PCA for Orthopedic Surgery Data on Hospitals

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Outline

1 PCA on Orthopedic Data

PCA on Orthopedic Data

We'll analyze data on orthopedic surgeries in hospitals. Download hospital.csv from Blackboard.

This file contains information compiled by a company that sells orthopedic devices to hospitals. The observations are 4703 U.S. hospitals. One goal is to better understand the client hospitals; e.g., what hospitals are similar in terms of their size, revenue, and numbers of various kinds of operations? What aspects differentiate the hospitals from one another?

Let's try to answer these questions with PCA

PCA on Orthopedic Data

Relevant variables:

Variable Name	Description
BEDS	Number of hospital beds
RBEDS	Number of rehab beds
OUTV	Number of outpatient visits
ADM	Administrative cost (thousands of $'s per year)
SIR	Revenue from inpatient
HIP95	Number of hip operations for 1995
KNEE95	Number of knee operations for 1995
TH	Indicator of teaching hospital
TRAUMA	Indicator of having a trauma unit
REHAB	Indicator of having a rehab unit
HIP96	Number of hip operations for 1996
KNEE96	Number of knee operations for 1996
FEMUR96	Number of femur operations for 1996

PCA on Orthopedic Data

- Type all your code into a script file, so that you can easily modify & rerun.

- Read the data into R.

- remove several variables that are not relevant to the analysis, and remove the binary variables, because WE CAN ONLY DO PCA ON CONTINUOUS VARIABLES:

  ```
  ortho2 = ortho[,-c(1:4,10:11,14:16)]
  ```

- restrict to just hospitals that made a purchase:

  ```
  ortho3 = ortho2[ ortho2[,”SIR”] > 0, ]
  ```

PCA on Orthopedic Data

- We will do PCA on the continuous variables. First, look at pairwise scatterplots for the remaining variables in the dataset.
- The relationships between these variables might look more linear if we log-transformed them, because many of the variables appear to be heavily right-skewed. Linear relationships between the variables will help PCA work better (although it is NOT a requirement/assumption of the method), because PCA finds a good low-dimensional linear transformation of the variables.

PCA on Orthopedic Data

- Log-transform ALL the variables:

  ortho4 = log(ortho3 + 1)

- Notice I added a small constant before taking the log, to avoid taking log(0). There's nothing magical about the number 1, except that it is very small relative to the scale of the variables in the dataset.

- Recheck the pairwise scatterplots. Better?

PCA on Orthopedic Data

Now we'll get the PCs.

- First STANDARDIZE THE VARIABLES:
  orthoStandard = scale(ortho4)

- Now do PCA:
  res = prcomp(orthoStandard)

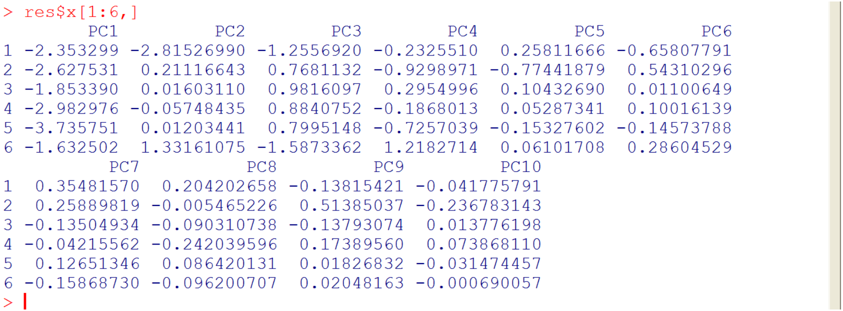
- The principal components are given in **res$x**. How many PCs are there total?

  A. 1-3
  B. 4-8
  C. 9-15
  D. 16+

  How many variables are there in orthoStandard?

