Text Processing: Latent semantic indexing

-Applied Multivariate Analysis-

Lecturer: Darren Homrighausen, PhD

A repeated theme

Often, dimension reduction can be 'layered' between our original representations (in this case bag of words) and techniques that operate on that representation (in this case, finding relevant documents)

Original representation

[dimension reduction]

Statistical method

How does this relate to text processing?

PCA and factor analysis interpretation

- PCA: Find the directions of greatest variance. This doesn't on its face seem like it maintains correlations, but observe:

$$var(aX_1 + bX_2) = a^2 Var(X_1) + b^2 Var(X_2) + 2abCov(X_1, X_2)$$

If we standardize the matrix, then this reduces to

$$var(aX_1 + bX_2) = a^2 + b^2 + 2abCov(X_1, X_2)$$

This gets maximized over $a^2 + b^2 = 1$.

  - If $Cov(X_1, X_2) \approx 0$, then this gets maximized by any $a^2 + b^2 = 1$ (it doesn't matter)
  - If $Cov(X_1, X_2) \approx 1$, then this gets maximized by setting $a = b = 1/\sqrt{2}$

- Factor analysis: Defined by maintaining correlations.

So, in either case, we are really maintaining correlations

Graphical example of the phenomenon

```
library(mvtnorm)
sigma   = matrix(c(1,sig,sig,1),nrow=2)
nsweep  = 1000
outcome = matrix(0,nrow=nsweep,ncol=2)
for(sweep in 1:nsweep){
x                = rmvnorm(200,c(0,0),sigma)
out.pca          = prcomp(x,center=T,scale=F)
outcome[sweep,] = out.pca$rotation[,1]
}
plot(outcome,xlab='PC1',ylab='PC2')
```

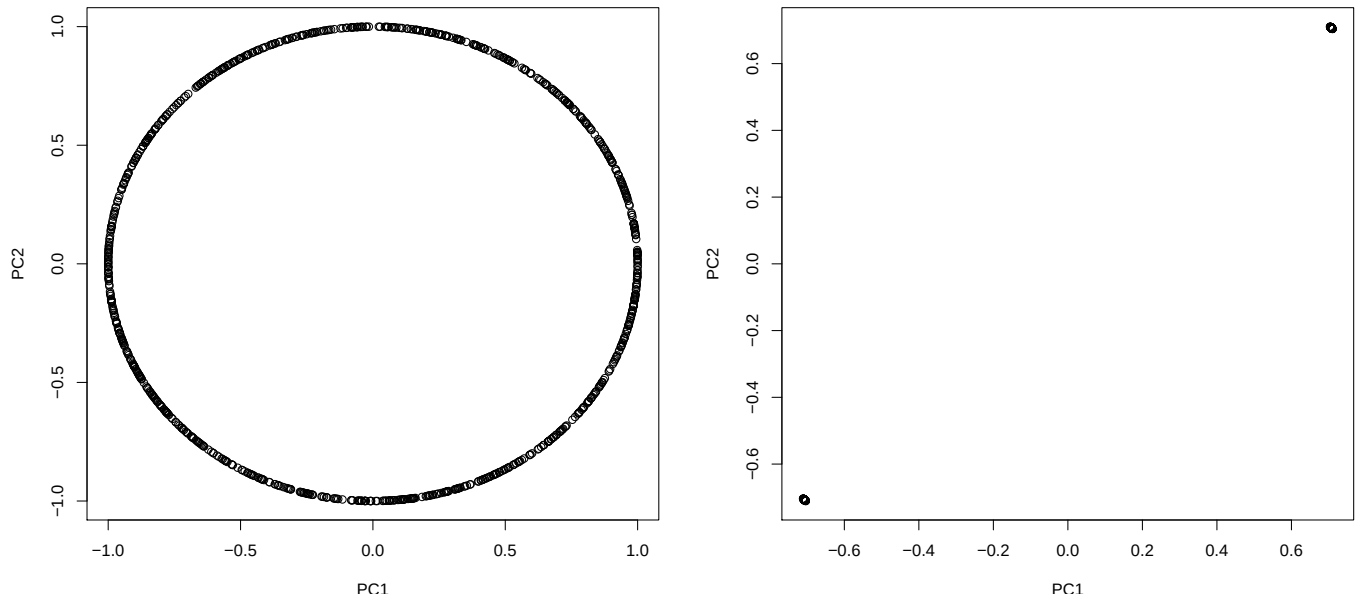


Figure: Left: `sig = 0`. Right: `sig = .999`

How does this apply to text processing?

