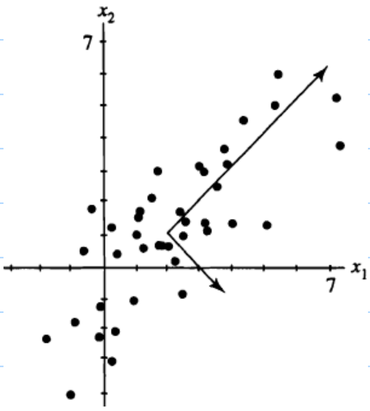
Principal Component Analysis (PCA)

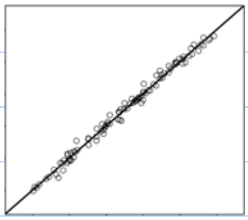
- PCA belongs to the class of *projection methods*, which choose one or few linear combinations of the original *p* variables to *maximize* some measure of "*interestingness*."

$$Y_1 = \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \longrightarrow \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \cdots & \cdots & & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nk} \end{bmatrix}$$



- In PCA, the goal is to reduce the *dimensionality* (**Q**: why?) of a data set comprised of a large number of interrelated variable, while retaining as much as possible the *variation* (**Q**: why?) present in the data

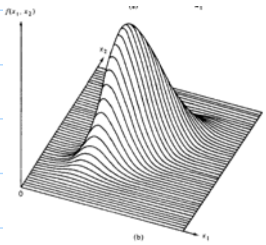


Q: 1-dim? or 2-dim?

- Population Principal Components
  - ➢ model: Let the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$.
  - ➢ First principal component = linear combination

$$Y_1 = \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

that maximizes $\mathrm{Var}(Y_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$

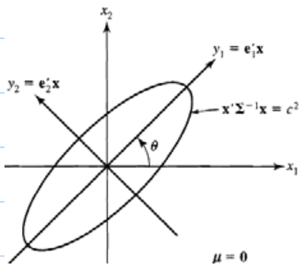


- ▪ note: $\mathrm{Var}(Y_1)$ can be increased by multiplying any $\mathbf{a}_1$ by some constant

$\Rightarrow$ restrict attention to coefficient vectors of unit length

$\Rightarrow$ $\mathbf{a}_1'\mathbf{X}$ is the projection of $\mathbf{X}$ on the direction of $\mathbf{a}_1$

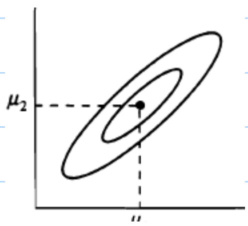
$\Rightarrow$ alternative: $\max_{\mathbf{a}\ne\mathbf{0}} \dfrac{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}}{\mathbf{a}'\mathbf{a}}$



- ➢ Second principal component = linear combination

$$Y_2 = \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

that maximizes $\mathrm{Var}(\mathbf{a}_2'\mathbf{X})$ subject to $\mathbf{a}_2'\mathbf{a}_2 = 1$ and $\mathrm{Cov}(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_2'\mathbf{X}) = 0$



- ▪ note: $\mathrm{Cov}(Y_1, Y_2) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_2$

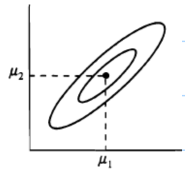
- ➢ At the ith step, ith principal component = linear combination $\mathbf{a}_i'\mathbf{X}$ that maximizes $\mathrm{Var}(\mathbf{a}_i'\mathbf{X})$ subject to $\mathbf{a}_i'\mathbf{a}_i = 1$ and $\mathrm{Cov}(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_k'\mathbf{X}) = 0$ for $k < i$
- ➢ The principal components are those *uncorrelated* linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variances are as large as possible.

➢ Finding the Principal Components

- **Result 8.1.** Let $\boldsymbol{\Sigma}$ be the covariance matrix associated with the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$. Let $\boldsymbol{\Sigma}$ have the eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the *ith principal component* is given by

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \qquad i = 1, 2, \ldots, p$$

With these choices,



$$\begin{aligned} \operatorname{Var}(Y_i) &= \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i \qquad && i = 1, 2, \ldots, p \\ \operatorname{Cov}(Y_i, Y_k) &= \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_k = 0 \qquad && i \neq k \end{aligned}$$

If some λ_i are equal, the choices of the corresponding coefficient vectors, $\mathbf{e}_i$, and hence Y_i, are not unique.

- **Result 8.2.** Let $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have covariance matrix $\boldsymbol{\Sigma}$, with eigenvalue–eigenvector pairs $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_1 = \mathbf{e}_1'\mathbf{X}$, $Y_2 = \mathbf{e}_2'\mathbf{X}, \ldots, Y_p = \mathbf{e}_p'\mathbf{X}$ be the principal components. Then

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \operatorname{Var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \operatorname{Var}(Y_i)$$

- ◆ $$\begin{pmatrix} \text{Proportion of total} \\ \text{population variance} \\ \text{due to } k\text{th principal} \\ \text{component} \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \ldots, p$$

  If most (for instance, 80 to 90%) of the total population variance, for large p, can be attributed to the first one, two, or three components, then these components can "replace" the original p variables without much loss of information.

