FINM 331: DATA ANALYSIS FOR FINANCE AND STATISTICS
FALL 2015
MATRIX APPROXIMATIONS, MEAN AND VARIANCE

1. Procrustes analysis

- we start with our first, and arguably the simplest, multivariate analysis tool
- given data matrices A and $B \in \mathbb{R}^{n \times p}$, we want to 'rotate' one to other so that the columns of A and B match up as much as possible
- to be precise, let $O(p)$ be the set of all $p \times p$ orthogonal matrices, note that these include both rotation and reflection matrices
- the Procrustes problem asks to find $Q \in O(p)$ such that

$$\min_{Q \in O(p)} \|A - BQ\|_F$$

- note that

$$\|A - BQ\|_F^2 = \operatorname{tr}(A^\mathsf{T} A) + \operatorname{tr}(B^\mathsf{T} B) - 2\operatorname{tr}(Q^\mathsf{T} B^\mathsf{T} A)$$

- so minimizing $\|A - BQ\|_F^2$ is equivalent to maximizing $\operatorname{tr}(Q^\mathsf{T} B^\mathsf{T} A)$
- let $B^\mathsf{T} A = U \Sigma V^\mathsf{T}$ be the SVD of $B^\mathsf{T} A$
- then writing $Z = V^\mathsf{T} Q^\mathsf{T} U$, we get

$$\operatorname{tr}(X^\mathsf{T} B^\mathsf{T} A) = \operatorname{tr}(X^\mathsf{T} U \Sigma V^\mathsf{T}) = \operatorname{tr}(Z\Sigma) = \sum_{i=1}^{p} z_{ii}\sigma_i \le \sum_{i=1}^{p} \sigma_i$$

  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.,

$$Q = UV^\mathsf{T}$$

- we have the following algorithm

  Algorithm: Orthogonal Procrustes Analysis
  Input: $A, B \in \mathbb{R}^{n \times p}$
  Step 1: compute $C \leftarrow B^\mathsf{T} A$;
  Step 2: compute left and right singular vectors of $C \to (U, V)$;
  Output: $Q \leftarrow UV^\mathsf{T}$

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

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

  then the optimal orthogonal matrix is

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

  which minimizes $\|A - BQ\|_F^2$
- exercise: what if we want orthogonally transform A to B instead?

Date: October 21, 2015, version 1.0. Comments, bug reports: `lekheng@galton.uchicago.edu`.

- a special case is to find the nearest orthogonal matrix Q to a given matrix $A \in \mathbb{R}^{n\times n}$

$$\min_{Q\in O(n)} \|A - Q\|_F$$

- if $A = U\Sigma V^{\mathsf{T}}$ is the SVD of A and

$$Q = UV^{\mathsf{T}},$$

  then

$$\|A - Q\|_F^2 = \|U(\Sigma - I)V^{\mathsf{T}}\|_F^2 = \|\Sigma - I\|_F^2 = (\sigma_1 - 1)^2 + \cdots + (\sigma_n - 1)^2$$

- other problems of this nature include finding a closest symmetric matrix to a given matrix $A \in \mathbb{R}^{n\times n}$

$$\min_{X^{\mathsf{T}}=X} \|A - X\|_F \tag{1.1}$$

- note that any square matrix can be written as a sum of a symmetric matrix and a skew-symmetric matrix

$$A = \frac{1}{2}(A + A^{\mathsf{T}}) + \frac{1}{2}(A - A^{\mathsf{T}})$$

- the solution to (1.1) is given by $X = \frac{1}{2}(A + A^{\mathsf{T}})$ (why?)
- more generally, Procrustes analysis allows for just orthogonal transformation but also translation and scaling of B to make it as close to A 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 A and B

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

  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 A and B

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

  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

2. sample mean

- the mean centering and scaling by standard deviation operations introduce are simple but exceptionally important in multivariate analysis
- here we will formally introduce these and other statistical terminologies that we will use throughout the course
- we will state everything in terms of matrices and vectors since these underlie all our multivariate analysis tools
- given a data matrix