PCA on Orthopedic Data

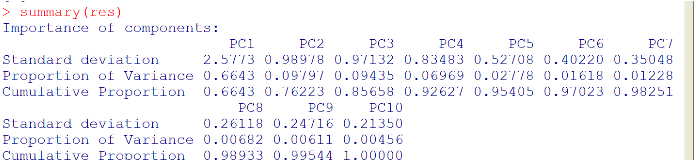
Here are the PCs for the first 6 hospitals:



```
> res$x[1:6,]
        PC1         PC2        PC3        PC4         PC5         PC6
1 -2.353299 -2.81526990 -1.2556920 -0.2325510  0.25811666 -0.65807791
2 -2.627531  0.21116643  0.7681132 -0.9298971 -0.77441879  0.54310296
3 -1.853390  0.01603110  0.9816097  0.2954996  0.10432690  0.01100649
4 -2.982976 -0.05748435  0.8840752 -0.1868013  0.05287341  0.10016139
5 -3.735751  0.01203441  0.7995148 -0.7257039 -0.15327602 -0.14573788
6 -1.632502  1.33161075 -1.5873362  1.2182714  0.06101708  0.28604529
          PC7          PC8         PC9         PC10
1  0.35481570  0.204202658 -0.13815421 -0.041775791
2  0.25889819 -0.005465226  0.51385037 -0.236783143
3 -0.13504934 -0.090310738 -0.13793074  0.013776198
4 -0.04215562 -0.242039596  0.17389560  0.073868110
5  0.12651346  0.086420131  0.01826832 -0.031474457
6 -0.15868730 -0.096200707  0.02048163 -0.000690057
> 
```

PCA on Orthopedic Data

- calling **summary(res)** gives a summary of the proportion of the variance captured by each PC (equal to $\frac{d_j^2}{\sum_{k=1}^{p} d_k^2}$).

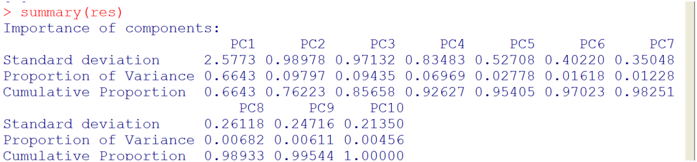


```
> summary(res)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5773 0.98978 0.97132 0.83483 0.52708 0.40220 0.35048
Proportion of Variance 0.6643 0.09797 0.09435 0.06969 0.02778 0.01618 0.01228
Cumulative Proportion  0.6643 0.76223 0.85658 0.92627 0.95405 0.97023 0.98251
                           PC8     PC9    PC10
Standard deviation     0.26118 0.24716 0.21350
Proportion of Variance 0.00682 0.00611 0.00456
Cumulative Proportion  0.98933 0.99544 1.00000
```

- How many PCs are needed to capture 90% of the variability in the original data? iClicker:

  A. 1
  B. 2
  C. 3
  D. 4
  E. 5+

PCA on Orthopedic Data

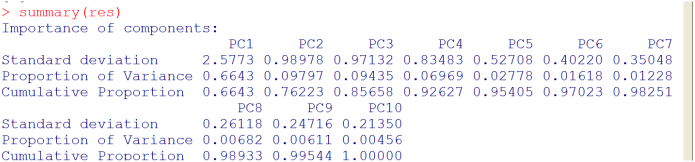


```
> summary(res)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5773 0.98978 0.97132 0.83483 0.52708 0.40220 0.35048
Proportion of Variance 0.6643 0.09797 0.09435 0.06969 0.02778 0.01618 0.01228
Cumulative Proportion  0.6643 0.76223 0.85658 0.92627 0.95405 0.97023 0.98251
                           PC8     PC9    PC10
Standard deviation     0.26118 0.24716 0.21350
Proportion of Variance 0.00682 0.00611 0.00456
Cumulative Proportion  0.98933 0.99544 1.00000
```

- What is the value of $\frac{d_2^2}{n-1}$?

  A. 0-2
  B. 2-8
  C. 8-30
  D. 30+

PCA on Orthopedic Data



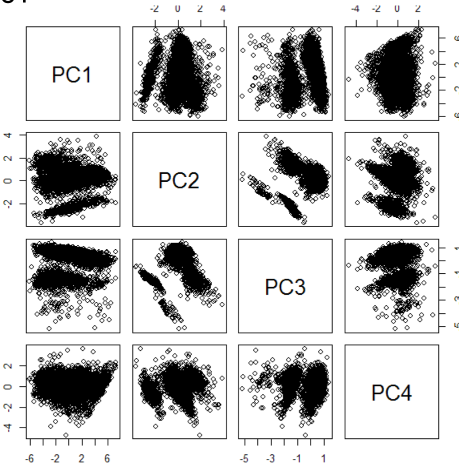
```
> summary(res)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5773 0.98978 0.97132 0.83483 0.52708 0.40220 0.35048
Proportion of Variance 0.6643 0.09797 0.09435 0.06969 0.02778 0.01618 0.01228
Cumulative Proportion  0.6643 0.76223 0.85658 0.92627 0.95405 0.97023 0.98251
                           PC8     PC9    PC10
Standard deviation     0.26118 0.24716 0.21350
Proportion of Variance 0.00682 0.00611 0.00456
Cumulative Proportion  0.98933 0.99544 1.00000
```

- What is the sample covariance of PC 1 & PC 3?

