FINM 331: MULTIVARIATE DATA ANALYSIS
FALL 2021
MATRIX THEORY BACKGROUND

1. SINGULAR VALUE DECOMPOSITION

- let $A \in \mathbb{R}^{n\times p}$, we can always write
$$A = U\Sigma V^\intercal \tag{1.1}$$
  - $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are both orthogonal matrices
$$U^\intercal U = I_n = UU^\intercal, \quad V^\intercal V = I_p = VV^\intercal$$
  - $\Sigma \in \mathbb{R}^{n\times p}$ is a diagonal matrix in the sense that $\sigma_{ij} = 0$ if $i \neq j$
  - if $n > p$, then Σ looks like
$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_p \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$$
  - if $n < p$, then Σ looks like
$$\Sigma = \begin{bmatrix} \sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & & \vdots \\ & & \sigma_n & 0 & \cdots & 0 \end{bmatrix}$$
  - if $n = p$, then Σ looks like
$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}$$
  - the diagonal elements of Σ, denoted σ_i, $i = 1, \ldots, p$, are all nonnegative, and can be ordered such that
$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \quad \sigma_{r+1} = \cdots = \sigma_{\min(n,p)} = 0$$
  - r is the rank of A
- this decomposition of A is called the *singular value decomposition*, or SVD
  - the values σ_i, for $i = 1, 2, \ldots, p$, are the *singular values* of A
  - the columns of U are the *left singular vectors*
  - the columns of V are the *right singular vectors*
- an alternative decomposition of A omits the singular values that are equal to zero:
$$A = U_r\Sigma_r V_r^\intercal$$
  - $U_r \in \mathbb{R}^{n\times r}$ is a matrix with orthonormal columns, i.e. satisfying $U_r^\intercal U_r = I_r$ (but not $U_r U_r^\intercal = I_n$!)

Date: October 6, 2021, version 1.0. Comments, bug reports: `lekheng@uchicago.edu`.

  - $V_r \in \mathbb{R}^{p\times r}$ is also a matrix with orthonormal columns, i.e. satisfying $V_r^\mathsf{T} V_r = I_r$ (but again not $V_r V_r^\mathsf{T} = I_p$!)
  - Σ_r is an $r \times r$ diagonal matrix with diagonal elements $\sigma_1, \ldots, \sigma_r$
  - again $r = \operatorname{rank}(A)$
  - the columns of U_r are the left singular vectors corresponding to the nonzero singular values of A, and form an orthonormal basis for the range of A
  - the columns of V_r are the right singular vectors corresponding to the nonzero singular values of A, and form an orthonormal basis for the cokernel of A
  - this is called the *condensed* or *compact* or *reduced* SVD
  - note that in this case, Σ_r is a square matrix
- the form in (1.1) is sometimes called the *full* SVD
- we may also write the reduced SVD as a sum of rank-1 matrices

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\mathsf{T} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\mathsf{T} + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r^\mathsf{T}$$

