Chapter 4
Principles Component Analysis (2)

Rung-Hung, Su

Fu Jen Catholic University
Department of Statistics and Information Science
Mail: 141637@mail.fju.edu.tw
Phone: 2905-3194 (SL474)

4.4 Issues Relating to the use of Principal Component Analysis

1) What effect does the type of data (i.e., mean-corrected or standardized data) have on principal components analysis.
2) Is principal components analysis the appropriate technique for forming the new variables? That is, what additional insights or parsimony is achieved by subjecting the data to principal components analysis?
3) How many principal components analysis should be retained? That is, how many new variables should be used for further analysis or interpretation?
4) How do we interpret the principal components (i.e., the new variables)?
5) How can principal components scores be used in future analysis?

4.4 Issues Relating to the use of Principal Component Analysis

1) Effect of type of data on principal components analysis.

For example

Assume that the main objective for the data given in Table 4.7 is to form a measure of the Consumer Price Index (CPI)

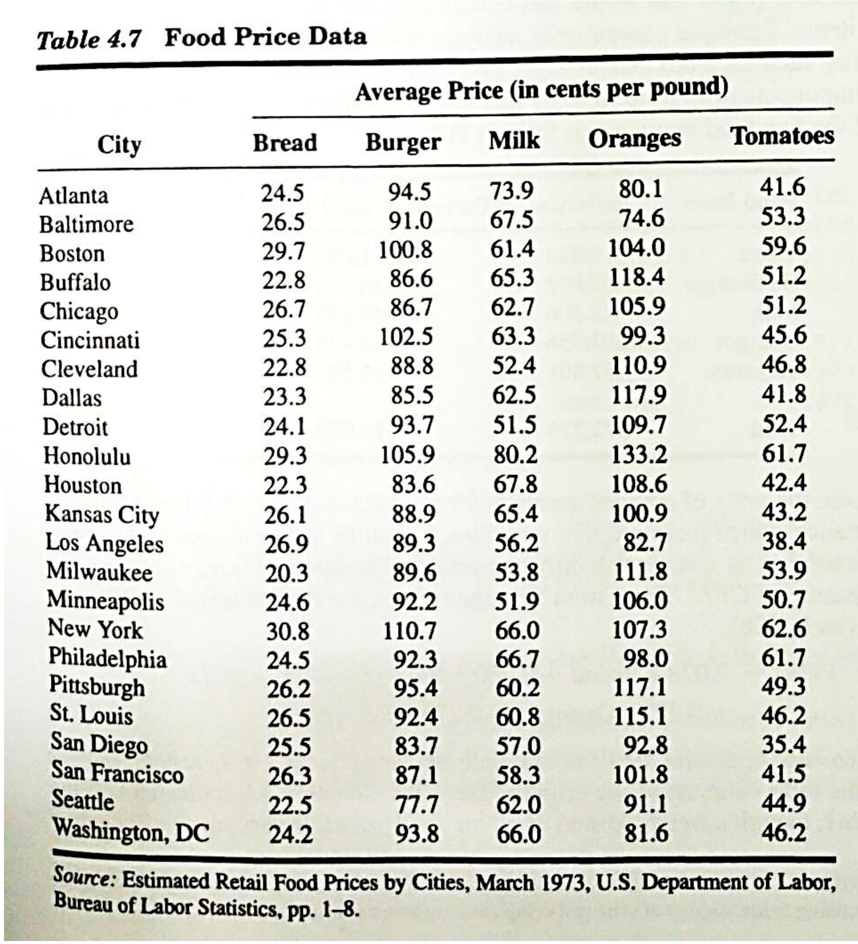


Table 4.7 Food Price Data

