A short presentation of PCA

by
Sergio Rojas
http://prof.usb.ve/srojas/

Abstract

In this presentation we'll be talking about the use of the statistical methodology known as Principal Component Analysis (PCA) to analyze data via an illustrative example related to the study of food price data in some US cities.

The end goal is to gain understanding, intuition and experience to explore the possibility of applying this technique to analyze the interactions among the big data sets, like the returns of stocks in the financial market.

It is expected that this sort of statistical analysis may lead to the discovery of nonlinear and non-Gaussian effects in the interactions between the return changes across the stocks. More importantly, it is also expected that profitable investment strategies and/or risk controlling strategies could be devised from this sort of analysis.

Outline

1 Introductory Remarks

- Data available for financial modeling (i.e. stock prices, food prices, financial ratios, etc.) is overwhelming huge.
- As a first approximation, the need for uncovering interesting patterns from such large data sets and the lack of appropriated theories explaining the observed behavior leads naturally to the application of statistical methods in the hope of turning such collection of data into useful information.
- The information and knowledge gained could result in: devising profitable investment/trading strategies; characterization and management of the risk involved in financial operations; defining or selecting appropriated financial measurements that best describes the available information, etc.

2 Statistical techniques in financial modeling

- **Multi-scale Decomposition (MSD)**: long term memory and fractional integration effects, the existence of trends and mean reverting behaviors, nonlinear effects, and the presence of hierarchical effects in the term structure of volatility.
- **Principal Component Analysis (PCA)**: Uncorrelate a set of variables. Commonly used to reduce the *dimensionality* of a set of variables.
- **Independent Component Analysis (ICA)**: reduce statistical dependency among variables.

2.1 Data Preprocessing

- ***Centering* or *mean-correcting***: Subtracting the mean of a set of observations from each observation. Most statistics and analysis performed on the data are **not affected** by mean correcting the data.
- ***Standardizing***: After centering a set of observations, transform them such that the new transformed data set have variance equal to unity. Some statistics and analysis performed on the data are **affected** by its standardization.

2.2 Representing data in matrix form

- Consider we have ***m*** variables (i.e. ***m*** stocks) each one containing ***n*** *centered* or *mean-corrected* observations (i.e. daily returns for ***n*** days) . This set of data could be represented in the following way:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_m} \end{pmatrix} \tag{1}$$

$$\mathbf{x_i} = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{in} \end{pmatrix} \quad i = 1, \cdots, m \tag{2}$$

$$\mathbf{X}\mathbf{X}^{\mathrm{T}} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_m \end{pmatrix} \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} & \mathbf{x}_2^{\mathrm{T}} & \cdots & \mathbf{x}_m^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1\mathbf{x}_1^{\mathrm{T}} & \mathbf{x}_1\mathbf{x}_2^{\mathrm{T}} & \cdots & \mathbf{x}_1\mathbf{x}_m^{\mathrm{T}} \\ \mathbf{x}_2\mathbf{x}_1^{\mathrm{T}} & \mathbf{x}_2\mathbf{x}_2^{\mathrm{T}} & \cdots & \mathbf{x}_2\mathbf{x}_m^{\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_m\mathbf{x}_1^{\mathrm{T}} & \mathbf{x}_m\mathbf{x}_2^{\mathrm{T}} & \cdots & \mathbf{x}_m\mathbf{x}_m^{\mathrm{T}} \end{pmatrix} \quad (3)$$

- $\mathbf{C}_{\mathbf{x}} = \mathbf{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}})$ is the **covariance** matrix among variables $\mathbf{x}_i$.

