2019 DA Lecture 6. Principal Component Analysis: Method and Applications

- PCA and SVD: Data Approximation
- Conventional approach to PCA. Finding maximum positive eigenvalue. Comparison to that model based
- Applications
  - Ranking
  - Visualization on 2D plane
  - Latent Semantic Indexing
  - Google ranking in networks
- Homework 5

2019 DA Lecture 6. Principal Component Analysis: Method and Applications

- PCA and SVD: Data Approximation
- Conventional approach to PCA. Finding maximum positive eigenvalue. Comparison to that model based
- Applications
  - Ranking
  - Visualization on 2D plane
  - Latent Semantic Indexing
  - Google ranking in networks
- Homework 5

Principal Component Analysis 3: Singular value decomposition: Approximation 1

Data *X*: The singular values sorted $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_r$

SVD: $X = \mu_1 z_1 c_1' + \mu_2 z_2 c_2' + \cdots \mu_r z_r c_r'$

Rank is a mathematical explication of the space dimension.

Problem: Given data matrix X=[x_{iv}], find a matrix Xp=[xp_{iv}] of rank $p < r$ to minimize the sum-of-squares difference

$$\|X - Xp\|^2 = \sum_{i,v} (x_{iv} - xp_{iv})^2$$

Solution: The first p singular triplets,

$$\mathbf{Xp} = \mu_1 z_1 c_1' + \mu_2 z_2 c_2' + \cdots \mu_p z_p c_p'$$

Principal Component Analysis 3: Singular value decomposition: Approximation 2

Data *X*: The singular values sorted $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_r$

SVD: $X = \mu_1 z_1 c_1' + \mu_2 z_2 c_2' + \cdots \mu_r z_r c_r'$

Matrix of first *p* singular triplets

$Xp = \mu_1 z_1 c_1' + \mu_2 z_2 c_2' + \cdots \mu_p z_p c_p'$, $p < r$

minimizes the sum-of-squares difference

$\|X - Xp\|^2 = \sum_{i,v}(x_{iv} - xp_{iv})^2$ **over all matrices of rank *p***

Data scatter decomposition

$$\|X\|^2 = \mu_1^2 + \mu_2^2 + \ldots + \mu_p^2 + \|X - Xp\|^2$$

where $\|X\|^2 = \sum_{i,v} x_{iv}^2$, **the data scatter;**

μ_k^2 **- the contribution of *k*-th singular triplet**

5. Principal Component Analysis

* =

PCA methodology: Summary

- Principal component is a rescaled singular triplet
- Principal components optimally approximate the data according to the least squares criterion
- Principal components are orthogonal to each other
- Principal component's contribution is proportional to its squared singular value
- Principal components change at any data transformation, centering included
- Principal components of X are highly related to eigenvalues and eigenvectors of A=X′X

2019 DA Lecture 6. Principal Component Analysis: Method and Applications

- PCA and SVD: Data Approximation
- **Conventional approach to PCA. Finding maximum positive eigenvalue. Comparison to that model based**
- Applications
  - Ranking
  - Visualization on 2D plane
  - Latent Semantic Indexing
  - Google ranking in networks
- Homework 5

5. Principal Component Analysis: Conventional approach, 1a

Conventional definition:

Given a data matrix *X*, first principal component (PC):

a weighted combination *z* of *X* features after centering, that is,

$$z = Y * c,$$

that has the maximum variance with respect to all normed *c*

Second PC is defined similarly except for a condition that it must be orthogonal to the First PC; Third PC must be orthogonal to both first and second PCs, etc.

5. Principal Component Analysis: Conventional approach, 1b

Equivalent: $Y = X - repmat(mean, N, 1)$

Find c in $z=Y*c$ **to maximize** $\|z\|^2 = \sum_i z_i^2$, $\|c\| = 1$.

Why? Variance(*z*)= $\|z\|^2/N$**!!! Indeed, mean** $\bar{z} = 0$**, because *Y* is centered.**

Equivalent: Find *c* to maximize $\lambda = \frac{c^T Y^T Y c}{c^T c}$

(Rayleigh quotient).

Solution: maximum eigenvalue λ **of** $A=Y^TY$ **and corresponding eigen-vector *c***

5. Principal Component Analysis: Conventional approach, 1c

Equivalent: $Y = X - repmat(mean, N, 1)$

Find c in $z=Y*c$ **to maximize** $\|z\|^2 = \sum_i z_i^2$, $\|c\| = 1$.

Solution: maximum eigenvalue λ **of** $A=Y^TY$ **and corresponding eigen-vector** c.

How to find them? $|A-\lambda E|=0$? **NO.**

Power method: 1. Take any c0.

2. Find a=Ac0, c1=a/$\|a\|$

3. c1=c0, output $\lambda = \|a\|$ and c1.

Otherwise, c0$\Leftarrow$ c1, go to 2.

5. Principal Component Analysis: Conventional approach, 1d

Y= X – repmat(mean, N, 1)

Find maximum eigenvalue λ of $A=Y^TY$ and corresponding eigen-vector *c*.

Power method: 1. Take any c0.

2. Find a=Ac0, c1=a/$\|a\|$

3. c1=c0, output λ= $\|a\|$ and c1.

Otherwise, c0$\Leftarrow$ c1, go to 2.

Why?

Take spectral decomposition

$$A = \lambda_1 c_1 c_1^T + \lambda_2 c_2 c_2^T + \ldots + \lambda_r c_r c_r^T$$

5. Principal Component Analysis: Conventional approach, 1e

Find maximum eigenvalue λ of $A=Y^TY$ and corresponding eigen-vector *c*.

$$A = \lambda_1 c_1 c_1^T + \lambda_2 c_2 c_2^T + \ldots + \lambda_r c_r c_r^T$$

Multiply by itself:

$$A^2 = \lambda_1^2 c_1 c_1^T + \lambda_2^2 c_2 c_2^T + \ldots + \lambda_r^2 c_r c_r^T$$

n-1 times

$$A^n = \lambda_1^n c_1 c_1^T + \lambda_2^n c_2 c_2^T + \ldots + \lambda_r^n c_r c_r^T$$

$$A^n = \lambda_1^n [c_1 c_1^T + (\lambda_2/\lambda_1)^n c_2 c_2^T + \ldots + (\lambda_r/\lambda_1)^n c_r c_r^T]$$

5. Principal Component Analysis: Conventional approach, 1f

Find maximum eigenvalue λ of $A=Y^TY$ and corresponding eigen-vector *c*.

$$A = \lambda_1 c_1 c_1^T + \lambda_2 c_2 c_2^T + \ldots + \lambda_r c_r c_r^T$$

$$A^n = \lambda_1^n [c_1 c_1^T + (\lambda_2/\lambda_1)^n c_2 c_2^T + \ldots + (\lambda_r/\lambda_1)^n c_r c_r^T]$$

All the parentheses () $\Rightarrow$ 0 at increasing n. Then $\frac{A^n}{\lambda_1^n}$ $converges\ to\ c_1 c_1^T$**: All columns converge to c1 !!!**

5. Principal Component Analysis: Conventional approach, 2

Actually Computing the First Principal Component:

1. Given a N×V data matrix ***X***, compute its centered version ***Y*** and the V×V feature covariance matrix $B=Y^TY/N$:
2. Find the first eigenvalue λ_1 and corresponding normed eigenvector c_1 so that $Bc_1 = \lambda_1 c_1$;
3. Compute the principal component

$$z = \frac{Yc_1}{\sqrt{N\lambda_1}}$$

The 2nd PC is computed in the same way based on a residual covariance matrix $\dot{B}=B- \lambda_1 c_1 c_1'$, etc.

5. Principal Component Analysis: Conventional approach, 3

Covariance matrix:

1. Given a N×V data matrix ***X***, compute its centered version Y and the V×V feature covariance matrix B:
   a. Center matrix X by finding, for each feature, its mean and subtracting it from all the feature values, Y=X-m(X)
   b. Compute square matrix A=Y′∗Y and divide it by N or N-1 (do the latter if you think that the result is going to be used as an estimate of the covariance matrix of a multivariate density function, I rather divide by N): B= Y′∗Y/N.