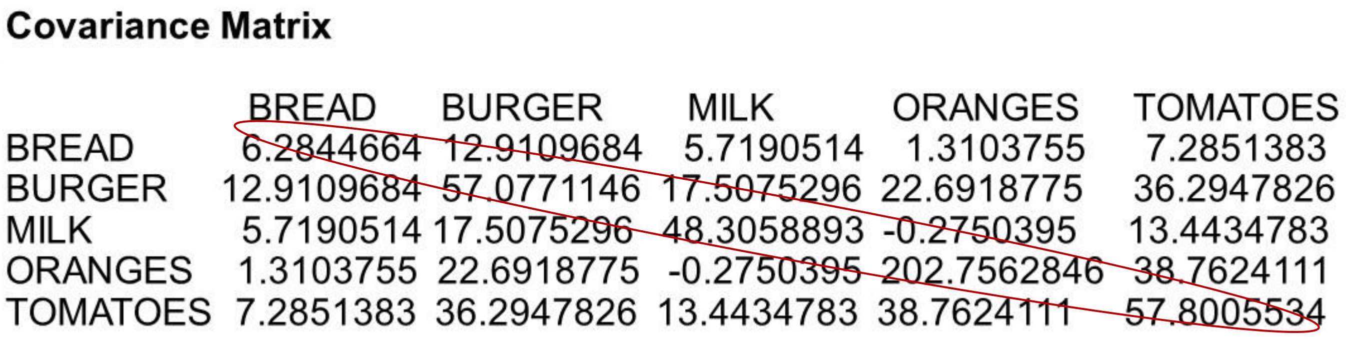
	Average Price (in cents per pound)				
City	Bread	Burger	Milk	Oranges	Tomatoes
Atlanta	24.5	94.5	73.9	80.1	41.6
Baltimore	26.5	91.0	67.5	74.6	53.3
Boston	29.7	100.8	61.4	104.0	59.6
Buffalo	22.8	86.6	65.3	118.4	51.2
Chicago	26.7	86.7	62.7	105.9	51.2
Cincinnati	25.3	102.5	63.3	99.3	45.6
Cleveland	22.8	88.8	52.4	110.9	46.8
Dallas	23.3	85.5	62.5	117.9	41.8
Detroit	24.1	93.7	51.5	109.7	52.4
Honolulu	29.3	105.9	80.2	133.2	61.7
Houston	22.3	83.6	67.8	108.6	42.4
Kansas City	26.1	88.9	65.4	100.9	43.2
Los Angeles	26.9	89.3	56.2	82.7	38.4
Milwaukee	20.3	89.6	53.8	111.8	53.9
Minneapolis	24.6	92.2	51.9	106.0	50.7
New York	30.8	110.7	66.0	107.3	62.6
Philadelphia	24.5	92.3	66.7	98.0	61.7
Pittsburgh	26.2	95.4	60.2	117.1	49.3
St. Louis	26.5	92.4	60.8	115.1	46.2
San Diego	25.5	83.7	57.0	92.8	35.4
San Francisco	26.3	87.1	58.3	101.8	41.5
Seattle	22.5	77.7	62.0	91.1	44.9
Washington, DC	24.2	93.8	66.0	81.6	46.2

Source: Estimated Retail Food Prices by Cities, March 1973, U.S. Department of Labor, Bureau of Labor Statistics, pp. 1–8.

4.4 Issues Relating to the use of Principal Component Analysis

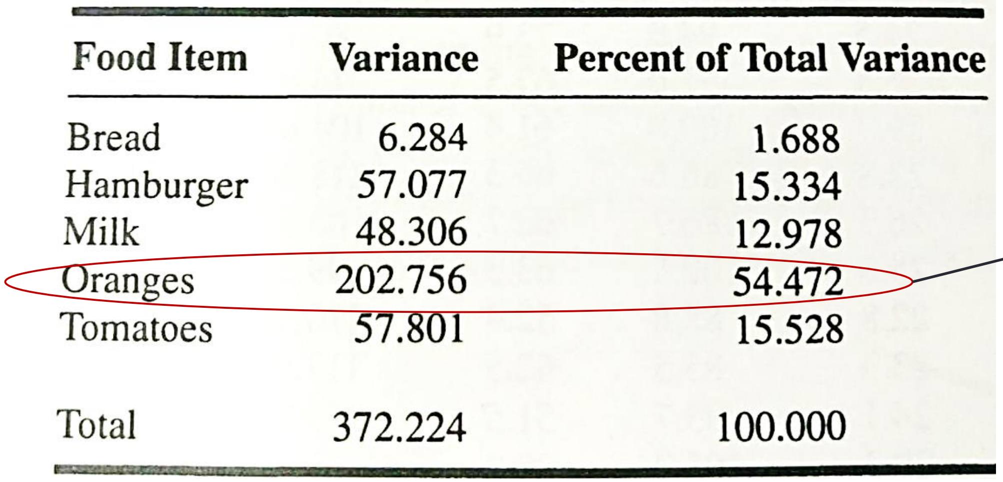
1) **Effect of type of data on principal components analysis.**