Think about the document-term matrix $\mathbb{X}$.

The columns correspond to the words. If two words w_1, w_2 commonly appear together then the w_1^{th} and w_2^{th} columns of $\mathbb{X}$ are correlated

When we have a large number of documents that are about a topic, it is common to have some, or most, of the documents using related, but not identical words

Therefore, if we were to search $\mathbb{X}$ with a query, we would miss some of the important documents

An example

Suppose we have a corpus of documents

We wish to search for documents containing agriculture

We can query $Y = (\text{“agriculture”})$

However, “agriculture” is not regularly explicitly mentioned in articles about agriculture

This is where correlations come in. Whenever agriculture is mentioned, it will occur very frequently along with many synonyms (“farming”, for instance)

This is where latent semantic indexing comes in

An example, given to us by an invisible hand

To see why it is called latent semantic *indexing*, observe the following

When a book is written, a list of terms (or topics) is written down and an index is formed saying where these terms appear. For example, here is the start to the entry for "Agriculture" in the index to *The Wealth of Nations*

AGRICULTURE, the labour of, does not admit of such subdivisions as manufactures, 6; this impossibility of separation, prevents agriculture from improving equally with manufactures, 6; natural state of, in a new colony, 92; requires more knowledge and experience than most mechanical professions, and yet is carried on without any restrictions, 127; the terms of rent, how adjusted between landlord and tenant, 144; is extended by good roads and navigable canals, 147; under what circumstances pasture land is more valuable than arable, 149; gardening not a very gainful employment, 152–3; vines the most profitable article of culture, 154; estimates of profit from projects, very fallacious, *ib.*; cattle and tillage mutually improve each other, 220; . . .

An example, given to us by an invisible hand

It is asking a lot for a computer to do this.

However, if we only want to get the pages where "agriculture" is the topic (like, 6, 92, 152–3, 220..), then we can make a document-term matrix out of the pages of the book.

This approach will fail if we search this document-term matrix directly.

However, asking for pages that contain highly correlated words (like "rent") should work very well

Latent semantic indexing (LSI)

If we have our document-term matrix $\mathbb{X}$, then we write $\mathbb{X} = UDV^\top$, where

- The matrix U is the concept-document matrix (and maps into the document space)
- The matrix V is the term-concept matrix (and maps into the term space)

V is the matrix of loadings of the original words

If we have our query Y, we can map it into the document space by thinking of it as a new row in $\mathbb{X}$

$$\mathbb{X} = UDV^\top \quad \text{is the same as} \quad \mathbb{X}VD^{-1} = U$$

Which means we transform Y as

$$YVD^{-1}$$

To center or not to center?

If we think about LSI as performing PCA, then we should technically do the SVD like

$$\mathbb{X} - \overline{\mathbb{X}} = UDV^\top$$

However, observe the following example

$$A = \begin{bmatrix} a & b & 0 \\ d & 0 & f \\ 0 & h & i \end{bmatrix}$$

What happens if we column center A?

To center or not to center?

If we think about LSI as performing PCA, then we should technically do the SVD like

$$\mathbb{X} - \overline{\mathbb{X}} = UDV^\top$$

However, observe the following example

$$A = \begin{bmatrix} a & b & 0 \\ d & 0 & f \\ 0 & h & i \end{bmatrix}$$

What happens if we column center A?

We lose sparsity!

To center or not to center?

If we think about LSI as performing PCA, then we should technically do the SVD like

$$\mathbb{X} - \overline{\mathbb{X}} = UDV^\top$$

However, observe the following example

$$A = \begin{bmatrix} a & b & 0 \\ d & 0 & f \\ 0 & h & i \end{bmatrix}$$

What happens if we column center A?

We lose sparsity!

The consensus is to column center if your data is small enough, otherwise, don't worry about it

Example dataset

Let's look at $D = 20$ Reuters news articles about crude oil production and importation.

