HKU BUSINESS SCHOOL
港大經管學院

Business Statistics

Principal Component Analysis

Weichen Wang

Assistant Professor

Innovation and Information Management

ISLR Chapter 12.1-12.3, 6.3

Customer Contact Center Workforce Management

Workforce Management: Optimization

Problem

Schedule the right agent at the right time
to serve the right customer at the right level
with the minimum cost

Demand

Customer arrival, conversation time, patience

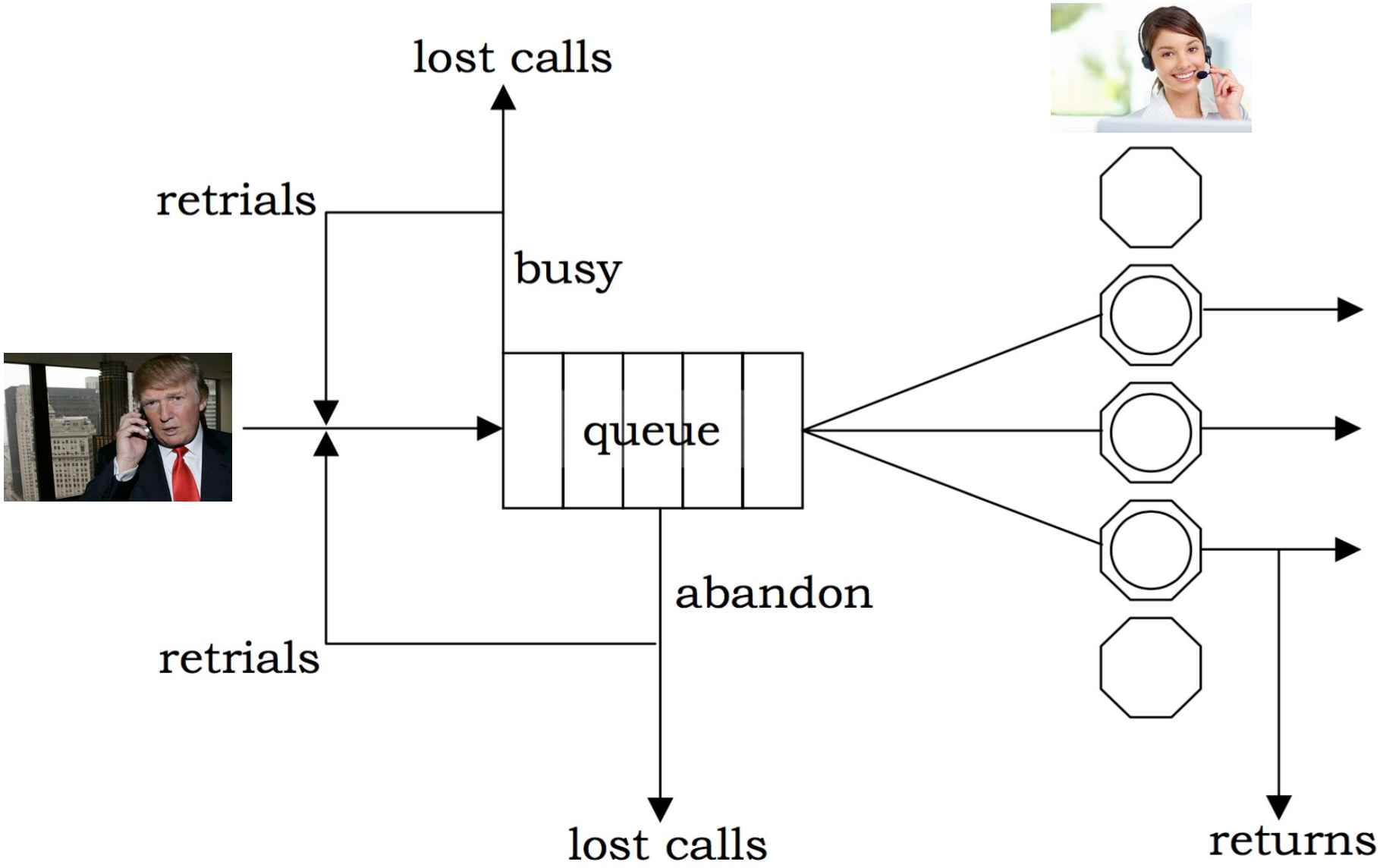
Supply

Agent skill set, number, shift

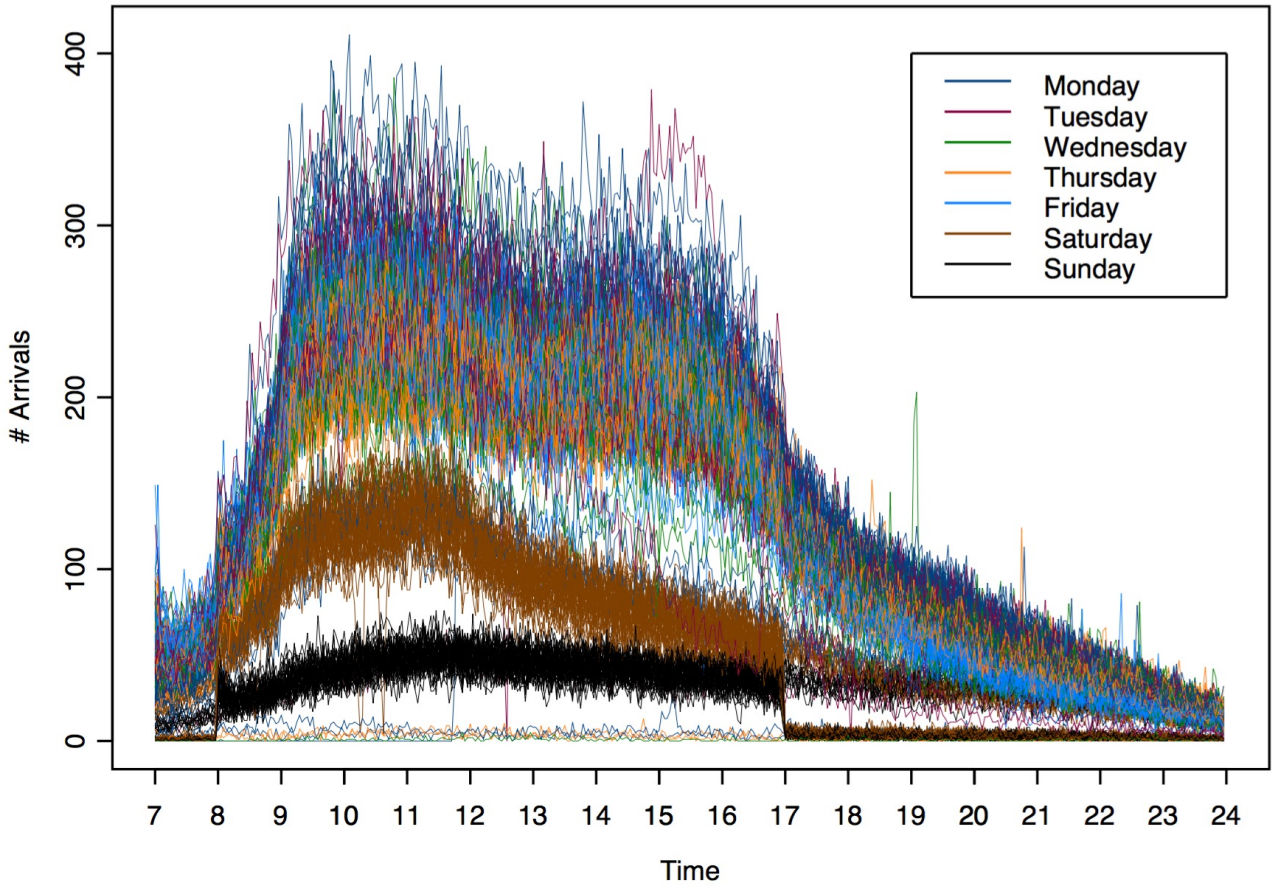
Quality of Service

Waiting time · abandonment · first call resolution

Queueing Model for A Single Call Type and Location



Demand - Call Center Arrivals: Key Features?



- Bank with a network of 4 call centers in northeast US
- 300K calls / day, 60K / day seeking agents
- 2.5-minute interval counts from 7am to 24am.

Stock Daily Closing Price 2000-2013

Altria (formerly Philip Morris; MO)
Apple (AAPL)
Automatic Data Processing (ADP)
Corrections Corporation of America (CXW)
Equifax (EFX)
Ford (F)
General Electric (GE)
Graham Holding Companies (GHC)
Proctor and Gamble (PG)
United States Steel (X)
Yahoo! (YHOO)
Amazon (AMZN)
Archer Daniels Midland (ADM)
Bank of America (BAC)
Dow Chemicals (DOW)
ExxonMobil (XOM)
Halliburton (HAL)
Goldman Sachs (GS)
Microsoft (MSFT)
Time Warner (TWX)
Walmart (WMT)
Yum! Brands (YUM)

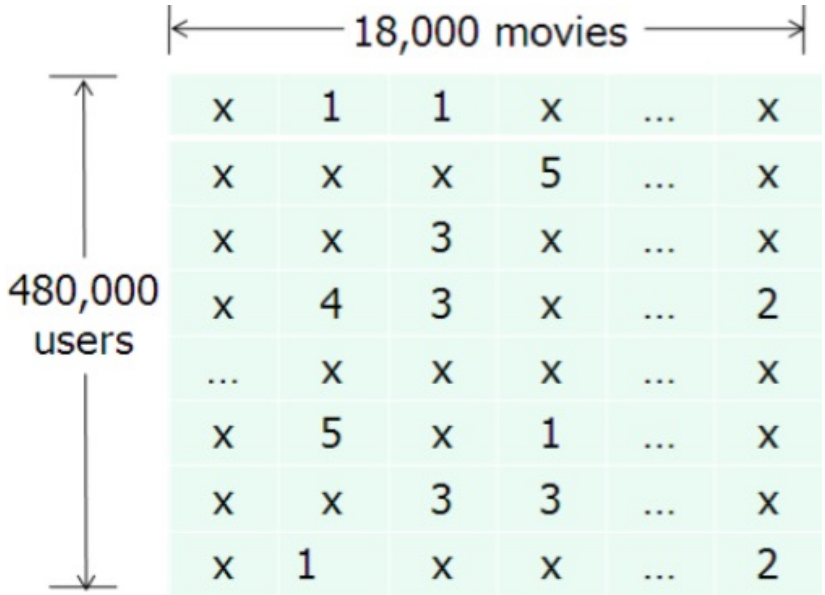
Netflix Movie Recommendation

- A US-based DVD retail company (1997 -)

- Good recommendation = happy customer = business value (Famous product: House of Cards, Narcos, Squid Game…)

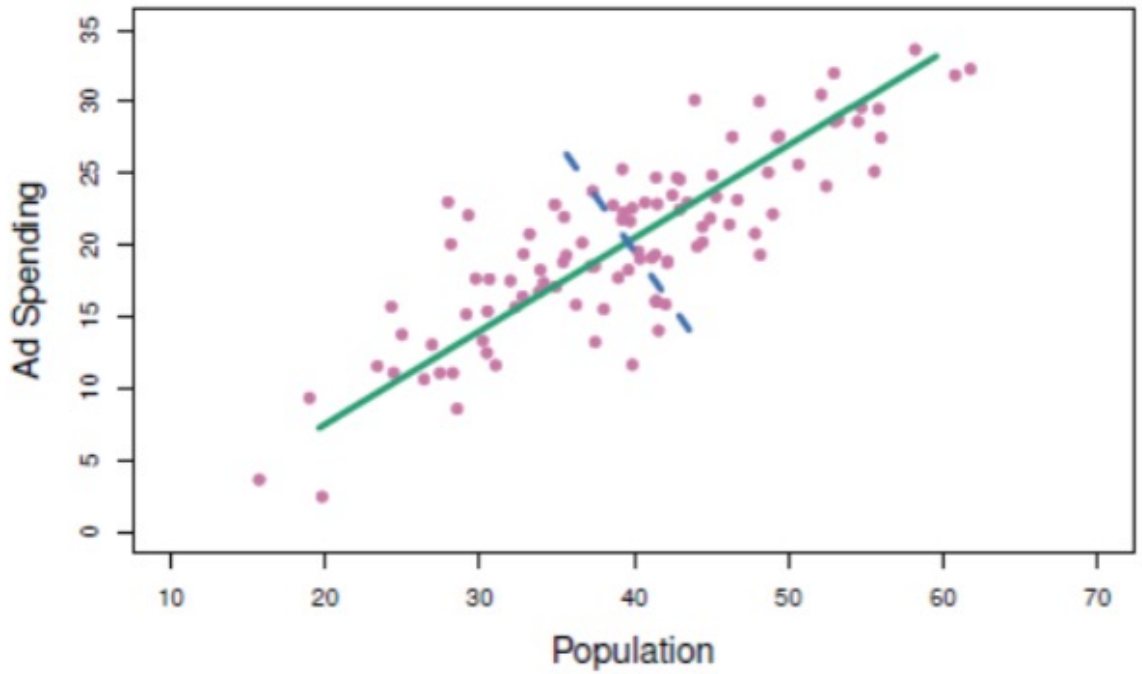
The Netflix Competition (2006-2009)

- Netflix offers $1M for an improved recommender algorithm
- Training data: 100 million ratings



- Test data
  - Last few ratings of each user (2.8 million)
- Winner BellKor's Pragmatic Theory, using a combination of > 800 models
  https://www.youtube.com/watch?v=ImpV70uLxyw
  - Two main classes: nearest neighbors and principal component analysis

Principal Component Analysis (PCA)



The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.

Popular unsupervised way for dimension reduction (Pearson, 1901; Hotelling, 1933)

The First Principal Component

- Consider a set of features: $X_1, X_2, \ldots, X_p$ on n individuals, and their normalized linear combination:

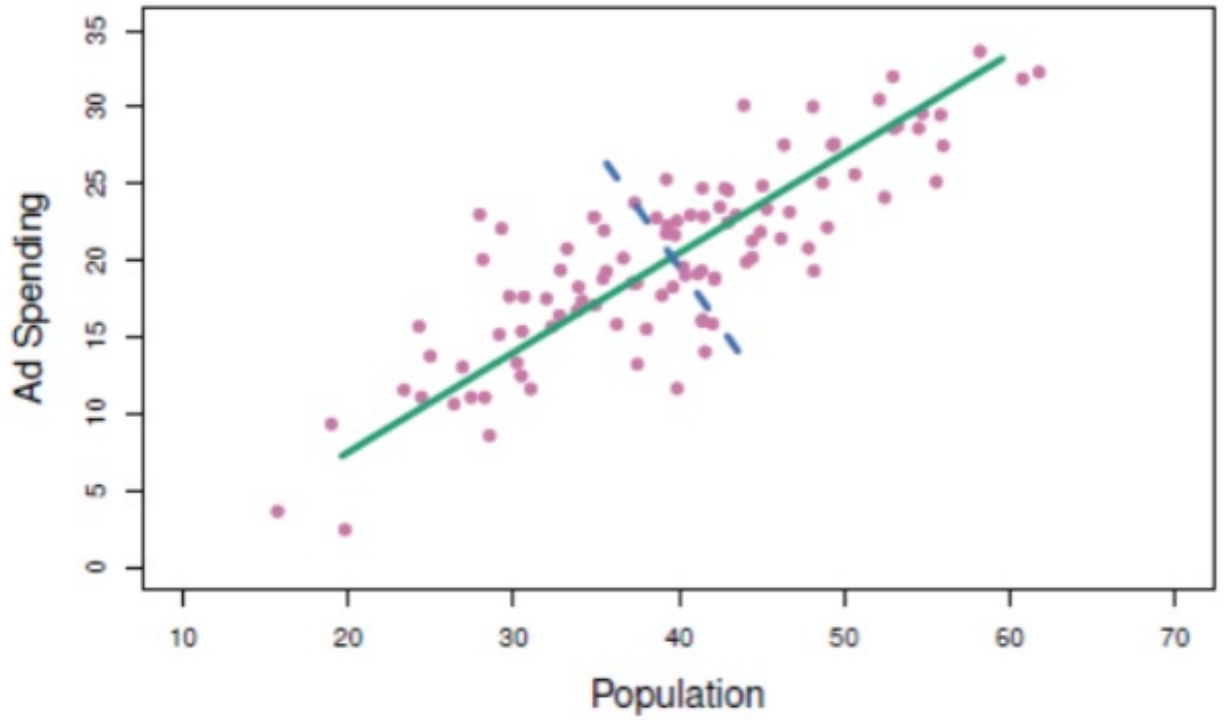
$$Z = \phi_1 X_1 + \phi_2 X_2 + \cdots + \phi_p X_p \text{ with } \sum_{j=1}^{p} \phi_j^2 = 1$$

- The first principal component (PC) is one such linear combination $Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p$ that has the largest variance

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

- The entries $z_{11}, z_{21}, \ldots, z_{n1}$ are the PC scores.
- The PC loading vector: $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^T$

Geometry of PCA



- The 1st PC loading vector ϕ_1 defines a direction in the feature space along which the data vary the most – the green line
- If we project the data points onto this direction, the project values are the PC scores in Z_1.

