Data Mining and Applied Linear Algebra

Moody T. Chu

North Carolina State University

MA325 @ North Carolina State University

Take Home Message

- Finding how a given number is made up by prime numbers — *Arithmetic factorization*.

$$854905477858447264886 4899 = (23, 631, 3923, 7901)^{(4,3,2,1)}.$$

- Finding how an observed data is composed of simple factors — *Matrix factorization*.

$$Y = AF.$$

Course Plan of This Module (9 lessons)

1. Overview (1 lesson)
2. Basic Model (3 lessons)
   - Homework
3. Singular Value Decomposition (2 lessons)
4. Computational Issues (1 lesson)
   - Homework
5. Link Analysis (2 lessons)
   - Project

Outline

Outline

Outline

Outline

Outline

Outline

Information Retrieval

Data mining is about extracting interesting information from raw data.

What constitutes “information"?

- Patterns of appearance.
- Association rules between sets of items.
- Clustering of the data points.
- Concepts or categories.
- Principal components or factors.
- . . .

What should be counted as “interesting"?

- Confidence and support.
- Information content.
- Unexpectedness.
- Actionability — The ability to suggest concrete and profitable decision-making.
- ...

Data Analysis

- An indispensable task in almost every discipline of science.
- Search for relationships between a set of externally caused and internal variables.
- Especially important in this era of information and digital technologies.
  - Massive amounts of data are generated at almost all levels of applications.

Inexact Data

- Data are collected from complex phenomena.
- Represent the integrated result of several interrelated variables.
- Variables are often less precisely defined.

Goal

- Interpretation.
  - Distinguish which variable is related to which and how the variables are related.
- Simplification.
  - Reduce the complexity and dimensionality.

EPA Date on Air Pollution

	1970	1975	1980	1985	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Carbon Monoxide	129444	116756	117434	117013	106438	99119	101797	99307	99790	103713	94057	101294	101459	96872	97441
Lead	221	160	74	23	5	5	4	4	4	4	4	4	4	4	4
Nitrogen Oxides	20928	22632	24384	23197	23892	24170	24338	24732	25115	25474	25052	26053	26353	26020	25393
Volatile Organic	30982	26080	26336	24428	22513	21052	21249	11862	21100	21682	20919	19464	19732	18614	18145
PM_{10}	13165	7677	7109	41397	40963	27881	27486	27249	27502	28756	25931	25690	25900	26040	23679
Sulfur Dioxide	31161	28011	25906	23658	23294	23678	23045	22814	22475	21875	19188	18859	19366	19491	18867
$PM_{2.5}$						7429	7317	7254	7654	7012	6909	7267	7065	6773	6773
Ammonia						4355	4412	4483	4553	4628	4662	4754	4851	4929	4963

Table: Annual pollutants estimates (in thousand short tons).

- Who should be blamed for emitting these pollutants?
- How much responsibility should each guilty party bear?

Pixels on Irises

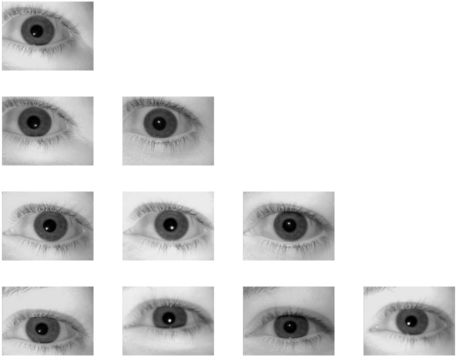


Figure: Intensity image of an iris

- Each iris is a 120×160 pixel grey-scale matrix.
- Can any intrinsic parts that make up these poses be identified?
- Can individual's biometric identification (fingerprint) be specified?

Basic Techniques

- Factor analysis:
  - Identify and test *constructs*, or *factors*, to explain the interrelationships among variables.
    - Each construct itself is a complex image, idea, or theory formed from a number of simpler elements.
- Cluster analysis:
  - Organize information about cases to form relatively *homogenous groups*, or *clusters*.
    - Group members should be highly internally homogenous and highly externally heterogenous.
- Two sides of the same coin!
  - Need a decision on how many factors/clusters to keep.
  - Need a measurement of similarity or dissimilarity.

