PCA (Principal Component Analysis)

- Dimension reduction (denoise)
- Capture group structure
- Data visulization (high dimension)

Case Study: ASVAB tests and AFQT

- Screening tests to join army
- ASVAB: 10 tests word, pwag, math, arith, science, coding ...

AFQT: Word, Pwag, math, arith

Goal: 1. How to summarize the tests to reveal one's intelligence or some abilities?

2. How to display 10 dimensional data
3. How AFQT is formed?

0. **A toy example:** Two sets of tests, X_1, X_2

Goal: to see how students do?

Case 1:

X_1: spread out

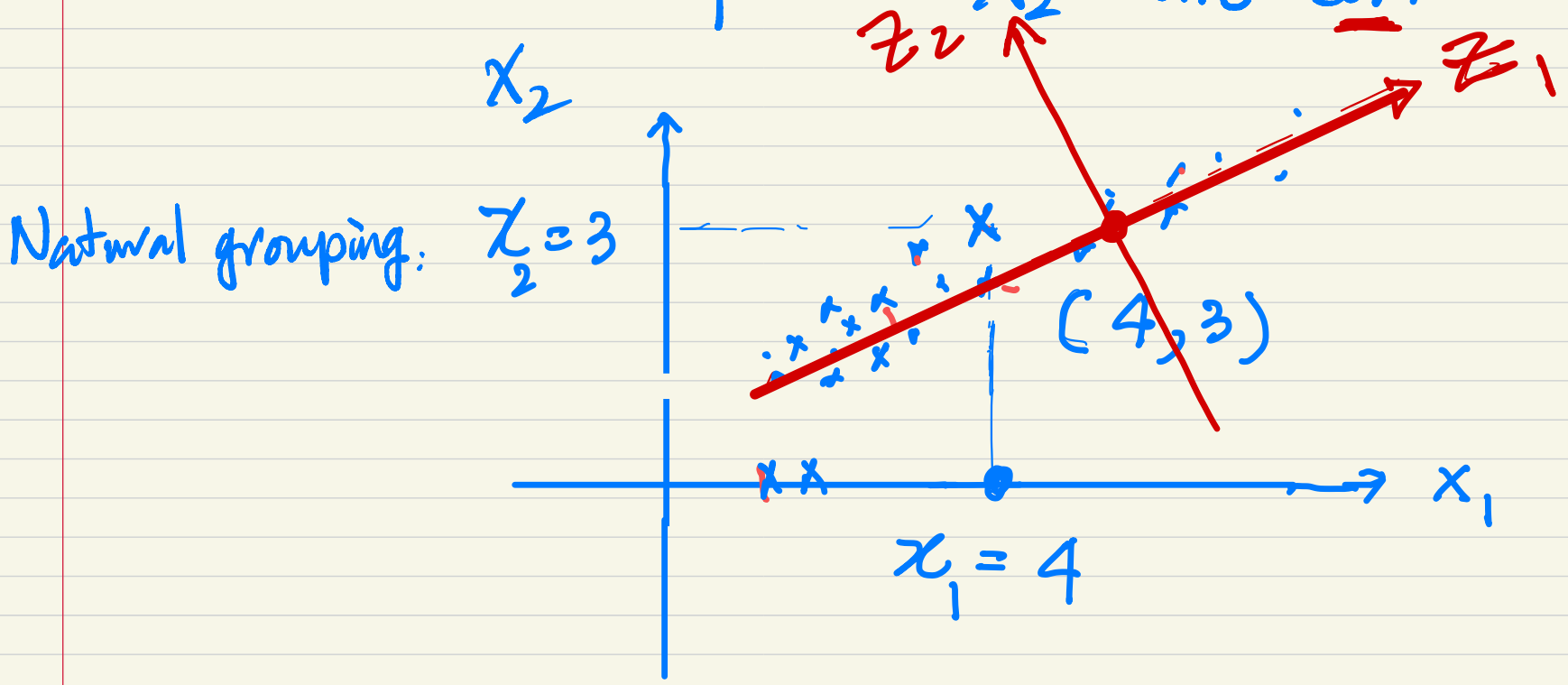
X_2: Every one got ≈ 30

too hard!

X_2 provides no information: throw it away

only uses x_1 ! X_1: Principal Component

Case 2: X_1 and X_2 are corr



optim 1: Keep both X_1 + X_2

PC1: $Z_1 = \varphi_{11} X_1 + \varphi_{21} X_2$ 2: Use X_1 alone

PC2: $Z_2 = \varphi_{12} X_1 + \varphi_{22} X_2$ 3: take weighted sum : Z_1

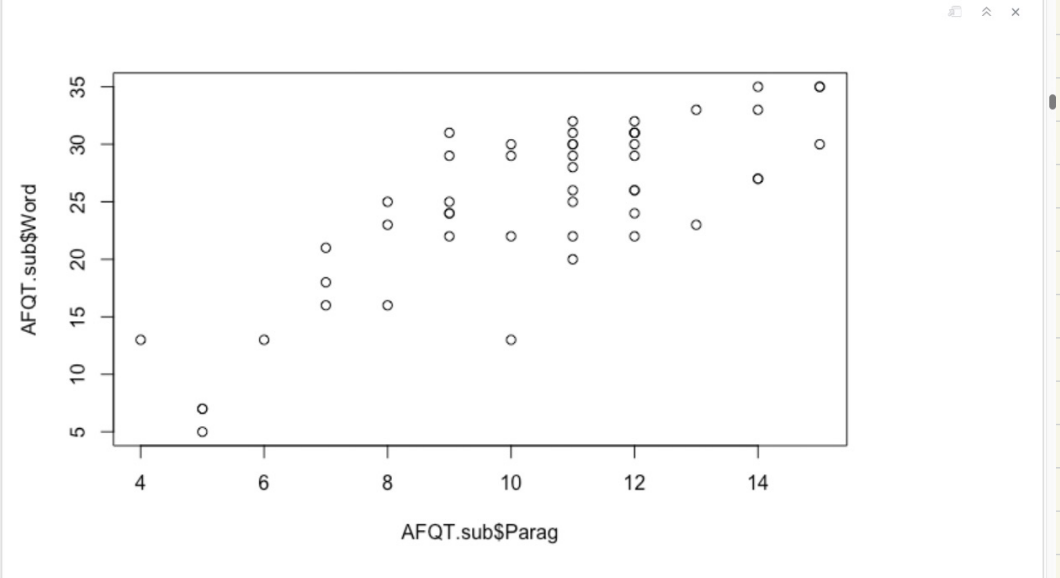
$\Rightarrow$ PCA

- How to do it ?
- How much informatin lost ?

