Reflection Matrices

A matrix $\mathbf{H} \in \mathbb{R}^{m \times m}$ is called a reflection matrix if

$$\mathbf{H} = \mathbf{I} - 2\mathbf{P},$$

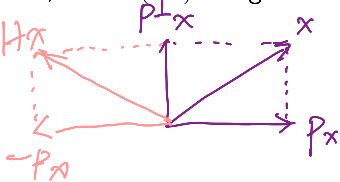
where $\mathbf{P}$ is an orthogonal projector (symmetric and idempotent)

Interpretation: Let $\mathbf{P}^{\perp} = \mathbf{I} - \mathbf{P}$ be the orthogonal complement projector

$$\mathbf{x} = \mathbf{P}\mathbf{x} + \mathbf{P}^{\perp}\mathbf{x}, \qquad \mathbf{H}\mathbf{x} = -\mathbf{P}\mathbf{x} + \mathbf{P}^{\perp}\mathbf{x}$$

The vector $\mathbf{Hx}$ is a reflected version of $\mathbf{x}$, with $\mathcal{R}(\mathbf{P}^{\perp})$ being the "mirror"

$$Hx = (I-2P)(Px + P^{\perp}x)$$
$$= Px - 2\underbrace{PP}_{=P \text{ idempotent}}x + P^{\perp}x - 2\underbrace{PP^{\perp}}_{=0}x$$
$$= -Px + P^{\perp}x$$



A reflection matrix is orthogonal:

$$HH^T = \mathbf{H}^T\mathbf{H} = (\mathbf{I} - 2\mathbf{P})(\mathbf{I} - 2\mathbf{P}) = \mathbf{I} - 4\mathbf{P} + 4\mathbf{P}^2 = \mathbf{I} - 4\mathbf{P} + 4\mathbf{P} = \mathbf{I}$$

H symmetric

Householder Reflections

Problem: Given $\mathbf{x} \in \mathbb{R}^m$, find an orthogonal $\mathbf{H} \in \mathbb{R}^{m \times m}$ such that

$$\mathbf{Hx} = \begin{bmatrix} \beta \\ \mathbf{0} \end{bmatrix} = \beta \mathbf{e}_1, \qquad \text{for some } \beta \in \mathbb{R}$$

Householder reflection: Let $\mathbf{v} \in \mathbb{R}^m$, $\mathbf{v} \neq \mathbf{0}$, and let

$$\mathbf{H} = \mathbf{I} - \frac{2}{\|\mathbf{v}\|_2^2} \mathbf{v}\mathbf{v}^T,$$

which is a reflection matrix with $\mathbf{P} = \mathbf{v}\mathbf{v}^T / \|\mathbf{v}\|_2^2$

symmetric and idempotent

$$P \cdot P = \frac{vv^T}{\|v\|_2^2} \cdot \frac{vv^T}{\|v\|_2^2} = \frac{v\,\underbrace{(v^T v)}_{=\|v\|_2^2}\,v^T}{\|v\|_2^4} = \frac{vv^T}{\|v\|_2^2} = P$$

Householder Reflections (cont'd)

$$= x - \frac{2}{\|v\|_2^2} v \underbrace{(v^T x)}_{\text{scalar}}$$

$$\mathbf{Hx} = \left(\mathbf{I} - \frac{2}{\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{v}^T\right)\mathbf{x} = \mathbf{x} - \frac{2\mathbf{v}^T\mathbf{x}}{\mathbf{v}^T\mathbf{v}}\mathbf{v} \quad //$$

We want $\mathbf{Hx}$ to be a multiple of $\mathbf{e}_1$. Hence, we require $\mathbf{v} \in \mathrm{span}\{\mathbf{x}, \mathbf{e}_1\}$
Let $\mathbf{v} = \mathbf{x} + \alpha\mathbf{e}_1$. Then,

first element in x

$$\mathbf{v}^T\mathbf{x} = \mathbf{x}^T\mathbf{x} + \alpha x_1, \quad \mathbf{v}^T\mathbf{v} = \mathbf{x}^T\mathbf{x} + 2\alpha x_1 + \alpha^2$$

It follows that

$$\mathbf{Hx} = \frac{\alpha^2 - \mathbf{x}^T\mathbf{x}}{\mathbf{x}^T\mathbf{x} + 2\alpha x_1 + \alpha^2}\mathbf{x} - 2\alpha\frac{\mathbf{v}^T\mathbf{x}}{\mathbf{v}^T\mathbf{v}}\mathbf{e}_1$$

