FINM 331: DATA ANALYSIS FOR FINANCE AND STATISTICS
FALL 2015
SAMPLE PRINCIPAL COMPONENTS ANALYSIS

1. POPULATION VERSUS SAMPLE

- what we discussed last time was *population* PCA, essentially the statistical principles underlying *sample* PCA, the technique that is used in practice
- in population PCA, we are given p random variables $X_1, \dots, X_p$ and we assemble it into a random vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

- in sample PCA, we are given n sample data $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ and we assemble it into a data matrix

$$X = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

- in statistics, a sample or data point

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \in \mathbb{R}^p$$

  is regarded as an observed value of the random vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

- we assume that we observe more than once so that subscript i in $\mathbf{x}_i$ indicates that this sample point comes from the ith observation
- so x_{ij}, the (i,j)th entry in the data matrix X, is the ith observation of the jth random variable X_j, $i = 1, \dots, n$, $j = 1, \dots, p$

2. LINEAR COMBINATIONS OF SAMPLE POINTS

- in population PCA, we talk about linear combinations of random variables $X_1, \dots, X_p$,

$$\mathbf{a}^\mathsf{T}\mathbf{X} = a_1 X_1 + \cdots + a_p X_p$$

- in sample PCA, we talk about linear combinations of their ith observed values $x_{i1}, \dots, x_{ip}$,

$$\mathbf{a}^\mathsf{T}\mathbf{x}_i = a_1 x_{i1} + \cdots + a_p x_{ip}$$

Date: November 4, 2015, version 1.0. Comments, bug reports: `lekheng@galton.uchicago.edu`.

- given samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$
- given any

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \in \mathbb{R}^p$$

the sample mean of $\mathbf{a}^\mathsf{T}\mathbf{x}_1, \ldots, \mathbf{a}^\mathsf{T}\mathbf{x}_n \in \mathbb{R}$ is

$$\frac{1}{n}(\mathbf{a}^\mathsf{T}\mathbf{x}_1 + \cdots + \mathbf{a}^\mathsf{T}\mathbf{x}_n) = \mathbf{a}^\mathsf{T}\left(\frac{\mathbf{x}_1 + \cdots + \mathbf{x}_n}{n}\right) = \mathbf{a}^\mathsf{T}\overline{\mathbf{x}} \in \mathbb{R}$$

this is just the sample mean of the entries of the vector

$$X\mathbf{a} = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{bmatrix} \mathbf{a} = \begin{bmatrix} \mathbf{x}_1^\mathsf{T}\mathbf{a} \\ \mathbf{x}_2^\mathsf{T}\mathbf{a} \\ \vdots \\ \mathbf{x}_n^\mathsf{T}\mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{a}^\mathsf{T}\mathbf{x}_1 \\ \mathbf{a}^\mathsf{T}\mathbf{x}_2 \\ \vdots \\ \mathbf{a}^\mathsf{T}\mathbf{x}_n \end{bmatrix} \in \mathbb{R}^n$$

and

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

- the sample variance of $\mathbf{a}^\mathsf{T}\mathbf{x}_1, \ldots, \mathbf{a}^\mathsf{T}\mathbf{x}_n \in \mathbb{R}$ is

$$\begin{aligned} \frac{1}{n-1}\sum_{i=1}^n (\mathbf{a}^\mathsf{T}\mathbf{x}_i - \mathbf{a}^\mathsf{T}\overline{\mathbf{x}})^2 &= \frac{1}{n-1}\sum_{i=1}^n \mathbf{a}^\mathsf{T}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\mathsf{T}\mathbf{a} \\ &= \mathbf{a}^\mathsf{T}\left[\frac{1}{n-1}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\mathsf{T}\right]\mathbf{a} \\ &= \mathbf{a}^\mathsf{T}S\mathbf{a} \end{aligned}$$

- the first equality above follows from

$$(\mathbf{x}^\mathsf{T}\mathbf{y})(\mathbf{z}^\mathsf{T}\mathbf{w}) = \mathbf{x}^\mathsf{T}(\mathbf{y}\mathbf{z}^\mathsf{T})\mathbf{w} \tag{2.1}$$

for any vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, $\mathbf{z}, \mathbf{w} \in \mathbb{R}^n$ — note that the LHS is a product of two scalars while the RHS is a product of a rank-1 matrix $\mathbf{y}\mathbf{z}^\mathsf{T} \in \mathbb{R}^{m\times n}$ with vectors $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{w} \in \mathbb{R}^n$

