Principal Components Analysis

Ying-Chao Hung
Department of Statistics
National Chengchi University
Email: hungy@nccu.edu.tw

Motivation

- Choose one or more linear combinations of the original variables (features, attributes) in the data set, while retaining as much as possible of the variation present in the data.

- Reveal interesting structure in low-dimensional plots

  ➔ For visualization purpose

Setup

- Data matrix:

$$X_{N\times m} = \begin{pmatrix} X_1 & X_2 & \cdots & X_m \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{matrix} 1 \\ \vdots \\ N \end{matrix}$$

- Variables: $X_1, X_2, \ldots, X_m$ (column vectors)
- Usually, variables are standardized so that

$$\text{Mean}(X_i) = 0 \quad \text{and} \quad \text{Var}(X_i) = 1 \quad \text{for all } i.$$

➔Remove the effect of units of variables.

Covariance/Correlation Matrix

- Correlation matrix:

  Since X_i are standardized,

$$R_{m\times m} = \frac{1}{N-1} X'X = \begin{pmatrix} 1 & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & 1 \end{pmatrix}$$

  (Note that $\frac{1}{N-1}$ is missing in the lecture note)

- Remark: $r_{ij} = \dfrac{\operatorname{cov}(X_i, X_j)}{\sqrt{\operatorname{var}(X_i)\operatorname{var}(X_j)}}$.

Approach

- Consider a set of variables Y_j such that

$$Y_j = \sum_{i=1}^{m} \beta_{ji} X_i = \beta_{j1} X_1 + \beta_{j2} X_2 + \cdots + \beta_{jm} X_m, \quad j = 1, \ldots, m.$$

In matrix form:

$$Y_{N\times m} = \begin{pmatrix} Y_1 & \cdots & Y_m \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{pmatrix} = X_{N\times m} B_{m\times m}$$

Question: What kind of linear combinations will capture most of the variation in the original data ?

Approach (continued)

That is, how to choose B to maximize the variance of Y ?

- Note that:

$$\max_B \mathrm{Var}(Y_j) = \max_B \mathrm{Var}\left(\sum_{i=1}^{m} X_i \beta_{ji}\right)$$

$$\text{(in matrix form)} = \max_B B'\left(\tfrac{1}{N-1} X'X\right)B = \max_B B'RB.$$

- To avoid a trivial solution of B (we can make the elements of B arbitrarily large), we request that

$$B'B = I_m$$

That is, B is an *orthonormal* matrix.

Rayleigh-Ritz Theorem

- The optimal solutions B are determined by the eigen-decomposition of R:

$$R = B\Lambda B'$$

$$= \underbrace{\begin{pmatrix} \beta_1 & \cdots & \beta_j & \cdots & \beta_m \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{pmatrix}}_{\text{Eigenvectors of } R} \underbrace{\begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & . & 0 \\ . & \ddots & \lambda_j & \ddots & 0 \\ . & . & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \lambda_m \end{pmatrix}}_{\text{Eigenvalues of } R} B'$$

Principal Components

$$Y = \begin{pmatrix} Y_1 & \cdots & Y_m \\ & & \end{pmatrix} = XB = \begin{pmatrix} X_1 & \cdots & X_m \\ & & \end{pmatrix} \begin{pmatrix} \beta_1 & \cdots & \beta_m \\ & & \end{pmatrix}$$

➔ $Y_1, \ldots, Y_m$ are called principal components.

Note: There are *m* principal components.

Properties of PCs

- $\text{Mean}(Y_j) = \text{Mean}\left(\sum_{i=1}^{m} X_i \beta_{ji}\right) = \sum_{i=1}^{m} \text{Mean}(X_i)\beta_{ji} = 0$ for all j.

- All PCs are uncorrelated (linearly independent, orthogonal):

$$\text{Cov}(Y) = \tfrac{1}{N-1} Y'Y = \tfrac{1}{N-1}(XB)'(XB) = B'(\tfrac{1}{N-1} X'X)B = B'RB$$

$$= B'(B\Lambda B')B = \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_m \end{pmatrix}.$$

➔ $\begin{cases} \text{Cov}(Y_i, Y_j) = 0 \text{ for all } i \neq j \\ \text{Var}(Y_j) = \lambda_j \end{cases}$

Properties of PCs (continued)

- Correlation of X with PCs Y:

$$\mathrm{Cor}(X_i, Y_j) = \frac{\mathrm{Cov}(X_i, Y_j)}{\sqrt{\mathrm{Var}(X_i)}\sqrt{\mathrm{Var}(Y_j)}} = \frac{\beta_{ji}\lambda_j}{1 \cdot \sqrt{\lambda_j}} = \beta_{ji}\sqrt{\lambda_j}.$$

Factor loadings

- Proportion of variance explained by PCs:

$$\mathrm{trace}(R) = m = \mathrm{trace}(B\Lambda B') = \mathrm{trace}(BB'\Lambda) = \mathrm{trace}(\Lambda)$$

➔ $$m = \sum_{j=1}^{m} \lambda_j = \sum_{j=1}^{m} \mathrm{Var}(Y_j)$$

➔ Each PC explains (λ_j / m) proportion of the total variance.

Connection to SVD of X

Data Matrix:

$$X_{N\times m} = U_{N\times m}\Sigma_{m\times m}V^{'}_{m\times m} \quad \text{where} \begin{cases} U,\ V : \text{orthonormal} \\ \Sigma = \begin{pmatrix} \lambda_1^* & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_{\mathrm{m}}^* \end{pmatrix} \end{cases}$$

➔ $$X'X = (U\Sigma V')'(U\Sigma V') = V\Sigma' U' U\Sigma V' = V\Sigma^2 V'$$
$$= (N-1)R = (N-1)B\Lambda B'$$

➔ $$\begin{cases} V = B \\ (\lambda_i^*)^2 = (N-1)\lambda_i \end{cases}$$

Using Covariance Matrix

If we decide not to standardize the data, can use the covariance matrix instead of correlation matrix.

Consider

$$\Sigma = \frac{1}{N-1} X'X, \text{ Mean}(X_i) = 0, \text{ but Var}(X_i) \neq 1.$$

➔ $\Sigma = \tilde{B}\tilde{\Lambda}\tilde{B}'$ and PCs $\tilde{Y} = X\tilde{B}$

➔ No trivial relationship between B and $\tilde{B}$ (or Y and $\tilde{Y}$).