The corpus has 860 words

```
"Diamond Shamrock Corp said that\neffective today it
had cut its contract prices for crude oil by\n1.50 dlrs a
barrel.\n    The reduction brings its posted price for
West Texas\nIntermediate to 16.00 dlrs a barrel, the
copany said.\n    \"The price reduction today was made
in the light of falling\noil product prices and a weak crude
oil market,\" a company\nspokeswoman said.\n    Diamond
is the latest in a line of U.S. oil companies that\nhave cut
its contract, or posted, prices over the last two days\nciting
weak oil markets.\n Reuter"
```

Example dataset: R

Let's look at

```
mydtm = as.matrix(dtm.stem)

out.pca = prcomp(mydtm)

out.lsi = out.pca$rotation

signif(sort(out.lsi[,1],decreasing=TRUE)[1:24],2)
signif(sort(out.lsi[,1],decreasing=FALSE)[1:24],2)
```

PC loadings, for Reuters documents

```
> signif(sort(out.lsi[,1],decreasing=TRUE)[1:24],2)
    january       cubic    fiscales petroliferos  yacimientos     billion
     0.5000      0.3200      0.3200       0.3200       0.3200      0.2500
   argentine         gas     metrers       metres      natural    produced
     0.1600      0.1600      0.1600       0.1600       0.1600      0.1600
    totalled     barrels       added   production          pct         mln
     0.1600      0.0940      0.0860       0.0850       0.0780      0.0770
      output      budget      riyals   abdul-aziz  expenditure     revenue
     0.0630      0.0061      0.0061       0.0051       0.0051      0.0041
> signif(sort(out.lsi[,1],decreasing=FALSE)[1:24],2)
   posted    canada  canadian      west     power       bbl   lowered  texaco
   -0.050    -0.044    -0.044    -0.041    -0.040    -0.040    -0.036  -0.033
    texas    brings effective       dlrs     grade     sweet  contract    ship
   -0.033    -0.032    -0.032    -0.031    -0.031    -0.031    -0.031  -0.030
    price   changed       pay  postings  decrease   company benchmark     feb
   -0.030    -0.028    -0.028    -0.028    -0.027    -0.027    -0.026  -0.026
```

These are the 24 largest and smallest loadings on the first PC

What story can be told here?

PC loadings, for Reuters documents

```
> signif(sort(out.lsi[,1],decreasing=TRUE)[1:24],2)
     january        cubic     fiscales petroliferos  yacimientos      billion
      0.5000       0.3200       0.3200       0.3200       0.3200       0.2500
   argentine          gas      metrers       metres      natural     produced
      0.1600       0.1600       0.1600       0.1600       0.1600       0.1600
    totalled      barrels        added   production          pct          mln
      0.1600       0.0940       0.0860       0.0850       0.0780       0.0770
      output       budget       riyals   abdul-aziz  expenditure      revenue
      0.0630       0.0061       0.0061       0.0051       0.0051       0.0041
> signif(sort(out.lsi[,1],decreasing=FALSE)[1:24],2)
   posted    canada  canadian      west     power       bbl   lowered  texaco
   -0.050    -0.044    -0.044    -0.041    -0.040    -0.040    -0.036  -0.033
    texas    brings effective       dlrs     grade     sweet  contract    ship
   -0.033    -0.032    -0.032    -0.031    -0.031    -0.031    -0.031  -0.030
    price   changed       pay  postings  decrease   company benchmark     feb
   -0.030    -0.028    -0.028    -0.028    -0.027    -0.027    -0.026  -0.026
```

These are the 24 largest and smallest loadings on the first PC

What story can be told here?

- Large loadings correspond to things related to the international market
- Negative loadings correspond to the American/Canadian market

PC loadings, for TMNT documents

```
> signif(sort(out.lsi[,1],decreasing=TRUE)[1:24],2)
     cool      rude      dude     parti   raphael     -foot       --x
 5.0e-01   5.0e-01   5.0e-01   5.0e-01   3.6e-02  -8.1e-05  -8.1e-05
  acclaim acknowledg       add     adequ  administ    aerial     ahead
-8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05
  albiera  albrecht alessandra      amaz ambrosiana    amount    analys
-8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05
> signif(sort(out.lsi[,1],decreasing=FALSE)[1:24],2)
    turtl     comic donatello     paint     ninja   florenc  leonardo    mutant
-0.0180   -0.0100   -0.0085   -0.0081   -0.0076   -0.0073   -0.0070   -0.0063
     seri      statu       art      voic      anim      tmnt      game   brother
-0.0061   -0.0047   -0.0047   -0.0045   -0.0044   -0.0043   -0.0039   -0.0038
  charact  shredder     duomo    medici     casey  splinter    teenag     penni
-0.0037   -0.0036   -0.0035   -0.0034   -0.0032   -0.0032   -0.0032   -0.0032
```

These are the 24 largest and smallest projections onto the first PC

What story can be told here?

