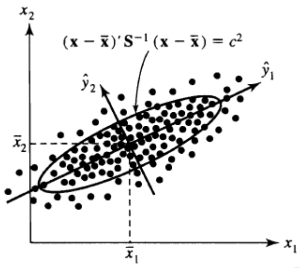
- Sample Principal Components

  - Suppose the data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ represent n independent drawings from some p-dimensional population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. These data yield the sample mean vector $\bar{\mathbf{x}}$, the sample covariance matrix $\mathbf{S}$, and the sample correlation matrix $\mathbf{R}$.
  - objective: construct uncorrelated linear combination of the measured characteristics that account for much of the variation in the sample
  - Finding principal components
    - replace population distribution by empirical distribution
    - replace Σ by $\mathbf{S}$ (or $\mathbf{S}_n$)
    - replay ρ by $\mathbf{R}$
    - then, the rests the same as above



- PCA based on $\mathbf{S}$

If $\mathbf{S} = \{s_{ik}\}$ is the $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, the ith sample principal component is given by

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{x} = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p, \qquad i = 1, 2, \ldots, p$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ and $\mathbf{x}$ is any observation on the variables $X_1, X_2, \ldots, X_p$. Also,

$$\text{Sample variance}(\hat{y}_k) = \hat{\lambda}_k, \quad k = 1, 2, \ldots, p$$

$$\text{Sample covariance}(\hat{y}_i, \hat{y}_k) = 0, \quad i \neq k$$

In addition,

$$\text{Total sample variance} = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

and

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, 2, \ldots, p$$

- principal component scores

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{x}, \qquad i = 1, 2, \ldots, p$$

$$\hat{y}_i = \hat{\mathbf{e}}_i'(\mathbf{x} - \bar{\mathbf{x}}), \qquad i = 1, 2, \ldots, p$$

- PCA based on ρ

If $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ are standardized observations with covariance matrix $\mathbf{R}$, the ith sample principal component is

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{z} = \hat{e}_{i1}z_1 + \hat{e}_{i2}z_2 + \cdots + \hat{e}_{ip}z_p, \qquad i = 1, 2, \ldots, p$$

where $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ is the ith eigenvalue-eigenvector pair of $\mathbf{R}$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Also,

$$\text{Sample variance}(\hat{y}_i) = \hat{\lambda}_i \qquad i = 1, 2, \ldots, p$$

$$\text{Sample covariance}(\hat{y}_i, \hat{y}_k) = 0 \qquad i \neq k$$

In addition,

$$\text{Total (standardized) sample variance} = \mathrm{tr}(\mathbf{R}) = p = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

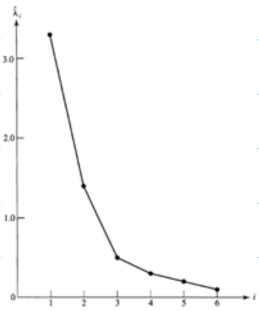
and

$$r_{\hat{y}_i, z_k} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i}, \qquad i, k = 1, 2, \ldots, p$$

- principal component scores

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{z}$$

- Number of Principal Components
  - **Q**: how many principal components to retain?
  - scree plot
  - amount of total variance explained (say, 70~90%)
  - for PCA based on $\boldsymbol{\rho}$, use the cutoff point 1 or 0.7 (i.e., keep PCs with $\hat{\lambda}_i \geq 1$ or 0.7)



- Using Principal Components to display multivariate data
  - In addition to plotting $X_1, \ldots, X_p$, plot $Y_1, \ldots, Y_k$ to check normality assumption and detect suspect observations
  - The last few PCs can help pinpoint suspect observations
- Large Sample Inference
  - **Q**: What are the distributions of the $(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$, $(\hat{e}_1, \ldots, \hat{e}_p)$, $(\hat{y}_1, \ldots, \hat{y}_p)$?
  - assumption
    - $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are a random sample from a normal population.
    - eigenvalues of $\boldsymbol{\Sigma}$ are distinct and positive, so that $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$.
  - asymptotic distribution

1. Let $\boldsymbol{\Lambda}$ be the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\boldsymbol{\Sigma}$, then $\sqrt{n}\,(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ is approximately $N_p(\mathbf{0}, 2\boldsymbol{\Lambda}^2)$.

2. Let

$$\mathbf{E}_i = \lambda_i \sum_{\substack{k=1 \\ k \neq i}}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \mathbf{e}_k \mathbf{e}_k'$$

then $\sqrt{n}\,(\hat{\mathbf{e}}_i - \mathbf{e}_i)$ is approximately $N_p(\mathbf{0}, \mathbf{E}_i)$.

3. Each $\hat{\lambda}_i$ is distributed independently of the elements of the associated $\hat{\mathbf{e}}_i$.

- A large sample $100(1-\alpha)\%$ confidence interval for λ_i is thus provided by

$$\frac{\hat{\lambda}_i}{(1 + z(\alpha/2)\sqrt{2/n})} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{(1 - z(\alpha/2)\sqrt{2/n})}$$

- Result 2 implies that the $\hat{\mathbf{e}}_i$'s are normally distributed about the corresponding $\mathbf{e}_i$'s for large samples. The elements of each $\hat{\mathbf{e}}_i$ are correlated, and the correlation depends to a large extent on the separation of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$.