The coefficient of $\mathbf{x}$ has to be zero, so that $\alpha^2 = \|\mathbf{x}\|_2^2$. Therefore,

$$\mathbf{v} = \mathbf{x} \mp \|\mathbf{x}\|_2\mathbf{e}_1 \quad \Longrightarrow \quad \mathbf{Hx} = \pm\|\mathbf{x}\|_2\mathbf{e}_1$$

The sign in the expression of $\mathbf{v}$ may be determined to be the one that maximizes $\|\mathbf{v}\|_2$ for the sake of numerical stability

Householder QR

1. Let $\mathbf{H}_1 \in \mathbb{R}^{m \times m}$ be the Householder reflection w.r.t. $\mathbf{a}_1$. Transform $\mathbf{A}$ as

$$\mathbf{A}^{(1)} = \mathbf{H}_1 \mathbf{A} = \begin{bmatrix} \times & \times & \dots & \times \\ 0 & \times & \dots & \times \\ \vdots & \vdots & & \vdots \\ 0 & \times & \dots & \times \end{bmatrix}$$

$$H_1 a_1 = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

2. Let $\tilde{\mathbf{H}}_2 \in \mathbb{R}^{(m-1)\times(m-1)}$ be the Householder reflection w.r.t. $\mathbf{A}^{(1)}(2:m,2)$ (marked red above). Transform $\mathbf{A}^{(1)}$ as

$$\mathbf{A}^{(2)} = \underbrace{\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{H}}_2 \end{bmatrix}}_{=\mathbf{H}_2} \mathbf{A}^{(1)} = \begin{bmatrix} \times & \times \quad \dots \times \\ \mathbf{0} & \tilde{\mathbf{H}}_2 \mathbf{A}^{(1)}(2:m,2:n) \end{bmatrix} = \begin{bmatrix} \times & \times & \times & \dots & \times \\ 0 & \times & \times & \dots & \times \\ \vdots & 0 & \times & \dots & \times \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \times & \dots & \times \end{bmatrix}$$

3. By repeating this, $\mathbf{A}$ is transformed to $\mathbf{R}$

Householder QR (cont'd)

WLOG, assume $m \geq n$

$\mathbf{A}^{(0)} = \mathbf{A}$
for $k = 1, \ldots, n-1$
  $\mathbf{A}^{(k)} = \mathbf{H}_k \mathbf{A}^{(k-1)}$, where

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{H}}_k \end{bmatrix}$$

  $\tilde{\mathbf{H}}_k$ is the Householder reflection of $\mathbf{A}^{(k-1)}(k:m, k)$
end

$\tilde{H}_k$ orthogonal matrix

H_k

- The above procedure results in

$$\mathbf{A}^{(n-1)} = \mathbf{H}_{n-1} \cdots \mathbf{H}_2 \mathbf{H}_1 \mathbf{A}, \quad \mathbf{A}^{(n-1)} \text{ is upper triangular}$$

- QR decomposition is obtained by letting $\mathbf{R} = \mathbf{A}^{(n-1)}$ and $\mathbf{Q} = (\mathbf{H}_{n-1} \cdots \mathbf{H}_2 \mathbf{H}_1)^T$
- A widely used method for QR decomposition

$A = QR$
$= Q(H_{n-1} \cdots H_1 A)$
$Q = (H_{n-1} \cdots H_1)^{-1}$ T

Householder QR (cont'd)

$\mathbf{A}^{(0)} = \mathbf{A}$
for $k = 1, \ldots, n-1$
    $\mathbf{A}^{(k)} = \mathbf{H}_k \mathbf{A}^{(k-1)}$, where

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{H}}_k \end{bmatrix}$$

    $\tilde{\mathbf{H}}_k$ is the Householder reflection of $\mathbf{A}^{(k-1)}(k:m, k)$
end

Complexity (for $m \geq n$):

- $\mathcal{O}(n^2(m - n/3))$ for $\mathbf{R}$ only
  - A direct implementation of the above pseudo-code does not lead to this complexity–Need to exploit the structures of $\mathbf{H}_k$ in the implementations
- $\mathcal{O}(m^2 n - mn^2 + n^3/3)$ if $\mathbf{Q}$ is also wanted
- See Section 5.2.2 of textbook