1. Simple case : X_1 = Parag $\Rightarrow \bar{x}_1 = 10.4$, $s_1 = 2.75$

X_2 = Word $\bar{x}_2 = 24.9$, $s_2 = 7.41$

$n = 50$



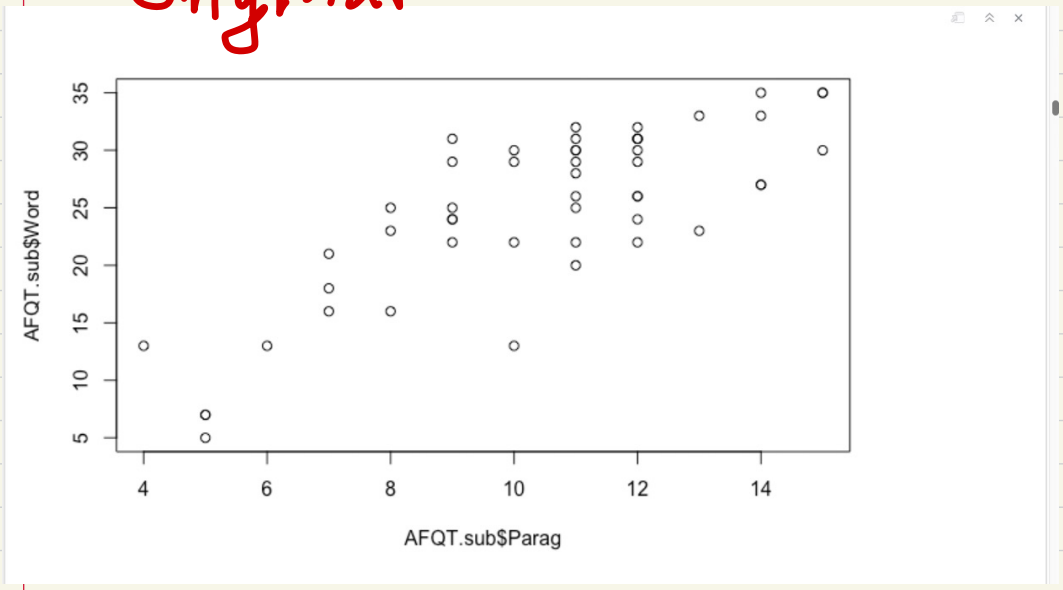
Goal : Find a new score

$$Z_1 = \varphi_{11} \cdot x_1 + \varphi_{21} \cdot x_2, \quad \varphi_{11}, \varphi_{21} \text{ unknown}$$

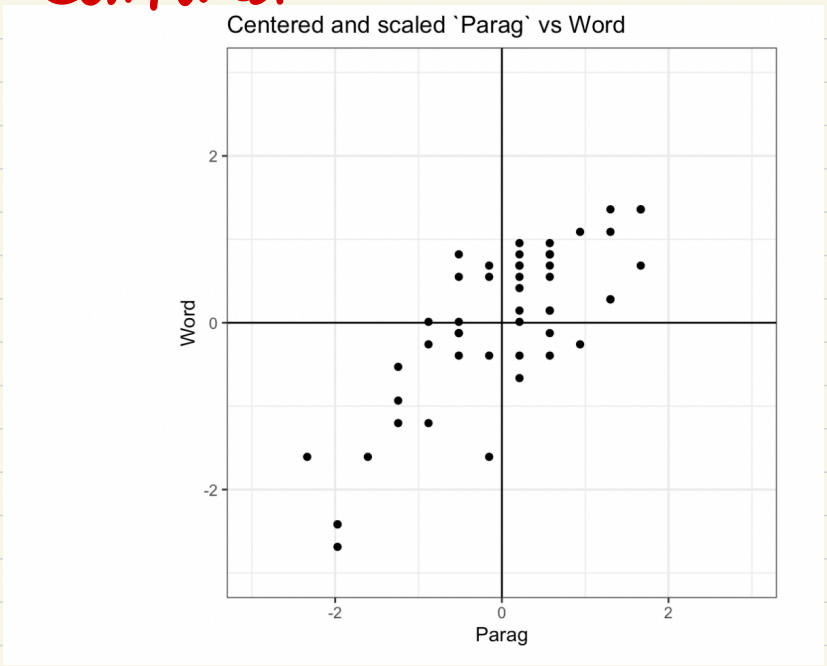
Z_1 : Informative

Q: Informative ?

