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

1. POPULATION MEAN AND VARIANCE

- the means and variances we introduced in the last chapter are sample means and sample variances, i.e., quantities that you actually compute from the data
- we now introduce the population analogues of these
- for most of you this section ought to be just a revision
- recall that a *random variable* is a real-valued function on a sample space, i.e., $X : \Omega \to \mathbb{R}$, where Ω is the sample space (the set of all possible outcomes)
- strictly speaking we should say 'measurable real-valued function' but in this course we just need a rough working notion of random variables
- intuitively a random variable is a quantity whose values are determined by chance
- a *random vector* is a vector whose coordinates are random variables

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

- $X_1, \dots, X_p$ are random variables and so $\mathbf{X}$ is a vector-valued function

$$\mathbf{X} : \Omega \to \mathbb{R}^p$$

- the *population mean* or *mean* of a random variable X_i is

$$E(X_i) = \int_{-\infty}^{\infty} x_i f(x_i)\, dx_i =: \mu_i \in \mathbb{R}$$

  where f_i is the probability density function of X_i
- the *population variance* or *variance* of a random variable X_i is

$$\operatorname{Var}(X_i) = E(X_i - \mu_i)^2 = \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f(x_i)\, dx_i =: \sigma_{ii} = \sigma_i^2 \in \mathbb{R}$$

  where f_i is the probability density function of X_i
- the *population covariance* or *covariance* of a pair of random variables X_i and X_j is

$$\operatorname{Cov}(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - \mu_i)(x_i - \mu_i) f_{ij}(x_i, x_j)\, dx_i dx_j =: \sigma_{ij} \in \mathbb{R}$$

  where f_{ij} is the joint probability density function of X_i and X_j
- the *population correlation* or *correlation* of a pair of random variables X_i and X_j is

$$\operatorname{Corr}(X_i, X_j) = \frac{E(X_i - \mu_i)(X_j - \mu_j)}{\sqrt{E(X_i - \mu_i)^2}\sqrt{E(X_j - \mu_j)^2}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} =: \rho_{ij} \in \mathbb{R}$$

Date: October 26, 2015, version 1.1. Comments, bug reports: `lekheng@galton.uchicago.edu`.

- the *population mean vector* or *mean vector* or *mean* of a random vector $\mathbf{X}$ is

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} =: \boldsymbol{\mu} \in \mathbb{R}^p$$

- the *population variance-covariance matrix* or *covariance matrix* or *covariance* of a random vector $\mathbf{X}$ is

$$\begin{aligned}
\operatorname{Cov}(\mathbf{X}) &= E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathsf{T}} \qquad (1.1)\\
&= E\left(\begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_p-\mu_p \end{bmatrix} [X_1-\mu_1, X_2-\mu_2, \dots, X_p-\mu_p]\right)\\
&= \begin{bmatrix} E(X_1-\mu_1)^2 & E(X_1-\mu_1)(X_2-\mu_2) & \cdots & E(X_1-\mu_1)(X_1-\mu_1) \\ E(X_2-\mu_2)(X_1-\mu_1) & E(X_1-\mu_1)^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ E(X_p-\mu_p)(X_1-\mu_1) & \cdots & \cdots & E(X_p-\mu_p)^2 \end{bmatrix}\\
&= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{p1} & \cdots & \cdots & \sigma_{pp} \end{bmatrix} =: \Sigma \in \mathbb{R}^{p\times p}
\end{aligned}$$

- the *population correlation matrix* or *correlation matrix* or *correlation* of a random vector $\mathbf{X}$ is

$$\operatorname{Corr}(\mathbf{X}) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho_{p1} & \cdots & \cdots & 1 \end{bmatrix} =: P \in \mathbb{R}^{p\times p}$$

- note that the correlation and covariance matrices are related by

$$P = V^{-1/2}\Sigma V^{-1/2}$$

where

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & & & \\ & \sqrt{\sigma_{22}} & & \\ & & \ddots & \\ & & & \sqrt{\sigma_{pp}} \end{bmatrix}$$

is the *population standard deviation matrix*

2. Linear combinations of random variables

- a linear combination of random variables $X_1, \dots, X_p$ can be expressed as an inner product

$$a_1X_1 + \cdots + a_pX_p = \mathbf{a}^{\mathsf{T}}\mathbf{X}$$

where

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

is the vector of coefficients and

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

is a random vector

- the expectation satisfies

$$E(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top E(\mathbf{X}) = \mathbf{a}^\top \boldsymbol{\mu}$$

- equivalently

$$E(a_1 X_1 + \cdots + a_p X_p) = a_1 E(X_1) + \cdots + a_p E(X_p)$$

- the variance satisfies

$$\operatorname{Var}(\mathbf{a}^\top \mathbf{X}) = \mathbf{a}^\top \operatorname{Cov}(\mathbf{X}) \mathbf{a} = \mathbf{a}^\top \Sigma \mathbf{a}$$

- equivalently

$$\operatorname{Var}(a_1 X_1 + \cdots + a_p X_p) = \sum_{i=1}^{p} a_i^2 \operatorname{Var}(X_i) + 2 \sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j)$$