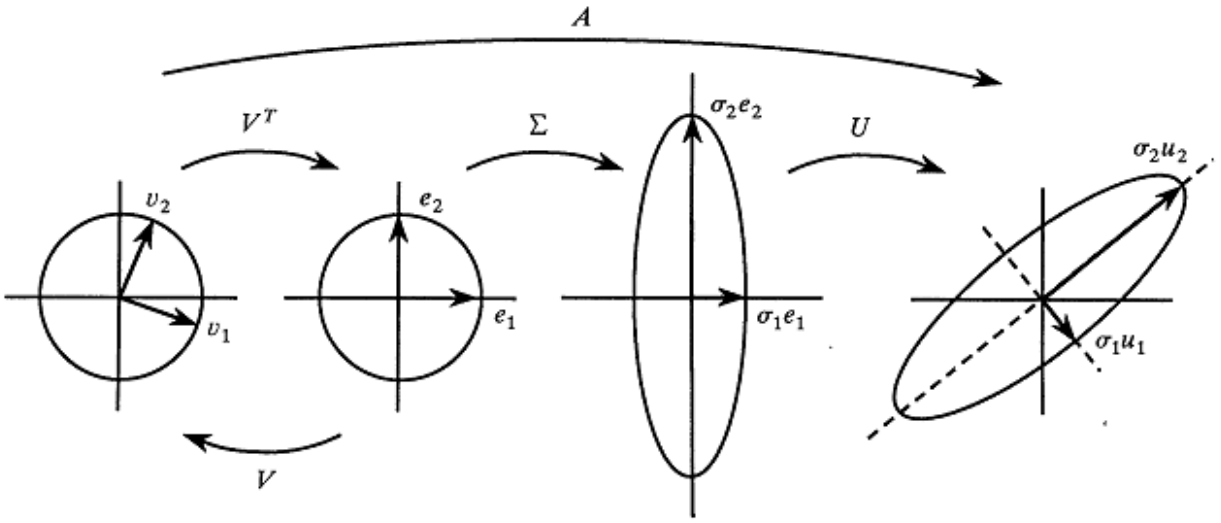
  - $U_r = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$, i.e. $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{R}^n$ are the left singular vectors of A
  - $V_r = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$, i.e. $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^p$ are the right singular vectors of A
  - note that for $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^n$ and $\mathbf{y} = [y_1, \ldots, y_p]^\mathsf{T} \in \mathbb{R}^p$,

$$\mathbf{x}\mathbf{y}^\mathsf{T} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_p \\ x_2y_1 & x_2y_2 & \cdots & x_2y_p \\ \vdots & \vdots & & \vdots \\ x_ny_1 & x_ny_2 & \cdots & x_ny_p \end{bmatrix}$$

  - if neither $\mathbf{x}$ nor $\mathbf{y}$ is the zero vector, then

$$\operatorname{rank}(\mathbf{x}\mathbf{y}^\mathsf{T}) = \operatorname{rank}(\mathbf{x}\mathbf{y}^\mathsf{T}) = 1$$

  - furthermore if $\operatorname{rank}(A) = 1$, then there exists $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^p$ so that $A = \mathbf{x}\mathbf{y}^\mathsf{T}$
- a common way to picture that SVD is that it decomposes any matrix A into a rotation V^T followed a scaling Σ, followed by another rotation U, i.e., the action $x \mapsto Ax$ can always be broken into a rotation $x \mapsto V^\mathsf{T}x =: y$, a scaling $y \mapsto \Sigma y =: z$, and another rotation $z \mapsto Uz$



2. PROPERTIES OF SVD

- the fact that any matrix has an SVD may found in the appendix
- the proof gives us alternative characterizations of singular values and singular vectors:

(i) in terms of eigenvalues and eigenvectors of an $(n+p)\times(n+p)$ symmetric matrix:

$$\begin{bmatrix} 0 & A \\ A^\intercal & 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} U & U \\ V & -V \end{bmatrix}\begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}\begin{bmatrix} U^\intercal & V^\intercal \\ U^\intercal & -V^\intercal \end{bmatrix}$$

(ii) in terms of a coupled system of equations

$$\begin{cases} A\mathbf{v} = \sigma\mathbf{u}, \\ A^\intercal\mathbf{u} = \sigma\mathbf{v} \end{cases}$$

- the following is yet another way to characterize them in terms of eigenvalues/eigenvectors of an $n\times n$ symmetric matrix and an $p\times p$ symmetric matrix

Lemma 1. *The square of the singular values of a matrix A are eigenvalues of $AA^\intercal$ and $A^\intercal A$. The left singular vectors of A are the eigenvectors of $AA^\intercal$ and the right singular vectors of A are the eigenvectors of $A^\intercal A$.*

Proof. From the relationships $A\mathbf{y} = \sigma\mathbf{x}$, $A^\intercal\mathbf{x} = \sigma\mathbf{y}$, we obtain

$$A^\intercal A\mathbf{y} = \sigma^2\mathbf{y}, \quad AA^\intercal\mathbf{x} = \sigma^2\mathbf{x}.$$