(*v*,*w*) entry in B: $b_{vw} = \frac{1}{N}\sum_{i=1}^{N}(x_{iv} - \overline{x_v})(x_{iw} - \overline{x_w})$

5. Principal Component Analysis: Conventional approach, 4

Covariance matrix:

Given a N×V data matrix *X*, its V×V feature covariance matrix *B*=[b_{vw}]:

$$b_{vw} = \frac{1}{N}\sum_{i=1}^{N}(x_{iv} - \overline{x_v})(x_{iw} - \overline{x_w}), \qquad \overline{x_v}, \overline{x_w} \text{ - means}$$

Correlation matrix:

If features in a N×V data matrix *X*, have been normalized by their standard deviations, then the covariances b_{vw} are correlation coefficients

$$b_{vw} = \frac{1}{\sigma_v \sigma_w}\sum_{i=1}^{N}(x_{iv} - \overline{x_v})(x_{iw} - \overline{x_w})/N,$$

σ_v, σ_w **- standard deviations**

5. Principal Component Analysis: Conventional approach, 5 Example

1. Given X and its centered version Y, compute the covariance matrix **B= Y′∗Y/N.**

$$X=\begin{bmatrix} 41 & 66 & 90 \\ 57 & 56 & 60 \\ 61 & 72 & 79 \\ 69 & 73 & 72 \\ 63 & 52 & 88 \\ 62 & 83 & 80 \end{bmatrix} \quad Y=\begin{bmatrix} -17.83 & -1.00 & 11.83 \\ -1.83 & -11.00 & -18.17 \\ 2.17 & 5.00 & 0.83 \\ 10.17 & 6.00 & -6.17 \\ 4.17 & -15.00 & 9.83 \\ 3.17 & 16.00 & 1.83 \end{bmatrix} \quad B=\begin{bmatrix} 76.14 & 16.33 & -31.97 \\ 16.33 & 110.67 & 6.17 \\ -31.97 & 6.17 & 101.47 \end{bmatrix}$$

2. Compute first eigenvalue and eigenvector of **B= Y′∗Y/N:**

>>[C, La]=eig(B);% eigenvalues in the descending order

>>l1=La(3,3) % λ1=124.85

>>c1=C(:,3);% $c1=\begin{bmatrix} -0.59 \\ -0.37 \\ 0.71 \end{bmatrix}$

5. Principal Component Analysis: Conventional approach, 6 Example

Y=

$$\begin{bmatrix} -17.83 & -1.00 & 11.83 \\ -1.83 & -11.00 & -18.17 \\ 2.17 & 5.00 & 0.83 \\ 10.17 & 6.00 & -6.17 \\ 4.17 & -15.00 & 9.83 \\ 3.17 & 16.00 & 1.83 \end{bmatrix}$$

λ1=124.85

$$\mathbf{c1}=\begin{bmatrix} \mathbf{-0.59} \\ \mathbf{-0.37} \\ \mathbf{0.71} \end{bmatrix}$$

3. Given **centered data Y, eigenvalue λ1 and eigenvector c1,** compute the Principal component scoring vector

$$z = \frac{Yc_1}{\sqrt{6*\lambda_1}}$$

$$Z=\begin{bmatrix} 0.71 \\ -0.28 \\ -0.09 \\ -0.46 \\ 0.37 \\ -0.24 \end{bmatrix}$$

>>z=Y*c1/sqrt(6*λ1);

5. Principal Component Analysis: Conventional approach, 7 Example

Y=

$$\begin{bmatrix} -17.83 & -1.00 & 11.83 \\ -1.83 & -11.00 & -18.17 \\ 2.17 & 5.00 & 0.83 \\ 10.17 & 6.00 & -6.17 \\ 4.17 & -15.00 & 9.83 \\ 3.17 & 16.00 & 1.83 \end{bmatrix}$$

Solution:

$\lambda 1 = 124.85$

$$c1 = \begin{bmatrix} -0.59 \\ -0.37 \\ 0.71 \end{bmatrix}$$

$$Z = \begin{bmatrix} 0.71 \\ -0.28 \\ -0.09 \\ -0.46 \\ 0.37 \\ -0.24 \end{bmatrix}$$

Compare to the solution found using the model-based approach:

$\mu =$ 27.37

$$C = \begin{matrix} 0.5933 \\ 0.3734 \\ -0.7131 \end{matrix}$$

$$Z = \begin{bmatrix} -0.7086 \\ 0.2836 \\ 0.0935 \\ 0.4629 \\ -0.3705 \\ 0.2391 \end{bmatrix}$$

Well: (c1,z) coincide with (c,z) up to the rounding error and multiplying by -1; $27.37 = \sqrt{6 * 124.85}$. **The same!!! Why?**

5. Principal Component A: Conventional approach, 8 Relation to that SVD based

Why leads the conventional PCA approach to the same scoring and loading vectors as the model-based PCA?

The former operates over the covariance matrix never used by the latter.

Explanation:

covariance matrix coincides, up to a constant factor, with matrix A=Y′∗Y, provided that Y is centered.

Matrix A is in the core of Singular triplets.

Working with eigenvectors of A is equivalent to working with singular vectors of Y.

5. Difference between Model-Based (MB) and Conventional approaches

i. MB PCA models data; Conventional is heuristic
ii. MB PCA derives PCA scoring is a weighted sum of the features; Conventional PCA presumes that
iii. MB PCA applies to any data-preprocessing option; Conventional PCA needs features centered
iv. MB PCA gives contributions to the data scatter; Conventional PCA does not
v. MB PCA approximates data by a low rank space in which a further search for a "base of simple structure" is possible; Conventional PCA can use only bases of eigenvectors

5. Table of Differences between Model-Based (MB) and Conventional approaches

Feature of PCA	Model Based	Conventional
Relation to data	model	heuristic
Weighted sum of features	derived	presumed
Data-preprocessing	any	centered data
Contributions to data scatter	yes	no
Search for a base of “simple structure”	possible	only eigen-vectors

2019 DA Lecture 6. Principal Component Analysis: Method and Applications

- PCA and SVD: Data Approximation
- Conventional approach to PCA. Finding maximum positive eigenvalue. Comparison to that model based
- Applications
  - Ranking
  - Visualization on 2D plane
  - Latent Semantic Indexing
  - Google ranking in networks
- Homework 5

5. Principal Component Analysis: Applications

- **Application of PCA to finding a hidden factor**
- **Application of PCA to data visualization**
- **Application of PCA to Dimension reduction (at the case of so-called Latent semantic indexing [LSI] in text analysis)**
- **Google ranking in networks**

5. Principal Component Analysis: Applications

- **Application of PCA to finding a hidden factor**
- Application of PCA to data visualization
- Application of PCA to Dimension reduction (at the case of so-called Latent semantic indexing [LSI] in text analysis)
- Google ranking in networks

5. Principal Component Analysis: Application 1: Hidden factor, 1

Typical problems of scoring according to unmeasurable features:

- **Measuring student talent**
- **Measuring manager performances**
- **Evaluation of regional wellbeing**
- **Ranking countries over life satisfaction**

Week 5. Principal Component Analysis: Application 1: Hidden factor, 2

Ranking individuals over life satisfaction:

Principal component of five indicators:

- 1. In most ways my life is close to my ideal. .84
- 2. The conditions of my life are excellent. .77
- **3. I am satisfied with my life .83**
- 4. So far I have gotten the important things I want in life. .72
- 5. If I could live my life over, I would change almost nothing. .61

(an earlier attempt, **ED DIENER, ROBERT A. EMMONS, RANDY J. LAR.SEM, and SHARON GRIFFIN, 1985**)

Week 5. Principal Component Analysis : Application 1: Hidden factor, 3

#	SEn	OOP	CI	Average
1	41	66	90	65.7
2	57	56	60	57.7
3	61	72	79	70.7
4	69	73	72	71.3
5	63	52	88	67.7
6	62	83	80	75.0

F. Galton: Set of students, marks over Software engineering, Object-Oriented Programming, Computational Intelligence

How can table help in deriving hidden talent scores?

