SVD used in LSI

Oracle Li

潛在語義分析（Latent Semantic Analysis），是語義學的一個新的分支。傳統的語義學通常研究字、詞的含義以及詞與詞之間的關係，如同義，近義，反義等等。潛在語義分析探討的是隱藏在字詞背後的某種關係，以字詞的使用環境作為最基本的參考。人們找到了一種簡單的數學模型，這種模型的輸入是由任何一種語言書寫的文獻構成的文庫，輸出是該語言的字、詞的一種數學表達（向量）。字、詞之間的關係乃至任何文章片斷之間的含義的比較就由這種向量之間的運算產生。

潛在語義學的觀念也被應用在資訊檢索上，所以有時潛在語義學也被稱為隱含語義索引（Latent Semantic Indexing，LSI）。

隱含語義索引 (latent semantic indexing)，簡稱 LSI，目的是探討隱藏在字詞背後的某種關係，參考字詞的使用環境。LSI 的運作理論建基於奇異值分解，也就是以向量空間為分析模型，利用基底來呈現語料庫中不同字詞以及文件之間的關係。

1.preprocess the data

Load in data searched “R” and “Bayesian” from NCCU library,which contained 101 books as our data for analysis.

The chararcters starting with “%T” is books names in the same row, after “/” is auther names

```
lib=read.csv("C:\\Users\\Gene\\Desktop\\Library\\lib.csv",header=F,colClasses="character")
head(lib)
```

```
##                                                                                   V1
## 1                                                                   %A Aizaki, Hideo
## 2 %T Stated preference methods using R / Hideo Aizaki, Tomoaki Nakatani, Kazuo Sato
## 3                                    %@ 9781439890479 (hardcover : acid-free paper)
## 4                                       %@ 1439890471 (hardcover : acid-free paper)
## 5                                                       %O "A Chapman & Hall book."
## 6                  %O Includes bibliographical references (pages 211-233) and index
```

Delete the unknown book name

```
## [1] "%T Mr. Sh-------n's apology to the town; with the reasons which unfortunately induced him to his late
```

take out the book names row

```
## [1] "%t stated preference methods using r / hideo aizaki, tomoaki nakatani, kazuo sato"
## [2] "%t growth curve analysis and visualization using r / daniel mirman"
## [3] "%t a primer in biological data analysis and visualization using r / gregg hartvigsen"
```

split out the books names

```
## [1] "stated preference methods using r"
## [2] "growth curve analysis and visualization using r"
## [3] "a primer in biological data analysis and visualization using r"
```

delete some meaningless words

```r
nam[c(1:11,15,19,45,72,86,110,132,160,163,164,191,199,240,244,248,258,263)]
```

```
##  [1] "(responsibility)"
##  [2] ":"
##  [3] "[electronic"
##  [4] "="
##  [5] "18,"
##  [6] "2010"
##  [7] "2014"
##  [8] "2014,"
##  [9] "3-8,"
## [10] "9"
## [11] "a"
## [12] "an"
## [13] "and"
## [14] "by"
## [15] "em,"
## [16] "for"
## [17] "in"
## [18] "ma,"
## [19] "non-"
## [20] "o."
## [21] "of"
## [22] "r."
## [23] "resource]"
## [24] "the"
## [25] "to"
## [26] "usa,"
## [27] "with"
## [28] "馬可夫鏈蒙地卡羅收斂的研究與貝氏漸進的表現"
```

wow is a 235x100 matrix,whose rows are vocabularies and columns are books called term-document matrix

```r
wow[1:10,1:10]
```

```
##              [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## accuracy        0    0    0    0    0    0    0    0    0     0
## adaptive        0    0    0    0    0    0    0    0    0     0
## advanced        0    0    0    0    0    0    0    0    0     0
## analysis        0    1    1    0    1    0    0    0    1     0
## analytics       0    0    0    0    0    0    0    0    0     0
## analyzing       0    0    0    0    0    0    0    0    0     0
## applications    0    0    0    0    0    0    0    0    0     0
## applied         0    0    0    0    0    0    0    0    0     0
## approach        0    0    0    0    0    0    0    0    0     0
## approaches      0    0    0    0    0    0    0    0    0     0
```

2.singular value decomposition

(1) explanation

a.