Therefore, if $\pm\sigma$ are eigenvalues of W, then σ^2 is an eigenvalue of both $AA^\intercal$ and $A^\intercal A$. Also

$$AA^\intercal = (U\Sigma V^\intercal)(V\Sigma^\intercal U^\intercal) = U\Sigma\Sigma^\intercal U^\intercal,$$
$$A^\intercal A = (V\Sigma^\intercal U^\intercal)(U\Sigma V^\intercal) = V\Sigma^\intercal\Sigma V^\intercal.$$

Note that $\Sigma^\intercal = \Sigma^\intercal$ since singular values are real. The matrices $\Sigma^\intercal\Sigma$ and $\Sigma\Sigma^\intercal$ are respectively $p\times p$ and $n\times n$ diagonal matrices with diagonal elements σ_i^2 and 0. □

- note that the matrix $A^\intercal A$ is symmetric and positive semidefinite, i.e. $\mathbf{x}^\intercal(A^\intercal A)\mathbf{x} \geq 0$ for all nonzero $\mathbf{x}\in\mathbb{R}^p$
- exercise: show that if a matrix M is symmetric positive semidefinite, then its EVD and SVD coincide
- as such, $A^\intercal A$ has EVD/SVD given by

$$A^\intercal A = V\Sigma V^\intercal$$

where V is a orthogonal matrix whose columns are the eigenvectors of $A^\intercal A$, and Σ is a diagonal matrix of the form

$$\begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_p^2 \end{bmatrix}$$

where each σ_i^2 is nonnegative and an eigenvalue of $A^\intercal A$

- the discussions above hold similarly for $AA^\intercal$
- the SVD is intimately related to the Frobenius norm
- exercise: show that the Frobenius norm is orthogonally invariant, i.e.,

$$\|UXV\|_F = \|X\|_F$$

for any orthogonal matrices $U\in\mathbb{R}^{n\times n}$, $V\in\mathbb{R}^{p\times p}$, and any matrix $X\in\mathbb{R}^{n\times p}$

- this yields an expression in terms of singular values

$$\|A\|_F = \|U\Sigma V^\intercal\|_F = \|\Sigma\|_F = \sqrt{\sigma_1^2+\cdots+\sigma_r^2}$$

where $r = \operatorname{rank}(A)$

- the SVD is something like a swiss army knife of linear algebra, matrix theory, and numerical linear algebra, you can do a lot with it
- in this course we will primarily be interested in two uses of the SVD

(i) computing projections
(ii) computing best rank-r approximations

- the good news is that unlike the Jordan canonical form, the SVD is actually computable
- there are two main methods to compute it: Golub–Reinsch and Golub–Kahan, we will look at these briefly later, right now all you need to know is that you can call MATLAB to give you the SVD, both the full and compact versions

3. basis and orthonormal basis

- we will review the notions of *basis* and *orthonormal basis*
- let $W \subseteq \mathbb{R}^p$ be a subspace
- a *basis* of W is a set of vectors $\mathbf{b}_1, \dots, \mathbf{b}_k$ such that every $\mathbf{w} \in W$ can be uniquely expressed

$$\mathbf{w} = \alpha_1 \mathbf{b}_1 + \cdots + \alpha_k \mathbf{b}_k \tag{3.1}$$

- you can show that $\mathbf{b}_1, \dots, \mathbf{b}_k$ is a basis if and only if they are linearly independent and $\operatorname{span}\{\mathbf{b}_1, \dots, \mathbf{b}_k\} = W$
- in particular $k = \dim W$
- an *orthonormal basis* $\mathbf{q}_1, \dots, \mathbf{q}_k$ of W is a basis with the additional property that the vectors are orthonormal, i.e.,

$$\mathbf{q}_i^{\mathsf{T}} \mathbf{q}_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \end{cases}$$