Data Collection

- Making observation, gathering and pre-processing data :
  - Let $Y = [y_{ij}] \in \mathbb{R}^{n \times \ell}$ denote the matrix of observed data.
    - Assume ℓ entities and n variable.
    - y_{ij} = *standard score* of entity j on variable i (raw scores are normalized to have mean 0 and standard deviation 1).
- Correlation matrix of all n variables:

$$R := \frac{1}{\ell} YY^{\top}. \tag{1}$$

Linear Model

- Assume that y_{ij} is a linearly weighted score of entity j on m factors.

$$Y = AF. \tag{2}$$

- $A = [a_{ik}] \in \mathbb{R}^{n\times m}$ (loading matrix).
  - a_{ik} = the influence of factor k on variable i.
- $F = [f_{kj}] \in \mathbb{R}^{m\times \ell}$ (scoring matrix).
  - f_{kj} = the response of entity j to factor k.

$$\begin{bmatrix} & y_{1j} & \\ & \vdots & \\ \dots & y_{ij} & \dots \\ & \vdots & \\ & y_{nj} & \end{bmatrix} = \underbrace{\begin{bmatrix} & & a_{1k} & & \\ & & \vdots & & \\ a_{i1} & \dots & a_{ik} & \dots & a_{im} \\ & & \vdots & & \\ & & a_{nk} & & \end{bmatrix}}_{\text{influence of factors}} \left.\begin{bmatrix} & f_{1j} & \\ & \vdots & \\ \dots & f_{kj} & \dots \\ & \vdots & \\ & f_{mj} & \end{bmatrix}\right\} \text{response to factors}$$

College Admission Criteria Analysis

- $Y =$ grades of ℓ college students (entities) on n fixed subjects (variables) at the end of their freshman year.

Subject / Student	*Akira*	*Kaya*	*Kenji*	*Taji*	…
Calculus	62	73	85	90	
Chemistry	90	75	63	40	
Physics	70	70	81	80	
History	66	82	88	71	…
⋮	⋮			⋮	

$n \times \ell$

Predictable Performance?

- Has the college selected students with the best potential?
  - Academic performance may depend on a number of factors.
    - What to be considered as admission criteria?
    - Family social status, finance, high school GPA, cultural background, and so on.
- Upon entering the college, students are asked to fill out questionnaires inquiring these factors of his/her background.
  - Individual responses are translated into scores and placed in the corresponding column of the scoring matrix F.

Compose the Criteria

- What is not clear to the educators/administrators,
  - How to choose the factors to compose the questionnaires?
  - How to weight each chosen factor to reflect the effect (loadings) on each particular subject?
- Even less information in practice,
  - No *a priori* knowledge about the number m.
  - No foresight about the character of underlying factors in A.
  - Do not even know the factor scores in F.
  - Only the data matrix Y is observable.
- Explaining the complex phenomena observed in Y, with the help of a minimal number of factors extracted from the data matrix, is the primary and most important goal of factor analysis.

Receptor Model

- A technique used by the air pollution research community.
- Based on the conservation law of mass.
  - Employ mass balance analysis to identify and apportion sources of airborne particulate matter in the atmosphere.
  - The relationships between p sources which contribute m chemical species to n samples lead to a *mass balance equation*.

Mass Balance Equation

$$y_{ij} = \sum_{k=1}^{p} a_{ik} f_{kj}.$$

- y_{ij} = the elemental concentration of the ith chemical measured in the jth sample.
- a_{ik} = the gravimetric concentration of the ith chemical in the kth source.
- f_{kj} = the airborne mass concentration that the kth source has contributed to the jth sample.

What to Look for?

- A typical scenario:
  - Only values of y_{ij} are observable.
  - Neither the sources are known nor the compositions of the local particulate emissions are measured.
- Critical questions:
  - Estimate the number p.
  - Determine the compositions a_{ik}, and the contributions f_{kj} of the sources.
- Nonnegativity requirement:
  - The source compositions a_{ik} and the source contributions f_{kj}, being mass concentrations, must all be nonnegative.

Image Articulation

- Forward problem:
  - Image articulation libraries are made up of images showing a composite object in many articulations and poses.
  - Straightforward application.
- Inverse problem:
  - Identify and classify intrinsic “parts” that make up the object being imaged by multiple observations.
  - Hard, may not be possible.

Mathematical Representation

$$\mathbf{y}_j = \sum_{k=1}^{p} \mathbf{a}_k f_{kj},$$

- $\mathbf{y}_j$ = m pixel values of one image.
- $\mathbf{a}_k$ = one basis part in $\mathbb{R}^m$.
- $[f_{1j}, \ldots, f_{pj}]^\top$ = biometric identification of the jth image.
- Nonnegativity requirement:
  - Basic parts, being images themselves, are necessarily nonnegative.
  - Superposition coefficients, being present of absent, are also necessarily nonnegative.

Factor Extraction

- Two additional assumptions:
  - All sets of factors being considered are uncorrelated.
  - Scores in F for each factor are normalized.

$$\frac{1}{\ell} F F^\top = I_m. \tag{3}$$

- Correlation matrix R is directly related to loading matrix A,

$$R = \frac{1}{\ell}(AF)(AF)^\top = AA^\top. \tag{4}$$

- Factor extraction of $Y \Leftrightarrow$ Matrix factorization of R.
  - Would like to use as few factors as possible.