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

  where n is the number of samples and p is the number of variables and

$$x_{ij} = \text{measured value of the } j\text{th variable on the } i\text{th sample}$$

- recall that these are know by various other names
  - samples = items = objects = subjects
  - variables = features = measurements = observations = outcomes = responses
- we denote the jth column vector of X as $\mathbf{x}_j \in \mathbb{R}^n$ and so we can also write

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] \in \mathbb{R}^{n \times p}$$

- we set $\mathbf{1}$ to be the 'vector of all ones,' i.e.,

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

- this vector could be in $\mathbb{R}^n$ or $\mathbb{R}^p$ depending on context
- the *sample mean* of the jth variable is the scalar

$$\overline{x}_j = \frac{x_{1j} + x_{2j} + \cdots + x_{nj}}{n} \in \mathbb{R}$$

  for $j = 1, \ldots, p$
- this can be expressed as

$$\overline{x}_j = \frac{1}{n}\mathbf{x}_j^{\mathsf{T}}\mathbf{1}$$

- the *sample mean* of X is the vector

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix} \in \mathbb{R}^p$$

- this can be expressed as

$$\overline{\mathbf{x}} = \frac{1}{n}X^{\mathsf{T}}\mathbf{1}$$

- the *matrix of means* is the rank-1 matrix

$$\mathbf{1}\overline{\mathbf{x}}^{\mathsf{T}} = \begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \end{bmatrix} \in \mathbb{R}^{n\times p}$$

- this can be expressed as a product of two matrices

$$\mathbf{1}\mathbf{x}^{\mathsf{T}} = \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}} X$$

where

$$\mathbf{1}\mathbf{1}^{\mathsf{T}} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

is often called the 'matrix of all ones'
- the *matrix of deviations* is defined as

$$X - \mathbf{1}\overline{\mathbf{x}}^{\mathsf{T}} = \begin{bmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \cdots & x_{1n} - \overline{x}_p \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & \cdots & x_{2n} - \overline{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \overline{x}_1 & x_{m2} - \overline{x}_2 & \cdots & x_{np} - \overline{x}_p \end{bmatrix} \in \mathbb{R}^{n\times p}$$

- this can be expressed as

$$X - \mathbf{1}\overline{\mathbf{x}}^{\mathsf{T}} = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}\right) X$$

where $I \in \mathbb{R}^{n\times n}$ is the $n \times n$ identity matrix
- a matrix of the form

$$I - \mathbf{x}\mathbf{y}^{\mathsf{T}}$$

is called a rank-1 perturbation of the identity matrix, in particular

$$I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}$$

is such a matrix
- *mean centering* is the operation of taking a data matrix to its matrix of deviations

$$X \to X - \mathbf{1}\overline{\mathbf{x}}^{\mathsf{T}}$$

3. Sample covariance

- the *sample variance* of the jth variable is

$$s_{jj} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2 \in \mathbb{R} \tag{3.1}$$

where $j = 1, \dots, p$
- the *sample standard deviation* of the jth variable is

$$s_j = \sqrt{s_{jj}} = \left[\frac{1}{n}\sum\nolimits_{i=1}^{n}(x_{ij} - \overline{x}_j)^2\right]^{1/2}$$

where $j = 1, \dots, p$

- the *sample covariance* of the ith and the jth variable is

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) \in \mathbb{R} \tag{3.2}$$

where $i \neq j$ and $i, j = 1, \dots, p$
- the *sample variance-covariance matrix* is

$$S_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \in \mathbb{R}^{p \times p}$$

which is a symmetric matrix since $s_{ij} = s_{ji}$ — note that S_n comprises p variances $s_{11}, \dots, s_{pp}$ and $p(p-1)/2$ covariances s_{ij}, $i \neq j$
- this can be expressed as

$$S_n = \frac{1}{n}(X - \mathbf{1}\overline{\mathbf{x}}^\intercal)^\intercal (X - \mathbf{1}\overline{\mathbf{x}}^\intercal) = \frac{1}{n} X^\intercal \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal \right) X$$