Original



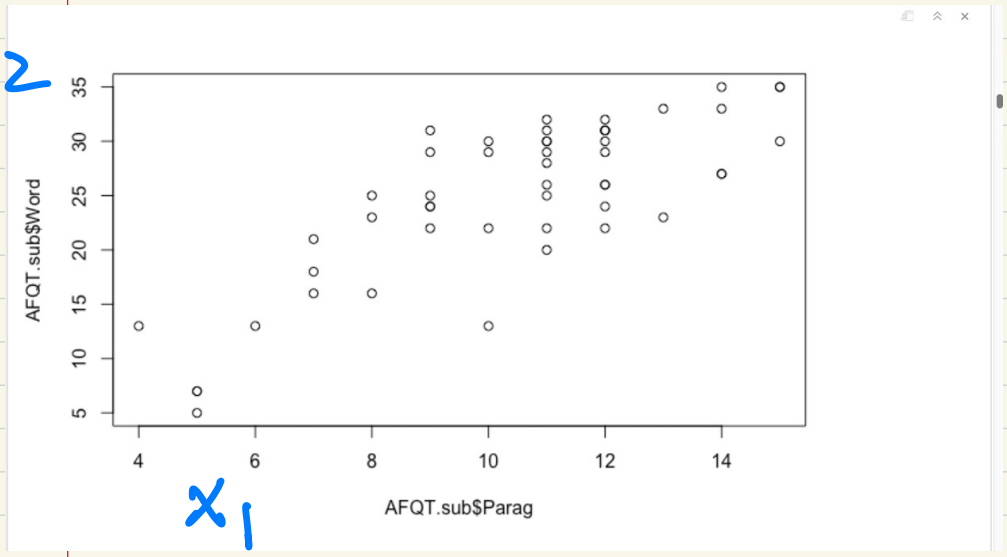
Centered & Scaled



First: Center and scale each

For simplicity consider $n = 50$ people

X_2



X_1

Original

Parag: Ave = 10.4 SD = 2.75

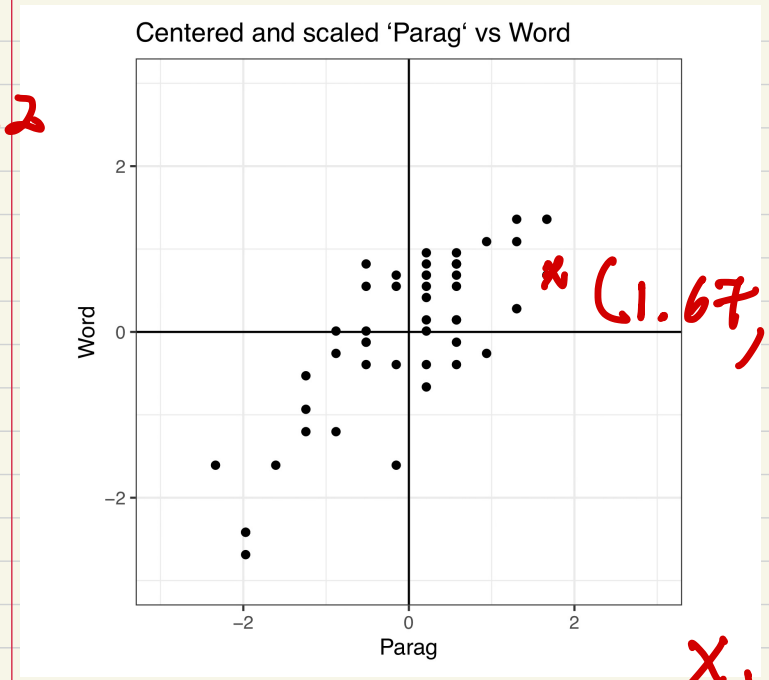
Word: Ave = 24.9 SD = 7.41

Centered & Scaled

$X_1 = \text{Parag\_center\_scaled} = \dfrac{\text{Parag} - 10.4}{2.75}$

$X_2 = \text{word\_center\_scaled} = \dfrac{\text{Word} - 24.9}{7.41}$

X_2

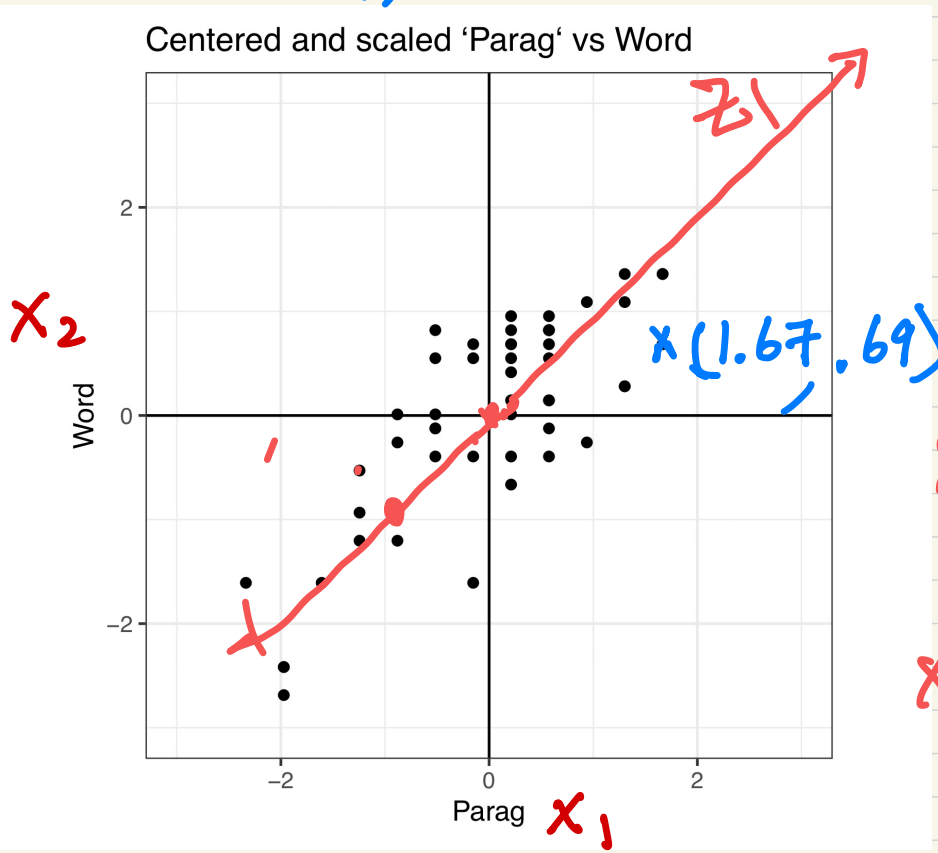


(1.67, .69)

X_1

Q: (1.67, .69) ?

