Multivariate Analysis of Canadian Consumer Price Index in 2013

Kun Yang

STAT 445

Simon Fraser University

Table of Contents

[Introduction 3](#_Toc384771890)

[How the Dataset Was Generated 3](#_Toc384771891)

[Insight from the Summary of the Data 3](#_Toc384771892)

[The Principal Component Method 4](#_Toc384771893)

[Factor Analysis 5](#_Toc384771894)

[Insight from the Factor Scores and Scatter Plots 6](#_Toc384771895)

[Conclusion 8](#_Toc384771896)

[What We Have Learned and Flaws Checked 8](#_Toc384771897)

[Reference List 9](#_Toc384771898)

[Appendix 9](#_Toc384771899)

[*The raw dataset* 9](#_Toc384771900)

[*R codes* 9](#_Toc384771901)

# Introduction

The Consumer Price Index (CPI) is “an indicator of changes in consumer prices experienced by the target population” (Government of Canada, 2013). The CPI reflects the price fluctuation and therefore provides an important reference for the government in drawing up economic policy. In our project, we will focus on the raw dataset first and then fit it into the principal component method and factor analysis. Finally, we will check the results of the analysis and see which region in Canada has a significantly high CPI. We will also provide some advice to the government on the economic policy.

How the Dataset Was Generated

The analysis will be based on the CPI of each Canadian province in 2013. We will gather the dataset from the Canadian government*.* Accordingly, “the time base is the period for which the CPI is equal to 100; currently this is the year 2002” (Government of Canada, 2013) There are eight variables that have a significant effect on the total CPI in the year 2013. The eight variables show as the following: clothing and footwear, household operations and furnishings, recreation, education and reading, alcoholic beverages and tobacco products, health and personal care, transportation, shelter, and food. The eight variables are set as columns of the table in Excel; the ten provinces of Canada are the rows of the table (see Appendix).

# Insight from the Summary of the Data

In order to get an overview of the dataset, we first check the summary of the raw dataset. We get Table 1 from R.

Table 1.

CF HF RER AT HP TR SF FO

Min. : 88.90 Min. :108.7 Min. : 97.7 Min. :131.4 Min. :109.5 Min. :119.9 Min. :113.3 Min. :127.4

1st Qu.: 93.80 1st Qu.:112.7 1st Qu.:106.4 1st Qu.:140.0 1st Qu.:113.3 1st Qu.:124.8 1st Qu.:126.8 1st Qu.:132.9

Median : 95.10 Median :114.5 Median :107.3 Median :146.8 Median :115.9 Median :126.5 Median :132.2 Median :133.9

Mean : 95.74 Mean :114.1 Mean :106.5 Mean :149.5 Mean :116.2 Mean :126.5 Mean :134.9 Mean :134.3

3rd Qu.: 97.47 3rd Qu.:115.4 3rd Qu.:107.7 3rd Qu.:158.5 3rd Qu.:118.2 3rd Qu.:128.9 3rd Qu.:145.3 3rd Qu.:137.5

Max. :104.60 Max. :119.8 Max. :112.6 Max. :168.3 Max. :127.1 Max. :131.7 Max. :156.3 Max. :139.6

Standard Deviation

4.632542 2.948069 3.784545 13.18552 4.81974 3.579649 13.02809 3.794792

From the minimum, mean, and maximum, we notice that the alcoholic beverages and tobacco have the highest CPI, followed by shelter in second place. In addition, these two variables also have the largest standard deviations. This means that the increasing rate of alcoholic beverages and tobacco prices and the price of shelter have large differences among different provinces. Moreover, the price of clothing and footwear and the price of recreation, education, and reading did not change a lot (based on the minimum, mean, and maximum). The government can pay more attention to the control of the increasing price of alcoholic beverages, tobacco, and shelter. Also, our analysis mainly focuses on these two variables.

# The Principal Component Method

We extracted the principal component by using R. Because each observation of the raw dataset has the same scale unit, we can first compute the variance-covariance matrix.

Table 2.