- why is an orthonormal basis nice?
- because the coefficients $\alpha_1, \dots, \alpha_k$ in (3.1) can be easily determined: any $\mathbf{w} \in W$ can be expressed as

$$\mathbf{w} = (\mathbf{w}^{\mathsf{T}} \mathbf{q}_1)\mathbf{q}_1 + (\mathbf{w}^{\mathsf{T}} \mathbf{q}_2)\mathbf{q}_2 + \cdots + (\mathbf{w}^{\mathsf{T}} \mathbf{q}_k)\mathbf{q}_k$$

- in other words, we always have

$$\alpha_i = \mathbf{w}^{\mathsf{T}} \mathbf{q}_i, \quad i = 1, \dots, k,$$

whenever $\mathbf{q}_1, \dots, \mathbf{q}_k$ is an orthonormal basis

4. projection matrices

- in general a matrix $P \in \mathbb{R}^{p \times p}$ is called a *projection* if $P^2 = P$ (this condition is also called idempotent in ring theory)
- if $P \in \mathbb{R}^{p \times p}$ is a projection and $\operatorname{im}(P) = W$, we say that P is a projection onto the subspace W
- a projection is called an orthogonal projection if it is also symmetric, i.e. an orthogonal projection is a matrix $P \in \mathbb{R}^{p \times p}$ satisfying
  (i) $P = P^{\mathsf{T}}$
  (ii) $P^2 = P$
- caveat: an orthogonal projection is in general *not* an orthogonal/orthogonal matrix (i.e., $P^{\mathsf{T}} \neq P^{-1}$) in fact, projections are usually non-invertible
- for example, the matrix

$$P_\alpha = \begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix}$$

is a projection for any $\alpha \in \mathbb{R}$, it projects a vector $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2$ onto the vector

$$\begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x + \alpha y \\ 0 \end{bmatrix} \in W \subseteq \mathbb{R}^2$$

where W is the subspace $\{[x, 0]^{\mathsf{T}} \in \mathbb{R}^2 : x \in \mathbb{R}\}$, i.e., the x-axis

- P_α is an orthogonal projection if and only if $\alpha = 0$ and in which case it projects a vector $\begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2$ onto the vector

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix} \in W \subseteq \mathbb{R}^2$$

- if $P \in \mathbb{R}^{p \times p}$ is a projection matrix, then $I - P$ is also a projection
- furthermore if $\operatorname{im}(P) = W$ and $\operatorname{im}(I - P) = W'$, then

$$\mathbb{R}^p = W \oplus W'$$

- if P is an orthogonal projection and $\operatorname{im}(P) = W$, then $\operatorname{im}(I - P) = W^\perp$
- we sometimes write P_W if we know the subspace P that projects onto

5. COMPUTING PROJECTIONS

- in practice what we given is *not* a projection matrix $P \in \mathbb{R}^{p \times p}$ but a subspace $W \subseteq \mathbb{R}^p$ that we need to project points in $\mathbb{R}^p$ onto
- so this is what we need to learn to do
- what does it mean to "give you a subspace W"?
- usually that means to give you a *basis* or an *orthonormal basis* of W
- suppose you are given W in the form of

$$W = \operatorname{span}\{\mathbf{q}_1, \dots, \mathbf{q}_k\}$$

  where $\mathbf{q}_1, \dots, \mathbf{q}_k$ is an orthonormal basis
- then it is easy write down a projection of $\mathbf{x} \in \mathbb{R}^p$ onto W — it is simply the vector

$$(\mathbf{x}^\mathsf{T}\mathbf{q}_1)\mathbf{q}_1 + (\mathbf{x}^\mathsf{T}\mathbf{q}_2)\mathbf{q}_2 + \cdots + (\mathbf{x}^\mathsf{T}\mathbf{q}_p)\mathbf{q}_p \in W$$