Geometry of PCA

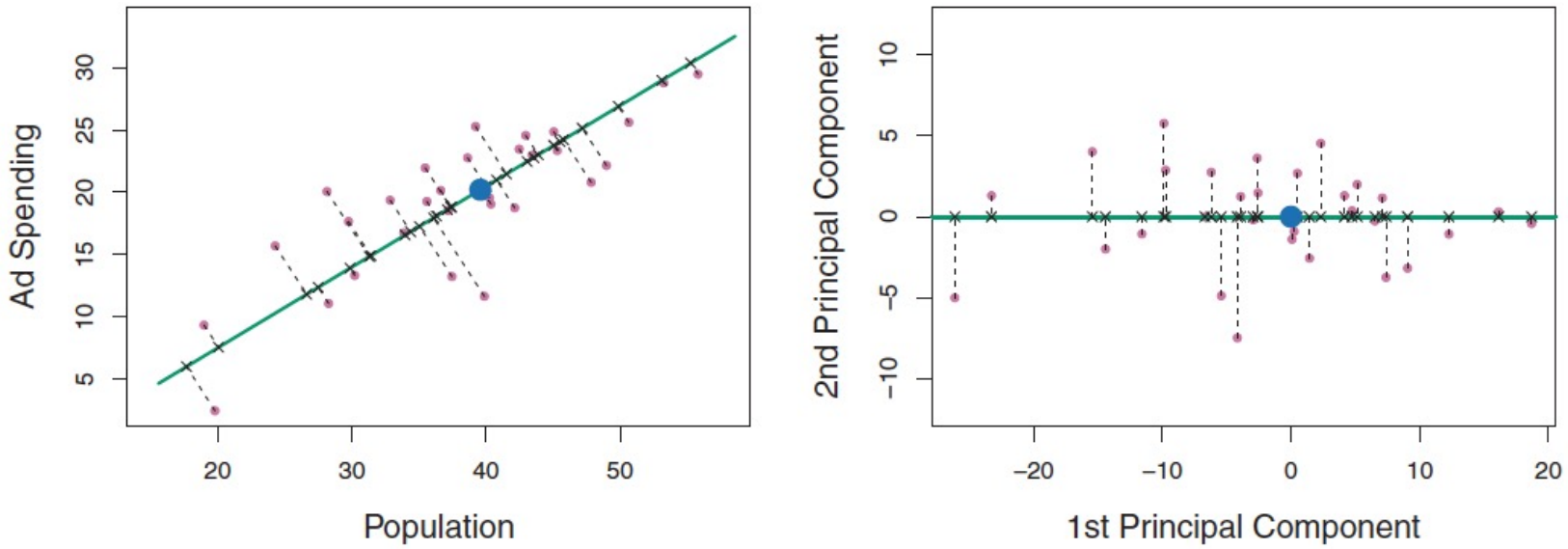


FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all* n *of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents* $(\overline{\texttt{pop}}, \overline{\texttt{ad}})$. Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the* x*-axis.*

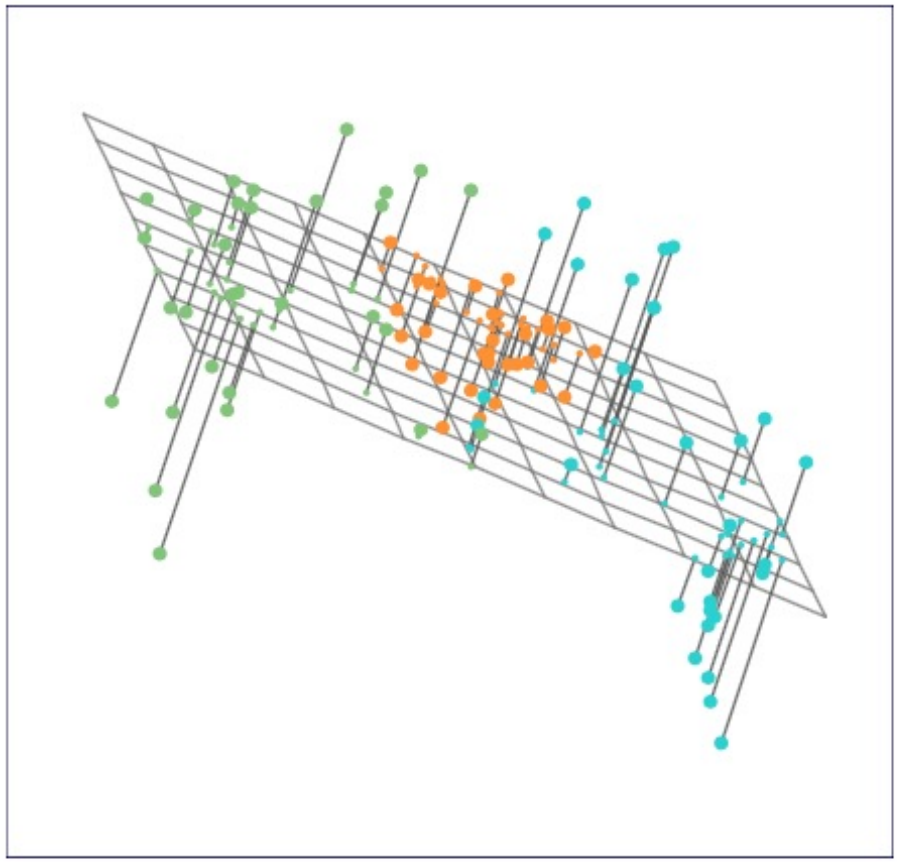
Higher Order Principal Components

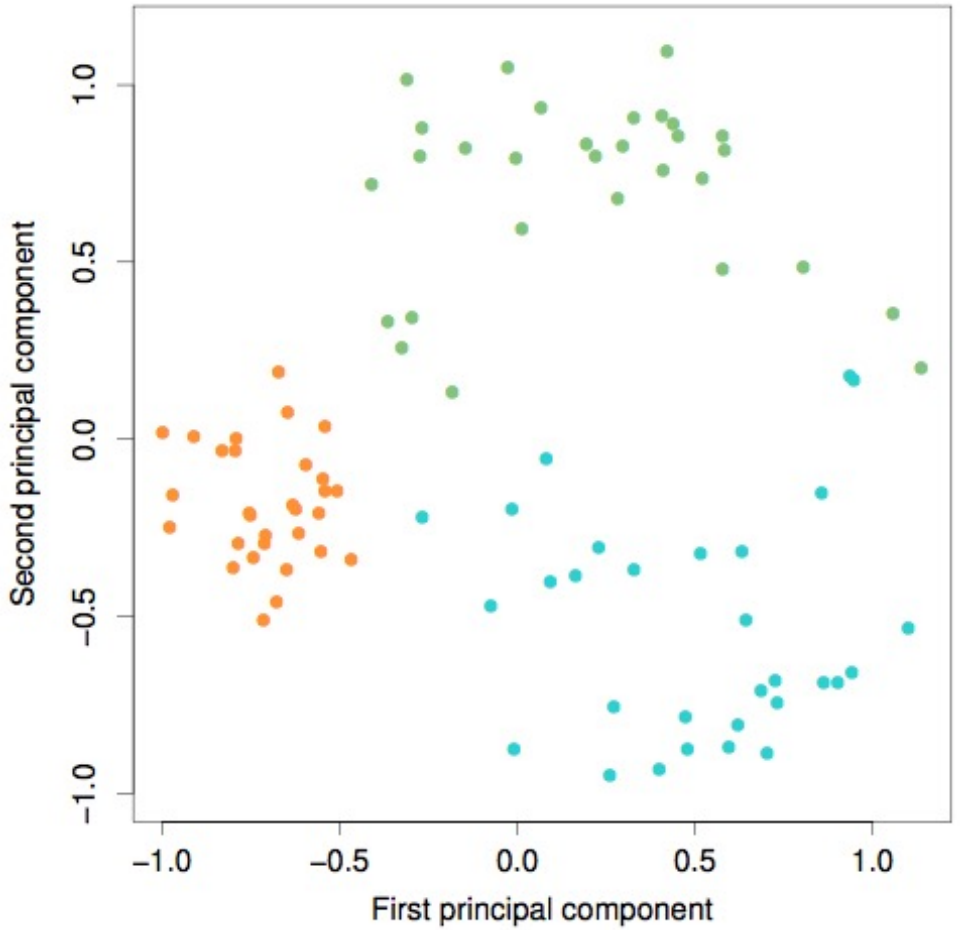
- The second principal component Z_2 is the normalized linear combination that has the maximal variance among all linear combinations that are *uncorrelated* with Z_1. (The blue dash line in page 11).

- The un-correlatedness is equivalent to the orthogonality between ϕ_1 and ϕ_2.

- Similarly,
  - The third principal component Z_3 has the maximum variance that are *uncorrelated* with both Z_1 and Z_2.
  - …
  - The coefficients in the linear combinations are the corresponding loadings.

Another Interpretation of Principal Components





- PCA finds hyperplanes closest to the observations!

Closest Hyperplanes

- The first M PC vectors give the following approximation.

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\phi_{jm}.$$

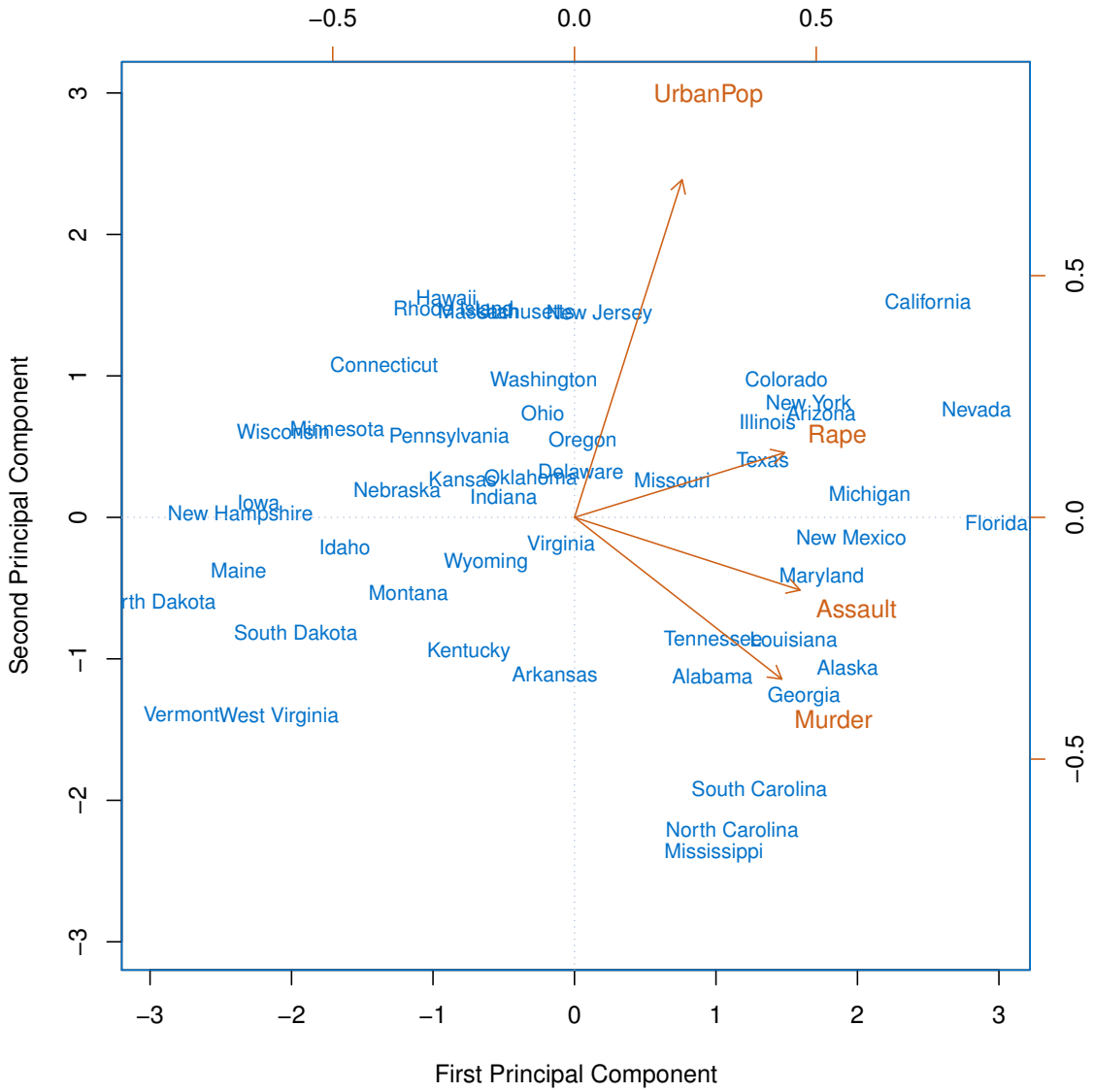
- Actually, $\hat{a}_{im} = z_{im}, \hat{b}_{jm} = \phi_{jm}$ does minimize the MSE.

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p}\sum_{i=1}^{n} \left(x_{ij} - \sum_{m=1}^{M} a_{im}b_{jm} \right)^2 \right\}.$$

USAarrests: 50 US states, 4 Variables

- Number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape
- UrbanPop: the % of the population in each state living in urban areas
- The crime variables are correlated.
- Can we summarize the 4 variables using a smaller number of "variables"?
- PCA was performed after standardizing each variable to have mean zero and standard deviation one.

USAarrests: Biplot



Biplot: Details

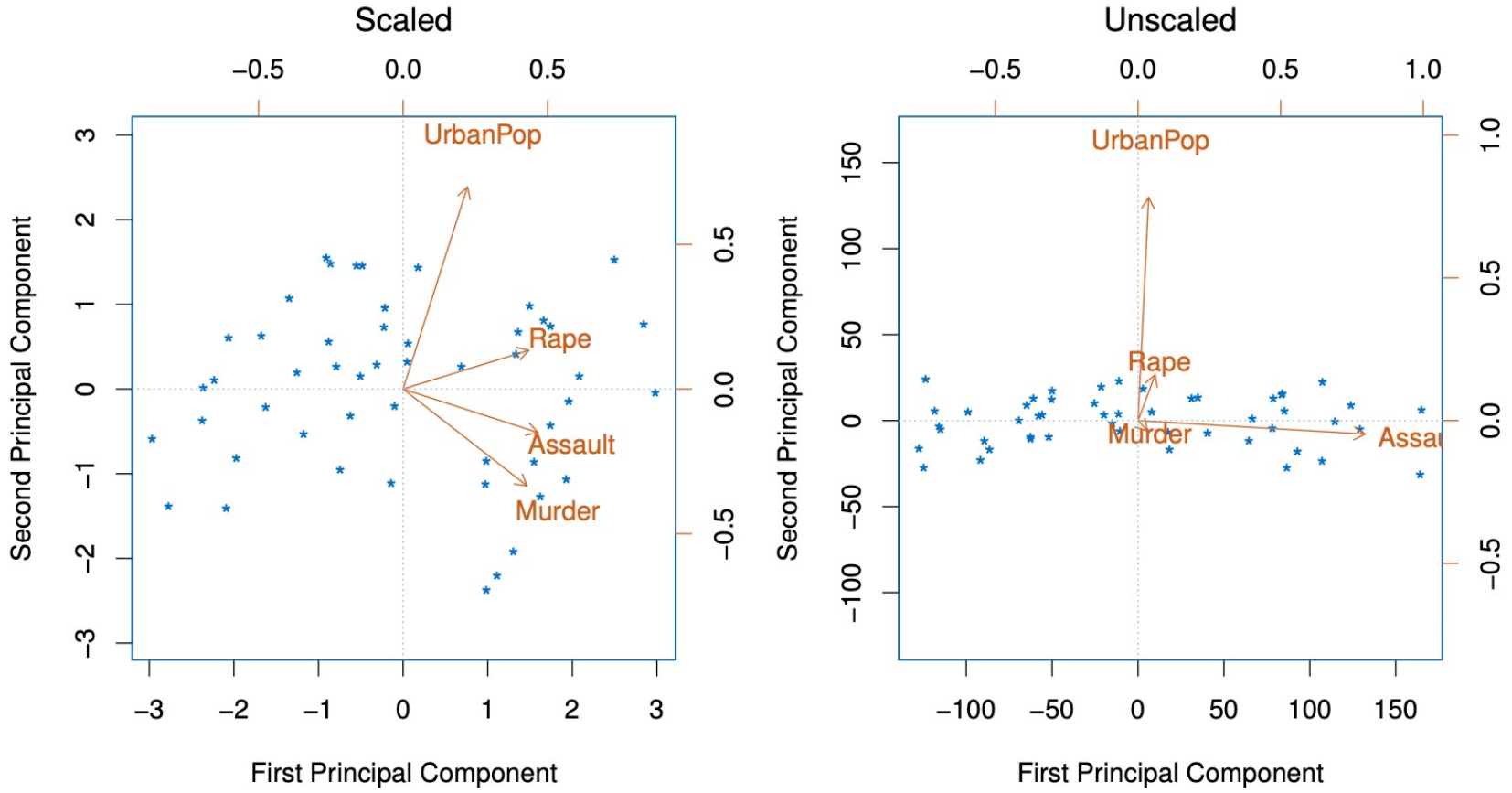
- Biplot shows both the PC scores and loadings.
- PC scores: the blue state names
- PC loading vectors: the orange arrows (with axes on the top and right)

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

- For example, the loading for Rape on PC1 is 0.54, and PC2 is 0.17 (center of the word)

Scaling the Variables Matters!

- Variables are measured in different units and can have different variances.
  - E.g., Murder, Rape, Assault and UrbanPop have variances of 19.0, 87.7, 6945.2 and 209.5



R Commands (Section 12.5)

PCA

```
> pr.out=prcomp(USArrests, scale=TRUE)
> names(pr.out)
[1] "sdev"     "rotation" "center"   "scale"    "x"
```

PC Loadings

```
> pr.out$rotation
             PC1    PC2    PC3    PC4
Murder   -0.536  0.418 -0.341  0.649
Assault  -0.583  0.188 -0.268 -0.743
UrbanPop -0.278 -0.873 -0.378  0.134
Rape     -0.543 -0.167  0.818  0.089
```

Principal components are only unique up to a sign change.

PC -> -PC; Loadings -> -Loadings

PC Scores

```
> dim(pr.out$x)
[1] 50  4
```

Biplot

```
> biplot(pr.out, scale=0)
```

Proportion Variance Explained

- Assuming that the variables are centered to have mean zero, the total variance is defined as

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

- The variance explained by the mth PC is

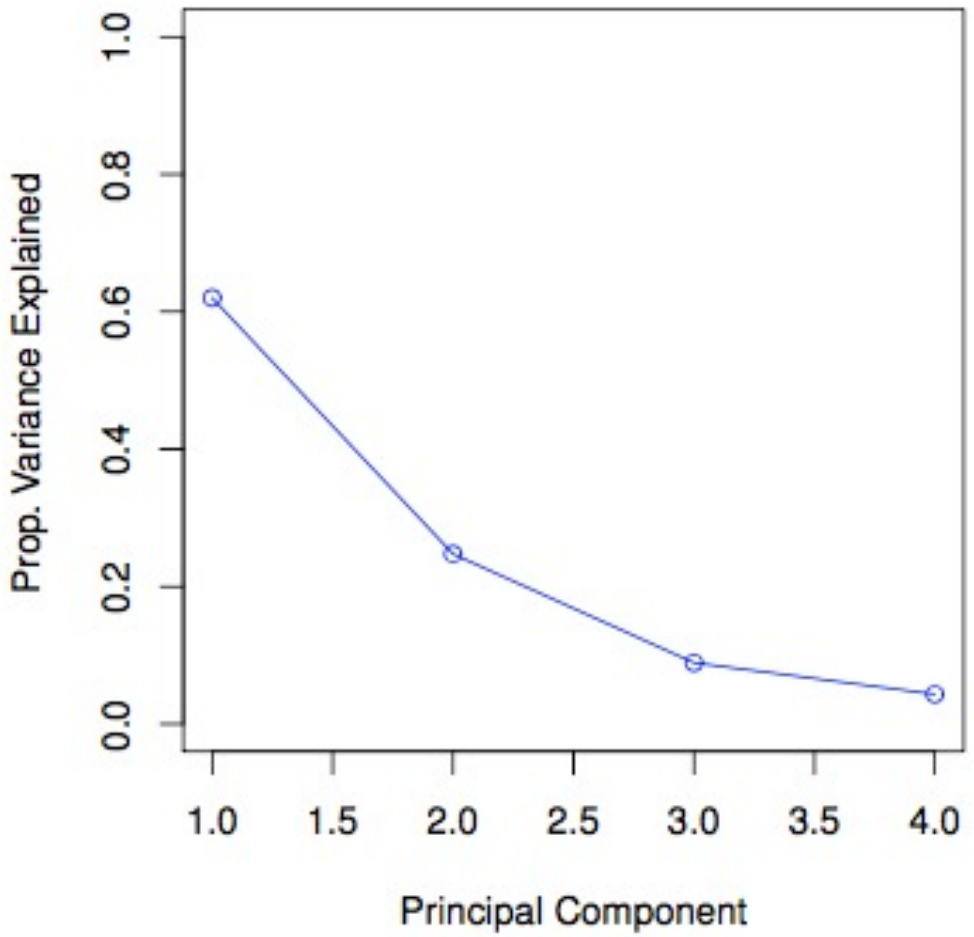
$$Var(U_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$$

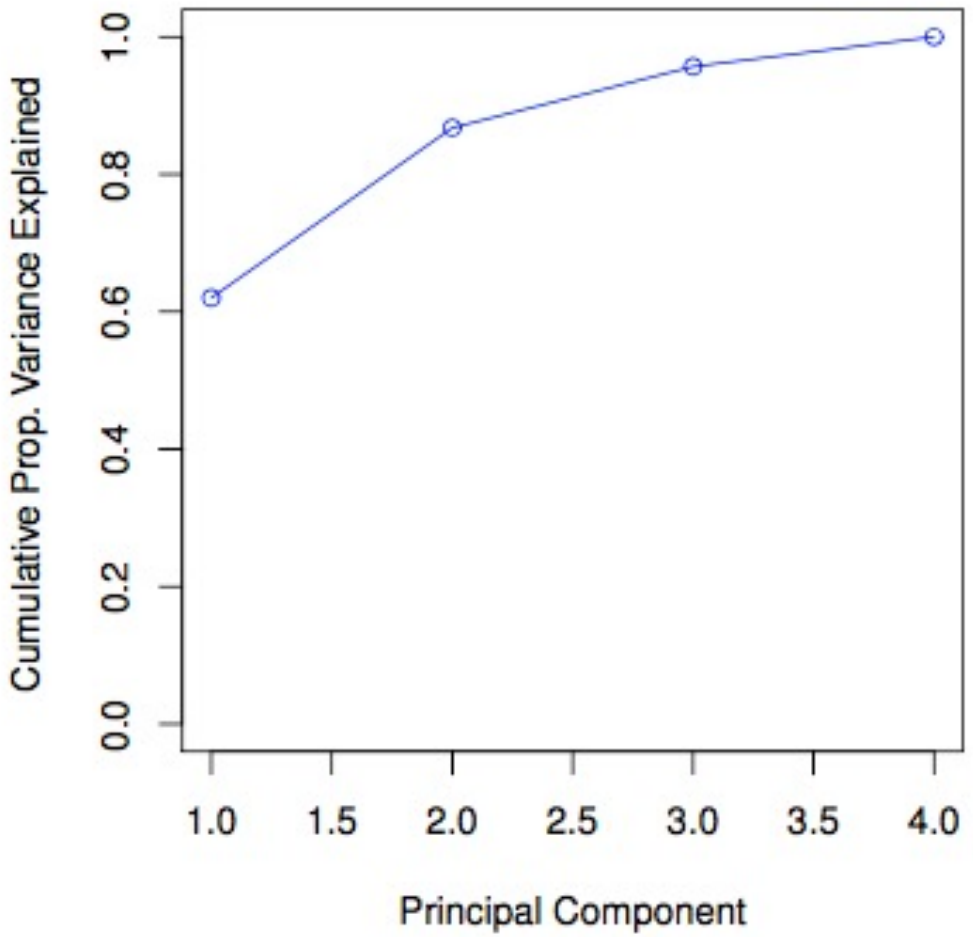
- PVE indicates the strength of each PC, and is defined as

$$PVE_m = \frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

$$\sum_{m=1}^{M} PVE_m = 1 - \frac{\sum_{j=1}^{p} \sum_{i=1}^{n} \left(x_{ij} - \sum_{m=1}^{M} z_{im} \phi_{jm} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2} = 1 - \frac{RSS}{TSS}$$

Scree Plot to Determine # of PCs





Elbow Hunting!