SAS was applied to the ***mean-corrected data***



Covariance Matrix

```
            BREAD      BURGER      MILK        ORANGES     TOMATOES
BREAD       6.2844664  12.9109684  5.7190514   1.3103755   7.2851383
BURGER     12.9109684  57.0771146  17.5075296  22.6918775  36.2947826
MILK        5.7190514  17.5075296  48.3058893  -0.2750395  13.4434783
ORANGES     1.3103755  22.6918775  -0.2750395  202.7562846 38.7624111
TOMATOES    7.2851383  36.2947826  13.4434783  38.7624111  57.8005534
```



Food Item	Variance	Percent of Total Variance
Bread	6.284	1.688
Hamburger	57.077	15.334
Milk	48.306	12.978
Oranges	202.756	54.472
Tomatoes	57.801	15.528
Total	372.224	100.000

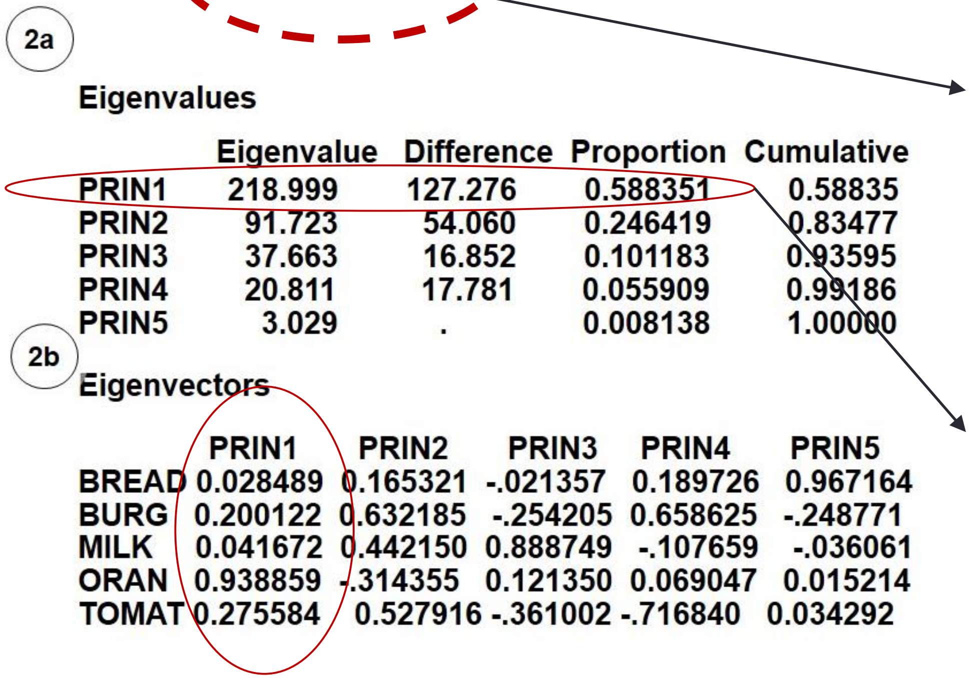
The variance of price of oranges account of a substantial portion of the total variance (almost 55%)

4.4 Issues Relating to the use of Principal Component Analysis

1) **Effect of type of data on principal components analysis.**

SAS was applied to the ***mean-corrected data***

$$\textbf{Prin1} = 0.228\times Bread + 0.200\times Buger + 0.042\times Milk$$
$$+ 0.939\times Orange + 0.276\times Tomatoes$$



```
2a

Eigenvalues

        Eigenvalue  Difference  Proportion  Cumulative
PRIN1     218.999     127.276    0.588351     0.58835
PRIN2      91.723      54.060    0.246419     0.83477
PRIN3      37.663      16.852    0.101183     0.93595
PRIN4      20.811      17.781    0.055909     0.99186
PRIN5       3.029        .       0.008138     1.00000

2b

Eigenvectors

        PRIN1     PRIN2     PRIN3     PRIN4     PRIN5
BREAD  0.028489  0.165321  -.021357  0.189726  0.967164
BURG   0.200122  0.632185  -.254205  0.658625  -.248771
MILK   0.041672  0.442150  0.888749  -.107659  -.036061
ORAN   0.938859  -.314355  0.121350  0.069047  0.015214
TOMAT  0.275584  0.527916  -.361002  -.716840  0.034292
```