  A. 0-2
  B. 2-8
  C. 8-30
  D. 30+

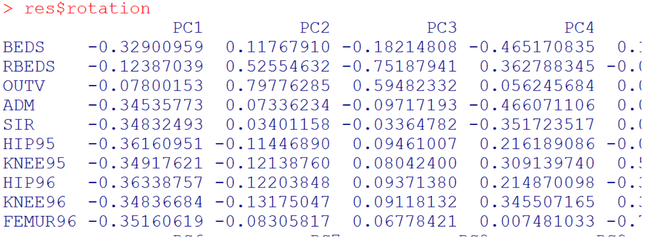
PCA on Orthopedic Data

Create pairwise scatterplot of the first four PCs. What do you see?



PCA on Orthopedic Data

Let’s try to understand what these groups might represent. Here are the PC loadings for the first 4 PCs (1st 4 columns of V matrix):



```
> res$rotation
                PC1         PC2         PC3          PC4
BEDS    -0.32900959  0.11767910 -0.18214808 -0.465170835  0.1
RBEDS   -0.12387039  0.52554632 -0.75187941  0.362788345 -0.0
OUTV    -0.07800153  0.79776285  0.59482332  0.056245684  0.0
ADM     -0.34535773  0.07336234 -0.09717193 -0.466071106  0.0
SIR     -0.34832493  0.03401158 -0.03364782 -0.351723517  0.0
HIP95   -0.36160951 -0.11446890  0.09461007  0.216189086 -0.0
KNEE95  -0.34917621 -0.12138760  0.08042400  0.309139740  0.5
HIP96   -0.36338757 -0.12203848  0.09371380  0.214870098 -0.3
KNEE96  -0.34836684 -0.13175047  0.09118132  0.345507165  0.3
FEMUR96 -0.35160619 -0.08305817  0.06778421  0.007481033 -0.7
```

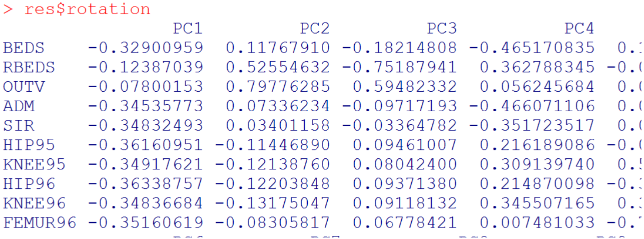
What do the first few PCs represent?

PCA on Orthopedic Data

What does this tell us about some of the groups we saw in the scatterplots?

PCA on Orthopedic Data

Here are the 1st 4 columns of *V* matrix again:

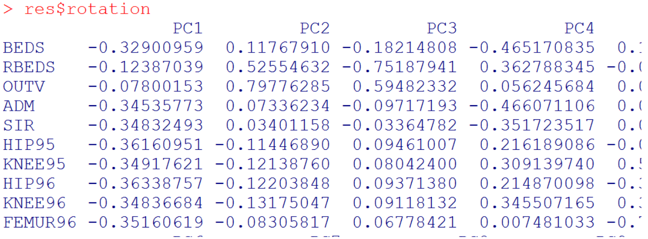


```
> res$rotation
                PC1         PC2         PC3          PC4
BEDS    -0.32900959  0.11767910 -0.18214808 -0.465170835  0.1
RBEDS   -0.12387039  0.52554632 -0.75187941  0.362788345 -0.0
OUTV    -0.07800153  0.79776285  0.59482332  0.056245684  0.0
ADM     -0.34535773  0.07336234 -0.09717193 -0.466071106  0.0
SIR     -0.34832493  0.03401158 -0.03364782 -0.351723517  0.0
HIP95   -0.36160951 -0.11446890  0.09461007  0.216189086 -0.0
KNEE95  -0.34917621 -0.12138760  0.08042400  0.309139740  0.5
HIP96   -0.36338757 -0.12203848  0.09371380  0.214870098 -0.1
KNEE96  -0.34836684 -0.13175047  0.09118132  0.345507165  0.1
FEMUR96 -0.35160619 -0.08305817  0.06778421  0.007481033 -0.7
```

Say we have a new observation with value of RBEDS that is one standard deviation below the average of RBEDS, and value of OUTV that is one standard deviation above the average of OUTV, and all other variables equal to the average value of that variable across hospitals. What would be the value of the second PC for this observation?