- given any

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \in \mathbb{R}^p$$

the sample covariance of $\mathbf{a}^\mathsf{T}\mathbf{x}_1, \ldots, \mathbf{a}^\mathsf{T}\mathbf{x}_n \in \mathbb{R}$ and $\mathbf{b}^\mathsf{T}\mathbf{x}_1, \ldots, \mathbf{b}^\mathsf{T}\mathbf{x}_n \in \mathbb{R}$ is

$$\begin{aligned} \frac{1}{n-1}\sum_{i=1}^n (\mathbf{a}^\mathsf{T}\mathbf{x}_i - \mathbf{a}^\mathsf{T}\overline{\mathbf{x}})(\mathbf{b}^\mathsf{T}\mathbf{x}_i - \mathbf{b}^\mathsf{T}\overline{\mathbf{x}}) &= \frac{1}{n-1}\sum_{i=1}^n \mathbf{a}^\mathsf{T}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\mathsf{T}\mathbf{b} \\ &= \mathbf{a}^\mathsf{T}\left[\frac{1}{n-1}\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\mathsf{T}\right]\mathbf{b} \\ &= \mathbf{a}^\mathsf{T}S\mathbf{b} \end{aligned}$$

where again we have used (2.1)

- in summary, for linear combinations,

$$\text{sample mean} = \mathbf{a}^\mathsf{T}\overline{\mathbf{x}}$$

$$\text{sample variance} = \mathbf{a}^\mathsf{T} S\mathbf{a} \tag{2.2}$$

$$\text{sample covariance} = \mathbf{a}^\mathsf{T} S\mathbf{b} \tag{2.3}$$

3. SAMPLE PRINCIPAL COMPONENTS ANALYSIS

- let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ be n independent drawings from p-dimensional population with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$
- recall: the sample mean of $\mathbf{x}_1, \dots, \mathbf{x}_n$ is

$$\overline{\mathbf{x}} = \frac{1}{n} X^\mathsf{T}\mathbf{1} \in \mathbb{R}^p$$

and the sample covariance of $\mathbf{x}_1, \dots, \mathbf{x}_n$ is

$$S = \frac{1}{n-1}(X - \mathbf{1}\overline{\mathbf{x}}^\mathsf{T})^\mathsf{T}(X - \mathbf{1}\overline{\mathbf{x}}^\mathsf{T}) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\mathsf{T} \in \mathbb{R}^{p\times p} \tag{3.1}$$

- sample PCA is analogous to population PCA in that sample principal components are linear combinations of sample points (instead of random variables) that have:
  (1) maximal sample variance
  (2) uncorrelated (i.e., zero covariance) with other sample principal components
- more specifically,
  - the *first sample principal component* is a unit vector $\mathbf{a}_1 = [a_{11}, \dots, a_{1p}]^\mathsf{T} \in \mathbb{R}^p$ such that the sample variance of

$$\mathbf{a}_1^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{a}_1^\mathsf{T}\mathbf{x}_n$$

is the maximum among all possible linear combinations, i.e., by (2.2),

$$\mathbf{a}_1 = \operatorname{argmax}\Big\{\mathbf{a}^\mathsf{T} S\mathbf{a} : \|\mathbf{a}\|_2 = 1\Big\}$$

  - the *second sample principal component* is a unit vector $\mathbf{a}_2 = [a_{21}, \dots, a_{2p}]^\mathsf{T} \in \mathbb{R}^p$ such that the sample variance of

$$\mathbf{a}_2^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{a}_2^\mathsf{T}\mathbf{x}_n$$

is the maximum among all possible linear combinations subjected to the constraint that,

$$\mathbf{a}_1^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{a}_1^\mathsf{T}\mathbf{x}_n \qquad \text{and} \qquad \mathbf{a}_2^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{a}_2^\mathsf{T}\mathbf{x}_n$$

are uncorrelated, i.e., by (2.2) and (2.3),

$$\mathbf{a}_2 = \operatorname{argmax}\Big\{\mathbf{a}^\mathsf{T} S\mathbf{a} : \|\mathbf{a}\|_2 = 1,\ \mathbf{a}_1^\mathsf{T} S\mathbf{a} = 0\Big\}$$