Prin 1 is very much affected by the price of orange, due to high variance of orange in data.

Prin 1 account for 58.8% of the total variance

4.4 Issues Relating to the use of Principal Component Analysis

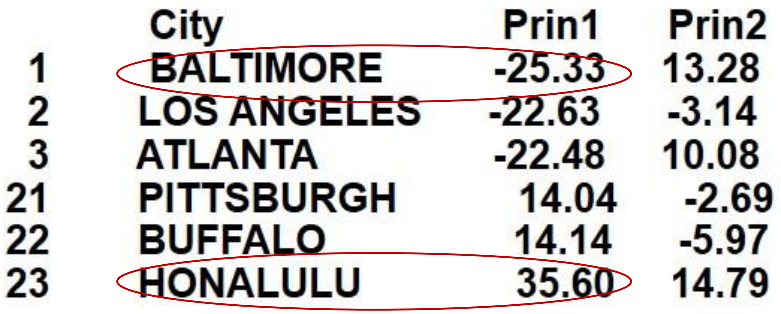
1) **Effect of type of data on principal components analysis.**

SAS was applied to the ***mean-corrected data***

Principal components scores of **Prin 1** suggest that

- Honolulu is the most expensive city
- Baltimore is the least expensive city

The main reason the price of orange dominates the formation of **Prin 1** is that there exists a wide variation in the price of orange across the cities.



	City	Prin1	Prin2
1	BALTIMORE	-25.33	13.28
2	LOS ANGELES	-22.63	-3.14
3	ATLANTA	-22.48	10.08
21	PITTSBURGH	14.04	-2.69
22	BUFFALO	14.14	-5.97
23	HONALULU	35.60	14.79

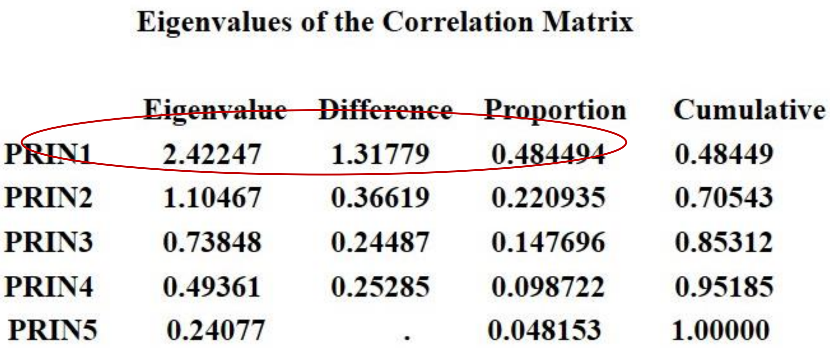
4.4 Issues Relating to the use of Principal Component Analysis

1) **Effect of type of data on principal components analysis.**

SAS was applied to the ***Standardized data*** (the wide variation can not be affected the weights)

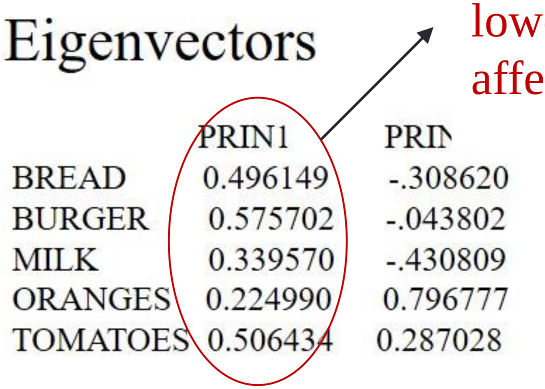
- Each variable accounts for 20% (1/5=0.2) of the total variance
- **Prin 1**, account for 48.44% (2.42247/5=0.4844)of the total variance

$$\boxed{\begin{aligned}\textbf{Prin1} &= 0.496\times Bread + 0.576\times Buger + 0.340\times Milk \\ &+ 0.225\times Orange + 0.506\times Tomatoes\end{aligned}}$$



Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	2.42247	1.31779	0.484494	0.48449
PRIN2	1.10467	0.36619	0.220935	0.70543
PRIN3	0.73848	0.24487	0.147696	0.85312
PRIN4	0.49361	0.25285	0.098722	0.95185
PRIN5	0.24077	.	0.048153	1.00000



Eigenvectors

	PRIN1	PRIN
BREAD	0.496149	-.308620
BURGER	0.575702	-.043802
MILK	0.339570	-.430809
ORANGES	0.224990	0.796777
TOMATOES	0.506434	0.287028

No special higher or lower weight affect **Prin 1**

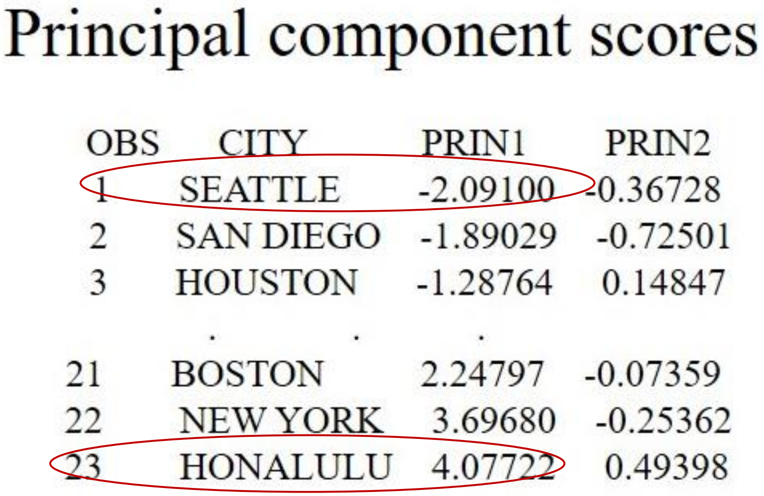
4.4 Issues Relating to the use of Principal Component Analysis

1) **Effect of type of data on principal components analysis.**

SAS was applied to the ***Standardized data*** (the wide variation can not be affected the weights)

Principal components scores of **Prin 1** suggest that

- Honolulu is the most expensive city
- Seattle is the least expensive city



Principal component scores

OBS	CITY	PRIN1	PRIN2
1	SEATTLE	-2.09100	-0.36728
2	SAN DIEGO	-1.89029	-0.72501
3	HOUSTON	-1.28764	0.14847
	.	.	.
21	BOSTON	2.24797	-0.07359
22	NEW YORK	3.69680	-0.25362
23	HONALULU	4.07722	0.49398

4.4 Issues Relating to the use of Principal Component Analysis

1) **Effect of type of data on principal components analysis.**

SAS was applied to the ***Standardized data*** (the wide variation can not be affected the weights)

Results:

- If the variances of the variable do indicate the important of a given variable, then **mean-corrected data** should be used.
- Otherwise, the variances of the variable do indicate the unimportant of a given variable, then **standardized data** should be used.

4.4 Issues Relating to the use of Principal Component Analysis

2) **Is Principal Components Analysis the Appropriate Technique?**

- If the objective is to form uncorrelated linear combinations then the decision will depend on the interpretability of the resulting principal components.
- If the principal components can not be interpreted, one should avoid principal components analysis for forming uncorrelated variables.
- *Substantial loss of information* depends on the purpose for which tte principal components will be used. (100 variables become 5 new variables to interpret the purpose/variance with 99%, but 1% is more important, the principal components analysis is not appropriate)

4.4 Issues Relating to the use of Principal Component Analysis

3) Number of Principal Component to Extract

① **Eigenvalue-greater-than-one rule**: In the case of *standardized data*, retain only those components whose eigenvalues are greater than one. (See ppt page 7, Eigenvalue of **Prin 1** and **Prin 2** are 2.42247>1 and 1.10467>1, respectively, so we choice 2 principal components).