Givens Rotations

Example: Let

$$\mathbf{J} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$

where $c = \cos(\theta), s = \sin(\theta)$ for some θ

$$\mathbf{y} = \mathbf{J}\mathbf{x} \Longleftrightarrow \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} cx_1 + sx_2 \\ -sx_1 + cx_2 \end{bmatrix}$$

rotate clockwise by θ

Observe that

- $\mathbf{J}$ is orthogonal
- $y_2 = 0$ if $\theta = \tan^{-1}(x_2/x_1)$, or if

$$c = \frac{x_1}{\sqrt{x_1^2 + x_2^2}}, \quad s = \frac{x_2}{\sqrt{x_1^2 + x_2^2}}$$

Givens Rotations (cont'd)

Givens rotations:

$$\mathbf{J}(i,k,\theta) = \begin{array}{cc} \begin{array}{ccccc} & i & & k & \\ & \downarrow & & \downarrow & \end{array} & \\ \begin{bmatrix} \mathbf{I} & & & & \\ & c & & s & \\ & & \mathbf{I} & & \\ & -s & & c & \\ & & & & \mathbf{I} \end{bmatrix} & \begin{array}{l} \\ \leftarrow i \\ \\ \leftarrow k \\ \\ \end{array} \end{array}$$

where $c = \cos(\theta)$, $s = \sin(\theta)$

- $\mathbf{J}(i,k,\theta)$ is orthogonal
- Let $\mathbf{y} = \mathbf{J}(i,k,\theta)\mathbf{x}$. Then,

$$y_j = \begin{cases} cx_i + sx_k, & j = i \\ -sx_i + cx_k, & j = k \\ x_j, & j \neq i, k \end{cases}$$

- y_k is forced to zero if we choose $\theta = \tan^{-1}(x_k/x_i)$

Givens QR

Example: Consider a 4×3 matrix.

$$\mathbf{A} = \begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix} \xrightarrow{\mathbf{J}_{1,2}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix} \xrightarrow{\mathbf{J}_{1,3}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ \times & \times & \times \end{bmatrix} \xrightarrow{\mathbf{J}_{1,4}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{bmatrix}$$

$$\xrightarrow{\mathbf{J}_{2,3}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & \times & \times \end{bmatrix} \xrightarrow{\mathbf{J}_{2,4}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{bmatrix} \xrightarrow{\mathbf{J}_{3,4}} \begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{R}$$

where $\mathbf{B} \xrightarrow{\mathbf{J}} \mathbf{C}$ means $\mathbf{C} = \mathbf{J}\mathbf{B}$; $\mathbf{J}_{i,k} = \mathbf{J}(i, k, \theta)$, with θ chosen to zero out the (k, i)th entry of the matrix transformed by $\mathbf{J}_{i,k}$

Givens QR (cont'd)

Givens QR ($m \geq n$): Perform a sequence of Givens rotations to annihilate the lower triangular parts of $\mathbf{A}$

$$\underbrace{(\mathbf{J}_{n,m} \dots \mathbf{J}_{n,n+2}\mathbf{J}_{n,n+1}) \dots (\mathbf{J}_{2m} \dots \mathbf{J}_{24}\mathbf{J}_{23})(\mathbf{J}_{1m} \dots \mathbf{J}_{13}\mathbf{J}_{12})}_{=\mathbf{Q}^T} \mathbf{A} = \mathbf{R}$$

where $\mathbf{R}$ is upper triangular and $\mathbf{Q}$ is orthogonal

- Complexity (for $m \geq n$): $\mathcal{O}(n^2(m - n/3))$ for $\mathbf{R}$ only (see Section 5.2.5 of textbook)
- Not as efficient as Householder QR for general (and dense) $\mathbf{A}$'s
  - The flop count for Householder QR is $2n^2(m - n/3)$
  - The flop count for Givens QR is $3n^2(m - n/3)$
- Givens QR can be faster than Householder QR if $\mathbf{A}$ has certain sparse structures and we exploit them

Matrix Computations
Chapter 3: Least-squares Problems and QR Decomposition

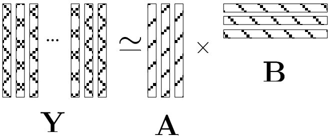
Section 3.4 Problems Related to Least Squares

Jie Lu
ShanghaiTech University

Matrix Factorization

Matrix Factorization: Given $\mathbf{Y} \in \mathbb{R}^{m \times n}$ and a positive integer $k < \min\{m, n\}$, solve

$$\min_{\mathbf{A} \in \mathbb{R}^{m \times k}, \mathbf{B} \in \mathbb{R}^{k \times n}} \|\mathbf{Y} - \mathbf{AB}\|_F^2$$



