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.

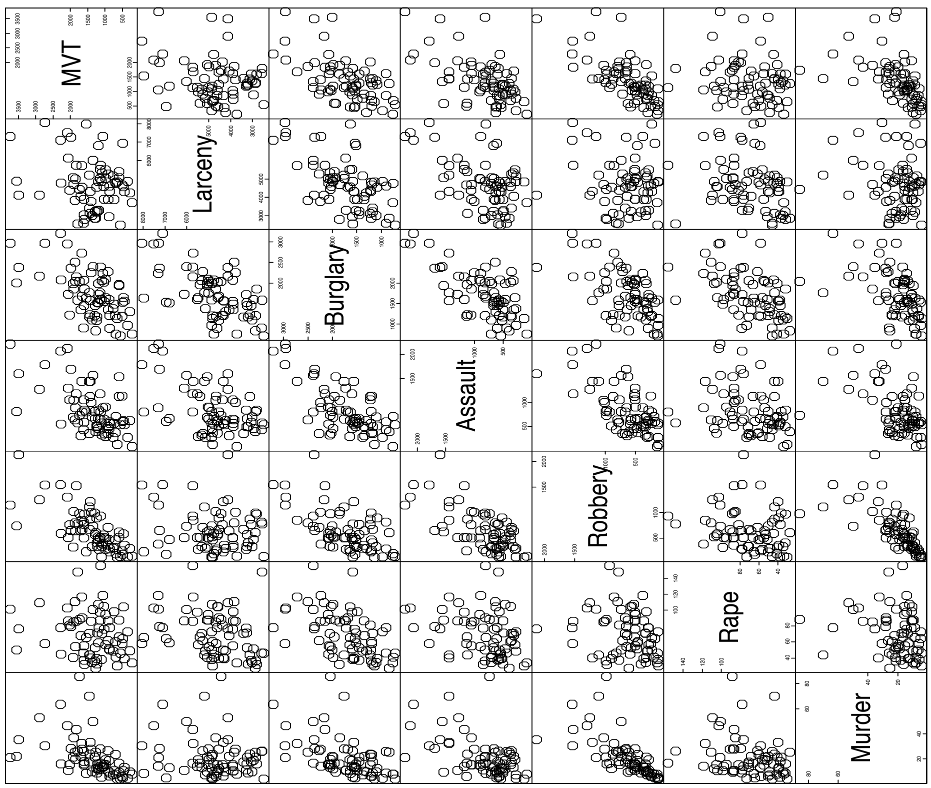

FIGURE 5. Scatterplot matrix of crime data

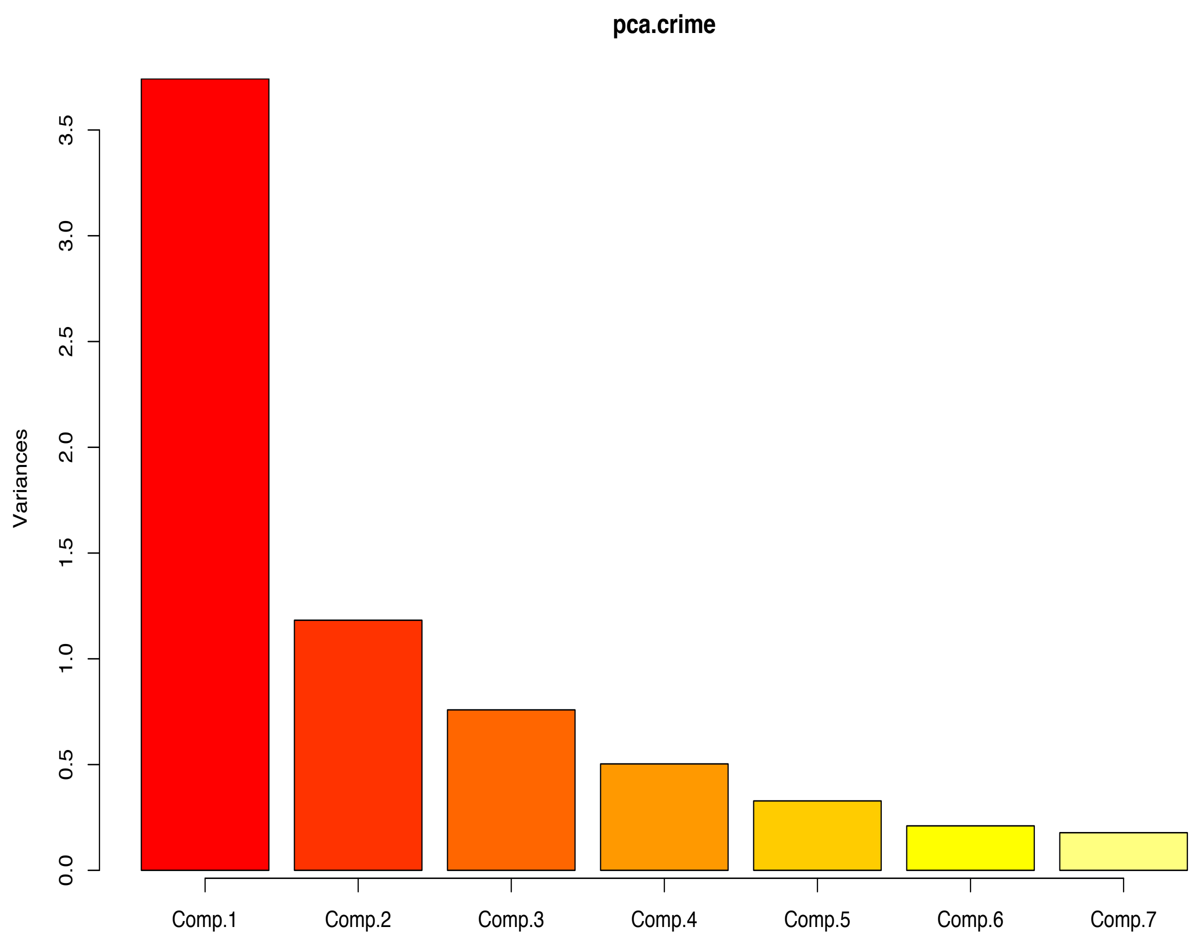