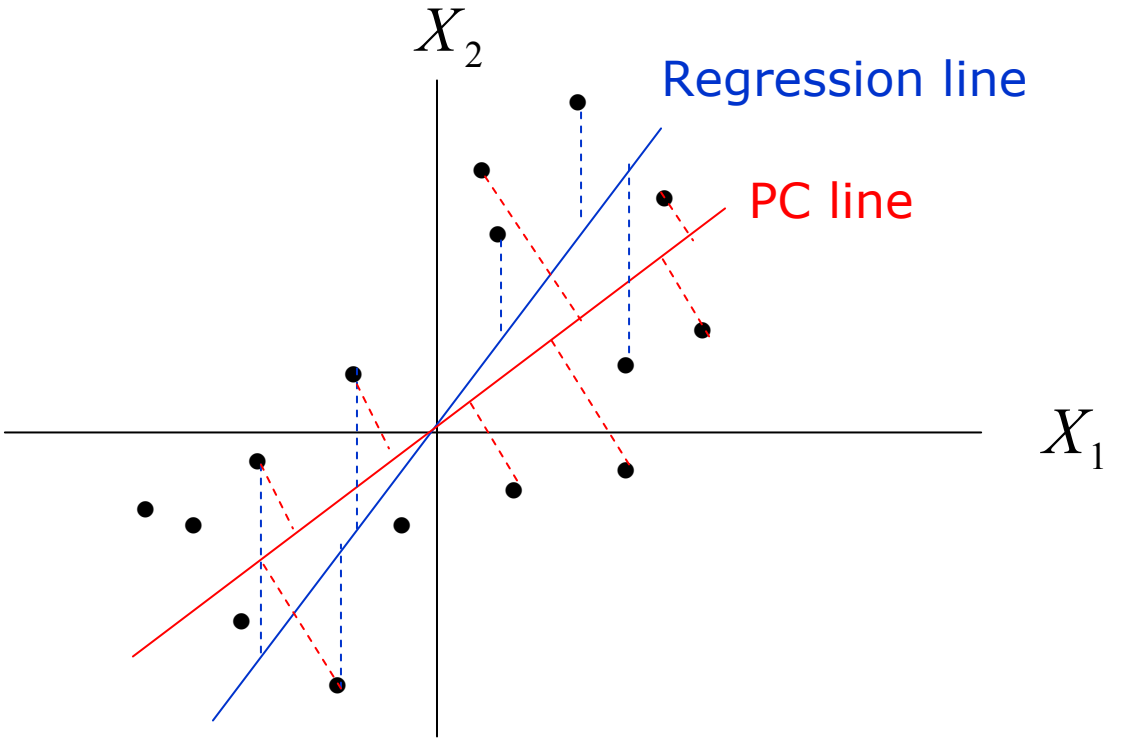
Proposition

- The first k PCs minimize the sum of squared distances from points to their projections onto any subspace of dimension at most k.

  ➔ That is, the first k PCs have the best rank-k reconstruction of original data.

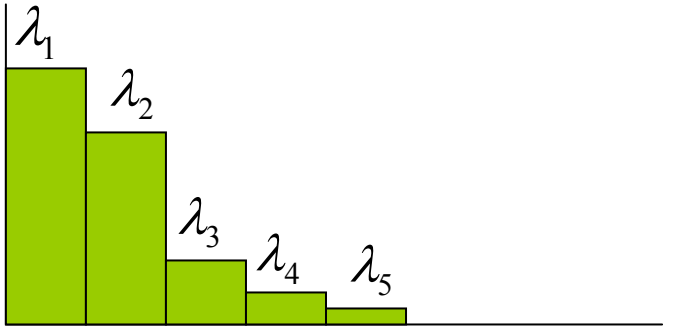
Note: The result comes from the Schmidt-Beltrami-Eckart-Young Theorem.

Geometry of PC Lines



How Many PCs to Use ?

- Scree Plot (Cattell, 1966)



- Cutoff point $\lambda_i < 1$ (Kaiser, 1960)
- Cutoff point $\lambda_i < 0.7$ (Jolliffe, 1971)
- $\left(\sum \lambda_i / m \right) > ?$ (explain 70 ~ 90% of total variance)

Variable Selection

- Note that # of PCs = # of variables

  ➔ Can discard some variables by keeping only those with high factor loadings.

- A quick examination of the correlation matrix R might help us find "colinearity" between variables.

  For example, if $r_{ij} = \mathrm{Cor}(X_i, X_j) = 0.97$, can exclude X_i or X_j from analysis.

Outlier Detection

- PCA is based on variances, so it is sensitive to outliers.
- How to identify outliers? (in high-dimensional space)
  ➔ Data visualization system (e.g. Ggobi) might help.
- How many outliers to exclude?
  ➔ Don't want to lose too much information of the raw data !
- Can use more "robust" measure to compute *R*.
  (e.g. using $l_1 - \text{norm}$ instead of $l_2 - \text{norm}$)

Biplots (by Gabriel, 1971)

- Provide a picture of both data points and variables in a common space.

- Consider the singular value decomposition (SVD) of data matrix:

$$X = U\Sigma V' = \left(U\Sigma^{\alpha}\right)\left(\Sigma^{1-\alpha}V'\right) = GH'$$

➔ $X(i,j) = G(i,:)\ H(:,j)$

row information (obejct infom.)

column information (variable inform.)

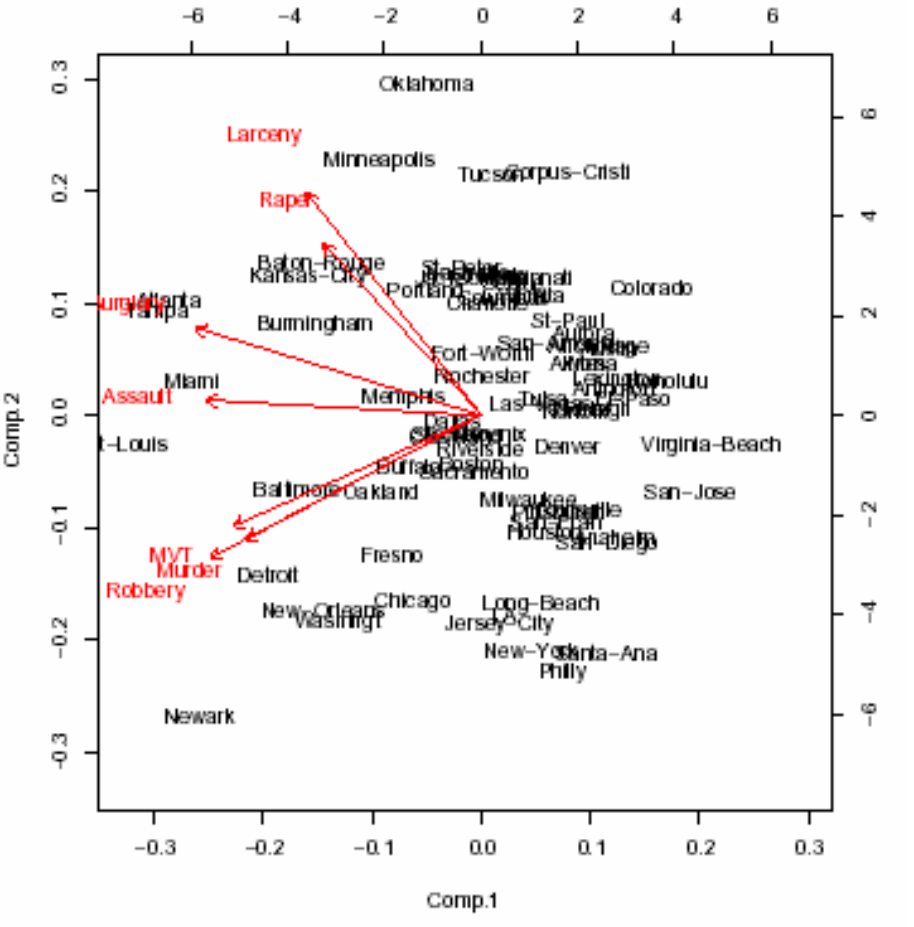
- Set $\alpha = 1$ ➔ $G = U\Sigma = Y =$ PC scores (why?)