3 Principal Component Analysis (PCA)

- **PCA** : Statistical procedure to obtain a set of **uncorrelated** variables ($\mathbf{y_i}$) by linear combination of given (known) correlated variables ($\mathbf{x_i}$):

$$
\begin{aligned}
\mathbf{y_1} &= v_{11}\mathbf{x_1} + v_{12}\mathbf{x_2} + \cdots + v_{1m}\mathbf{x_m} \\
&= \textstyle\sum_{j=1}^{m} v_{1j}\mathbf{x_j} = \mathbf{v_1}\mathbf{X} \\
\mathbf{y_2} &= v_{21}\mathbf{x_1} + v_{22}\mathbf{x_2} + \cdots + v_{2m}\mathbf{x_m} \\
&= \textstyle\sum_{j=1}^{m} v_{2j}\mathbf{x_j} = \mathbf{v_2}\mathbf{X} \\
\vdots \quad & \quad \vdots \quad\quad \vdots \quad\quad\quad \vdots \\
\mathbf{y_m} &= v_{m1}\mathbf{x_1} + v_{m2}\mathbf{x_m} + \cdots + v_{mm}\mathbf{x_m} \\
&= \textstyle\sum_{j=1}^{m} v_{mj}\mathbf{x_j} = \mathbf{v_m}\mathbf{X}
\end{aligned} \tag{4}
$$

$$
\mathbf{v_i} = \begin{pmatrix} v_{i1} & v_{i2} & \cdots & v_{im} \end{pmatrix} \quad i = 1, \cdots, m \tag{5}
$$

subject to the conditions:

- The variances of the new variables are maximized.
- The components of the mixing matrix are subject to:

$$\|\mathbf{v_i}\|^2 = \sum_{j=1}^{m} v_{ij}^2 = 1 \quad i = 1, \ldots, m \tag{6}$$

$$\mathbf{v_i} \cdot \mathbf{v_j} = \sum_{k=1}^{m} v_{ik} v_{jk} = 0 \quad \text{for all } i \neq j \tag{7}$$

- $\mathbf{C_y} = \mathbf{E}(\mathbf{y_i y_i^T}) = \mathbf{E}(\mathbf{v_i X X^T v_i^T}) = \mathbf{v_i E(X X^T) v_i^T} = \mathbf{v_i C_x v_i^T}$, is the variance for each new variable $\mathbf{y_i}$.

- New problem is to find $\mathbf{v_i}$ such that new variance $\mathbf{v_i C_x v_i^T}$ is maximum over all possible linear combinations that can be formed subject to $\|\mathbf{v_i}\|^2 = \mathbf{v_i v_i^T} = 1$.
- Using *Lagrangian Multipliers* technique, the solution can be found in the following way:
  - $F = \mathbf{v_i C_x v_i^T} - \lambda(\mathbf{v_i v_i^T} - \mathbf{1})$
  - $\frac{\partial F}{\partial \mathbf{v_i}} = 2\mathbf{v_i C_x} - \mathbf{2}\lambda\mathbf{v_i} = \mathbf{0}$
  - $|\mathbf{C_x} - \lambda\mathbf{I}| = 0$
- That $\lambda_i = \mathbf{v_i C_x v_i^T}$ is obtained from the following:
  - $\mathbf{v_i C_x} - \lambda_\mathbf{i}\mathbf{v_i} = \mathbf{0}$
  - $(\mathbf{v_i C_x} - \lambda_\mathbf{i}\mathbf{v_i})\mathbf{v_i^T} = \mathbf{0}$
  - $\mathbf{v_i C_x v_i^T} = \lambda_\mathbf{i}$ because $\mathbf{v_i v_i^T} = \mathbf{1}$

4 PCA Illustrative Example: Food Prices [5]

4.1 The data

- The data for this purpose comprise the prices of five food products (bread, burger, milk, oranges, and tomatoes) in several cities of the US as in March of 1973 (Fig.- 1, page 14).
- PCA is performed on mean corrected data (table on page 13; Fig.- 2, page 15).
- Basic descriptive statistics (i.e. correlations and variances) of the mean corrected data is shown in Fig.- 3, page 17 and Fig.- 4, page 18 respectively.

- Mean corrected data set.

