LINEAR ALGEBRA FOR DATA SCIENCE IN R

Principal Component Analysis

Eric Eager
Data Scientist at Pro Football Focus

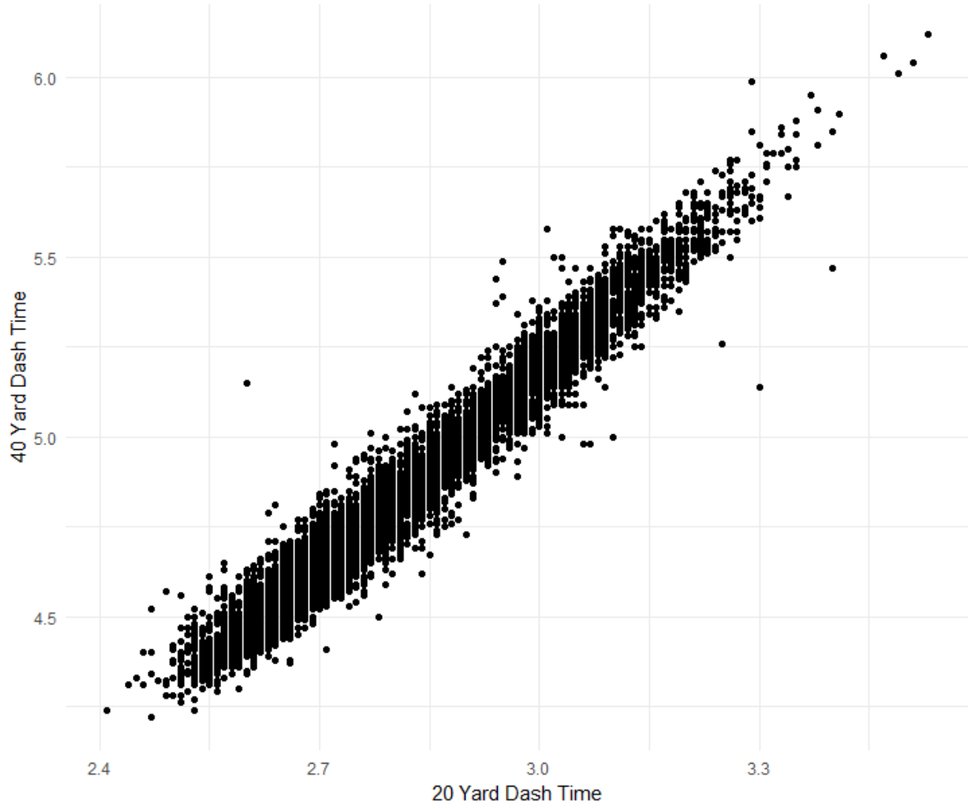
Big Data

```
> head(combine)

> head(select(combine, height:shuttle))
  height weight forty vertical bench broad_jump three_cone shuttle
1     71    192  4.38     35.0    14        127       6.71    3.98
2     73    298  5.34     26.5    27         99       7.81    4.71
3     77    256  4.67     31.0    17        113       7.34    4.38
4     74    198  4.34     41.0    16        131       6.56    4.03
5     76    257  4.87     30.0    20        118       7.12    4.23
6     78    262  4.60     38.5    18        128       7.53    4.48
```

```
> nrow(combine)
[1] 2885
```

Big Data - Redundancy



Principal Component Analysis

- One of the more-useful methods from applied linear algebra
- Non-parametric way of extracting meaningful information from confusing data sets
- Uncovers hidden, low-dimensional structures that underlie your data
- These structures are more-easily visualized and are often interpretable to content experts

Principal Component Analysis - Motivating Example

LINEAR ALGEBRA FOR DATA SCIENCE IN R

Let's practice!

LINEAR ALGEBRA FOR DATA SCIENCE IN R

The Linear Algebra Behind PCA

Eric Eager
Data Scientist at Pro Football Focus

Theory

The matrix A^T, the *transpose* of A, is the matrix made by interchanging the rows and columns of A.

If your data set is in a matrix A, and the mean of each column has been subtracted from each element in a given column, then the i, jth element of the matrix

$$\frac{A^T A}{n-1},$$

where n is the number of rows of A, is the *covariance* between the variables in the ith and jth column of the data in the matrix.