對應第 j 個奇異值，uj 稱為左奇異向量，vj 稱為右奇異向量

t(wow) * wow * vj= dj^2 * vj

wow * t(wow) * uj= dj^2 * uj

b.

rank(wow)=r

wow=d1 * u1 * t(v1) + d2 * u2 * t(v2) + … + dr * ur * t(vr)

上式 uj 和 vj 是 U 和 V 的行向量

c. do SVD

```
wowow=svd(wow)
names(wowow)
```

```
## [1] "d" "u" "v"
```

(2) sigular value(奇異值)

由大到小排列的 100 個奇異值 100x100 diagnol matrix

```
head(wowow$d)
```

```
## [1] 8.880785 7.265193 5.320462 4.463646 4.445275 3.894539
```

(3) term-concept matrix(字詞—概念矩陣)

U 的行向量構成一正交正規向量集合，故可作為“字詞空間的基底”235x100 matrix

```
wowow$u[1:5,1:5]
```

```
##                   [,1]        [,2]         [,3]         [,4]        [,5]
## accuracy  -0.008540039  0.01814153 -0.017404873  0.006853416 -0.01694220
## adaptive  -0.008149952  0.01694777 -0.016886340  0.006849284 -0.01158875
## advanced  -0.019067636  0.03878509 -0.015263768  0.003461674 -0.05333428
## analysis  -0.374586302 -0.04671234  0.385514852  0.410012142  0.52324990
## analytics -0.012956879 -0.01539498  0.004668582 -0.035606249 -0.01840545
```

(4) concept-documenr matrix(概念—文件矩陣)

V 的行向量也構成一正交正規向量集合，可作為“文件空間的基底” 100x100 matrix

```
wowow$v[1:5,1:5]
```

```
##              [,1]         [,2]        [,3]       [,4]         [,5]
## book1 -0.09344747 -0.071572455 -0.09588300 0.11830419 -0.139851048
## book2 -0.12559866 -0.097268319  0.07894659 0.20231612  0.009236766
## book3 -0.17042514 -0.123703385  0.10651401 0.04083655  0.063623877
## book4 -0.09156810 -0.072284406 -0.00795892 0.09827969 -0.207275003
## book5 -0.04915688 -0.001690364  0.10340314 0.15841256  0.186909881
```