	Bread	Burger	Milk	Oranges	Tomatoes
ATLANTA	-0.79130	2.64348	11.60435	-22.89130	-7.16522
BALTIMORE	1.20870	-0.85652	5.20435	-28.39130	4.53478
BOSTON	4.40870	8.94348	-0.89565	1.00870	10.83478
BUFFALO	-2.49130	-5.25652	3.00435	15.40870	2.43478
CHICAGO	1.40870	-5.15652	0.40435	2.90870	2.43478
CINCINNATI	0.00870	10.64348	1.00435	-3.69130	-3.16522
CLEVELAND	-2.49130	-3.05652	-9.89565	7.90870	-1.96522
DALLAS	-1.99130	-6.35652	0.20435	14.90870	-6.96522
DETROIT	-1.19130	1.84348	-10.79565	6.70870	3.63478
HONALULU	4.00870	14.04348	17.90435	30.20870	12.93478
HOUSTON	-2.99130	-8.25652	5.50435	5.60870	-6.36522
KANSAS CITY	0.80870	-2.95652	3.10435	-2.09130	-5.56522
LOS ANGELES	1.60870	-2.55652	-6.09565	-20.29130	-10.36522
MILWAUKEE	-4.99130	-2.25652	-8.49565	8.80870	5.13478
MINNEAPOLIS	-0.69130	0.34348	-10.39565	3.00870	1.93478
NEW YORK	5.50870	18.84348	3.70435	4.30870	13.83478
PHILADELPHIA	-0.79130	0.44348	4.40435	-4.99130	12.93478
PITTSBURGH	0.90870	3.54348	-2.09565	14.10870	0.53478
ST LOUIS	1.20870	0.54348	-1.49565	12.10870	-2.56522
SAN DIEGO	0.20870	-8.15652	-5.29565	-10.19130	-13.36522
SAN FRANCISCO	1.00870	-4.75652	-3.99565	-1.19130	-7.26522
SEATTLE	-2.79130	-14.15652	-0.29565	-11.89130	-3.86522
WASHINGTON DC	-1.09130	1.94348	3.70435	-21.39130	-2.56522

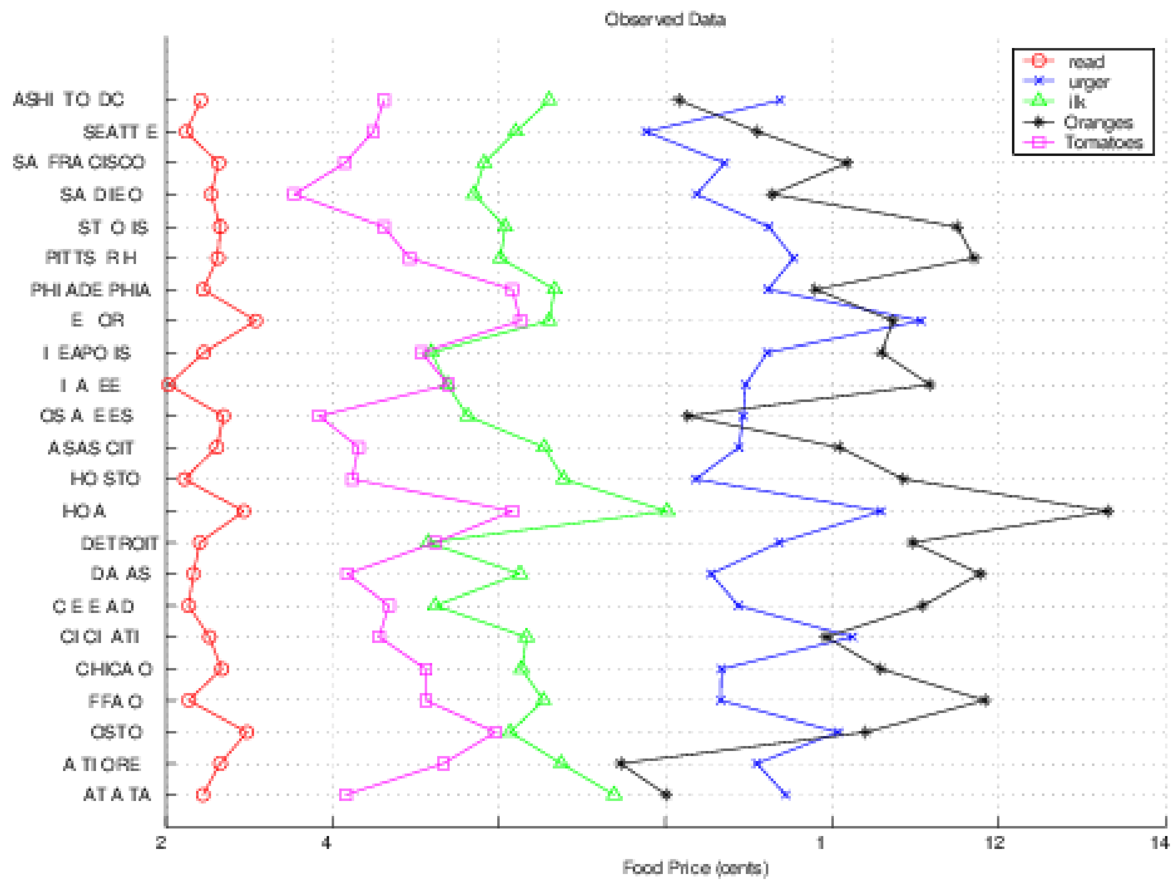
Observed Data
read
urger
ilk
Oranges
Tomatoes
ASHI TO DC
SEATT E
SA FRA CISCO
SA DIE O
ST O IS
PITTS R H
PHI ADE PHIA
E OR
I EAPO IS
I A EE
OS A E ES
A SAS CIT
HO STO
HO A
DETROIT
DA AS
C E E A D
CI CI ATI
CHICA O
FFA O
OSTO
A TI ORE
AT A TA
2
4
1
12
14
Food Price (cents)