- this actually gives us the projection matrix $P_W \in \mathbb{R}^{p \times p}$ too
- first observe that

$$(\mathbf{x}^\mathsf{T}\mathbf{q}_i)\mathbf{q}_i = (\mathbf{q}_i^\mathsf{T}\mathbf{x})\mathbf{q}_i = \mathbf{q}_i(\mathbf{q}_i^\mathsf{T}\mathbf{x}) = (\mathbf{q}_i\mathbf{q}_i^\mathsf{T})\mathbf{x}$$

- therefore we may write

$$\begin{aligned} P_W\mathbf{x} &= (\mathbf{x}^\mathsf{T}\mathbf{q}_1)\mathbf{q}_1 + (\mathbf{x}^\mathsf{T}\mathbf{q}_2)\mathbf{q}_2 + \cdots + (\mathbf{x}^\mathsf{T}\mathbf{q}_k)\mathbf{q}_k \\ &= (\mathbf{q}_1\mathbf{q}_1^\mathsf{T})\mathbf{x} + (\mathbf{q}_2\mathbf{q}_2^\mathsf{T})\mathbf{x} + \cdots + (\mathbf{q}_k\mathbf{q}_k^\mathsf{T})\mathbf{x} \\ &= \big(\mathbf{q}_1\mathbf{q}_1^\mathsf{T} + \mathbf{q}_2(\mathbf{q}_2^\mathsf{T} + \cdots + \mathbf{q}_k\mathbf{q}_k^\mathsf{T})\mathbf{x} \end{aligned}$$

  which means that

$$P_W = \mathbf{q}_1\mathbf{q}_1^\mathsf{T} + \mathbf{q}_2(\mathbf{q}_2^\mathsf{T} + \cdots + \mathbf{q}_k\mathbf{q}_k^\mathsf{T}$$

  as a sum of rank-one matrices or alternatively

$$P_W = Q_k Q_k^\mathsf{T}$$

  where

$$Q_k \coloneqq [\mathbf{q}_1, \dots, \mathbf{q}_k] \in \mathbb{R}^{p \times k}$$

- in this course are almost exclusively interested in the case $k = 2$ (sometimes $k = 3$) because this is the case that can be plotted on screen or on a piece of paper
- plotting the projection of $\mathbf{x} \in \mathbb{R}^p$ onto $W = \operatorname{span}\{\mathbf{p}, \mathbf{q}\}$ where $\mathbf{p}$ and $\mathbf{q}$ are orthonormal means to plot on a graph where the x-axis is in the direction of $\mathbf{p}$ and the y-axis is in the direction of $\mathbf{q}$
- in other words, we plot the point with coordinates

$$(\mathbf{x}^\mathsf{T}\mathbf{p}, \mathbf{x}^\mathsf{T}\mathbf{q}) \in \mathbb{R}^2$$

- more generally, to plot the projections of n points, $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$, we simply plot the n points
$$\{(\mathbf{x}_i^\mathsf{T}\mathbf{p}, \mathbf{x}_i^\mathsf{T}\mathbf{q}) \in \mathbb{R}^2 : i = 1, \dots, n\}$$
on a plane, i.e., a 2-dimensional graph whose x-axis is labelled $\mathbf{p}$ and y-axis is labelled $\mathbf{q}$
- we will have more to say about this and how SVD is involved when we discuss principal components analysis

6. matrix approximation problems

- remember that in multivariate data analysis, data is given to us in the form of a matrix $X \in \mathbb{R}^{n\times p}$
- sometimes it is enough to simply treat X as a matrix and immediately start to analyze it — this often involve some *matrix approximation* or *matrix nearness problems*
- we will discuss three:
  - finding nearest orthogonal matrix: given $X \in \mathbb{R}^{p\times p}$, solve
$$\min_{Y^\mathsf{T}Y=I} \|X - Y\|_F$$
  - finding nearest symmetric matrix: given $X \in \mathbb{R}^{p\times p}$, solve
$$\min_{Y^\mathsf{T}=Y} \|X - Y\|_F$$
  - finding nearest rank-r matrix: given $X \in \mathbb{R}^{n\times p}$, solve
$$\min_{\operatorname{rank}(Y)\le r} \|X - Y\|_F$$
- these are often used to preprocess data — for example, if we know the matrix we are given ought to have certain properties (e.g., covariance and correlation matrices must be symmetric) but do not because of measurement or rounding errors, these techniques allow us to transform the data matrix X into a form with the required property
- more sophisticated techniques will require some statistics, i.e., where we view X as observations of some random variables
- we start with our first, and arguably the simplest, multivariate analysis tool

