DATA ANALYSIS

JR, 2018/03/07
EE16B, Spring 2018

THE SINGULAR VALUE DECOMPOSITION (SVD)

→ ANY $n \times m$ real matrix A can be decomposed as:

$$A = U \Sigma V^T \quad , \quad \text{(called the SVD of } A\text{)} \tag{1}$$

(A: $n \times m$; U: $n \times n$; Σ: $n \times m$ diagonal; V^T: $m \times m$)

→ where $UU^T = U^TU = I$, $VV^T = V^TV = I$ (such matrices are called UNITARY) (2)

→ and Σ is diagonal – it looks like this: $n \times m$ $\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{bmatrix}$ or $n \times m$ $\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ & & \end{bmatrix}$ (3)

→ and moreover: $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_m \geq 0$ are all real and positive. (4)

→ $\{\sigma_i\}$ are called the SINGULAR VALUES of A.

– NOTE that the SVD is NOT the same as eigendecomposition.

– (why not? answer this for yourself before moving on)

– Why is the SVD useful? How can we calculate the SVD?

– We'll look into these questions below, in due course.

– First, we need to establish some mathematical preliminaries:

– unitary matrices:

$$\underbrace{\begin{bmatrix} \leftarrow \vec{u}_1^T \rightarrow \\ \leftarrow \vec{u}_2^T \rightarrow \\ \vdots \\ \leftarrow \vec{u}_n^T \rightarrow \end{bmatrix}}_{U^T} \underbrace{\begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{u}_1 & \vec{u}_2 & \cdots & \vec{u}_m \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}}_{U} = \underbrace{\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}}_{I} \tag{5}$$

– From the above, you can see that:

→ $\vec{u}_1^T \vec{u}_1 = 1 \Rightarrow \|\vec{u}_1\| = 1$; $\vec{u}_2^T \vec{u}_2 = 1$, etc, ..., $\vec{u}_m^T \vec{u}_m = 1 \Rightarrow \|\vec{u}_i\| = 1$ (6)

→ also $\vec{u}_1^T \vec{u}_2 = 0$, $\vec{u}_1^T \vec{u}_3 = 0$, $\vec{u}_2^T \vec{u}_3 = 0$, etc $\Rightarrow$ $\vec{u}_i^T \vec{u}_j = 0$ if $i \neq j$ (7)

→ any set of vectors $\{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_n\}$ satisfying (6) & (7) is called **ORTHONORMAL**.

→ if $\|\vec{u}_i\| \neq 1$, (necessarily) then they are called **ORTHOGONAL**

↳ $\vec{u}_i$ should not be $\vec{0}$, of course.

→ An important property of orthogonal sets of vectors:

→ any set of orthogonal vectors is linearly independent. (8)

→ proof: given an orthogonal set of vectors $\{\vec{p}_1, \dots, \vec{p}_n\}$

→ suppose they are linearly dependent: i.e. ∃ (there exists) same $\alpha_1, \dots, \alpha_n$, not all zero, s.t.

$$\alpha_1 \vec{p}_1 + \alpha_2 \vec{p}_2 + \cdots + \alpha_n \vec{p}_n = \vec{0} \quad (9)$$

→ pre-multiply (10) by $\vec{p}_1^T$; then we get

$\alpha_1 \|\vec{p}_1\| = 0$ ($\vec{p}_1^T \vec{p}_2$, $\vec{p}_1^T \vec{p}_3$, etc. are all 0 due to orthogonality)

→ since $\|\vec{p}_1\| \neq 0$, $\alpha_1 = 0$.

→ similarly, you can pre-multiply by $\vec{p}_2^T, \vec{p}_3^T, \dots, \vec{p}_n^T$, to get $\alpha_2 = \alpha_3 = \alpha_4 \cdots = \alpha_n = 0$

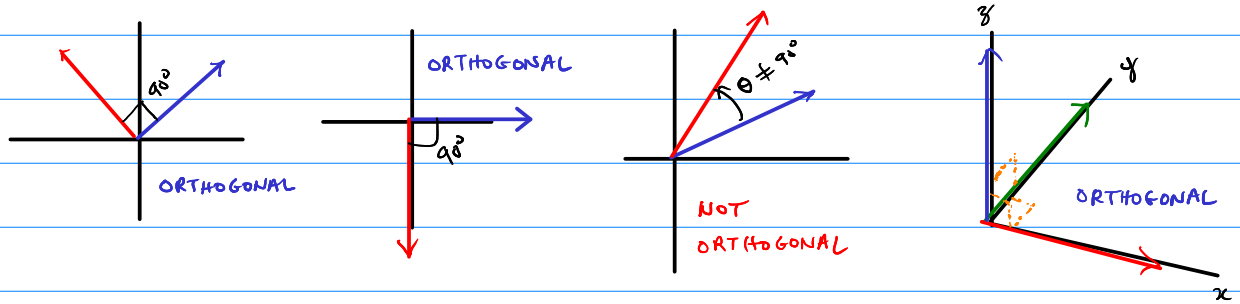
→ hence we have a contradiction to our assumption of linear dependence.

→ Due to (8), <u>any</u> set of orthogonal vectors $\{\vec{p}_1, \dots, \vec{p}_n\}$, each in $\mathbb{R}^n$, constitutes a <u>basis</u> for $\mathbb{R}^n$. (10)

↳ i.e., you can express any $\vec{x} \in \mathbb{R}^n$ as a linear combination of $\vec{p}_1, \dots, \vec{p}_n$.

— Geometrical interpretation of orthogonal vectors:

→ in 2D & 3D, orthogonal vectors are at right angles to each other:



(11)

→ in 4D and higher, also the same, though difficult to visualize for most.

→ One reason orthonormal bases are nice is that it is easy to express vectors as a linear combination of such bases.

→ Suppose you have a vector $\vec{x} \in \mathbb{R}^n$, and you wish to express it as a linear combination of an orthonormal set of vectors $\vec{p}_1, \vec{p}_2, \dots, \vec{p}_n \in \mathbb{R}^n$

→ i.e, write $\vec{x} = \alpha_1 \vec{p}_1 + \alpha_2 \vec{p}_2 + \dots + \alpha_n \vec{p}_n$; we would like to find (12)
the scalars $\alpha_1, \alpha_2, \dots, \alpha_n$.

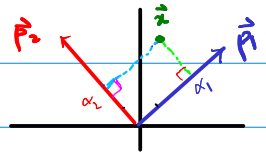
→ It turns out that:

$$\alpha_1 = \vec{p}_1^T \vec{x} \ (\text{or } \vec{x}^T \vec{p}_1) \ ; \ \alpha_2 = \vec{p}_2^T \vec{x} \ ; \ \alpha_3 = \vec{p}_3^T \vec{x} \ ; \ \dots \ ; \ \alpha_n = \vec{p}_n^T \vec{x} . \quad (13)$$

→ Proof: left as a small exercise. Hint: pre-multiply (12) by $\vec{p}_i^T$.

→ Operations like (13) are called a projection of $\vec{x}$ onto the basis (14)
$\{\vec{p}_1, \vec{p}_2, \dots, \vec{p}_n\}$

→ Projection has a nice geometric interpretation:



(15)

→ A very important and useful property of orthonormal bases is the following:

— the vector of projections of any $\vec{x}$ onto any orthonormal basis has the same norm as $\vec{x}$. (15.1)

— Proof: Take any $\vec{x} \in \mathbb{R}^n$ and project it onto some orthonormal basis $\{\vec{p}_1, \dots \vec{p}_n\}$, as in (12) and (13). Then $\vec{x} = \alpha_1 \vec{p}_1 + \alpha_2 \vec{p}_2 + \dots + \alpha_n \vec{p}_n$, with $\alpha_i = \vec{p}_i^T \vec{x}$ (15.2)

$$\Rightarrow \|\vec{x}\|^2 = \vec{x}^T \vec{x} = (\alpha_1 \vec{p}_1^T + \alpha_2 \vec{p}_2^T + \dots + \alpha_n \vec{p}_n^T)(\alpha_1 \vec{p}_1 + \alpha_2 \vec{p}_2 + \dots + \alpha_n \vec{p}_n) \quad (15.3)$$

$$\Rightarrow \|\vec{x}\|^2 = \alpha_1^2 \|\vec{p}_1\|^2 + \alpha_2^2 \|\vec{p}_2\|^2 + \dots + \alpha_n^2 \|\vec{p}_n\|^2 \ (\because \vec{p}_i^T \vec{p}_j = 0 \text{ if } i \neq j, \text{ by orthogonality}) \quad (15.4)$$

$$\Rightarrow \|\vec{x}\|^2 = \alpha_1^2 + \alpha_2^2 + \dots + \alpha_n^2 = \|\vec{\alpha}\|^2, \text{ if you define} \quad (15.5)$$

$$\rightarrow \vec{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_n]^T = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \quad (15.6)$$

→ Note: (15.6) and (13) can be written compactly as:

→ $\vec{\alpha} = P^T \vec{x}$ (15.7)

→ where $P \triangleq \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{p}_1 & \vec{p}_2 & \cdots & \vec{p}_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$ (15.8)

→ A useful property of P is that its rows are also orthonormal. (15.85)

→ $P^T P = I$ (from definition of P, i.e., (15.8) and (5))

→ $P^T = P^{-1} \Rightarrow P P^T = P P^{-1} = I$ (15.9)

→ $\begin{bmatrix} \leftarrow & \vec{q}_1 & \rightarrow \\ \leftarrow & \vec{q}_2 & \rightarrow \\ & \vdots & \\ \leftarrow & \vec{q}_n & \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{q}_1 & \vec{q}_2 & \cdots & \vec{q}_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & 1 & & 0 \\ & & \ddots & \\ & 0 & & 1 \end{bmatrix}$ (15.91)

rows of P

P $\qquad$ P^T

$\Rightarrow \quad \vec{q}_i^T \vec{q}_j = \begin{cases} 1, & \text{if } i=j \\ 0, & \text{otherwise} \end{cases}$ (15.92)

→ Another useful property of a set of orthonormal vectors $\{\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_r\}$, for any $r \leq n$, is that

$$\underbrace{\|\vec{p}_1\|^2}_{1} + \underbrace{\|\vec{p}_2\|^2}_{1} + \cdots + \underbrace{\|\vec{p}_r\|^2}_{1} = \sum_{i=1}^{r} \sum_{j=1}^{n} p_{ji}^2 = r. \quad (15.93)$$

— Here, p_{ji} represents the $(\text{row}=j, \text{col}=i)^{\text{th}}$ entry of the matrix (15.94)

$\begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \vec{p}_1 & \vec{p}_2 & \cdots & \vec{p}_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$

OUTER PRODUCTS OF VECTORS AND RANK-1 MATRICES

→ To best appreciate what SVDs can do for us, we need to know about outer products of vectors.

→ Suppose you have two vectors $\vec{x} \in \mathbb{R}^n$ & $\vec{y} \in \mathbb{R}^m$, then

$\vec{x}\vec{y}^T$ is called an outer product (16)

→ $\vec{y}\vec{x}^T$ is also a (different) outer product.

→ outer products are better appreciated when you see them in expanded form:

$$\underbrace{\vec{x}}_{n}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \underbrace{\vec{y}^T}_{m}\begin{bmatrix} y_1, y_2, \cdots, y_m \end{bmatrix} = \overbrace{\begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_m \\ x_2y_1 & x_2y_2 & \cdots & x_2y_m \\ \vdots & & & \\ x_ny_1 & x_ny_2 & \cdots & x_ny_m \end{bmatrix}}^{m} \Big\} n \qquad (17)$$

→ So we see that $\vec{x}\vec{y}^T$ is an $n \times m$ matrix.

→ not to be confused with $\vec{x}^T\vec{y}$ or $\vec{y}^T\vec{x}$ – which are scalars.

→ $\vec{x}^T\vec{y}$ and $\vec{y}^T\vec{x}$ are called INNER PRODUCTS or DOT PRODUCTS.

→ moreover, $\vec{x}\vec{y}^T$ is a RANK-1 matrix (18)

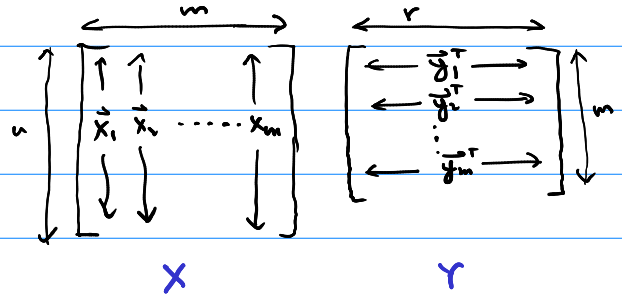
– each row of (17) is just a scaled copy of any other one.

– and also, each column of (17) is a scaled copy of any other column.

– this is easy to see if you write (17) like this:

$$\underbrace{\vec{x}}_{n}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \underbrace{\vec{y}^T}_{m}\begin{bmatrix} y_1, y_2, \cdots, y_m \end{bmatrix} = \begin{bmatrix} \longleftarrow x_1\vec{y}^T \longrightarrow \\ \longleftarrow x_2\vec{y}^T \longrightarrow \\ \vdots \\ \longleftarrow x_n\vec{y}^T \longrightarrow \end{bmatrix} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ y_1\vec{x} & y_2\vec{x} & \cdots & y_m\vec{x} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \qquad (19)$$

— One reason outer products are useful is that matrix multiplication can be expressed as a sum of outer products:



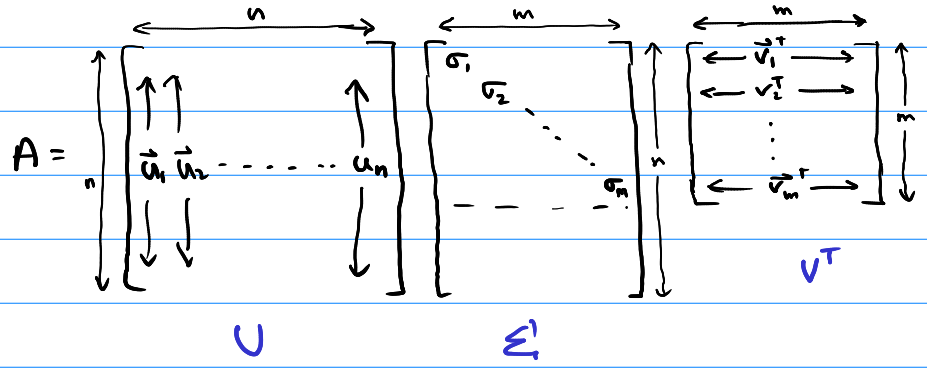
$$= \vec{x}_1\vec{y}_1^T + \vec{x}_2\vec{y}_2^T + \cdots + \vec{x}_m\vec{y}_m^T = \sum_{i=1}^{m} \vec{x}_i\vec{y}_i^T \qquad (20)$$

each one is a nxr rank-1 matrix

— to convince yourself of this: just try a few small examples.

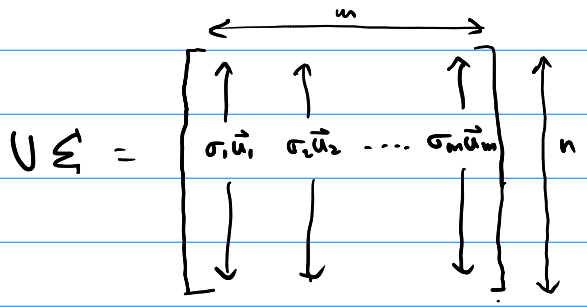
BACK TO SVDs – we are now ready to use the above mathematical machinery to explore (1) in more detail.

→ Rewriting (1) in expanded form:



(21)

→ Notice that $U\Sigma$ above expands to:



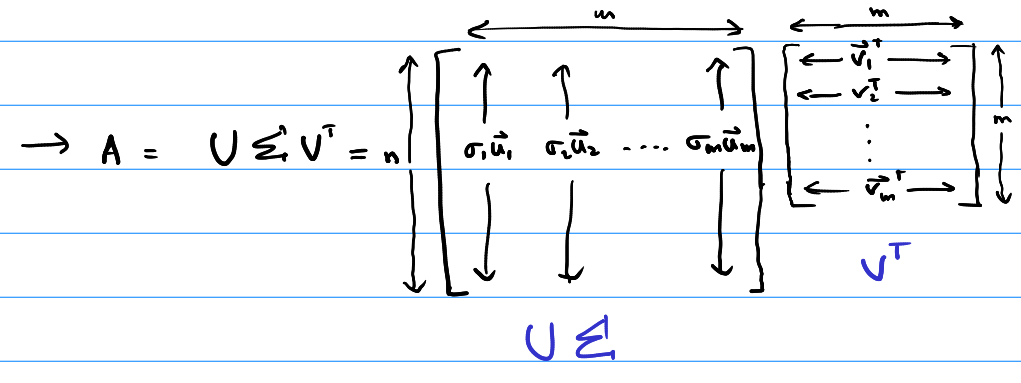
(22)

– an interesting observation:

– from (22), we see that $\vec{u}_{m+1}, \ldots, \vec{u}_n$ are not even involved.
– which means that they don't affect (1) or (21).
– so in fact, $\vec{u}_{m+1}$ to $\vec{u}_n$ can be any vectors that complete $\vec{u}_1, \ldots, \vec{u}_m$ into an orthonormal basis for $\mathbb{R}^n$.