$values

[1] 206.5961215 170.7129826 28.3981380 18.4465799 8.2106636 3.9605564 1.8715526 0.3110722

$vectors

PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8

CF -0.0922924034 -0.16736143 -0.4446284 -0.559382995 0.64489645 0.17075611 -0.008714134 -0.08799702

HF -0.0475705124 -0.12743462 0.3641921 -0.220892574 0.19595479 -0.23179026 0.152550635 0.82744869

RER -0.0008788315 -0.09171616 -0.4676890 -0.407092466 -0.58595518 -0.38484808 0.336118842 0.05198737

AT -0.8071315496 -0.47454679 0.1057134 -0.008298649 -0.20947705 0.10179389 -0.235977525 -0.04660238

HP 0.0186435197 0.26070195 0.3749321 -0.531593163 -0.30256480 0.62720215 0.146734375 -0.04541552

TR 0.0819603510 -0.02134278 0.4937544 -0.398943757 0.08459593 -0.56527056 -0.230630369 -0.45825738

SH -0.5370265767 0.80110117 -0.1060555 -0.017332779 0.10555291 -0.20979093 -0.015875371 0.05371691

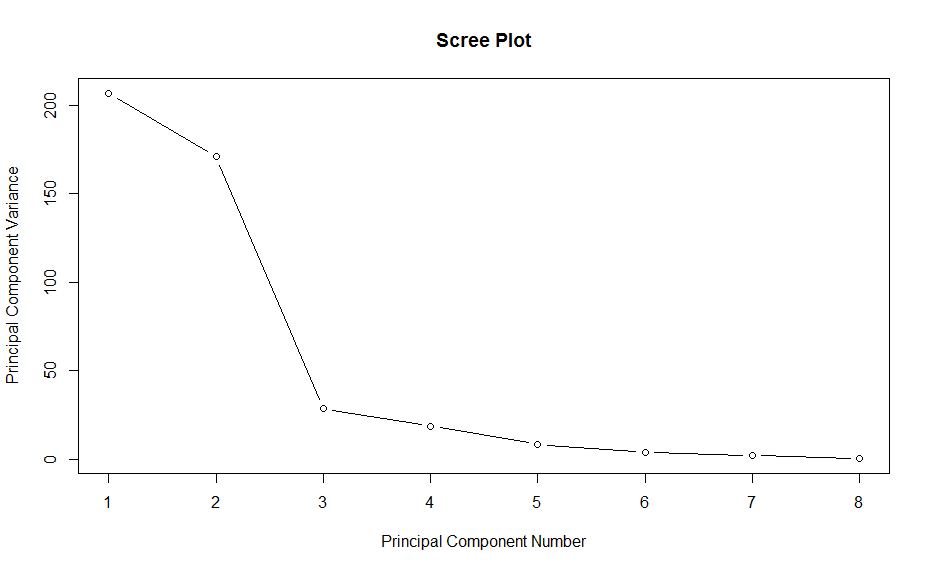
FO -0.2056560816 -0.10936155 0.2101097 0.174516918 0.22055121 -0.04147296 0.856164545 -0.29624792

Standardized eigenvalues:

0.4711345712 0.3893044422 0.0647608700 0.0420667214 0.0187241049 0.0090318977 0.0042680043 0.0007093883

The sum of the first two of these standardized eigenvalues is 0.86. Hence, the first two principal components account for about 86% of the overall variance. The first component is mainly negatively driven by the CPI of the alcoholic beverages and tobacco products, and it also negatively measures the CPI of shelter. The second principal component is positively driven by the CPI of shelter and is negatively driven by the CPI of the alcoholic beverages and tobacco products. Based on the principal component method, we cannot find a meaningful conclusion. Therefore, we want to use the factor analysis next.