A 2-D Biplot

- Consider $k = 2$,

$$\underset{N\times m}{X} = \underset{N\times m}{G}\ \underset{m\times m}{H'}$$

$$\begin{array}{c} \\ 1 \\ \vdots \\ N \end{array} \begin{pmatrix} Y_1 & Y_2 \\ \cdot & \cdot \\ \vdots & \vdots \\ \cdot & \cdot \end{pmatrix} \qquad \begin{array}{c} \\ 1 \\ \vdots \\ m \end{array} \begin{pmatrix} Z_1 & Z_2 \\ \cdot & \cdot \\ \vdots & \vdots \\ \cdot & \cdot \end{pmatrix}$$



Interpret the Result

- Is your answer trustworthy?
  (outlier detection, variable selection, select the # of PCs, etc)
- How do you interpret the PCs?
  (from the factor loadings)
- How do you interpret the relationships between objects, variables, and both.
  (the biplot provides a good visual)

Some Remarks

- PCA is a projection method
- PCA utilizes information contained in the 2nd moment (i.e. variance)
- Non-linear structure might be missed
- Linear combinations may not be meaningful if the variables do not have comparable quantities (e.g. 0.9Salary + 0.3Age = ?)
- PCA is most useful when applied to correlated variables representing one or more common domains.
- Some large sample properties

  (e.g. how to estimate λ_i and B when N and m are large?)

Large Sample Properties of PCs

- Assume dataset X is a large sample (huge # of obs.) from a normal distribution with correlation matrix R, where

$$R = B\Lambda B'$$

Q1: How do we estimate Λ ?

Q2: How accurate is the estimated proportion (total variance explained):

$$\hat{\theta} = (\hat{\lambda}_1 + \hat{\lambda}_2) \Big/ \left(\sum_{i=1}^{m} \hat{\lambda}_i \right)$$

Bootstrap Sampling and Inference

- Assume all eigenvalues of R are distinct and positive.

  (remember all covariance/correlation matrices are symmetric and positive semidefinite.)

- Denote the bootstrap data of size N by $X^*_{N\times m}$ and the corresponding eigenvalues and eigenvectors by $\hat{\Lambda}$, $\hat{\beta}$.

  Then we have:

$$\begin{cases} \sqrt{N}(\hat{\Lambda}-\Lambda) \to N(0, 2\Lambda^2) \\ \sqrt{N}(\hat{\beta}_j-\beta_j) \to N(0, V_j) \\ \qquad\qquad \text{where } V_j = \lambda_j \sum_{k\neq j} \frac{\lambda_k}{(\lambda_k-\lambda_j)^2} \beta_k \beta_k^{'} \\ \text{Distribution of } \hat{\lambda}_j \text{ is indep. of } \hat{\beta}_j \end{cases}$$

Simulation Result

- For each generated X^*, we can compute R^*, and then Λ^*.
- Plot the distribution (histogram) of λ_i^* and

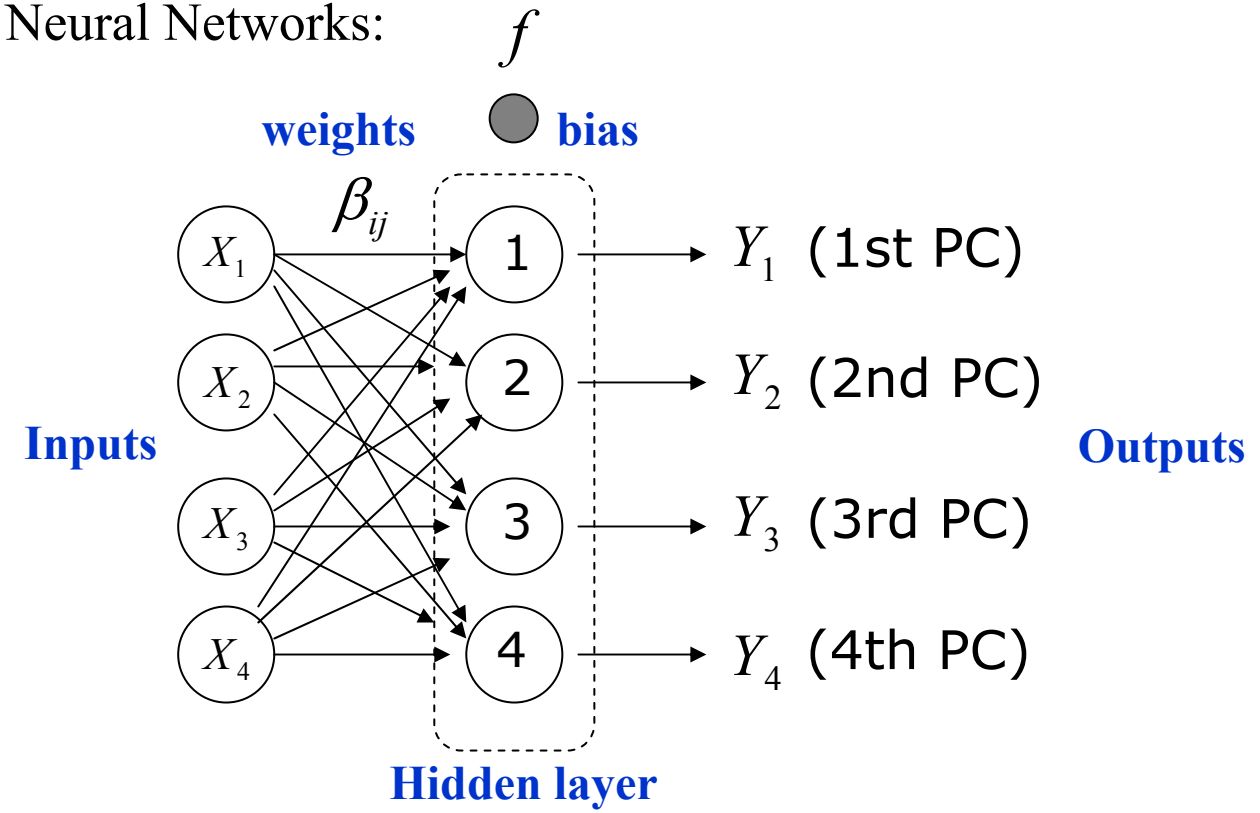
$$\theta^* = (\lambda_1^* + \lambda_2^*) \Big/ \left(\sum_{i=1}^{m} \lambda_i^* \right)$$

with 10000 replications (see examples in the lecture note).

- Can construct confidence intervals for the desired quantity.

Connect to Neural Networks

- Feed-forward Neural Networks:



➔ $Y_j = f_j(\alpha_j + \sum_{i=1}^{4} \beta_{ij} X_j) = f_j(\sum_{i=1}^{4} \beta_{ij} X_j)$, bias $= 0$ and f is linear.

7. **Example: Crime data**

In this example we analyze using PCA the crime data set. The data give crime rates per 100,000 people for the 72 largest US cities in 1994.

The variables are:

1) Murder
2) Rape
3) Robbery
4) Assault
5) Burglary
6) Larceny
7) Motor Vehicle Thefts

A scatterplot matrix of the data is given in Figure 1.

The first 3 PCs account for 54%, 17% and 11% of total variance respectively, and in total for 82%. So, it suffices to look at a 2-dim or a 3-dim representation of the results (the *screeplot* of the eigenvalues/variances of the PCs is given in Figure 2).