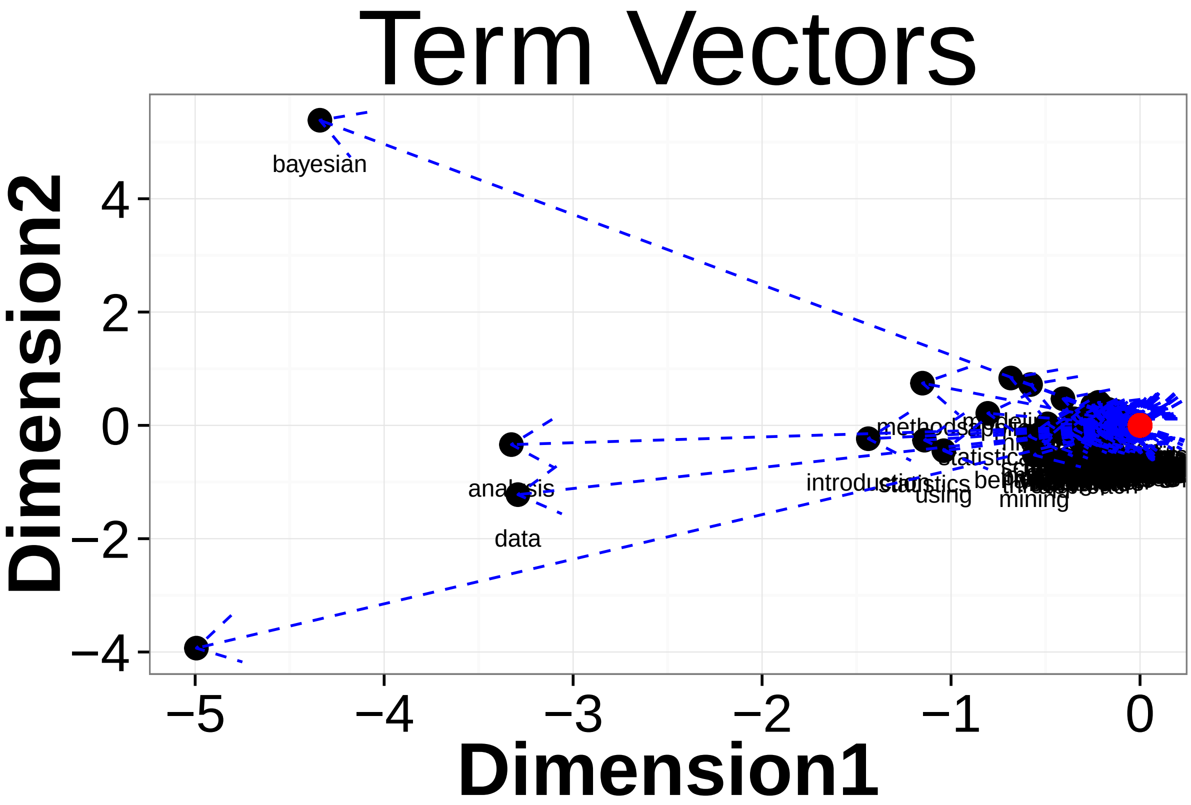
U 的行為 term vectors

V 的行為 document vectors

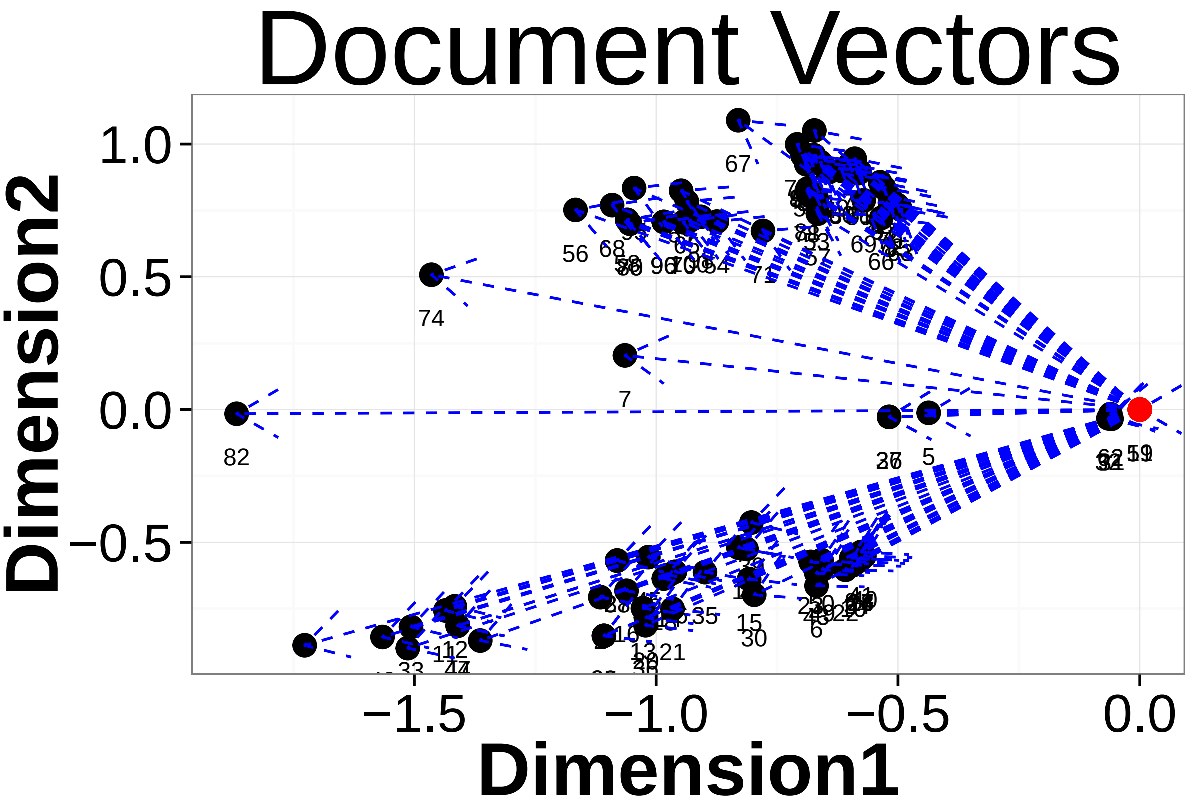
3.some easy visualization

(1) 2 Dimension(ggplot)

plot 2D of Term Vectors(U)