Figure 1:

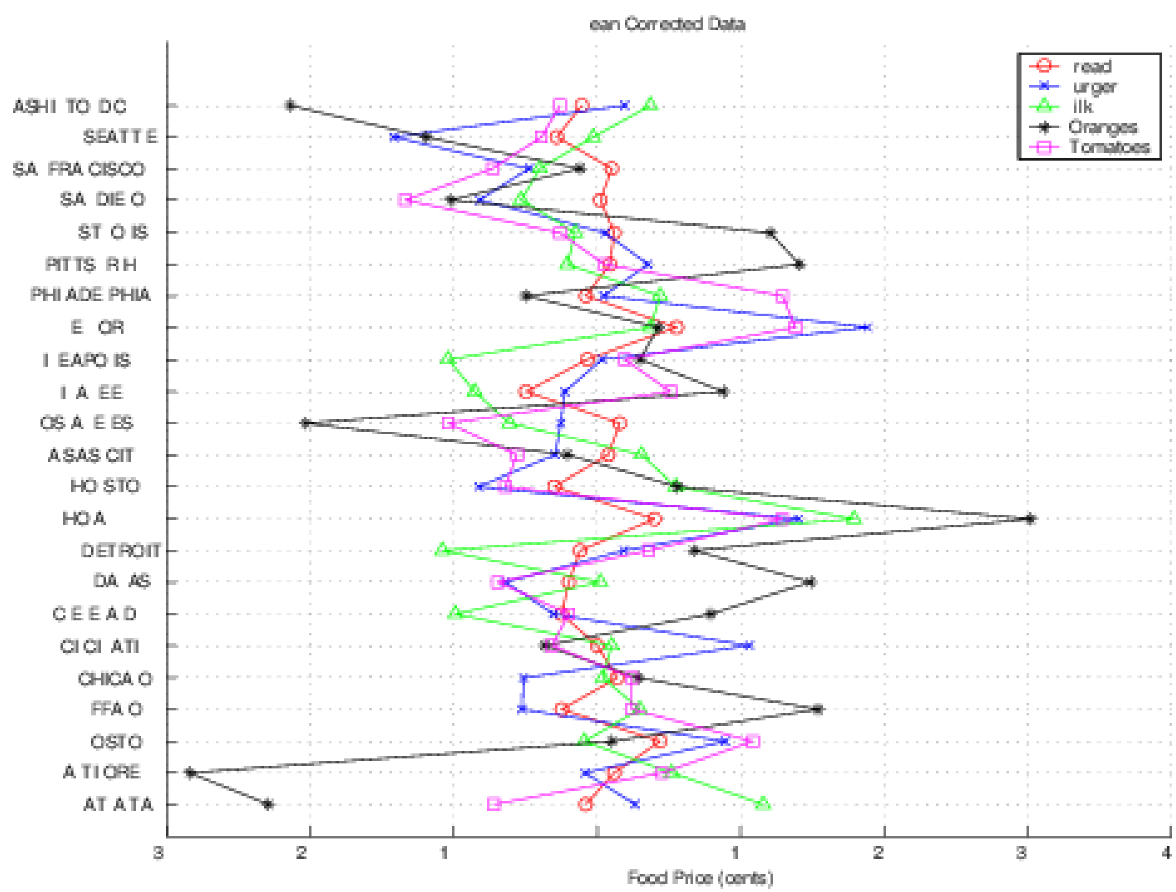
ean Corrected Data
read
urger
ilk
Oranges
Tomatoes
ASHI TO DC
SEATT E
SA FRA CISCO
SA DIE O
ST O IS
PITTS R H
PHI ADE PHIA
E OR
I EAPO IS
I A EE
OS A E ES
A SAS CIT
HO STO
HO A
DETROIT
DA AS
C E E A D
CI CI ATI
CHICA O
FFA O
OSTO
A TI ORE
AT A TA
3
2
1
1
2
3
4
Food Price (cents)