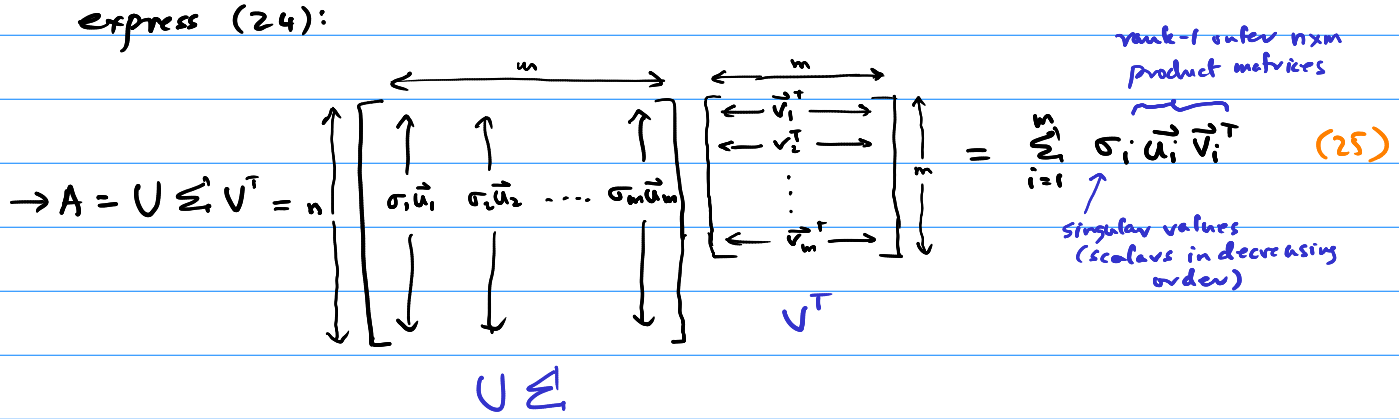
(23)

→ In view of (22), (21) becomes:



(24)

→ now we can use the outer product form of matrix multiplication (20) to express (24):



→ Consider $\vec{u}_i \vec{v}_i^T$ for any $i \in 1, \ldots, m$.

→ recall that $\|\vec{u}_i\| = \|\vec{v}_i\| = 1$, because they are members of orthonormal bases.

→ let's write $\vec{u}_i \vec{v}_i^T$ in expanded form, as in (19):

$$\rightarrow R_i \triangleq \vec{u}_i \vec{v}_i^T = \begin{bmatrix} v_{i1}\vec{u}_i & v_{i2}\vec{u}_i & \cdots & v_{im}\vec{u}_i \end{bmatrix} \quad (26)$$

→ Now, we would like to find the Frobenius norm (denoted $\|R_i\|_F$) of R_i:

→ the Frobenius norm is simply the square root of the sum of all the entries of R_i. (27)

→ the sum of the squares of all the entries of R_i is the sum of the squares of the norms of each column.

→ i.e., $\|R_i\|_F^2 = v_{i1}^2 \|\vec{u}_i\|^2 + v_{i2}^2 \|\vec{u}_i\|^2 + \cdots + v_{im}^2 \|\vec{u}_i\|^2$

$$\left[\text{Frobenius norm of } R_i\right]^2 = (v_{i1}^2 + v_{i2}^2 + \cdots + v_{im}^2)\|\vec{u}_i\|^2$$

$$= \|\vec{v}_i\|^2 \|\vec{u}_i\|^2$$

$\Rightarrow \|R_i\|_F^2 = 1$ (due to orthonormality of $\vec{u}_i$ and $\vec{v}_i$) (28)

→ what (28) and (25) are telling us is:

→ any matrix A can be decomposed into a sum of rank 1 matrices, each of unit Frobenius norm, weighted by the singular values. (29)

→ i.e., we can think of $\sigma_1 \vec{u}_1 \vec{v}_1^T$ (where σ_1 = "weight", $\vec{u}_1$ = "size 1" — ie, Frobenius norm = 1) as the "most important rank-1 component" of A, since $\sigma_1 \geq \sigma_2, \sigma_3, \ldots, \sigma_m$ (30)

→ $\sigma_2 \vec{u}_2 \vec{v}_2^T$ is the "2nd most important rank-1 component" (31)

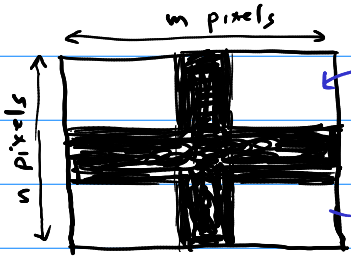
("size 1")

→ similarly for $\sigma_3 \vec{u}_3 \vec{v}_3^T$, $\sigma_4 \vec{u}_4 \vec{v}_4^T$ and so on.

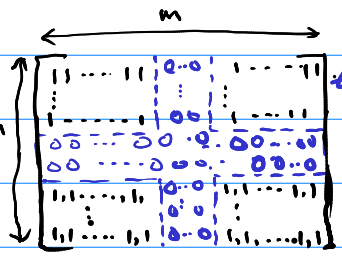
— The uses of (29) – (31) are perhaps most easily seen by applying SVD to images

---

— Suppose you have a simple B&W image like this:



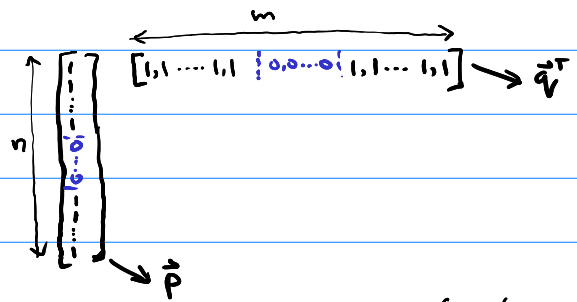
, captured by the matrix $A =$



(32)

— A is a rank 1 matrix, since it can be written as

$A =$



(33)

where $\sigma_1 = \|\vec{p}\| \, \|\vec{q}\|$ (34)

$\vec{u}_1 = \frac{\vec{p}}{\|\vec{p}\|}$, $\vec{v}_1 = \frac{\vec{q}}{\|\vec{q}\|}$

— an SVD of A would have only one significant component, $\sigma_1 \vec{u}_1 \vec{v}_1^T$, where

→ $\sigma_2, \sigma_3, \ldots \sigma_m$ would all be 0.

- More generally, images composed of "simple rectangular patterns" will have only a few significant singular values.
  - see the flag examples in the slides
- Images that are not so simple — especially if they have non-rectangular shapes in them — typically have a larger number of significant singular values.
- the rank-1 matrices $\vec{u}_1\vec{v}_1^T$, $\vec{u}_2\vec{v}_2^T$, ..., etc., often have some visual interpretation or significance — they are sometimes called "features"

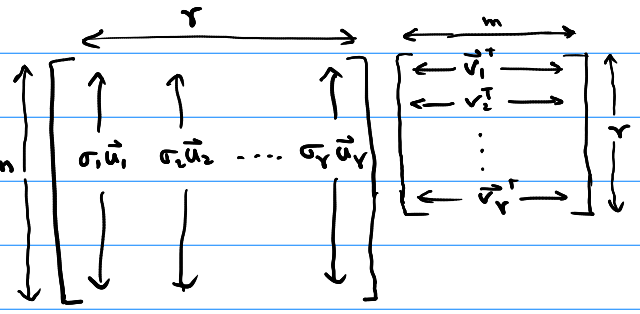
- (Perhaps more surprisingly) almost every image that makes instinctive sense to human beings — eg, faces, pictures of scenery, etc. — can be well approximated visually if you approximate it using only the most significant singular values.
  - More precisely, suppose you have an image matrix $A \in \mathbb{R}^{n \times m}$
    → take an SVD of it (recall (25)):

$$\rightarrow A = U \Sigma V^T = n \underbrace{\overset{m}{\begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \sigma_1\vec{u}_1 & \sigma_2\vec{u}_2 & \cdots & \sigma_m\vec{u}_m \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}}}_{U\Sigma} \underbrace{\overset{m}{\begin{bmatrix} \leftarrow \vec{v}_1^T \rightarrow \\ \leftarrow \vec{v}_2^T \rightarrow \\ \vdots \\ \leftarrow \vec{v}_m^T \rightarrow \end{bmatrix}} m}_{V^T} \qquad (35)$$

→ Suppose you decide that only the first r singular values are significant, with $r < m$.

→ for images, a good rule of thumb for deciding r is to ensure that $\sigma_r \simeq 10^{-2}\sigma_1$, i.e., that all singular values 2 orders of magnitude, or more less than σ_1 are negligible.

→ then, <u>approximate</u> (35), keeping only the first r columns of U and rows of V:

$\rightarrow A \simeq A_r \triangleq$ 

(36)

→ if $r \ll m$ (as it often is), and the approximate image still looks good, then we can achieve a way to represent the image with less data than storing every pixel — i.e, image compression.

→ example: see Michel's image in the slides, eg, for $r = 50$.