R Commands (Section 12.5)

```
> pr.var <- pr.out$sdev^2
> pr.var
[1] 2.480 0.990 0.357 0.173
```

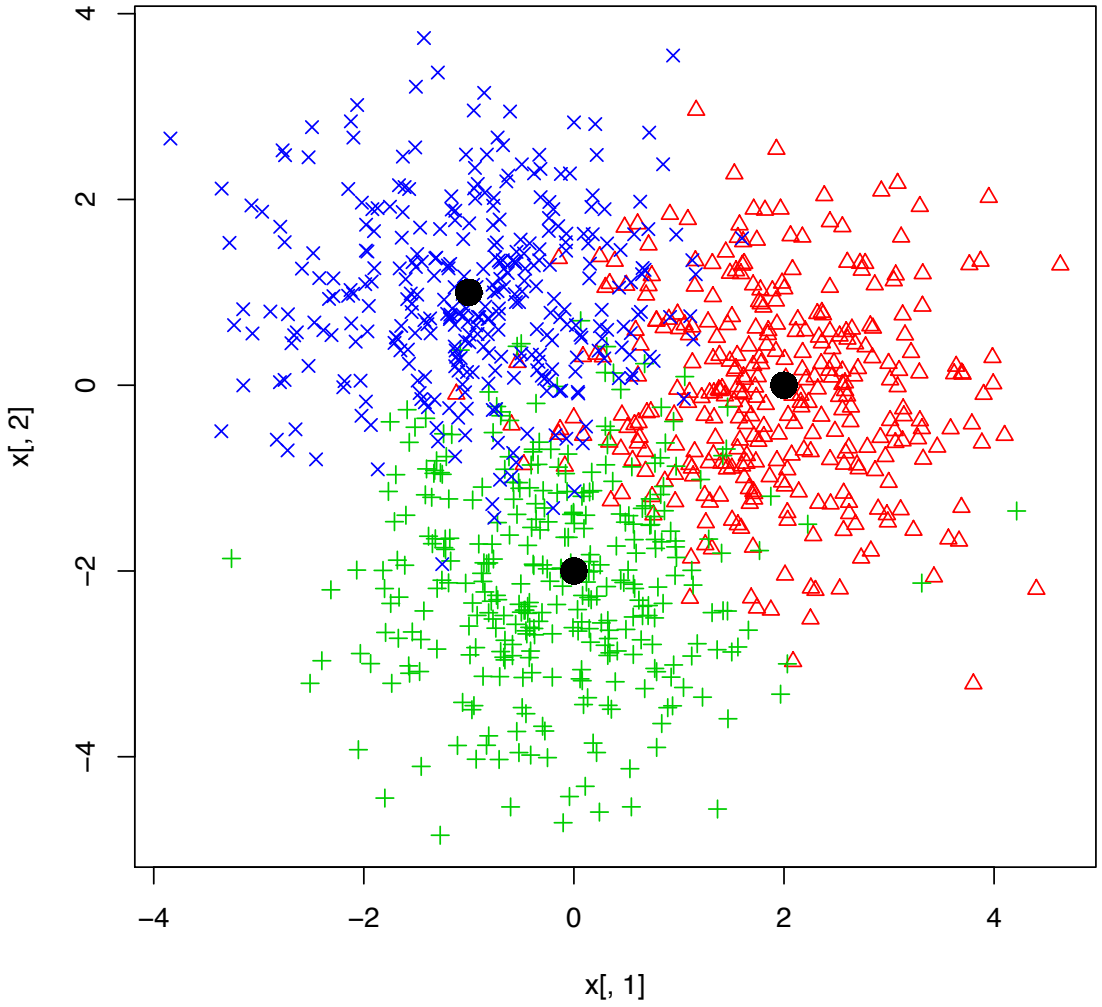
```
> pve <- pr.var / sum(pr.var)
> pve
[1] 0.6201 0.2474 0.0891 0.0434
```

```
> plot(pve, xlab = "Principal Component",
    ylab = "Proportion of Variance Explained", ylim = c(0, 1),
    type = "b")
> plot(cumsum(pve), xlab = "Principal Component",
    ylab = "Cumulative Proportion of Variance Explained",
    ylim = c(0, 1), type = "b")
```

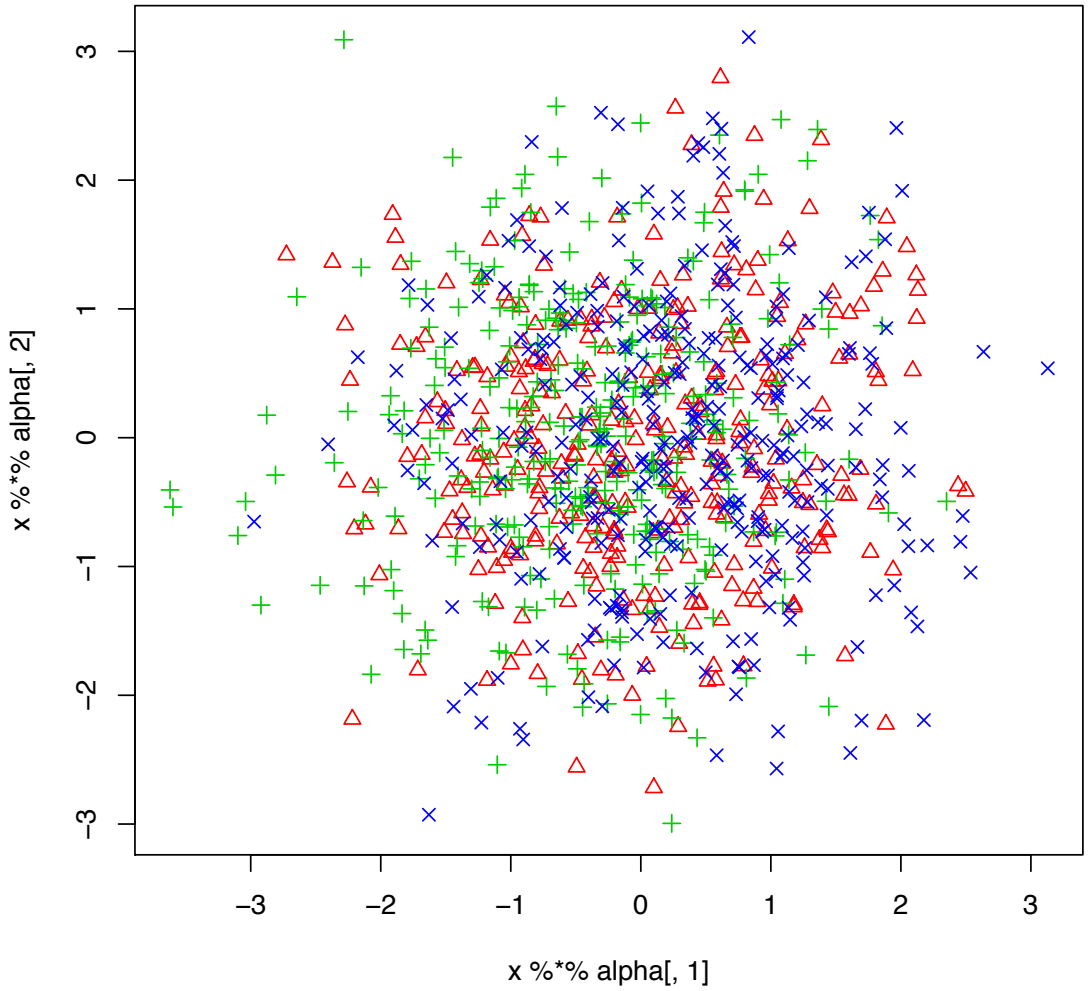
Simulation Example: Three-class Gaussian

- p-dimensional feature space ($p = 128$)
- Class label Y: $P(Y = 1) = P(Y = 2) = P(Y = 3) = \frac{1}{3}$
- Class 1:
  - Gaussian with mean $\theta_1 = [2, 0, 0, \ldots, 0]$
  - $X|Y = 1 \sim N(\theta_1, I)$
- Class 2:
  - Gaussian with mean $\theta_2 = [0, -2, 0, \ldots, 0]$
  - $X|Y = 2 \sim N(\theta_2, I)$
- Class 3:
  - Gaussian with mean $\theta_3 = [-1, 1, 0, \ldots, 0]$
  - $X|Y = 3 \sim N(\theta_3, I)$

