## I.9 Principal components and the best low rank approximation

**Definition** A norm $\|\cdot\|$ on a vector space $V$ is a function such that

(1) $\|v\| \ge 0$ for all $v \in V$, where equality holds if and only if $v = 0$.

(2) $\|u+v\| \le \|u\| + \|v\|$ for all $u, v \in V$.

(3) $\|\mu v\| = |\mu|\|v\|$ for any scalar $\mu$ and $v \in V$.

**Example** For $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$, $\|x\| = (x^T x)^{1/2}$, $\|x\|_{\sup} = \max\{|x_j| : 1 \le j \le n\}$, $\|x\|_{\ell_1} = |x_1| + \cdots + |x_n|$ are norm functions.

**Definition** A norm on $M_{m,n}$ is unitarily invariant if a norm on $M_{m,n}$ satisfies $\|UAV\| = \|A\|$ for any $A \in M_{m,n}$, unitary $U \in M_m$ and $V \in M_n$. In other words, the norm depend on the singular values of $A$ only.

**Examples** Let $A = (a_{ij}) = \sum_{j=1}^r s_j u_j v_j^* \in M_{m,n}$

Spectral norm: $\|A\|_2 = \max_{x \ne 0} \|Ax\|/\|x\| = s_1(A)$.

Frobenius norm: $\|A\|_F = (\sum_{i,j} |a_{ij}|^2)^{1/2} = (\mathrm{tr}\, A^*A)^{1/2} = (\sum_{j=1}^j s_1^2 + \cdots + s_r^2)^{1/2}$.

Nuclear norm or trace norm: $\|A\|_N = \sum_{j=1}^r s_j(A)$.

**Theorem** Suppose $\|\cdot\|$ is a unitarily invariant norm on $M_{m,n}$. Let $A = \sum_{j=1}^r s_j u_j v_j^*$. If $B = \sum_{j=1}^k s_j u_j v_j^*$, where $s_j = 0$ if $j > r$, then $\|A - B\| \le \|A - X\|$ for any $X \in M_{m,n}$ of rank at most $k$.

*Proof for the spectral norm.* We may focus on the case when $k < r$. Else, we can let $B = A$ be the optimizer. Suppose $B \in M_{m,n}$ has rank at most $k$. Then there is a unit vector $x = (a_1, \dots, a_k)^t$ such that $B[v_1|\cdots|v_{k+1}]x = 0$ because $B[v_1|\dots|v_{k+1}]$ has rank at most $k$. Then $\|x\| = 1$, $\|y\| = 1$ for $y = [v_1|\cdots|v_{k+1}]x$ and

$$\begin{aligned}
\|A - B\|_2 &\ge \|(A-B)y\| = \|A(a_1 v_1, \dots, a_{k+1} v_{k+1})\| \\
&= \|\sum_{j=1}^{k+1} s_j a_j u_j\| = \{\sum_{j=1}^{k+1} (s_j^2 |a_j|^2)\}^{1/2} \\
&\ge s_{j+1} = \|A - \sum_{j=1}^{k} s_j u_j v_j^*\|_s.
\end{aligned}$$

Here, we get $\|A - B\| = s_{k+1}$ if $B = \sum_{j=1}^k s_j u_j v_j^*$. $\square$

*Proof for the Frobenius norm.* Suppose $B$ has rank at most $k$ such that

$$\|A-B\|_F \le \|A-X\|_F \quad \text{for any } X \in M_{m,n} \text{ of rank at most } k.$$

Then there are $U \in U_m, V \in U_n$ and $D = \operatorname{diag}(d_1,\dots,d_k) \in M_k$ with $d_1 \ge \cdots \ge d_k \ge 0$ such that $UBV = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$, and $\|A-B\|_F = \|UAV - UBV\|_F$.