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Standard deviation	1.948	1.095	0.877	0.714
Proportion of Variance	0.54	0.17	0.11	0.073
Cumulative Proportion	0.54	0.71	0.82	0.89

The optimal weights (the matrix B) of the crime variables on the first 3 components are given next.

	PC1	PC2	PC3
Murder	0.370	-.339	0.202
Rape	0.249	0.466	0.782
Robbery	0.426	-.387	0.079
Assault	0.434	0.042	-.282
Burglary	0.449	0.238	0.015
Larceny	0.276	0.605	-.492
MVT	0.390	-.302	-.134

Looking at these numbers we see that the first component can be interpreted as an overall measure of crime activity, and looking at the picture of the first 2 PCs we see that cities such as St. Louis, Atlanta, Tampa Bay, Newark, Detroit, Miami, etc can be characterized as "dangerous," while cities such as Virginia Beach, San Jose, Colorado, Honolulu, etc as "safe" (remember these results correspond to 1994 data). The second PC distinguishes between cities with high rape and larceny incidents (and to some degree burglaries) and

cities with high murder, robbery and MVT incidents. So, on the bottom of the picture we find cities such as Newark, Jersey City, Philadelphia, Santa Ana, Detroit, Washington DC, NYC, Chicago and Long Beach, characterized by relatively more murder, robbery and MVT crimes, while in Oklahoma, Corpus Cristi, Tucson and Minneapolis rapes and larcenies are more frequent. However, you should be careful on how far you should go with such an interpretation. It is safe to make such statements for Newark and Detroit, which score high on the first component as well. But the story is not that clear between Washington and Santa Ana, since Santa Ana according to its score on the first component is a relatively "safe" city. On the other hand, the second component allows you to distinguish between Corpus Cristi and Santa Ana, that score similarly on the first component. So, the picture tells you that there are many more rapes in Corpus Cristi compared to Santa Ana or NYC, while more murders in the latter two cities. Finally, the third PC basically contrasts cities with lots of rapes vs cities with lots of larcenies, but since it accounts for 10of the total variance, you should be cautious and not make a big deal out of this component.

Next we examine the biplot (Figure 8) in order to look at a joint representation of data points and variables. The arrows on the biplot indicate where you can find cities with high values on a particular variable. The picture of the biplot I gave you in the handout has the signs of the weights reversed on the second PC (that's why Oklahoma is at the bottom of that picture and Newark at the northeast corner). Keeping this in mind, we see that V3 (rape) and V7 (larceny) point towards Minneapolis and Oklahoma, which is consistent with the discussion above.