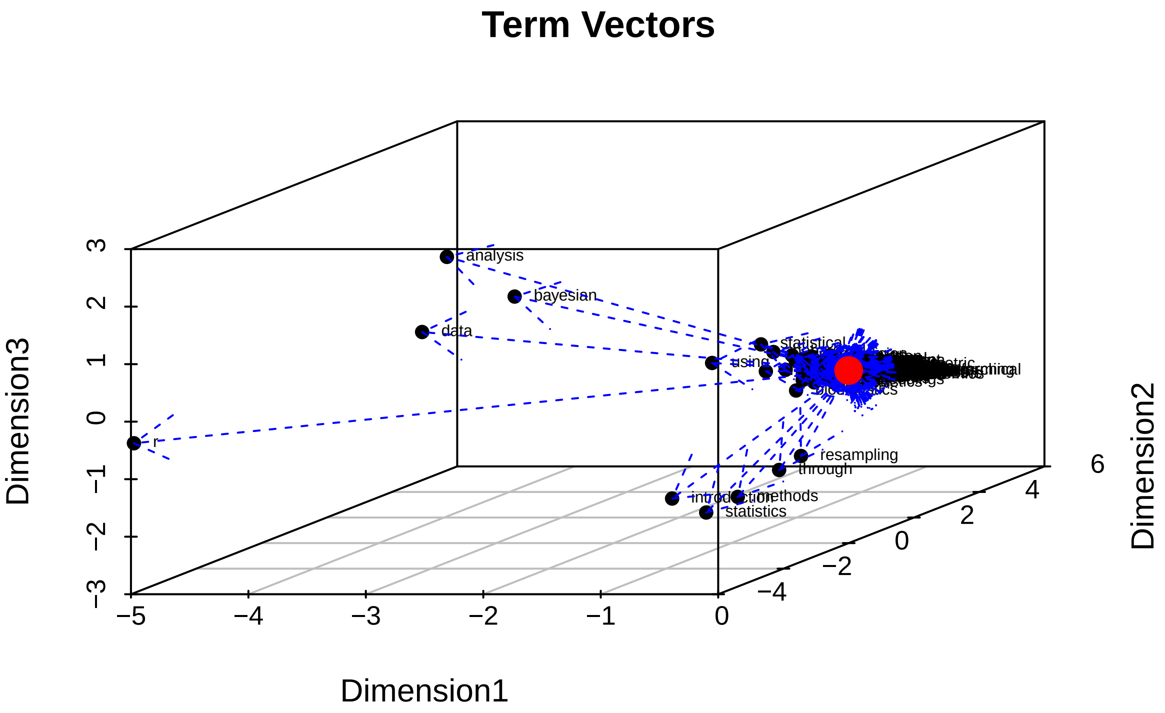


plot 2D of Document Vectors(v)