What is wrong with the average (see the table)?

Will apply PCA to the data to find the First singular triplet

5. Principal Component Analysis : Application 1: Hidden factor, 4

#	SE	OOP	CI	Mean
1	41	66	90	65.7
2	57	56	60	57.7
3	61	72	79	70.7
4	69	73	72	71.3
5	63	52	88	67.7
6	62	83	80	75.0

How can table help in deriving hidden talent scores?

What is wrong with the average (see the table)?

Apply PCA to the data as is, **no centering,** to find the First singular triplet

$$X=\begin{bmatrix} 41 & 66 & 90 \\ 57 & 56 & 60 \\ 61 & 72 & 79 \\ 69 & 73 & 72 \\ 63 & 52 & 88 \\ 62 & 83 & 80 \end{bmatrix}$$

```
>> [Z,Mu,C]=svd(X);
>> z=Z(:,1)  % First singular 6D scoring vector
>> Mu=Mu(1,1)  % maximum singular value
>> ds=sum(sum(X.*X)) % data scatter
>> z = -z; c = -c;
```

5. Principal Component Analysis : Application 1: Hidden factor, 5

#	SE	OOP	CI	Mean
1	41	66	90	65.7
2	57	56	60	57.7
3	61	72	79	70.7
4	69	73	72	71.3
5	63	52	88	67.7
6	62	83	80	75.0

How can table help in deriving hidden talent scores?

What is wrong with the average (see the table)?

Apply PCA to the data to find the First singular triplet

$$\mathbf{X}=\begin{bmatrix} 41 & 66 & 90 \\ 57 & 56 & 60 \\ 61 & 72 & 79 \\ 69 & 73 & 72 \\ 63 & 52 & 88 \\ 62 & 83 & 80 \end{bmatrix}$$

$$\mathbf{c}'=[0.49 \quad 0.57 \quad 0.66]$$

$$\mathbf{Z}=\begin{matrix} 0.40 \\ 0.34 \\ 0.42 \\ 0.42 \\ 0.41 \\ 0.45 \end{matrix}$$

$$\mu=291.39$$

$$X \approx \mu\, z\, c'$$

How can one rescale z to convert it to 0 – 100 scale?

5. Principal Component Analysis : Application 1: Hidden factor, 6

#	SE	OOP	CI	Mean
1	41	66	90	65.7
2	57	56	60	57.7
3	61	72	79	70.7
4	69	73	72	71.3
5	63	52	88	67.7
6	62	83	80	75.0

What is wrong with the average?

Apply PCA to the data to find the First singular triplet

$$Z = \begin{pmatrix} 0.40 \\ 0.34 \\ 0.42 \\ 0.42 \\ 0.41 \\ 0.45 \end{pmatrix} \quad c = \begin{pmatrix} 0.49 \\ 0.57 \\ 0.66 \end{pmatrix} \quad \mu = 291.39$$

$$X \approx \mu z c'$$

$$z = Xc/\mu \ (*)$$

How can one rescale z to convert it to 0 – 100 scale?

Use (*)-like equation

$$Z=(.49*Se+.57*OOP+0.66*CI)*\alpha$$

Find α from OUR WISH that Z=100 at all subject marks being 100:

$$100=(.49*100+.57*100+0.66*100)*\alpha$$

$$\alpha=100/[(.49+.57+.66)100]=0.5813$$

5. Principal Component Analysis :

Application 1: Hidden factor, 7

#	SE	OOP	CI	Mean	PC z
1	41	66	90	65.7	68.0
2	57	56	60	57.7	57.8
3	61	72	79	70.7	71.5
4	69	73	72	71.3	71.5
5	63	52	88	67.7	68.9
6	62	83	80	75.0	75.8

What is wrong with the average?

Apply PCA to the data to find the First singular triplet

Rescale z to 0–100 scale

PCA: $Z=(.49*Se+.57*OOP+0.66*CI)*\alpha$

Found $\alpha=0.5813$ from OUR WISH that Z=100 at all subject marks being 100. The FINAL equation:

$$Z=0.29*Se+0.33*OOP+0.38*CI$$

$$Mean=0.33*Se+0.33*OOP+0.33*CI$$

5. Principal Component Analysis : Application 1: Hidden factor, 8

#	SE	OOP	CI	Mean	PC z
1	41	66	90	65.7	68.0
2	57	56	60	57.7	57.8
3	61	72	79	70.7	71.5
4	69	73	72	71.3	71.5
5	63	52	88	67.7	68.9
6	62	83	80	75.0	75.8

What is wrong with the mean?

Mean=0.33*Se+0.33*OOP+0.33*CI,

Almost as good as the PCA factor; yet no contribution to the data scatter!

Hidden factor derivation:

1. Select base features and convert them into the same scale, say, 0 to 100, to form matrix X
2. Find the first singular triplet (μ, z, c)
3. Use equation $z=Xc\alpha$ to rescale z, typically to 0-100 scale, as explained

$$Z=0.29*Se+0.33*OOP+0.38*CI$$

4. Apply the equation to the used and new entities

5. Principal Component Analysis :

Application 1: Hidden factor, 9

#	SE	OOP	CI	Mean	PC z
1	41	66	90	65.7	68.0
2	57	56	60	57.7	57.8
3	61	72	79	70.7	71.5
4	69	73	72	71.3	71.5
5	63	52	88	67.7	68.9
6	62	83	80	75.0	75.8

Interpretation:

Z=0.29*Se+0.33*OOP+0.38*CI

Weight of Se is lowered while Weight of CI is increased in comparison to the mean

All 4 of Iris dataset features relate to the specimen's size.

Find hidden factor size for Iris.

1. Normalize features so that maximum gets a value of 100:

```
>>ma=max(iris);
>>X=iris*100./repmat(ma, 150, 1)
```

2. Find first singular triplet with positive loadings:

```
>>[z,mu,c]=svd(X);
>>c1=-c(:,1)
```

3. Determine factor α

```
>> alpha=1/sum(c);
```

5. Principal Component Analysis :
Application 1: Hidden factor, 10

Interpretation:

All 4 of Iris dataset features relate to the specimen's size.

Find hidden factor size for Iris.

1. **Normalize features so that maximum gets a value of 100:**
2. **Find first singular triplet with positive loadings:**

$$c1' = [0.58 \quad 0.53 \quad 0.46 \quad 0.42]$$

3. **Determine factor α**

>> alpha=1/sum(c); α=0.504

4. **Compute the PCA hidden factor score vector**

>>z= 0.29*SL+ 0.27*SW+ 0.23*PL+ 0.21PW

5. **Determine its contribution to the data scatter**

>>100*mu(1,1)^2/sum(sum(x.*x));% 94.05%

5. Principal Component Analysis : Application 1: Hidden factor, 11

Interpretation:

All 4 of Iris dataset features relate to the specimen's size.

Loadings: this is what is interpreted first: c1′= [0.58 0.53 0.46 0.42]
All positive, even in spite that SL and SW correlate negatively.
Sepal sizes play here more important role that the petal sizes, by about (.58+.53)/(.46+.42)=1.26, 26%.

Also, the contribution is quite good: 94.05%

Also, compare the 50:50:50 distribution of Iris according to the first PC found, with the taxa

	T1	T2	T3
z1	47	3	0
z2	3	40	7
z3	0	7	43,

rather compatible!