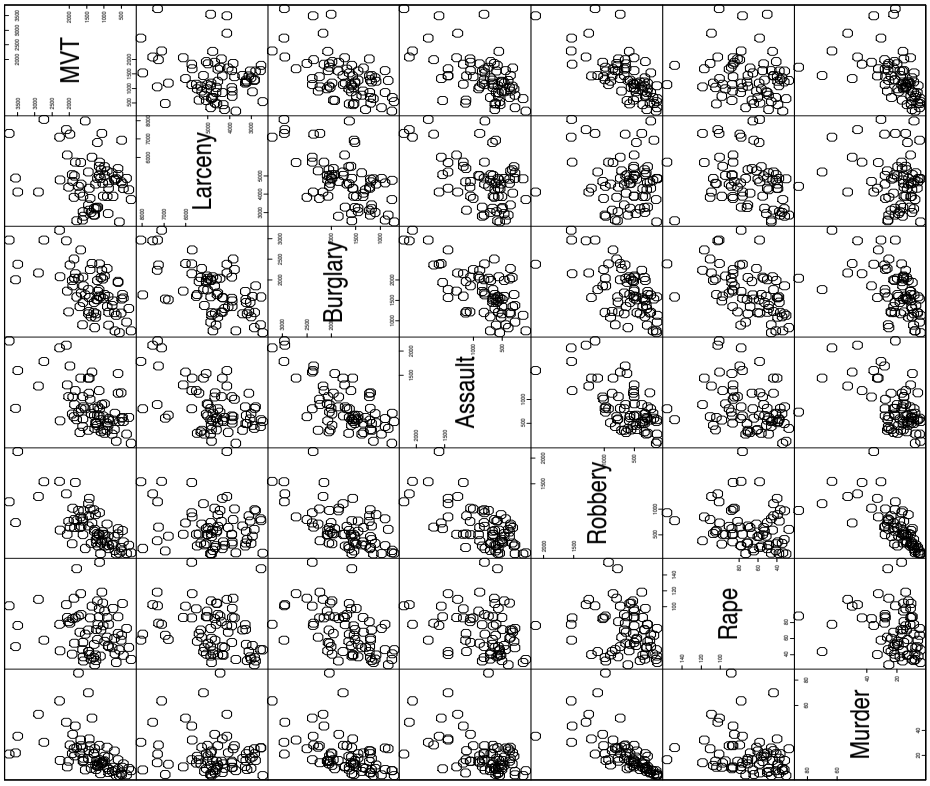


FIGURE 5. Scatterplot matrix of crime data

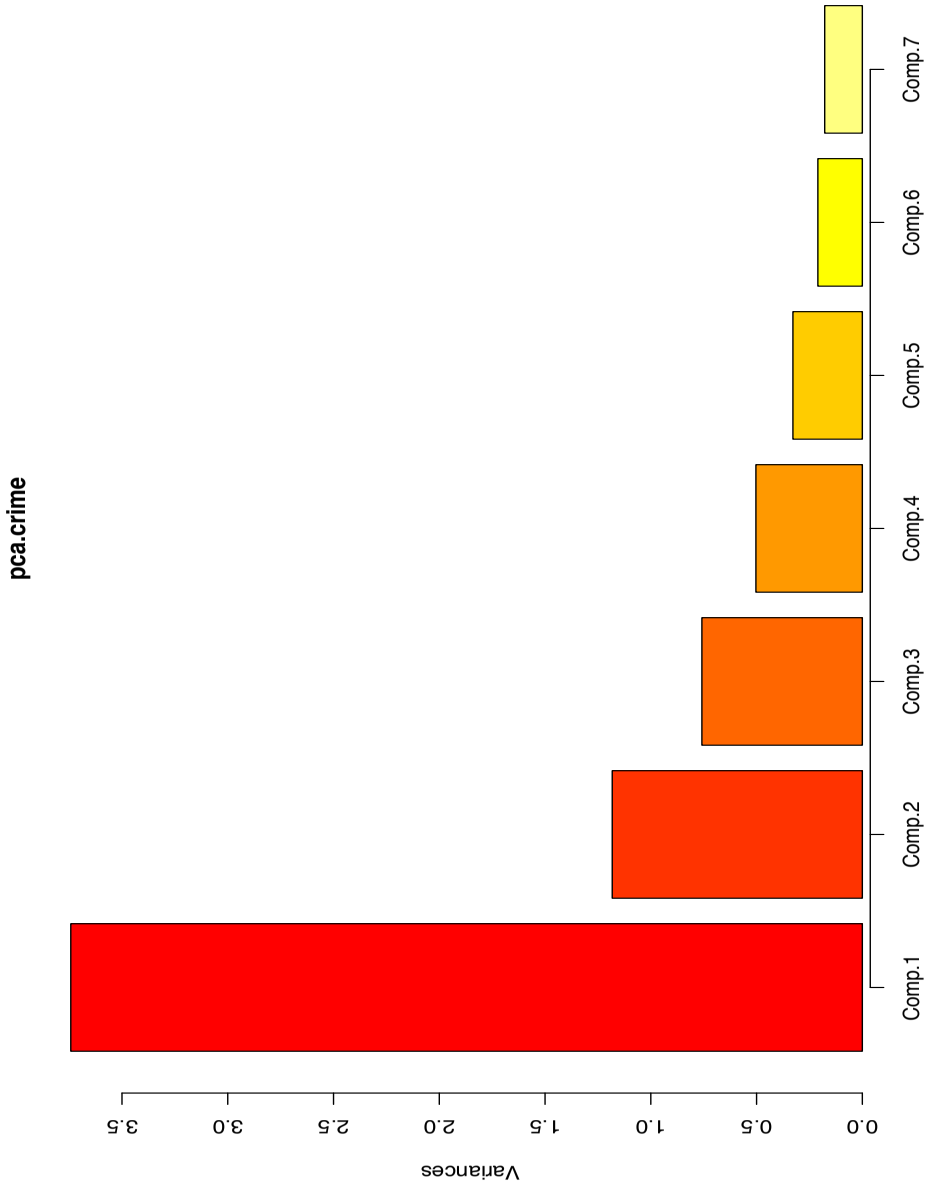


FIGURE 6. Screeplot of eigenvalues of crime data

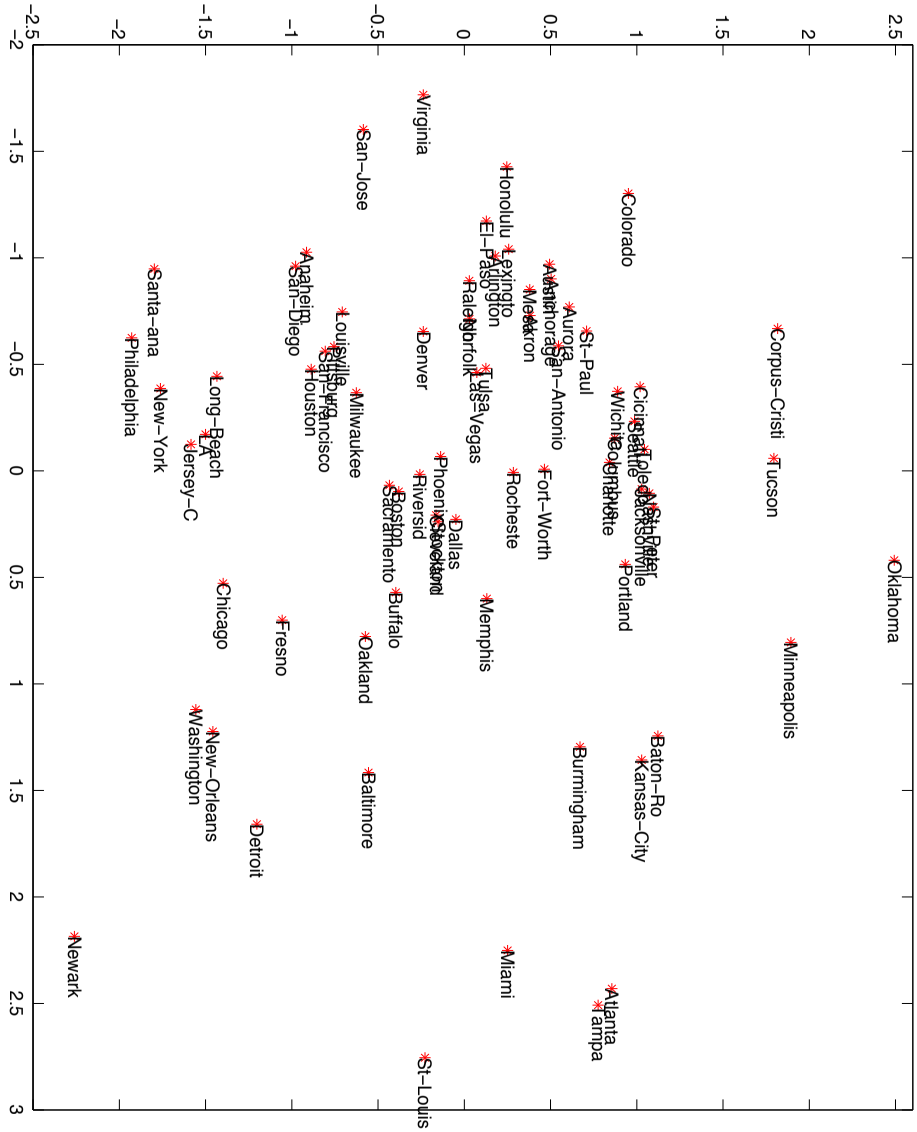


FIGURE 7. First two PCs

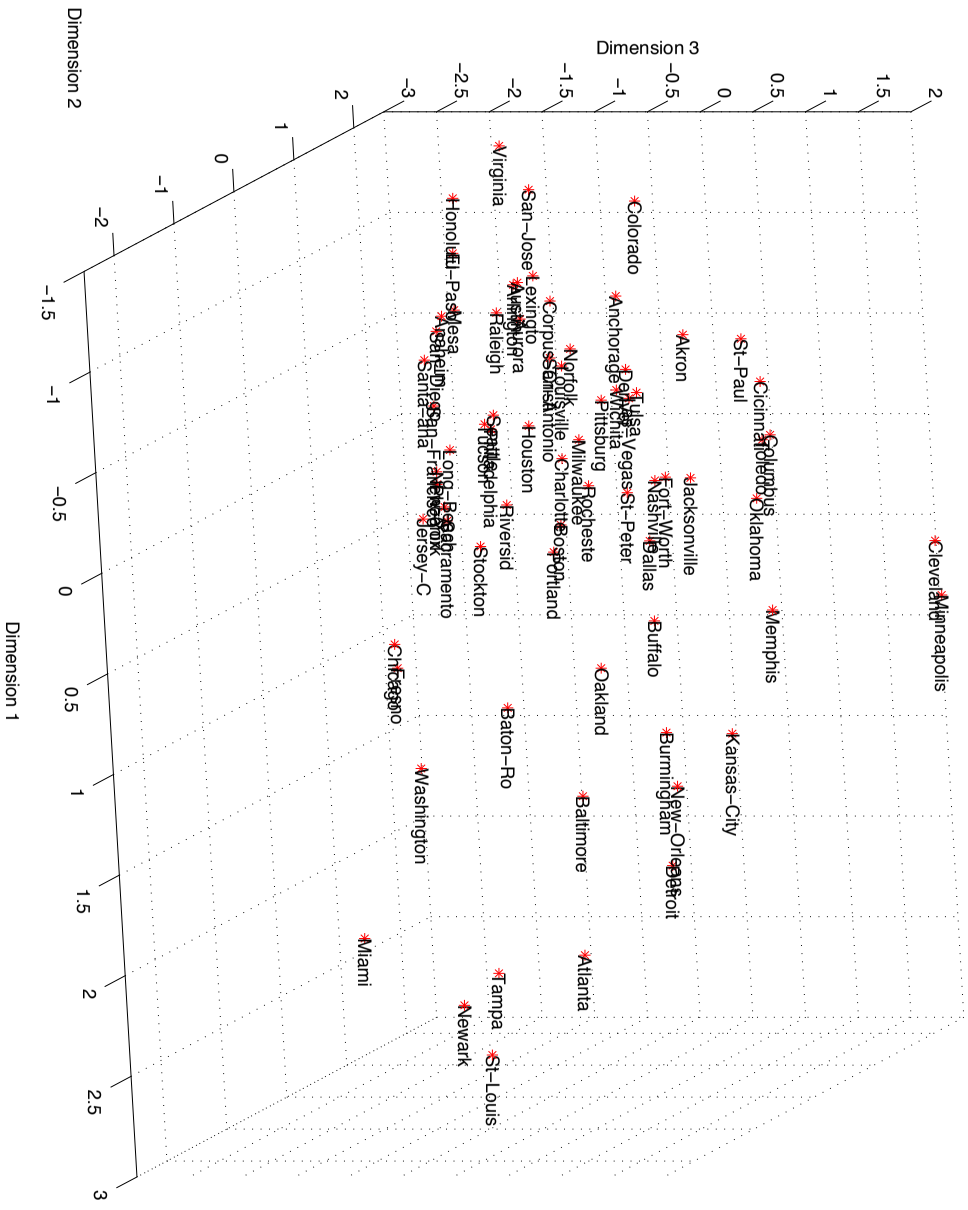


FIGURE 8. First three PCs

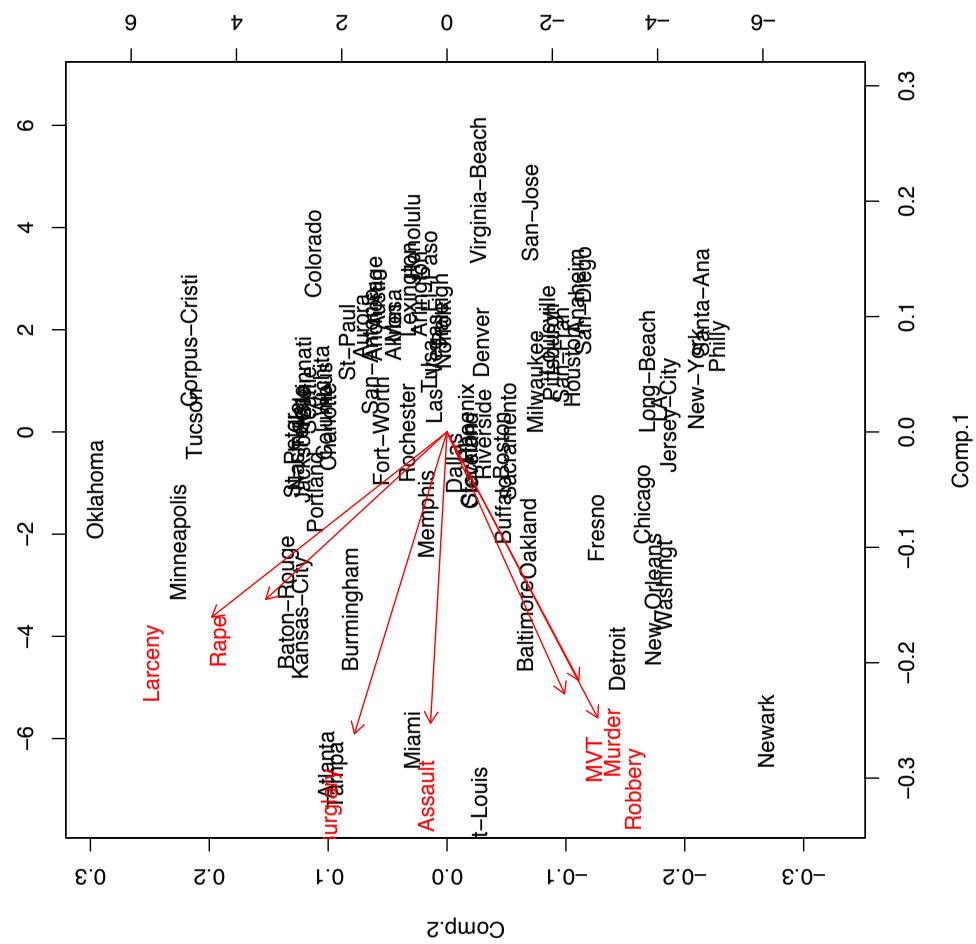


FIGURE 9. Biplot of crime data

PCA of Canadian Temperatures

The data represent average monthly temperatures (in degrees Celsius) for 35 weather stations in Canada (this data set has been also analyzed in Ramsey and Silverman's book *Functional Data Analysis*, 1997, New York: Springer). A time series plot of the data is given next. It can be seen that the data follow a regular pattern, with colder on average

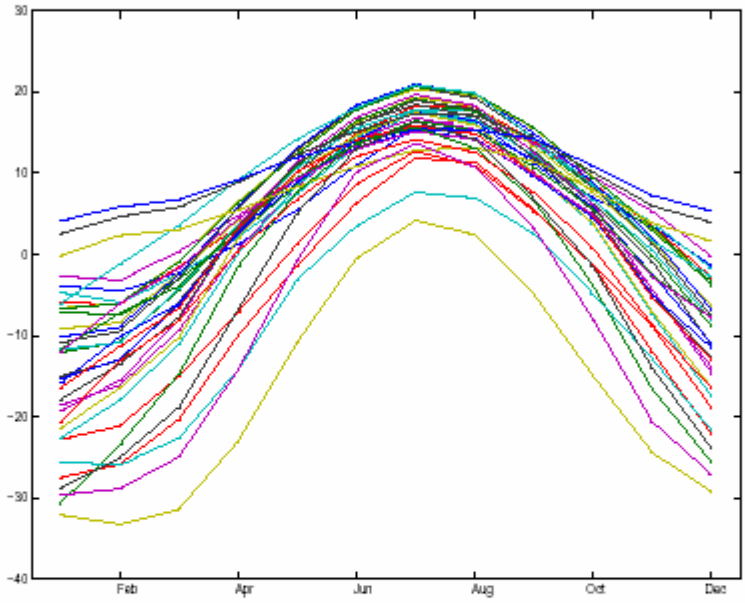


FIGURE 1. Average monthly temperatures in degrees C

temperatures in the winter and spring months and warmer in the summer and fall months. We will use PCA to uncover the basic features in this data set.

```
Importance of components:
                          Comp.1     Comp.2     Comp.3
Standard deviation     3.1447615 1.2037888 0.46514235
Proportion of Variance 0.8483661 0.1243107 0.01856007
Cumulative Proportion  0.8483661 0.9726768 0.99123686
```

It can be seen that the first 3 components capture over 99% of the variation in the data set. Moreover, the type of variation captured by the first PC strongly dominates all other types of variation. In order to interpret the components we examine their loadings next.