Let $UAV = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$. Then $B_1 = U^* \begin{pmatrix} A_{11} & A_{12} \\ 0 & 0 \end{pmatrix} V^* \in M_{m,n}$ rank at most $k$ so that

$$\|A_{11}\|_F^2 + \|A_{12}\|_F^2 + \|A_{21}\|_F^2 + \|A_{22}\|_F^2 = \|A-B\|_F^2 \le \|A-B_1\|_F^2 = \|A_{21}\|_F^2 + \|A_{22}\|_F^2.$$

Thus, $A_{11}$ and $A_{12}$ are zero matrices. Similarly, $B_2 = U^* \begin{pmatrix} A_{11} & 0 \\ A_{21} & 0 \end{pmatrix} V^* \in M_{m,n}$ has rank at most $k$ so that

$$\|A_{11}\|_F^2 + \|A_{12}\|_F^2 + \|A_{21}\|_F^2 + \|A_{22}\|_F^2 = \|A-B\|_F^2 \le \|A-B_1\|_F^2 = \|A_{12}\|_F^2 + \|A_{22}\|_F^2.$$

So, $A_{21}$ is also a zero matrix. Hence, $UAV = \operatorname{diag}(d_1,\dots,d_k) \oplus A_{22}$, where the list $d_1,\dots,d_k$ include at most $k$ nonzero singular values of $A$, and $A_{22}$ has the remaining nonzero singular values $\sigma_1,\dots,\sigma_\ell$, where $\ell \ge r-k$. Hence,

$$\|A-B\|_F^2 = \sum_{j=1}^{\ell} \sigma_j^2 \ge \sum_{j=k+1}^{r} s_j^2,$$

where the equality holds when $B = \sum_{j=1}^k s_j u_j v_j^*$. □

**Principal Component Analysis**

- Let $A$ be $m \times n$, each column is a sample with $m$ measurements.
- Normalize the means to 0 for all measurements. So, sum of columns is the zero vector in $\mathbb{R}^m$.
- The variances are the diagonal entries of $AA^T \in M_m$.
- The co-variances are the off-diagonal entries of $AA^T \in M_n$.
- The sample co-variance matrix is $S = AA^T/(n-1)$.

For example, when $m = 2$, the best rank one approximation of $S$ is $s_1 u_1 v_1^t$. So, the slop of the line is the ratio of the second entry of $u_1$ to the first entry of $u_1$.
If the rank of $AA^T$ is low, then the two measurement is closely related, i.e., almost agree on a linear relation.
We can extend the idea to higher dimension data set recorded as $A \in M_{m,n}$. One can use $k$-dimensional hyperplane to approximate the data.

- The total variance is

$$T = \operatorname{tr} A^T A/(n-1) = \operatorname{tr} S/(n-1) = (\sum_{j=1}^{r} s_j^2)/(n-1),$$

  where $s_1 \geq \cdots \geq s_r > 0$ are the singular values of $A$.
- The first $k$ singular vectors capture more information than other vectors; $u_j$ is referred to as the $j$th principal component of the data that accounts for the fraction $\sigma_j^2/T$ of the variance.
- The effective rank $k$ of $A$ or $S$ is the number of singular values larger than certain threshold so that the other part in the singular value decomposition is considered as noise in the data.

- Note that the line is different from finding the best fit $y = ax + b$. In that case, we want to find best $a, b$ such that $ax_i + b = y_i$ for $i = 1, \ldots, n$ without centering the data. We consider $\tilde{A}(a, b)^T = (y_1, \ldots, y_n)^T$ and find the least square solution:

$$\tilde{A}^T \tilde{A}(a, b)^T = \tilde{A}^T (y_1, \ldots, y_n)^T.$$

  This is known as standard least square.

- In our case, we consider the centered data, and

$$\|A^T\|_F^2 = \|A^T u_1\|_F^2 + \cdots + \|A^T u_m\|_F^2$$

  so that

  **the sum of squared distances from the data points to $u_1, \ldots, u_k$ is a minimum**.

There are interesting discussion of the Hilbert matrix

$$H = [a_{ij}] = [1/(i + j - 1)]$$

and the zero-one matrix representing the picture of square, triangle, circe, etc. See pp. 78-79.

### 1.10 Rayleigh Quotients and Generalized Eigenvalues

Let $S = S^* \in M_n$ be complex Hermitian or real symmetric. For a nonzero $x \in \mathbb{F}^n$, define the

$$\textbf{Rayleigh quotient} \qquad \frac{x^*Sx}{x^*x} = \frac{x^*}{\|x\|} S \frac{x}{\|x\|},$$

which is a function of $n$ variables $x = (x_1, \ldots, x_n)^t$.