7. nearest orthogonal matrix

- given $X \in \mathbb{R}^{p\times p}$, we want to find the nearest orthogonal matrix Y to it
$$\min_{Y^\mathsf{T}Y=I} \|X - Y\|_F$$
- note that this problem only makes sense if X is a square matrix
- we claim that if $X = U\Sigma V^\mathsf{T}$ is the SVD of X, then the solution is given by
$$Y = UV^\mathsf{T}$$
- first note that
$$\|X - Y\|_F^2 = \operatorname{tr}(X^\mathsf{T}X) + \operatorname{tr}(Y^\mathsf{T}Y) - 2\operatorname{tr}(Y^\mathsf{T}X)$$
- so minimizing $\|X - Y\|_F^2$ is equivalent to maximizing $\operatorname{tr}(Y^\mathsf{T}X)$
- let $X = U\Sigma V^\mathsf{T}$ be the SVD of X
- then writing $Z = V^\mathsf{T}Y^\mathsf{T}U$, we get
$$\operatorname{tr}(Y^\mathsf{T}X) = \operatorname{tr}(Y^\mathsf{T}U\Sigma V^\mathsf{T}) = \operatorname{tr}(Z\Sigma) = \sum_{i=1}^p z_{ii}\sigma_i \le \sum_{i=1}^p \sigma_i$$
where the last inequality follows since Z is an orthogonal matrix and so $z_{ii} \le 1$

- the upper bound is attained by making $Z = I$, i.e.,
$$Y = UV^\mathsf{T}$$
- observe also that
$$\|X - Y\|_F^2 = \|U(\Sigma - I)V^\mathsf{T}\|_F^2 = \|\Sigma - I\|_F^2 = (\sigma_1 - 1)^2 + \cdots + (\sigma_n - 1)^2$$
- later on we look at a more sophisticated variant of this problem
$$\min_{Y^\mathsf{T}Y = I} \|X - AY\|_F$$
with two given matrices $X, A \in \mathbb{R}^{n\times p}$, which need not be square matrices

8. nearest symmetric matrix

- other problems of this nature include finding a closest symmetric matrix to a given matrix $X \in \mathbb{R}^{p\times p}$
$$\min_{Y^\mathsf{T} = Y} \|X - Y\|_F \tag{8.1}$$
- note that any square matrix can be written as a sum of a symmetric matrix and a skew-symmetric matrix
$$X = \frac{1}{2}(X + X^\mathsf{T}) + \frac{1}{2}(X - X^\mathsf{T})$$
- the solution to (8.1) is given by $Y = \frac{1}{2}(X + X^\mathsf{T})$
- you will be asked to prove this in the homework
- there is also a more sophisticated variant of this problem
$$\min_{Y^\mathsf{T} = Y} \|X - AY\|_F$$
with two given matrices $X, A \in \mathbb{R}^{n\times p}$, which need not be square matrices
- the solution in this case is more complicated and will not be discussed in this course

9. nearest rank-r matrix

- given $X \in \mathbb{R}^{n\times p}$, we want to find $Y \in \mathbb{R}^{n\times p}$ of rank not more than r so that $\|X - Y\|_F$ is minimized
- in notations, we want
$$\min_{\operatorname{rank}(Y)\le r} \|X - Y\|_F \tag{9.1}$$
- such an Y is called a best rank-r approximation to X
- if $r \ge \operatorname{rank}(X)$, then clearly $Y = X$ and the problem is trivial, so we shall always assume that $r < \operatorname{rank}(X)$
- the proof of the next result may be found in the appendix