First: Center and scale each x_1, x_2



x_2 x_1 Z_1 $x(1.67, .69)$

For simplicity consider $n = 50$ people

Original

Parag: Ave $= 10.4$ SD $= 2.75$

Word: Ave $= 24.9$ SD $= 7.41$

Centered & Scaled

$$x_1 = \text{Parag\_center\_scaled} = \frac{\text{Parag} - 10.4}{2.75}$$

$$x_2 = \text{word\_center\_scaled} = \frac{\text{Word} - 24.9}{7.41}$$

$$PC1: \; Z_1 = \varphi_{11} x_1 + \varphi_{21} x_2$$

Z_1: a new score

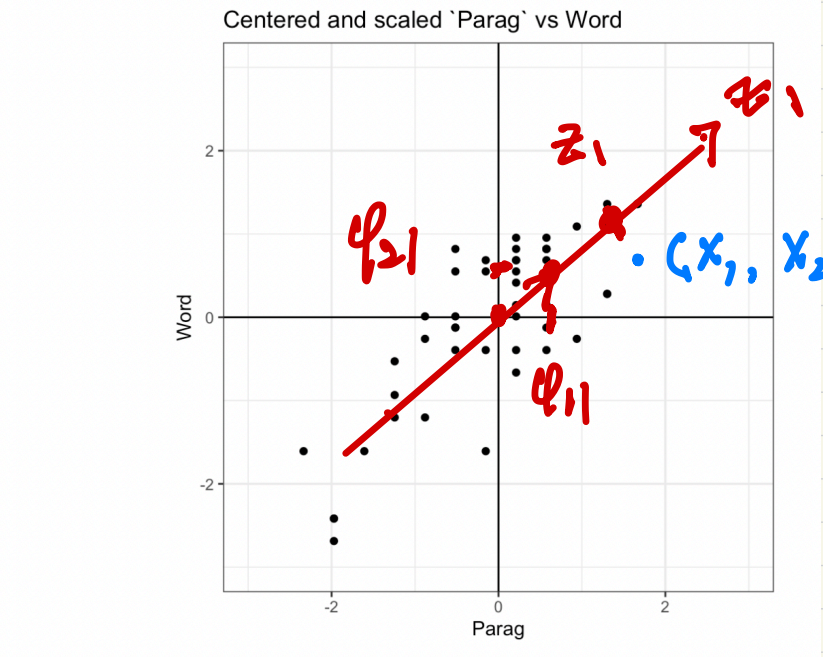
(x_1, x_2): sits on the line

$(\varphi_{11}, \varphi_{21})$: Loading

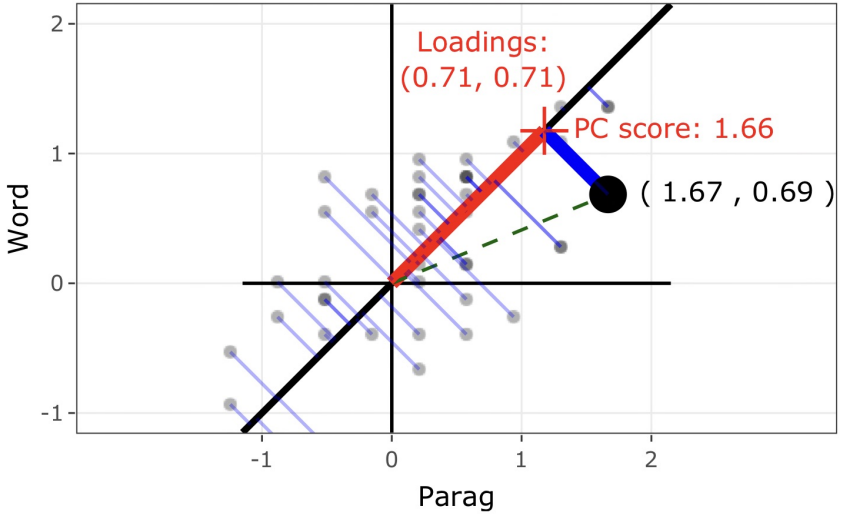
$$\max_{\varphi_{11}, \varphi_{21}} \; Var(Z_1) = \max_{\varphi_{11}, \varphi_{21}} Var(\varphi_{11} x_1 + \varphi_{21} x_2)$$

$$\varphi_{11}^2 + \varphi_{21}^2 = 1$$

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2$$



Comments : Interpretations



In the above plot we want to demonstrate the following beautiful geometric interpretation of PCA.

Fact 1: A line which minimizes the total squared distance must go through the origin (or sample means)

Fact 2: By the Pythagorean theorem, for any point:

$$\text{PC score}^2 + \text{Perpendicular distance}^2 = \text{Distance to origin}^2$$