- more generally, suppose we transform random variables $X_1, \ldots, X_p$ into q different linear combinations

$$\begin{aligned} Y_1 &= a_{11} X_1 + \cdots + a_{1p} X_p \\ Y_2 &= a_{21} X_1 + \cdots + a_{2p} X_p \\ &\vdots \\ Y_q &= a_{q1} X_1 + \cdots + a_{qp} X_p \end{aligned}$$

- this can be written as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{q1} & \cdots & \cdots & a_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

or simply

$$\mathbf{Y} = A\mathbf{X}$$

- the expectation satisfies

$$E(\mathbf{Y}) = E(A\mathbf{X}) = AE(\mathbf{X})$$

or sometimes

$$\mu_{\mathbf{Y}} = A\mu_{\mathbf{X}}$$

where $\mu_{\mathbf{X}}$ denotes the mean of the random vector $\mathbf{X}$

- the covariance satisfies
$$\operatorname{Cov}(\mathbf{Y}) = \operatorname{Cov}(A\mathbf{X}) = A \operatorname{Cov}(\mathbf{X}) A^{\mathsf{T}} \tag{2.1}$$
or sometimes
$$\Sigma_{\mathbf{Y}} = A \Sigma_{\mathbf{X}} A^{\mathsf{T}}$$
where $\Sigma_{\mathbf{X}}$ denotes the covariance of the random vector $\mathbf{X}$

3. PRINCIPAL COMPONENTS ANALYSIS: STATISTICAL PRINCIPLES

- what we present below is often called *population principal components analysis* — it explains the statistical principles behind what we are doing when we apply PCA to data, which is often called *sample principal components ananlysis*
- the statistical motivation behind PCA is the following: suppose we have random variables $X_1, \ldots, X_p$ and we want to form linear combinations
$$\begin{aligned} Y_1 &= a_{11}X_1 + \cdots + a_{1p}X_p \\ Y_2 &= a_{21}X_1 + \cdots + a_{2p}X_p \\ &\vdots \\ Y_p &= a_{p1}X_1 + \cdots + a_{pp}X_p \end{aligned}$$
so that
  (1) $Y_1, \ldots, Y_p$ are *uncorrelated*, i.e.,
$$\operatorname{Cov}(Y_i, Y_j) = 0, \quad i \neq j, \ i, j = 1, \ldots, p$$
  (2) $Y_1, \ldots, Y_p$ have variances as large as possible, i.e.,
$$\operatorname{Var}(Y_i) \text{ is maximized}, \quad i = 1, \ldots, p$$
- $Y_1, \ldots, Y_p$ are called the *population principal components* of $X_1, \ldots, X_p$
- to get these $Y_1, \ldots, Y_p$, we find them one by one
  - the first principal component $Y_1 = \mathbf{a}_1^{\mathsf{T}}\mathbf{X}$ is such that
$$\mathbf{a}_1 = \operatorname{argmax}\{\operatorname{Var}(\mathbf{a}^{\mathsf{T}}\mathbf{X}) : \|\mathbf{a}\|_2 = 1\}$$
    note that once we get $\mathbf{a}_1$, we get Y_1
  - the second principal component $Y_2 = \mathbf{a}_2^{\mathsf{T}}\mathbf{X}$ is such that
$$\mathbf{a}_2 = \operatorname{argmax}\{\operatorname{Var}(\mathbf{a}^{\mathsf{T}}\mathbf{X}) : \|\mathbf{a}\|_2 = 1, \ \operatorname{Cov}(\mathbf{a}^{\mathsf{T}}\mathbf{X}, \mathbf{a}_1^{\mathsf{T}}\mathbf{X}) = 0\}$$
    note that we need $\mathbf{a}_1$ in order to find $\mathbf{a}_2$ and thus Y_2
  - the kth principal component $Y_k = \mathbf{a}_k^{\mathsf{T}}\mathbf{X}$ is such that
$$\mathbf{a}_k = \operatorname{argmax}\{\operatorname{Var}(\mathbf{a}^{\mathsf{T}}\mathbf{X}) : \|\mathbf{a}\|_2 = 1, \ \operatorname{Cov}(\mathbf{a}^{\mathsf{T}}\mathbf{X}, \mathbf{a}_1^{\mathsf{T}}\mathbf{X}) = \cdots = \operatorname{Cov}(\mathbf{a}^{\mathsf{T}}\mathbf{X}, \mathbf{a}_{k-1}^{\mathsf{T}}\mathbf{X}) = 0\} \tag{3.1}$$
    note that we need $\mathbf{a}_1, \ldots, \mathbf{a}_{k-1}$ in order to find $\mathbf{a}_k$ and thus Y_k
- a word on the notation used:
$$\max\{f(\mathbf{x}) : \text{ some conditions on } \mathbf{x} \in \mathbb{R}^p\}$$
is a real number $f_{\max} \in \mathbb{R}$, the maximal possible value of $f(\mathbf{x})$ over all $\mathbf{x}$ satisfying the conditions
$$\operatorname{argmax}\{f(\mathbf{x}) : \text{ some conditions on } \mathbf{x} \in \mathbb{R}^p\}$$
is a vector $\mathbf{x}_{\max} \in \mathbb{R}^p$ satisfying the conditions and that attains the value $f_{\max}$, i.e.,
$$f(\mathbf{x}_{\max}) = f_{\max}$$
- this applies also to real-valued functions f defined on matrices or scalars or any other quantities