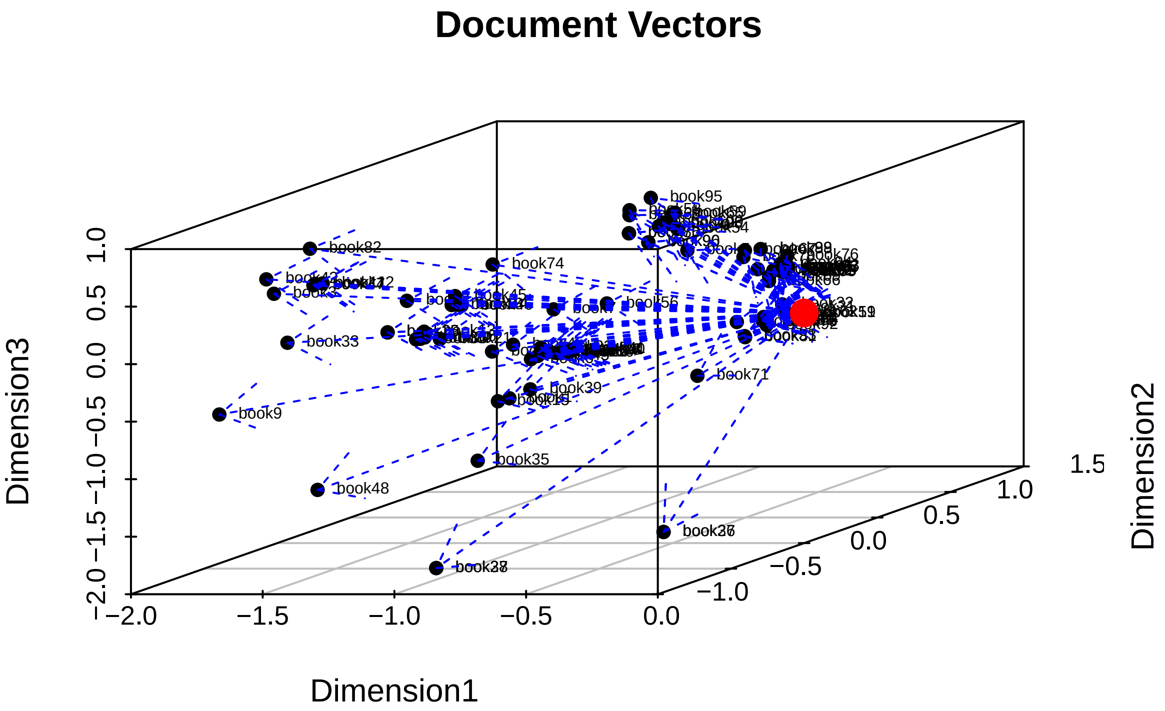


(2) 3 Dimension (scatterplot3d)

plot 3D of Term Vectors(U)



plot 3D of Document Vectors(v)



4. a pakage for LSA called lsa

this package can do all the things above and below

also can selected NO. Dimension ,textmatrix ,and query

wow made before contain documents in colums, terms in rows

```r
library(lsa)
hi=lsa(wow)
names(hi)
```

```
## [1] "tk" "dk" "sk"
```

“tk” 235*27

“dk” 100*27

“sk” 27*27

M = T S t(D)

拿之前的 wowow 來做也可以算出 27 個維度

```r
dimcalc_share()(wowow$d)
```

```
## [1] 27
```

so,we use 27D

5. Interpret

(1)function to caculate the angle of two vector

```
angle <- function(x,y){
      x=as.matrix(x)
      y=as.matrix(y)
      dot.prod <- t(x) %*% y
      norm.x <- norm(x,type="2")
      mat=dot.prod / norm.x
      theta <- sapply(mat,acos)
      as.numeric(theta)
}
```

(2) 字詞的相似度問題：字詞 ui 列和 uj 列有多相似？

→ 算夾角 statistics and methods 的夾角

```
angle(hi$tk[202,],hi$tk[124,])
```

```
## [1] 1.564347
```

statistics and data 的夾角

```
angle(hi$tk[202,],hi$tk[49,])
```

```
## [1] 1.6244
```

從圖看出 statistics and methods 明顯比 statistics and data 近夾角亦然

算倆倆的距離

```
www=matrix(0,ncol=235,nrow=235)
for(i in 1:235){
      for(j in i:235){
            www[j,i]=angle(hi$tk[i,],hi$tk[j,])
      }
      www[i,i]=0
}
```

最小夾角

```
wwwnu=as.numeric(www)
wwwnu=round(wwwnu,4)
table(wwwnu)[2]
```

```
## 0.8381
##      2
```

```
which(wwwnu==0.8381)
```

```
## [1]  359 8114
```

359 =235*1 + 124

8114 =235*34 + 124

```
www[124,2]
```

```
## [1] 0.8380532
```

```
www[124,35]
```

```
## [1] 0.8380532
```

```
nam[c(2,35,124)]
```

```
## [1] "adaptive" "clinical" "methods"
```

adaptive” “clinical” “methods” 此三詞最近? 何種擾動?

(3) 文件的相似度問題：文件 vi 列和 vj 列有多相似？

→ 算夾角

2: growth curve analysis and visualization using r

3: a primer in biological data analysis and visualization using r

```
angle(hi$dk[2,],hi$dk[3,])
```

```
## [1] 1.146402
```

2: growth curve analysis and visualization using r

51: innovation without r&d [electronic resource] : heterogeneous innovation patterns of non-r&d-performing firms in the german manufacturing industry

```
angle(hi$dk[2,],hi$dk[51,])
```

```
## [1] 1.570796
```

6.query

文件查詢問題：給出若干查詢字詞，哪些是最相關聯的文件？

Equation 1: AT = (U S VT)T = V S UT

Equation 2: AT U S-1 = V S UT U S-1

Equation 3: V[m,] = AT[m,] U S-1

For a given document vector d Equation 3 can be rewritten as

Equation 4: d = AT[m,] U S-1

Since in LSI a query is treated just as another document then the query vector is given by

Equation 5: q* = qT U S-1

Thus, in the reduced k-dimensional space we can write

Equation 6: q* = qT Uk Sk-1

qT is what we want to query

q* can be seen as kind of document vector, so we caulate angle of q * and all document vector

predict (new data) (tk) (sk-1)

predict 1 * n n * k k * k

"An Introduction to Statistical Learning with Applications in R"