Figure 2:

4.2 **PCA** Results

- The following table show the weights (eigenvector components) to find each PC in the form:
$PC_i = c_{i1} Bread + c_{i2} Burguer + c_{i3} Milk + c_{i4} Oranges + c_{i5} Tomatoes$

	Bread	Burger	Milk	Oranges	Tomatoes
PC_1	0.02849	0.20012	0.04167	0.93886	0.27558
PC_2	-0.16532	-0.63218	-0.44215	0.31435	-0.52792
PC_3	0.02136	0.25420	-0.88875	-0.12135	0.36100
PC_4	-0.18973	-0.65862	0.10766	-0.06905	0.71684
PC_5	0.96716	-0.24877	-0.03606	0.01521	0.03429

- Correlations before and after **PCA** are shown in Fig.- 3, page 17.
- Variances before and after **PCA** are shown in Fig.- 4, page 18 and Fig.- 5, page 19 respectively.

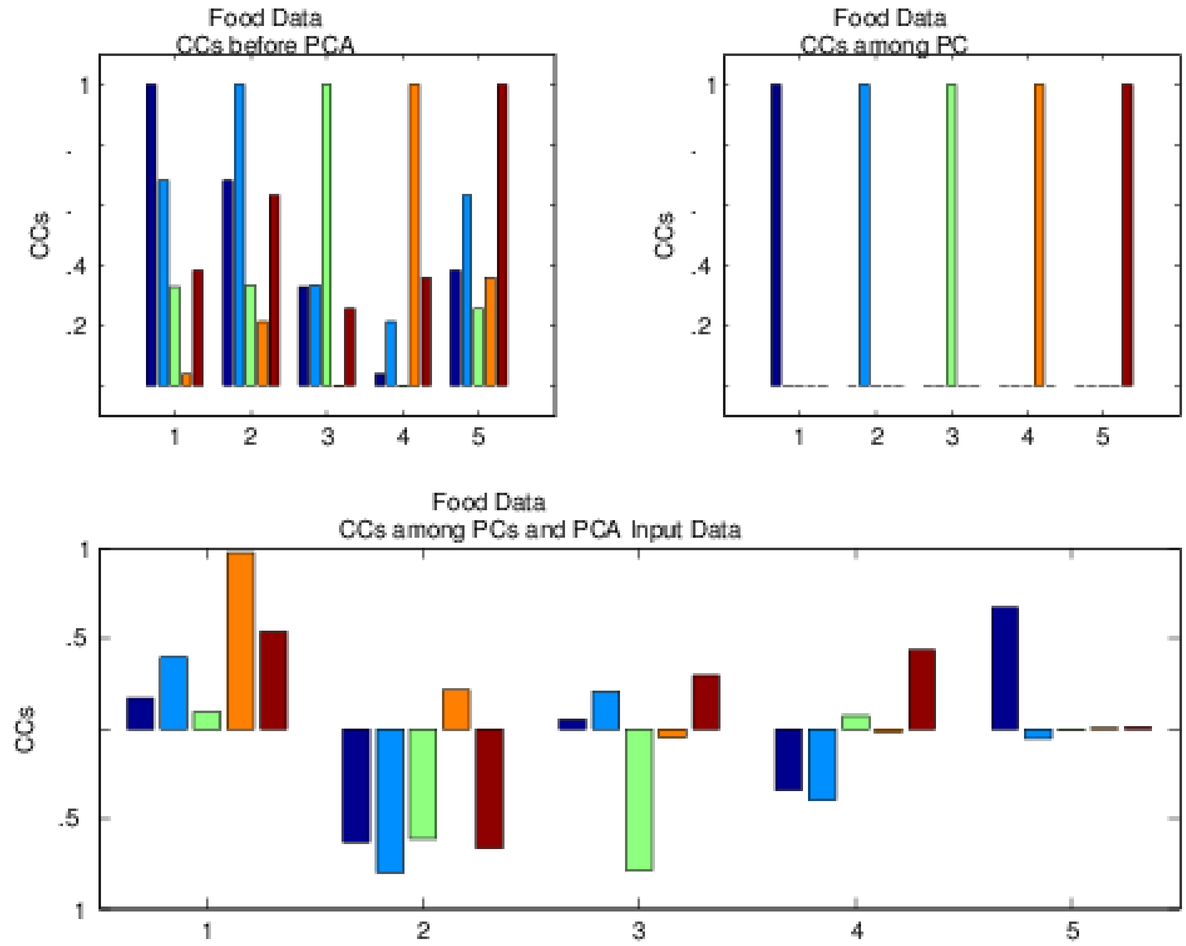
Food Data
CCs before PCA
CCs
1
.4
.2
1 2 3 4 5
Food Data
CCs among PC
CCs
1
.4
.2
1 2 3 4 5
Food Data
CCs among PCs and PCA Input Data
CCs
1
.5
.5
1
1 2 3 4 5

Figure 3:

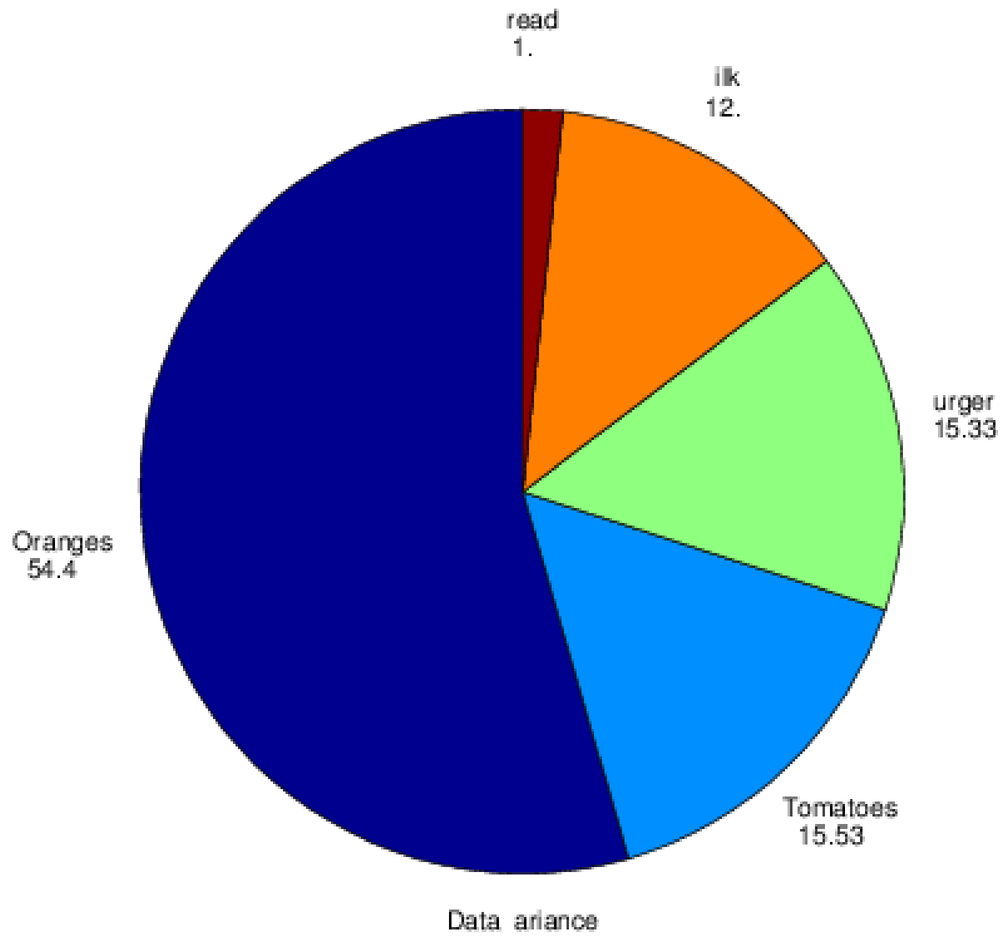
read
1.
ilk
12.
urger
15.33
Tomatoes
15.53
Oranges
54.4
Data ariance