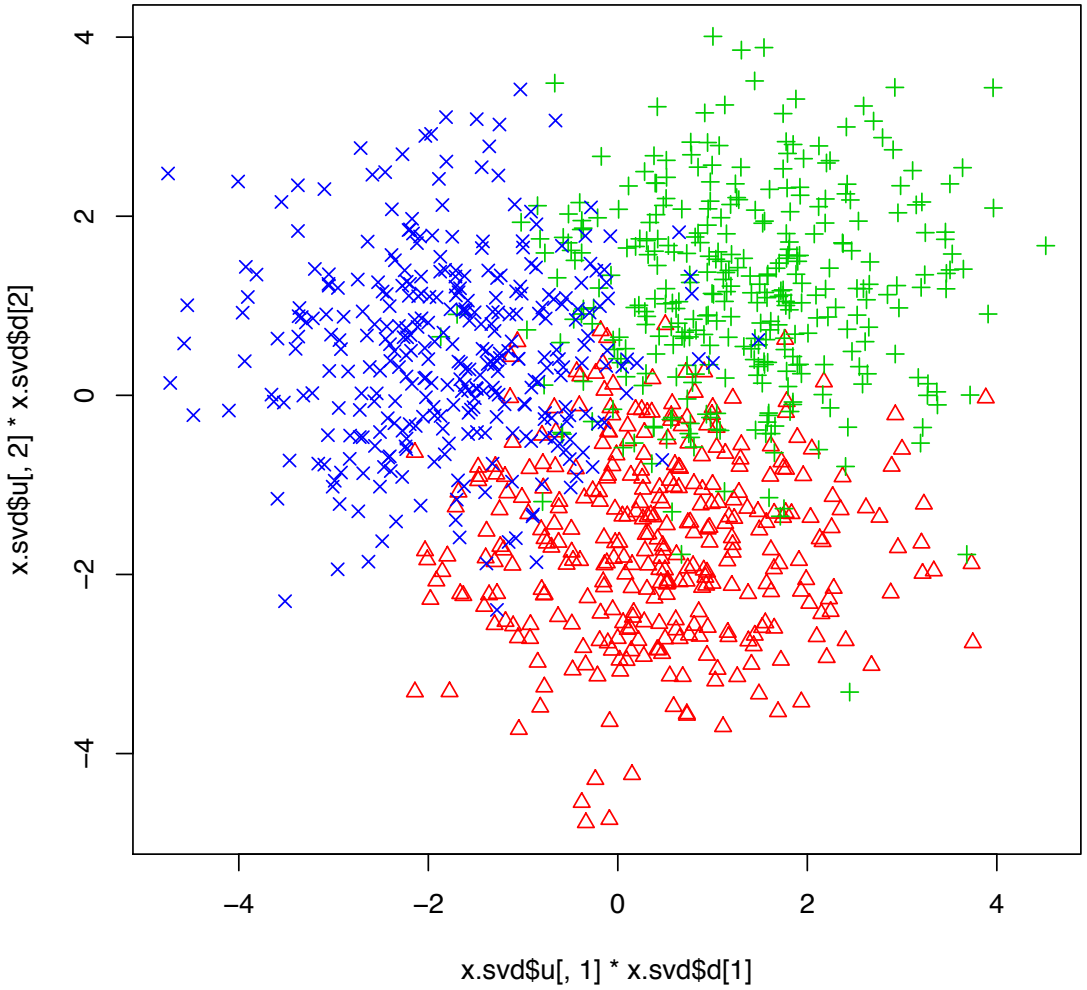
Scatter Plot: Oracle Projection



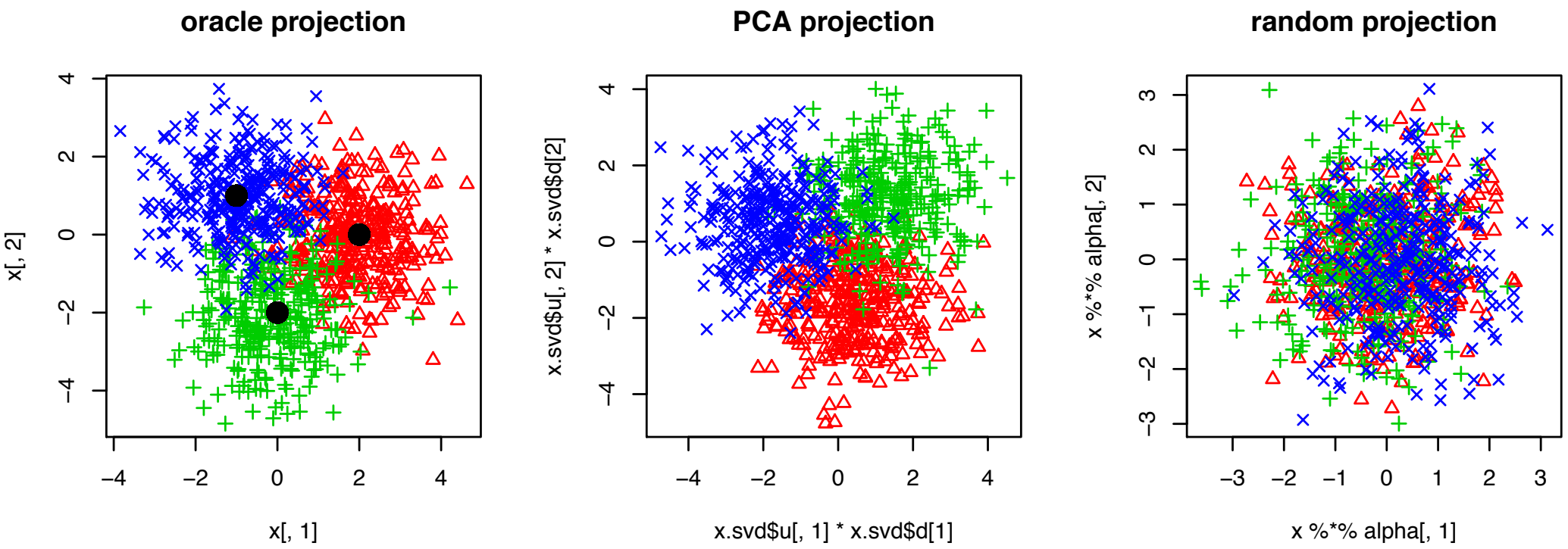
Scatter Plot: Random Projection



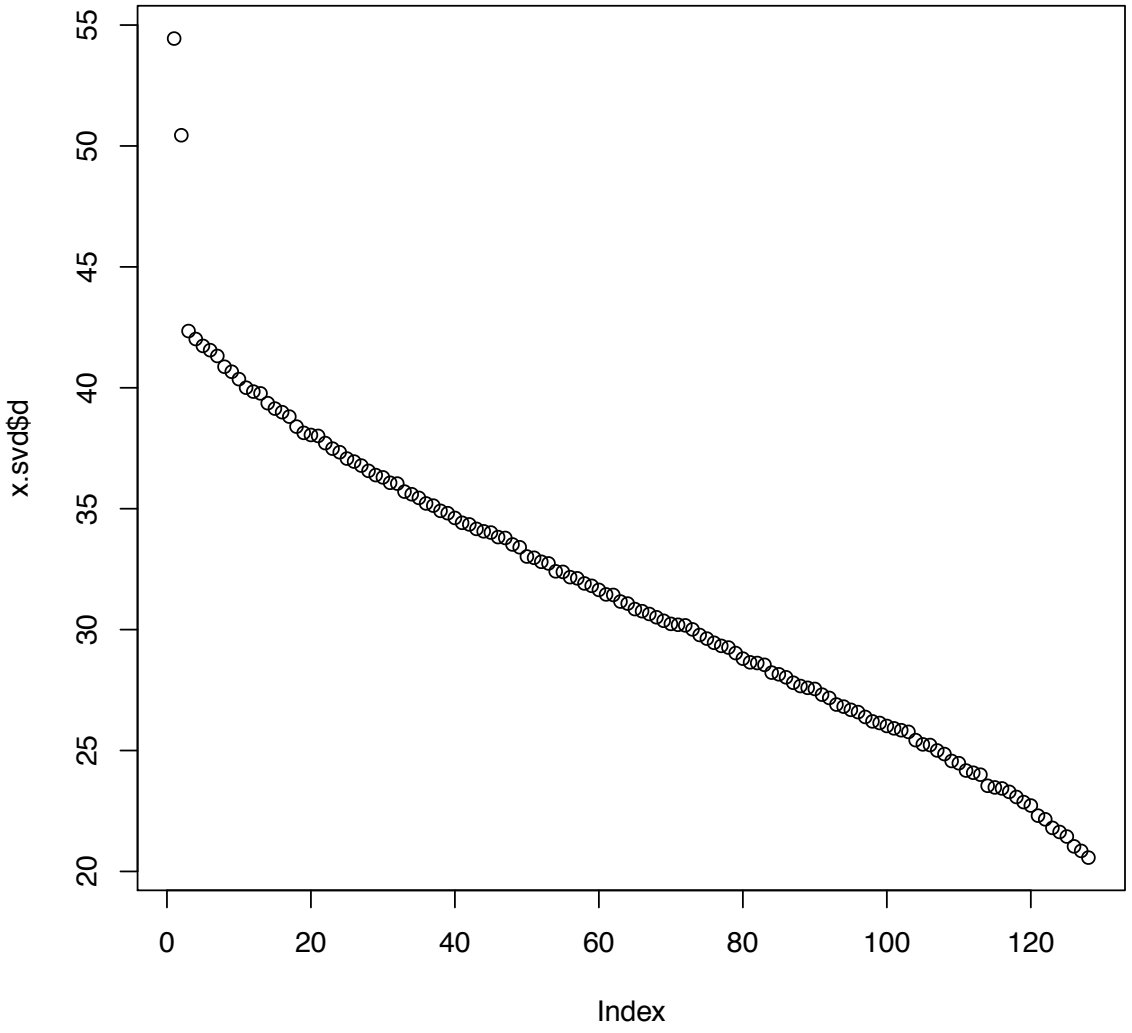
Scatter Plot: PCA Projection



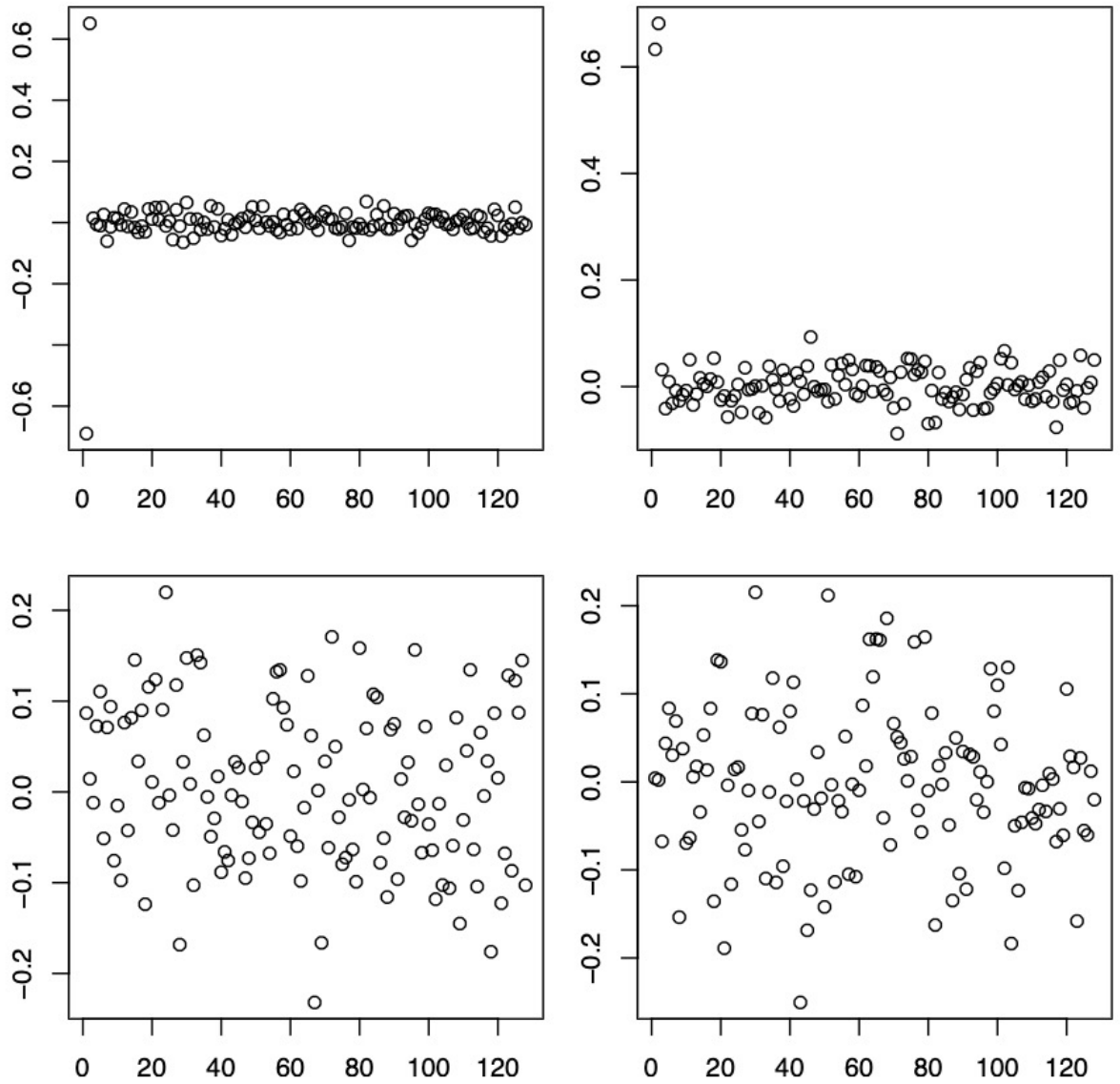
Scatter Plot: All Projection



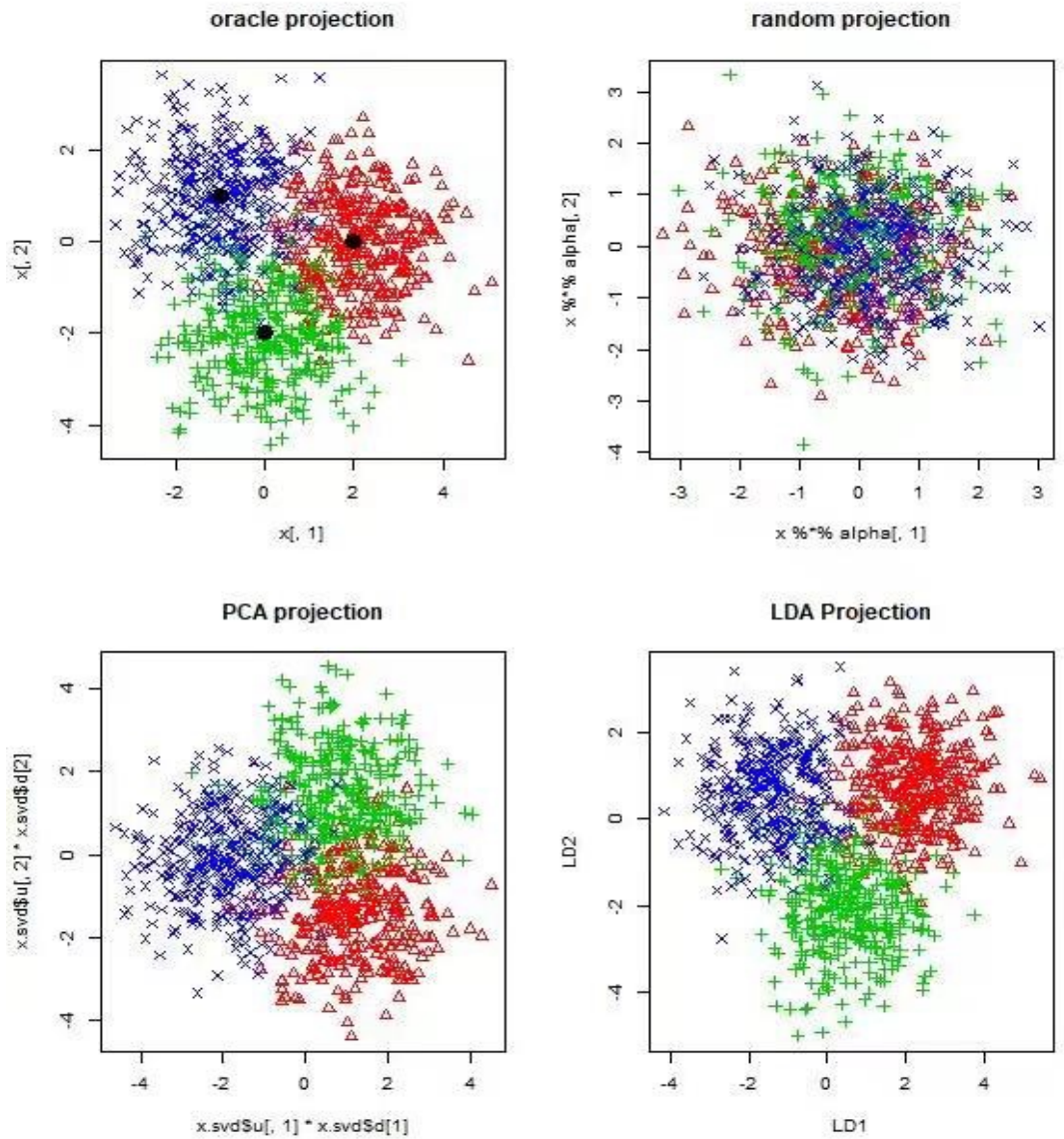
PCA: Scree Plot



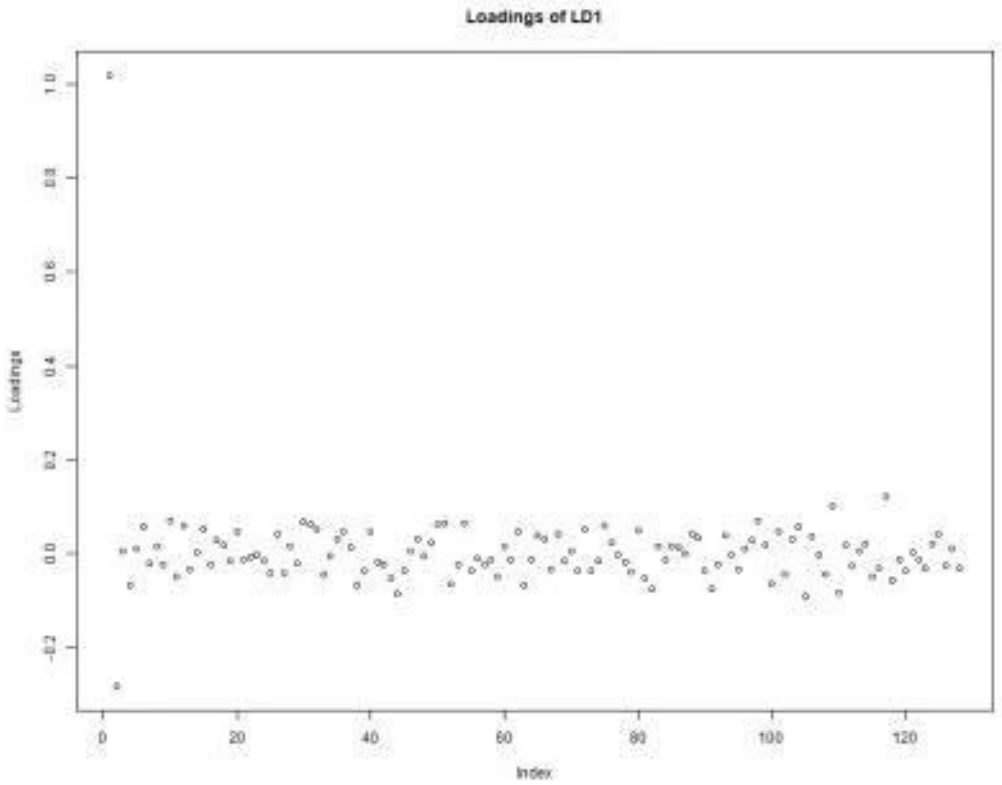
PC Loadings

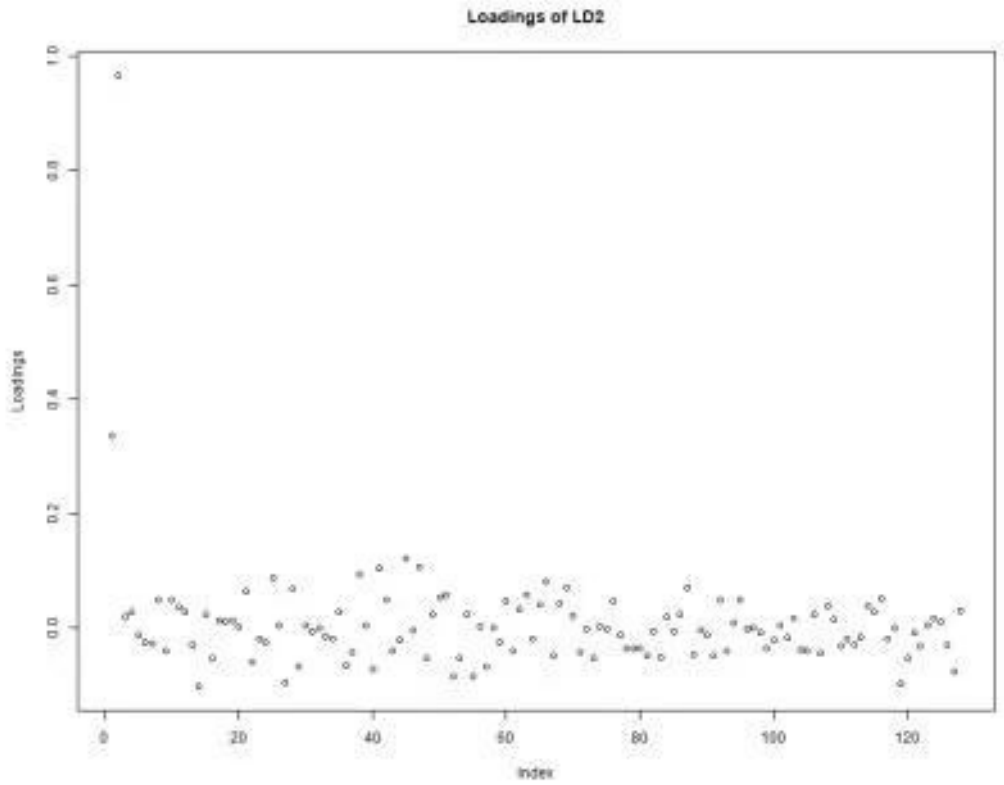

0.6
0.4
0.2
−0.2
−0.6
0.6
0.4
0.2
0.0
0.2
0.1
0.0
−0.1
−0.2
0.2
0.1
0.0
−0.1
−0.2
0
20
40
60
80
100
120

PCA, LDA

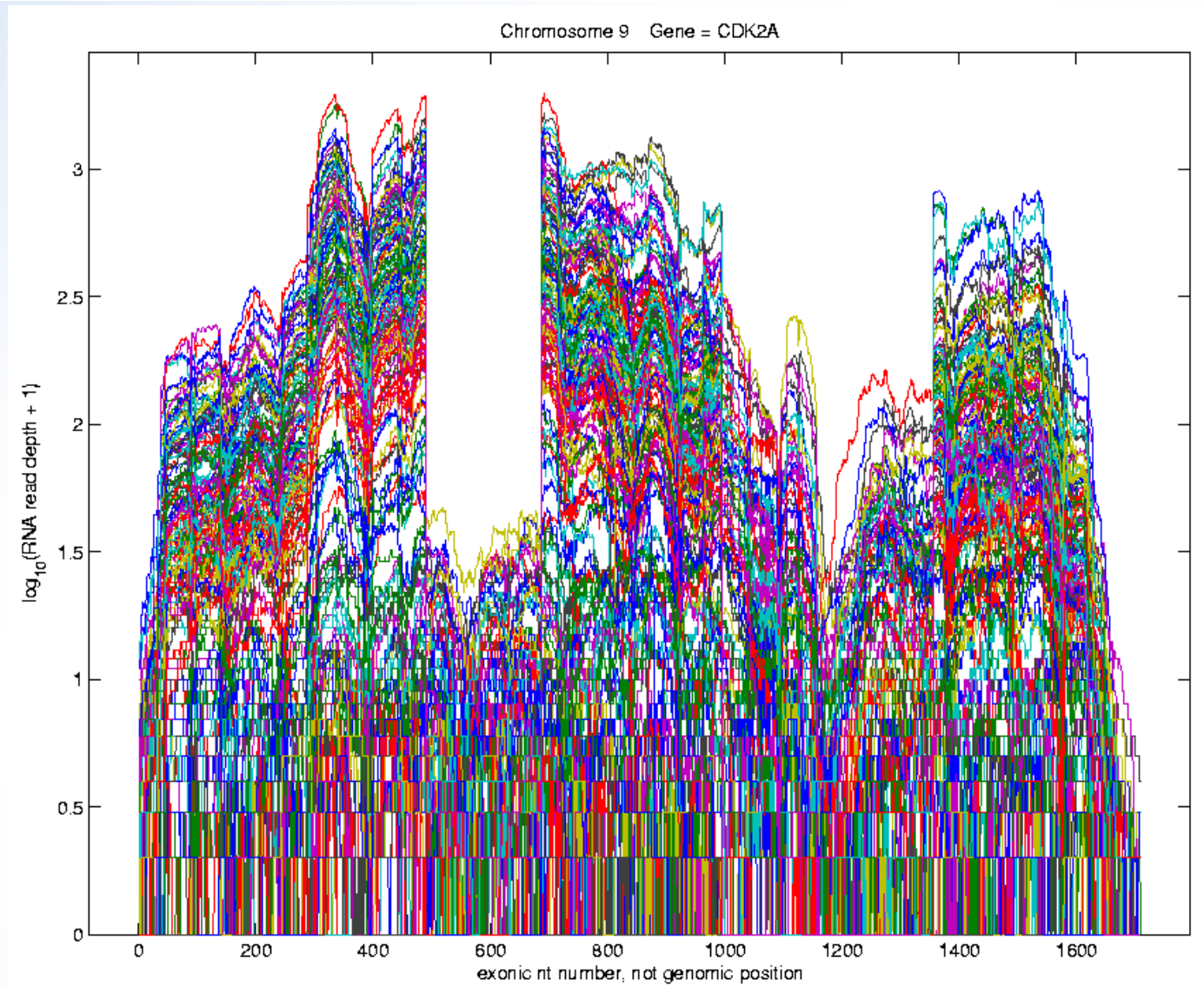

oracle projection
x[, 2]
x[, 1]
random projection
x %*% alpha[, 2]
x %*% alpha[, 1]
PCA projection
x.svd$u[, 2] * x.svd$d[2]
x.svd$u[, 1] * x.svd$d[1]
LDA Projection
LD2
LD1

