Matrix Computations
Chapter 1 Introduction
Section 1.1 Course Overview

Jie Lu
School of Information Science and Technology
ShanghaiTech University

Course Information

- Prerequisites: Mathematical analysis/Calculus, Linear algebra

- Textbook: *Matrix Computations* by G. Golub and C. Van Loan
- References: See course syllabus

- Instructor: Jie Lu (SIST 1D-201B, lujie@shanghaitech.edu.cn)
- Please contact me via email (not WeChat)

- Office hour: SIST 1D-201B, Tuesday 5pm-6pm or by appointment

Course Overview

- Course organization
  - *Classroom lectures* (Week 1-12): fundamentals and theory for matrix computations and analysis
  - *Computer programming experiments* (Week 13-16): algorithm implementation and course project

- Matrix arises in an abundance of engineering fields such as
  - machine learning, computer vision, data mining
  - control and robotics, signal processing, communications
  - circuits, power systems, biomedical image processing

- After learning the course, you are expected to
  - understand principles and implementations of various tools for matrix computations and analysis
  - apply appropriate tools in real-world engineering problems
  - create novel tools for solving problems of recent interest at research level

Course Evaluation

- Assignments (30%)
- Midterm exam (40%)
- Project (including lab simulations) (30%)

- Classroom attendance (online after Week 12) is mandatory. May cause score penalty unless you ask for a leave with an acceptable reason

- Cheating and plagiarism are prohibitive. May cause course failure or more severe penalty

Test Yourself

Ready for this course?

- Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. How to express the distance between $\mathbf{x}$ and $\mathbf{y}$?

- Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{y} \in \mathbb{R}^m$. How many solutions can $\mathbf{y} = \mathbf{A}\mathbf{x}$ have? When does $\mathbf{y} = \mathbf{A}\mathbf{x}$ have no solution?

- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$. If $\mathbf{A}\mathbf{x} = \mathbf{0}$ has a nonzero solution, what can you say about the eigenvalues of $\mathbf{A}$?

A Taste of Matrix Computations and Analysis

Linear System of Equations

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{y} \in \mathbb{R}^m$

- How to solve $\mathbf{y} = \mathbf{A}\mathbf{x}$?
- How sensitive is the solution $\mathbf{x}$ to the errors of $\mathbf{A}$ and $\mathbf{y}$?
- How to solve $\mathbf{y} = \mathbf{A}\mathbf{x}$ when there is no exact solution?
- How to solve $\mathbf{y} = \mathbf{A}\mathbf{x}$ when n is huge?
- How to find a sparse solution to $\mathbf{y} = \mathbf{A}\mathbf{x}$?

Exploit problem structure to develop efficient, robust, scalable solvers

Least-squares Problem

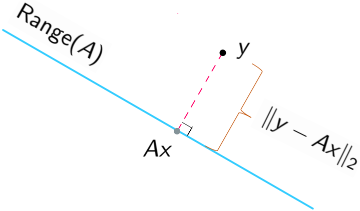
Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ $(m \geq n)$ and $\mathbf{y} \in \mathbb{R}^m$, consider

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

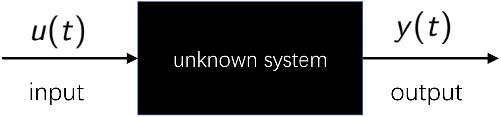
- Overdetermined, more equations than unknowns, often no solution

Approximately solve $\mathbf{y} = \mathbf{A}\mathbf{x}$ by minimizing

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$$



System Identification



- Measure input and output signals at $t = 0, 1, \dots, N$
- *System Identification*: Find a reasonable system model based on measured I/O data

System Identification (cont'd)

Example: Moving average (MA) model with $n << N$ delays
predicted output $\hat{y}(t) = h_0 u(t) + h_1 u(t-1) + \cdots + h_n u(t-n)$

$$\begin{bmatrix} \hat{y}(n) \\ \hat{y}(n+1) \\ \vdots \\ \hat{y}(N) \end{bmatrix} = \begin{bmatrix} u(n) & u(n-1) & \cdots & u(0) \\ u(n+1) & u(n) & \cdots & u(1) \\ \vdots & \vdots & \vdots & \vdots \\ u(N) & u(N-1) & \cdots & u(N-n) \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_n \end{bmatrix}$$

Model prediction error

$$\mathbf{e} = [y(n) - \hat{y}(n), y(n+1) - \hat{y}(n+1), \cdots, y(N) - \hat{y}(N)]^T$$

Least-squares identification: Choose $h_0, \ldots, h_n$ that minimize $\|\mathbf{e}\|_2$

Linear Prediction

Let $\{\mathbf{y}_t\}_{t\geq 0}$ be a time series

Autoregressive (AR) model:

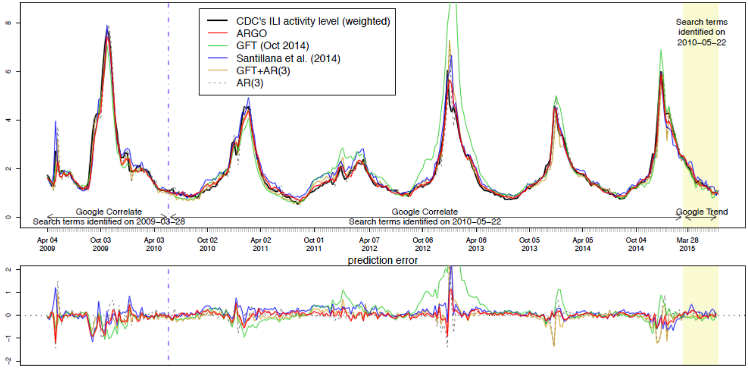
$$\mathbf{y}_t = \alpha_1 \mathbf{y}_{t-1} + \alpha_2 \mathbf{y}_{t-2} + \cdots + \alpha_q \mathbf{y}_{t-q} + \mathbf{v}_t, \quad t = 0, 1, 2, \ldots$$

where v_t represents (unknown) noise or modeling error

Determine the coefficients $\alpha_1, \ldots, \alpha_q$.

How to formulate this into a least-squares problem?

Real-time Prediction of Flu Activity



- S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using Google search data via ARGO," *Proc. of the National Academy of Sciences*, vol. 112, no. 47, pp. 14473–14478, 2015.

Sparse Recovery

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m < n$ and $\mathbf{y} \in \mathbb{R}^{m}$

Find a sparse solution $\mathbf{x} \in \mathbb{R}^{n}$ (with as many 0 elements as possible)

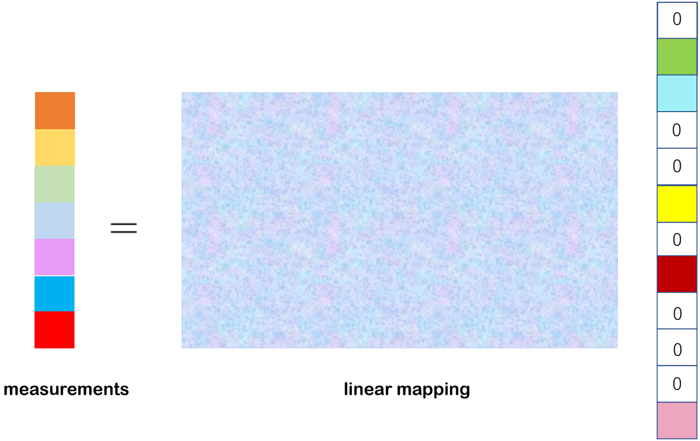


Image Reconstruction from Incomplete Frequency Info

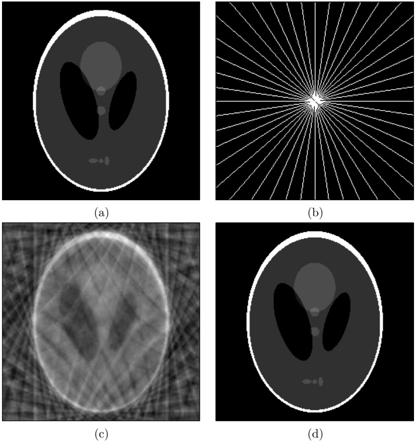


Figure (a) The Logan-Shepp phantom test image. (b) Sampling “domain” in the frequency plane; Fourier coefficients are sampled along 22 approximately radial lines. (c) Minimum energy reconstruction obtained by setting unobserved Fourier coefficients to zero. (d) Reconstruction obtained by minimizing the total-variation. The reconstruction is an exact replica of (a). **Reference**: E. J. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” *IEEE Trans. Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

Eigenvalue Problem

Given $\mathbf{A} \in \mathbb{R}^{n \times n}$, find $\lambda \in \mathbb{C}$ and $\mathbf{v} \in \mathbb{C}^n$ such that

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$$

λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{v}$ is an eigenvector associated with λ

Eigendecomposition: For $\mathbf{A} = \mathbf{A}^T = \mathbb{R}^{n \times n}$, $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$$

where $\mathbf{Q}$ is an orthogonal matrix ($\mathbf{Q}^T = \mathbf{Q}^{-1}$) and $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal entries are $\mathbf{A}$'s eigenvalues

PageRank

An algorithm used by Google to rank the pages of a search result

More important webpages are likely to receive more links from other websites

Determine the importance of each webpage based on the quality and quantity of links pointing to it