Figure 4:

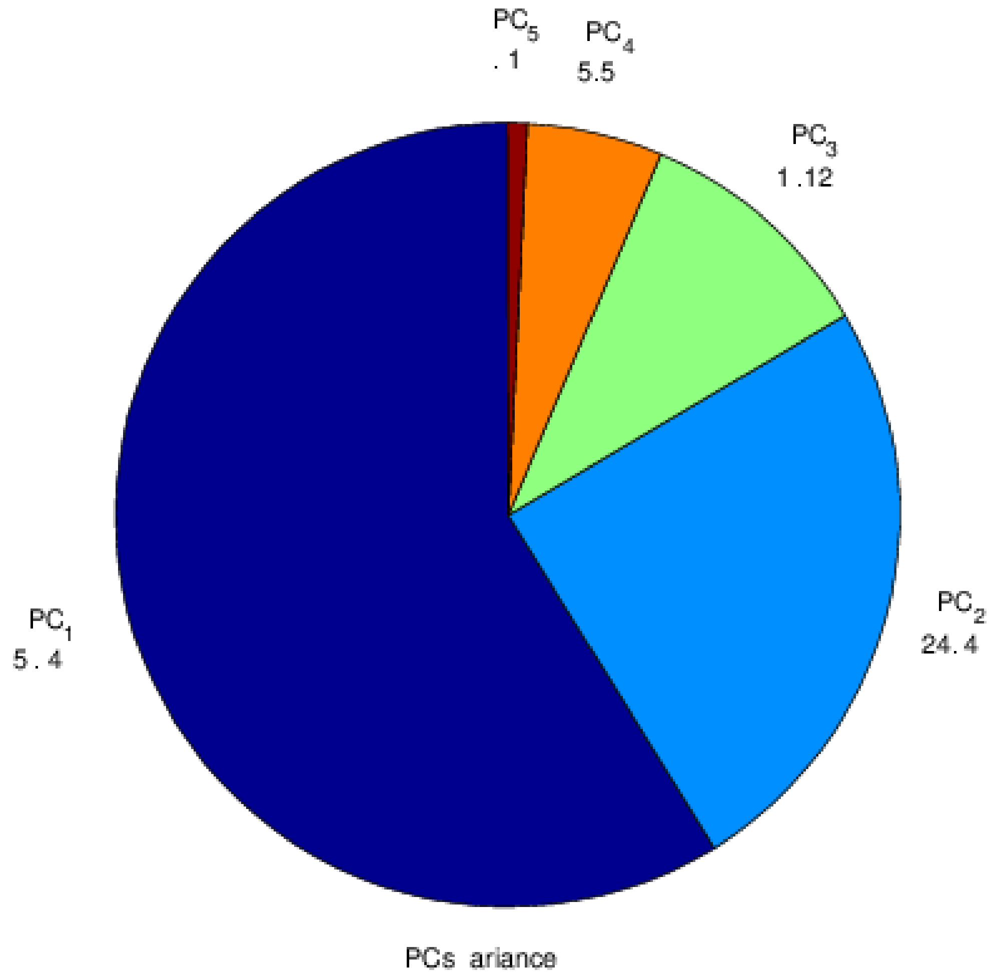

Figure 5:

- PC values for each city.