LDA Loadings

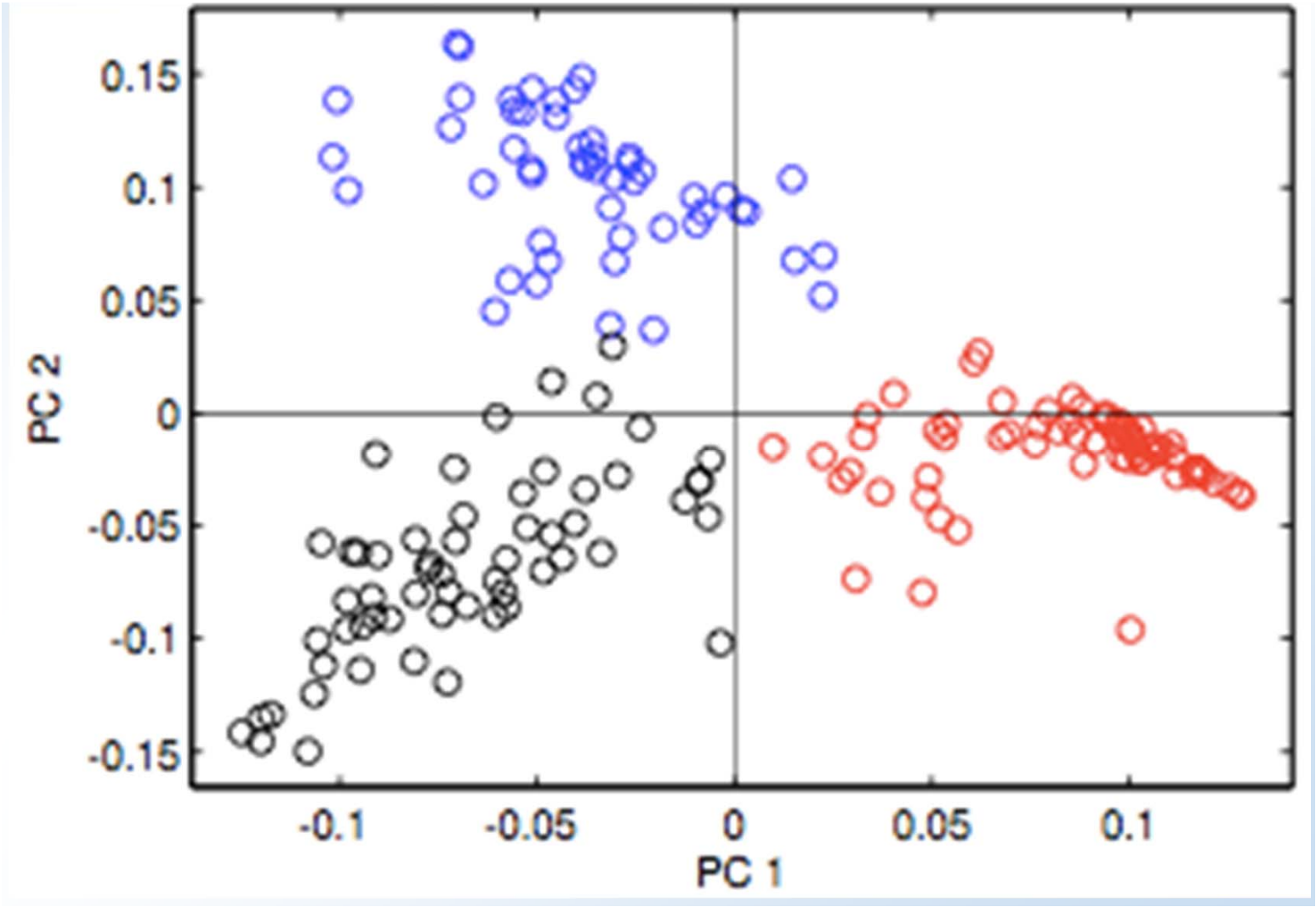




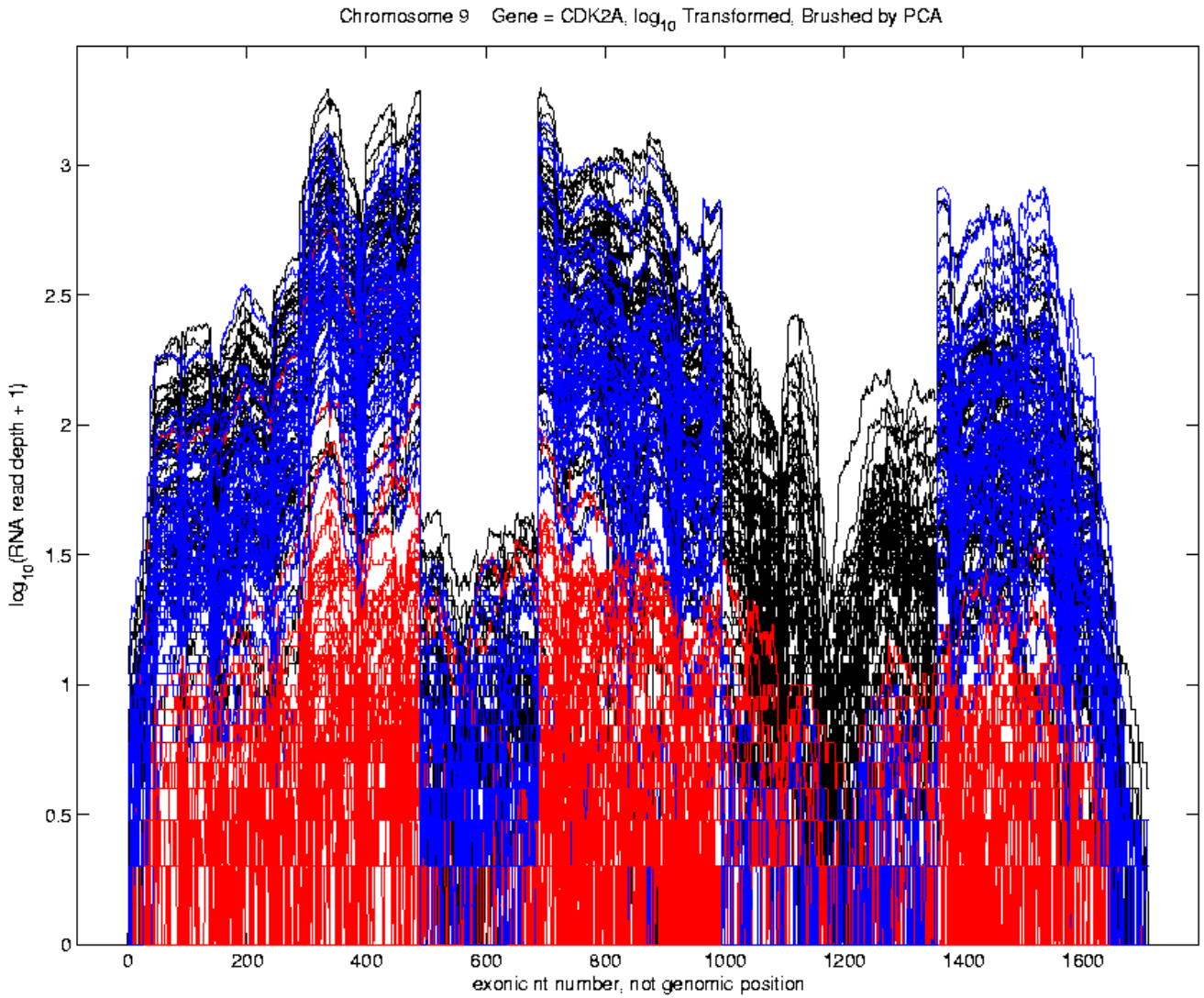
Gene CDK2A RNASeq: 180 Samples, 1700 Locations



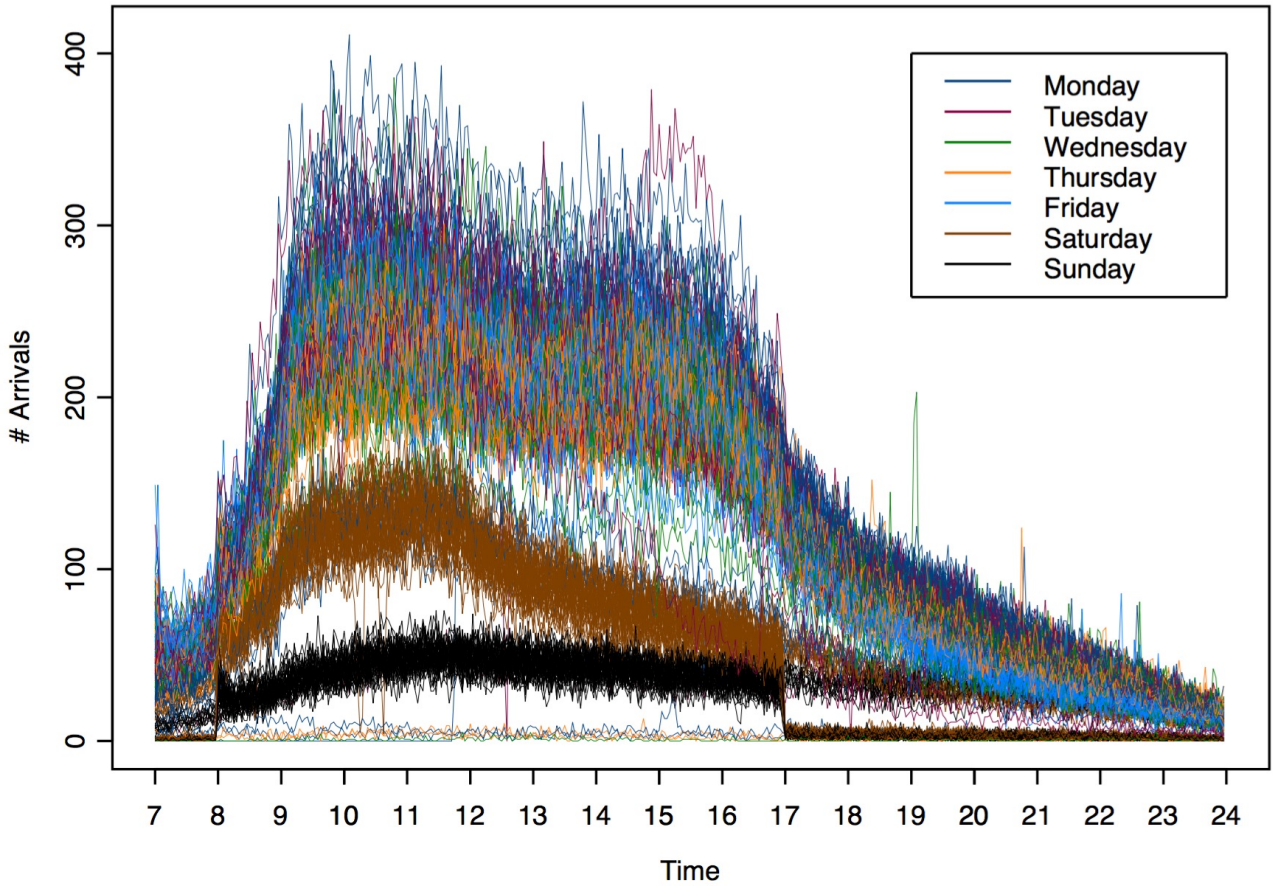
Clustering on the PCA Scores



Back to the Gene Profiles

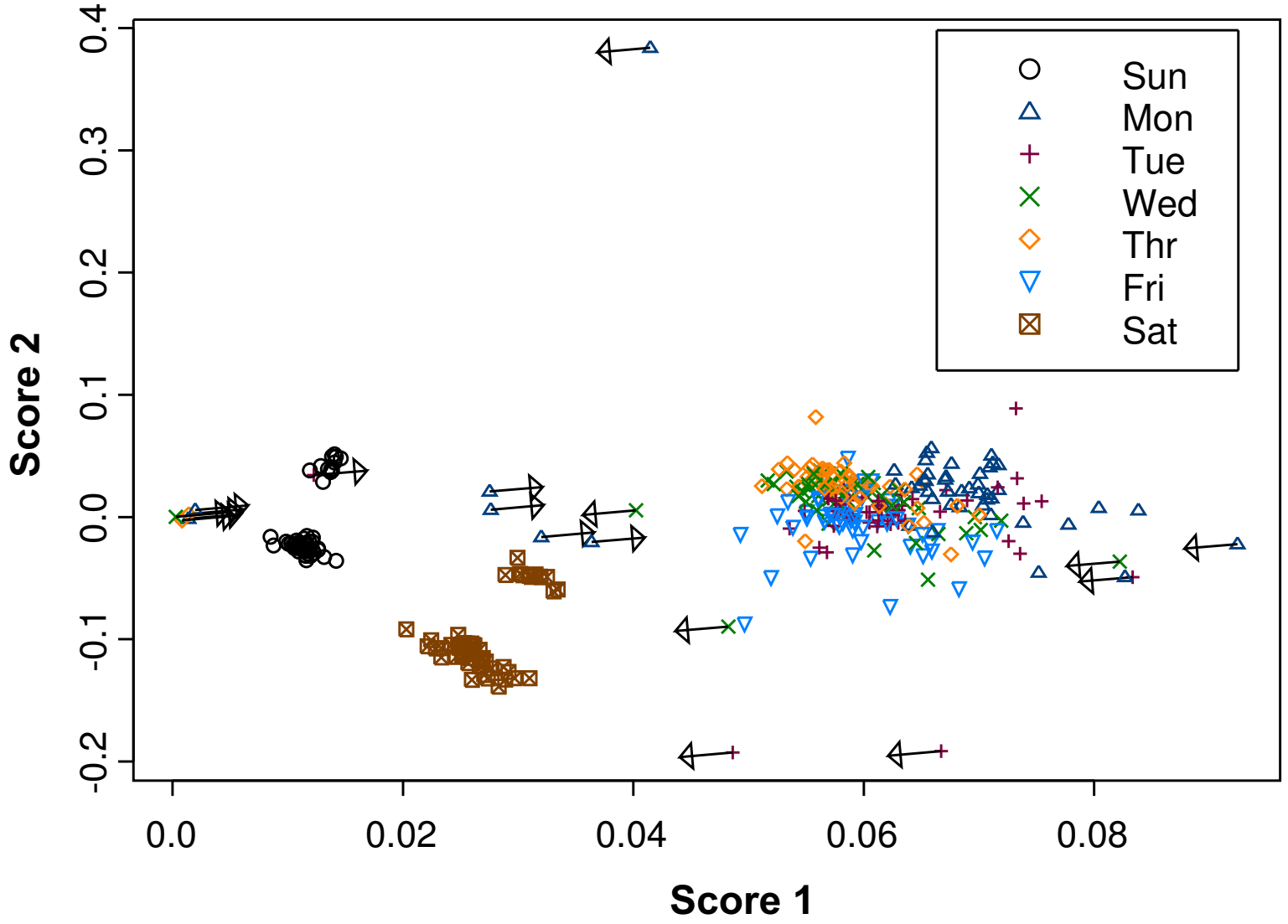


Call Center Arrivals



- Bank with a network of 4 call centers in northeast US
- 300K calls / day, 60K / day seeking agents
- 2.5-minute interval counts from 7am to 24am.

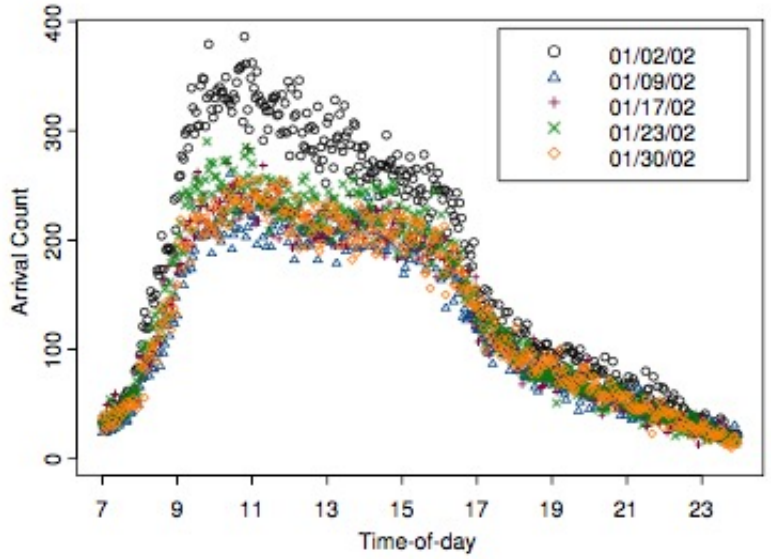
Scatter Plot of PC1 and PC2

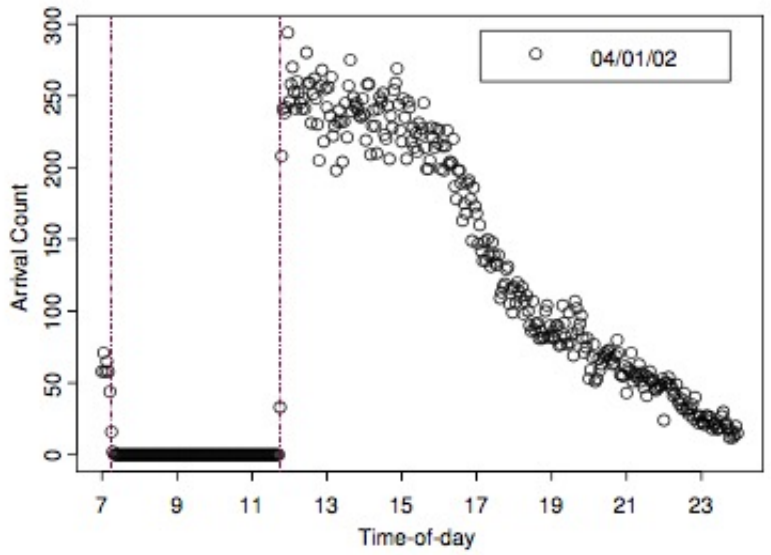


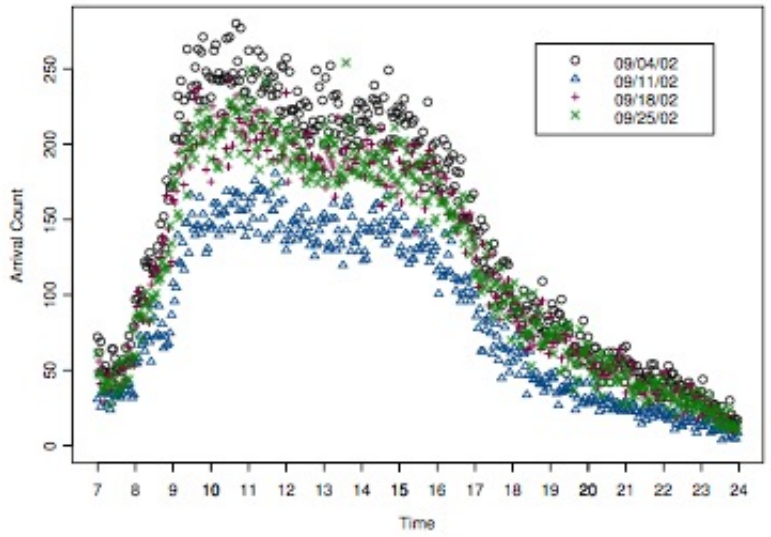
18 Anomalies

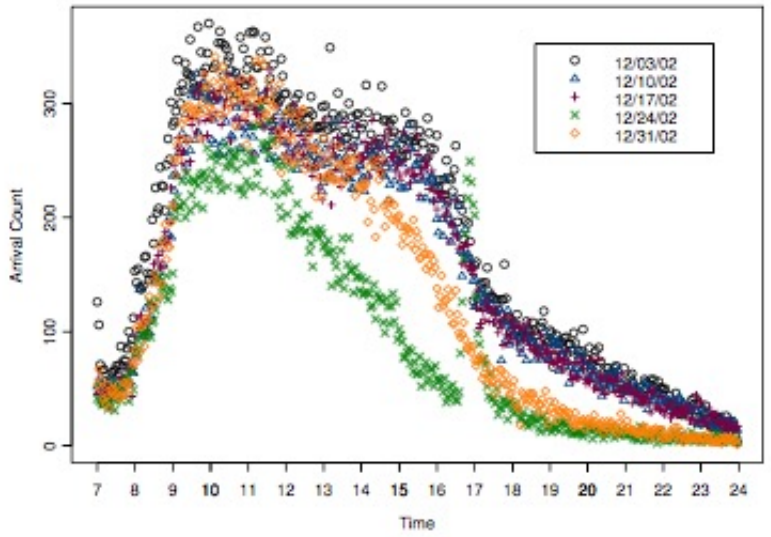
Reason	Date	
Holiday	Jan 1	New Year's Day
	Jan 21	MLK Day
	Feb 18	Washington's Birthday
	May 27	Memorial Day
	Jul 4	Independence Day
	Sep 2	Labor Day
	Oct 14	Columbus Day
	Nov 11	Veterans' Day
	Nov 28	Thanksgiving Day
	Dec 25	Christmas Day
Holiday related	Jan 2	After New Year's Day
	Dec 2/3	After Thanksgiving
	Dec 24	Christmas Eve
	Dec 31	New Year's Eve
System related/other	Apr 1	System error
	Jul 24	System error
	Dec 24	System error
	Sep 11	"Emotional" day