Interpretation of Loading Matrix A

- a_{i*} = influence from factors in current list on variable i.
  - $\|a_{i*}\|$ = the communality of variable i.
    - Small $\|a_{i*}\| \Rightarrow$ Variable i is of little consequence to current factors.
- a_{*k} = correlations of variables with kth factor.
  - $\|a_{*k}\|$ = the significance of factor k.
    - Variables with high loadings are more “like" the factor.
    - Variables with lows loadings are unlike the factor.
    - Smaill $\|a_{*k}\| \Rightarrow$ Factor k is negligible.

Tasks to Do in Factor Analysis

- Rewrite loadings of variables over some *newly selected* factors.
  - Fewer factors.
  - Manifest more clearly correlation between variables and factors.
- Represent the loading of each variable (row of A) as a single point in the factor space $\mathbb{R}^m$.
  - What if these points cluster around a certain direction?
  - How to find the clustering direction?

What Is Going On?

- Determine new factors as columns of the orthogonal matrix

$$V := [\mathbf{v}_1, \ldots, \mathbf{v}_m] \in \mathbb{R}^{m \times m}. \quad (5)$$

  - Factor loadings with respect to $V \equiv$ Change of basis.

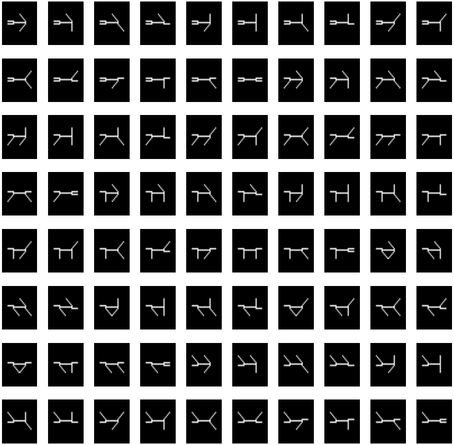
$$Y = AF = \underbrace{(AV)}_{B} \underbrace{(V^T F)}_{G}.$$

  - $R = AA^T = BB^\top$ is independent of factors selected.
    - Prefer to concentrate significance of factors on “fewer" columns of B.
    - Lower rank approximation of R.
- Retrieve B directly without reference to any particular loading matrix A.

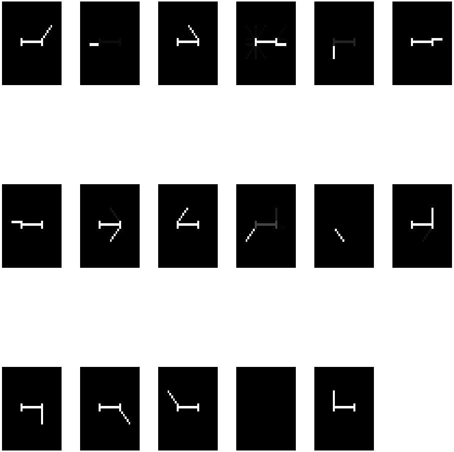
Swimmer Database

- A set of black-and-while stick figures satisfying the so called *separable factorial articulation criteria*.
- Each figure consists of a “torso" of 12 pixels in the center and four “limbs" of six pixels that can be in any one of four positions.
- With limbs in all possible positions, there are a total of 256 figures of dimension 32×32 pixels.
- Can the parts be recovered?

Eighty Swimmers



Seventeen Parts



Pollutant Decomposition

- Assume four principal sectors across the national economy.
  1. Fuel combustion
  2. Industrial Processes:
     - Chemical and allied product manufacturing
     - Metals processing
     - Petroleum and related industries
     - Other industrial processes
     - Solvent utilization
     - Storage and transport
     - Waste disposal and recycling
  3. Transportation
  4. Miscellaneous
- Each subsector contributes certain degree of pollution.

Scenario I: Who Is Doing What Damages?

	1970	1975	1980	1985	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Fuel	41754	40544	43512	41661	40659	39815	39605	40051	38926	38447	36138	36018	35507	34885	34187
Industrial	48222	32364	29615	22389	21909	21120	20900	21102	21438	21467	21190	17469	17988	17868	20460
Transportation	125637	121674	117527	119116	107978	100877	106571	105114	106328	108125	99642	106069	104748	103523	100783
Miscellaneous	10289	6733	10589	46550	46560	45877	42572	40438	41501	45105	39752	43829	46487	42467	39836

- Assume total emissions F from each sector is available.
- Determine a nonnegative matrix A of size 8×4 that solves the optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|Y - AF\|_F^2, \\ \text{subject to} & A \geq 0, \quad \text{and} \quad \sum_{i=1}^{8} a_{ij} = 1, \quad j = 1, \ldots 4. \end{array} \tag{6}$$

  - Each column of A represents the best fitting percentage distribution of pollutants from the emission of the corresponding sector.
  - This is a convex programming problem and the global minimizer is unique.

