (near 0) |

<Figure 6>

By conducting PCA, we reduced 11 dimensions to 3 dimensions and made each variable uncorrelated. This would make further statistical analysis and computation much more convenient and easy. However, 3 Principal Components chosen only explains 53% of the data's variance. This means we drop almost half of information from original data. Hence, the data have a big trade-off between dimensionality reduction and information gain. This tells us that the data we have may not be appropriate for Principal Component analysis. This may come from the fact that original data itself doesn't show distinct correlations between variables. This helps us understand why second Principal Component and third Principal Component have coefficients concentrated on one specific variable.

## Ⅳ. Conclusion

PCA not only makes it possible to analyze the data with much fewer dimensions but also makes the variables from transformed data to be uncorrelated with each other. This result itself is very useful but what's more meaningful is that we can further conduct other statistical methodologies with Principal Components with much more ease than the original data.

However, PCA is not a panacea. As seen from the example above, some data suffer great information loss from dimensionality reduction. Also, interpretation issue can arise. Sometimes the interpretation of Principal Components itself can be very hard, (normally interpretation gets harder for later Principal Components) but also if we conduct further investigation using Principal Components, interpretation of the results should be different since what we used is totally different variable from original data. Hence a researcher should think thoroughly about this trade-off and whether PCA would be appropriate for his/her research. As the old saying goes, 'There is no free lunch".