- the *total sample variance* is

$$\operatorname{tr}(S_n) = s_{11} + s_{22} + \cdots + s_{pp} \in \mathbb{R}$$

- the *generalized sample variance* is

$$\det(S_n) \in \mathbb{R}$$

- the *sample variance matrix* is

$$D = \operatorname{diag}(s_{11}, s_{22}, \dots, s_{pp}) = \begin{bmatrix} s_{11} & & & \\ & s_{22} & & \\ & & \ddots & \\ & & & s_{pp} \end{bmatrix} \in \mathbb{R}^{p \times p}$$

- the *sample standard deviation matrix* is

$$D^{1/2} = \operatorname{diag}(\sqrt{s_{11}}, \sqrt{s_{22}}, \dots, \sqrt{s_{pp}}) = \operatorname{diag}(s_1, s_2, \dots, s_p) = \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_p \end{bmatrix} \in \mathbb{R}^{p \times p}$$

note that $s_{ii} = s_i^2$
- *scaling by standard deviation* can be expressed as

$$XD^{1/2} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_p \end{bmatrix} = \begin{bmatrix} x_{11}/s_1 & x_{12}/s_2 & \cdots & x_{1p}/s_p \\ x_{21}/s_1 & x_{22}/s_2 & \cdots & x_{2p}/s_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}/s_1 & x_{n2}/s_2 & \cdots & x_{np}/s_p \end{bmatrix} \in \mathbb{R}^{p \times p}$$

- the *sample correlation coefficient* is

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_j)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}$$

where $i \neq j$ and $i, j = 1, \dots, p$
- r_{ij} is also known as *sample cross correlation coefficient* or *Pearson's product moment correlation coefficient*

- the *sample correlation matrix* is

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{p\times p}$$

which is a symmetric matrix since $r_{ij} = r_{ji}$
- this can be expressed as

$$R = D^{-1/2} S_n D^{-1/2}$$

- we often see *unbiased* versions of the quantities above where the factor $1/n$ is replaced by $1/(n-1)$
- for example, the unbiased versions of (3.1) and (3.2) would be

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2 \qquad \text{and} \qquad \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)(x_{kj}-\overline{x}_j) \tag{3.3}$$

- these are called the *unbiased sample variance/covariance*
- the *unbiased sample variance-covariance matrix* is

$$S = \left(\frac{n}{n-1}\right) S_n \in \mathbb{R}^{p\times p}$$

- in practice, the biased and unbiased versions of these quantities aren't that different — the reason being that in the applications we consider the sample size n would be large enough as to render the difference insignificant, i.e.,

$$\frac{n}{n-1} \approx 1$$

- but for consistency, let us decide to use the unbiased versions from now on
- note that there is only one version of the correlation coefficients and the correlation matrix — regardless of whether we use (3.1) and (3.2) or (3.3), we get the same value for r_{ij} and thus R

4. Best rank-r approximation

- we now introduce the computational basis behind *principal components analysis* (PCA), which we will introduce in the next lecture
- given $A \in \mathbb{R}^{n\times p}$, we want to find $X \in \mathbb{R}^{n\times p}$ of rank not more than r so that $\|A - X\|$ is minimized
- in notations, we want

$$\min_{\operatorname{rank}(X)\le r} \|A - X\| \tag{4.1}$$

- such an X is called a best rank-r approximation to A or a rank-r projection of A
- if $r \ge \operatorname{rank}(A)$, then clearly $X = A$ and the problem is trivial
- so we shall always assume that $r < \operatorname{rank}(A)$
- we will see how to construct such an X explicitly when the norm $\|\cdot\|$ is orthogonally invariant, i.e., satisfying

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

for all $X \in \mathbb{R}^{n\times p}$ whenever U and V are orthogonal matrices
- we will start with the classical case where $\|\cdot\|$ is the matrix 2-norm or spectral norm

Theorem 1 (Eckart–Young)**.** *Let the* SVD *of A be*

$$A = \sum_{i=1}^{\operatorname{rank}(A)} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}, \quad \sigma_1 \geq \cdots \geq \sigma_r > 0.$$