FIGURE 6. Screeplot of eigenvalues of crime data

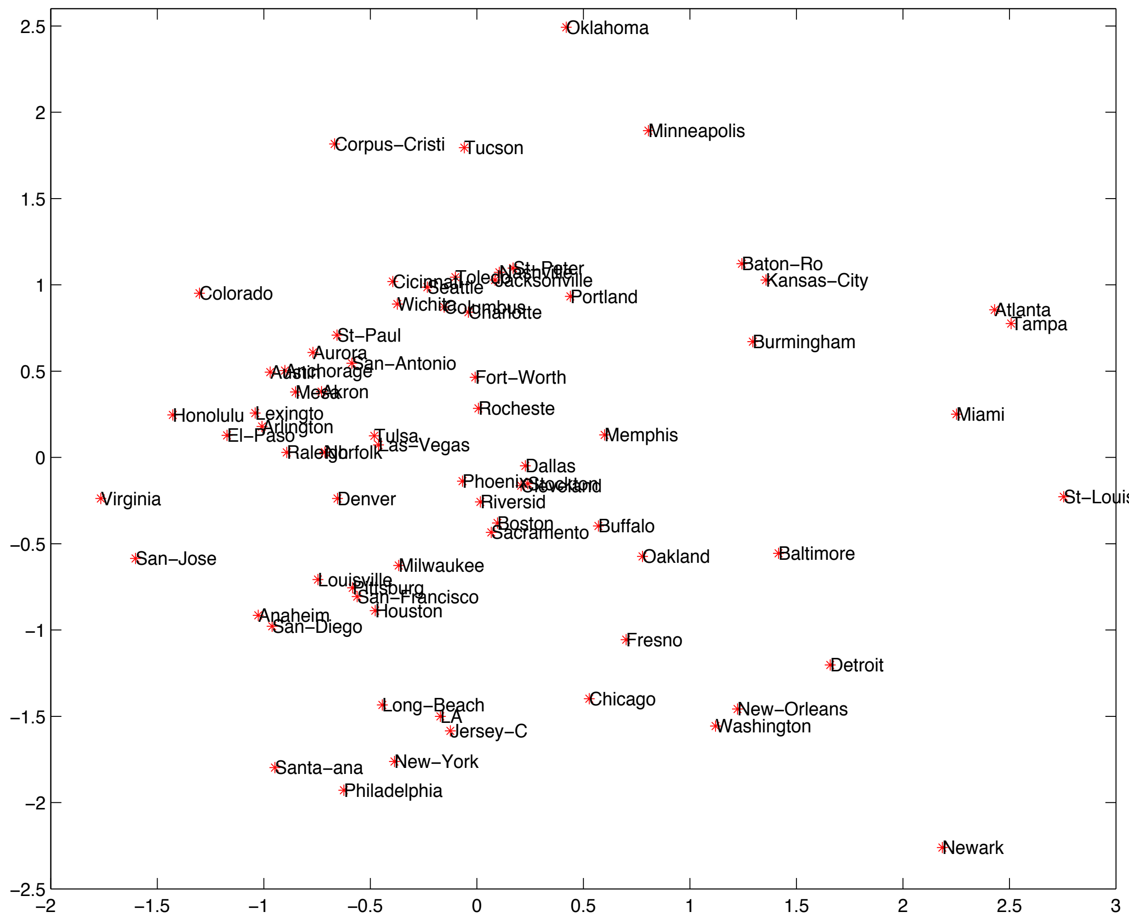

FIGURE 7. First two PCs

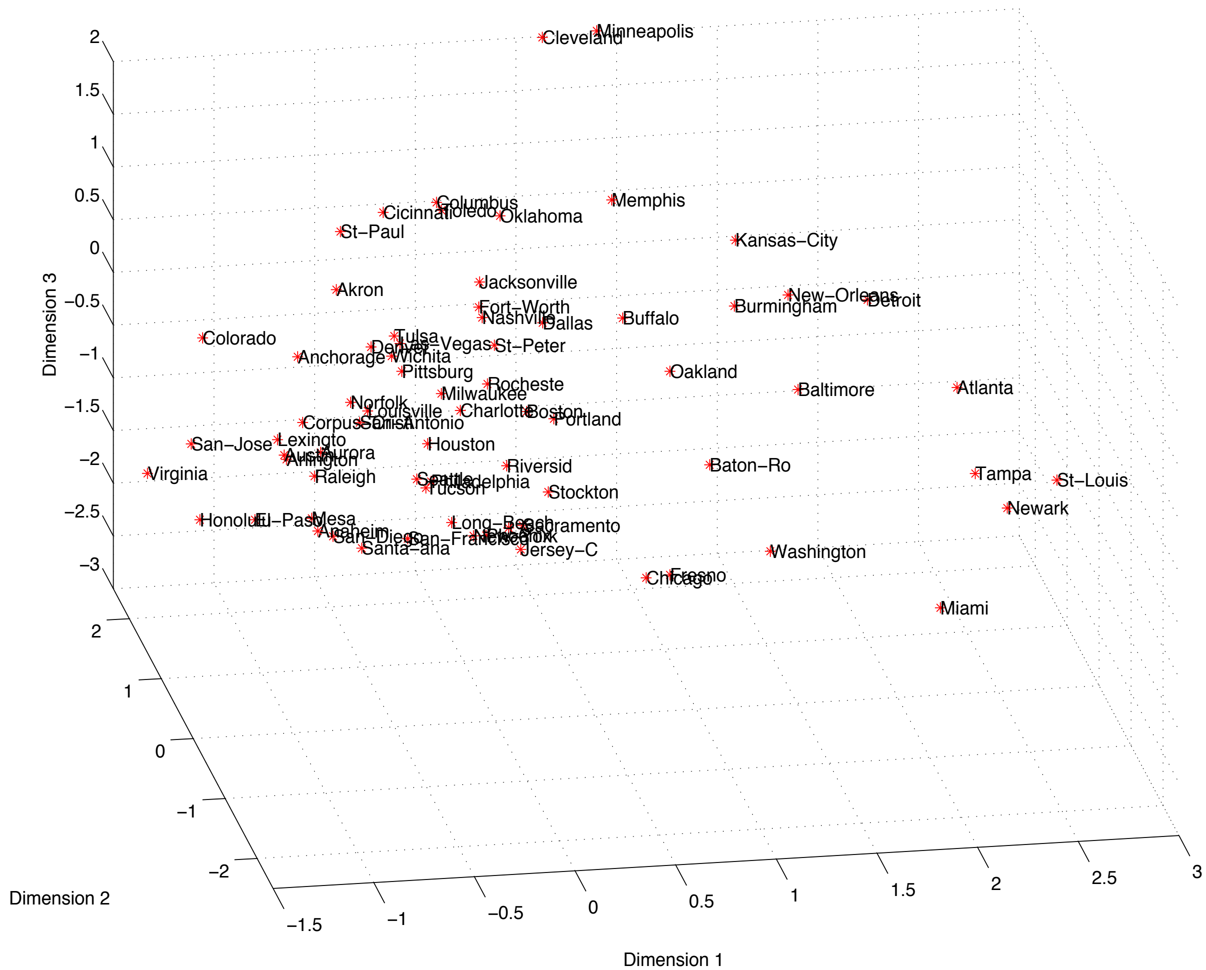

FIGURE 8. First three PCs