```
       Comp.1      Comp.2     Comp.3
Jan 0.2724928  0.38864197 -0.1669129
Feb 0.2840193  0.32068851  0.2823290
Mar 0.3024111  0.17776321  0.2837999
Apr 0.3043605 -0.05148195  0.4575479
May 0.2906974 -0.24656985  0.4464331
Jun 0.2663673 -0.41940374  0.1242702
Jul 0.2600561 -0.44041802 -0.2575670
Aug 0.2843447 -0.31342729 -0.3382340
Sep 0.3086486 -0.08632847 -0.2394660
Oct 0.3083248  0.04508856 -0.1589914
Nov 0.2960035  0.22103108 -0.3196341
Dec 0.2812635  0.35302259 -0.1492018
```

The first PC captures the average trend in the data. Hence, weather stations with high scores will have much warmer than average winters combined with warm summers. From figure 2 we see that Vancouver and Victoria receive the highest scores, while Resolute in the high arctic receives the lowest one. The second PC captures the positive contribution of the winter/spring months and the negative contribution of the summer/fall months, thys corresponding to a measure of uniformity of temperature throughout the year. Low scores go to prairie stations such as Winnipeg that have hot summers and cold winters, while weather stations on the Pacific coast (e.g. Prince Rupert) exhibit very uniform temperatures throughout the year. Finally, the third PC corresponds to a time shift combined with an overall increase in temperature between summer and winter.