— The intuition that you can "throw away" all the terms $\sigma_i \vec{u}_i \vec{v}_i^T$ that have relatively small σ_i (compared to σ_1, or the top few singular values) can be concretized mathematically, as follows:

— The difference in Frobenius norm (27) between A & A_r (in (36)) is $\sqrt{\sigma_{r+1}^2 + \sigma_{r+2}^2 + \cdots + \sigma_m^2}$ (36·1)

— Proof: (36) is the same as: $A_r = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T$, with $r < m$. (36·2)

→ Since $A = \sum_{i=1}^{m} \sigma_i \vec{u}_i \vec{v}_i^T$ (by (25)), we have $A - A_r = \sum_{i=r+1}^{m} \sigma_i \vec{u}_i \vec{v}_i^T$ (36·3)

→ We now write out the k^{th} column of $(A - A_r)$, where $1 \le k \le m$:

→ Let $\vec{q}_k \in \mathbb{R}^n$ be this column, then $\vec{q}_k = \sum_{i=r+1}^{m} \sigma_i v_{ki} \vec{u}_i$, (36·4)

where v_{ki} denotes the k^{th} element of the column vector $\vec{v}_i$.

→ illustration: let $k = 2$ (i.e, the 2nd column):

→ 2nd col of $\sigma_1 \vec{u}_1 \vec{v}_1^T = \sigma_1 v_{21} \vec{u}_1$ (using the notation of (15·44))

→ 2nd col of $\sigma_2 \vec{u}_2 \vec{v}_2^T = \sigma_2 v_{22} \vec{u}_2$

⋮

→ 2nd col of $\sigma_r \vec{u}_r \vec{v}_r^T = \sigma_r v_{2r} \vec{u}_r$

⋮

→ 2nd col of $\sigma_m \vec{u}_m \vec{v}_m^T = \sigma_m v_{2m} \vec{u}_m$

(36·45)

$\Rightarrow \vec{q}_2 \equiv$ 2nd column of $\sum_{i=r+1}^{m} \sigma_i \vec{u}_i \vec{v}_i^T = \sum_{i=r+1}^{m} \sigma_i v_{2i} \vec{u}_i$

→ Since $\{\vec{u}_i\}$ are orthonormal, then by (15.1) – (15.6), we have

→ $$\|\vec{q}_k\|^2 = \sum_{i=v+1}^{m} \sigma_i^2 v_{ki}^2 \qquad (36.5)$$

→ If we sum (36.5) over all $k = 1, \dots, m$, (i.e, all the columns of $A - A_v$), we get

→ this is just $\|\vec{v}_i\|^2 = 1$

→ $$\sum_{k=1}^{m} \|\vec{q}_k\|^2 = \sum_{k=1}^{m} \sum_{i=v+1}^{n} \sigma_i^2 v_{ki}^2 = \sum_{i=v+1}^{m} \sigma_i^2 \sum_{k=1}^{m} v_{ki}^2 \qquad (36.6)$$

→ $$\sum_{k=1}^{m} \|\vec{q}_k\|^2 = \sum_{i=v+1}^{m} \sigma_i^2 \qquad (36.7)$$

→ Finally, we need to simply observe that:

→ $\|A - A_v\|_F^2$ = sum of the norm-squares of the columns $= \sum_{k=1}^{m} \|\vec{q}_k\|^2$ (36.8)

— Putting together (36.8) and (36.7), we get what we set out to prove, i.e.,

$$\|A - A_v\|_F = \sqrt{\sum_{i=v+1}^{m} \sigma_i^2} \qquad (36.9)$$

— This result (36.1) or (36.9) is very significant, since it exactly quantifies the error (in Frobenius norm) between the original matrix A and an approximation truncated to the top v singular values.

→ An immediate corollary is that the relative error is:

$$\epsilon_v \triangleq \frac{\|A - A_v\|_F}{\|A\|_F} = \sqrt{\frac{\sum_{i=v+1}^{m} \sigma_i^2}{\sum_{i=1}^{m} \sigma_i^2}} = \frac{1}{\sqrt{1 + \frac{\sum_{i=1}^{v} \sigma_i^2}{\sum_{i=v+1}^{m} \sigma_i^2}}} \qquad (36.95)$$

— thus, if the squared sums of the truncated SVs is very much smaller than that of the non-truncated ones, the error is very small.

— in the slides: show the error for the various examples.

SVDs for IDENTIFYING "FEATURES" IN TABULAR DATA

— SVDs are also useful for analysing tabular data.

— for example, suppose we have a matrix representing grades in an exam.

$A \in \mathbb{R}^{6\times 8}$ (more generally, data samples = rows)

QUESTIONS (or more generally: types of data = columns)

STUDENTS	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
S1	1	2	1	3	2	1	0	1
S2	10	9	1	10	10	2	10	8
S3	5	4	2	6	5	0	3	2
S4	9	8	3	9	10	1	10	9
S5	0	0	0	1	3	2	2	2
S6	10	10	3	10	10	2	10	10

(37)

— we run an SVD on the data, to get:

$$\rightarrow A = \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{v}_2^T + \cdots + \sigma_m \vec{u}_m \vec{v}_m^T \qquad (38)$$

— unlike for the case of images, the "features" $\vec{u}_1 \vec{v}_1^T$, $\vec{u}_2 \vec{v}_2^T$ are often less intuitive for general data.

— However, the "top columns" of U (i.e., $\vec{u}_1, \vec{u}_2$, etc.) and the "top rows" of V^T (i.e., $\vec{v}_1^T, \vec{v}_2^T$, etc.) can often be interpreted meaningfully, as follows: (39)

→ $\vec{u}_1$ (which is of length $n=6$ in the above example), can be thought of as the "strongest column (or data type) feature" (or Question feature in the above example). (40)

→ $\vec{u}_2$ would be the 2nd strongest data type/col. feature.

→ with this interpretation, you can project each of the data type (Question) columns onto $\vec{u}_1$ through $\vec{u}_n$.

For example, the Q4 column would be

$$\begin{bmatrix} 3 \\ 10 \\ 6 \\ 9 \\ 1 \\ 10 \end{bmatrix}_{Q4} = \alpha_{41}\vec{u}_1 + \alpha_{42}\vec{u}_2 + \cdots + \alpha_{46}\vec{u}_6, \quad (41)$$

→ where $\alpha_{4i} = \vec{u}_i^T \begin{bmatrix} 3 \\ 10 \\ 6 \\ 9 \\ 1 \\ 10 \end{bmatrix}_{Q4}$ is the strength of the i^{th} data type (Question) feature for Q4. (42)

→ we can find all the αs (42) for all the columns of the matrix via

→ $\hat{A} = U^T A$, (43)

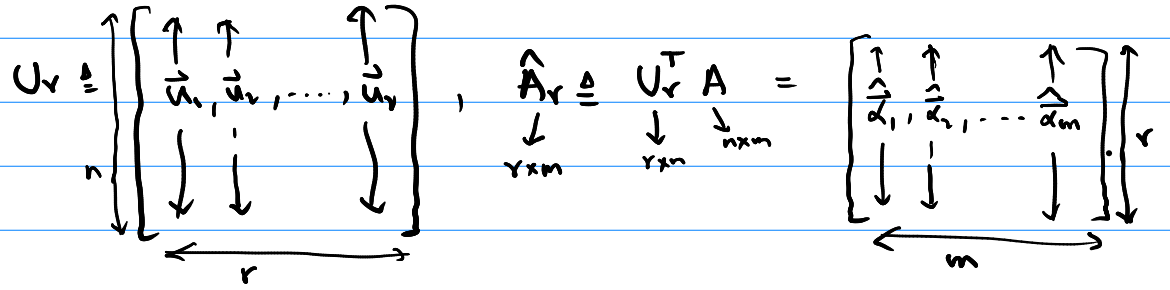
→ where $\hat{A} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \vec{\alpha}_1 & \vec{\alpha}_2 & \cdots & \vec{\alpha}_m \\ \vdots & \vdots & & \vdots \end{bmatrix}$ (44)

→ and the k^{th} column of A can be recovered as $U\vec{\alpha}_k$ (45)

— i.e., the entries of $\vec{\alpha}_k$ $(a_{1k}, a_{2k}, \ldots, a_{nk})$ are the projections in (42) and (41)

→ so, in fact, A can be recovered as $A = U\hat{A}$ (46)

→ Now, suppose we choose to retain only the 1st $r < m$ "column features", i.e, we project onto only the first r columns of U:



(47)

→ and we reconstruct A from $\hat{A}_r$ similarly, using only the first r columns of U, i.e.,

$$A_r \triangleq U_r \hat{A}_r \tag{48}$$

→ what is the re-construction error $\|A - A_r\|_F$?

→ It is straightforward to show that this error is:

$$\|A - A_r\|_F = \epsilon_r = \sqrt{\sum_{i=r+1}^{m} \sigma_i^2} \quad \text{(exactly as in (36.9))} \tag{49}$$

— the proof is the same

→ (49), interpreted in words, is:

→ If we represent all the columns of A by projection onto only the first r "column features" (eg, Questions or types of data), then we can recover all the columns of A with total mean-squared (Frobenius) error given by (49). (50)

→ i.e., if $\sigma_{r+1}, \ldots, \sigma_m$ are "negligible" compared to $\sigma_1, \ldots, \sigma_r$, this error can be small → in the sense that ϵ_r in (39.95) is $\ll 1$

→ So again, if we can truncate with small error to recover "column features", we can achieve good data compression.