PCA on Orthopedic Data

Here are the 1st 4 columns of *V* matrix again:

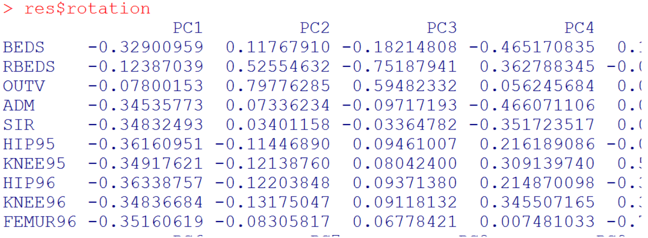


```
> res$rotation
                PC1         PC2         PC3          PC4
BEDS    -0.32900959  0.11767910 -0.18214808 -0.465170835  0.1
RBEDS   -0.12387039  0.52554632 -0.75187941  0.362788345 -0.0
OUTV    -0.07800153  0.79776285  0.59482332  0.056245684  0.0
ADM     -0.34535773  0.07336234 -0.09717193 -0.466071106  0.0
SIR     -0.34832493  0.03401158 -0.03364782 -0.351723517  0.0
HIP95   -0.36160951 -0.11446890  0.09461007  0.216189086 -0.0
KNEE95  -0.34917621 -0.12138760  0.08042400  0.309139740  0.5
HIP96   -0.36338757 -0.12203848  0.09371380  0.214870098 -0.3
KNEE96  -0.34836684 -0.13175047  0.09118132  0.345507165  0.3
FEMUR96 -0.35160619 -0.08305817  0.06778421  0.007481033 -0.7
```

Give an example of a DIFFERENT loadings matrix that is also correct for the orthopedic dataset.

PCA on Orthopedic Data

Here are the 1st 4 columns of *V* matrix again:

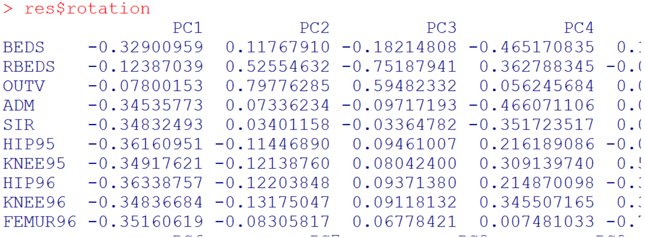


```
> res$rotation
                PC1         PC2         PC3          PC4
BEDS    -0.32900959  0.11767910 -0.18214808 -0.465170835  0.1
RBEDS   -0.12387039  0.52554632 -0.75187941  0.362788345 -0.0
OUTV    -0.07800153  0.79776285  0.59482332  0.056245684  0.0
ADM     -0.34535773  0.07336234 -0.09717193 -0.466071106  0.0
SIR     -0.34832493  0.03401158 -0.03364782 -0.351723517  0.0
HIP95   -0.36160951 -0.11446890  0.09461007  0.216189086 -0.0
KNEE95  -0.34917621 -0.12138760  0.08042400  0.309139740  0.5
HIP96   -0.36338757 -0.12203848  0.09371380  0.214870098 -0.3
KNEE96  -0.34836684 -0.13175047  0.09118132  0.345507165  0.3
FEMUR96 -0.35160619 -0.08305817  0.06778421  0.007481033 -0.7
```

What is the Euclidean norm (length) of the third column of this matrix?

PCA on Orthopedic Data

Here are the 1st 4 columns of *V* matrix again:



```
> res$rotation
                PC1         PC2         PC3          PC4
BEDS    -0.32900959  0.11767910 -0.18214808 -0.465170835  0.1
RBEDS   -0.12387039  0.52554632 -0.75187941  0.362788345 -0.0
OUTV    -0.07800153  0.79776285  0.59482332  0.056245684  0.0
ADM     -0.34535773  0.07336234 -0.09717193 -0.466071106  0.0
SIR     -0.34832493  0.03401158 -0.03364782 -0.351723517  0.0
HIP95   -0.36160951 -0.11446890  0.09461007  0.216189086 -0.0
KNEE95  -0.34917621 -0.12138760  0.08042400  0.309139740  0.5
HIP96   -0.36338757 -0.12203848  0.09371380  0.214870098 -0.3
KNEE96  -0.34836684 -0.13175047  0.09118132  0.345507165  0.3
FEMUR96 -0.35160619 -0.08305817  0.06778421  0.007481033 -0.7
```

What is the inner product (dot product) of the second and fourth columns of this matrix?