Solution : $\phi_{11} = .71, \phi_{21} = .71$

PC1 : $Z_1 = .71 X_1 + .71 X_2$

ex. $(X_1, X_2) = (1.67, .69)$

PC1 : $Z_1 = .71 \times 1.67 + .71 \times .69$

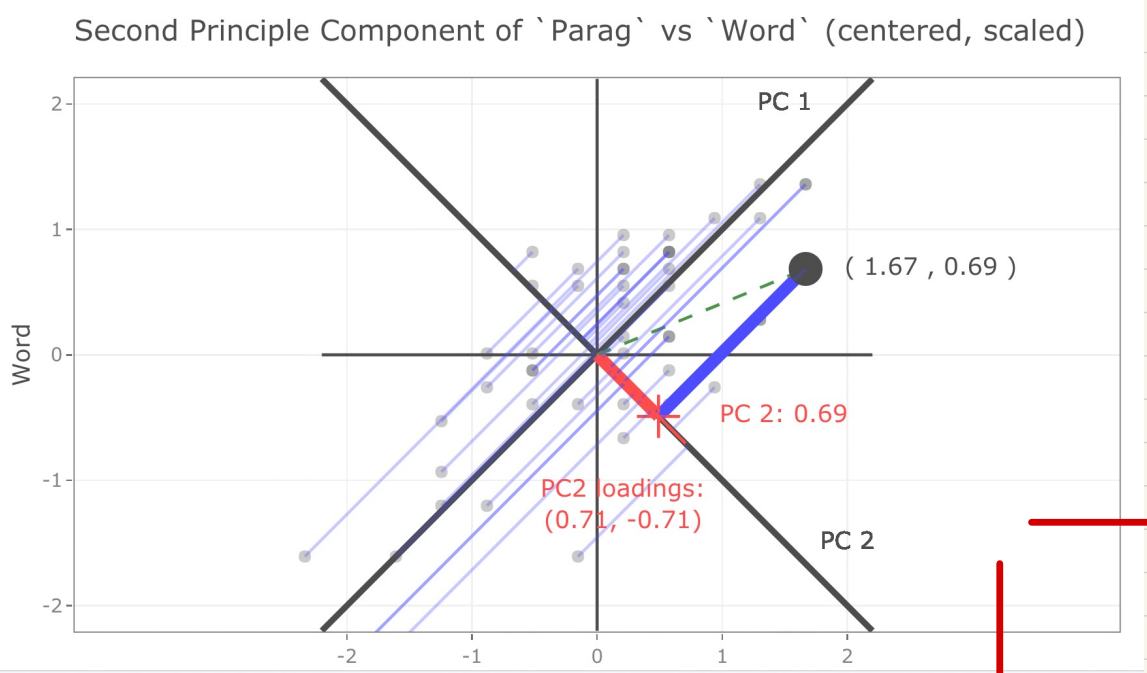
$= 1.66$

Summary :- PC1 : linear combination of X_1, X_2

Interpretation I : Var(Z_1) is the largest

Interpretation II : The total perpendicular distance^2 is the smallest

More PC's



Given PC1, Look for PC2:

$$Z_2 = \phi_{12} X_1 + \phi_{22} X_2$$

$$\max Var(Z_2)$$

$$\phi_{12}^2 + \phi_{22}^2 = 1$$

$$(\phi_{11}, \phi_{21}) \perp (\phi_{12}, \phi_{22})$$

Solution Loadings

PC2 : $\phi_{12} = .71$ $\phi_{22} = -.71$

PC2 : $Z_2 = .71 X_1 - .71 X_2$

$(X_1, X_2) = (1.67, .69)$ $PC2 = Z_2 = .71 \times 1.67 - .71 \times .69$

$= .7$

How to get PC1 ?

R-func : prcomp( )

components "PC1" and "PC2":

```{r}
pc.parag.word <- prcomp(parag.word.scaled.centered)
names(pc.parag.word)
pc.parag.word$rotation  # loadings
#pc.parag.word$x  # PC scores
data.frame(parag.word, parag.word.scaled.centered,
pc.parag.word$x)[1:10, ] #parag.word: original data
```

	PC1	PC2
Parag	0.707	0.707
Word	0.707	-0.707

Description: df [10 × 6]

centered / scaled

	Parag <int>	Word <int>	Parag.1 <dbl>	Word.1 <dbl>	PC1 <dbl>	PC2 <dbl>
p1	9	22	-0.517	-0.3938	-0.644	-0.0868
p2	9	25	-0.517	0.0108	-0.358	-0.3729
p3	10	29	-0.153	0.5503	0.281	-0.4972
p4	15	35	1.666	1.3595	2.140	0.2169
p5	14	33	1.302	1.0898	1.692	0.1504
p6	11	26	0.211	0.1457	0.252	0.0462
p7	11	30	0.211	0.6852	0.634	-0.3353
p8	9	24	-0.517	-0.1241	-0.453	-0.2776
p9	12	26	0.575	0.1457	0.509	0.3035
p10	12	24	0.575	-0.1241	0.319	0.4942

Recover all PC1's.

P1: $PC1 = .707 \times (-.517) + .707 \times (-.394)$

$= .644$

skim_variable	numeric.mean	numeric.sd
<chr>	<dbl>	<dbl>
Word	24.9	7.41
Parag	10.4	2.75
Math	12.2	5.69
Arith	17.0	7.19

from the original data.

$$PC1 = .707\left(\frac{9 - 10.4}{2.75}\right) + .707\left(\frac{22 - 24.9}{7.41}\right)$$

$$= \frac{.707}{2.75} \times 9 + \frac{.707}{7.41} \times 22 - c$$

Weighted Sum of Parag & Word

Fancy word

PC1, PC2

Summary: Turn $(X_1, X_2) \rightarrow (Z_1, Z_2)$

- Z_1, Z_2 are linear comb. of X_1 and X_2
- $Var(Z_1) \geq Var(Z_2) \geq \begin{cases} Var X_1 \\ Var X_2 \end{cases}$
- $Z_1 \perp Z_2$. $Cor(Z_1, Z_2) = 0$
- No information lost using Z_1, Z_2
- May use Z_1 alone if $Var(Z_2)$ is much smaller

2. AFQT: Dim 4 : $x_1, x_2, x_3, x_4 \Rightarrow z_1, z_2, z_3, z_4$

We are ready to extend to PC's for more variables.

AFQT : x_1, x_2, x_3, x_4.

PC's : $z_1 = \varphi_{11} x_1 + \varphi_{21} x_2 + \varphi_{31} x_3 + \varphi_{41} x_4$

$\vdots$

$z_4 = \varphi_{14} x_1 + \varphi_{24} x_2 + \varphi_{34} x_3 + \varphi_{44} x_4$

R-functin:

prcomp( )

3. Goto lecture

1) What does each PC score mean?
2) How many PCs should be used?
3) What interesting facts can be revealed by PCs?

1) Interpretations of PC's

a) skim(data.AFQT) %>% select(skim_variable, numeric.mean, numeric.sd)

skim_variable <chr>	numeric.mean <dbl>	numeric.sd <dbl>
Word	24.9	7.41
Parag	10.4	2.75
Math	12.2	5.69
Arith	17.0	7.19

default = FALSE

b) Prcomp(data.AFQT, scale = True, center = True)

↑ default

rotation : loadings

x : PC scores.

```{r loading}
pc.4.loading <- pc.4$rotation   #pc.4$x
knitr::kable(pc.4.loading)
```

	PC1	PC2	PC3	PC4
Word	0.509	-0.442	0.365	-0.642
Parag	0.500	-0.546	-0.308	0.598
Math	0.496	0.483	-0.652	-0.310
Arith	0.494	0.523	0.589	0.368

All Centered, Scaled

PC1 = .5 Word + .5 Parag + .5 Math + .5 Arith

= .5(Word + Parag + Math + Arith)

= .5 Sum (4 tests) !!!