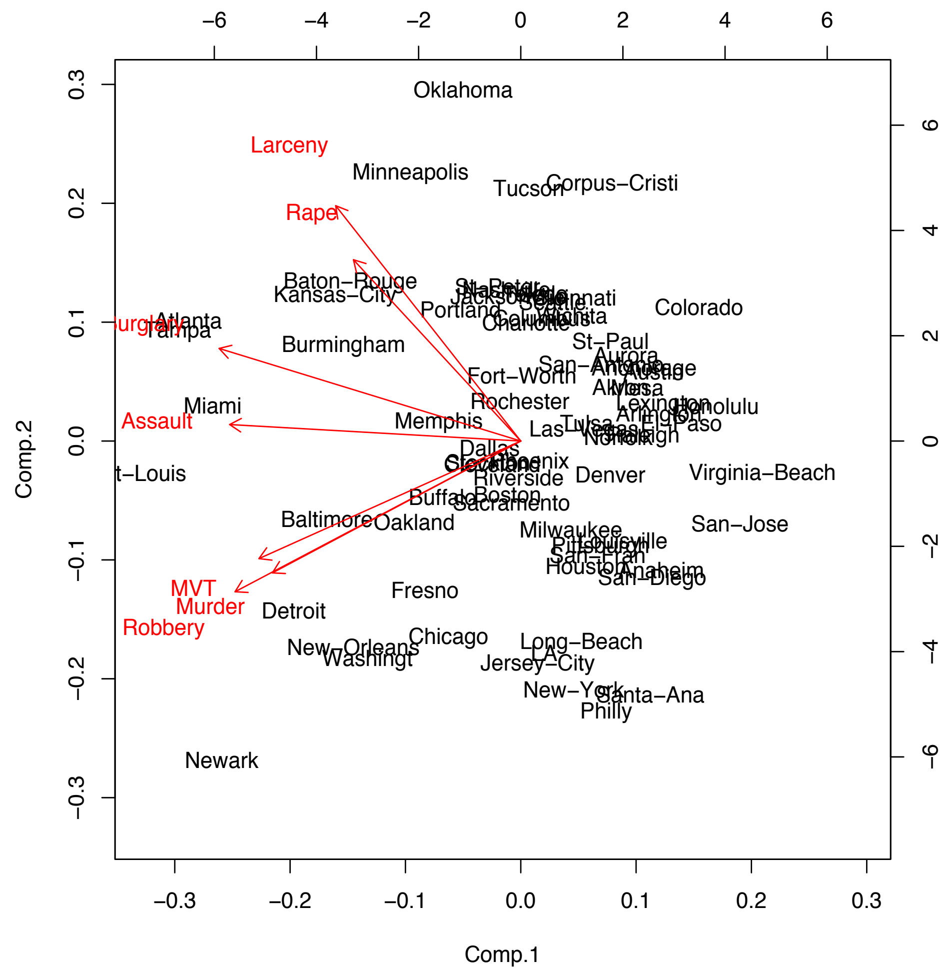

FIGURE 9. Biplot of crime data

PCA of Canadian Temperatures

12

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

$n=35$
$p=12$