→ Note: we can, similarly, define row features, in exact analogy with column features, as above.

→ left as an exercise

→ if you truncate to r row features, what is the Frobenius norm error?

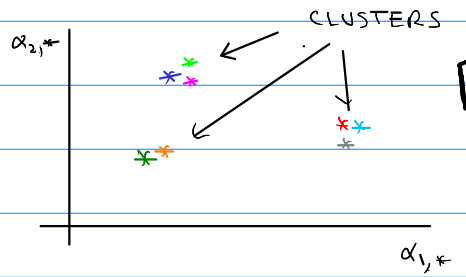
CLUSTERING USING COLUMN (or row) FEATURES

→ The projections on column features (eg, in (41)-(44), and (47)), if plotted against each other graphically, can reveal meaningful clusters in the data.

→ For example, consider $\hat{A}_r$ in (47):

$$\hat{A}_r = \left[\vec{\alpha}_1, \vec{\alpha}_2, \ldots, \vec{\alpha}_m \right] = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{r1} & \alpha_{r2} & \cdots & \alpha_{rm} \end{bmatrix} \quad \begin{matrix} \leftarrow m \rightarrow \\ \\ \\ \\ \updownarrow r \end{matrix} \qquad (51)$$

— Plot the first 2 entries of each column of (51) as a 2-D point:



[each * is a column of (51), corresponding to a column of A (47)] (52)

→ The clusters reveal that the corresponding columns of A have similar values of "column features" (since their projections onto the features are similar).

→ For our example in (37), columns 3 & 6 (Q3 & Q6) might fall in a cluster, since most students did poorly in both.

→ See the slides for more precise examples.

PRINCIPAL COMPONENT ANALYSIS (PCA)

- PCA is another useful way of looking at matrices/tables of data like (37).

- To understand PCA, we first need to understand the concepts of co-variance and correlation matrices of data.

- In the following, it is helpful (though not essential) to think of $A \in \mathbb{R}^{n\times m}$ as consisting of m columns, each column corresponding to n samples of some type of data (like position, velocity, etc.) (52.5)
  → It is also helpful to think $n > m$, i.e, more samples than types of data.

- COVARIANCE MATRICES

  - Suppose you are given $A \in \mathbb{R}^{n\times m}$
    - for each column of A, find its average
      - i.e, average over all the rows of the column:

$$\mu_k = \text{average of } k^{\text{th}} \text{ column } (1\le k\le m) = \frac{1}{n}\sum_{i=1}^{n} a_{ik} \tag{53}$$

  - define $\vec{\mu}^T$ to be the row vector of all the column averages:

$$\vec{\mu}^T = [\mu_1, \mu_2, \dots, \mu_m] = \frac{1}{n}\underbrace{[1, 1, \dots, 1]}_{n} A \tag{54}$$

  - now, subtract μ_k from the k^{th} column of A, and do this for all columns $k=1,\dots,m$. The result can be written as

$$\tilde{A} = A - \begin{bmatrix} \vdots \\ \vdots \\ \vdots \\ \vdots \end{bmatrix} \overbrace{[\mu_1, \dots, \mu_m]}^{\vec{\mu}^T} \tag{55}$$

  ← length n col. vector of 1s.

— $\tilde{A}$ as defined in (55) is the same data, but with the mean of each column subtracted from that column (55·1).

— i.e., each column of $\tilde{A}$ has zero mean, i.e.,

$$\underbrace{[1, 1, 1, \cdots, 1]}_{n}\,\tilde{A} = \vec{0}^T \tag{55·2}$$

— $\tilde{A}$ is sometimes called the mean-centered data matrix

→ now, form the matrix

$$S \triangleq \frac{1}{n}\underset{m\times n}{\tilde{A}^T}\underset{n\times m}{\tilde{A}} \tag{56}$$

mxm

→ S is called the co-variance matrix (cf A, or $\tilde{A}$)

— Example: Suppose $A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$; (57·1)

— $\mu_1 = \frac{1+2+3}{3} = 2$; $\mu_2 = \frac{4+5+6}{3} = 5$

— $\tilde{A} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$; (57·2)

— $S = \frac{1}{3}\tilde{A}^T\tilde{A} = \frac{1}{3}\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$ (57·3)

— PROPERTIES OF COVARIANCE MATRICES

— S is square and symmetric (i.e., $S^T = S$) (58)

- both follow immediately from the definition (56)

— the diagonal entries of S are greater than or equal to 0 (59)

- Let $\vec{a}_i$ be the i^{th} column of $\tilde{A}$, then the i^{th} diagonal entry of S is $\frac{1}{n}\vec{a}_i^T\vec{a}_i = \frac{1}{n}\|\vec{a}_i\|^2 \geq 0$

— Suppose we write S as

$$\rightarrow S = \begin{bmatrix} s_1^2 & s_{12} & s_{13} & \cdots & s_{1m} \\ s_{21} & s_2^2 & s_{23} & \cdots & s_{2m} \\ \vdots & & \ddots & & \\ s_{m1} & s_{m2} & \cdots & \cdots & s_m^2 \end{bmatrix} \tag{60}$$

→ then $s_{ij} = s_{ji}, \ \forall \ i \neq j$ (61)

↳ due to symmetry, i.e., (58)

→ and $|s_{ij}| \leq s_i s_j, \ \forall \ i \neq j$ (62)

→ proof:

→ Let $\vec{a}_i$ and $\vec{a}_j$ be the i^{th} and j^{th} columns of $\tilde{A}$, respectively. (62.1)

→ then $s_{ij} = \frac{\vec{a}_i^T \vec{a}_j}{n}$, from definition of S (56). (62.2)

→ we also have: $s_i^2 = \frac{1}{n} \vec{a}_i^T \vec{a}_i = \frac{\|\vec{a}_i\|^2}{n}$, and $s_j^2 = \frac{1}{n} \vec{a}_j^T \vec{a}_j = \frac{1}{n}\|\vec{a}_j\|^2$ (62.3)

→ The famous Cauchy-Schwarz inequality states that for any 2 vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$:

$$\|\vec{x}^T \vec{y}\|^2 \leq \|\vec{x}\|^2 \|\vec{y}\|^2 \tag{62.4}$$

→ Applying it to $\vec{a}_i$ and $\vec{a}_j$, we get:

$$\|\vec{a}_i^T \vec{a}_j\|^2 \leq \|\vec{a}_i\|^2 \|\vec{a}_j\|^2 \tag{62.5}$$

$$\Rightarrow \frac{\|\vec{a}_i^T \vec{a}_j\|^2}{n^2} \leq \frac{\|\vec{a}_i\|^2}{n} \frac{\|\vec{a}_j\|^2}{n} \tag{62.6}$$

$$\Rightarrow |s_{ij}|^2 \leq s_i^2 s_j^2 \Rightarrow |s_{ij}| \leq s_i s_j \tag{62.7}$$

→ The diagonal entries of S (i.e., s_i^2) are the VARIANCES of the columns of A (62.8)

→ Follows directly from the definitions of S and of variance:

→ The variance of the k^{th} column of A is defined as:

→ $\sigma_k^2 \triangleq \frac{\|\vec{a}_k\|^2}{n}$, where $\vec{a}_k$ is the k^{th} column of $\tilde{A}$ (62.81)

→ which is the same as s_i^2 (by (62.3))

— From Covariance Matrices to Correlation Matrices

— The CORRELATION MATRIX (of A or $\tilde{A}$) is defined to be:

$$R \triangleq \begin{bmatrix} 1 & \frac{s_{12}}{s_1 s_2} & \frac{s_{13}}{s_1 s_3} & \cdots & \frac{s_{1m}}{s_1 s_m} \\ \frac{s_{21}}{s_2 s_1} & 1 & \frac{s_{23}}{s_2 s_3} & \cdots & \frac{s_{2m}}{s_2 s_m} \\ \vdots & & & & \\ \frac{s_{m1}}{s_m s_1} & \cdots & \cdots & & 1 \end{bmatrix}, \text{ i.e, } R_{ij} = \frac{s_{ij}}{s_i s_j} \qquad (63)$$

— Due to (62), we immediately see that the off-diagonal terms of R are ≤ 1. (64)

— $R_{ij} \triangleq \frac{s_{ij}}{s_i s_j}$ is called the correlation between the i^{th} and j^{th} columns of A (or $\tilde{A}$) (65)

— Example: continuing from (57.1) and (57.3), we see immediately that

$$R = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \text{ i.e, the correlation between the 2 columns is 1.} \qquad (66)$$

— Another example: if $A = \begin{bmatrix} 1 & 6 \\ 2 & 5 \\ 3 & 4 \end{bmatrix}$, $R = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ (67)

— here, the correlation between the 2 columns is -1.