  - the *kth sample principal component* is a unit vector $\mathbf{a}_k = [a_{k1}, \dots, a_{kp}]^\mathsf{T} \in \mathbb{R}^p$ such that the sample variance of

$$\mathbf{a}_k^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{a}_k^\mathsf{T}\mathbf{x}_n$$

is the maximum among all possible linear combinations subjected to the constraints that,

$$\mathbf{a}_j^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{a}_j^\mathsf{T}\mathbf{x}_n \qquad \text{and} \qquad \mathbf{a}_k^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{a}_k^\mathsf{T}\mathbf{x}_n$$

are uncorrelated for all $j = 1, \dots, k-1$, i.e., by (2.2) and (2.3),

$$\mathbf{a}_k = \operatorname{argmax}\Big\{\mathbf{a}^\mathsf{T} S\mathbf{a} : \|\mathbf{a}\|_2 = 1,\ \mathbf{a}_1^\mathsf{T} S\mathbf{a} = \cdots = \mathbf{a}_{k-1}^\mathsf{T} S\mathbf{a} = 0\Big\}$$

Theorem 1. *Let $S \in \mathbb{R}^{p\times p}$ be the sample covariance matrix as given in* (3.1) *and let its eigenvalue decomposition be*

$$S = Q\Lambda Q^{\mathsf{T}}$$

where $Q = [\mathbf{q}_1, \dots, \mathbf{q}_p] = [q_{ij}] \in \mathbb{R}^{p\times p}$ is an orthogonal matrix of eigenvectors of S and $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$ with

$$\lambda_1 \geq \cdots \geq \lambda_p \geq 0$$

is a diagonal matrix of eigenvalues of S. The kth sample principal component of $\mathbf{x}_1, \dots, \mathbf{x}_n$ is given by

$$\mathbf{q}_k, \quad k = 1, \dots, p.$$

Furthermore the sample variance of $\mathbf{q}_k^{\mathsf{T}}\mathbf{x}_1, \dots, \mathbf{q}_k^{\mathsf{T}}\mathbf{x}_n$ is λ_k, $k = 1, \dots, p$, while the sample covariance of $\mathbf{q}_i^{\mathsf{T}}\mathbf{x}_1, \dots, \mathbf{q}_i^{\mathsf{T}}\mathbf{x}_n$ and $\mathbf{q}_j^{\mathsf{T}}\mathbf{x}_1, \dots, \mathbf{q}_j^{\mathsf{T}}\mathbf{x}_n$ is 0 for all $i \neq j$, $i, j = 1, \dots, p$.

- the total sample variance is

  $$s_{11} + \cdots + s_{pp} = \lambda_1 + \cdots + \lambda_p$$

  since $\operatorname{tr}(S) = \operatorname{tr}(Q\Lambda Q^{\mathsf{T}}) = \operatorname{tr}(\Lambda)$
- the *proportion of total sample variance* due to the kth principal component is defined as

  $$\frac{\lambda_k}{\lambda_1 + \cdots + \lambda_p}$$

  and it is often expressed as a percentage
- there is also a variant of sample PCA with the sample correlation matrix in place of the sample covariance matrix
- this is equivalently to first *standardizing* the sample or data points $\mathbf{x}_1, \dots, \mathbf{x}_n$, i.e., mean centering + scaling by standard deviation

  $$\mathbf{z}_i = D^{-1/2}(\mathbf{x}_i - \overline{\mathbf{x}}) = \begin{bmatrix} \dfrac{x_{i1} - \overline{x}_1}{\sqrt{s_{11}}} \\ \dfrac{x_{i2} - \overline{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \dfrac{x_{ip} - \overline{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \in \mathbb{R}^p, \qquad i = 1, \dots, n$$