② **Scree Plot**: In the case of *mean-corrected* and *standardized data* Plot the percent of variance accounted for by each principal component and look for an elbow.

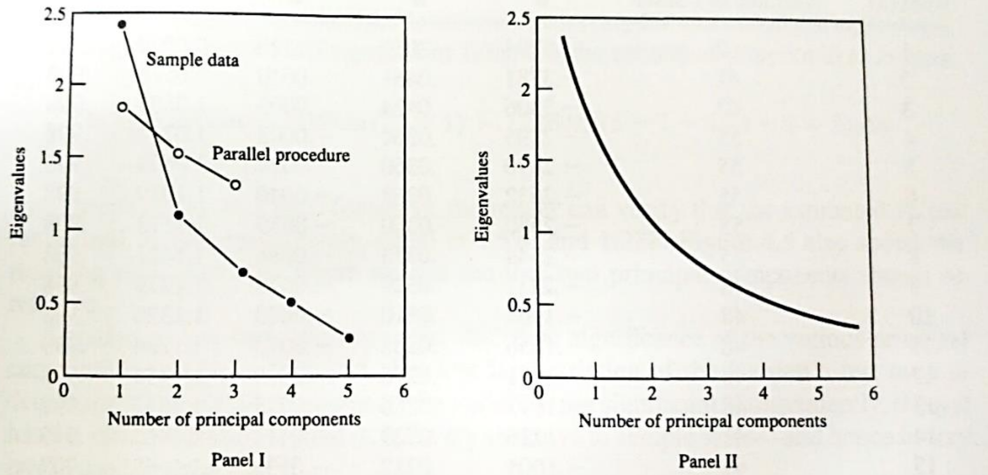


Figure 4.5 Scree plots. Panel I, Scree plot and plot of eigenvalues from parallel analysis. Panel II, Scree plot with no apparent elbow.

4.4 Issues Relating to the use of Principal Component Analysis

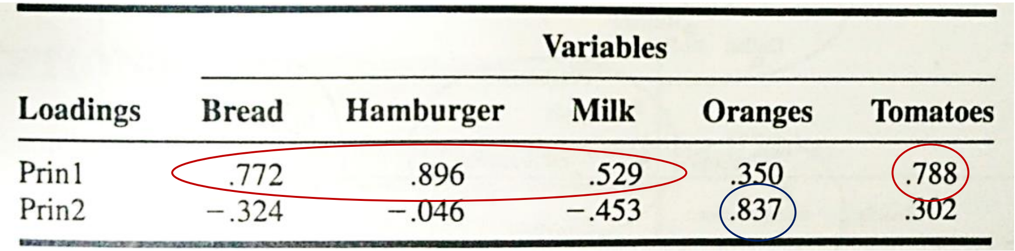
4) Interpreting Principal Components

① By using the loading, the principal components can be interpreted.

② In the standardized case for the example on Table 4.7 (ppt page 3), the loadings for the first two principal components are as follows:

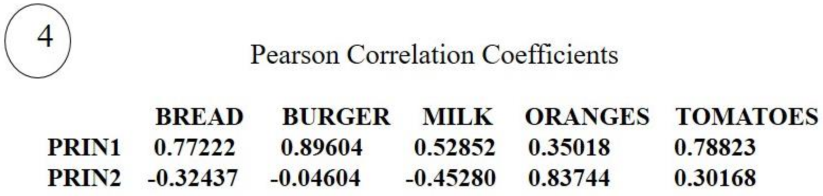
③ The higher the loading of a variable, the more influence it has in the formation of the principal component score and vise versa, Traditionally, the loading >=0.5 as the cutoff point.

④ The first principal component represents the price index for nonfruit items, and second represents the price index of the fruit item. (Loading >=0.5)



	Variables				
Loadings	Bread	Hamburger	Milk	Oranges	Tomatoes
Prin1	.772	.896	.529	.350	.788
Prin2	−.324	−.046	−.453	.837	.302

④



```
                        Pearson Correlation Coefficients

            BREAD      BURGER      MILK      ORANGES    TOMATOES
PRIN1     0.77222     0.89604    0.52852    0.35018    0.78823
PRIN2    -0.32437    -0.04604   -0.45280    0.83744    0.30168
```

4.4 Issues Relating to the use of Principal Component Analysis

5) Use of Principal Components Scores

① The principal components scores can be plot for further interpreting the results.