Also called low-rank matrix approximation

- $\mathrm{rank}(\mathbf{AB}) \leq k$

Principal Component Analysis

Aim: Given a collection of data points $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^m$, perform a low-dimensional representation

$$\mathbf{y}_i = \mathbf{A}\mathbf{b}_i + \mathbf{c} + \mathbf{v}_i, \quad i = 1, \ldots, n,$$

where $\mathbf{A} \in \mathbb{R}^{m \times k}$ is a basis matrix, $\mathbf{b}_i \in \mathbb{R}^k$ is the coefficient for $\mathbf{y}_i$, $\mathbf{c} \in \mathbb{R}^m$ is the base or mean in statistics terms, and $\mathbf{v}_i$ is noise or modeling error

Principal component analysis (PCA):

1. Choose $\mathbf{c} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$
2. Let $\bar{\mathbf{y}}_i = \mathbf{y}_i - \mathbf{c}$, and solve

$$\min_{\mathbf{A},\mathbf{B}} \|\bar{\mathbf{Y}} - \mathbf{A}\mathbf{B}\|_F^2$$

3. we may want a semi-orthogonal $\mathbf{A}$

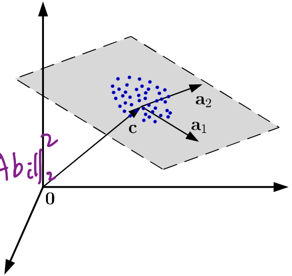
$\bar{y}_i = Ab_i + v_i$

$\downarrow$

$\min \|\bar{y}_i - Ab_i\|_2^2$

$\bar{Y} = [\bar{y}_1 \cdots \bar{y}_n]$

$B = [b_1 \cdots b_n]$



Applications: dimensionality reduction, visualization of high-dimensional data, compression, extraction of meaningful features from data, etc.

- Example of senate voting: `http://livebooklabs.com/keeppies/c5a5868ce26b8125`

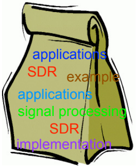
Topic Modeling

Aim: Discover thematic information or topics from a large collection of documents (e.g., books, articles, news, blogs)

Bag-of-words representation: Represent each document as a vector of word counts

… In fact, we will soon see that the implementation of SDR can be very easy, which allows signal processing practitioners to quickly test the viability of SDR in their applications. Several highly successful applications will be showcased as examples ……

a document



bag of words

$$\mathbf{y} = \begin{array}{c|l} \textbf{count} & \textbf{term} \\ \left[\begin{array}{c} 0 \\ 2 \\ 2 \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{array}\right] & \begin{array}{l} \text{efficiency} \\ \text{applications} \\ \text{SDR} \\ \text{communications} \\ \text{example} \\ \text{signal processing} \\ \vdots \\ \text{implementation} \end{array} \end{array}$$

bag–of–words representation

Topic Modeling (cont'd)

- Let n be the number of documents

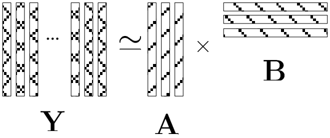
- Let $\mathbf{y}_i \in \mathbb{R}^m$ be the bag-of-words representation of the ith document
- $\mathbf{Y} = [\ \mathbf{y}_1, \ldots \mathbf{y}_n\ ] \in \mathbb{R}^{m \times n}$ is called the term-document matrix

- Hypotheses:[1]
  - If documents have similar columns vectors in $\mathbf{Y}$ or similar usage of words, they tend to have similar meanings
  - The topic of a document will probabilistically influence the author's choice of words when writing the document

[1] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of Artificial Intelligence Research*, vol. 37, pp. 141–188, 2010.