	PC_1	PC_2	PC_3	PC_4	PC_5
ATLANTA	-22.47627	-10.08457	-9.46707	-3.89736	-2.43537
BALTIMORE	-25.32582	-13.27837	0.26507	6.10615	0.91798
BOSTON	5.81098	-11.38954	6.95261	0.87390	2.45825
BUFFALO	14.13986	5.96502	-5.05044	4.93961	-0.89225
CHICAGO	2.42689	2.47721	-1.11410	4.71699	2.75840
CINCINNATI	-2.16580	-6.66357	1.11849	-8.91766	-2.84029
CLEVELAND	5.78854	10.24312	6.29540	-0.53441	-1.23935
DALLAS	10.75737	12.62102	-6.16363	-1.43599	-0.36401
DETROIT	7.18531	3.99488	10.53587	-0.00805	-0.99478
HONALULU	35.59706	-14.78944	-11.25329	-0.89601	0.64095
HOUSTON	2.00347	8.40385	-10.03319	1.64796	-1.17054
KANSAS CITY	-3.93639	2.64334	-5.24855	-1.71695	1.18303
LOS ANGELES	-22.62697	3.13873	3.52247	-5.30683	1.74753
MILWAUKEE	8.73738	6.06638	7.65502	4.59115	-3.64960
MINNEAPOLIS	2.97377	4.41798	9.64503	-0.03506	-0.26705
NEW YORK	11.94021	-20.41029	6.08704	-3.43729	1.04650
PHILADELPHIA	-0.87177	-10.49444	1.45665	9.94902	-0.66684
PITTSBURGH	14.04114	2.68905	1.26365	-3.32265	0.30590
ST LOUIS	10.74230	5.27855	-0.90221	-3.42321	1.18399
SAN DIEGO	-15.09848	11.31543	-0.95061	-4.11468	1.80854
SAN FRANCISCO	-4.21030	8.06785	-0.11465	-2.61453	2.03568
SEATTLE	-15.15433	7.84415	-3.34785	7.87190	0.51929
WASHINGTON DC	-20.27814	-8.05634	-1.15172	-1.03601	-2.08594

5 How one could use these PCA results?

- These results could be use to quantify how expensive or cheap are the analyzed city's food items. This is done by looking at the values of the PC for each city. For instance based on the values for PC_1, **Honolulu** is the most expensive city, while **Baltimore** is the less expensive one (see values on page 20).
- One can use a few PC to represent the initial data without a *substantial loss of information*, whatever it means.
- If the idea is to have a set of orthogonal uncorrelated variables, then **PCA** is the way to go.
- keep in mind that **PCA** is affected by the variability of the data. If such variability is not important, one could standardize the initial data before applying **PCA** to it.

References

[1] Jolliffe, I. T. (2010), Principal component analysis, *2nd edition*, Springer.

[2] Hyvärinen, A. and Oja, E. (2000), Independent component analysis: algorithms and applications, *Neural Networks*, **13**, 411-430.

[3] Rojas, S. and Moody, J. (2001), Cross-sectional analysis of the returns of iShares MSCI Index Funds using Independent Component Analysis, *OGI CSE610 internal report*, Oregon Graduate Institute of Science and Technology.

[4] Rojas, S., Candanoza Santos, C. and Guevara Jordán, J. M. (2008), Procesamiento de señales vía análisis de componentes independientes, en "Desarrollo y Avances en Métodos Numéricos para Ingeniería y Ciencias Aplicadas", editado por L. Martino, V. Carrera, G. Larrazábal y M. Cerrolaza (2008), p. PS-21

[5] Sharma, S. (1996), Applied Multivariate Techniques, Wiley.