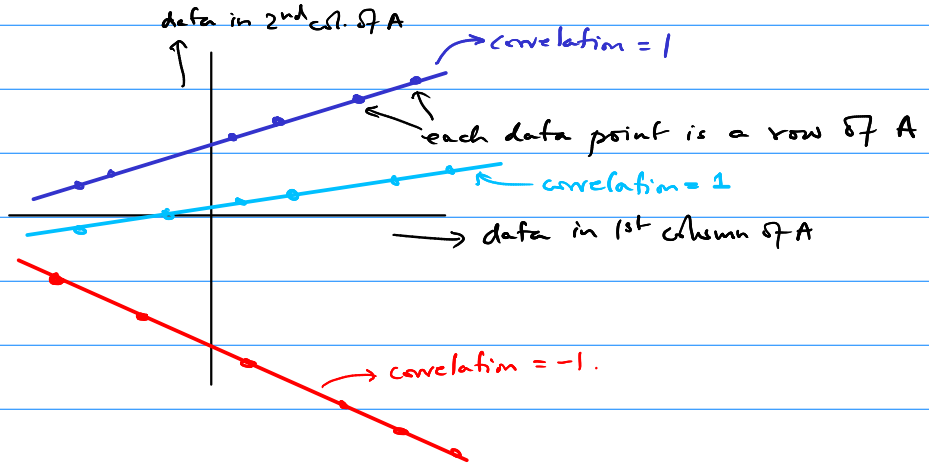
— Correlations have a "nice" property:

→ if every column of $\tilde{A}$ is simply a scaled version of a single vector $\vec{a}$, then every entry of R is either 1 or -1. (68)

→ proof: left as an exercise

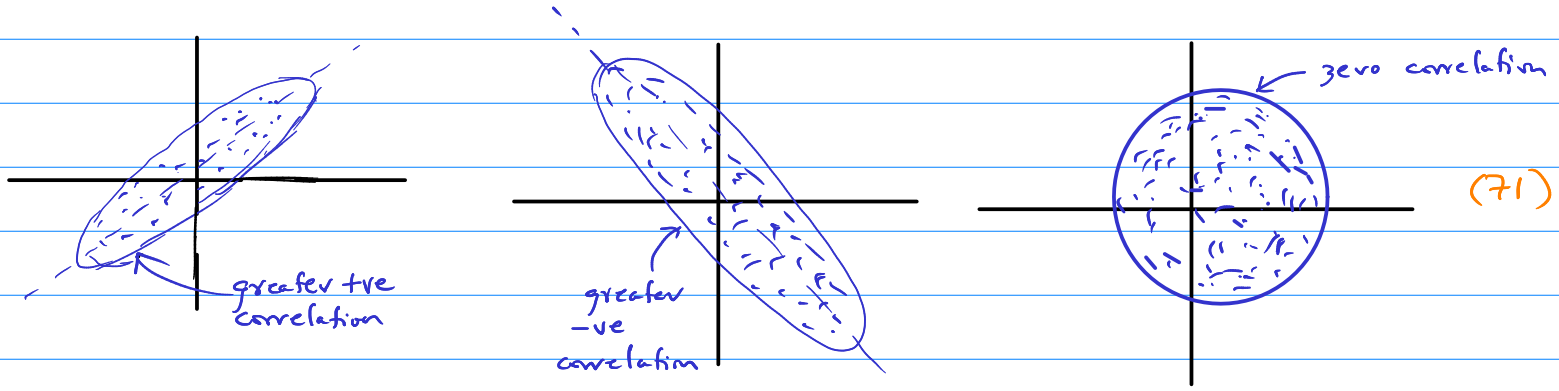
— geometrical interpretation of (68) in 2D:

— If the data in A lie on a straight line, then the correlation is 1 if the straight line has +ve slope, and −1 if it has −ve slope. (69)



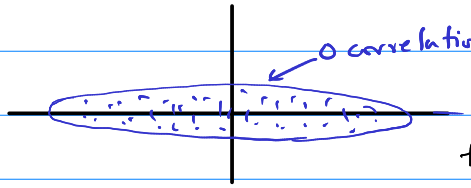
(70)

— if the data is more scattered — but still clustered around a line — correlations still can give you a rough idea how closely clustered around a line the data is. The following figures provide a rough guide, assuming that the samples are uniformly spread within the bounding shapes.



(71)

— Correlations do not, however, provide a clear idea of the tilt of the line around which the data is clustered, nor even a reliable notion of the spread of the data. For example, the following also has zero correlation:



even though it is much more clustered along a line than the circular spread in (71) (73)

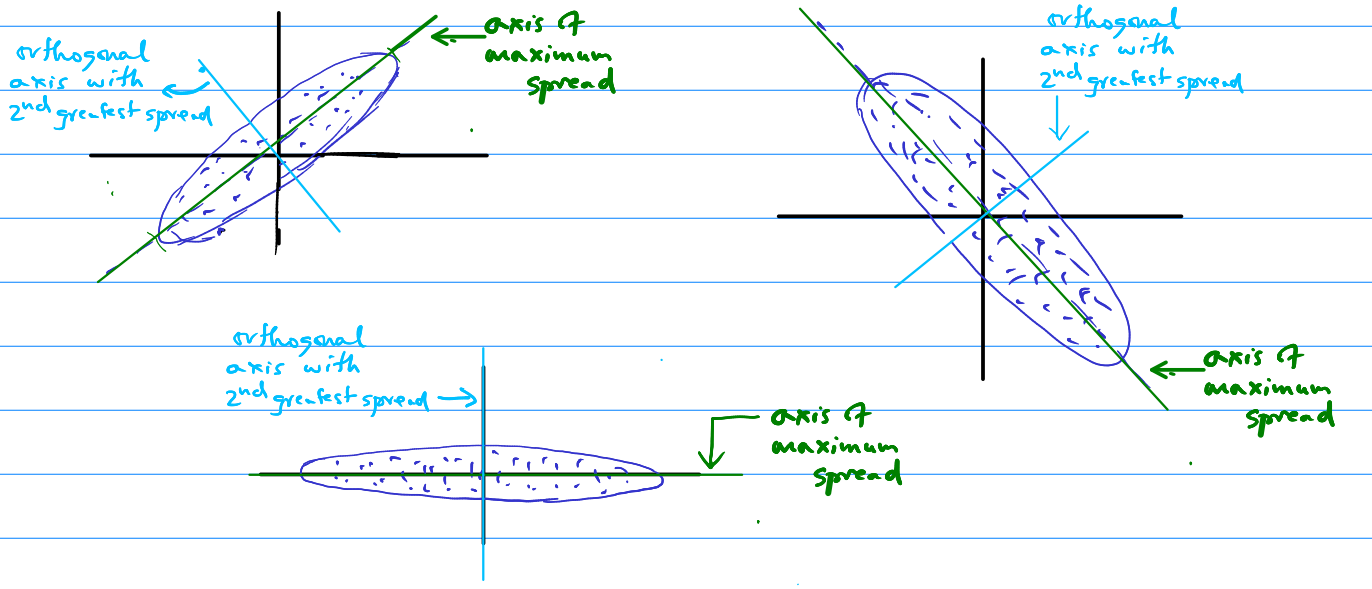
— There are more examples in the slides.

— IT IS IN THIS CONTEXT THAT PCA IS MUCH MORE INFORMATIVE:

— PCA identifies the "main axes" around which the data clusters (74)

— it provides quantitative information about the spread of the data along each of these main axes.

— For example, PCA will correctly identify these axes in the following examples:



(75)

— PCA is basically nothing more than eigen-decomposition of the covariance matrix S in (60) (76)

— because the covariance matrix has special structure and properties, so do the eigenvectors and eigenvalues — which is what results in the ability of PCA to find the axes corresponding to the greatest data spreads (77)

— To understand PCA properly, we first need to derive certain properties of (real) symmetric and co-variance matrices:

(not necessarily in the form $\tilde{A}^T\tilde{A}$)

→ If S is <u>any</u> (real) symmetric matrix, then (78)

1. The eigenvalues of S are all real. (79)

— proof: if λ is an eigenvalue of S, then $S\vec{p} = \lambda\vec{p}$ (79.1)
for some $\vec{p}$ (eigenvector of S)

→ since S is <u>real</u>, we have $S\vec{p} = \bar{\lambda}\bar{\vec{p}}$ (79.2)

→ since S is <u>symmetric</u>, we have $\vec{p}^T S = \lambda\vec{p}^T$ (79.3)

$\therefore \vec{p}^T S\vec{p} = \bar{\lambda}\,\vec{p}^T\bar{\vec{p}} = \bar{\lambda}\,\|\vec{p}\|^2$ (by 79.2) (79.4)

$= \lambda\vec{p}^T\vec{p} = \lambda\|\vec{p}\|^2$ (by 79.3) (79.5)