Hence, the ith element of the diagonal of $\frac{A^T A}{n-1}$ is the *variance* of the ith column of the matrix.

Theory

```
> A
     [,1] [,2]
[1,]    1    2
[2,]    2    4
[3,]    3    6
[4,]    4    8
[5,]    5   10
```

```
> A[, 1] <- A[, 1] - mean(A[, 1])
> A[, 2] <- A[, 2] - mean(A[, 2])
>
> A
     [,1] [,2]
[1,]   -2   -4
[2,]   -1   -2
[3,]    0    0
[4,]    1    2
[5,]    2    4
```

Theory

```
> t(A)%*%A/(nrow(A) - 1)
     [,1] [,2]
[1,]  2.5    5
[2,]  5.0   10
```

```
> cov(A[, 1], A[, 2])
[1] 5
```

```
> var(A[, 1])
[1] 2.5
> var(A[, 2])
[1] 10
```

PCA

- The eigenvalues $\lambda_1, \lambda_2, ...\lambda_n$ of $\frac{A^T A}{n-1}$ are real, and their corresponding eigenvectors are *orthogonal*, or point in distinct directions.
- The *total variance* of the data set is the sum of the eigenvalues of $\frac{A^T A}{n-1}$.
- These eigenvectors $v_1, v_2, ..., v_n$ are called the *principal components* of the data set in the matrix A.
- The direction that v_j points in can explain λ_j of the total variance in the data set. If λ_j, or a subset of $\lambda_1, \lambda_2, ...\lambda_n$ explain a significant amount of the total variance, there is an opportunity for dimension reduction.

Example

```
> eigen(t(A)%*%A/(nrow(A) - 1))
eigen() decomposition
$`values`
[1] 12.5  0.0

$vectors
          [,1]       [,2]
[1,] 0.4472136 -0.8944272
[2,] 0.8944272  0.4472136
```

LINEAR ALGEBRA FOR DATA SCIENCE IN R

Let's practice!

LINEAR ALGEBRA FOR DATA SCIENCE IN R

Performing PCA in R

Eric Eager
Data Scientist at Pro Football Focus

NFL Combine Data

```
> head(select(combine, height:shuttle))
> head(A)
  height weight forty vertical bench broad_jump three_cone shuttle
1     71    192  4.38     35.0    14        127       6.71    3.98
2     73    298  5.34     26.5    27         99       7.81    4.71
3     77    256  4.67     31.0    17        113       7.34    4.38
4     74    198  4.34     41.0    16        131       6.56    4.03
5     76    257  4.87     30.0    20        118       7.12    4.23
6     78    262  4.60     38.5    18        128       7.53    4.48
```

NFL Combine Data

```
> prcomp(A)
Standard deviations (1, .., p=8):
[1] 46.7720885  6.6356959  4.7108443  2.2950226  1.6430770  0.2513368  0.121690

Rotation (n x k) = (8 x 8):
                   PC1           PC2           PC3           PC4           PC5
height      0.042047079 -0.061885367  0.1454490039 -0.1040556410 -0.980792060  0
weight      0.980711529 -0.130912788  0.1270100265  0.0193388930  0.066908382 -0
forty       0.006112061  0.012525260  0.0025260713 -0.0021291637  0.004096693  0
vertical   -0.062926466 -0.333556369  0.0398922845  0.9366594549 -0.074901137  0
bench       0.088291423 -0.313533433 -0.9363461471 -0.0745692157 -0.107188391  0
broad_jump -0.156742686 -0.876925849  0.2904565302 -0.3252903706  0.126494599  0
three_cone  0.007468520  0.014691994  0.0009057581  0.0003320888  0.020902644  0
shuttle     0.004518826  0.009863931  0.0023111814 -0.0094052914  0.004010629  0
```

```
> summary(prcomp(A))
Importance of components:
                            PC1      PC2      PC3      PC4      PC5      PC6      PC7
Standard deviation     46.7721 6.63570 4.71084 2.29502 1.64308 0.25134 0.12169 0
Proportion of Variance  0.9672 0.01947 0.00981 0.00233 0.00119 0.00003 0.00001 0
Cumulative Proportion   0.9672 0.98663 0.99644 0.99877 0.99996 0.99999 0.99999 1
```