Then for any $r \in \{1, \dots, \operatorname{rank}(A) - 1\}$, a solution to (4.1) when $\|\cdot\| = \|\cdot\|_2$ is given by

$$X = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{T}}.$$

Furthermore, we have

$$\min_{\operatorname{rank}(X) \leq r} \|A - X\|_2 = \sigma_{r+1}. \tag{4.2}$$

In matrix form, we have

$$X = U \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} V^{\mathsf{T}}, \tag{4.3}$$

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

Proof. Suppose there is a $B \in \mathbb{R}^{n \times p}$ with $\operatorname{rank}(B) \leq r$ and $\|A - B\|_2 < \sigma_{r+1}$. Then by the rank-nullity theorem

$$\operatorname{rank}(B) + \dim(\ker(B)) = p$$

and so

$$\dim(\ker(B)) \geq p - r.$$

Let $\mathbf{w} \in \ker(B)$. Then $B\mathbf{w} = \mathbf{0}$ and so

$$\|A\mathbf{w}\|_2 = \|(A - B)\mathbf{w}\|_2 \leq \|A - B\|_2 \|\mathbf{w}\|_2 < \sigma_{r+1} \|\mathbf{w}\|_2. \tag{4.4}$$

Let $\mathbf{w} \in W := \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_{r+1}\}$. Then $\mathbf{w} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_{r+1} \mathbf{v}_{r+1}$. Rewriting this in matrix form

$$\mathbf{w} = V_{r+1} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{r+1} \end{bmatrix} = V_{r+1} \boldsymbol{\alpha}$$

where $V_{r+1} = [\mathbf{v}_1, \dots, \mathbf{v}_{r+1}] \in \mathbb{R}^{n \times r}$, i.e., the first $r + 1$ columns of V.

$$\|A\mathbf{w}\|_2^2 = \|U \Sigma V^{\mathsf{T}} V_{r+1} \boldsymbol{\alpha}\|_2^2 = \left\| \Sigma \begin{bmatrix} I_{r+1} \\ O \end{bmatrix} \boldsymbol{\alpha} \right\|_2^2 = \left\| \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{r+1} \end{bmatrix} \right\|_2^2$$

$$= \sum_{i=1}^{r+1} \sigma_i^2 |\alpha_i|^2 \geq \sigma_{r+1}^2 \sum_{i=1}^{r+1} |\alpha_i|^2 = \sigma_{r+1}^2 \|\mathbf{w}\|_2^2.$$

Hence if $\mathbf{w} \in W$, then

$$\|A\mathbf{w}\|_2 \geq \sigma_{r+1} \|\mathbf{w}\|_2. \tag{4.5}$$

But since $\dim(\ker(B)) \geq n-r$ and $\dim(W) = r+1$, the two subspaces must intersect nontrivially, i.e., $\dim(\ker(B) \cap W) \geq 1$ and so there exists a non-zero vector $\mathbf{w} \in \ker(B) \cap W$. Such a vector would satisfy both (4.4) and (4.5), a contradiction. Hence our original assumption is false: There is no rank-r matrix B that could beat the bound in (4.2). On the other hand it is easy to verify that the choice of X in (4.3) satisfies (4.2):

$$\|A - X\|_2 = \left\| U \begin{bmatrix} 0 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 0 & & & & & & \\ & & & \sigma_{r+1} & & & & & \\ & & & & \ddots & & & & \\ & & & & & \sigma_{\operatorname{rank}(A)} & & & \\ & & & & & & 0 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 0 \end{bmatrix} V^{\mathsf{T}} \right\|_2 = \sigma_{r+1}.$$

$\square$

- the generalization of Eckart–Young theoem to any arbitrary orthogonally invariant norm is due to Mirsky and this theorem is sometimes also called the Eckart–Young–Mirsky theorem
- note that the general theorem only says that (4.3) is the best rank-r approximation of A, the value in (4.2) would in general be different
- for example if we use the Frobenius norm

$$\min_{\operatorname{rank}(X) \leq r} \|A - X\|_F = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_{\operatorname{rank}(A)}^2}.$$