將此 document 轉換成 query 形式

The query vector

```
head(t(qu),10)
```

```
##               [,1]
## accuracy         0
## adaptive         0
## advanced         0
## analysis         0
## analytics        0
## analyzing        0
## applications     1
## applied          0
## approach         0
## approaches       0
```

轉到參考基底

query * 單字的基底 * 奇異值的倒數

1 * 235 235 * 27 27 * 27 =1 * 27

```
(que=qu %*% hi$tk %*% solve(diag(hi$sk)))
```

```
##            [,1]        [,2]        [,3]       [,4]       [,5]       [,6]
## [1,] -0.1006644 -0.06065994 -0.08249454 0.02560868 -0.0494527 0.06323846
##           [,7]      [,8]      [,9]      [,10]        [,11]        [,12]
## [1,] 0.1144352 0.0189292 0.0534555 -0.1275054 2.800878e-16 -0.003230793
##            [,13]      [,14]      [,15]       [,16]       [,17]       [,18]
## [1,] -0.01586559 0.09413535 -0.1228509 -0.06291404 -0.01121483 -0.05053008
##           [,19]       [,20]      [,21]      [,22]        [,23]      [,24]
## [1,] -0.1754365 -0.05519499 0.03160384 -0.0576115 -0.009671259 0.08293052
##           [,25]       [,26]       [,27]
## [1,] 0.03381124 -0.02274268 -0.06379997
```

再計算 que 和 “dk” 100*27 100 本書的夾角

求出最小夾角

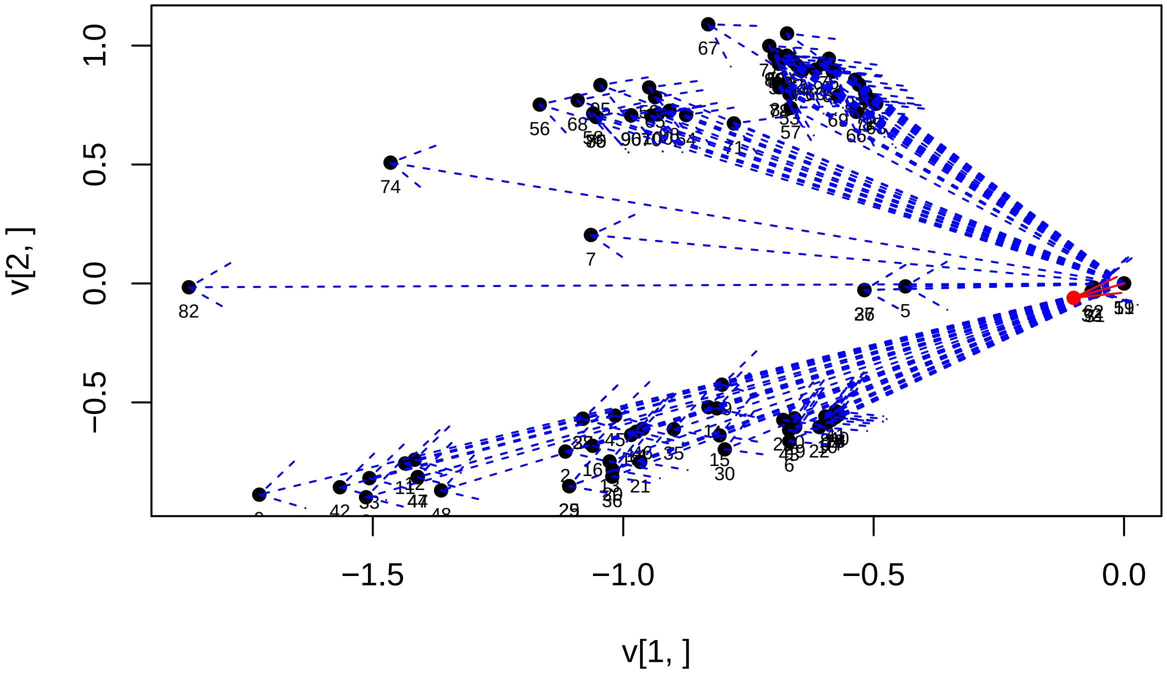
“bayesian item response modeling theory and applications”

“An Introduction to Statistical Learning with Applications in R”

```
all.angle=angle(t(que),t(hi$dk))
m=which.min(all.angle)
rbind(aa[m],q)
```

```
##   [,1]
##   "%t bayesian item response modeling [electronic resource] : theory and applications / by jean-paul fox"
## q "An Introduction to Statistical Learning with Applications in R"
```

畫畫看，但是在前兩個維度很荒謬



terms too much