$\Rightarrow \lambda = \bar{\lambda} \Rightarrow \lambda$ is real

2. A set of <u>real</u> eigenvectors $\{\vec{p}_1, \vec{p}_2, \cdots, \vec{p}_m\}$ of S can be found (80)

→ Pf: Suppose we have <u>some</u> set of eigenvectors $\{\vec{b}_1, \cdots, \vec{b}_m\}$ of S, not necessarily real. Then $S\vec{b}_i = \lambda_i\vec{b}_i$. (80.1)

→ Let $\vec{b}_i = \vec{p}_i + j\vec{q}_i$, where $\vec{p}_i$ and $\vec{q}_i$ are the real and imaginary parts of $\vec{b}_i$. (80.2)

→ Then $S\vec{b}_i = S\vec{p}_i + jS\vec{q}_i = \lambda_i(\vec{p}_i + j\vec{q}_i) = \lambda_i\vec{p}_i + j(\lambda_i\vec{q}_i)$ (80.3)

→ Since λ_i is real (from (79)), we have

$S\vec{p}_i = \lambda_i\vec{p}_i$, i.e., $\{\vec{p}_1, \cdots, \vec{p}_m\}$, all real, are eigenvectors of S. (80.4)

3. The eigenvectors in (80) are orthogonal (81)

– Pf: $S\vec{p}_i = \lambda_i \vec{p}_i$, $i = 1, \ldots, m$, with λ_i & $\vec{p}_i$ real ((79) & (80))

→ from symmetry of S, we have:

$$\vec{p}_j^{\,T} S = \lambda_j \vec{p}_j^{\,T}, \quad \text{for } j = 1, \ldots, m \tag{82.1}$$

$$\rightarrow \lambda_j \vec{p}_j^{\,T} \vec{p}_i = \vec{p}_j^{\,T} S \vec{p}_i = \vec{p}_j^{\,T} \lambda_i \vec{p}_i = \lambda_i \vec{p}_j^{\,T} \vec{p}_i \tag{82.2}$$

$$\rightarrow \lambda_j \vec{p}_j^{\,T} \vec{p}_i = \lambda_i \vec{p}_j^{\,T} \vec{p}_i \Rightarrow \vec{p}_j^{\,T} \vec{p}_i (\lambda_j - \lambda_i) = 0 \tag{82.3}$$

→ If we assume that all the eigenvalues are distinct, then $\lambda_i - \lambda_j \neq 0$ if $i \neq j$, hence $\vec{p}_j^{\,T} \vec{p}_i = 0$ if $i \neq j$ (82.4)

→ Even if we have repeated eigenvalues, it can be shown (though we won't do so here) that $\vec{p}_j$ and $\vec{p}_i$ can be found such that they are orthogonal (82.5)

4. A real orthonormal set of eigenvectors of S can be found. (83)

Pf: from (80), we have $\{\vec{p}_1, \cdots, \vec{p}_m\}$ real, and from (82), they are orthogonal.

→ then $\left\{ \frac{\vec{p}_1}{\|\vec{p}_1\|}, \frac{\vec{p}_2}{\|\vec{p}_2\|}, \ldots, \frac{\vec{p}_m}{\|\vec{p}_m\|} \right\}$ are orthonormal (83.1)

→ we can then just redefine $\{\vec{p}_1, \cdots, \vec{p}_m\}$ in (80) to be orthonormal (83.4)

→ Now, if S is a real symmetric matrix that can be written in the form $S = B^T B$, for some matrix B (84)

→ then the eigenvalues of S are all ≥ 0 (85)

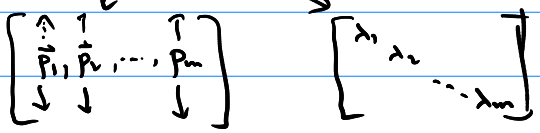
→ Pf: $S\vec{p} = \lambda \vec{p} \Rightarrow B^T B \vec{p} = \lambda \vec{p}$ (85.1)

$$\Rightarrow \vec{p}^{\,T} B^T B \vec{p} = \lambda \|\vec{p}\|^2 \Rightarrow (B\vec{p})^T (B\vec{p}) = \lambda \|\vec{p}\|^2 \tag{85.2}$$

$$\Rightarrow \|B\vec{p}\|^2 = \lambda \|\vec{p}\|^2 \Rightarrow \lambda = \frac{\|B\vec{p}\|^2}{\|\vec{p}\|^2} \geq 0 \tag{85.3}$$

— Having established (78) – (85), we can state PCA:

→ Eigendecompose $S = P \Lambda P^{-1}$ (86)



→ since the eigenvectors are orthonormal (83), we have

$P P^T = P^T P = I$, i.e, $P^{-1} = P^T$, aka P is unitary (87)

$\Rightarrow S = P \Lambda P^T$ (88)

with $\lambda_1, \lambda_2, \cdots, \lambda_m \geq 0$ (from (85)) (89)

(i.e, the eigenvectors $\vec{p}_i$ of the co-variance matrix S)

→ The columns of P are called the principal components (90)

(of S or $\tilde{A}$ or A)

— the principal components have unit norm and are orthogonal to each other (due to (87)) (91)

→ An important property of the principal components is that if you project the (rows of the) mean-centered data $\tilde{A}$ onto them, the projected data becomes un-correlated, i.e., the covariance matrix of the projected data is diagonal — in fact, the covariance matrix is simply $= \Lambda$ (in (88)): (92)

— Proof: the projected rows of $\tilde{A}$ onto P are: $F \triangleq \underset{n\times m}{\tilde{A}} \underset{m\times m}{P}$ (92.1)

— the i^{th} column of F ($i = 1, \cdots, m$) consists of the projection of every row of $\tilde{A}$ onto the i^{th} principal component $\vec{p}_i$. (92.2)

— note that F is mean-centered, since

$\underbrace{[1, 1, 1, \cdots, 1]}_{n} F = \underbrace{[1, 1, 1, \cdots, 1]}_{n} \tilde{A} P = \vec{0}$, $\because \tilde{A}$ is mean-centered (from (55.2)) (92.3)

– therefore the covariance matrix of F is:

$$\rightarrow S_F = \frac{1}{n} F^T F = P^T \frac{\tilde{A}^T \tilde{A}}{n} P = P^T S P \quad (92.4)$$

→ using (88), we get $S_F = \underbrace{P^T P}_{I} \Lambda \underbrace{P^T P}_{I} = \Lambda$ (92.5)

→ hence, using (62.8) and (92.5), the variance of the i^{th} column of the projected data F is simply λ_i. (93)

→ Another key property of PCA is this: (94)

→ pick any unit-norm $\vec{p} \in \mathbb{R}^m$ (94.05)

→ project the mean-centered data $\tilde{A}$ onto $\vec{p}$: $(\underset{n \times m}{\tilde{A}} \vec{p})$ (94.1)

↳ size n vector, zero-mean

→ find the variance of the projected data: $\frac{1}{n} (\tilde{A}\vec{p})^T (\tilde{A}\vec{p})$ (94.2)

→ of all unit-norm $\vec{p}$, the one that maximizes this variance is $\vec{p}_1$, i.e, the principal component corresponding to the largest eigenvalue, λ_1; and, the maximum variance is simply λ_1 (94.3)

– proof: first, assume that $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_m \geq 0$ (94.35)

→ denote by σ_p^2 the variance in (94.2), i.e,

$$\rightarrow \sigma_p^2 \triangleq \frac{1}{n} \left(\tilde{A}\vec{p}\right)^T \left(\tilde{A}\vec{p}\right) = \vec{p}^T \frac{\tilde{A}^T \tilde{A}}{n} \vec{p} = \vec{p}^T S \vec{p} = \vec{p}^T P \Lambda P^T \vec{p}$$

$$= (P^T \vec{p})^T \Lambda (P^T \vec{p}) \quad (94.4)$$

→ express $\vec{p}$ in the basis of eigenvectors $\vec{p}$:

$$\vec{p} = \alpha_1 \vec{p}_1 + \alpha_2 \vec{p}_2 + \cdots + \alpha_m \vec{p}_m, \text{ or } \vec{p} = P\vec{\alpha}, \quad (94.5)$$

where $\vec{\alpha} \triangleq \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}$, with (94.6)

$$\alpha_i = \vec{p}_i^T \vec{p}, \text{ or } \vec{\alpha} = P^T \vec{p} \quad (94.7)$$

→ then, (94.4) becomes

$$\sigma_p^2 = \vec{\alpha}^T \Lambda \vec{\alpha} = \sum_{i=1}^{m} \alpha_i^2 \lambda_i \qquad (94.8)$$

→ note that in (94.7), $P(P^T$ is orthonormal, hence (15.1) applies, and we have $\|\vec{\alpha}\| = \|\vec{p}\| = 1$ (using (94.05)) (94.9)

→ i.e., $\sum_{i=1}^{m} \alpha_i^2 = 1 \Rightarrow \alpha_1^2 = 1 - \sum_{i=2}^{m} \alpha_i^2$ (94.10)