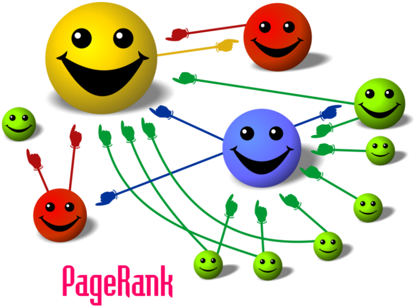


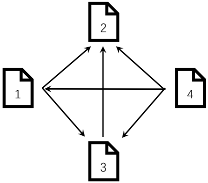
Figure: PageRank. Source: Wikipedia

PageRank (cont'd)

Let v_i be the importance score of page $i = 1, \dots, n$, $\mathcal{L}_i$ be the set of pages containing a link to page i, and c_j be the number of outgoing links from page j

$$\sum_{j \in \mathcal{L}_i} \frac{v_j}{c_j} = v_i, \quad \forall i = 1, \dots, n$$

Example:



$$\begin{array}{ll} \mathcal{L}_1 = \{4\} & c_1 = 2 \\ \mathcal{L}_2 = \{1, 3, 4\} & c_2 = 0 \\ \mathcal{L}_3 = \{1, 4\} & c_3 = 1 \\ \mathcal{L}_4 = \emptyset & c_4 = 3 \end{array}$$

$$\begin{bmatrix} 0 & 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 1 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}$$

Principal Component Analysis

Example: Rate various universities according to the following criteria:

- research funding
- publication
- research impact
- faculty-student ratio
- teaching excellence
- internationalization
- employability
- employer recognition
- livability
- student activities

Graphical displays are not helpful. Need 120 three-dimensional scatterplots

We hope to reduce the number of variables to a few interpretable linear combinations of the data

Each linear combination will correspond to a principal component

Principal Component Analysis (Cont'd)

Suppose we have a random vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$ with population variance-covariance matrix

$$\text{var}(\mathbf{x}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix}$$

whose eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$

Consider the linear combinations (principal components)

$$\begin{aligned} y_1 &= \eta_{11}x_1 + \eta_{12}x_2 + \cdots + \eta_{1n}x_n \\ y_2 &= \eta_{21}x_1 + \eta_{22}x_2 + \cdots + \eta_{2n}x_n \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ y_n &= \eta_{n1}x_1 + \eta_{n2}x_2 + \cdots + \eta_{nn}x_n \end{aligned}$$

Principal Component Analysis (Cont'd)

1st principal component y_1: determine $\eta_1 = [\eta_{11}, \ldots, \eta_{1n}]^T$ by

$$\begin{array}{ll} \max_{\eta_1} & \mathrm{var}(y_1) = \sum_{\ell=1}^{n} \sum_{s=1}^{n} \eta_{1\ell}\eta_{1s}\sigma_{\ell s} = \eta_1^T \boldsymbol{\Sigma} \eta_1 \\ \text{s.t.} & \eta_1^T \eta_1 = 1 \end{array}$$

y_1 has the maximum variance among all the linear combinations of $x_1, \ldots, x_n$. Account for as much variation in the data as possible

2nd principal component y_2: determine $\eta_2 = [\eta_{21}, \ldots, \eta_{2n}]^T$ by

$$\begin{array}{ll} \max_{\eta_2} & \mathrm{var}(y_2) = \sum_{\ell=1}^{n} \sum_{s=1}^{n} \eta_{2\ell}\eta_{2s}\sigma_{\ell s} = \eta_2^T \boldsymbol{\Sigma} \eta_2 \\ \text{s.t.} & \eta_2^T \eta_2 = 1 \\ & \mathrm{cov}(y_1, y_2) = \sum_{\ell=1}^{n} \sum_{s=1}^{n} e_{1\ell} e_{2s}\sigma_{\ell s} = \eta_1^T \boldsymbol{\Sigma} \eta_2 = 0 \end{array}$$

Account for as much of the remaining variation as possible, not correlated with y_1

Principal Component Analysis (Cont'd)

***i*th principal component** y_i: determine $\eta_i = [\eta_{i1}, \ldots, \eta_{in}]^T$ by

$$\begin{aligned} \max_{\eta_i} \quad & \mathrm{var}(y_i) = \sum_{\ell=1}^{n} \sum_{s=1}^{n} \eta_{i\ell}\eta_{is}\sigma_{\ell s} = \eta_i^T \boldsymbol{\Sigma} \eta_i \\ \text{s.t.} \quad & \eta_i^T \eta_i = 1 \\ & \mathrm{cov}(y_j, y_i) = \sum_{\ell=1}^{n} \sum_{s=1}^{n} \eta_{j\ell}\eta_{is}\sigma_{\ell s} = \eta_j^T \boldsymbol{\Sigma} \eta_i = 0, \\ & \qquad\qquad\qquad\qquad \forall j = 1, \ldots, i-1 \end{aligned}$$