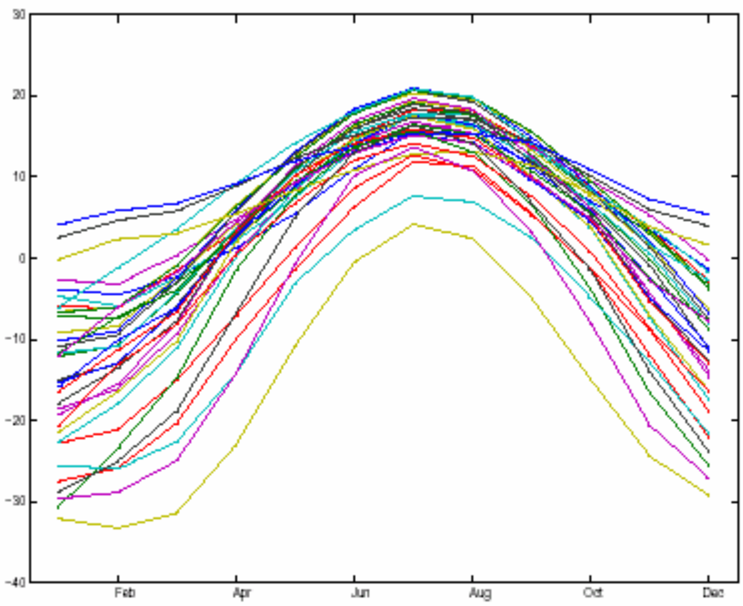

FIGURE 1. Average monthly temperatures in degrees C

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

冬天溫度較高
夏天溫度較低
会跑到图上方