1. The maximum value is attained at an eigenvector of the largest eigenvalue of $S$.

2. The minimum value is attained at an eigenvector of the smallest eigenvaleu of $S$.

3. All other eigenvalues are saddle points.

*Proof.* (1) and (2): Let $S = S^*$. Then

(3): In the real case, if $Sx = \lambda_j x$, then $x^T Sx = \lambda_j$ so that $\frac{\partial R}{\partial x_j} = 0$ and the directional derivatives in the direction $x + ty$ may increase or decrease depending on $y$ is the eigenvector of $\lambda_1$ or $\lambda_n$.

**Generalized eigenvalues and eigenvectors**

In applications such as statistics, engineering, etc., one considers the

$$\textbf{Generalized Rayleigh Quotient} \qquad \frac{x^*Sx}{x^*Mx},$$

for a given $S^* = S$ and $M = M^*$ and a nonzero $x \in \mathbb{F}^n$.

**Remark** We usually assume that $M$ is positive definite, i.e., $x^*Mx > 0$ for all nonzero $x \in \mathbb{F}^n$. Else, we may have $x^*Mx = 0$, and $R(x)$ is not well defined.

So, to find $\max R(x)$ is the equivalent to finding

$$\begin{aligned}
& \max\{x^*Sx : x^*Mx = 1\} \\
= \; & \max\{y^*M^{-1/2}SM^{-1/2}y : y = M^{1/2}x, y^*y = 1\} \\
= \; & \max\{y^*M^{-1/2}SM^{-1/2}y : y^*y = 1\} \\
= \; & \lambda_1(M^{-1/2}SM^{-1/2}) \\
= \; & \lambda_1(M^{-1}H).
\end{aligned}$$

So, we can solve $M^{-1/2}SM^{-1/2}x = \lambda x$, or we can solve $M^{-1}Sx = \lambda x$, i.e., $Sx = \lambda Mx$.

Note that

(a) $\det(M^{-1/2}SM^{-1/2} - \lambda I) = 0$ and $\det(S - \lambda M) = 0$ have the same roots.

(b) $M^{-1/2}SM^{-1/2}x = \lambda x$ if and only if $Sy = \lambda My$ for $y = M^{-1/2}x$.

(c) If $Hy_1 = \lambda_1 My_1, Hy_2 = \lambda_2 My_2$ for $\lambda_1 \neq \lambda_2$, then $y_1^*My_2 = 0$, i.e., $y_1, y_2$ are $M$-orthogonal.

**Example** (p.83-84) Consider

$$R(x) = \frac{x^T S x}{x^T M x}, \qquad x \neq 0,$$

with $S = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix}$, $M = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.

We may consider $M^{-1/2} S M^{-1/2}$, and compute

$$\det(M^{-1/2} S M^{-1/2} - \lambda I) = \det \begin{pmatrix} 4-\lambda & -\sqrt{2} \\ -\sqrt{2} & 2-\lambda \end{pmatrix} = 0.$$

It is better to compute

$$\det(S - \lambda M) = \det \begin{pmatrix} 4-\lambda & -2 \\ -2 & 4-2\lambda \end{pmatrix} = 0.$$

Eigenvalues are: $3 \pm \sqrt{3}$.

For $\lambda_1 = 3 + \sqrt{3}$, we solve

$$(S - \lambda_1 M) y_1 = 0, \; y_1 = c_1 \begin{bmatrix} 2 \\ 1+\sqrt{3} \end{bmatrix}.$$

For $\lambda_2 = 3 - \sqrt{3}$, we solve

$$(S - \lambda_2 M) y_2 = 0, \; y_2 = c_2 \begin{bmatrix} 2 \\ 1-\sqrt{3} \end{bmatrix}.$$

Note that $y_1^T y_2 = c_1 c_2 2 \neq 0$, and $y_1^T M y_2 = 0$.

**An application**

**Separating vector**

In machine learning/statistics, we want to do the following.

Given two groups of data (points) in $v_i \in \mathbb{R}^2$, we want to find a separating vector $v \in \mathbb{R}^2$ so that $v^T x_i > c$ will be group I data, $v^T x_i < c$ will be group II data.