2019 DA Lecture 6. Principal Component Analysis: Method and Applications

- PCA and SVD: Data Approximation
- Conventional approach to PCA. Finding maximum positive eigenvalue. Comparison to that model based
- Applications
  - Ranking
  - **Visualization on 2D plane**
  - Latent Semantic Indexing
  - Google ranking in networks
- Homework 5

«Market Towns UK West Country» (45x6 table), 1

		Pop.Res	PSchool	Doctor	Bank	Petrol	PostOf
1	'Ashburton'	3660	1	0	2	2	1
2	'Bere Alston'	2362	1	0	1	0	1
3	'Bodmin'	12553	5	2	6	5	2
4	'Brixham'	15865	7	3	5	3	5
5	'Buckfastleigh’	2786	2	1	1	2	1
		. . . . .					
43	'Torpoint’	6929	2	1	7	1	4
44	'Totnes’	18966	9	3	19	5	7
45	'Truro'	5291	1	1	5	1	1

«Market Towns UK West Country» (45x6 table), 2

Three popular data standardization methods:

	Pop.Res	PSchool	Doctor	Bank	Petrol	PostOf
Mean	7351.356	3.0222	1.3778	4.3111	2.0444	2.6222
Std	6193.246	2.7344	1.3019	4.3840	1.6370	2.1137
Max	23801	13	4	19	7	9
Min	2040	1	0	0	0	1

$y_{iv} = \frac{x_{iv} - meanv}{std_v}$ - z-scoring

$\boldsymbol{y_{iv} = \frac{x_{iv} - meanv}{max_v - min_v}}$ **- range normalization**

$y_{iv} = \frac{x_{iv} - minv}{max_v - minv}$ - ranking normalization

«Market Towns UK West Country» (45x6 table), 3 z-scoring

	Pop.Res	PSchool	Doctor	Bank	Petrol	PostOf
'Ashburton'	-0.7643	-0.8021	-1.1577	-0.6438	-0.2031	-0.7561
'Bere Alston'	-0.9671	-0.8021	-1.1577	-0.8155	-1.2860	-0.7561
'Bodmin'	0.6253	0.4813	0.5262	0.0429	1.4214	-0.3240
'Brixham'	1.1428	1.1230	1.3682	-0.1288	0.3384	0.9721
'Buckfastleigh’	-0.9009	-0.4813	-0.3157	-0.8155	-0.2031	-0.7561
'Torpoint'	-0.2535	-0.4813	-0.3157	0.2146	-0.7445	0.5401
'Totnes'	1.6273	1.7646	1.3682	2.2747	1.4214	1.8362
'Truro'	-0.5095	-0.8021	-0.3157	-0.1288	-0.7445	-0.7561

«Market Towns UK West Country» (45x6 table), 4 range normalization

	Pop.Res	PSchool	Doctor	Bank	Petrol	PostOf
'Ashburton'	-0.2946	-0.3125	-0.4583	-0.2083	-0.0750	-0.2917
'Bere Alston'	-0.3728	-0.3125	-0.4583	-0.2639	-0.4750	-0.2917
'Bodmin'	0.2410	0.1875	0.2083	0.0139	0.5250	-0.1250
'Brixham'	0.4405	0.4375	0.5417	-0.0417	0.1250	0.3750
'Buckfastleigh’	-0.3472	-0.1875	-0.1250	-0.2639	-0.0750	-0.2917
'Torpoint'	-0.0977	-0.1875	-0.1250	0.0694	-0.2750	0.2083
'Totnes'	0.6272	0.6875	0.5417	0.7361	0.5250	0.7083
'Truro'	-0.1964	-0.3125	-0.1250	-0.0417	-0.2750	-0.2917

«Market Towns UK West Country» (45x6 table), 5

Ranking normalization

	Pop.R	PScho	Doct	Bank	Petr	PostOf
'Ashburton'	7.8174	0	0	5.56	40.00	0
'Bere Alston'	0	0	0	0	0	0
'Bodmin'	61.37	50.00	66.67	27.78	100.00	16.67
'Brixham'	81.32	75.00	100.00	22.22	60.00	66.67
'Buckfastleigh’	2.55	12.50	33.33	0	40.00	0
'Torpoint'	27.51	12.50	33.33	33.33	20.00	50.000
'Totnes'	100.00	100.00	100.00	100.00	100.00	100.00
'Truro'	17.64	0	33.33	22.22	20.00	0

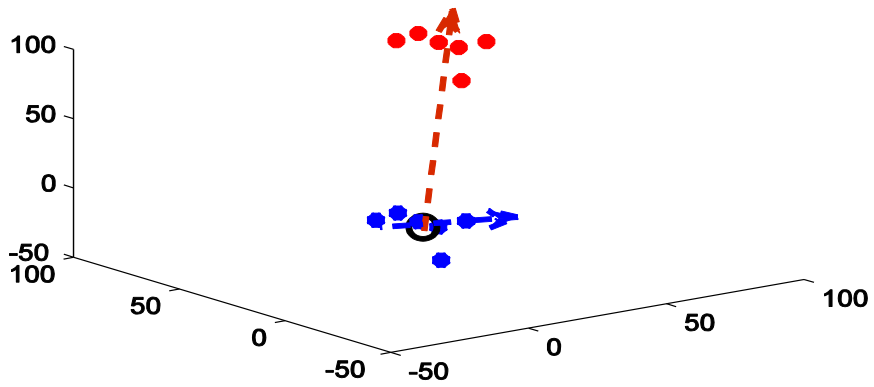
5. Principal Component Analysis: Application 2: Data Visualization,1

How can one visualize a dataset X on a plane so that every entity is mapped to a 2D point according to the data structure?

Just take the two first singular triplets: this is the best 2D approximation of X possible.

5. Principal Component Analysis: Application 2: Data Visualization,2

Caveat, 1. A figure showing the first PCA component before and after the data X is preprocessed into Y by centering, that is, subtracting the column means from the columns.



Red dots – raw data X

Blue – data Y centered

Circle – space origin

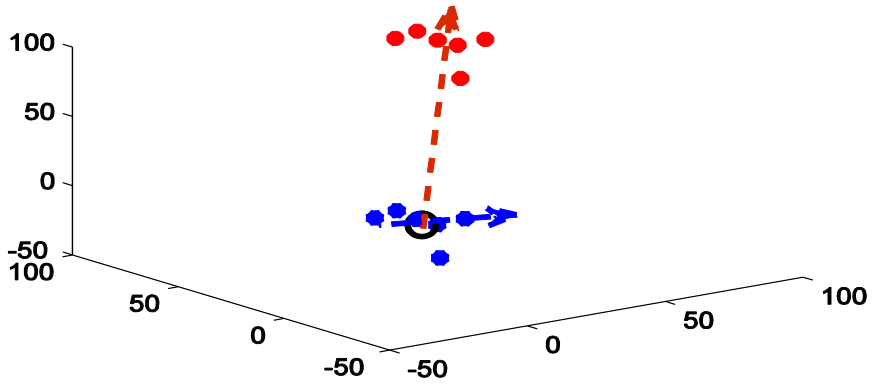
Red arrow – PC for X

Blue arrow – PC for Y

Because all PCs must go through the origin, the data structure is better seen if it is looked at against a backdrop of the means, that is, centered.

5. Principal Component Analysis: Application 2: Data Visualization,2

Caveat, 2. All PCs must go through the origin,



Red dots – raw data X
Blue – data Y centered
Circle – space origin
Red arrow – PC for X
Blue arrow – PC for Y

Appl1. Hidden factor: No Centering is best to integrate the features

Appl2. Visualization: Centering is a Must; data structure is better seen from a backdrop of the center.

5. Principal Component Analysis: Application 2: Data Visualization, 3

Procedure.