PC1 : Weighted Sum of 4 tests.

$$PC2 \approx .5(\text{Math} + \text{Arith} - (\text{Word} + \text{Parag}))$$

$$= .5(\text{Sum(Math)} - \text{Sum(English)})$$

!!!

Remark 1: In the original scores

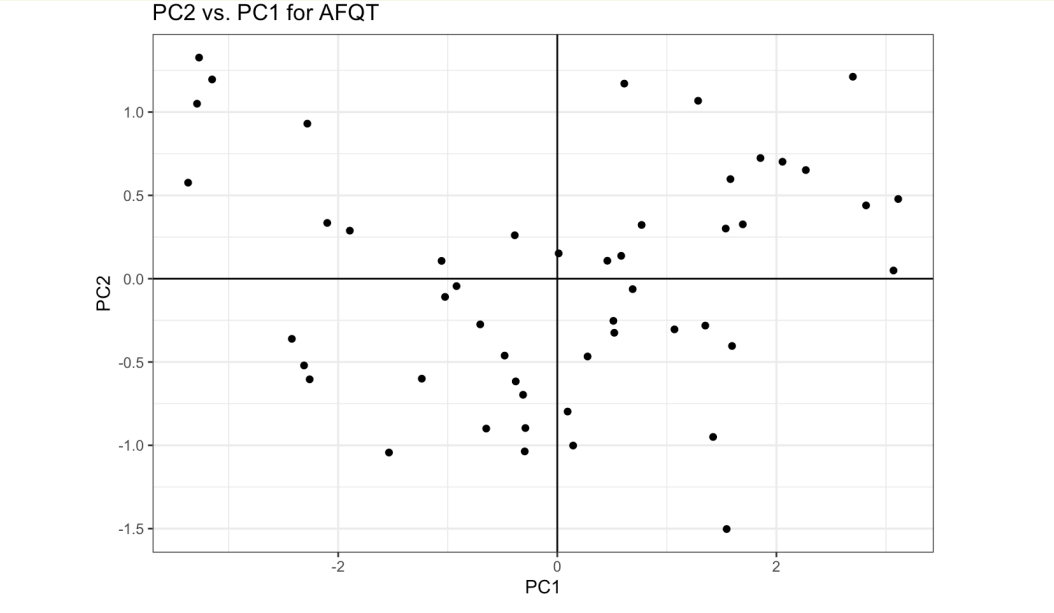
$$PC1 = .5\left(\left(\frac{\text{Math}}{5.69} + \frac{\text{Arith}}{7.19}\right) - \left(\frac{\text{Word}}{7.41} + \frac{\text{Parag}}{2.75}\right)\right) + c_1$$

inverse prop to each sd!

Known constant

Remark 2: Loadings are unique upto signs.

c) plot(PC1, PC2) : informative



2) How many PC's should be used?

a) PVE: Proportion of Variance Explained

Var(PC) / Total Variances

```{r}
summary(pc.4)$importance  #notice it is from summary()  n
```

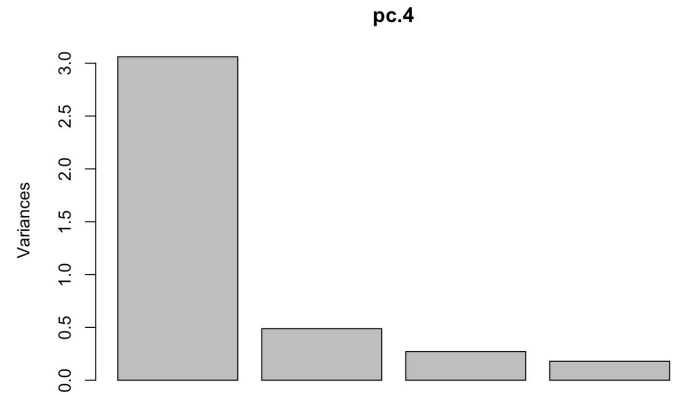
```
                          PC1   PC2    PC3    PC4
Standard deviation     1.750 0.699 0.5208 0.4238
Proportion of Variance 0.765 0.122 0.0678 0.0449
Cumulative Proportion  0.765 0.887 0.9551 1.0000
```

→ 3.06, .49, .27, .18

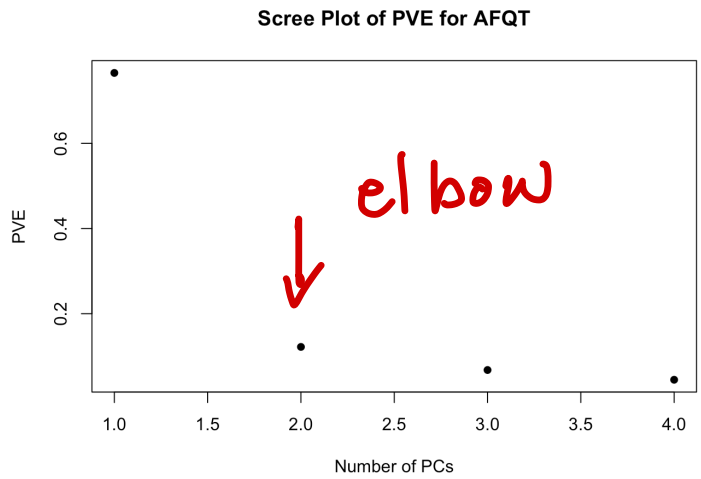
b) plot(pc.4)

Scree plots

```
plot(pc.4) # variances of each pc
```




```
plot(summary(pc.4)$importance[2, ],  # PVE
     ylab="PVE",
     xlab="Number of PCs",
     pch = 16,
     main="Scree Plot of PVE for AFQT")
```