# Factor Analysis



Call:

factanal(factors = 3, covmat = COV, n.obs = 80)

Uniquenesses:

CF HF RER AT HP TR SF FO

0.753 0.005 0.842 0.121 0.282 0.058 0.005 0.132

Loadings:

Factor1 Factor2 Factor3

CF 0.408 -0.284

HF 0.549 0.831

RER -0.191 -0.348

AT 0.937

HP -0.356 0.213 0.739

TR -0.205 0.930 0.185

SH 0.149 -0.439 0.883

FO 0.927

Factor1 Factor2 Factor3

SS loadings 2.395 1.839 1.569

Proportion Var 0.299 0.230 0.196

Cumulative Var 0.299 0.529 0.725

Test of the hypothesis that 3 factors are sufficient.

The chi square statistic is 79.15 on 7 degrees of freedom.

The p-value is 2.05e-14

Figure 1.

Based on the scree plot in Figure 1, we first chose three factors. However, the p-value of the three factors was 2.05e-14, so it was obviously insufficient. Thus, we tried to use four factors, and we got new output from R (see Figure 2).

Call:

factanal(factors = 4, covmat = COV, n.obs = 80)

Uniquenesses:

CF HF RER AT HP TR SH FO

0.415 0.005 0.288 0.082 0.276 0.079 0.005 0.046

Loadings:

Factor1 Factor2 Factor3 Factor4

CF 0.298 -0.132 0.692

HF 0.555 0.807 -0.188

RER -0.126 -0.149 -0.137 0.809

AT 0.914 0.284

HP -0.319 0.313 0.695 -0.203

TR -0.166 0.933 -0.148

SH 0.183 -0.322 0.913 -0.154

FO 0.970

Factor1 Factor2 Factor3 Factor4

SS loadings 2.353 1.748 1.391 1.311

Proportion Var 0.294 0.219 0.174 0.164

Cumulative Var 0.294 0.513 0.687 0.851

Test of the hypothesis that 4 factors are sufficient.

The chi square statistic is 14.37 on 2 degrees of freedom.

The p-value is 0.000757

Figure 2.

The p-value is not sufficient on 0.05 levels but it is better than choosing three factors. We tried to choose five factors but it was not allowed by R. Thus, we can do the following interpretation based on the four outputs. **Factor one** puts maximum weight on food () and also places large weights on the variable of alcoholic beverages and tobacco (). Thus, the first factor reflects the information of the CPI of the edibles (or food in general). **Factor two** puts significant and maximum weight on transportation () and also puts large weights on household operations and furnishings (). **Factor three** puts significant and maximum weight on shelter () and also places large weights on health and personal care (). Thus, the third factor reflects the price index on shelter and health care aspects. **Factor four** puts significant and maximum weight on recreation, education, and reading () and also places large weights on clothing and footwear (). We can say this factor reflects the expenditure on clothing, recreation, and education in general. These four factors collectively accounted for 85.1% of the total variation. In addition, the result of the rotation matrix is the same (see Appendix). Thus, the economic policy can be adjusted based on further analysis of these four factors.

# Insight from the Factor Scores and Scatter Plots

In order to do further analysis, we need to compute the factor scores of each Canadian province. Table 3 shows the results of using the “Bartlett” method (we do not remember whether we have studied this from our lecture; we just picked up this method from R based on its description). “Bartlett” given Bartlett’s weighted least-squares scores (Swiss Federal Institute of Technology Zurich, n.d.).

Table 3.

Factor1 Factor2 Factor3 Factor4

AB -0.80587280 0.48226578 2.1831571 0.320479903

BC -1.81916732 -0.08580589 -1.2223669 1.558161885

MB 0.02591584 -0.19050349 -0.2788008 0.508913381

NB 1.15623365 -0.36912618 -1.0560383 0.006099167

NL -0.05310562 -0.38498521 0.7706060 -0.950200675

NS 1.32427398 -0.54810994 -0.1666494 0.073739126

ON -0.66828698 0.99074589 -0.2648828 -0.479876627

PE 1.25806576 1.58852799 0.3588330 1.258749714

QC -0.36784089 0.62307596 -0.8242865 -2.355378566

SK -0.05021561 -2.10608492 0.5004286 0.059312692

From the chapter on “Insight from the Summary of the Data,” we know that the variables of alcoholic beverages, tobacco, and shelter have the highest CPI and are quite distinct in different provinces. Therefore, I would like to construct a score plot for *factor one (the edibles) and factor three (shelter and health care)* (see Figure 3).