Comparing NMF and Averaging Results

- Using existing software, such as FMINCON in MATLAB, to find the best fitting distribution A_{opt} to Problem (6).
- The average distribution A_{avg} would have to be obtained by extensive efforts in gathering *itemized* pollutant emissions of each sector per year.

	Fuel	Industrial	Transportation	Miscellaneous
Carbon Monoxide	0.1535	0.3116	0.7667	0.3223
Lead	0.0001	0.0002	0.0002	0
Nitrogen Oxides	0.2754	0.0417	0.1177	0.0113
Volatile Organic	0.0265	0.4314	0.0908	0.0347
PM_{10}	0.0368	0.0768	0.0074	0.4911
Sulfur Dioxide	0.4923	0.0996	0.0112	0.0012
$PM_{2.5}$	0.0148	0.0272	0.0043	0.0761
Ammonia	0.0007	0.0115	0.0016	0.0634

Table: Average distribution of pollutants from sectors.

	Fuel	Industrial	Transportation	Miscellaneous
Carbon Monoxide	0.1925	0.3400	0.8226	0.0090
Lead	0	0.0000	0	0.0000
Nitrogen Oxides	0.0631	0	0.1503	0.1524
Volatile Organic	0.3270	0.2759	0.0272	0
PM_{10}	0.0000	0.1070	0.0000	0.6198
Sulfur Dioxide	0.4174	0.2771	0.0000	0
$PM_{2.5}$	0.0000	0.0000	0	0.1326
Ammonia	0.0000	0	0	0.0862

Table: Optimal distribution of pollutants from sectors with fixed emission estimates.

Serious Discrepancies

- In A_{opt} that 32.70% emissions from the fuel burning contribute to the volatile organic compounds whereas A_{avg} counts only 2.65%.
- In A_{opt} that only 6.31% emissions from the fuel goes to the nitrogen oxides whereas A_{avg} count 27.54%.
- Estimates from the best fitting A_{opt} is inconsistent with the scientific truth. Why?

We discussed in this module a simple linear model mimicking how a centralized data matrix could be factorized in order to retrieve important information. This homework asks you to think about some data in our mundane lives where interesting information can be mined.

1. Describe **two** possible data sets where ”interesting information” might be mined. It will be most fitting if the linear model we described can be applied. If your data sets are for a different model, you need to brief describe what the model is about.
   (a) (20 pts) If the data sets are available over the network but are too large to be downloaded, list their complete URL’s. Make sure that you give credits to the original sources by giving references, if the data set is not your own.
   (b) (20 pts) Provide a short description for each data set. For example, if your data is a matrix, then describe what each dimension or entry represents. You may use a shortened/reduced data set to demonstrate your point.
   (c) (10 pts) Describe what information you want to retrieve. For example, in the linear model, what factors are you looking for.
2. What to submit:
   - Typeset your report. You may prepare your report in whatever format, but your report should be in a PDF file for submission.
   - Your report should be submitted electronically. Further information will be given later.
3. Some possible repositories:
   - https://www.nature.com/sdata/policies/repositories
   - https://archive.ics.uci.edu/ml/datasets.php
   - http://oad.simmons.edu/oadwiki/Data_repositories

Singular Value Decomposition

- Any matrix $A \in \mathbb{R}^{m \times n}$ enjoys a singular value decomposition (SVD)

$$A = U \Sigma V^\top$$

  where
  - $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal.
  - $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal.
- This is an important matrix factorization known before mathematical theory was complete and across many fields.
  - Often is the first computational step in many numerical algorithms.
  - Also is the first conceptual step in many theoretical studies.

Variational Property

- Given $A \in \mathbb{R}^{m \times n}$ ($m \geq n$), the image of the unit sphere in $\mathbb{R}^n$ under A is a hpyerellipse (of dimension n) in $\mathbb{R}^m$.
  - $\mathbf{u}_i \in \mathbb{R}^m$ = unit directions of the principal semiaxes of the hyperellipse = *left singular vectors* of A.
  - $\mathbf{v}_i \in \mathbb{R}^n$ = unit directions of the preimage of $\mathbf{u}_i$ = *right singular vectors* of A.
  - σ_i = length of the principal semiaxes of the hyperellipse = *singular values* of A.
- Rewrite the relationship as:

$$Av_i = \sigma_i u_i. \tag{7}$$

  - It is clear that all $_i$, $i = 1, \ldots, n$ are mutually orthogonal.
  - It can be shown that all $\mathbf{v}_i$, $i = 1, \ldots n$ are also mutually orthogonal.