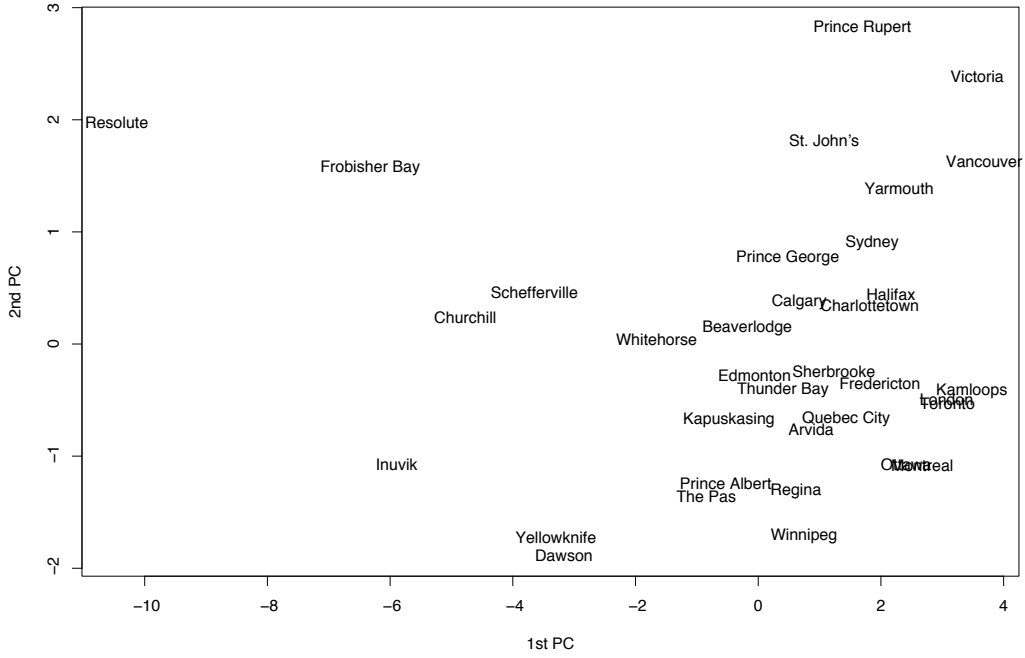


FIGURE 2. Plot of first 2 PCs

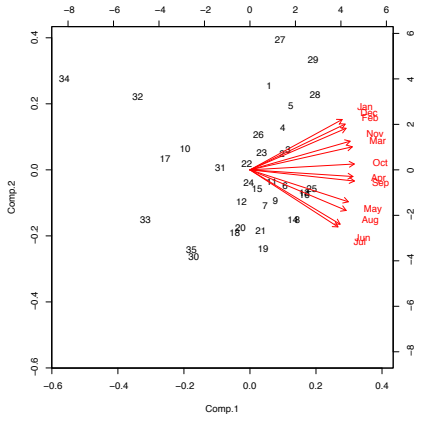
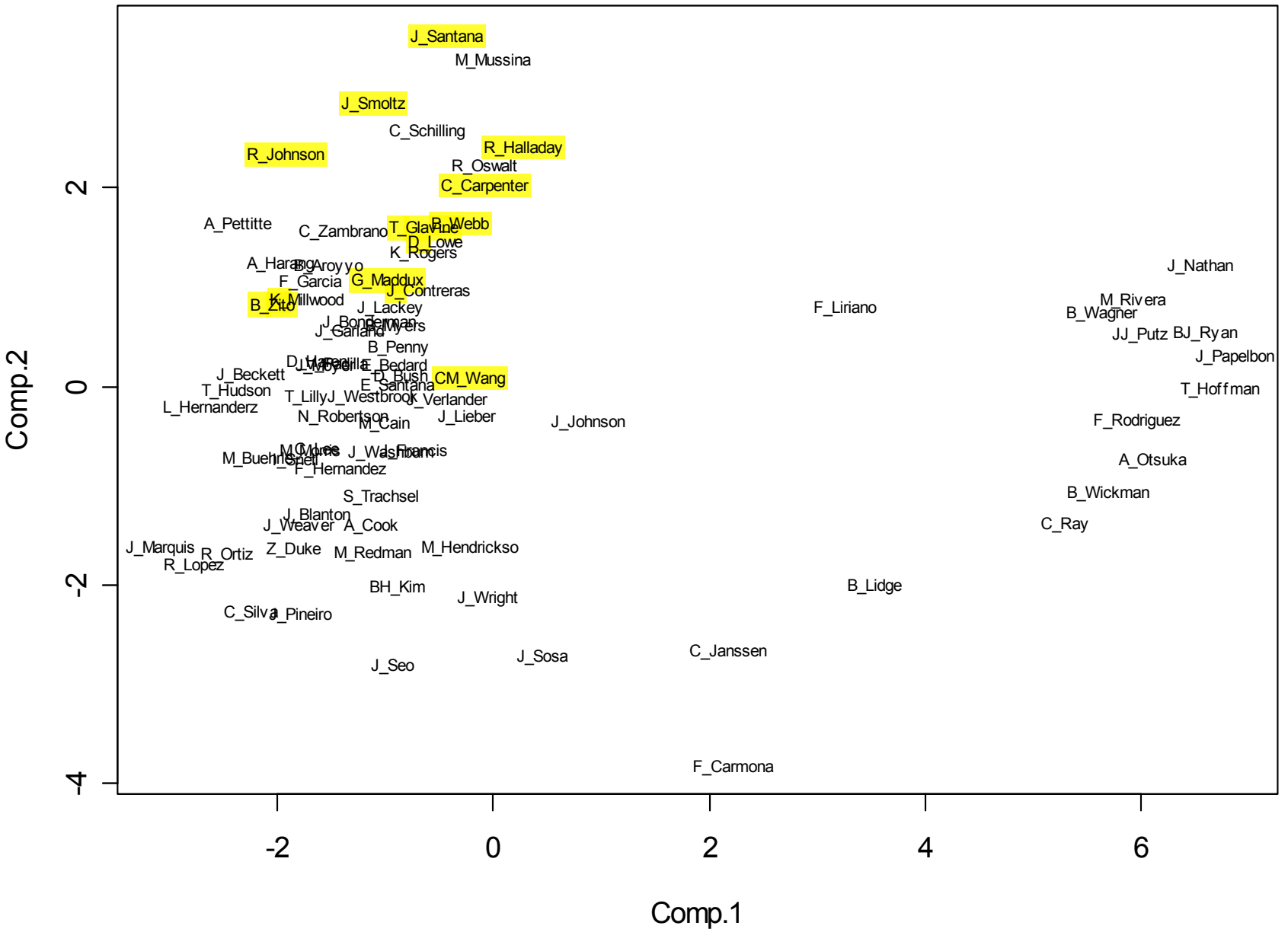