② From the following principal components scores plot for standardized data in the CPI example, there are five groups cities.

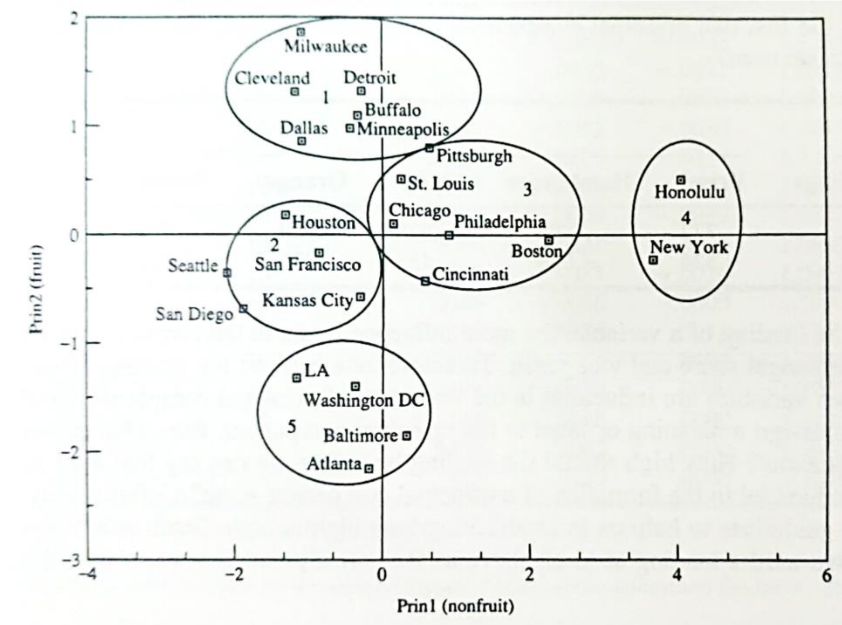


Figure 4.6 Plot of principal components scores.

Homework

4.4 File FOODP.doc gives the average price in cents per pound of five food item in 24 U.S. Cities

a) Using principal components analysis, define price index measure(s) based on the five food items.