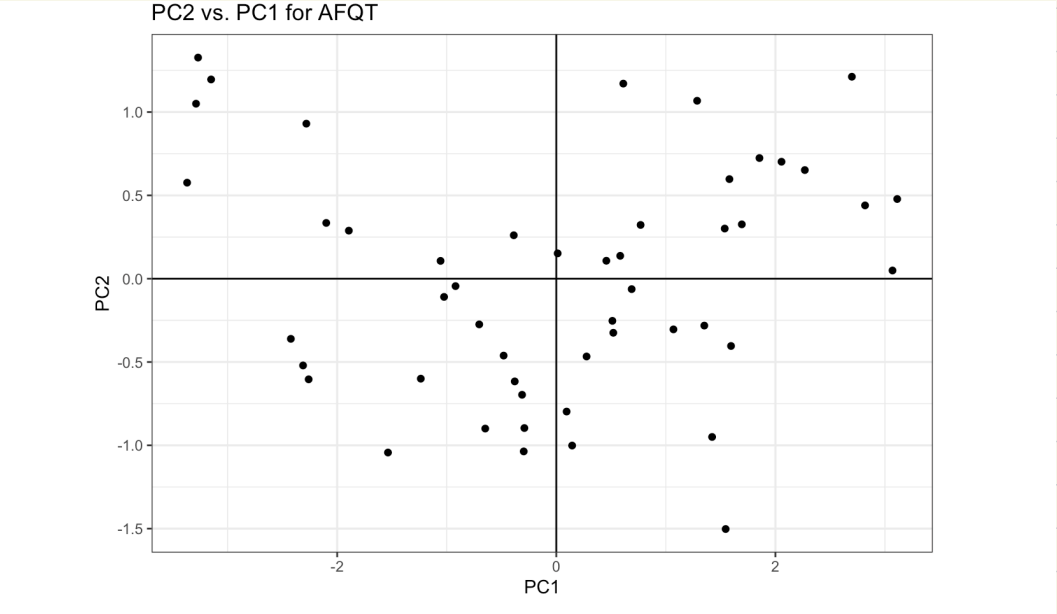


Elbow rule:

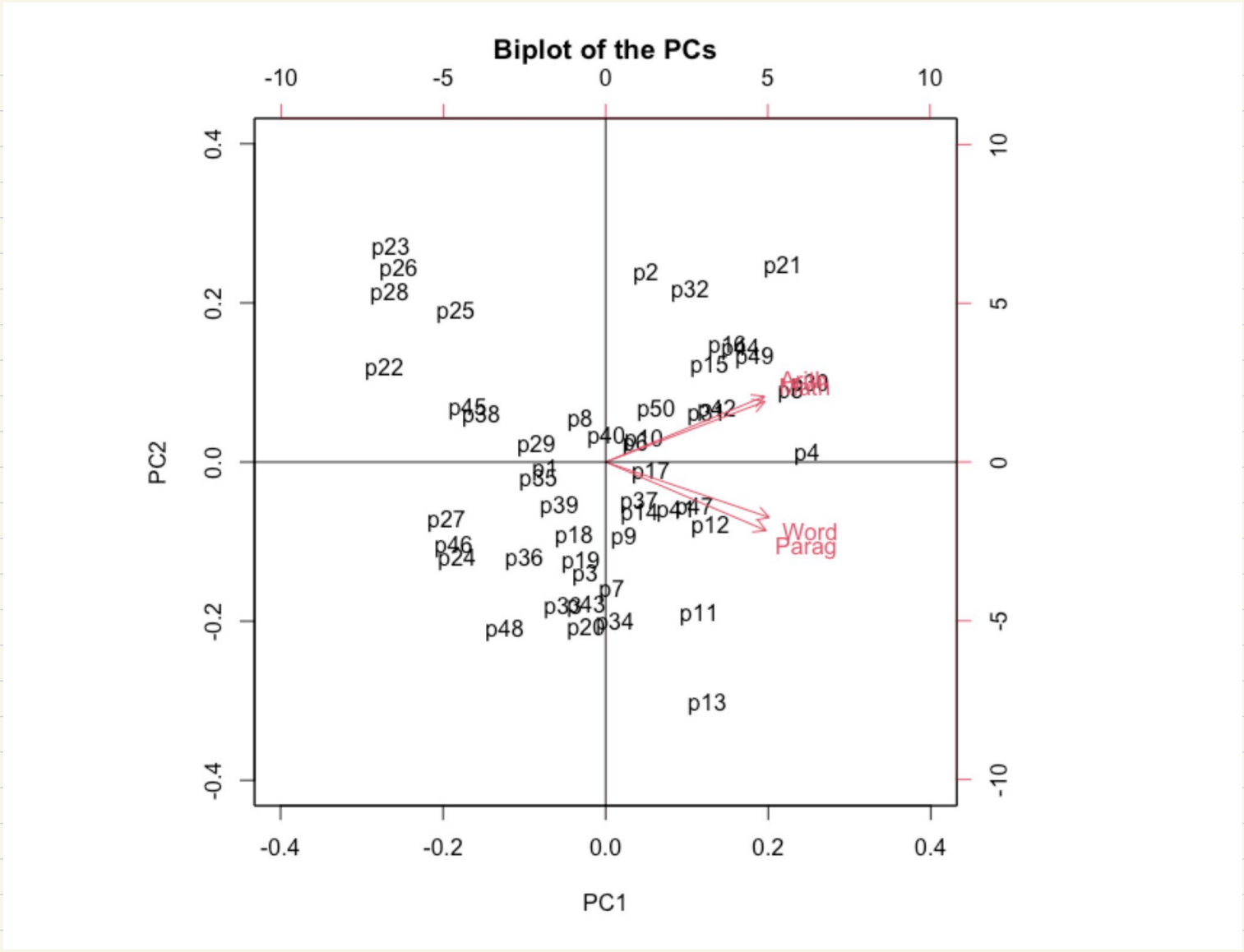
Point: the curve levels off afterwards

So we take two PC's

c) plot (PC1, PC2) : see the variabilities of PC1, vs. PC2

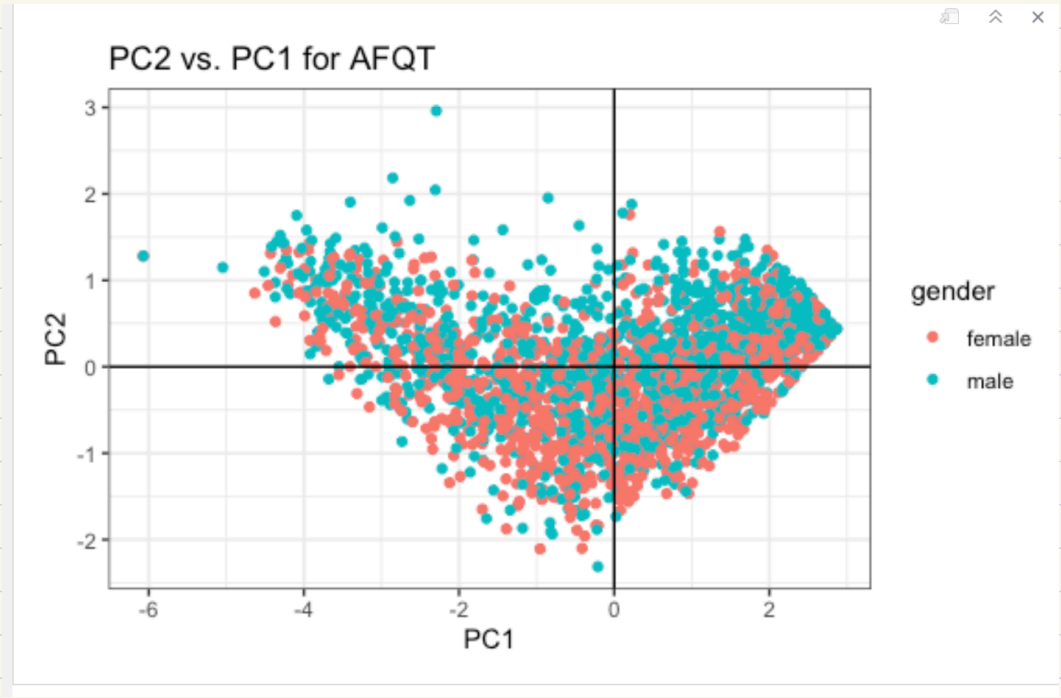


d) Biplot


Biplot of the PCs
PC1
PC2
-10
-5
0
5
10
-0.4
-0.2
0.0
0.2
0.4
p23
p26
p28
p25
p22
p2
p21
p32
p15
p49
p45
p38
p8
p50
p42
p29
p40
p10
p4
p1
p35
p17
p39
p37
p14
p47
p12
p27
p46
p24
p18
p9
p36
p19
p3
p7
p43
p11
p48
p20
p34
p13
Word
Parag

3) a) Interesting findings?

Differences between males & females??



b) How is AFQT created??

4. a) How to find the loadings $\varphi_1 = \begin{pmatrix} \varphi_{11} \\ \varphi_{21} \end{pmatrix}$?

Facts: Let $\bar{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \\ x_{501} & x_{502} \end{pmatrix}_{n \times 2}$

Then φ_1 is the leading eigen-vector of

$$\frac{(\bar{X}^T \cdot \bar{X})}{n-1}$$ cov. matrix

i.e. $$\left(\frac{\bar{X}^T \cdot \bar{X}}{n-1}\right) \varphi_1 = \lambda_1 \cdot \varphi_1$$

↑ largest eigenvalue

↖ eigenvector

Prof:

$$\max_{\|\varphi_1\|^2=1} \quad Var(Z_1) = \varphi_1^T \cdot \frac{\bar{X}^T \cdot \bar{X}}{n-1} \cdot \varphi_1$$

$$\Rightarrow \quad \frac{\bar{X}^T \cdot \bar{X}}{n-1} \cdot \varphi_1 = \lambda \varphi_1$$

- λ_1 : largest eigen-value
- $\boxed{Var(Z_1) = \lambda_1}$

b) PCA secret: SVD (Singular Value Decomposition)

Take AFQT: X_1 = Word, X_2 = Paragr, X_3 = Math
X_4 = Arith

Centered, Scaled

The data matrix

$$\underline{X} = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 \\ \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

$n = 50$, $p = 4$

Fact 1: SVD: There exist $v_1, \cdots v_4$

$u_1, u_2, u_3, u_4 : n \times 1 = 50 \times 1$, $|d_1| > |d_2| > \cdots |d_4|$ 4x1

$$X = d_1 \cdot u_1 \cdot v_1^T + d_2 u_2 v_2^T + d_3 u_3 v_3^T + d_4 u_4 v_4^T$$

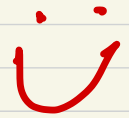
- $u_1 \cdots u_4$ Orthonormal
- $v_1, \cdots v_4$ " "

Fact 2: PC's $\Leftrightarrow$ SVD

- (v_1, v_2, v_3, v_4): PC loadings.

- $PC1 = d_1 u_1 \leftarrow \bar{X} \cdot v_1$
  $PC2 = d_2 u_2 \leftarrow \bar{X} \cdot v_2$
  $\vdots$
  $PC4 = d_4 u_4 \leftarrow \bar{X} \cdot v_4$

- $\lambda_1 = d_1^2/n-1 \quad = Var(PC1)$
  $\vdots \qquad\qquad\qquad\quad \vdots$
  $\lambda_4 = d_4^2/n-1 \quad = Var(PC4)$



5. Recap PCA

Given a data with variables

$x_1, x_2, \cdots$ say $x_{p=10}$

- PC's are linear Comb. of $x_1, \cdots x_{10}$

$$PC1 = \varphi_{11} x_1 + \varphi_{21} x_2 \cdots + \varphi_{10,1} x_{10}$$

Properties

1. - $(\varphi_{11}, \varphi_{21}, \cdots \varphi_{10,1})$: loadings
   - Var(PC1) maximized
   - loadings / PC1 unique up to the sign

- ϕ_{11} prop Cor $(PC1, X_1)$

⋮

2. PC2, PC3 ... PC10 max #.

$$Var(PC_1) \geq Var(PC_2) \geq \cdots \geq Var(PC_{10})$$

3. - All loadings are uncorr.
   - All PC's ≥

4. How many PC's ?

PVE. $$\frac{Var(PC_1)}{Var(PC_1) + \cdots + Var(PC_{10})}$$

5. How to get PC loadings / PC's

- Prcomp( ) → plot
  → biplot
- svd( ) : v loadings, pc = d.u
- irlba( ) much faster

more to come in module_clustering