Topic Modeling (cont'd)

Problem: Apply matrix factorization to a term-document matrix $\mathbf{Y}$



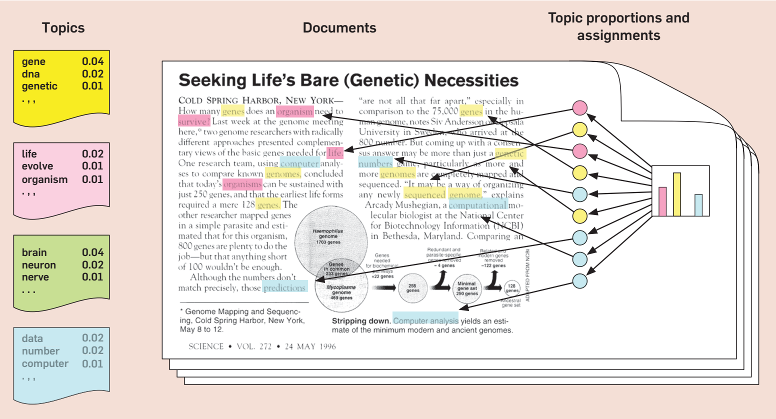
$\mathbf{A}$ is called a term-topic matrix and $\mathbf{B}$ is called a topic-document matrix

Interpretation:

- Each column $\mathbf{a}_i$ of $\mathbf{A}$ represents a theme topic (e.g., local affairs, foreign affairs, politics, sports)
- $\mathbf{y}_i \approx \mathbf{A}\mathbf{b}_i$: each document is postulated as a linear combination of topics
- Matrix factorization aims at discovering topics from the documents

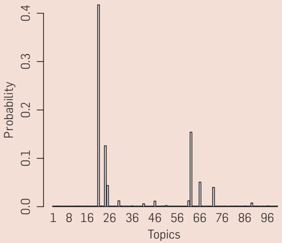
Topic modeling via matrix factorization has been used in or is tightly connected to information retrieval, natural language processing, machine learning; document clustering, classification and retrieval; latent semantic analysis, latent semantic indexing: finding similarities of documents, similarities of terms, etc.

Topic Modeling (cont'd)



Source: D. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

Topic Modeling (cont'd)



"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Topics found in a real set of documents. The document set consists of 17,000 articles from the journal *Science*. The topics are discovered using a technique called *latent Dirichlet allocation*, which is not the same as, but has strong connections to, matrix factorization [Blei'12]

Matrix Factorization

Problem:

$$\min_{\mathbf{A}\in\mathbb{R}^{m\times k},\mathbf{B}\in\mathbb{R}^{k\times n}} \|\mathbf{Y}-\mathbf{AB}\|_F^2$$

The problem has non-unique solutions

- If $(\mathbf{A}^\star, \mathbf{B}^\star)$ is an optimal solution to the problem, then $(\mathbf{A}^\star\mathbf{Q}^{-1}, \mathbf{Q}\mathbf{B}^\star)$ is also an optimal solution for any nonsingular $\mathbf{Q} \in \mathbb{R}^{k\times k}$
- The non-uniqueness of solution makes it a bad formulation for problems such as topic modeling

The problem is non-convex, but can be solved by singular value decomposition (beautifully)

It can also be solved by LS approach

Alternating LS for Matrix Factorization

Alternating LS (ALS): Given a starting point $(\mathbf{A}^{(0)}, \mathbf{B}^{(0)})$, do

$$\mathbf{A}^{(i+1)} = \arg\min_{\mathbf{A}\in\mathbb{R}^{m\times k}} \|\mathbf{Y} - \mathbf{A}\mathbf{B}^{(i)}\|_F^2 \qquad (1)$$

$$\mathbf{B}^{(i+1)} = \arg\min_{\mathbf{B}\in\mathbb{R}^{k\times n}} \|\mathbf{Y} - \mathbf{A}^{(i+1)}\mathbf{B}\|_F^2 \qquad (2)$$

for $i = 0, 1, 2, \ldots$, and stop when a termination criterion is satisfied

Make a mild assumption that $\mathbf{A}^{(i)}, \mathbf{B}^{(i)}$ have full rank at every i

Look at (2) first.

$$\min_{B\in\mathbb{R}^{k\times n}} \|Y - A^{(i+1)}B\|_F^2 = \min_{b_1,\cdots,b_n\in\mathbb{R}^k} \sum_{j=1}^{n} \|y_j - A^{(i+1)}b_j\|_2^2$$

$$\Leftrightarrow \min_{b_j\in\mathbb{R}^k} \|y_j - A^{(i+1)}b_j\|_2^2, \quad \forall j = 1, \cdots, n$$

LS

$$b_j^{(i+1)} = \left[(A^{(i+1)})^T A^{(i+1)}\right]^{-1} (A^{(i+1)})^T y_j$$

$$B^{(i+1)} = \left[(A^{(i+1)})^T A^{(i+1)}\right]^{-1} (A^{(i+1)})^T Y$$