- the *standardized data matrix* is then

  $$Z = \begin{bmatrix} \mathbf{z}_1^{\mathsf{T}} \\ \mathbf{z}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{z}_n^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} \dfrac{x_{11} - \overline{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{12} - \overline{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{1p} - \overline{x}_p}{\sqrt{s_{pp}}} \\ \dfrac{x_{21} - \overline{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{22} - \overline{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{2p} - \overline{x}_p}{\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{x_{n1} - \overline{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{n2} - \overline{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{np} - \overline{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \in \mathbb{R}^{n\times p} \tag{3.2}$$

- note that

  $$\text{sample covariance matrix of } Z = \frac{1}{n-1}Z^{\mathsf{T}}Z = R = \text{ sample correlation matrix of } X$$

- so performing sample PCA using the standardized data matrix is equivalent to using R in place of S in Theorem 1

4. scree plots

- the most basic way to use sample PCA is to do a scree plot, i.e., plotting the points $\{(i, \lambda_i) \in \mathbb{R}^2 : i = 1, \dots, p\}$ and connecting them by straight line segments
- note that we always arrange eigenvalues in nonincreasing order

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

- a scree plot gives the eigenvalue profile of the sample covariance matrix S and tells where most of the variation of the data lies
- e.g. if the first three eigenvalues are much larger than the rest, then it tells us that we should probably focus on the pairwise scatter plots of the first three principal components
- see the slides for actual examples

5. principal components scatter plots

- this is by far the most common use of PCA
- we will assume in the following that we have already mean centered our data

$$X_c = X - \mathbf{1}\overline{\mathbf{x}}^{\mathsf{T}} = \begin{bmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{bmatrix}$$

- we project our (mean centered) sample points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ onto 2- or 3-dimensional spaces spanned by pairs or triples of sample principal components
- let $\mathbf{q}_1, \dots, \mathbf{q}_p \in \mathbb{R}^p$ be the principal components of $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$, i.e., the eigenvectors of S
- since $\mathbf{q}_1, \dots, \mathbf{q}_p$ form an orthonormal basis for $\mathbb{R}^p$, we have

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

  for each $i = 1, \dots, n$
- the projection of $\mathbf{x}_i$ onto $W = \operatorname{span}\{\mathbf{q}_k\}$ the subspace spanned by the kth sample principal component (often we will just say 'projection onto the kth principal component' for short) is

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

- to plot the projections of $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ onto the kth principal component, we simply plot the n points

$$\{\mathbf{x}_i^{\mathsf{T}}\mathbf{q}_k \in \mathbb{R} : i = 1, \dots, n\}$$

  on a line, i.e., a 1-dimensional graph with one axis labelled as $\mathbf{q}_k$
- the projection of $\mathbf{x}_i$ onto $W = \operatorname{span}\{\mathbf{q}_j, \mathbf{q}_k\}$ the subspace spanned by the jth and kth sample principal components (often we will just say 'projection onto the jth and kth principal components' ror short) is

$$P_W \mathbf{x}_i = (\mathbf{x}_i^{\mathsf{T}}\mathbf{q}_j)\mathbf{q}_j + (\mathbf{x}_i^{\mathsf{T}}\mathbf{q}_k)\mathbf{q}_k$$

- to plot the projections of $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ onto the jth and kth principal components, we simply plot the n points

$$\{(\mathbf{x}_i^{\mathsf{T}}\mathbf{q}_j, \mathbf{x}_i^{\mathsf{T}}\mathbf{q}_k) \in \mathbb{R}^2 : i = 1, \dots, n\}$$

  on a plane, i.e., a 2-dimensional graph whose x-axis is labelled $\mathbf{q}_j$ and y-axis is labelled $\mathbf{q}_k$

- the projection of $\mathbf{x}_i$ onto $W = \operatorname{span}\{\mathbf{q}_j, \mathbf{q}_k, \mathbf{q}_\ell\}$ the subspace spanned by the jth, kth, and ℓth sample principal components (often we will just say 'projection onto the jth, kth, and ℓth principal components' for short) is

$$P_W \mathbf{x}_i = (\mathbf{x}_i^\mathsf{T} \mathbf{q}_j)\mathbf{q}_j + (\mathbf{x}_i^\mathsf{T} \mathbf{q}_k)\mathbf{q}_k + (\mathbf{x}_i^\mathsf{T} \mathbf{q}_\ell)\mathbf{q}_\ell$$