We can apply the same idea to data in $\mathbb{R}^n$.

Let $m_1, m_2 \in \mathbb{R}^n$ and $\Sigma_1, \Sigma_2$ be the means values of the data and the co-variance matrices of the two populations. We want to maximize the separating ratio:

$$\begin{aligned} R &= \frac{(x^T m_1 - x^T m_2)^2}{x^T \Sigma_1 x + x^T \Sigma_2 x} \\ &= \frac{x^T (m_1 - m_2)(m_1 - m_2)^T x}{x^T (\Sigma_1 + \Sigma_2) x}. \end{aligned}$$

So, the optimal separating vector is

$v = M^{-1}(m_1 - m_2)$ with $S = \Sigma_1 + \Sigma_2$.

(Next homework!)

**Generalized eigenvalue equal to $\infty$**

If $M$ is singular, we may consider $x \neq 0$ such that $\alpha Hx = \beta Mx$ and write $\lambda = \beta/\alpha$.

If $\alpha \neq 0$, then $\lambda = \beta/\alpha$.

If $Mx = 0 \neq Sx$, we set $\alpha = 0 \neq \beta$, then $\lambda = \beta/\alpha = \infty$.

If both $Mx = 0 = Sx$, then $\lambda$ is undertermined.

**Generalized SVD**

**Theorem** Assume $A \in M_{m_1,n}, B \in M_{m_2,n}$ such that $T = \begin{bmatrix} A \\ B \end{bmatrix}$ has rank $n$. Then there exist unitary $U_1 \in M_{m_1}, U_2 \in M_{m_2}$, and an invertible $Z \in M_n$ such that $A = U_1\Sigma_1 Z, B = U_2\Sigma_2 Z$ such that $\Sigma_1, \sigma_2$ only have nonzero (nonnegative) entries at the diagonal positions satisfying

$$\Sigma_1^*\Sigma_1 + \Sigma_2^*\Sigma_2 = I_n.$$

*Proof.* Find SVD of $T$ so that

$$T = W\Sigma V^*,$$

where $\Sigma, V \in M_n$ is unitary and $W = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \in M_{m,n}$ has orthonormal columns with $m = m_1 + m_2$.

Let $W_1 = U_1\Sigma_1 W_0$, where $U_1 \in M_{m_1}$, $W_0 \in M_n$ are unitary, and $\Sigma_1 \in M_{m_1,n}$.

Now, $\begin{bmatrix} U_1^* & \\ & I_{m_2} \end{bmatrix} WW_0^* = \begin{bmatrix} \Sigma_1 \\ W_2W_0^* \end{bmatrix}$ has orthogonal columns.

So, $W_2W_0$ has orthogoanl columns and equals $U_2\Sigma_2$, where $U_2 \in M_{m_2}$ is unitary, $\Sigma_2 \in M_{m_2,n}$. If $Z = W_0^*\Sigma V^*$, then

$$T = \begin{bmatrix} U_1\Sigma_1 & \\ & U_2\Sigma_2 \end{bmatrix} Z.$$

The result follows. □

**Remark** In the book and next homework, we use the idea of congruence. We say that $A, B \in M_n$ are congruent if there is an invertible $Z \in M_n$ such that $A = Z^T BZ$.

## I.11 Norms of vectors and functions of matrices

A norm $|\cdot|$ on $\mathbb{F}^n$ is a real-valued function satisfying

(a) $|x| \ge 0$ for all $x \in \mathbb{F}^n$,

where the quality holds if and only if $x = 0 \in \mathbb{F}^n$.

(b) $|cx| = |c| \cdot |x|$ for any $c \in \mathbb{F}$ and $x \in \mathbb{F}^n$.

(c) $|x + y| \le |x| + |y|$ for all $x, y \in \mathbb{F}^n$.

**Example** Let $p \ge 1$. For $x = (x_1, \dots, x_n)^T \in \mathbb{F}^n$, define

$$\ell^p(x) = (\sum_{j=1} |x_j|^p)^{1/p}.$$

Also, define

$$\ell^\infty(x) = \max\{|x_j| : 1 \le j \le n\}.$$

Consider the geometry of $\ell^p$, including $p \in (0, 1)$.