Relation to Eigenvalues

- Recall that $A^T A \in \mathbb{R}^{n \times n}$ is symmetric and positive semi-definite.
  - $A^T A$ has a complete set of eigenvectors.
  - All eigenvalues of $A^T A$ are nonnegative.
  - Denote the positive eigenvalues of $A^T A$ by $\sigma_1^2 \geq \ldots \geq \sigma_r^2 > 0$.
    - It can be proved that $r = \text{rank}(A)$.
  - Denote the normalized (and orthogonal) eigenvector of $A^T A$ associated with σ_i^2 by v_i
- Some important observations:
  - The two matrices $A^T A$ and AA^T have the same positive eigenvalues.
  - Av_i is an eigenvector of AA^T associated with eigenvalue σ_i^2.
  - The vector $u_i := Av_i / \sigma_i$ is a normalized eigenvector of AA^T.

Completion

- Let $V := [v_1, \ldots, v_n] \in \mathbb{R}^{n\times n}$ whose columns v_i are orthonormal eigenvectors of $A^T A$.
- Define $U := [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m}$ where
  - For $j = 1, \ldots, r$, $u_j := Av_j/\sigma_j$, and
  - For $j = r+1, \ldots, m$, $\{u_{r+1}, \ldots, u_m\}$ are orthonormal eigenvectors corresponding to the zero eigenvalue of AA^T.
- Define $\Sigma := \operatorname{diag}\{\sigma_1, \ldots, \sigma_r\}$.
- With U, Σ and V given above, it must be true that

$$A = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^T. \tag{8}$$

  - Write $U = [U_1, U_2],\ V = [V_1, V_2]$.
  - Observe that

$$U^T AV = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} A[V_1, V_2] = \begin{bmatrix} U_1^T AV_1 & U_1 AV_2 \\ U_2^T AV_1 & U_2^T AV_2 \end{bmatrix}.$$

  - Note that $AV_2 = 0$, $U_2^T AV_1 = U_2^T U_1 \Sigma = 0$ and $U_1^T AV_1 = \Sigma$ by the choice of U.

Decompose a Random Variable

- Let $\mathcal{X} \in \mathbb{R}^n$ denote a random column vector.

$$cov(\mathcal{X}) := \mathcal{E}[(\mathcal{X} - \mathcal{E}[\mathcal{X}])(\mathcal{X} - \mathcal{E}[\mathcal{X}])^\top] = \sum_{j=1}^{n} \lambda_j \mathbf{u}_j \mathbf{u}_j^\top.$$

  - $+\mathbf{u}_1, \ldots, \mathbf{u}_n$ are deterministic and orthonormal in $\mathbb{R}^n$.

- The random column vector $\mathcal{X}$ can be expressed as

$$\mathcal{X} = \sum_{j=1}^{n} (\mathbf{u}_j^T \mathcal{X}) \mathbf{u}_j.$$

  - Each coefficient $\alpha_j := \mathcal{X}^T u_j$ itself is a random variable.

Random Coefficients

$$
\begin{aligned}
\mathcal{E}[\alpha] &= U^T \mathcal{E}[\mathcal{X}], \\
cov(\alpha) &= \text{diag}\{\lambda_1, \dots, \lambda_n\}.
\end{aligned}
$$

- The randomness of $\mathcal{X}$ is due to the randomness of α.
- Variance measures the unpredictability of a random variable.
- Random coefficients α_j are mutually stochastically independent.

Ranking the Randomness

- Larger eigenvalue $\lambda_j \Rightarrow$ Larger variance of $\alpha_j \Rightarrow$ More randomness in the direction $\mathbf{u}_j$.
- Rank the importance of corresponding eigenvectors $\mathbf{u}_j$ as *essential* components for the variable $\mathcal{X}$ according to the magnitude of λ_j.
  - If truncation is necessary, those eigenvectors corresponding to smaller variances should be thrown away first.

Low Dimensional Approximation

- Best approximate $\mathcal{X}$ by a unbiased variable $\tilde{\mathcal{X}}$.
  - $\tilde{\mathcal{X}}$ is limited to an *m*-dimensional subspace with $m < n$.
  - $\mathcal{E}[\|\mathcal{X} - \hat{\mathcal{X}}\|^2]$ is minimized.
- Among *all* unbiased variables restricted to *any* *m*-dimensional subspaces in $\mathbb{R}^n$,

$$\hat{\mathcal{X}} := \sum_{j=1}^{r} (\mathbf{u}_j^T \mathcal{X}) \mathbf{u}_j \tag{9}$$

is the best linear minimum-variance estimate of $\mathcal{X}$.