1. Center data into matrix Y by subtracting the column means from the columns; normalize if needed.
2. Compute two first singular triplets:

```
>> [Z,Mu,C]=svd(Y);
>> z1=z(:,1)*sqrt(Mu(1,1));   %Quiz: Why is that?
>> z2=z(:,2)*sqrt(Mu(2,2));% sqrt(Mu) to fit in the PCA model
```

3. Determine the proportion of the variance taken into account:

```
>> p=100*(Mu(1,1)^2+Mu(2,2)^2)/sum(sum(Y.*Y));
```

4. Visualize the data

```
>> plot(z1, z2, 'k.');
```

5. Interpret the axes by looking at the loadings and their signs.

«Market Towns UK West Country»: PCA

- Centered, Range normalized
- Loadings

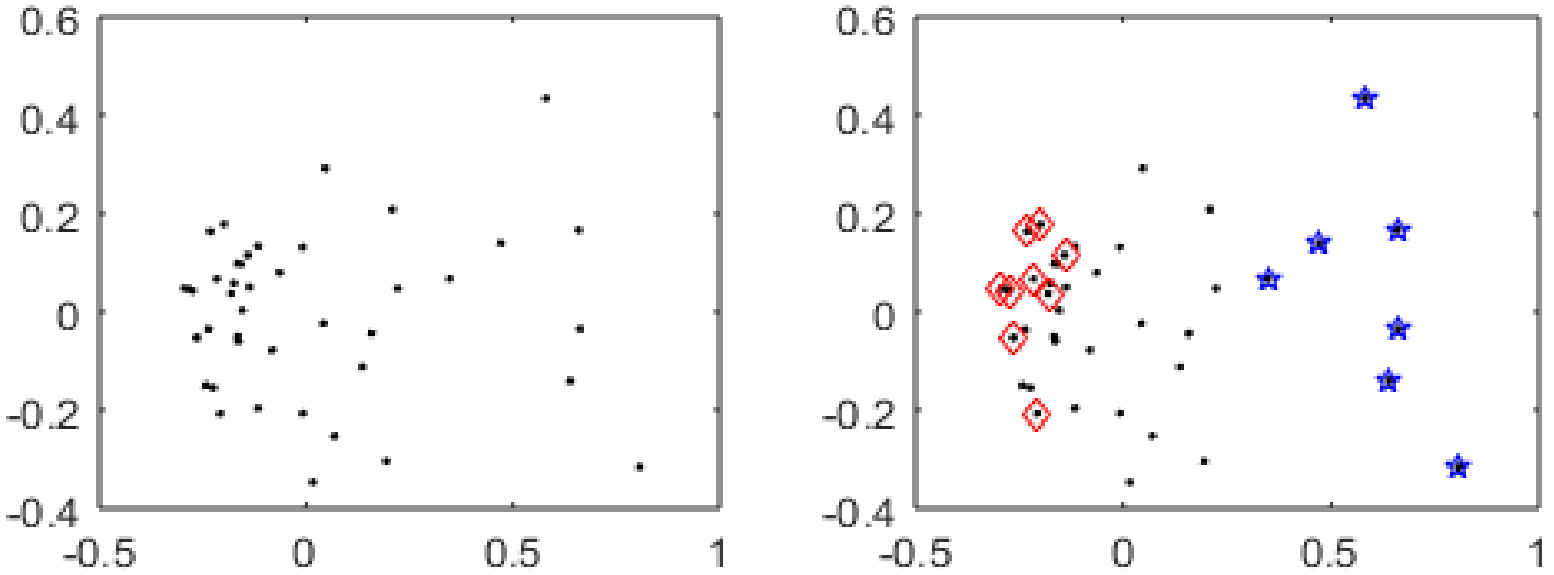
	Pop.Res	**PSchool**	**Doctor**	**Bank**	**Petrol**	**PostOf**
c_1'	**0.4722**	0.3562	**0.5109**	0.3466	0.3192	**0.4087**
c_2'	-0.0445	-0.3163	**0.4946**	-0.0463	-0.7387	**0.3249**

Contribution to the Variance: **83.6%**

Interpretation: c_1: Level of town development

c_2 : Doctors versus Cars (“Resort area” ?)

«Market Towns UK WC» visualized



Matlab Code: subplot(1,2,1);plot(z1,z2,'k.');
tt=find(ti(:,1)>15000); tts=find(ti(:,1)<2500);
subplot(1,2,2);plot(z1,z2,'k.', z1(tt),z2(tt), 'bp', z1(tts), z2(tts), 'rd');
On the right: towns with 15000 or more residents, in blue, less than 2500 residents, in red.
QUESTION: WHY are reds on the left and blues on the right?

5. Principal Component Analysis: Application 2: Data Visualization, 4

Iris dataset example.

1. Center data into matrix Y by subtracting the column means from the columns.

```
>>Y=iris-repmat(mean(iris), 150, 1); % No normalization
```

1. Compute two first singular triplets:

```
>> [Z,Mu,C]=svd(Y);
>> z1=z(:,1)*sqrt(Mu(1,1));
>> z2=z(:,2)*sqrt(Mu(2,2));% sqrt(Mu) to fit in the PCA model
```

3. Determine the proportion of the variance taken into account:

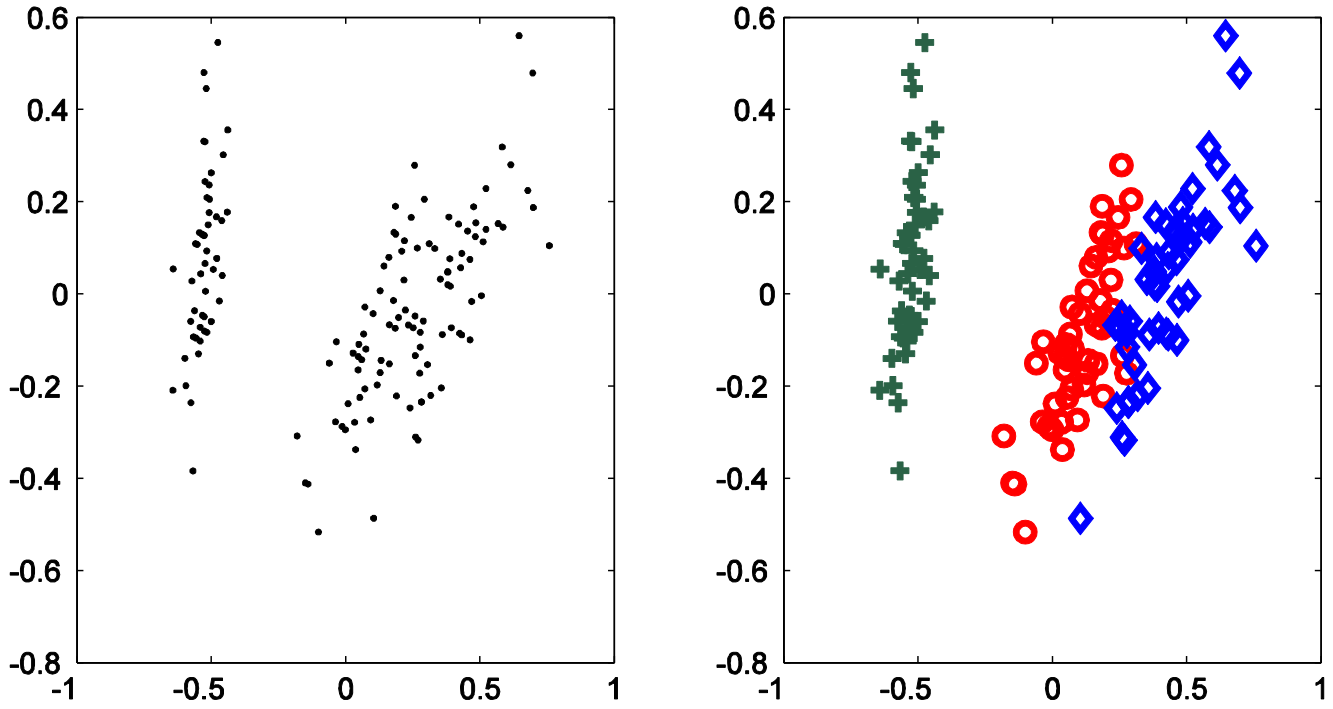
```
>> p=100*(Mu(1,1)^2+Mu(2,2)^2)/sum(sum(Y.*Y)); % p=97.8%, good!
```

5. Principal Component Analysis: Application 2: Data Visualization, 5

Iris dataset example.

4. Visualize the Iris data in two subplots, as a whole and taxon-wise.

```
>>subplot(1,2,1); plot(z1, z2, 'k.');
>>subplot(1,2,2);
>>plot(z1(1:50),z2(1:50),'g+',z1(51:100),z2(51:100),'ro',z1(101:150),z2(101:150),'bd');
```

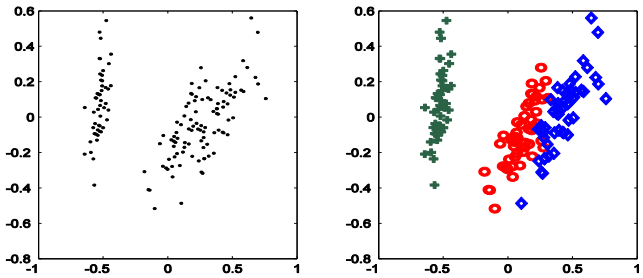


5. Principal Component Analysis:

Application 2: Data Visualization, 6

Iris dataset example.