- to plot the projections of $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ onto the jth, kth, and ℓth principal components, we simply plot the n points

$$\{(\mathbf{x}_i^\mathsf{T} \mathbf{q}_j, \mathbf{x}_i^\mathsf{T} \mathbf{q}_k, \mathbf{x}_i^\mathsf{T} \mathbf{q}_\ell) \in \mathbb{R}^3 : i = 1, \dots, n\}$$

  in 3-space, i.e., a 3-dimensional graph whose x-axis is labelled $\mathbf{q}_j$, y-axis is labelled $\mathbf{q}_k$, z-axis is labelled $\mathbf{q}_\ell$
- in principle we can keep going but in reality, the most useful case is when we project onto sample two principal components
- if we need to project onto three or more sample principal components, a commonly used strategy is to do a pairwise scatter plots
- e.g. if we want to project onto the first three sample principal components, we plot three 2-dimensional graphs

$$\{(\mathbf{x}_i^\mathsf{T} \mathbf{q}_1, \mathbf{x}_i^\mathsf{T} \mathbf{q}_2) \in \mathbb{R}^2 : i = 1, \dots, n\}, \quad \{(\mathbf{x}_i^\mathsf{T} \mathbf{q}_1, \mathbf{x}_i^\mathsf{T} \mathbf{q}_3) \in \mathbb{R}^2 : i = 1, \dots, n\},$$
$$\{(\mathbf{x}_i^\mathsf{T} \mathbf{q}_2, \mathbf{x}_i^\mathsf{T} \mathbf{q}_3) \in \mathbb{R}^2 : i = 1, \dots, n\}$$

- see the slides for actual examples

6. singular value decomposition

- the proper way to do sample pca, to obtain principal component scattter plots, or the biplots discussed below, is to use the SVD
- you should *never* compute the EVD on S
- you should *never* compute the projection matrix P_W
- you should *never* compute the points $(\mathbf{x}_i^\mathsf{T} \mathbf{q}_j, \mathbf{x}_i^\mathsf{T} \mathbf{q}_k)$, $i = 1, \dots, n$
- what you should do is to compute the SVD of the mean centered data matrix

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

- we will assume in the following that our data has already been mean centered, i.e.,

$$X_c = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{bmatrix}$$

- note that if

$$X_c = U \Sigma V^\mathsf{T}$$

  is the condensed SVD, then

$$S = \frac{1}{n-1} X_c^\mathsf{T} X_c = \frac{1}{n-1} V \Sigma^\mathsf{T} \Sigma V^\mathsf{T} = V \begin{bmatrix} \sigma_1^2/(n-1) & & & \\ & \sigma_2^2/(n-1) & & \\ & & \ddots & \\ & & & \sigma_p^2/(n-1) \end{bmatrix} V^\mathsf{T}$$

- comparing with $S = Q \Lambda Q^\mathsf{T}$ we see that

$$Q = V, \quad \Lambda = \frac{1}{n-1} \Sigma^\mathsf{T} \Sigma$$

or

$$\mathbf{q}_k = \mathbf{v}_k, \quad \lambda_k = \frac{1}{n-1}\sigma_k^2, \quad k = 1, \dots, p$$

- in other words, the sample principal components are the right singular vectors of X_c
- as in Theorem 1, the sample variance of $\mathbf{v}_k^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{v}_k^\mathsf{T}\mathbf{x}_n$ is $\sigma_k^2/(n-1)$, $k = 1, \dots, p$, while the sample covariance of $\mathbf{v}_i^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{v}_i^\mathsf{T}\mathbf{x}_n$ and $\mathbf{v}_j^\mathsf{T}\mathbf{x}_1, \dots, \mathbf{v}_j^\mathsf{T}\mathbf{x}_n$ is 0 for all $i \neq j$, $i, j = 1, \dots, p$
- note that mean centering does not affect the values of sample variance and covariance, the sample covariance matrix of the rows of X and the sample covariance matrix of the rows of X_c are exactly the same
- in factor analysis parlance,
  - the right singular vector matrix $V \in \mathbb{R}^{p\times p}$ is called the *loading matrix*
  - the right singular vectors $\mathbf{v}_1, \dots, \mathbf{v}_p \in \mathbb{R}^p$ are called the *loading vectors* or just *loadings*
  - the product of the left singular matrix with the singular value matrix $T = U\Sigma \in \mathbb{R}^{n\times p}$ is called the *score matrix*
  - the left singular vectors scaled by the corresponding singular values $\sigma_1\mathbf{u}_1, \dots, \sigma_p\mathbf{u}_p \in \mathbb{R}^n$ are called the *score vectors* or just *scores*
  - the *score vectors*, i.e., columns of the score matrix T, are often abbreviated as $\mathbf{t}_i = \sigma_i\mathbf{u}_i$, $i = 1, \dots, p$
- another issue with the normal equations is the loss of information when we roundoff
- there are several reasons why is the SVD method is better
- the first reason is that forming $X_c^\mathsf{T}X_c$ in general results in loss of accuracy since computers have finite precision and floating point numbers are rounded off
- for example if