See the pictures of the set of vectors in $\mathbb{R}^2$ with norm 1 for $\ell^2, \ell^1, \ell^\infty$.

**In applications**, one often consider in $\mathbb{R}^2$,

$$\min |v| \qquad \text{subject to} \quad a_1 v_1 + a_2 v_2 = 1.$$

Then the $\ell^1$ and $\ell^\infty$ norms often produce "sparse" solutions.

**Example** $\max |v|$ subject to $3v_1 + 4v_2 = 1$.

This leads to the idea of consider "$\ell^0$ norm", and sparse solution of $Ax = b$ by studying

$$\min \ell^0(x) \qquad \text{subject to } Ax - b = 0.$$

In practice, we consider

$$\min \ell^1(x) \qquad \text{subject to } Ax - b = 0.$$

This technique is known as compressed sensing.

**Inner products, norms, and angles**

Recall: the inner product of $x, y \in \mathbb{F}^n$ is defined by

$$(x, y) = y^* x.$$

The $\ell^2$-norm of $x \in \mathbb{F}^n$ is

$$\|x\| = \ell^2(x) = (x, x)^{1/2} = (x^* x)^{1/2}.$$

We have the Cauchy-Schwartz inequality:

$$|(x, y)| \leq \|x\| \|y\|.$$

*Proof.* Let $e^{i\theta}$ be such that $e^{-i\theta}(x, y) = (x, e^{i\theta} y) = |(x, y)|$. Then for any $t \in \mathbb{R}$,

$$\begin{aligned} 0 &\leq (x + te^{i\theta} y, x + te^{i\theta} y) \\ &= t^2 (y, y) + t(x, e^{i\theta} y) + t(e^{i\theta} y, x) + (x, x) \\ &= (y, y) t^2 + 2t|(x, y)| + (x, x). \end{aligned}$$

For a quadratic equation that $at^2 + bt + c = 0$ has no distinct real roots, we have $b^2 \leq 4ac$. Thus,

$$(x, x)(y, y) \geq |(x, y)|^2. \qquad \square$$

For real vectors $x, y \in \mathbb{R}^n$,

$$(x, y) = \|x\| \|y\| \cos \theta.$$

**$S$-norm**
Sometimes, certain coordinates of the vectors carry more weights. A norm will be defined accordingly.

**For example,**

$$\|(x_1, x_2)^T\|^2 = \frac{1}{a^2}|x_1|^2 + \frac{1}{b^2}|x_2|^2 = (x_1, x_2)D(x_1, x_2)^T$$

for some $a \geq b > 0$ and $D = \text{diag}\,(1/a^2, 1/b^2)$.
The set of norm one vectors is the ellipse

$$\mathcal{E} = \{(x_1, x_2)^T : x^T DX = 1\}.$$

More generally, if $S$ is a positive definite matrix, we can define the $S$-norm on $\mathbb{R}^n$ by

$$\|v\|_S^2 = v^T Sv$$

corresponding to the $S$-inner product

$$(v, w)_S = v^T Sw.$$

**General Banach spaces**

One may consider general vector spaces with norms.

One can define norms on infinite dimensional vector spaces with some care.

**Example** Consider the sequence space / infinite vectors

$$\{(a_1, a_2, \dots) : a_1, a_2, \dots \in \mathbb{F}\}.$$

- One may define $\ell^1, \ell^2, \ell^\infty$ norms with convergence.
- We have $\ell^1 \subsetneq \ell^2 \subsetneq \ell^\infty$.
- Consider the subset of all sequences with finitely many nonzero terms.
- Then all the norms are well-defined. But the limit of the sequence $\{v_1, v_2, \dots\}$ with

  $v_n = (1, 1/2, 1/3, \dots, 1/n, 0, \dots, 0)$ is not in the set.

A linear space with a norm so that every sequence of vector has a limit in the space is a Banach space.

One may consider the space of functions.

- We can consider continuous function $f : [0,1] \to \mathbb{R}$.