PC loadings, for TMNT documents

```
> signif(sort(out.lsi[,1],decreasing=TRUE)[1:24],2)
     cool      rude      dude     parti    raphael      -foot        --x
 5.0e-01   5.0e-01   5.0e-01   5.0e-01   3.6e-02  -8.1e-05  -8.1e-05
  acclaim acknowledg       add     adequ  administ     aerial     ahead
-8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05
  albiera  albrecht alessandra      amaz ambrosiana     amount     analys
-8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05  -8.1e-05
> signif(sort(out.lsi[,1],decreasing=FALSE)[1:24],2)
    turtl     comic donatello     paint     ninja   florenc  leonardo    mutant
-0.0180  -0.0100  -0.0085  -0.0081  -0.0076  -0.0073  -0.0070  -0.0063
     seri      statu       art       voic      anim      tmnt      game   brother
-0.0061  -0.0047  -0.0047  -0.0045  -0.0044  -0.0043  -0.0039  -0.0038
  charact  shredder     duomo    medici     casey  splinter    teenag     penni
-0.0037  -0.0036  -0.0035  -0.0034  -0.0032  -0.0032  -0.0032  -0.0032
```

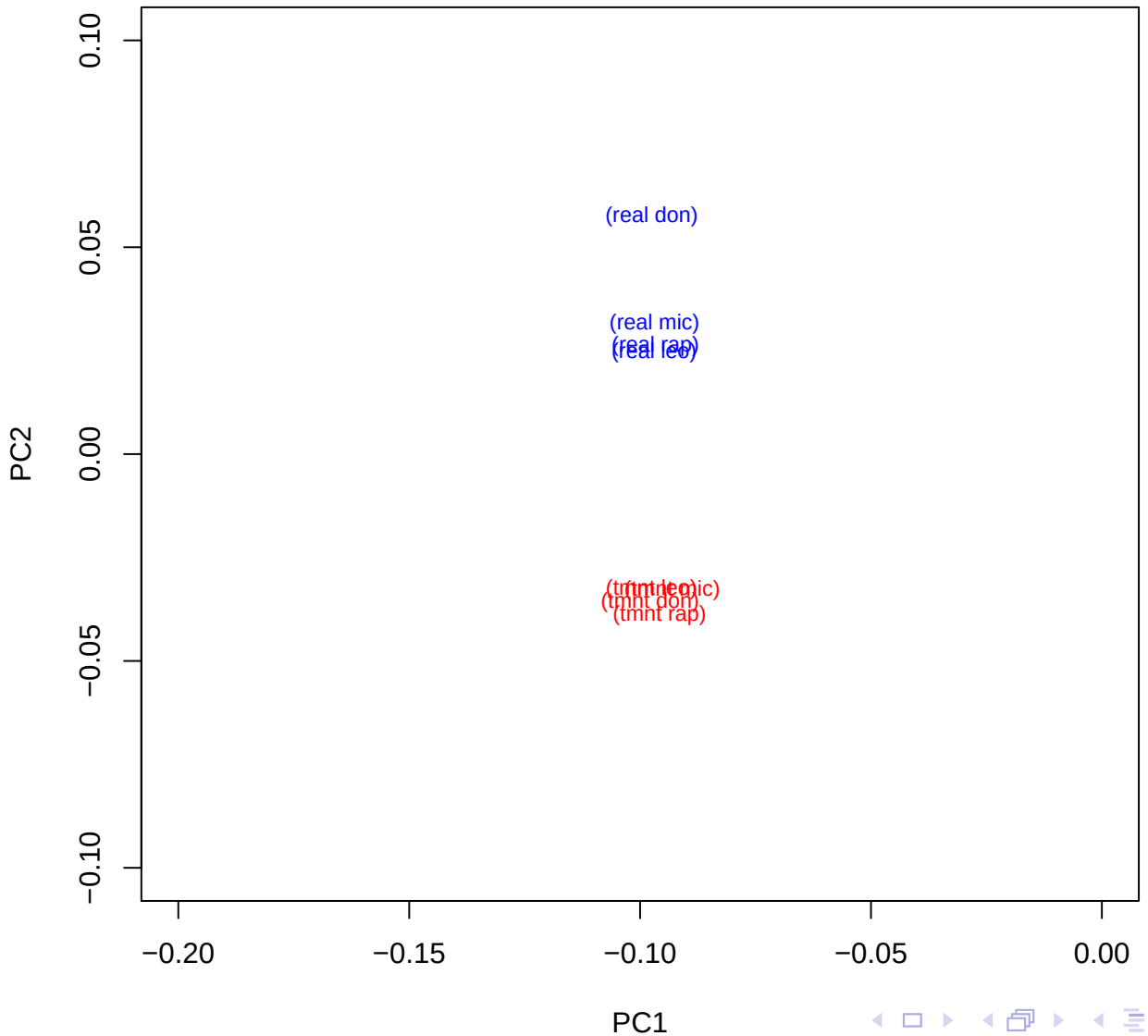
These are the 24 largest and smallest projections onto the first PC

What story can be told here?