Theorem 1 (Eckart–Young)**.** *Let the* SVD *of X be*
$$X = \sum_{i=1}^{\operatorname{rank}(X)} \sigma_i \mathbf{u}_i \mathbf{v}_i^\mathsf{T}, \quad \sigma_1 \ge \cdots \ge \sigma_r > 0.$$
Then for any $r \in \{1, \ldots, \operatorname{rank}(X) - 1\}$, a solution to (9.1) *is given by*
$$Y = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\mathsf{T}.$$
Furthermore, we have
$$\min_{\operatorname{rank}(Y)\le r} \|X - Y\|_F = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_{\operatorname{rank}(X)}^2} \tag{9.2}$$

In matrix form, we have

$$Y = U \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} V^{\mathsf{T}}, \tag{9.3}$$

where $X = U\Sigma V^{\mathsf{T}}$ *is the* SVD *of* X.

10. Procrustes analysis

- the problem of finding a nearest orthogonal matrix is a special case of a well-known multivariate data analysis technique called *orthogonal Procrustes analysis*
- given two data matrices X and $A \in \mathbb{R}^{n\times p}$, we want to 'rotate' one to other so that the columns of X and A match up as much as possible
- orthogonal matrices may be thought of as rotations and reflections
- note that unlike the nearest orthogonal matrix problem, the matrices X and A can be rectangular
- the Procrustes problem asks to find $Y \in \mathbb{R}^{p\times p}$ such that

$$\min_{Y^{\mathsf{T}}Y=I} \|X - AY\|_F$$

- the proof is very similar to the case $A = I$ and can be found in the appendix
- we have the following algorithm

Algorithm: Orthogonal Procrustes Analysis
INPUT: $X, A \in \mathbb{R}^{n\times p}$
STEP 1: compute $C \leftarrow A^{\mathsf{T}}X$;
STEP 2: compute left and right singular vectors of $C \to (U, V)$;
OUTPUT: $Y \leftarrow UV^{\mathsf{T}}$

- for example suppose we want to rotate (more accurately, to orthogonally transform) the matrix A to X where

$$X = \begin{bmatrix} 1.2 & 2.1 \\ 2.9 & 4.3 \\ 5.2 & 6.1 \\ 6.8 & 8.1 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix},$$

  then the optimal orthogonal matrix is

$$Y = \begin{bmatrix} 0.9999 & 0.0126 \\ -0.0126 & 0.9999 \end{bmatrix}$$

  which minimizes $\|X - AY\|_F^2$
- exercise: what if we want to orthogonally transform X to A instead?
- more generally, Procrustes analysis allows for not just orthogonal transformation but also translation and scaling of A to make it as close to X as possible
- these are usually done separately because it is computationally very difficult to do all three operations jointly (NP-hard)

- for translation, we just mean center our two data matrices, i.e., apply the following operations to both X and A

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \to \begin{bmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \cdots & x_{1p} - \overline{x}_p \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & \cdots & x_{2p} - \overline{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 & \cdots & x_{np} - \overline{x}_p \end{bmatrix}$$

  where

$$\overline{x}_j = \frac{x_{1j} + x_{2j} + \cdots + x_{nj}}{n}, \quad j = 1, \ldots, p$$

- note that this creates a matrix where each column has mean 0
- for scaling, we scale our data matrices by the standard deviation, i.e., apply the following operations to both X and A

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \to \begin{bmatrix} x_{11}/s_1 & x_{12}/s_2 & \cdots & x_{1p}/s_p \\ x_{21}/s_1 & x_{22}/s_2 & \cdots & x_{2p}/s_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}/s_1 & x_{n2}/s_2 & \cdots & x_{np}/s_p \end{bmatrix}$$

  where

$$s_j = \left[\frac{1}{n} \sum\nolimits_{i=1}^{n} (x_{ij} - \overline{x}_j)^2 \right]^{1/2}, \quad j = 1, \ldots, p$$

- further reading: Section 12.9 in Johnson–Wichern, Section 14.7 in Mardia–Kent–Bibby