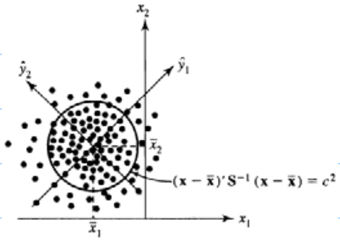
• Some important issues

- Analyses of principal components are more of a means to an end rather than an end in themselves, because they frequently serve as intermediate steps in much larger investigation.
- PCs with variance almost zero (i.e., $\hat{\lambda}_i \approx 0$)

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{x} = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p \approx c, \ c: \text{a constant}$$

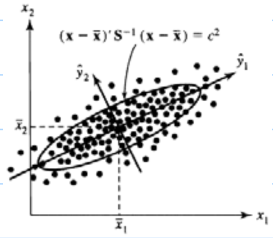
$\Rightarrow$ an unusually small value for eigenvalue can indicate a linear dependency in the data set

➢ equal eigenvalues (i.e., $\lambda_i = \lambda_{i+1} = \cdots = \lambda_j$)
- population principal component
- sample principal component



➢ selecting a subset of variables $X_1, \ldots, X_p$
- PCA does not quite reduce the data since PCs are linear combinations of all variables $X_1, \ldots, X_p$ so all variables are in some sense still needed. The technique, however, can help us discard some variables by keeping only those with the highest factor loading

➢ connection to the singular value decomposition (SVD)



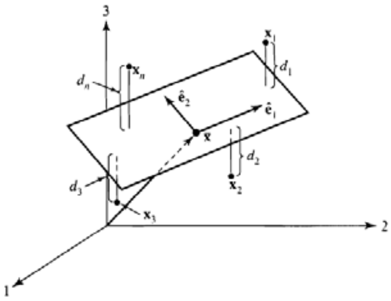
- Let us apply a SVD to the matrix $\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$

$$(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')_{n\times p} = U_{n\times p} L_{p\times p} V'_{p\times p}$$

where $U'U = I_m$, $V'V = I_m$, and L is a diagonal matrix. Then,

$$(n-1)\mathbf{S} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}') = VLU'ULV' = V(LL)V' = VL^2V'$$

$$\mathbf{S} = V\left(\tfrac{1}{n-1}L^2\right)V' = V\left(\tfrac{1}{\sqrt{n-1}}L\right)^2 V' \equiv VL^{*2}V'$$



which means
- ◆ columns of V: eigenvectors of $\boldsymbol{S}$
- ◆ $\frac{l_i^2}{n-1}$: eigenvalues of $\boldsymbol{S}$, where l_i: singular value
- ◆ $\mathbf{Y} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')V = UL$: PC scores

- the matrix $\mathbf{B}$ that minimize

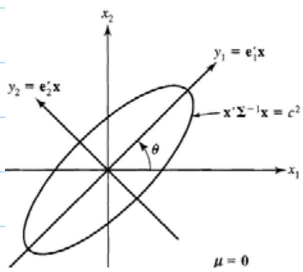
$$tr[(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}} - \mathbf{B})(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}} - \mathbf{B})']$$

over all $n\times p$ matrices $\mathbf{B}$ having rank no greater than k ($k < p$) is

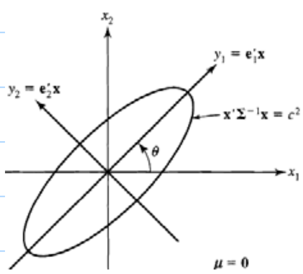
$$\mathbf{B} = \qquad\qquad \equiv UL_kV'$$

- UL_k represents the first k PC scores

➢ geometry of PC line (plane)
- the PC line (plane) minimizes the sum of squared *orthogonal distances* from each data point to the line (plane)



$$\mathbf{x}_j = (\mathbf{x}_j'\hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1 + (\mathbf{x}_j'\hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2 + \cdots + (\mathbf{x}_j'\hat{\mathbf{e}}_p)\hat{\mathbf{e}}_p$$
$$= \hat{y}_{j1}\hat{\mathbf{e}}_1 + \hat{y}_{j2}\hat{\mathbf{e}}_2 + \cdots + \hat{y}_{jp}\hat{\mathbf{e}}_p$$



- (c.f.) the least square line in regression minimizes the sum of *vertical distances* from the data point to the line

- drawbacks of PCA
  - PCA only utilizes information contained in the second moments
  - nonlinear structure may be missed
  - linear combination of variables may not be meaningful especially if the variables do not represent comparable quantities
  - outliers may distort the results

❖ **Reading**: Textbook, 8.1, 8.2, 8.3, 8.4, 8.5, Supplement 8A