Several Anomalous Days

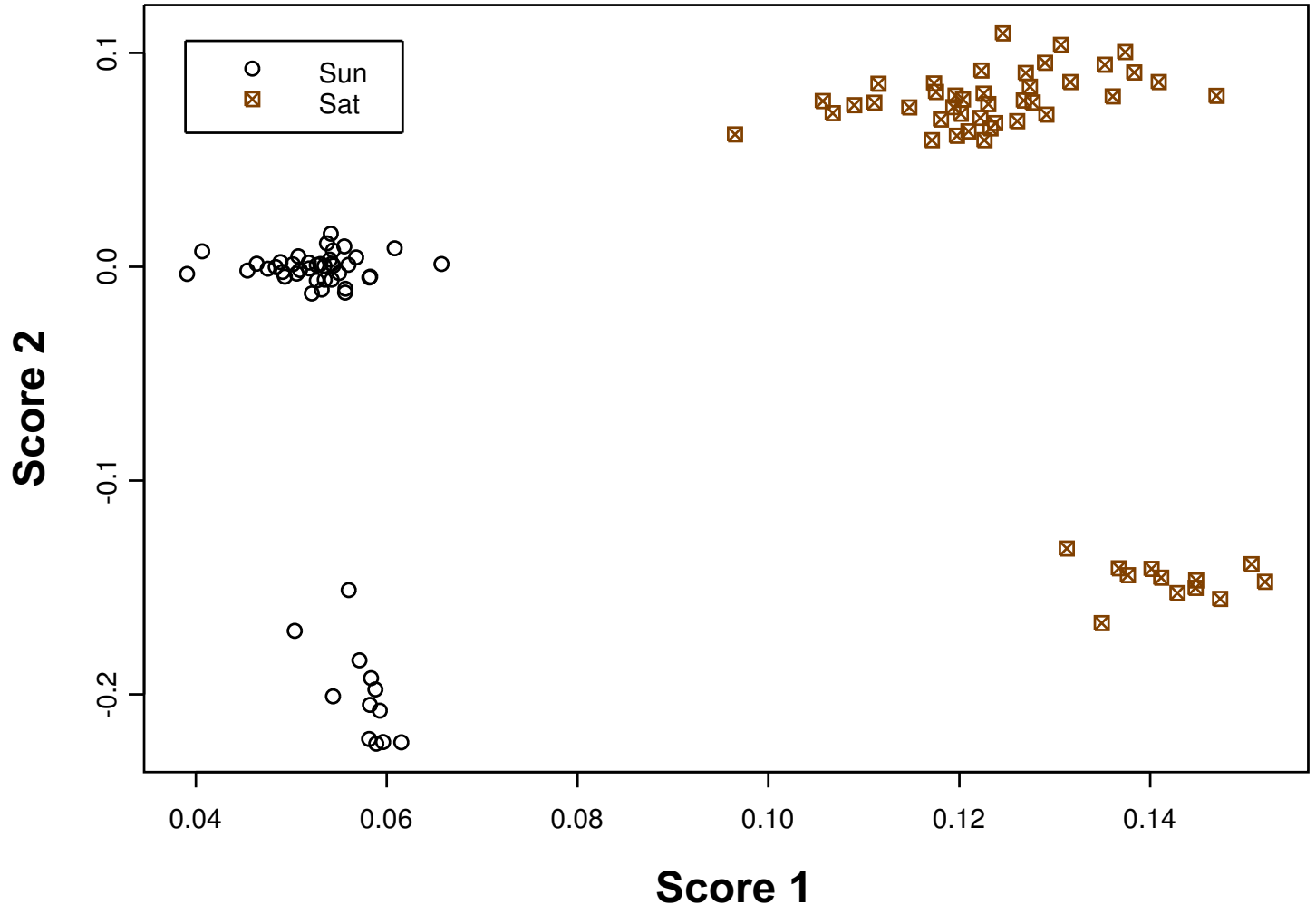




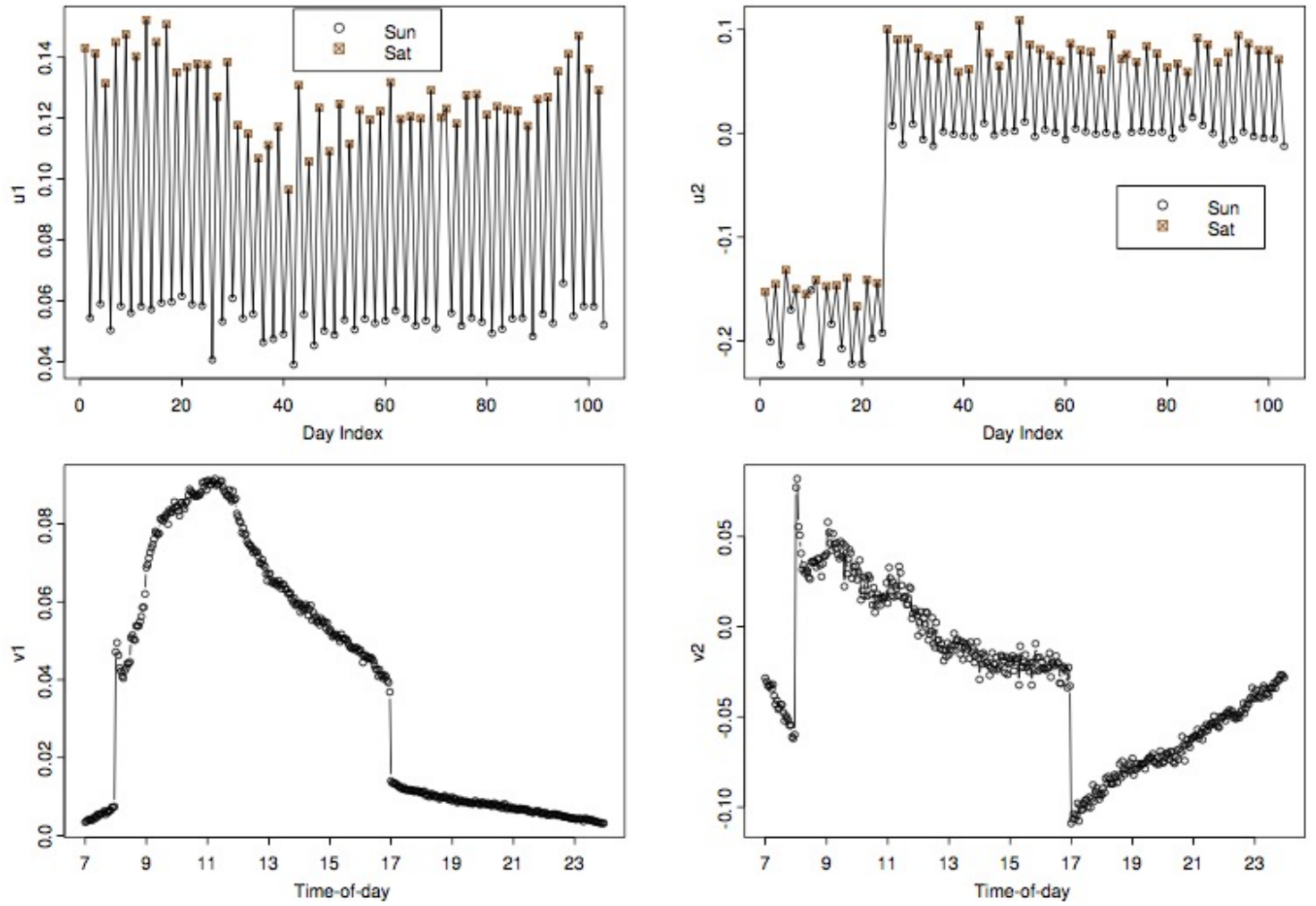




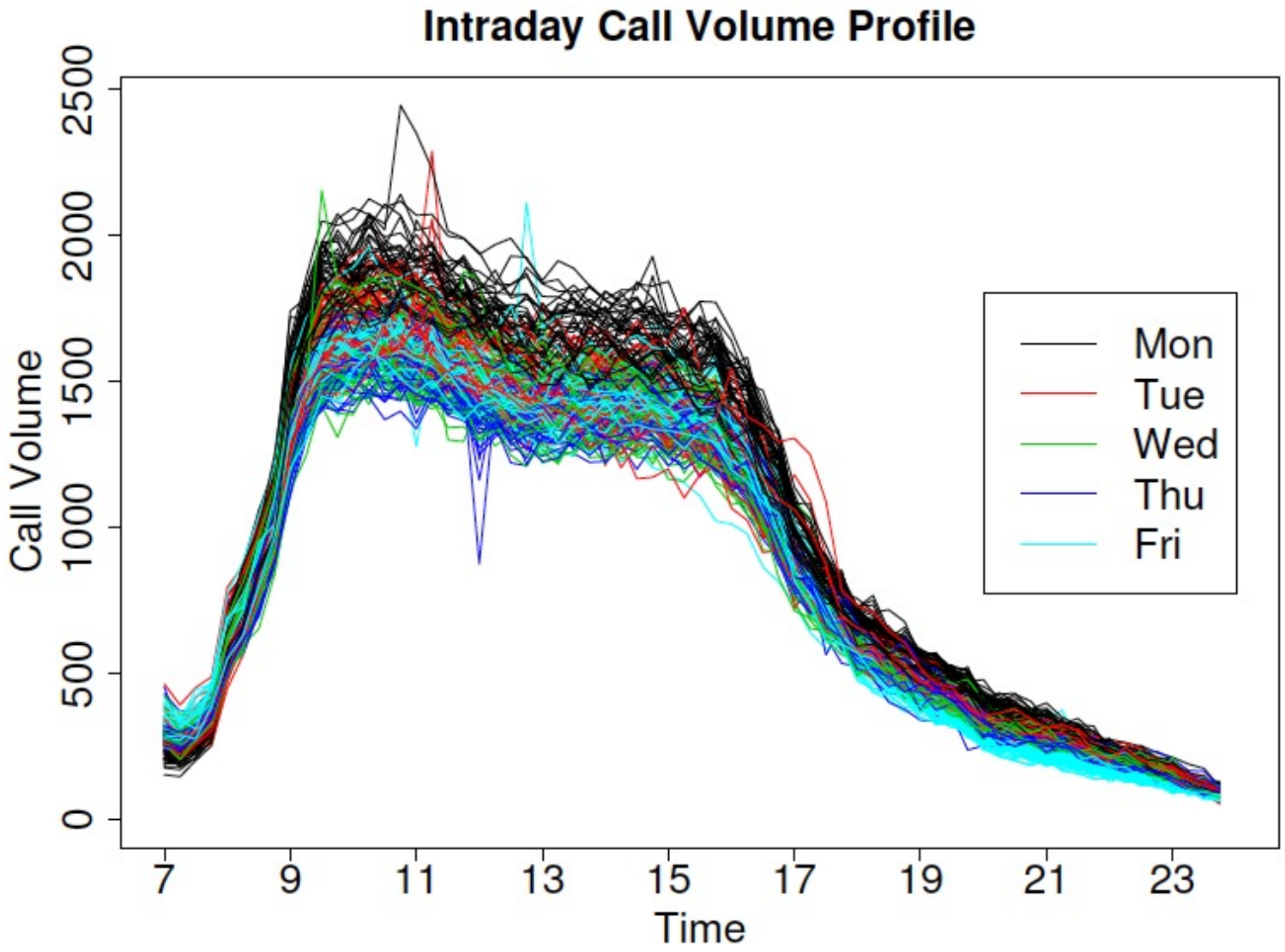
Two Saturday/Sunday Clusters (Based on PC Scores)



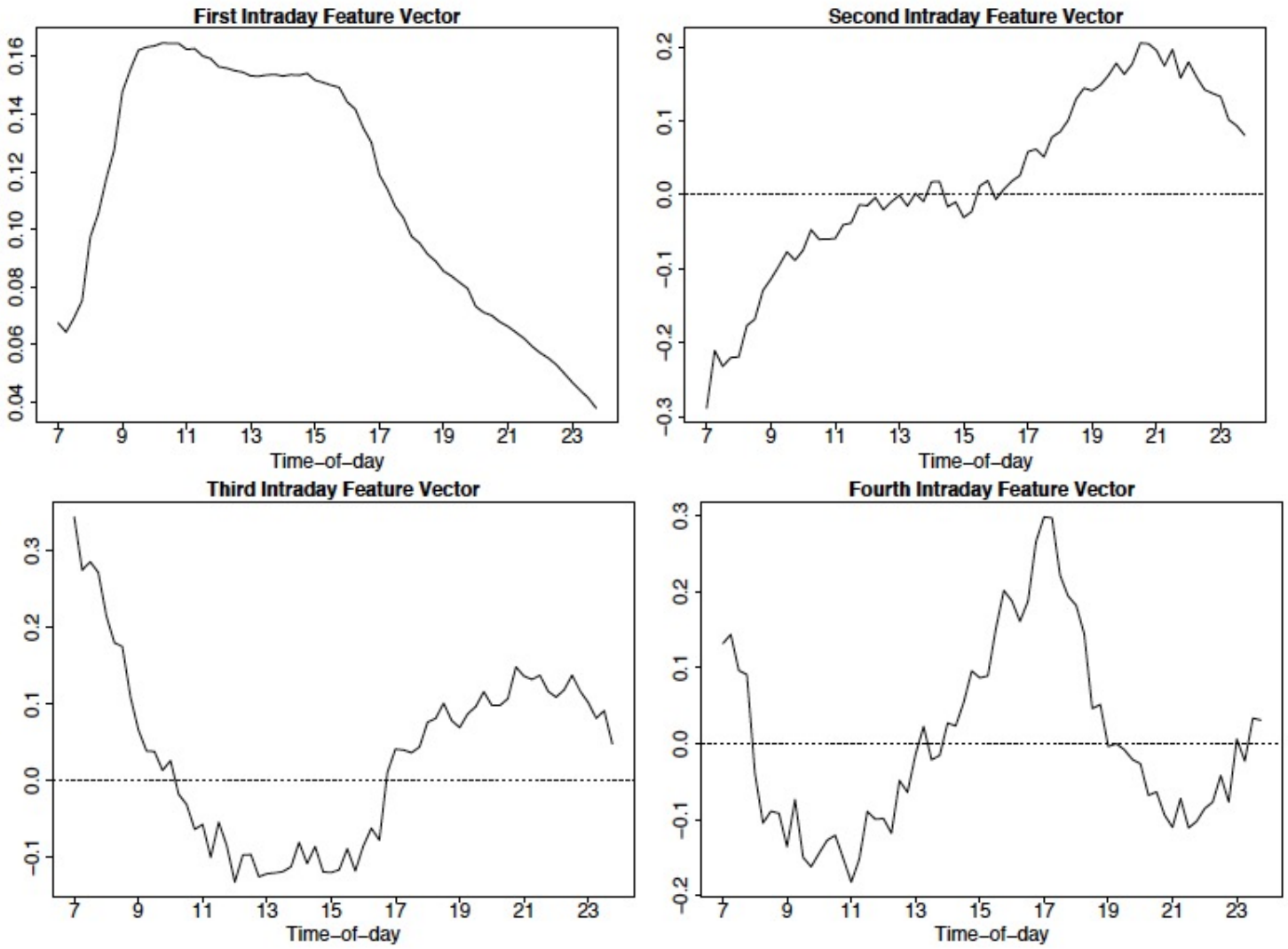
Weekend Days: Plots of PC Score (U) and Loading Vector (V)



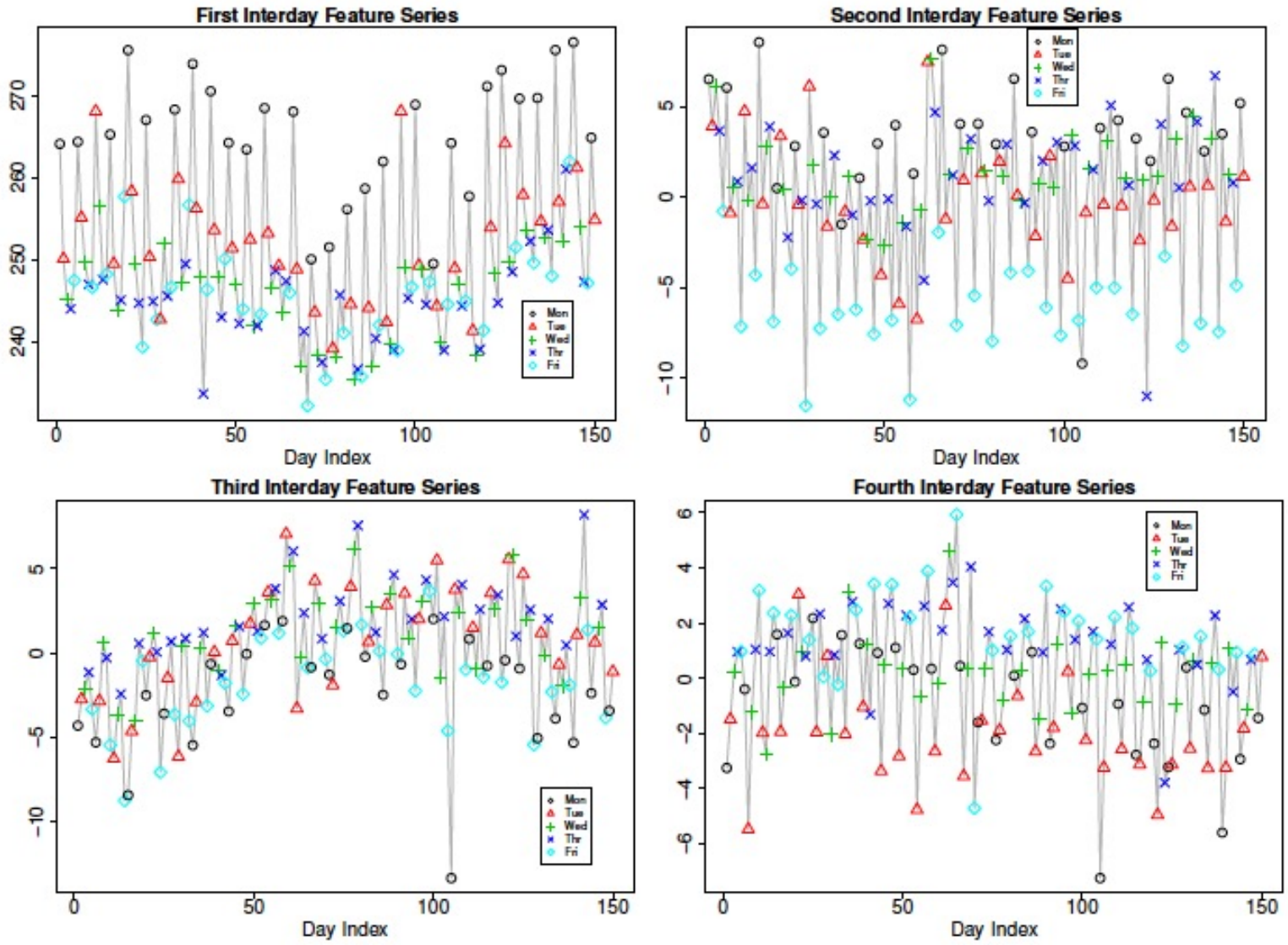
Weekday Call Center Arrivals (After Cleaning)