Truncation in Sample Space

- Collect ℓ random samples of $\mathcal{X}$.
  - Samples are recorded in a $n \times \ell$ matrix X.
  - Law of large numbers $\Rightarrow$ Can recoup stochastic properties of $\mathcal{X}$ from X with large enough ℓ.
- How to retrieve a sample data matrix from X to represent the minimum-variance approximation $\hat{\mathcal{X}}$ of $\mathcal{X}$?
  - Spectral decomposition of sample covariance:

$$R = \frac{XX^\top}{\ell} = \sum_{i=1}^{n} \mu_i \mathbf{u}_i \mathbf{u}_i^\top. \tag{10}$$

  - Projection of $\mathcal{X}$ to $\hat{\mathcal{X}}$ $\Rightarrow$

$$\hat{X} := \sum_{j=1}^{r} (\mathbf{u}_j^\top X)\mathbf{u}_j. \tag{11}$$

Truncated SVD

- *Low dimension estimate* $\hat{\mathcal{X}}$ to (continuous) random variable $\mathcal{X} \Rightarrow$ *Low rank approximation* $\hat{X}$ to (discrete) random sample matrix X.
- The singular value decomposition of X:

$$X = U\Sigma V^\top = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top \qquad (12)$$

  - Eigenvectors of R = Left singular vectors $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$.
  - Singular values $\sigma_i = \sqrt{\ell \mu_i}$ in the same ordering as eigenvalues μ_i.
  - TSVD of X = $\sum_{i=1}^{m} (\sigma_i \mathbf{v}_i^\top) \mathbf{u}_i$.
- The TSVD of a give data matrix X representing random samples of an *unknown* random variable $\mathcal{X}$ has a statistical meaning.

Computational Challenges

- Must mine through very large scale of data.
  - Matrix factorization becomes increasingly difficult.
- Data set changes dynamically.
  - Adding or deleting information requires updating or downdating current factorization.
- No obvious way to determine optimal rank m.
- Additional constraints on data for feasibility and interpretability.
  - Nonnegativity.
  - Algebraic variety.
  - Binary.
- Need structured low rank approximation.

Continuous Acquisition

- An inevitable task.
  - Robots crawl the web.
  - Software automation.
- Most search engines prepare database continually.
  - Index documents.
  - Mine and retrieve information.
  - Store the data in an organized way for quick reference when needed.

Ranking Retrieved Information

- A query usually can bring up deluging information.
  - Must be sorted again to reveal the most relevant pages.
- Link analysis help to tackle this ranking problem.
  - Eigenvector computation.

Power Iteration

- Given a matrix $A \in \mathbb{C}^{n \times n}$,
  - Begin with an arbitrary $\mathbf{x}^{(0)} \in \mathbb{C}^n$.
  - Generate the sequence $\{\mathbf{x}^{(k)}\}$ until convergence by

$$\mathbf{w}^{(k)} := A\mathbf{x}^{(k-1)};$$
$$\mathbf{x}^{(k)} := \frac{\mathbf{w}^{(k)}}{\|\mathbf{w}^{(k)}\|_\infty}.$$

- The normalization is for the purpose of avoiding overflow or underflow.
  - Any norm can be used for the normalization. The sup-norm is particularly convenient.

Delayed Normalization

- The normalization needs not be done at every step because

$$
\begin{aligned}
\mathbf{x}^{(k)} &= \frac{A\mathbf{x}^{(k-1)}}{\|A\mathbf{x}^{(k-1)}\|_\infty} &= \frac{A\left(\frac{\mathbf{w}^{(k-1)}}{\|\mathbf{w}^{(k-1)}\|_\infty}\right)}{\|A\left(\frac{\mathbf{w}^{(k-1)}}{\|\mathbf{w}^{(k-1)}\|_\infty}\right)\|_\infty} \\
&= \frac{A^2\mathbf{x}^{(k-2)}}{\|A^2\mathbf{x}^{(k-2)}\|_\infty} &= \frac{A^k\mathbf{x}^{(0)}}{\|A^k\mathbf{x}^{(0)}\|_\infty}.
\end{aligned}
$$

Dominant Eigenpair

- Assume A is diagonalizable
  - Eigenvalues are arranged as $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$.
  - Corresponding eigenvectors are $x_1, \ldots x_n$.
- Write $x^{(0)} = \sum_{i=1}^{n} \alpha_i x_i$.
  - Note that

$$
\begin{aligned}
Ax^{(0)} &= \sum_{i=1}^{n} \alpha_i \lambda_i x_i \\
A^k x^{(0)} &= \sum_{i=1}^{n} \alpha_i \lambda_i^k x_i = \lambda_1^k \left(\alpha_1 x_1 + \sum_{i=2}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k x_i \right).
\end{aligned}
$$

  - Assume $\alpha_1 \neq 0$. (This is guaranteed if $x^{(0)}$ is selected randomly.)