NFL Combine Data

```
> head(prcomp(A)$x[, 1:2])
            PC1        PC2
[1,] -62.005067  -2.654645
[2,]  48.123290   6.693433
[3,]   3.732016   1.283046
[4,] -56.823742  -9.764098
[5,]   4.213670  -3.779862
[6,]   6.924978 -15.530509
```

```
> head(cbind(combine[, 1:4], prcomp(A)$x[, 1:2]))
             player position       school year        PC1        PC2
1   Jaire Alexander       CB   Louisville 2018 -62.005067  -2.654645
2       Brian Allen        C Michigan St. 2018  48.123290   6.693433
3      Mark Andrews       TE     Oklahoma 2018   3.732016   1.283046
4         Troy Apke        S     Penn St. 2018 -56.823742  -9.764098
5 Dorance Armstrong     EDGE       Kansas 2018   4.213670  -3.779862
6         Ade Aruna       DE       Tulane 2018   6.924978 -15.530509
```

Things to Do After PCA

- Data wrangling/quality control
- Data visualization
- Unsupervised learning (clustering)
- Supervised learning (for prediction or explanation)
- Much more!

Example - Data Visualization

LINEAR ALGEBRA FOR DATA SCIENCE IN R

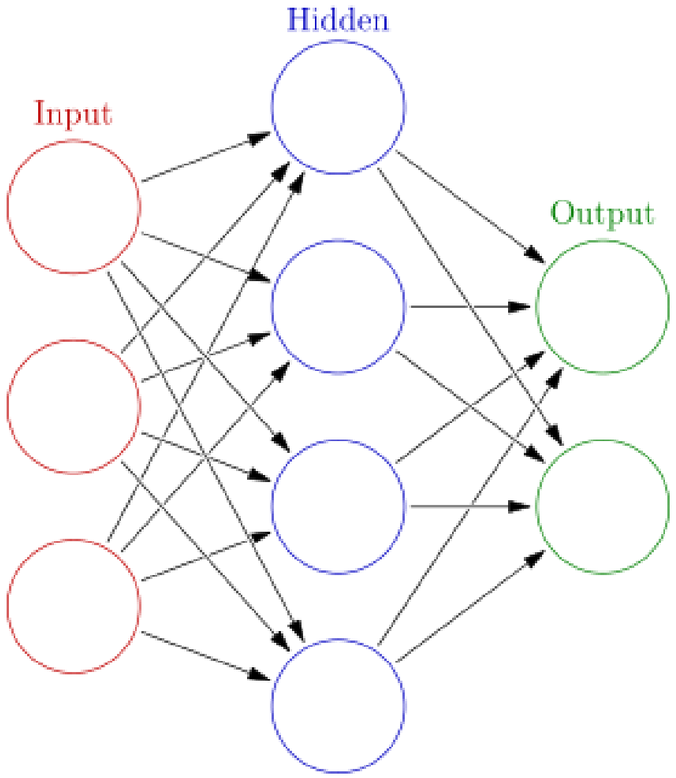
Let's practice!

LINEAR ALGEBRA FOR DATA SCIENCE IN R

Congratulations!

Eric Eager
Data Scientist at Pro Football Focus

Chapter 1 - Vectors and Matrices



Chapter 2 - Matrix-Vector Equations

Teams	Johns Hopkins	F & M	Gettysburg	Dickinson	McDaniel
Johns Hopkins	-	Loss, 12 - 14	Win 49-35	Win 49-0	Win 49-7
F & M	Win, 14 - 12	-	Loss, 31-38	Win 36-28	Win 35-10
Gettsyburg	Loss 35-49	Win, 38-31	-	Loss 13-23	Win 35-3
Dickinson	Loss 0-49	Loss 28-36	Win 23-13	-	Win 38-31
McDaniel	Loss 7-49	Loss 10-35	Loss 3-35	Loss 31-38	-

Chapter 3 - Eigenvalues and Eigenvectors

Chapter 4 - Principal Component Analysis

Going Further

- Introduction to Data
- Working with Data in the *tidyverse*
- Foundations of Probability in R
- Exploratory Data Analysis
- Data Visualization with ggplot2 (Parts 1 and 2)
- Case Studies!

LINEAR ALGEBRA FOR DATA SCIENCE IN R

Thank You!