You must be creative at interpretation



5. Interpret the axes by looking at the loadings and their signs:

c1′= [0.3614 -0.0845 0.8567 0.3583]

c2′ = [0.6566 0.7302 -0.1734 -0.0755]

First PC: all components should be positive to express the size of iris. Unfortunately, c_2 is small negative (probably because the Sepal sizes correlate negative), and Petal Length component c_3 is much higher than the others. Still, arguably "specimen's size". Taxa ordered along.

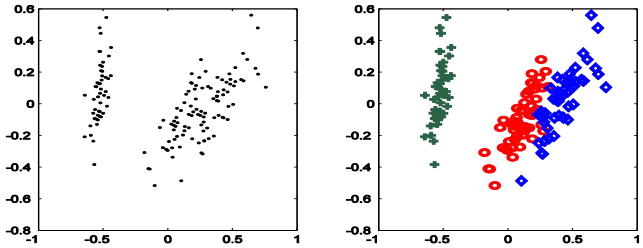
Second PC: Sepal sizes high +, petal sizes low –. Arguably, "Sepal size short of petal size". Has nothing to do with taxa.

5. Principal Component Analysis:

Application 2: Data Visualization, 7

Iris dataset example.

You must be creative at interpretation



5. Interpret the axes by looking at the loadings and their signs:

First PC z1: arguably "specimen's size". Taxa ordered along the axis on the drawing.

Define three classes in the ascending order z1:

>>[vz,iz]=sort(z1); pc{1}=iz(1:50); pc{2}=iz(51:100); pc{3}=iz(101:150);

Contingency table:

	T1	T2	T3	
pc{1}	47	3	0	50
pc{2}	3	40	7	50
pc{3}	0	7	43	50
	50	50	50	150

Indeed PC1, the "size" works along taxa, all errors among neighbors

2019 DA Lecture 6. Principal Component Analysis: Method and Applications

- PCA and SVD: Data Approximation
- Conventional approach to PCA. Finding maximum positive eigenvalue. Comparison to that model based
- Applications
  - Ranking
  - Visualization on 2D plane
  - **Latent Semantic Indexing**
  - Google ranking in networks
- Homework 5

Week 5. Principal Component Analysis:
Application 3: Dimension reduction with Latent Semantic Indexing, 1

Information retrieval :

Given a set of keywords, retrieve all the relevant documents

Example: Keywords "fuel & relief"

Merriam-Webster

fuel

a material used to produce heat by burning
a source of sustenance or incentive
to give support or strength

relief

removal of something oppressive
a mode of sculpture
elevations of a land surface

Ambiguity (Polysemy): What meanings to pick?

Week 5. Principal Component Analysis: Application 3: Latent Semantic Indexing, 2

Information retrieval :

Given a set of keywords, retrieve all the relevant documents

Example: Keywords "fuel & relief"

Merriam-Webster

fuel

a material used to produce heat by burning
a source of sustenance or incentive
to give support or strength

relief

removal of something oppressive
a mode of sculpture
elevations of a land surface

LSI principle: **"words used in the same contexts tend to have similar meanings"**

is well caught with the SVD of document-to-keyword data

5. Principal Component Analysis: Application 3: Latent Semantic Indexing, 3

A database of 12×10 newspaper article-to-keyword items. Article labels: F for feminism, E for entertainment, and H for household.

Problem: **Properly retrieve articles using keywords**

Article	drink	equal	fuel	play	popular	price	relief	talent	tax	woman
F1	1	2	0	1	2	0	0	0	0	2
F2	0	0	0	1	0	1	0	2	0	2
F3	0	2	0	0	0	0	0	1	0	2
F4	2	1	0	0	0	2	0	2	0	1
E1	2	0	1	2	2	0	0	1	0	0
E2	0	1	0	3	2	1	2	0	0	0
E3	1	0	2	0	1	1	0	3	1	1
E4	0	1	0	1	1	0	1	1	0	0
H1	0	0	2	0	1	2	0	0	2	0
H2	1	0	2	2	0	2	2	0	0	0
H3	0	0	1	1	2	1	1	0	2	0
H4	0	0	1	0	0	2	2	0	2	0

5. Principal Component Analysis: Application 3: Latent Semantic Indexing, 4

Problem: Properly retrieve articles using keywords

Article	drink	equal	fuel	play	popular	price	relief	talent	tax	woman
F1	1	2	0	1	2	0	0	0	0	2
F2	0	0	0	1	0	1	0	2	0	2
F3	0	2	0	0	0	0	0	1	0	2
F4	2	1	0	0	0	2	0	2	0	1
E1	2	0	1	2	2	0	0	1	0	0
E2	0	1	0	3	2	1	2	0	0	0
E3	1	0	2	0	1	1	0	3	1	1
E4	0	1	0	1	1	0	1	1	0	0
H1	0	0	2	0	1	2	0	0	2	0
H2	1	0	2	2	0	2	2	0	0	0
H3	0	0	1	1	2	1	1	0	2	0
H4	0	0	1	0	0	2	2	0	2	0

QUERY

Fuel
E1
E3
H1-H4
Precision=4/6=67%
Recall=4/4=100%

Fuel & Price
E3
H1-H4
Precision=4/5=80%
Recall=4/4=100%

No way to get 100% Precision by using crisp queries

5. Principal Component Analysis: Application 3: Latent Semantic Indexing, 5

Problem: Properly retrieve articles using keywords

Article	drink	equal	fuel	play	popular	price	relief	talent	tax	woman
F1	1	2	0	1	2	0	0	0	0	2
F2	0	0	0	1	0	1	0	2	0	2
F3	0	2	0	0	0	0	0	1	0	2
F4	2	1	0	0	0	2	0	2	0	1
E1	2	0	1	2	2	0	0	1	0	0
E2	0	1	0	3	2	1	2	0	0	0
E3	1	0	2	0	1	1	0	3	1	1
E4	0	1	0	1	1	0	1	1	0	0
H1	0	0	2	0	1	2	0	0	2	0
H2	1	0	2	2	0	2	2	0	0	0
H3	0	0	1	1	2	1	1	0	2	0
H4	0	0	1	0	0	2	2	0	2	0
Query	0	0	1	0	0	1	1	0	1	0

QUERY

Fuel Price Relief Tax

LSI:

Embed into a subspace of a few first singular triplets so that location of words relates to the similarity of contexts

5. Principal Component Analysis: Application 3: Latent Semantic Indexing, 6

TF-IDF normalization

Article	drink	equal	fuel	play	popular	price	relief	talent	tax	woman
F1	1	2	0	1	2	0	0	0	0	2
F2	0	0	0	1	0	1	0	2	0	2
F3	0	2	0	0	0	0	0	1	0	2
F4	2	1	0	0	0	2	0	2	0	1
E1	2	0	1	2	2	0	0	1	0	0
E2	0	1	0	3	2	1	2	0	0	0
E3	1	0	2	0	1	1	0	3	1	1
E4	0	1	0	1	1	0	1	1	0	0
H1	0	0	2	0	1	2	0	0	2	0
H2	1	0	2	2	0	2	2	0	0	0
H3	0	0	1	1	2	1	1	0	2	0
H4	0	0	1	0	0	2	2	0	2	0
DF	5	5	6	7	7	8	5	6	4	5
IDF	0.88	0.88	0.69	0.54	0.54	0.41	0.88	0.69	1.10	0.88
Query	0	0	0.69	0	0	0.41	0.88	0	1.10	0