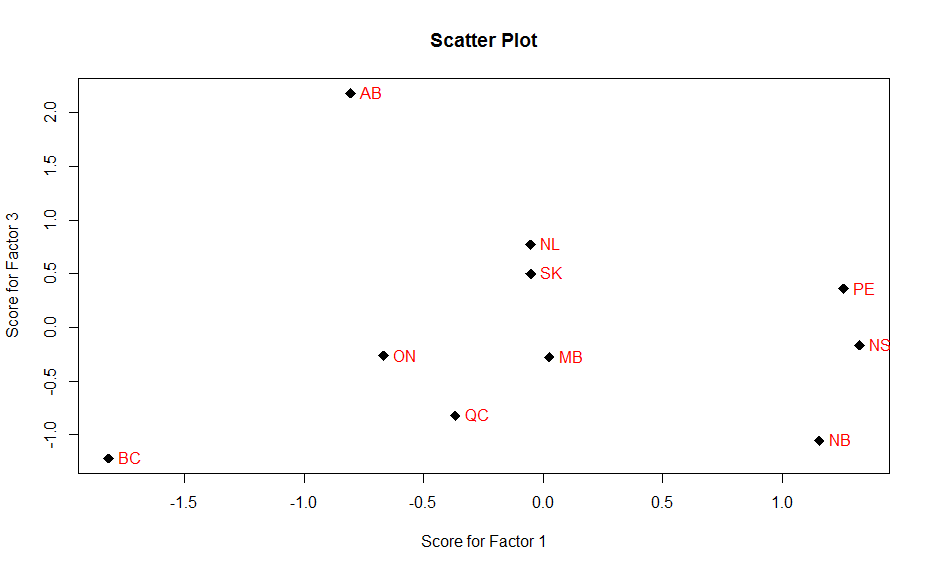


Figure 3.

From the above plot, we find that the provinces of Nova Scotia, Prince Edward Island, and New Brunswick got very high scores from the *first factor (x-axis)*. This means that the price of edibles (or CPI of foods) in these three provinces increased more rapidly than in the other provinces. British Colombia (BC) had an apparent negative score in factor one. This means that the price of edibles in BC increased at an extremely low rate compared to the three negative variables in factor one in BC (recreation, education and reading, health and personal care, and transportation).

From the axis of *factor three (y-axis)*, we found that only Alberta stands out with an extremely high CPI for shelter and health care. Most of the other provinces have a score between 1.0 and -1.0. However, we found again that BC has the smallest score in factor two. It reflects that the price of shelter and health care have not increased at a high rate compared to the three negative variables in factor three in BC. ( Clothing and footwear, household operations and furnishings, and recreation, education, and reading) So, compared to the results of BC in factor one, the recreation, education, and reading is the common variable and it is the major reason for the CPI increase in BC.

# Conclusion

In this project, we used the first principal method and fact analysis to process the data. Based on the results, we can give some advice to the government on how to depress the CPI. In the Atlantic Provinces of Canada (Nova Scotia, Prince Edward Island, and New Brunswick), the CPI mainly was driven by edibles (food, alcohol, and tobacco). The government may exercise more control over the price of edibles in order to depress the CPI. However, in the western provinces, the CPI of Alberta mainly was driven by the prices of shelter and health care. The government may want to pay more attention to rental prices and medical expenses in this region. In BC, the CPI mainly was driven by recreation, education, and reading. Thus, the government could do more work on controlling the prices of books and tuition.

# What We Have Learned and Flaws Checked

In order to do this project, we practiced the principal component method and factor analysis. We learned how to fit the factor loading back and get the factor scores by using the “Bartlett” method. There are some flaws we noticed. The scree plot shows that three factors can be enough, but the p-value is insufficient. We chose four factors, but the p-value was still insufficient. We had no other choice but to choose four factors. It was difficult to interpret the meaning of some factors, especially factors two and four. And we did not check the outliers because the dataset has already been processed by government of Canada. There will be no outlier we can delete.