- There is a substantial amount of overlap on the first PC

Let's look at a plot

Plot of PC scores for TMNT



PC loadings, for TMNT documents, second component

```
> signif(sort(out.lsi[,2],decreasing=TRUE)[1:24],2)
   florenc      statu      paint      duomo        art   basilica     medici
      0.22       0.15       0.14       0.13       0.11       0.08       0.08
     museo equestrian    madonna     execut        del      bronz     firenz
      0.07       0.07       0.07       0.07       0.07       0.06       0.06
    cosimo     judith    lorenzo    prophet      penni     vasari    centuri
      0.06       0.06       0.06       0.06       0.06       0.05       0.05
> signif(sort(out.lsi[,2],decreasing=FALSE)[1:24],2)
   turtl    comic    ninja   mutant     seri     voic     anim     tmnt   game
  -0.420   -0.240   -0.180   -0.150   -0.140   -0.100   -0.100   -0.100 -0.092
 charact  brother shredder    casey   teenag splinter     foot      sai  mirag
  -0.087   -0.083   -0.081   -0.077   -0.074   -0.074   -0.067   -0.066 -0.065
    movi    mikey   episod    mutat     clan     raph
  -0.063   -0.061   -0.056   -0.049   -0.046   -0.045
```

These are the 24 largest and smallest projections onto the first PC

What story can be told here?

PC loadings, for TMNT documents, second component

```
> signif(sort(out.lsi[,2],decreasing=TRUE)[1:24],2)
   florenc      statu      paint      duomo        art   basilica     medici
     0.220      0.150      0.140      0.130      0.110      0.083      0.080
     museo equestrian    madonna     execut        del      bronz     firenz
     0.074      0.073      0.071      0.070      0.069      0.063      0.062
    cosimo     judith    lorenzo    prophet      penni     vasari    centuri
     0.059      0.059      0.057      0.057      0.055      0.054      0.054
> signif(sort(out.lsi[,2],decreasing=FALSE)[1:24],2)
   turtl    comic    ninja   mutant     seri     voic     anim     tmnt   game
  -0.420   -0.240   -0.180   -0.150   -0.140   -0.100   -0.100   -0.100 -0.092
 charact  brother shredder    casey   teenag splinter     foot      sai  mirag
  -0.087   -0.083   -0.081   -0.077   -0.074   -0.074   -0.067   -0.066 -0.065
    movi    mikey   episod    mutat     clan     raph
  -0.063   -0.061   -0.056   -0.049   -0.046   -0.045
```

These are the 24 largest and smallest projections onto the first PC

What story can be told here?

- Positive values are related to the renaissance artists
- Negative values are related to the TMNTs

Distance to query using LSI

Original representation

[dimension reduction]

Statistical method

- Form the (normalized) document-term matrix $\mathbb{X}$
- Compute its LSI $\mathbb{X} = UDV^\top$
- Get Y into LSI via $\tilde{Y} = YVD^{-1}$
- Choose a K
- Find distances for documents $d = 1, \ldots, D$

$$\text{distance}(d, \tilde{Y}) = ||U_{d,1:k} - \tilde{Y}||_2$$

Distance to query using LSI

```
(tmnt leo) 0.9711807
(tmnt rap) 0.7696525
(tmnt mic) 0.7669749
(tmnt don) 0.9718710
(real leo) 0.9711512
(real rap) 0.9709391
(real mic) 0.9709492
(real don) 0.9734319
query      0.0000000
```

Wrap-up

Latent semantic indexing seeks to use correlations to help with document queries

This can be accomplished by forming the SVD of the document-term matrix $\mathbb{X}$. Alternatively, factor analysis is commonly used

The documents and the query get projected into a lower dimensional space where distances are computed

Some complications

- How do the normalization techniques we talked about affect this reduction?
- What happens if we want to make a new query, do we need to reform the entire SVD?
- How does factor analysis compare to PCA?