$$X_c = \begin{bmatrix} 1 & 1 \\ \epsilon & 0 \end{bmatrix}, \qquad X_c^\mathsf{T}X_c = \begin{bmatrix} 1+\epsilon^2 & 1 \\ 1 & 1 \end{bmatrix},$$

  and ϵ is so small that your computer rounds off $1 + \epsilon^2$ to 1, then you end up with a rank-deficient matrix

$$\mathrm{fl}(X_c^\mathsf{T}X_c) = \begin{bmatrix} \mathrm{fl}(1+\epsilon^2) & \mathrm{fl}(1) \\ \mathrm{fl}(1) & \mathrm{fl}(1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

  here $\mathrm{fl}(x)$ means the floating point representation of x
- note that for the SVD method, we work directly with X_c and do not need to form $X_c^\mathsf{T}X_c$ so we don't face this problem
- the second reason is that the condition number

$$\kappa(X_c^\mathsf{T}X_c) \approx \kappa(X_c)^2$$

  the larger the condition number, the less accurate our computation — so computing the SVD of X_c would in general be a better conditioned problem than computing the EVD of $X_c^\mathsf{T}X_c$

7. BEST WAY TO DO PCA

- the third, and most important reason, is that the projections of $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ onto *any* principal component can be read off directly from the SVD of X_c and there's no need to compute any projections
- more specifically, the projections of $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ onto any $\mathbf{v}_i$ can be read off from the entries of the score matrix $T = U\Sigma$

- let us see why

$$X_c = \begin{bmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \cdots & u_{np} \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_p \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^\mathsf{T} \\ \mathbf{v}_2^\mathsf{T} \\ \vdots \\ \mathbf{v}_n^\mathsf{T} \end{bmatrix} = \begin{bmatrix} \sigma_1 u_{11} & \sigma_2 u_{12} & \cdots & \sigma_p u_{1p} \\ \sigma_1 u_{21} & \sigma_2 u_{22} & \cdots & \sigma_p u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 u_{n1} & \sigma_2 u_{n2} & \cdots & \sigma_p u_{np} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^\mathsf{T} \\ \mathbf{v}_2^\mathsf{T} \\ \vdots \\ \mathbf{v}_n^\mathsf{T} \end{bmatrix}$$

- taking transpose

$$[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] \begin{bmatrix} \sigma_1 u_{11} & \sigma_1 u_{21} & \cdots & \sigma_1 u_{p1} \\ \sigma_2 u_{12} & \sigma_2 u_{22} & \cdots & \sigma_2 u_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p u_{1p} & \sigma_p u_{2p} & \cdots & \sigma_p u_{np} \end{bmatrix}$$

  and equating columns on LHS and RHS gives

$$\begin{aligned} \mathbf{x}_1 &= \sigma_1 u_{11} \mathbf{v}_1 + \sigma_2 u_{12} \mathbf{v}_2 + \cdots + \sigma_p u_{1p} \mathbf{v}_p \\ \mathbf{x}_2 &= \sigma_1 u_{21} \mathbf{v}_1 + \sigma_2 u_{22} \mathbf{v}_2 + \cdots + \sigma_p u_{2p} \mathbf{v}_p \\ &\vdots \\ \mathbf{x}_n &= \sigma_1 u_{n1} \mathbf{v}_1 + \sigma_2 u_{n2} \mathbf{v}_2 + \cdots + \sigma_p u_{np} \mathbf{v}_p \end{aligned}$$