# Reference List

Government of Canada. (2014). *Consumer Price Index, by province.* Retrieved from http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/econ09a-eng.htm

Government of Canada. (2013). *Data quality, concepts and methodology: Data quality,*

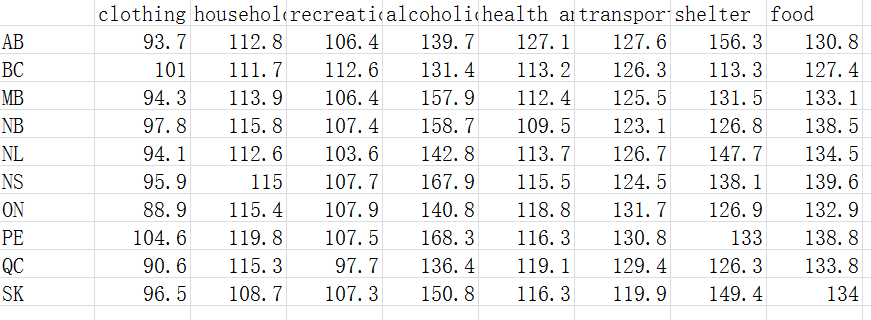
*concepts and methodology.* Retrieved from <http://www.statcan.gc.ca/pub/62-001-x/>

2013011/technote-notetech2-eng.htm

Swiss Federal Institute of Technology Zurich (n.d.). *Factor analysis*. Retrieved from <http://stat.ethz.ch/R-manual/R-patched/>library/stats/html/factanal.html

# Appendix

*The raw dataset*



*R codes*

data<-read.csv("445 Project.csv")

summary(data[2:9])

sd(data[,2]) ………………………. sd(data[,9])

###Principal components analysis

###The eigenvalues and eigenvectors:

library("MVA")

(COV=cov(data[,2:9]))

eig=eigen(COV)

colnames(eig$vectors)=c('PC1','PC2','PC3','PC4','PC5','PC6','PC7','PC8')

rownames(eig$vectors)=c('CF','HF','RER','AT','HP','TR','SF','FO')

###Then the rescaled eigenvalues:

(eig$values/sum(eig$values))

plot(1:length(eig$values),eig$values, main="Scree Plot",

xlab="Principal Component Number", ylab="Principal Component Variance",type="b")

###factor analysis (analysis with 4 factors)

(FA=factanal(covmat=COV,factors=4,n.obs=80) )

L=FA$loadings

load<-factanal(~data[,2]+data[,3]+data[,4]+data[,5]+data[,6]+data[,7]+data[,8]+data[,9],

factors=4,scores ="Bartlett")$scores

rownames(load)=data[,1]

load

x<-load[,1]

y<-load[,3]

plot(x,y,main="Scatter Plot",xlab="Score for Factor 1",ylab="Score for Factor 3",cex=1.5,pch=18)

text(x,y,row.names(load), cex=1, pos=4, col="red")

> (FA=factanal(covmat=COV,factors=4,n.obs=80，rotation="varimax") )

Call:

factanal(factors = 4, covmat = COV, n.obs = 80, rotation = "varimax")

Uniquenesses:

CF HF RER AT HP TR SH FO

0.415 0.005 0.288 0.082 0.276 0.079 0.005 0.046

Loadings:

Factor1 Factor2 Factor3 Factor4

CF 0.298 -0.132 0.692

HF 0.555 0.807 -0.188

RER -0.126 -0.149 -0.137 0.809

AT 0.914 0.284

HP -0.319 0.313 0.695 -0.203

TR -0.166 0.933 -0.148

SH 0.183 -0.322 0.913 -0.154

FO 0.970

Factor1 Factor2 Factor3 Factor4

SS loadings 2.353 1.748 1.391 1.311

Proportion Var 0.294 0.219 0.174 0.164

Cumulative Var 0.294 0.513 0.687 0.851

Test of the hypothesis that 4 factors are sufficient.

The chi square statistic is 14.37 on 2 degrees of freedom.

The p-value is 0.000757