Convergence

- As $k \to \infty$, the vector $A^k x^{(0)}$ *behaves like* $\alpha_1 \lambda_1^k x_1$ in the sense that contributions from $x_2, \ldots x_n$ becomes less and less significant.
  - Normalization makes $x^{(k)} \to \frac{\alpha_1 \lambda_1^k}{|\alpha_1 \lambda_1^k|} \frac{x_1}{\|x_1\|_\infty}$.
  - The sequence $\{x^{(k)}\}$ converges to an eigenvector associated with the eigenvalue λ_1.
  - Also, $w^{(k+1)} = Ax^{(k)} \to \lambda_1 x^{(k)}$. So $\frac{w^{(k+1)})_j}{x_j^{(k)}} \to \lambda_1$.
- The rate of convergence of power method depends on the ratio $\frac{\lambda_2}{\lambda_1}$.

HITS Algorithm

- Given a query, assume that n Web pages have been matched through some search mechanism.
- For each page P_i,
  - $\mathbb{I}_i$ = set of pages linking into P_i.
    - a_i = authority score.
  - $\mathbb{O}_i$ = set of pages linking out of P_i.
    - h_i = hub score.
- Starting with $h_i^{(0)} = \frac{1}{n}$, the pages compete for their authorities and hub reputations.

$$a_i^{(k)} = \sum_{j:P_j \in \mathbb{I}_i} h_j^{(k-1)}, \quad h_i^{(k)} = \sum_{j:P_j \in \mathbb{O}_i} a_j^{(k)}. \tag{13}$$

Score Evolution

- L = the adjacency matrix.

$$L_{ij} = \begin{cases} 1 & \text{if } P_j \in \mathbb{O}_i, \\ 0 & \text{otherwise.} \end{cases}$$

  - Successive refinement.

$$\mathbf{a}^{(k)} = L^\top \mathbf{h}^{(k-1)}, \quad \mathbf{h}^{(k)} = L\mathbf{a}^{(k)}. \tag{14}$$

  - Recursion.

$$\mathbf{a}^{(k)} = (L^\top L)\mathbf{a}^{(k-1)}, \quad \mathbf{h}^{(k)} = (LL^\top)\mathbf{h}^{(k-1)}. \tag{15}$$

- With appropriate normalization, this algorithm amounts to the power method.
  - Computes the dominant eigenvector.
  - The limit points provide a ranking of importance for each page.

PageRank

- For each page P_i,
  - $|\mathbb{O}_i|$ = number of out lines from P_i.
  - r_i = page rank.

$$r_i^{(k)} := \sum_{j:P_j \in \mathbb{I}_i} \frac{r_j^{(k-1)}}{|\mathbb{O}_j|}. \tag{16}$$

- H = the modified adjacency matrix.

$$H_{ij} = \begin{cases} \frac{1}{|\mathbb{O}_i|} & \text{if } P_j \in \mathbb{O}_i, \\ 0 & \text{otherwise.} \end{cases}$$

  - H is row stochastic?!
  - Probability distribution (row) vector $\mathbf{r}^{(k)} = [r_1^{(k)}, \ldots r_n^{(k)}]$.
  - Random walk on the hyperlinks.

$$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} H. \tag{17}$$

Google Matrix

- Technical issues:
  - Dead end page $\Rightarrow \mathbb{O}_i$ is empty.
  - Cyclic traps $\Rightarrow$ No convergence.
- Modify hyperlink matrix H to

$$G = \alpha \underbrace{\left(H + \frac{\mathbf{a}\mathbf{1}^\top}{n}\right)}_{\text{remove dangles}} + (1-\alpha) \underbrace{\frac{\mathbf{1}\mathbf{1}^\top}{n}}_{\text{enforce irreducibility}}, \qquad (18)$$

  - $\alpha \in [0, 1]$ is a parameter.
- G is row stochastic, irreducible and aperiodic.
  - The stationary distribution vector $\mathbf{r} = \mathbf{r}G$ exists and is unique.
- $\mathbf{r}$ provides a ranking of importance for each page.

Computation

- Google matrix G has indexed billions of pages and the size is constantly growing.
  - Iterative method is perhaps the only choice of method.
  - Power method converges at the rate of its second largest eigenvalue $|\lambda_2|$.

$$|\lambda_2| = \alpha.$$

- It has been said that Google uses $\alpha = .85$.
  - PageRank is within 10^{-4} accuracy by 50 iterations, regardless the size of the matrix.
- Link structure over the web is extremely dynamical.
  - The PageRank needs update periodically.
  - The mechanism of effectively updating an old PageRank is still an ongoing research endeavor.