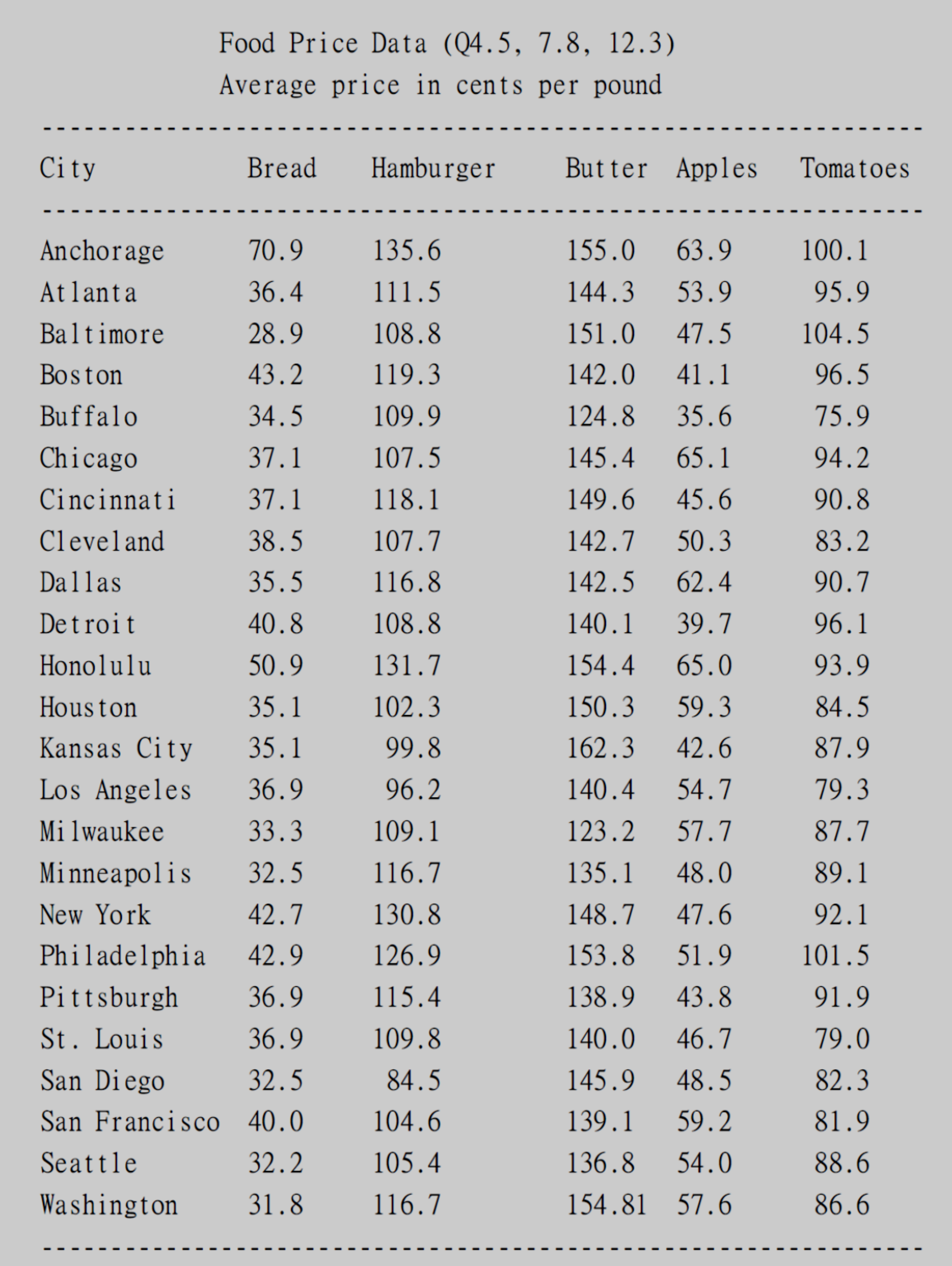
b) Identify the most and least expensive cities (based on the above price index measures). Do the most and least expensive cities change when standardized data are used to define price index measures? Why?

c) Plot the data using principal components scores and identify distinct groups of cities. How are these groups different from each other?

P.S. You can use R, STATISTICA, SAS, Excel to run your team report.

Due date 3/27, paper and file are need.

Homework



Food Price Data (Q4.5, 7.8, 12.3)
Average price in cents per pound

City	Bread	Hamburger	Butter	Apples	Tomatoes
Anchorage	70.9	135.6	155.0	63.9	100.1
Atlanta	36.4	111.5	144.3	53.9	95.9
Baltimore	28.9	108.8	151.0	47.5	104.5
Boston	43.2	119.3	142.0	41.1	96.5
Buffalo	34.5	109.9	124.8	35.6	75.9
Chicago	37.1	107.5	145.4	65.1	94.2
Cincinnati	37.1	118.1	149.6	45.6	90.8
Cleveland	38.5	107.7	142.7	50.3	83.2
Dallas	35.5	116.8	142.5	62.4	90.7
Detroit	40.8	108.8	140.1	39.7	96.1
Honolulu	50.9	131.7	154.4	65.0	93.9
Houston	35.1	102.3	150.3	59.3	84.5
Kansas City	35.1	99.8	162.3	42.6	87.9
Los Angeles	36.9	96.2	140.4	54.7	79.3
Milwaukee	33.3	109.1	123.2	57.7	87.7
Minneapolis	32.5	116.7	135.1	48.0	89.1
New York	42.7	130.8	148.7	47.6	92.1
Philadelphia	42.9	126.9	153.8	51.9	101.5
Pittsburgh	36.9	115.4	138.9	43.8	91.9
St. Louis	36.9	109.8	140.0	46.7	79.0
San Diego	32.5	84.5	145.9	48.5	82.3
San Francisco	40.0	104.6	139.1	59.2	81.9
Seattle	32.2	105.4	136.8	54.0	88.6
Washington	31.8	116.7	154.81	57.6	86.6