TF = Term frequency, data entry

DF = Document frequency

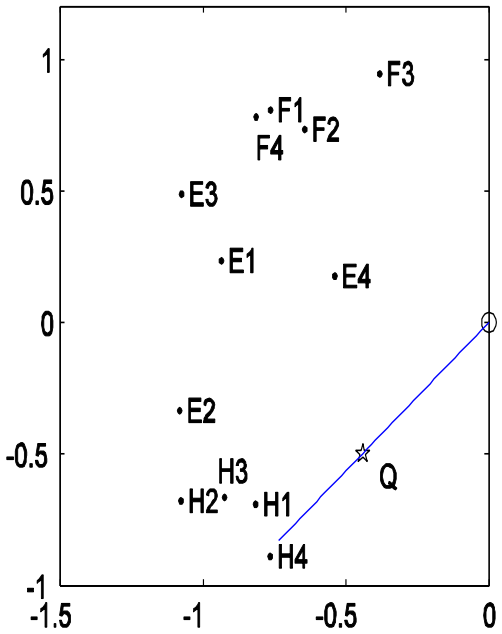
N= Number of documents

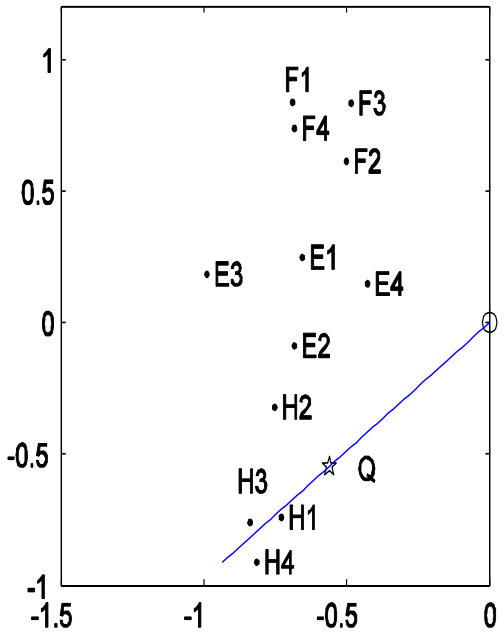
IDF = Log(N/DF)

TF-IDF(F1,equal)= 2*0.88=1.76

5. Principal Component Analysis: Application 3: Latent Semantic Indexing, 7

2D LSI representation





Specify LSI space dimension p

Pick first p singular triplets [Zp,Mp,Cp]

Visualize $Zp*Mp^{½}$

Define

$Q=Cp*Mp^{-1/2}*QUERY$

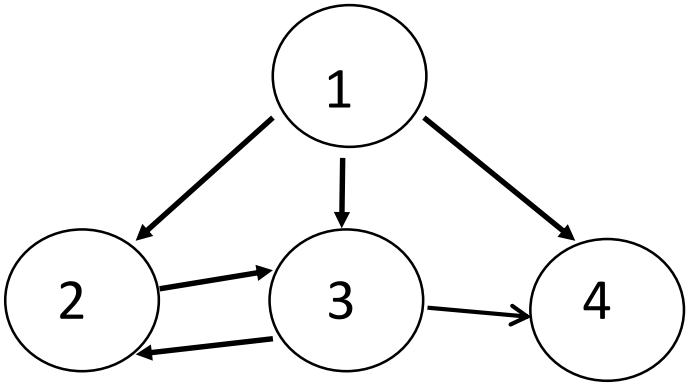
One can see: With both normalizations, Q is closer to H, whichever distance is taken, Euclidean or Cosine

2019 DA Lecture 6. Principal Component Analysis: Method and Applications

- PCA and SVD: Data Approximation
- Conventional approach to PCA. Finding maximum positive eigenvalue. Comparison to that model based
- Applications
  - Ranking
  - Visualization on 2D plane
  - Latent Semantic Indexing
  - **Google ranking in networks**
- Homework 5

Google Ranking in Networks

Network ≡ Graph, a different data type (entity-to-entity rather than entity-to-feature)



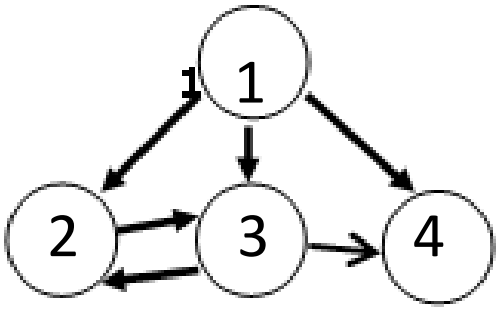
Which node is the leader? To answer, let us represent the graph by a matrix.

Google Ranking in Networks

1

2 3 4

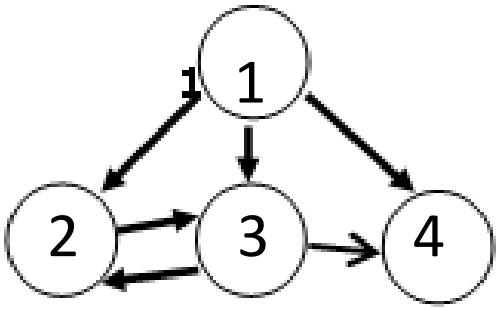
Google Ranking in Networks, 1



1	**0**	**1/3**	**1/3**	**1/3**
2	**0**	**0**	**1**	**0**
3	**0**	**1/2**	**0**	**1/2**
4	**0**	**0**	**0**	**0**
	1	**2**	**3**	**4**

= Q

Stochastic matrix of equal probabilities – well, not exactly (row 4)

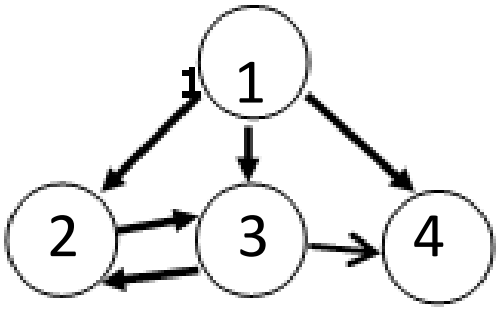


1	0	1/3	1/3	1/3
2	0	0	1	0
3	0	1/2	0	1/2
4	1/4	1/4	1/4	1/4

= Q’

1 2 3 4

Change 0 rows for all equal probabilities 1/N – Stochastic matrix indeed!



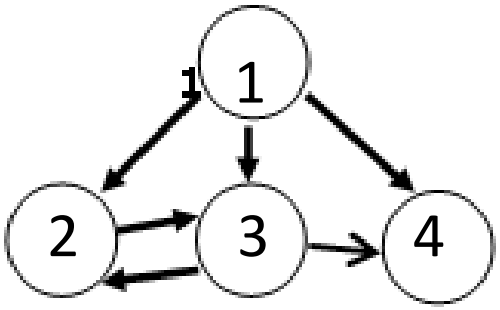
We want the transfer matrix be all positive!

1	**0**	**1/3**	**1/3**	**1/3**
2	**0**	**0**	**1**	**0**
3	**0**	**1/2**	**0**	**1/2**
4	**1/4**	**1/4**	**1/4**	**1/4**

= Q'

1 2 3 4

How come?



Teleportation stochastic N×N matrix - all 1/N:

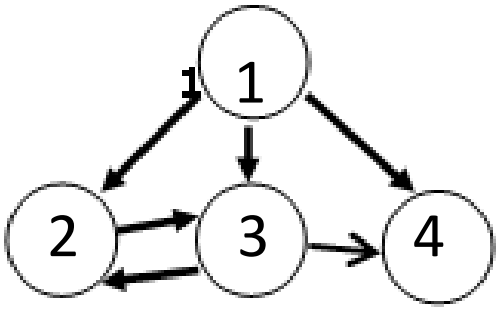
1	1/4	1/4	1/4	1/4
2	1/4	1/4	1/4	1/4
3	1/4	1/4	1/4	1/4 = E
4	1/4	1/4	1/4	1/4
	1	2	3	4

We want the transfer matrix be all positive!