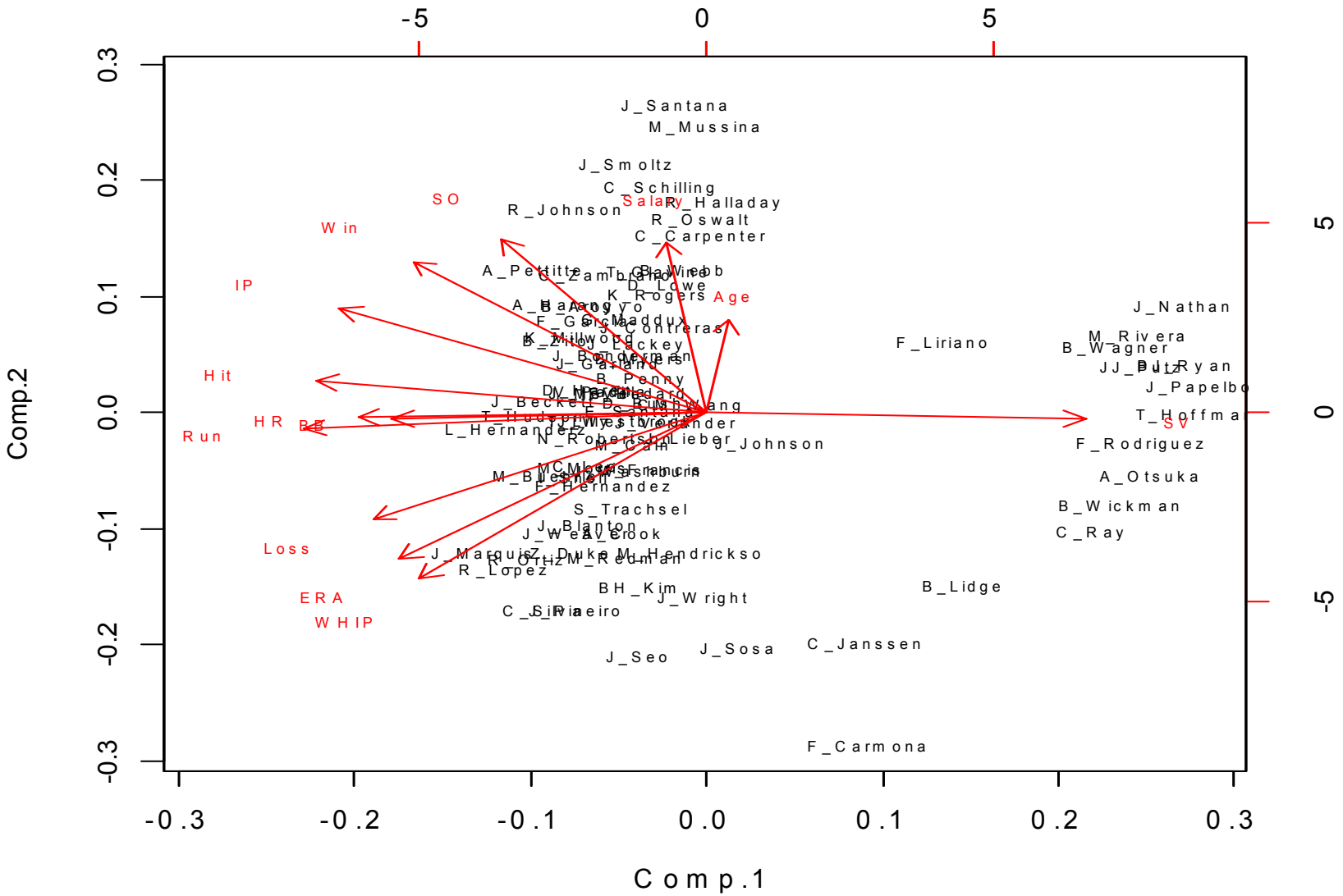
FIGURE 3. Biplot of first 2 PCs

The first 2 PCs for the MLB_2006 Data:

	Comp.1	Comp.2
» IP	-0.33144988	0.25766727
» Hit	-0.35121669	0.07944714
» Run	-0.36287019	-0.03985854
» HR	-0.31349296	-0.01116535
» BB	-0.28321439	-0.01712735
» SO	-0.18567727	0.42916036
» Win	-0.26425623	0.37433141
» Loss	-0.30068679	-0.26245398
» SV	0.34110227	-0.01635911
» WHIP	-0.25905795	-0.41158375
» ERA	-0.27685802	-0.36174289
» Age	0.02024442	0.23119198
» Salary	-0.03718710	0.42205197

» ➔ 72.35% variance explained.


J_Santana
M_Mussina
J_Smoltz
C_Schilling
R_Johnson
R_Halladay
R_Oswalt
C_Carpenter
A_Pettitte
C_Zambrano
T_Glavine
R_Webb
D_Lowe
K_Rogers
A_Harang
J_Arroyo
F_Garcia
G_Maddux
J_Contreras
K_Millwood
B_Zito
J_Lackey
J_Garland
B_Penny
J_Beckett
E_Bedard
E_Bush
CM_Wang
T_Hudson
E_Santana
T_Lilly
J_Westbrook
J_Verlander
L_Hernanderz
N_Robertson
M_Cain
J_Lieber
J_Johnson
M_Buehrle
J_Washburn
J_Francis
F_Hernandez
S_Trachsel
J_Blanton
J_Weaver
A_Cook
J_Marquis
R_Ortiz
Z_Duke
M_Redman
M_Hendrickso
R_Lopez
BH_Kim
J_Wright
C_Silva
J_Pineiro
J_Seo
J_Sosa
C_Janssen
F_Carmona
B_Lidge
F_Liriano
J_Nathan
M_Rivera
B_Wagner
JJ_Putz
BJ_Ryan
J_Papelbon
T_Hoffman
F_Rodriguez
A_Otsuka
B_Wickman
C_Ray
Comp.2
Comp.1
-4
-2
0
2
-2
0
2
4
6


Comp.2
Comp.1
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
-5
0
5
SO
Win
IP
Hit
HR
Run
Loss
ERA
WHIP
Salary
Age
SV
J_Santana
M_Mussina
J_Smoltz
C_Schilling
R_Johnson
R_Halladay
R_Oswalt
C_Carpenter
A_Pettitte
T_Glavine
D_Lowe
K_Rogers
J_Nathan
M_Rivera
B_Wagner
J_Papelbo
F_Liriano
J_Johnson
F_Rodriguez
A_Otsuka
B_Wickman
C_Ray
S_Trachsel
J_Weaver
J_Marquis
D_Duke
M_Hendrickson
R_Ortiz
M_Redman
R_Lopez
BH_Kim
J_Wright
C_Silva
B_Lidge
C_Janssen
J_Seo
J_Sosa
F_Carmona