- the solution to (3.1) for all $k = 1, \dots, n$ can be explicitly computed from the covariance matrix of $\mathbf{X}$ — the vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ are simply the eigenvectors of $\mathrm{Cov}(\mathbf{X})$

Theorem 1. *Let $\Sigma = \mathrm{Cov}(\mathbf{X}) \in \mathbb{R}^{p\times p}$ be the population covariance matrix as given in* (1.1) *and let its eigenvalue decomposition be*

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

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

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

is a diagonal matrix of eigenvalues of Σ. The kth population principal component of $X_1, \dots, X_p$ is given by

$$Y_k = \mathbf{q}_k^{\mathsf{T}}\mathbf{X}, \quad k = 1, \dots, p.$$

Furthermore

$$\mathrm{Var}(Y_k) = \mathbf{q}_k^{\mathsf{T}}\Sigma\mathbf{q}_k = \lambda_k \qquad \text{and} \qquad \mathrm{Cov}(Y_i, Y_j) = \mathbf{q}_i^{\mathsf{T}}\Sigma\mathbf{q}_j = 0$$

for all $i \neq j$, $i, j, k = 1, \dots, p$.

- the total population variance stays unchange

$$\sum_{i=1}^{p} \mathrm{Var}(X_i) = \sigma_{11} + \cdots + \sigma_{pp} = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(Y_i)$$

  since $\mathrm{tr}(\Sigma) = \mathrm{tr}(Q\Lambda Q^{\mathsf{T}}) = \mathrm{tr}(\Lambda)$
- the *proportion of total population 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
- for example, when we say things like the "first two principal components account for 90% of the variance," we mean that

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \cdots + \lambda_p} \times 100\% = 90\%$$

- the correlation coefficient between Y_i and X_j is

$$\mathrm{Corr}(Y_i, X_j) = \frac{\mathrm{Cov}(Y_i, X_j)}{\sqrt{\mathrm{Var}(Y_i)}\sqrt{\mathrm{Var}(X_j)}} = \frac{\sqrt{\lambda_i}q_{ij}}{\sqrt{\sigma_{jj}}}$$

  where q_{ij} is the (i, j)th entry of Q
- there is also a variant of population PCA with the correlation matrix in place of the covariance matrix
- this is equivalently to first *standardizing* the random variables $X_1, \dots, X_p$, i.e., mean centering + scaling by standard deviation

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_1}}, \dots, Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_p}}$$

- in vector form, this is just

$$\mathbf{Z} = V^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$$

- by (2.1), we see that the covariance matrix of $\mathbf{Z}$ is

$$\mathrm{Cov}(\mathbf{Z}) = V^{-1/2}\Sigma V^{-1/2} = P,$$

  the correlation matrix of $\mathbf{X}$
- so we get the following from Theorem 1

Corollary 1. *Let the eigenvalue decomposition of* $P = \operatorname{Corr}(\mathbf{X})$ *be*

$$P = Q\Lambda Q^{\mathsf{T}}.$$

Then the kth principal component of $Z_1, \dots, Z_p$ *is given by*

$$Y_k = \mathbf{q}_k^{\mathsf{T}}\mathbf{Z} = \mathbf{q}_k^{\mathsf{T}} V^{-1/2}(\mathbf{X} - \boldsymbol{\mu}).$$

- the total population variance is
$$\sum_{i=1}^{p} \operatorname{Var}(Y_i) = \sum_{i=1}^{p} \operatorname{Var}(Z_i) = p$$
- the correlation coefficient between Y_i and Z_j is
$$\operatorname{Corr}(Y_i, Z_j) = \frac{\operatorname{Cov}(Y_i, Z_j)}{\sqrt{\operatorname{Var}(Y_i)}\sqrt{\operatorname{Var}(Z_j)}} = \sqrt{\lambda_i}q_{ij}$$
where q_{ij} is the (i, j)th entry of Q
- in reality we are rarely given a bunch of random variables and asked to find the population principal components
- we are usually given data in the form of a data matrix $X \in \mathbb{R}^{n\times p}$
- what we do is *sample* PCA — essentially using sample variance S in place of population variance Σ, sample mean $\overline{\mathbf{x}}$ in place of population mean $\boldsymbol{\mu}$, sample standard deviation $D^{-1/2}$ in place of population standard deviation $V^{-1/2}$, etc, in what we do above
- but there is one more important difference — we will use the SVD of the data matrix $X \in \mathbb{R}^{n\times p}$ instead of the EVD of its sample covariance matrix $S \in \mathbb{R}^{p\times p}$