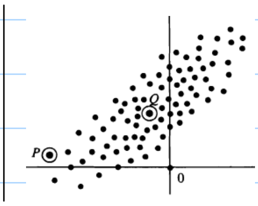
- Each component of the coefficient vector $\mathbf{e}_i' = [e_{i1}, \ldots, e_{ik}, \ldots, e_{ip}]$ also merits inspection. The magnitude of e_{ik} measures the importance of the kth variable to the ith principal component, irrespective of the other variables. In particular, e_{ik} is proportional to the correlation coefficient between Y_i and X_k.

- **Result 8.3.** If $Y_1 = \mathbf{e}_1'\mathbf{X}$, $Y_2 = \mathbf{e}_2'\mathbf{X}, \ldots, Y_p = \mathbf{e}_p'\mathbf{X}$ are the principal components obtained from the covariance matrix $\boldsymbol{\Sigma}$, then

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, 2, \ldots, p$$

are the correlation coefficients between the components Y_i and the variables X_k. Here $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ are the eigenvalue–eigenvector pairs for $\boldsymbol{\Sigma}$.

  - ◆ $e_{ik}\sqrt{\lambda_i}$ is called *factor loading*



- the i^{th} principal component scores
  - ◆ $Y_i = \mathbf{e}_i'\mathbf{X}$
  - ◆ $Y_i = \mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu})$

➢ Principal Components Obtained from Standardized Variables

- **Q**: What if one variables is measured in the millions whereas the others are measured in tens? or one variable has much larger scales than other variable?

  $\Rightarrow$ The 1st PC will essentially be just that variable

- Principal components may also be obtained for the standardized variables

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}},\ Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}},\ \ldots,\ Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

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

$$\text{Cov}(\mathbf{Z}) = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1} = \boldsymbol{\rho}$$

  where $\boldsymbol{\rho}$ is the *correlation* matrix of $\mathbf{X}$.

- **Result 8.4.** The ith principal component of the standardized variables $\mathbf{Z}' = [Z_1, Z_2, \ldots, Z_p]$ with $\text{Cov}(\mathbf{Z}) = \boldsymbol{\rho}$, is given by

$$Y_i = \mathbf{e}_i'\mathbf{Z} = \mathbf{e}_i'(\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu}), \qquad i = 1, 2, \ldots, p$$

  Moreover,

$$\sum_{i=1}^{p} \text{Var}(Y_i) = \sum_{i=1}^{p} \text{Var}(Z_i) = p \quad \Rightarrow$$

  all variables equally important

  and

$$\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i} \qquad i, k = 1, 2, \ldots, p$$

  In this case, $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ are the eigenvalue–eigenvector pairs for $\boldsymbol{\rho}$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

- Note: *the* $(\lambda_i, \mathbf{e}_i)$ *derived from* $\boldsymbol{\Sigma}$ *are, in general, not the same as the ones derived from* $\boldsymbol{\rho}$.

➢ Principal Components for Covariance Matrices with Special Structures

- $$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix}$$

  $\mathbf{e}_i' = [0, \ldots, 0, 1, 0, \ldots, 0]$, with 1 in the ith position.

  $(\sigma_{ii}, \mathbf{e}_i)$ is the ith eigenvalue–eigenvector pair

- $$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} \qquad \boldsymbol{\rho} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

  $$\lambda_1 = 1 + (p-1)\rho \qquad \mathbf{e}_1' = \left[\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}\right]$$

  $$\lambda_2 = \lambda_3 = \cdots = \lambda_p = 1 - \rho$$

  $$\mathbf{e}_2' = \left[\frac{1}{\sqrt{1 \times 2}}, \frac{-1}{\sqrt{1 \times 2}}, 0, \ldots, 0\right]$$

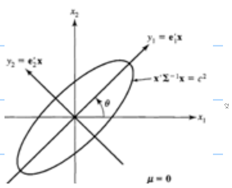
  $$\mathbf{e}_3' = \left[\frac{1}{\sqrt{2 \times 3}}, \frac{1}{\sqrt{2 \times 3}}, \frac{-2}{\sqrt{2 \times 3}}, 0, \ldots, 0\right]$$

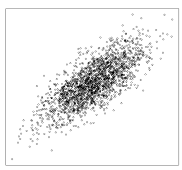
  $$\vdots \qquad \vdots$$

➢ Suppose $\mathbf{X}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- contour of its pdf is the ellipsoid defined by

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$$





❖ **Reading**: Textbook, 8.1, 8.2