Modifications

- PageRank is a useful tool in many Web search technologies and beyond.
  - Spam detection.
  - Crawler configuration.
  - Trust networks.
- PageRank can also been modified to correspond to different configurations.
  - HostRank: Compress webpages within a specific domain into one host, form a hostgraph, and apply the PageRank model to the smaller hostgraph.
    - Reduce both the number of iterations and the work per iteration $\Longrightarrow$ acceleration.
    - Global PageRank $\approx$ Local PageRank $\times$ HostRank.
  - Assign different link weights for internal or external linkages.
- (Question) How would these alternatives be described mathematically and how would they change the results?

PageRank Linear System

- The ultimate goal of the random walk on the hyperlinks is to solve

$$\mathbf{r} = \mathbf{r}G$$

where $\mathbf{r}$ is a row vector and $\|\mathbf{r}\|_1 = 1$.

- Abbreviate $\mathbf{v} := \frac{\mathbf{1}}{n}$, denote $\mathbf{r}^\top = \mathbf{x}$ and rewrite the system as

$$\begin{aligned}\left(\alpha H^\top + \alpha \mathbf{v}\mathbf{a}^\top + (1-\alpha)\mathbf{v}\mathbf{1}^\top\right)\mathbf{x} &= \mathbf{x}, \\ (\underbrace{I - \alpha H^\top}_{R} - \underbrace{\alpha \mathbf{v}\mathbf{a}^\top}_{\text{rank one}})\mathbf{x} &= (1-\alpha)\mathbf{v}.\end{aligned}$$

Sherman-Morrison Formula

Theorem

Suppose $A \in \mathbb{R}^{n\times n}$ is invertible and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are column vectors. Then

1. *$A + \mathbf{u}\mathbf{v}^\top$ is invertible if and only if $1 + \mathbf{v}^\top A^{-1}\mathbf{u} \neq 0$.*
2. *The inverse of the rank-1 updated matrix is given by*

$$(A + \mathbf{u}\mathbf{v}^\top)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1 + \mathbf{v}^\top A^{-1}\mathbf{u}}. \tag{19}$$

(Remark:) The proof is straightforward, but to get the formula for the first time is not. How to be the first discoverer?

Back to the PageRank Linear System

- $R = I - \alpha H^\top$ is invertible because $\alpha < 1$.
- Identify $-\alpha \mathbf{v}\mathbf{a}^\top$ as the rank-1 update and obtain

$$(R - \alpha \mathbf{v}\mathbf{a}^\top)^{-1} = R^{-1} + \frac{R^{-1}\mathbf{v}\mathbf{a}^\top R^{-1}}{\frac{1}{\alpha} + \mathbf{a}^\top R^{-1}\mathbf{v}}.$$

- The PageRank vector $\mathbf{x}$ should be given by

$$\mathbf{x} = (1-\alpha)\left(R^{-1} + \frac{R^{-1}\mathbf{v}\mathbf{a}^\top R^{-1}}{\frac{1}{\alpha} + \mathbf{a}^\top R^{-1}\mathbf{v}}\right)\mathbf{v}.$$

A Simplified System

$$(I - \alpha H^\top)\mathbf{y} = \mathbf{v}.$$

- Do a rearrangement:

$$\mathbf{x} = (1-\alpha)\left(1 + \frac{\mathbf{a}^\top \mathbf{y}}{\frac{1}{\alpha} + \mathbf{a}^\top \mathbf{y}}\right)\mathbf{y}.$$

- Suffices to solve for **y** only. Then obtain **x** by

$$\mathbf{x} = \frac{\mathbf{y}}{\|\mathbf{y}\|_1}.$$

  - Take advantage of the sparsity of H itself.
  - Many effective iterative methods available for tackling large sparse linear systems.
  - Updating and downdating remain challenging.

Exploiting the Dangles

- Most pages are dangles, causing zero rows in H.
- If H has many zeros rows, then further reduction is possible.
  - Separate dangling from non-dangling pages.

$$\begin{bmatrix} I - \alpha H_1^\top & 0 \\ -\alpha H_2^\top & I \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}.$$

  - Solve for $\mathbf{y}_1$ first from

$$(I - \alpha H_1^T)\mathbf{y}_1 = \mathbf{v}_1.$$

  - Then $\mathbf{y}_2 = \mathbf{v}_2 + \alpha H_2^\top \mathbf{y}_1$.

Conclusion

- Have shown only the tip of the iceberg.
- Applied linear algebra plays a fundamental role in data mining.
- This is a world changing application.
- Many open areas for further study.