Call Center Arrival: PC Loading Curve, i.e. Intraday Feature Vector



Call Center Arrival Data: PC Scores

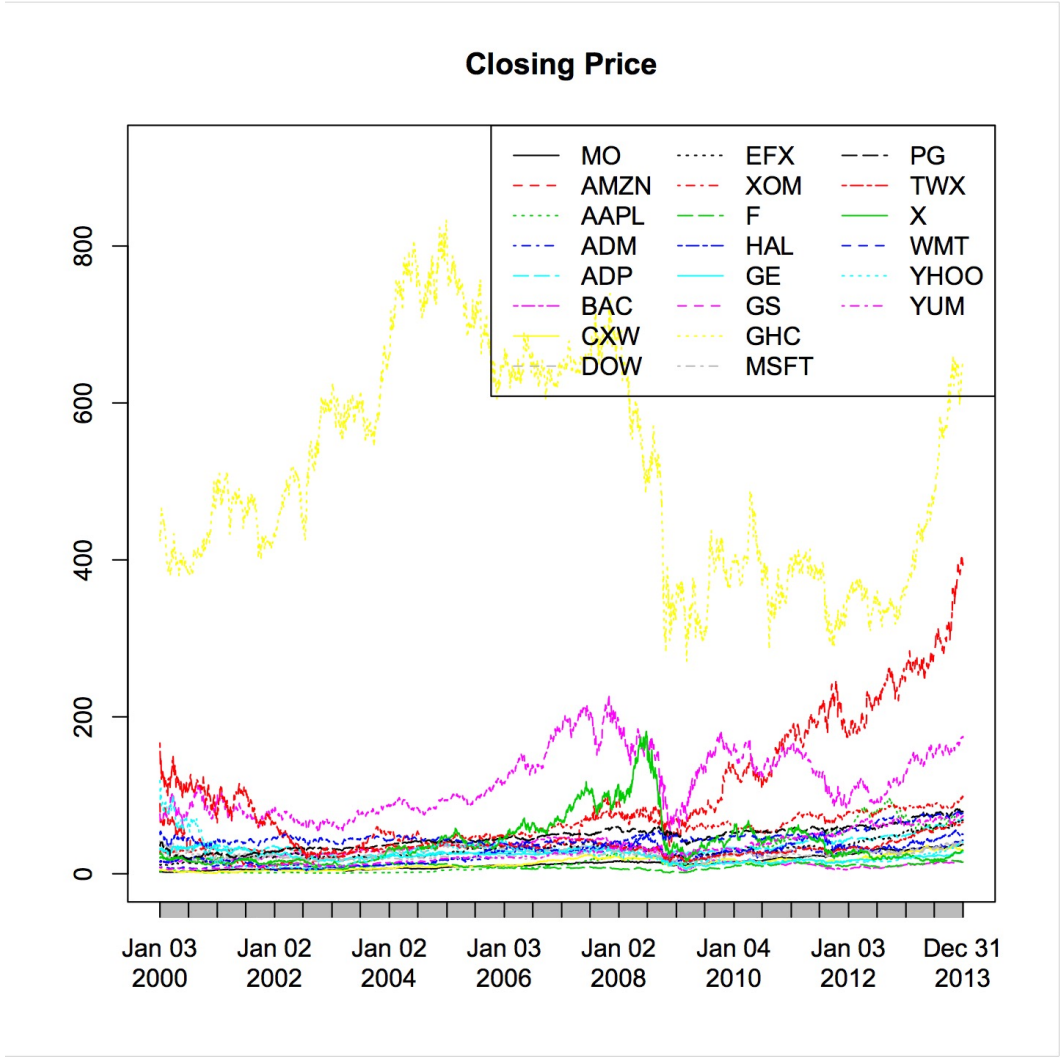


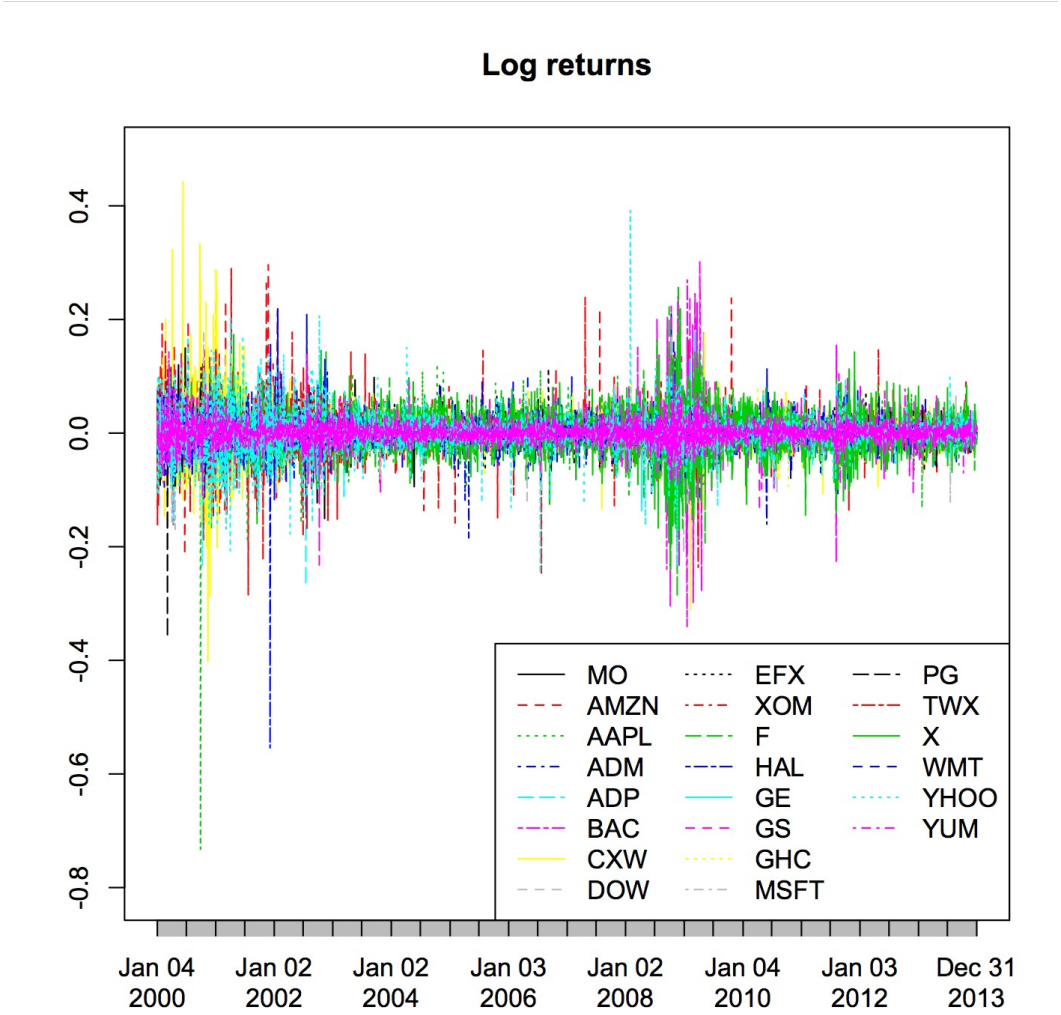
Stock Daily Closing Price 2000-2013

Altria (formerly Philip Morris; MO)
Apple (AAPL)
Automatic Data Processing (ADP)
Corrections Corporation of America (CXW)
Equifax (EFX)
Ford (F)
General Electric (GE)
Graham Holding Companies (GHC)
Proctor and Gamble (PG)
United States Steel (X)
Yahoo! (YHOO)

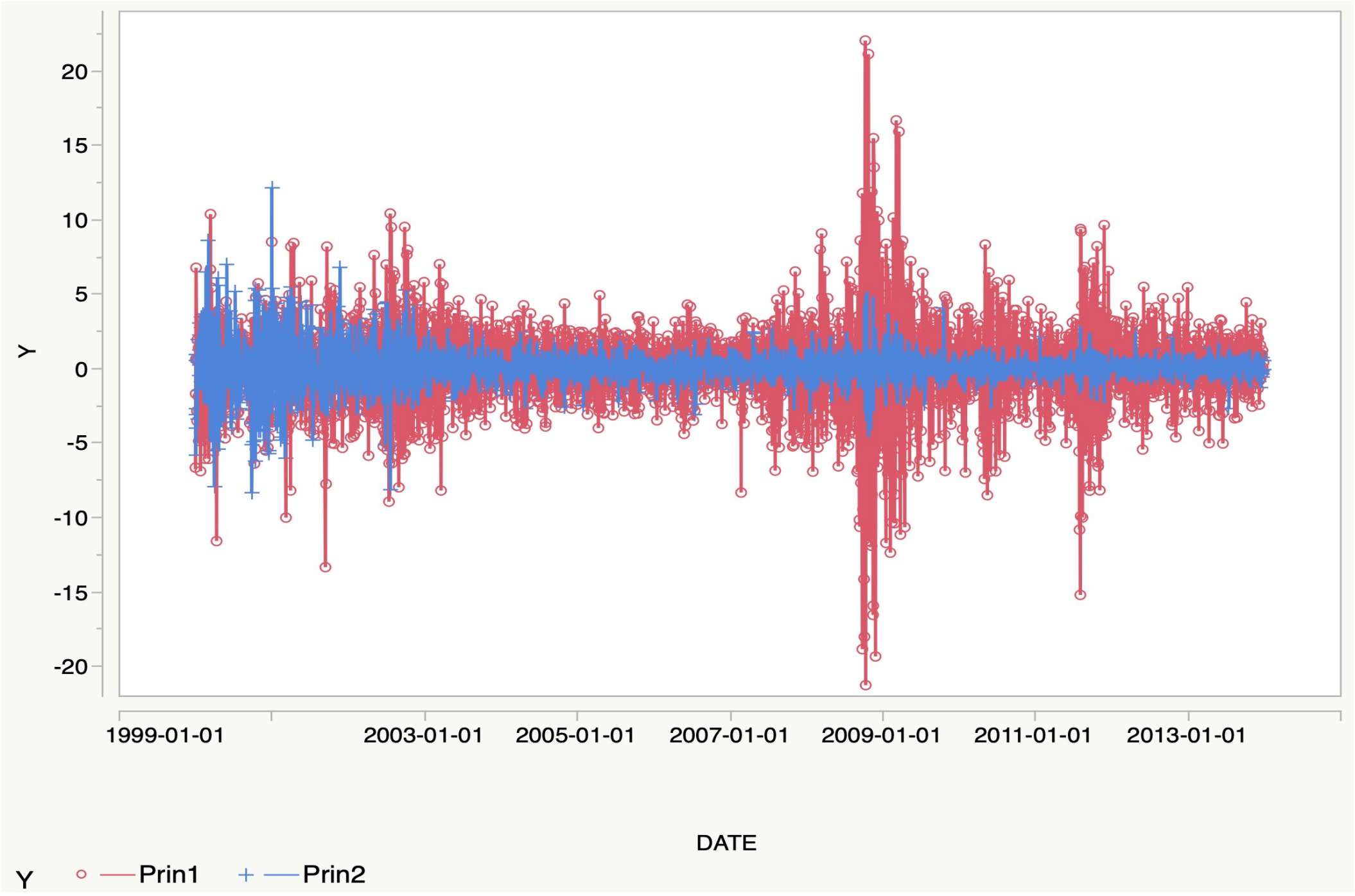
Amazon (AMZN)
Archer Daniels Midland (ADM)
Bank of America (BAC)
Dow Chemicals (DOW)
ExxonMobil (XOM)
Halliburton (HAL)
Goldman Sachs (GS)
Microsoft (MSFT)
Time Warner (TWX)
Walmart (WMT)
Yum! Brands (YUM)

Closing Price and Returns

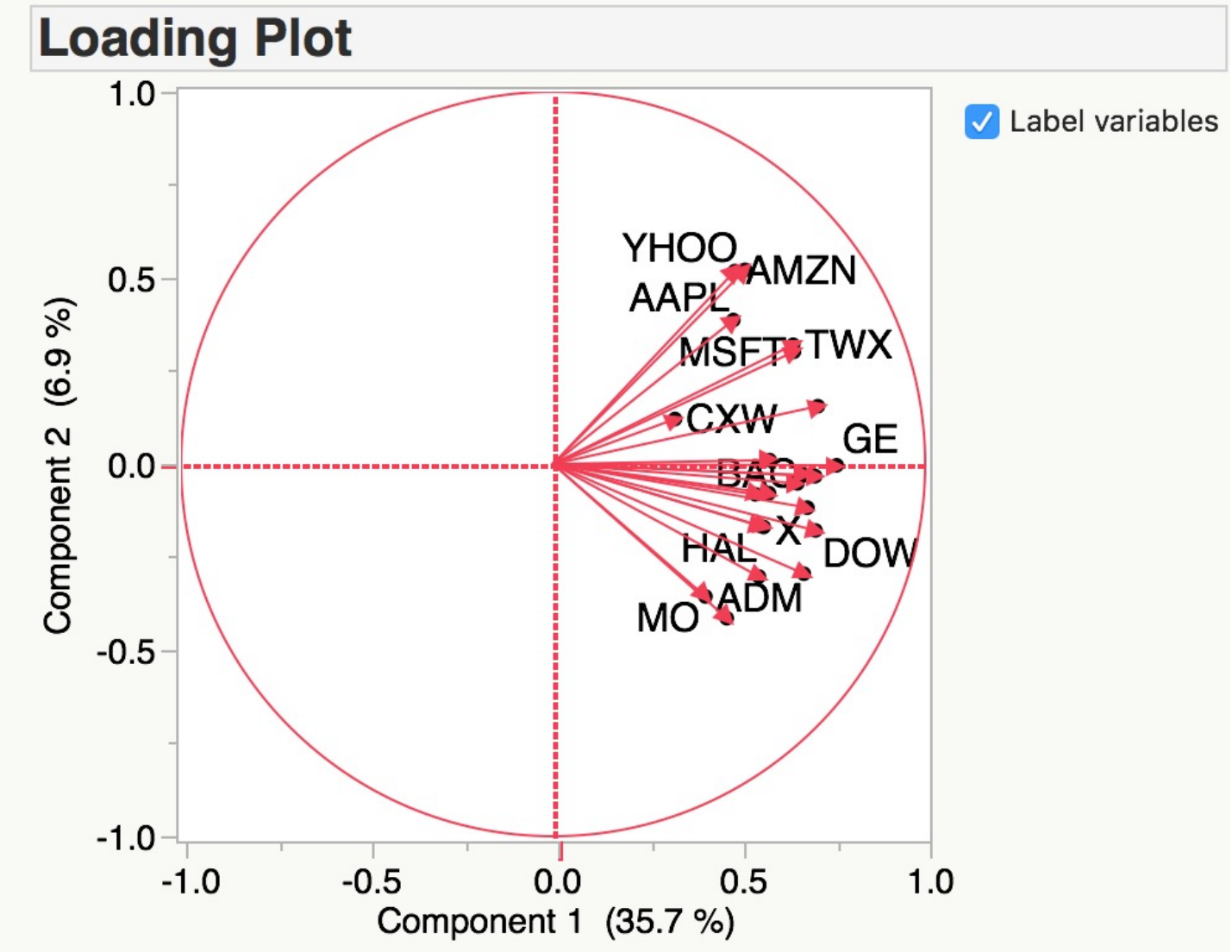




adjusted closing price = adjusted for dividend payments and stock splits


20
15
10
5
0
-5
-10
-15
-20
Y
1999-01-01
2003-01-01
2005-01-01
2007-01-01
2009-01-01
2011-01-01
2013-01-01
DATE
Y
Prin1
Prin2

PCA of Log-Return



Clustering after PCA

Amazon (AMZN)
Apple (AAPL)
Corrections Corporation of America (CXW)
Goldman Sachs (GS)
Microsoft (MSFT)
Time Warner (TWX)
Yahoo! (YHOO)

United States Steel (X)
Dow Chemicals (DOW)
ExxonMobil (XOM)
Halliburton (HAL)
Equifax (EFX)
Ford (F)
Archer Daniels Midland (ADM)
Graham Holding Companies (GHC)
General Electric (GE)
Bank of America (BAC)

Altria (formerly Philip Morris; MO)
Proctor and Gamble (PG)
Walmart (WMT)
Yum! Brands (YUM)
Automatic Data Processing (ADP)

Usage of PCA

- Data visualization
- EDA
- Clustering
- Dimension reduction for regression / classification
- Missing data imputation

Model Selection

- Subset Selection (Section 6.1)
  - Best subset, forward, backward, hybrid
  - AIC, BIC, Cp, Adjusted R2
- Shrinkage (Section 6.2)
  - Ridge, LASSO, Elastic Net
- Dimension Reduction (Section 6.3)
  - Principal Components Regression
  - Partial Least Squares

Principal Components Regression

$$y_i = \beta_0 + \sum \beta_j x_{ij} + \epsilon_i$$

Usual least squares may be inappropriate if

- x is high dimensional, especially in comparison to sample size
- x is highly correlated
- main focus is on future prediction

In these cases, we would like to find some way to reduce the amount of covariate informaion

- Variable selection procedures
- Principal components regression

Principal Components Regression

$$y_i = \beta_0 + \sum \beta_j x_{ij} + \epsilon_i$$

Instead, consider $\mathbf{u}_j$ the jth principal component of $\mathbf{x}$.

Then model

$$y_i = \beta'_0 + \sum_{j=1}^{p'} \beta'_j u_{ij} + \epsilon_i$$

for some $p' < p$.

Principal Components Regression

$$y_i = \beta'_0 + \sum_{j=1}^{p'} \beta'_j u_{ij} + \epsilon_i$$

Advantages:

- u_{ij} are uncorrelated – stability of estimates
- dimension reduction
- stable variable selection

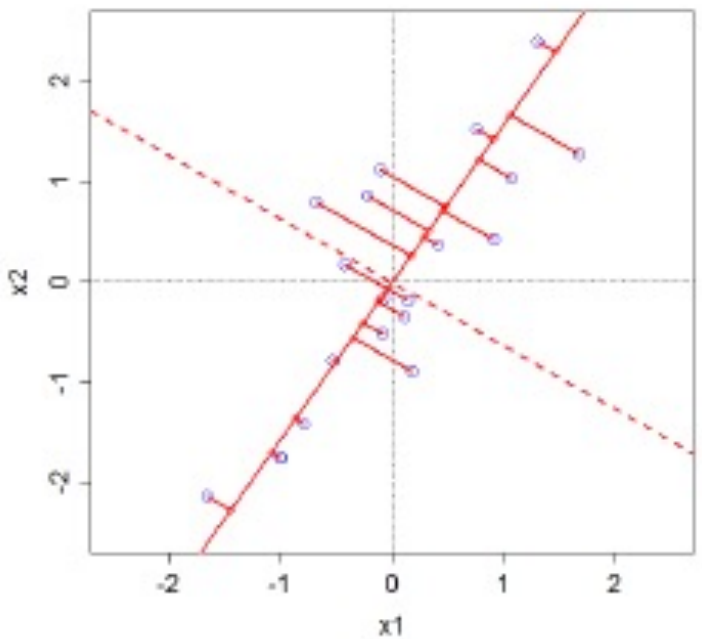
Choosing p'

- Use enough PCs to caputure 90% of variation
- Maximize adjusted R^2
- Do variable selection (ie, not necessarily leading PCs)

Principal Components Regression

$$y_i = \beta'_0 + \sum_{j=1}^{p'} \beta'_j u_{ij} + \epsilon_i$$

This is a bet that most variation in y is in direction of large variation in x.



Bias: variation in y due to PCs not included in the model

If $n > p$, we can consider all p Principle Components.

Partial Least Squares

- In PCR, we regress Y on directions that best represent the predictors $X_1, \ldots, X_p$. These directions are chosen to be PCs $Z_1, \ldots, Z_M$ in an unsupervised way.

- In PLS, we hope to choose the directions using Y.
  - $Z_1 = \psi_{11}X_1 + \psi_{21}X_2 + \cdots + \psi_{p1}X_p$ where $\psi_{j1} \propto corr(Y, X_j)$
  - $Z_2 = \cdots$ un-correlated with Z_1

- While the supervised dimension reduction of PLS can reduce bias, it also has the potential to increase variance.