  Then $L^1[0,1]$ is a Banach space with

  $$\|f\|_1 = \int_0^1 |f(x)| dx.$$

  $L^2[0,1]$ is an inner product space with

  $$(f, g) = \int_0^1 f(x)g(x) dx.$$

  $L^\infty[0,1]$ is a Banach space with

  $$\|f\|_\infty = \sup\{|f(x)| : x \in [0,1]\}.$$

  If the function $f$ is continuously differentiable on $[0,1]$, one may define

  $$\|f\| = \|f\|_C + \|\frac{d}{dx} f\|_C.$$

**Norms on matrices**

- One may define norms on matrices.
- Frobenius norm $\|\cdot\|_F$.
- Trace norm (nuclear norm) $\|A\|_{\mathrm{t}r}$ .
- Operator norm $\|A\|_2$.
- A norm $\|\cdot\|$ on $M_n$ is a matrix norm if $\|AB\| \le \|A\|\|B\|$.

**Theorem** Let $A, B \in M_n$. Then

$$\|AB\|_F \le \|A\|_F\|B\|_F.$$

The equality holds if and only if $A = xy^*$ and $B = yz^*$ for some $x, y, z \in \mathbb{F}^n$.

Proof. Suppose $A = U\Sigma V^*$ with unitary $U, V \in M_n$ and $\Sigma = \mathrm{diag}\,(a_1, \dots, a_n)$. If $V^*BU = (b_{ij})$, then

$$\begin{aligned}
\|AB\|_F^2 &= \|U\Sigma V^*B\|_F^2 = \|\Sigma V^*BU\|_F^2 \\
&= a_1^2 \sum_{j=1}^n |b_{1j}|^2 + \cdots + a_n^2 \sum_{j=1}^n |b_{nj}|^2 \\
&\le a_1^2\|B\|_F^2 \le \|A\|_F^2\|B\|_F^2.
\end{aligned}$$

The equality holds if and only if $a_1 = \|A\|_F$ and only the first row of $V^*BU$ is nonzero. Equivalently, $A = xy^*$ and $B = yz^*$. (Verify!) □

**Induced norms**

Let $|\cdot|$ be a norm on $\mathbb{F}^n$. Define the induced norm $\|\cdot\|$ on $M_n$ corresponding to $|\cdot|$ by

$$\|A\| = \sup\{|Ax| : x \in \mathbb{F}^n, |x| \leq 1\}.$$

- Induced norms are always multiplicative.

- The spectral norm $\|A\|_2$ is the induced norm corresponding to $\ell^2$, i.e., $\|A\|_2 = \max\{\ell^2(Ax) : \ell^2(x) \leq 1\}$.

- The row sum norm is induced by $\ell^\infty$ and equals

$$\|A\|_\infty = \max\{\sum_{j=1}^{n} |a_{ij}| : 1 \leq i \leq n\}.$$

  Proof. Note that $\|A\|_\infty = \ell^\infty(Av)$ is attained at $v = (e^{i\theta_1}, \ldots, e^{i\theta_n})^T$ such that ....

- The column sum norm is induced by $\ell^1$ and equals

$$\|A\|_1 = \max\{\sum_{i=1}^{n} |a_{ij}| : 1 \leq j \leq n\}.$$

  Proof. Note $Av = \sum_{j=1}^n v_j A_j \leq \sum_{j=1}^n |v_j| \ell^1(A_j)$ if $A = [A_1 \cdots A_n]$. Choose $v = e_j$ corresponding to the column with maximum column sum.

- $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$.
  Proof. Let $v$ with $\ell^\infty(v) = 1$ be such that

$$\|A\|_2^2 = \ell^2(A^T Av) \leq \|A^T\|_\infty \ell^\infty(Av) \leq \|A\|_1 \|A\|_\infty.$$

- The spectral radius of $A \in M_n$ defined by

$$\lambda_{\max}(A) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}$$

  is not a norm.

**Theorem** Let $A \in M_n$. Then $A^k \to 0$ if and only if $\lambda_{\max}(A) < 1$.
Proof. If $\lambda_{\max}(A) \geq 1$, then $Av = \lambda v$ with $|\lambda| \geq 1$. So, $A^k v = \lambda^k v \not\to 0$.
If $\lambda_{\max}(A) < 1$ and $A$ distinct eigenvalues, then $A^k = SD^kS^{-1} \to S0S^{-1} = 0$.
In general, there is unitary $U$ such that $U^*AU = T$ is in upper triangular form (by an induction argument).
Then we can find $D_m$ with all diagonal entries less than $1/m$ such that $T_m = T + D_m$ has distinct eigenvalues.
Hence, $A_m = UT_mU^*$ has distinct eigenvalues all have size less than 1 so that $A_m^k \to 0$.
Now, $\{A_1^k, A_2^k, \ldots\} \to A^k$. Thus, $A^k \to 0$.