→ using (94.10) in (94.8), we get:

$$\sigma_p^2 = \alpha_1^2 \lambda_1 + \sum_{i=2}^{m} \alpha_i^2 \lambda_i = \lambda_1 \left(1 - \sum_{i=2}^{m} \alpha_i^2\right) + \sum_{i=2}^{m} \alpha_i^2 \lambda_i$$

$$\Rightarrow \sigma_p^2 = \lambda_1 + \sum_{i=2}^{m} \alpha_i^2 (\lambda_i - \lambda_1) \qquad (94.11)$$

→ now, we would like to <u>maximize</u> σ_p^2 in (94.11) over α_i

→ hence, $\frac{\partial \sigma_p^2}{\partial \alpha_i}$ must be 0 for $i = 2, \cdots, m$, (94.12)

→ AND: $\frac{\partial^2 \sigma_p^2}{\partial \alpha_i^2}$ must be ≤ 0, for $i = 2, \cdots, m$ (94.13)
↳ for it to be a <u>maximum</u>

→ (94.12) yields $2 \alpha_i (\lambda_i - \lambda_1) = 0$, $i = 2, \ldots, m$ (94.14)

$\Rightarrow \alpha_i = 0$, $i = 2, \cdots, m$, satisfies (94.14) (94.15)

$\Rightarrow \alpha_1 = 1$, $\alpha_2 = 0$, $\alpha_3 = 0$, $\cdots$ $\alpha_m = 0$ (94.16)
↳ using (94.10)

→ (94.13) yields: $2(\lambda_i - \lambda_1) \leq 0$ for $i = 2, \cdots, m$ (94.17)

– which is true in view of (94.35)

– hence the solution (94.16) is a <u>maximum</u> (94.18)

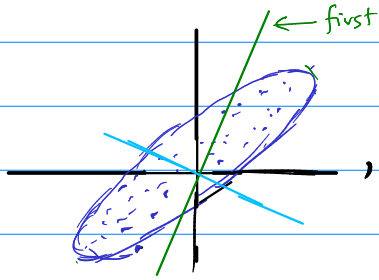
→ from (94.16), we have:

→ $\vec{\alpha} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, $\vec{p} = \vec{p}_1$ (94.5), and $\sigma_p = \lambda_1$ (94.8) (94.19)

— The result (94) is important because it tells us that the principal component (or axis) $\vec{p}_1$ is the direction of maximum data variance (or spread). (95)

— which is exactly what is depicted graphically in (75)

— it also rules out the possibility of situations like



, because the axis is not along the direction of maximum spread. (96)

→ Similar to (94), it can also be shown that if $\vec{p}$ in (94.05) is restricted to being orthogonal to $\vec{p}_1$, then $\vec{p} = \vec{p}_2$ maximizes the variance of $\tilde{A}$ projected onto $\vec{p}$. (97)

— and $\vec{p}_3$ does the same if $\vec{p}$ is restricted to being orthogonal to both $\vec{p}_1$ & $\vec{p}_2$; and similarly for $\vec{p}_4$, $\vec{p}_5$, etc.

THE CONNECTION BETWEEN PCA and the SVD

— Suppose we have an SVD of $\tilde{A}$: $\tilde{A} = U \Sigma V^T$ (98)

($\tilde{A}$: $n \times m$, U: $n \times n$, Σ: $n \times m$, V^T: $m \times m$)

→ the co-variance matrix $S = \frac{1}{n} \tilde{A}^T \tilde{A} = \frac{1}{n} V \Sigma^T U^T U \Sigma V^T$

$$= V \left(\frac{\Sigma^T \Sigma}{n} \right) V^T \quad (99)$$

(V: unitary/orthonormal; $\frac{\Sigma^T \Sigma}{n}$: $m \times m$ diagonal)

→ comparing (99) with (83), we see that if we set

$$P \equiv V, \quad \Lambda \equiv \frac{\Sigma^T \Sigma}{n}, \quad (100)$$

the SVD yields a PCA of $\tilde{A}$

→ we don't even need to form the covariance matrix S to get (101)
V and Σ from the SVD

→ from (100), the PCA variances λ_i are given in terms of the singular values by:

$$\lambda_i = \frac{\sigma_i^2}{n} \quad (102)$$

COMPUTING SVDs USING EIGENDECOMPOSITION

— (99) and (100) also immediately suggest that if we have an eigendecomposition of a matrix of the form B^TB, then then eigenvalues and eigenvectors can provide the SVD of B. (103)

— Suppose we are given some $n \times m$ matrix B, whose SVD we wish to calculate.

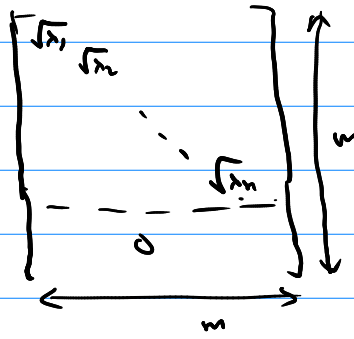
→ let us form $S \overset{\text{nxm}}{\triangleq} B^TB$ (no need to mean-center or divide by n) (104)

→ then eigendecompose $S = P \Lambda P^T$ (105)

— note that $PP^T = I$ (i.e, orthonormal) due to (83) (106)

— and that λ_i (diagonal entries of Λ) are all ≥ 0 (due to (84)) (107)

→ Define $\Sigma =$



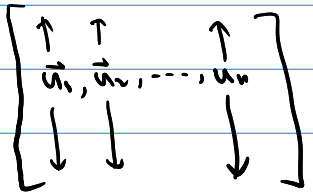
or $\sigma_i \triangleq \sqrt{\lambda_i}$ (108)

→ Define $\vec{v}_i = \vec{p}_i$, $i = 1, \cdots, m$, $\Rightarrow V^T = P^T$, with $V^TV = I$ (109)

→ Define $\vec{u}_i \triangleq \dfrac{B\vec{v}_i}{\sigma_i}$, $i = 1, \cdots, m$ (110)

→ for $i = m+1, \cdots, n$, choose <u>any</u> $\vec{u}_i$ that complete the orthonormal basis $\{\vec{u}_1, \vec{u}_2, \cdots, \vec{u}_n\}$ (111)

→ Define $U \triangleq$



(112)

→ Then $B = U \Sigma V^T$ is an SVD of B (113)

→ proof:

– we already have V is unitary and Σ has diagonal elements that are real and ≥ 0 (using (109) & (108)) (114)

– from (110), we have $\|\vec{u}_i\|^2 = \dfrac{\vec{v}_i^T B^T B \vec{v}_i}{\sigma_i^2}$, $i = 1, \dots, m$

$$= \frac{\vec{p}_i^T P \Lambda P^T \vec{p}_i}{\lambda_i} \quad \text{(from (104),(105),(108))} \qquad (115)$$

→ note that $P^T \vec{p}_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i^{th}$ (116)

→ hence (115) becomes

$$\|\vec{u}_i\|^2 = \frac{[0, 0, \dots, \underset{\uparrow\, i^{th}}{1}, 0, 0, 0]\, \Lambda\, [0, 0, \dots, \underset{\uparrow\, i^{th}}{1}, 0, 0, 0]^T}{\lambda_i}$$

$$= \frac{\lambda_i}{\lambda_i} = 1 \qquad (117)$$

→ also from (110),

$$\vec{u}_i^T \vec{u}_j = \frac{v_i^T B^T B v_j}{\sigma_i \sigma_j} = \vec{v}_i^T P \Lambda P^{-1} \vec{v}_j$$

$$= \frac{1}{\sigma_i \sigma_j} [0, \dots, \underset{\uparrow\, i^{th}}{1}, 0, \dots 0] \begin{bmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_i & & \\ & & & \lambda_j & \\ & & & & \lambda_m \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix} \leftarrow j^{th}$$

$$= \frac{1}{\sigma_i \sigma_j} [0, 0, \dots \underset{\downarrow\, i^{th}}{\lambda_i}, 0, 0 \dots 0] \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix} \leftarrow j^{th}$$

$$= 0 \text{ if } i \neq j, \quad i, j \in 1, \dots, m \qquad (118)$$

→ therefore, the first m columns of U are orthonormal

→ since the remaining cols are chosen (per (111)) to complete the orthonormal basis, we have $U^T U = I$ (119)

→ With (119) and (114), U, Σ and V^T satisfy all the criteria for an SVD. (120)

→ Now, $U\Sigma' = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \sigma_1 \vec{u}_1 & \sigma_2 \vec{u}_2 & \cdots & \sigma_m \vec{u}_m \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$ (121)

$\Rightarrow U\Sigma' = (\text{using (110)}) \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ B\vec{v}_1 & B\vec{v}_2 & \cdots & B\vec{v}_m \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = BV$ (122)

→ since $VV^T = I$ (from (109) & (106)), we have

$$- U\Sigma' V^T = BVV^T = B \quad (123)$$