PCR vs PLS directions

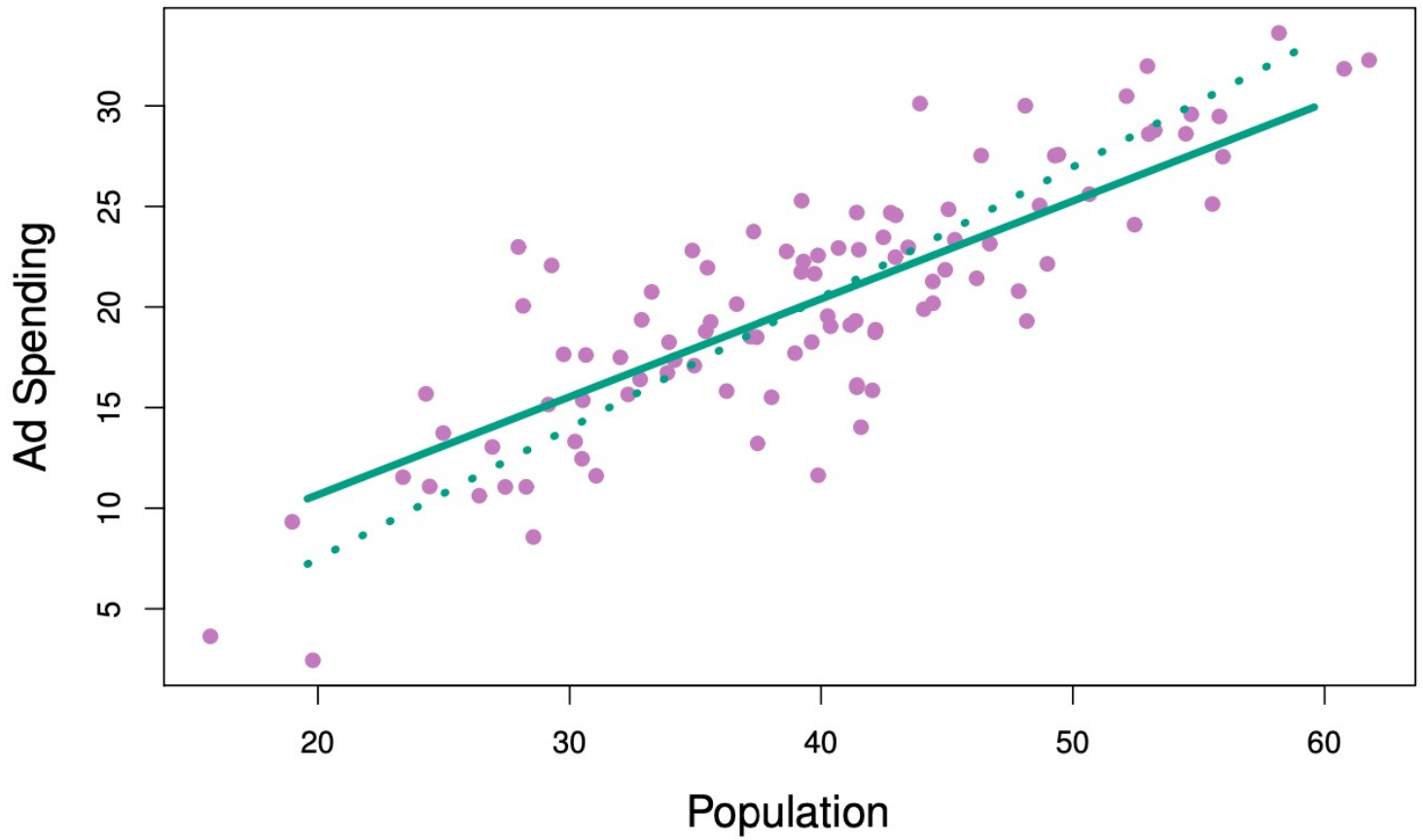


FIGURE 6.21. *For the advertising data, the first PLS direction (solid line) and first PCR direction (dotted line) are shown.*

Netflix Problem / Recommender System: Missing Data and Matrix Completion

	Jerry Maguire	Oceans	Road to Perdition	A Fortunate Man	Catch Me If You Can	Driving Miss Daisy	The Two Popes	The Laundromat	Code 8	The Social Network	...
Customer 1	•	•	•	•	4	•	•	•	•	•	...
Customer 2	•	•	3	•	•	•	3	•	•	3	...
Customer 3	•	2	•	4	•	•	•	•	2	•	...
Customer 4	3	•	•	•	•	•	•	•	•	•	...
Customer 5	5	1	•	•	4	•	•	•	•	•	...
Customer 6	•	•	•	•	•	2	4	•	•	•	...
Customer 7	•	•	5	•	•	•	•	3	•	•	...
Customer 8	•	•	•	•	•	•	•	•	•	•	...
Customer 9	3	•	•	•	5	•	•	1	•	•	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

- Can we impute the missing values / complete the data matrix with only a small proportion of observations?

Principal Components with Missing Values

- PCA finds the closest M-dim hyperplane

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - \sum_{m=1}^{M} a_{im}b_{jm} \right)^2 \right\}.$$

- With missing values, we solve

$$\underset{\mathbf{A}\in\mathbb{R}^{n\times M},\mathbf{B}\in\mathbb{R}^{p\times M}}{\text{minimize}} \left\{ \sum_{(i,j)\in\mathcal{O}}\left(x_{ij} - \sum_{m=1}^{M} a_{im}b_{jm} \right)^2 \right\},$$

- We can estimate a missing observation by $\hat{x}_{ij} = \sum_{m=1}^{M} \hat{a}_{im}\hat{b}_{jm}$

- For Netflix problem, $\hat{a}_{im}$ represents the strength of the ith user belonging to the group of customers who enjoys movies of the mth genre; $\hat{b}_{jm}$ represents the strength of the jth movie belonging to the mth genre.

Iterative Algorithm for Matrix Completion

1. Create a complete data matrix $\tilde{\mathbf{X}}$ of dimension $n \times p$ of which the (i, j) element equals

$$\tilde{x}_{ij} = \begin{cases} x_{ij} & \text{if } (i,j) \in \mathcal{O} \\ \bar{x}_j & \text{if } (i,j) \notin \mathcal{O}, \end{cases}$$

2. Repeat steps (a)–(c) until the objective (12.14) fails to decrease:

   (a) Solve

$$\underset{\mathbf{A} \in \mathbb{R}^{n \times M}, \mathbf{B} \in \mathbb{R}^{p \times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p} \sum_{i=1}^{n} \left(\tilde{x}_{ij} - \sum_{m=1}^{M} a_{im} b_{jm} \right)^2 \right\} \qquad (12.13)$$

   by computing the principal components of $\tilde{\mathbf{X}}$.

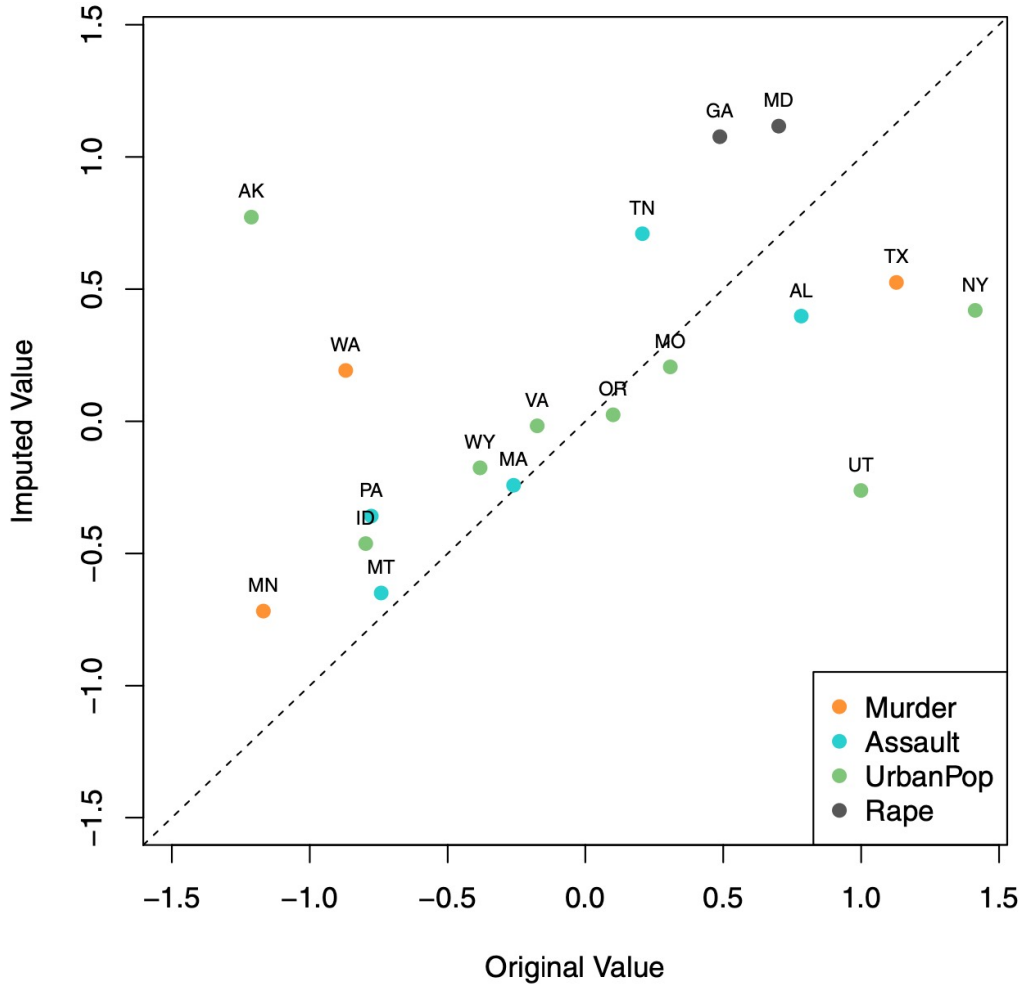
   (b) For each element $(i, j) \notin \mathcal{O}$, set $\tilde{x}_{ij} \leftarrow \sum_{m=1}^{M} \hat{a}_{im} \hat{b}_{jm}$.

   (c) Compute the objective

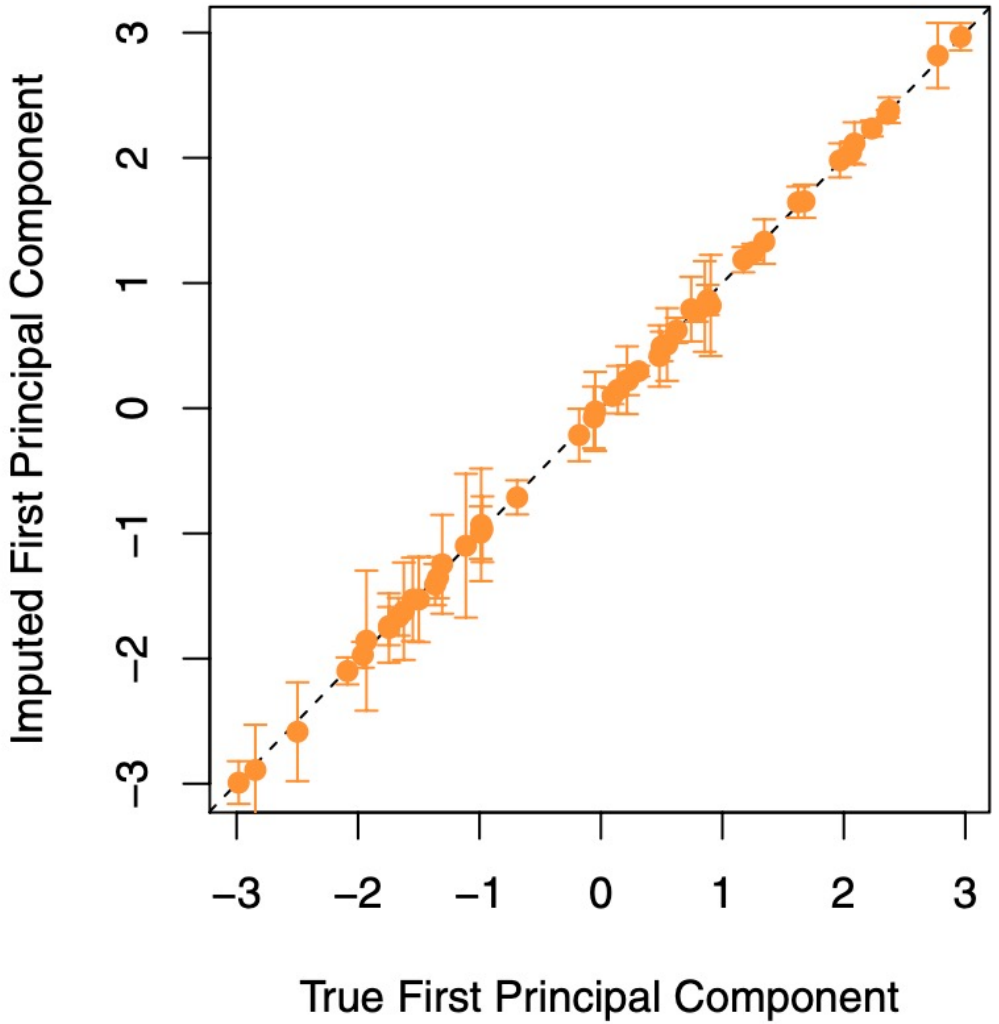
$$\sum_{(i,j) \in \mathcal{O}} \left(x_{ij} - \sum_{m=1}^{M} \hat{a}_{im} \hat{b}_{jm} \right)^2 . \qquad (12.14)$$

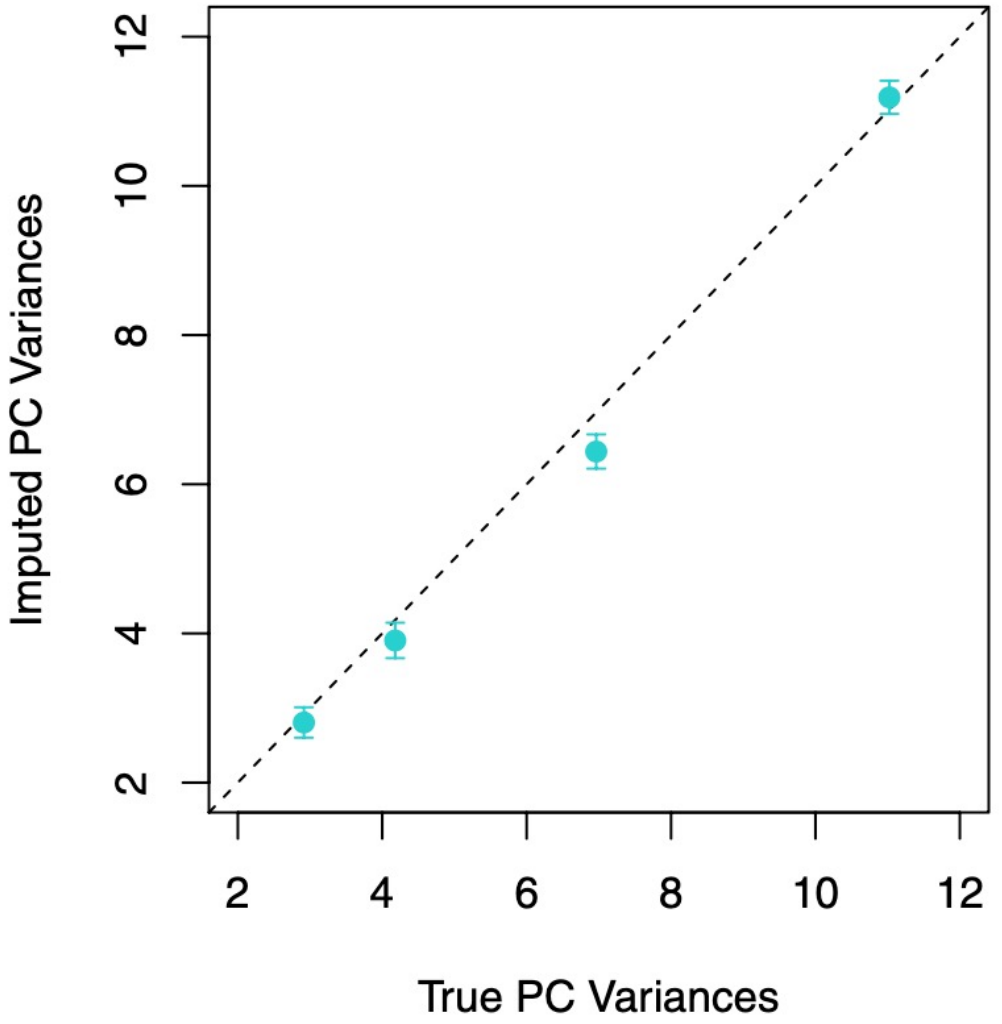
Performance of the Algorithm

- Randomly mask 20 rows out of 50 states in USArrests data.



Performance of the Algorithm





Summary

- PCA is a nice way for dimension reduction, data visualization
  - Identify key features, clusters, outliers
- It involves all variables; hence no variable selection; interpretation may be difficult
- In PCR, latter PCs may be more useful for predicting Y
  - No use of Y when performing PCA
- Supervised PCA: use Y to supervise PCA
  - Proposed by former PhD student Dr. LI Gen (Columbia) of Prof. SHEN Haipeng
  - https://sites.google.com/view/ligen

Computation of PCA: Singular Value Decomposition

- History: Beltrami (1873), Jordan (1874)
- Consider the $n \times p$ feature matrix $\boldsymbol{X}$ with $rank(\boldsymbol{X}) = r$.

- Then, the singular value decomposition (SVD) of $\boldsymbol{X}$ is

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \sum_{k=1}^{r} s_k U_k V_k^T$$

- Left singular vectors: $U_1, U_2, \ldots, U_r$ with $U_i^T U_j = \delta_{ij}$

- Right singular vectors: $V_1, V_2, \ldots, V_r$ with $V_i^T V_j = \delta_{ij}$

- Descending singular values: $s_1 \geq s_2 \geq \cdots \geq s_r > 0$

Connection between PCA and SVD

- Assume $\boldsymbol{X}$ is column-centered
- Via SVD,

$$\boldsymbol{X} = \boldsymbol{USV}^T$$

- Then,

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{VSU}^T\boldsymbol{USV}^T = \boldsymbol{VS}^2\boldsymbol{V}^T$$

- Note that $\boldsymbol{X}^T\boldsymbol{X}$ is proportional to the covariance matrix of $\boldsymbol{X}$

- Hence,
  - Columns of $\boldsymbol{US}(= \boldsymbol{XV})$: the PCs
  - Columns of $\boldsymbol{V}$: the PC loading vectors
  - Squared singular values: $(n \times)$ variances of the PCs
  - The PCs are naturally ordered

Another View of SVD: Low-Rank Approximation

- Eckart and Young (1936)
- For an integer $K \leq r$, consider an arbitrary rank-K matrix $\boldsymbol{X}^*$
- Consider the squared distance between $\boldsymbol{X}$ and $\boldsymbol{X}^*$, measured by the Frobenius norm:

$$||\boldsymbol{X} - \boldsymbol{X}^*||_F^2 = tr\{(\boldsymbol{X} - \boldsymbol{X}^*)(\boldsymbol{X} - \boldsymbol{X}^*)^T\} = \sum_{i,j}(x_{ij} - x_{ij}^*)^2$$

- Then, SVD provides the best rank-K approximation of $\boldsymbol{X}$:

$$\boldsymbol{X}^{(K)} \equiv \sum_{k=1}^{K} s_k U_k V_k^T = argmin_{\{\boldsymbol{X}^*: rank(\boldsymbol{X}^*)=K\}}||\boldsymbol{X} - \boldsymbol{X}^*||_F^2$$

SVD: Bilinear Model

- Consider rank-1 approximation.
- Any rank-1 matrix ben be written as

$$s\boldsymbol{u}\boldsymbol{v}^T$$

where s: a positive scalar, $\boldsymbol{u}$: norm-1 vector, $\boldsymbol{v}$: norm-1 vector.

- The first SVD triplet $\{s_1, \boldsymbol{u}_1, \boldsymbol{v}_1\}$ is the solution for

$$min_{\{s,\ \boldsymbol{u},\ \boldsymbol{v}\}} ||\boldsymbol{X} - s\boldsymbol{u}\boldsymbol{v}^T||_F^2$$

- Bilinear model
- Same for rank-K approximation

SVD: Alternating Least Squares

- Equivalently, the first SVD triplet $\{s_1, \boldsymbol{u}_1, \boldsymbol{v}_1\}$ is the solution to

$$min_{\{s,\ \boldsymbol{u},\ \boldsymbol{v}\}} ||\boldsymbol{X} - s\boldsymbol{u}\boldsymbol{v}^T||_F^2 = \sum_i (\boldsymbol{x}_{(i)} - su_i\boldsymbol{v}^T)^2$$

or

$$min_{\{s,\ \boldsymbol{u},\ \boldsymbol{v}\}} ||\boldsymbol{X} - s\boldsymbol{u}\boldsymbol{v}^T||_F^2 = \sum_j (\boldsymbol{x}_j - sv_j\boldsymbol{u})^2$$

with $\boldsymbol{x}_{(i)}$: i-th row of $\boldsymbol{X}$, $\boldsymbol{x}_j$: j-th column of $\boldsymbol{X}$

- Two least-squares regression models

- Alternating least squares (Gabriel and Zamir, 1979)