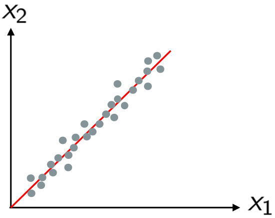
Account for as much of the remaining variation as possible, not correlated with previous principal components

Solution: Each η_i is an eigenvector associated with the eigenvalue λ_i of $\boldsymbol{\Sigma}$ and $\mathrm{var}(y_i) = \lambda_i$

Note: $\boldsymbol{\Sigma}$ is often unknown, but we may estimate $\boldsymbol{\Sigma}$ by the sample variance-covariance matrix

Principal Component Analysis (Cont'd)

Dimension Reduction: When there is a strong correlation between some x_i's, the data may more or less fall on a line or plane with lower dimension



The total variation of x is characterized by

$$\text{trace}(\mathbf{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

The ith principal component accounts for $\frac{\lambda_i}{\lambda_1+\cdots+\lambda_n}$ of the total variation

Only retain the first $k < n$ principal components

- Small k leads to simple interpretation of the data
- yet we need sufficiently large (as close to 1 as possible) $\frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_n}$

Singular Value Decomposition

Not very matrix has an eigendecomposition, but every matrix has a singular value decomposition (SVD)

SVD: $\mathbf{A} \in \mathbb{R}^{m \times n}$ can be decomposed into

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

where $\mathbf{U} \in \mathbb{R}^{m \times m}$, $\mathbf{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is such that the $(i, i)-$entry is a (nonnegative) singular value of $\mathbf{A}$.

Low-rank Approximation

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank k and $r \in \{1, \dots, k-1\}$, find $\hat{\mathbf{A}} \in \mathbb{R}^{m \times n}$ with $\text{rank}(\hat{\mathbf{A}}) \leq r$ such that $\|\mathbf{A} - \hat{\mathbf{A}}\|_2$ or $\|\mathbf{A} - \hat{\mathbf{A}}\|_F$ is minimum

Solution: Truncated SVD

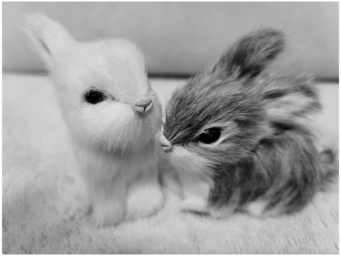
SVD of $\mathbf{A}$:

$$\mathbf{A} = \mathbf{U\Sigma V}^T = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

where $\sigma_1 \geq \cdots \geq \sigma_k > \sigma_{k+1} = \cdots = \sigma_{\min\{m,n\}} = 0$ are the singular values of $\mathbf{A}$

$$\hat{\mathbf{A}} = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

Image Compression



original image, size: 639 x 853

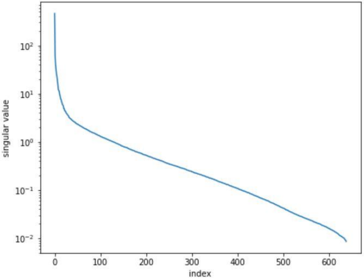
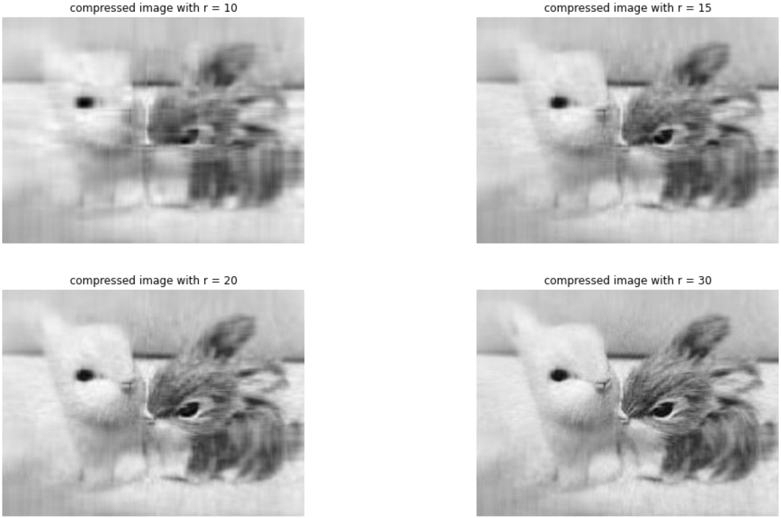


Image compression (cont'd)



compressed image with r = 10

compressed image with r = 15

compressed image with r = 20

compressed image with r = 30