**Factoring matrices and tensors**

**1. (Sparse) Nonnegative matrix factorization**

- Consider $\min \|A = UV\|_F^2$, $U, V$ are nonnegative.
- Preferably, $UV$ are low rank.
- Even if $A$ is symmetric with nonnegative entries, we may not be able to write $A = B^T B$ for some $B \geq 0$.
- In general, even if $A$ is doubly nonnegative, i.e., positive semidefinite with nonnegative entries, it may not be possible to write $A = B^T B$ for $B \geq 0$.
- Matrices of the form $B^T B$ for a nonnegative matrix $B$ is called completely positive.
- The factorization $A = BC$ may not be unique.
- We aim at $\min \|A = BC\|_F^2$, $B, C \geq 0$ (and sparse).
- In application of facial recognition, we assume columns of $A$ are facial features of people.
- Then $A = BC$ means that we choose a few "typical" faces, the columns of $B$, "generating" all other faces.
- In application of indexing of documents, the columns of $A$ are the key words in a document on a topic.
- Then $A = BC$ means that we choose a collections of topics, columns of $B$, that "generate" all other topics.

**Alternating factorization method:**

- Start with some nonnegative $U_1$, find a nonnegative $V_1$, say, column by column, to minimize $\|A - U_1 V_1\|_F$.
- Use the matrix $V_1$ to find a nonnegative $U_2$, say, row by row, to minimize $\|A - U_2 V_1\|_F$.
- Repeat the process to get $U_N, V_N$ for (large) $N$ so that the improvement is small for the next iterate.
- Some practice in the next homework.

**Sparsity**

- For sparsity, we consider $\min \|A - UV\|_F$ for $U, V \geq 0$,
  $Y = UVV^T - AV^T \geq 0$ with $Y_{ij}$ or $U_{ij} = 0$ for all $i, j$,
  $Z = U^T UV - U^T A \geq 0$ with $Z_{ij}$ or $V_{ij} = 0$ for all $i, j$.
- We can also use the ideas in compressed sensing to find sparse solution for

  LASSO $\quad \min(\|Ax - b\| + \lambda \ell^1(x))$,

  or $\quad \min(\|Ax - b\| + \lambda \ell^1(x)) + \beta \ell^2(x)$.

**CP Tensor decomposition**

- One can arrange data in vectors, matrices, or tensors.
  $$x \in \mathbb{R}^n, \; A \in \mathbb{R}^3, \; T \in \mathbb{R}^{3\times 4\times 2}.$$
- For a black and white picture, each pixel (entry of the $m\times n$ matrix) has a black and white level from 0 (black) to 255 (white), $256 = 2^8$ levels.

  For a color picture, we need the three colors - red, green, blue, each has a certain intensity scale. So, we need to store the data as a $m \times n \times 3$ tensor.
- In machine learning, we often consider $w = Av$ with $A \in \mathbb{R}^{m\times n}$ and $v \in \mathbb{R}^n$ so that
  $$T_{ijk} = \frac{\partial w_i}{\partial A_{jk}} = v_k \delta_{ij}.$$
- A rank one tensor $T \in \mathbb{R}^{m\times n\times p}$ has the form
  $$T = T_{ijk} = \mathbf{a} \circ \mathbf{b} \circ \mathbf{c} = (a_i b_j c_k),$$
  $\mathbf{a} = (a_1, \dots, a_m)^T, \mathbf{b} = (b_1, \dots, b_n)^T, \mathbf{c} = (c_1, \dots, c_p)^t$.
- It is challenging problem to find the tensor rank $R$ of $T = (T_{ijk})$, i.e., the smallest $R$ such that
  $$T = \sum_{j=1}^{R} \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i$$
- In practice, we consider $\min \|A - \sum_{j=1}^{R} \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i\|$.