- taking inner products with $\mathbf{v}_1, \dots, \mathbf{v}_p \in \mathbb{R}^p$ and using their othogonality, we see that

$$\mathbf{x}_i^\mathsf{T} \mathbf{v}_1 = \sigma_1 u_{i1}, \ \mathbf{x}_i^\mathsf{T} \mathbf{v}_2 = \sigma_2 u_{i2}, \dots, \mathbf{x}_i^\mathsf{T} \mathbf{v}_p = \sigma_p u_{ip}$$

  for $i = 1, \dots, n$
- in other words, all the information that we need to draw a principal components scatter plot can be read off from the score matrix $T = U\Sigma$
- for example, to plot the projections of $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ onto the jth and kth sample principal components, we simply plot the n points

$$\{(\sigma_j u_{ij}, \sigma_k u_{ik}) \in \mathbb{R}^2 : i = 1, \dots, n\}$$

  on a plane; to plot the projections of $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ onto the jth, kth, and ℓth sample principal components, we simply plot the n points

$$\{(\sigma_j u_{ij}, \sigma_k u_{ik}, \sigma_\ell u_{i\ell}) \in \mathbb{R}^3 : i = 1, \dots, n\}$$

  in 3-space
- the fourth reason in favor of the SVD approach is that it allows us to swap the role of samples and variables and perform a variable PCA or better, a *biplot*

8. VARIABLE PCA

- sometimes it is not clear what should be regarded as samples and what should be regarded as variables
- for example in the stock data set in the slides, we originally regarded the stocks as samples and the weekly returns as variables but there is no reason why we could not have regarded the weeks as samples and the stock prices as variables
- switching role of samples and variables is just taking the data matrix $X \in \mathbb{R}^{n \times p}$ and transposing it into $X^\mathsf{T} \in \mathbb{R}^{p \times n}$
- the SVD of X_c^T is obtained by taking transpose of $X_c = U\Sigma V^\mathsf{T}$, i.e.,

$$X_c^\mathsf{T} = V\Sigma^\mathsf{T} U^\mathsf{T}$$

- let $X_c = [\mathbf{y}_1, \dots, \mathbf{y}_n] \in \mathbb{R}^{p \times n}$ where $\mathbf{y}_i \in \mathbb{R}^p$ is the ith column of X_c

- then the same argument as in the previous section gives

$$[\mathbf{y}_1, \mathbf{y}_2 \dots, \mathbf{y}_p] = [\mathbf{u}_1, \mathbf{u}_2 \dots, \mathbf{u}_p] \begin{bmatrix} \sigma_1 v_{11} & \sigma_1 v_{21} & \cdots & \sigma_1 v_{p1} \\ \sigma_2 v_{12} & \sigma_2 v_{22} & \cdots & \sigma_2 v_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p v_{1p} & \sigma_p v_{2p} & \cdots & \sigma_p v_{pp} \end{bmatrix}$$

  and equating columns on LHS and RHS gives

$$\begin{aligned} \mathbf{y}_1 &= \sigma_1 v_{11} \mathbf{u}_1 + \sigma_2 v_{12} \mathbf{u}_2 + \cdots + \sigma_p v_{1p} \mathbf{u}_p \\ \mathbf{y}_2 &= \sigma_1 v_{21} \mathbf{u}_1 + \sigma_2 v_{22} \mathbf{u}_2 + \cdots + \sigma_p v_{2p} \mathbf{u}_p \\ &\vdots \\ \mathbf{y}_p &= \sigma_1 v_{p1} \mathbf{u}_1 + \sigma_2 v_{p2} \mathbf{u}_2 + \cdots + \sigma_p v_{pp} \mathbf{u}_p \end{aligned}$$

  taking inner products with $\mathbf{u}_1, \dots, \mathbf{u}_p \in \mathbb{R}^n$ and using their othogonality, we see that

$$\mathbf{y}_i^{\mathsf{T}} \mathbf{u}_1 = \sigma_1 v_{i1}, \ \mathbf{y}_i^{\mathsf{T}} \mathbf{u}_2 = \sigma_2 v_{i2}, \dots, \mathbf{y}_i^{\mathsf{T}} \mathbf{u}_p = \sigma_p v_{ip}$$