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.

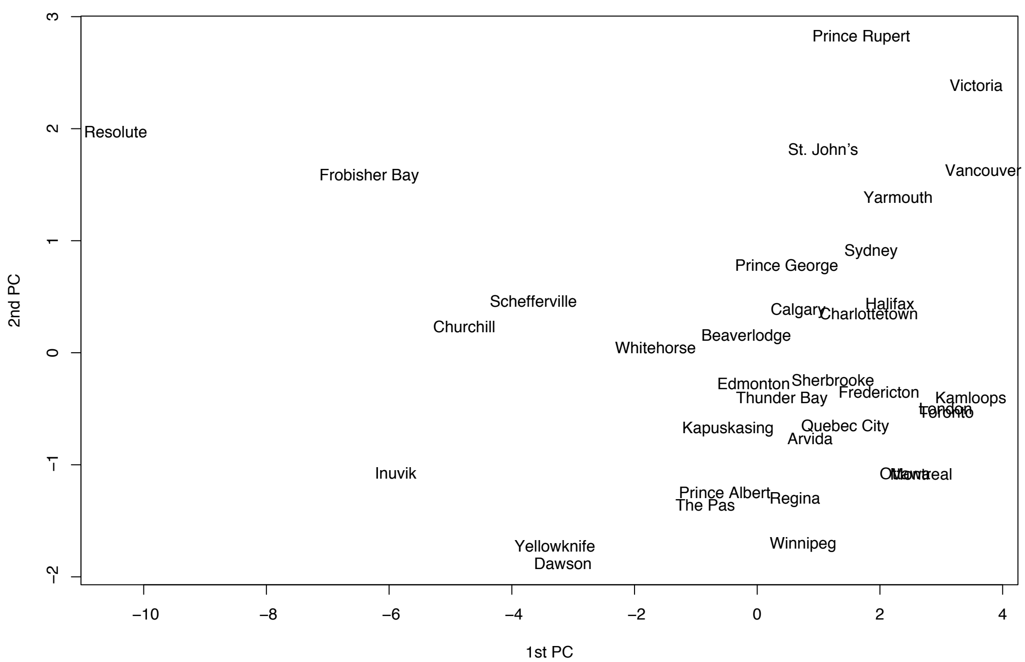

FIGURE 2. Plot of first 2 PCs

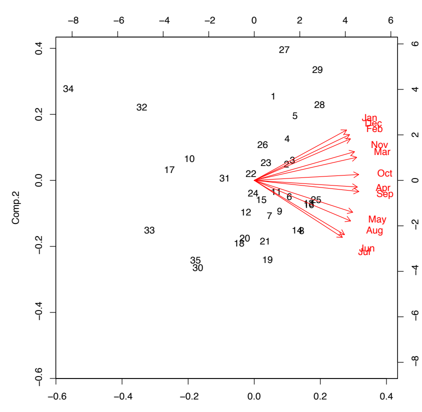

FIGURE 3. Biplot of first 2 PCs