1	0	1/3	1/3	1/3	
2	0	0	1	0	
3	0	1/2	0	1/2	= Q’
4	1/4	1/4	1/4	1/4	
	1	2	3	4	

MIX Q’ and E:

$P=\alpha Q'+(1-\alpha)E$



1	0	1/3	1/3	1/3
2	0	0	1	0
3	0	1/2	0	1/2
4	1/4	1/4	1/4	1/4

= Q'

1 2 3 4

Teleportation stochastic N×N matrix - all 1/N:

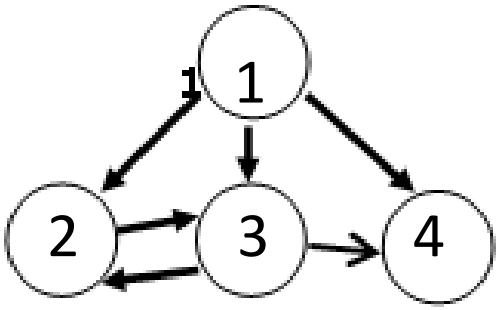
1	1/4	1/4	1/4	1/4
2	1/4	1/4	1/4	1/4
3	1/4	1/4	1/4	1/4
4	1/4	1/4	1/4	1/4

= E

1 2 3 4

At α =0.85, P=αQ'+(1-α)E=

1	.0375	.3208	.3208	.3208
2	.0375	.0375	.8875	.0375
3	.0375	.4625	.0375	.4625
4	.25	.25	.25	.25



Node *i=1,2,…, N* importance score r_i

Assume r_i is exercised at all related nodes according to *P* probabilities:

$$r_i = p_{i1} r_1 + p_{i2} r_2 + \ldots + p_{iN} r_N$$

$$r = Pr$$

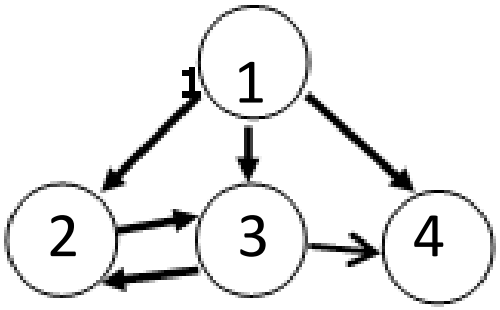
1	|	0	1/3	1/3	1/3
2	|	0	0	1	0
3	|	0	1/2	0	1/2
4	|	1/4	1/4	1/4	1/4
		1	2	3	4

= Q'

At α =0.85, $P=\alpha Q'+(1-\alpha)E$=

1	|	.0375	.3208	.3208	.3208
2	|	.0375	.0375	.8875	.0375
3	|	.0375	.4625	.0375	.4625
4	|	.25	.25	.25	.25

Google Ranking in Networks, 7



Node *i=1,2,…, N* importance score r_i

Assume r_i is exercised at all related nodes according to *P* probabilities:

$$r_i = p_{i1} r_1 + p_{i2} r_2 + \ldots + p_{iN} r_N$$

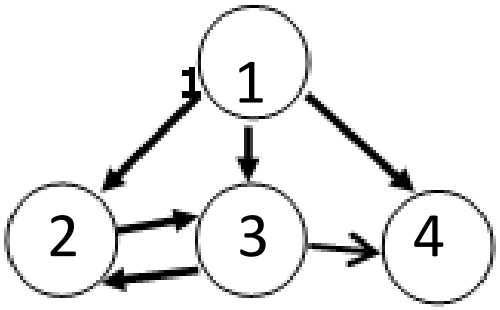
$$r = Pr$$

How can this be solved?

Perron-Frobenius theorem

For any positive square matrix, its maximum eigenvalue is positive and strictly greater than modules of the other eigenvalues, the corresponding eigenvector is positive too.

Google Ranking in Networks, 8



Node *i=1,2,…, N* importance score r_i

Assume r_i is exercised at all related nodes according to *P* probabilities:

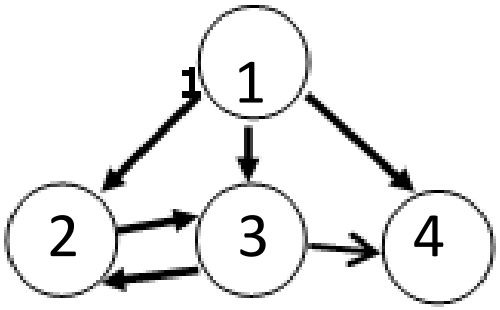
$$r_i = p_{i1} r_1 + p_{i2} r_2 + \ldots + p_{iN} r_N$$

$$r = Pr$$

For any stochastic matrix, its maximum eigenvalue is 1.

This implies convergence of Google ranking algorithm (spectral decomposition analogous to that in slide 14]

Google Ranking in Networks, 9



Node	0	1	5	10	14	15
1	0.25	0.0906	0.0961	0.0957	0.0958	0.0958
2	0.25	0.2677	0.2732	0.2742	0.2742	0.2742
3	0.25	0.3740	0.3576	0.3558	0.3559	0.3559
4	0.25	0.2677	0.2732	0.2742	0.2742	0.2742

Convergence of the node ranks from n=0 to n=15.

Unexpectedly, 1 is not the leader at all! Moreover, It is less relevant!

5. Principal Component Analysis: Method, Model, Application

Summary

- Theoretic introduction: Summarization versus Correlation: **Summarization is similar to correlation if all features are considered target; the data standardization issue is important yet to be explored.**
- Matrix operations: **Be cautious, matrices are not numbers!**
- Matrix spectrum, singular value decomposition, approximation: **Considering a square matrix as a mapping, eigenvectors represent axes that are mapped onto themselves. For symmetric matrices these axes are mutually orthogonal. A rectangular matrix can be treated similarly if both direct and inverse mapping are considered. Luckily this amounts to multiplication of matrices and using one more, semi positive definite, property.**

5. Principal Component Analysis: Method, Model, Application

Summary

- Hidden factor model. Its solution. Principal components, loadings, contributions. **This all comes from the SVD theory because of somewhat over simplistic character of the model, just a product of row-related item and a column-related item.**
- Conventional PCA criterion and method. Relation between the model-based and conventional approaches. Covariance and correlation matrix. **Amazingly, the covariance and correlation matrices are closely related to a matrix of direct-inverse mapping, which provides for getting the same results. Yet the model-based criterion has a number of good properties.**

5. Principal Component Analysis: Method, Model, Application

Summary

- **Finding a hidden factor with PCA:** No feature centering, though a normalization to 0-100 scale may be advisable; both contribution to data scatter and interpretation can be of interest.

- **Data visualization with PCA:** Feature centering is a must to present the data structure against the center's backdrop; both contribution to data scatter and interpretation can be of interest.

5. Principal Component Analysis: Method, Model, Application

Summary

- **Latent Semantic Indexing:** Effectively addressing the word polysemy problem for Information Retrieval with PCA by shifting from crisp interpretations to soft Euclidean or Cosine distances; simultaneously reduces dimension from thousands to hundreds.
- **Google ranking:** Base of a most effective browser.
- Many more applications exist, first of all, in signal and image compression.

2019 DA Lecture 6. Principal Component Analysis: Method and Applications

- PCA and SVD: Data Approximation
- Conventional approach to PCA. Finding maximum positive eigenvalue. Comparison to that model based
- Applications
  - Ranking
  - Visualization on 2D plane
  - Latent Semantic Indexing
  - Google ranking in networks
- **Homework 5**

Homework 5: PCA/SVD

- In your data set, select a subset of 3-6 features related to the same aspect and explain your choice
- Standardize the selected subset; compute its data scatter and SVD; determine contributions of all the principal components to the data scatter, naturally and per cent
- Visualize the data using two first principal components at the standardization with two versions of normalization: (a) range normalization and (b) z-scoring. At these visualizations, use a distinct shape/color for points representing a pre-specified by you group of objects. Also, apply the conventional PCA for finding two first principal components and visualization; compare to the results at z-scoring. Comment on which of the normalizations is better and why.
- Compute and interpret a hidden ranking factor behind the selected features. The factor should be expressed in a 0-100 rank scale (as well as the features – ranking normalization).