  for $i = 1, \dots, n$
- we will call $\mathbf{u}_1, \dots, \mathbf{u}_p \in \mathbb{R}^n$ the *variable principal components*
- in other words, all the information that we need to draw a variable principal components scatter plot can be read off from ΣV^{T}
- for example, to plot the projections of $\mathbf{y}_1, \dots, \mathbf{y}_p \in \mathbb{R}^n$ onto the jth and kth variable principal components, we simply plot the p variable points

$$\{(\sigma_j v_{ij}, \sigma_k v_{ik}) \in \mathbb{R}^2 : i = 1, \dots, p\}$$

  on a plane; to plot the projections of $\mathbf{y}_1, \dots, \mathbf{y}_p \in \mathbb{R}^n$ onto the jth, kth, and ℓth variable principal components, we simply plot the p variable points

$$\{(\sigma_j v_{ij}, \sigma_k v_{ik}, \sigma_\ell v_{i\ell}) \in \mathbb{R}^3 : i = 1, \dots, p\}$$

  in 3-space

9. BIPLOTS

- this is the powerful use of PCA combining both the sample PCA (i.e., PCA on X_c) and variable PCA (i.e., PCA on X_c^{T})
- a *biplot* essentially plots both sample points and variable points on the same graph by aligning the sample principal components and variable principal components
- if you use the SVD method, you immediately have all the required information for making a biplot
- for example, to form the biplot of $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ onto the jth and kth principal components, we simply plot the n sample points projected onto the jth and kth sample principal components

$$\{(\sigma_j u_{ij}, \sigma_k u_{ik}) \in \mathbb{R}^2 : i = 1, \dots, n\}$$

  and the p variable points onto the jth and kth variable principal components

$$\{(v_{ij}, v_{ik}) \in \mathbb{R}^2 : i = 1, \dots, p\}$$

  on the plane
- we label the x-axis $\mathbf{t}_j/\mathbf{v}_j$ and label the y-axis $\mathbf{t}_k/\mathbf{v}_k$ to indicate that each axis represents two different quantities — both scores ($\mathbf{t}_j$ and $\mathbf{t}_k$) and loadings ($\mathbf{v}_j$ and $\mathbf{v}_k$)
- the projected sample points $(\sigma_j u_{ij}, \sigma_k u_{ik})$ and projected variable points (v_{ij}, v_{ik}) should be plotted in different colors or with different symbols so that one can distinguish them
- the extension to a biplot with three principal components is straightforward

- there is one caveat: in biplot we have to make a choice whether to scale the sample principal component or variable simple components, i.e., either we plot

$$\{(\sigma_j u_{ij}, \sigma_k u_{ik}) \in \mathbb{R}^2 : i = 1, \dots, n\} \quad \text{and} \quad (v_{ij}, v_{ik}) \in \mathbb{R}^2 : i = 1, \dots, p\} \tag{9.1}$$

or

$$\{(u_{ij}, u_{ik}) \in \mathbb{R}^2 : i = 1, \dots, n\} \quad \text{and} \quad (\sigma_j v_{ij}, \sigma_k v_{ik}) \in \mathbb{R}^2 : i = 1, \dots, p\} \tag{9.2}$$

we need to decide 'where to put the singular values'
- it depends on whether you want to preserve the scale of the sample principal components (first case, more common) or the variable principal components (second case, less common)
- we will often apply biplots to sample correlation matrix instead sample covariance matrix, i.e., standardize our data as in (3.2) first before doing SVD, this reduces the effect of the scaling by singular values
- in this case, when we have standardized our data, it is sometimes to drop the singular values altogether and plot

$$\{(u_{ij}, u_{ik}) \in \mathbb{R}^2 : i = 1, \dots, n\} \quad \text{and} \quad (v_{ij}, v_{ik}) \in \mathbb{R}^2 : i = 1, \dots, p\} \tag{9.3}$$

- you can always try all three cases (9.1), (9.2), (9.3) to see which one reveals more interesting information — this is the whole point of exploratory data analysis (which is what the tools in this course is primarily for), you try different things and hopefully one or more of them would show you features hidden behind the raw data