STAT 3690 Lecture 21

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Mar 21, 2022

Dimension reduction

- p-dimensional $\mathbf{X} = [X_1, \dots, X_p]^\top \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Looking for a transformation $h : \mathbb{R}^p \to \mathbb{R}^s$ with $s \leq p$ such that $h(\mathbf{X})$ retains "as much information as possible" about $\mathbf{X}$

Population principal component analysis (PCA)

- Population PCA (based upon covariance matrix $\boldsymbol{\Sigma}$)
  - Looking for a linear transformation $h(\mathbf{X}) = \mathbf{X}^\top \mathbf{W}$ with $\mathbf{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_s]_{p\times s}$ and $\boldsymbol{w}_j \in \mathbb{R}^p$ such that

    $$\boldsymbol{w}_j^\top \boldsymbol{w}_j = 1 \text{ and } \mathbf{X}^\top \boldsymbol{w}_j \text{ has the maximal variance and is uncorrelated with } \mathbf{X}^\top \boldsymbol{w}_1, \dots, \mathbf{X}^\top \boldsymbol{w}_{j-1},$$

    i.e.,

    $$\boldsymbol{w}_1 = \arg\max_{\boldsymbol{w}\in\mathbb{R}^p} \mathrm{var}(\mathbf{X}^\top \boldsymbol{w}) \text{ subject to } \boldsymbol{w}_1^\top \boldsymbol{w}_1 = 1$$

    and, for $j \geq 2$,

    $$\boldsymbol{w}_j = \arg\max_{\boldsymbol{w}\in\mathbb{R}^p} \mathrm{var}(\mathbf{X}^\top \boldsymbol{w})$$

    $$\text{subject to} \qquad \boldsymbol{w}_j^\top \boldsymbol{w}_j = 1 \text{ and } \mathrm{cov}(\mathbf{X}^\top \boldsymbol{w}_j, \mathbf{X}^\top \boldsymbol{w}_{j'}) = 0 \text{ for } j' = 1, \dots, j-1$$

  - (PCA Theorem) Let $\lambda_1 \geq \dots \geq \lambda_p$ be eigenvalues of $\boldsymbol{\Sigma}$. Then the above $\boldsymbol{w}_j$ is the eigenvector corresponding to λ_j.

① To maximize $var(X^T w) = w^T \Sigma w$ subject to $w^T w = 1$

By the Lagrange multipliers, we may maximize the unconstrained problem

$$\phi(w, \theta) = w^T \Sigma w - \theta(w^T w - 1)$$

Let $\frac{\partial}{\partial w}\phi(w,\theta) = 2\Sigma w - 2\theta w = 0$

$\frac{\partial}{\partial \theta}\phi(w,\theta) = w^T w - 1 = 0$. Then,

the desired maximizer, say (θ^*, w^*), satisfies that

$$\begin{cases} \Sigma w^* = \theta^* w^* \\ w^{*T} w^* = 1 \end{cases}$$

$\therefore$ (θ^*, w^*) is an (eigenvalue, eigenvector) - pair for Σ

and $var(X^T w^*) = w^{*T}\Sigma w^* = \theta^* w^{*T} w^* = \theta^*$

$\therefore$ θ^* must be the first eigenvalue of Σ, say λ_1

$\therefore$ w^* is the eigenvector corresponding to λ_1, say w_1

② To maximize $var(X^T w) = w^T \Sigma w$ subject to $w^T w = 1$ and $cov(X^T w, X^T w_1) = w^T \Sigma w_1 = 0$

By the Lagrange multipliers, we may consider maximizing

$$\phi(w, \theta_1, \theta_2) = w^T \Sigma w - \theta_1(w^T w - 1) - \theta_2 w^T \Sigma w_1$$

$\frac{\partial}{\partial w}\phi(w, \theta_1, \theta_2) = 2\Sigma w - 2\theta_1 w - \theta_2 \Sigma w_1 = 0$

Let $\frac{\partial}{\partial \theta_1}\phi(w, \theta_1, \theta_2) = w^T w - 1 = 0$.

$\frac{\partial}{\partial \theta_2}\phi(w, \theta_1, \theta_2) = w^T \Sigma w_1 = 0$

Then the maximizer $(w^*, \theta_1^*, \theta_2^*)$ satisfies that

$$\begin{cases} 2\Sigma w^* - 2\theta_1^* w^* - \theta_2^* \Sigma w_1 = 0 & ⓐ \\ w^{*T} w^* = 1 & ⓑ \\ w^{*T} \Sigma w_1 = 0 & ⓒ \end{cases}$$

Plug $\Sigma w_1 = \lambda_1 w_1$ into ⓒ and obtain $\lambda_1 w^{*T} w_1 = 0 (\Leftrightarrow w^{*T} w_1 = 0)$

Plug $\Sigma w_1 = \lambda_1 w_1$ into ⓐ and obtain that

$2\Sigma w^* - 2\theta_1^* w^* - \theta_2^* \Sigma w_1 = 0$

$\Rightarrow \underbrace{2 w_1^T \Sigma w^*}_{=0} - \underbrace{2\theta_1 w_1^T w^*}_{=0} - \theta_2^* \underbrace{w_1^T \Sigma w_1}_{=\lambda_1} = 0$

$\Rightarrow \theta_2^* = 0$

$\Rightarrow \Sigma w^* = \theta_1^* w^*$

$\Rightarrow (\theta_1^*, w^*)$ is an (eigenvalue, eigenvector) - pair of Σ

$\because$ $var(X^T w^*) = w^{*T} \Sigma w^* = \theta_1^*$ & $w^{*T} w_1 = 0$

$\therefore$ θ_1^* is the 2nd largest eigenvalue of Σ, say λ_2

$\therefore$ w^* is the eigenvector corresponding to λ_2, say w_2.

- Vocabulary
  - $\boldsymbol{w}_j$: the jth vector of loadings
  - $Z_j = (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{w}_j \sim N(0, \lambda_j)$: the jth principal component (PC) of $\mathbf{X}$
- Identities
  - $\boldsymbol{w}_j^\top \boldsymbol{w}_{j'} = 1$ if $j = j'$ and 0 otherwise, i.e., $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p\}$ is an orthogonal basis of $\mathbb{R}^p$
    - $\mathbf{X} = \boldsymbol{\mu} + \sum_{j=1}^p Z_j \boldsymbol{w}_j$ (reconstruct the original $\mathbf{X}$ through loadings and PCs)
  - $\text{cov}(Z_j, Z_{j'}) = \boldsymbol{w}_j^\top \boldsymbol{\Sigma} \boldsymbol{w}_{j'} = \lambda_j$ if $j = j'$ and 0 otherwise
  - $\sum_{j=1}^p \text{var}(Z_j) = \sum_{j=1}^p \lambda_j = \text{tr}(\boldsymbol{\Sigma}) = \sum_{j=1}^p \text{var}(X_j)$
  - Z_j contributes $\lambda_j / \sum_{j=1}^p \lambda_j \times 100\%$ of the overall variance
    - Scree plot: displaying the amount of variation in each PC
    - Stopping rule (to determine s)

$$s = \min\{k \in \mathbb{Z}^+ : \sum_{j=1}^k \lambda_j / \sum_{j=1}^p \lambda_j \geq 90\% \text{ (or another preset threshhold)}\}$$

```
options(digits = 2)
Sigma <- matrix(
  c(10, 5, 1,
    5, 6, 5,
    1, 5, 8),
  ncol = 3)

# pca based upon covariance matrix
pca1 = eigen(Sigma, symmetric = T)
pca1$vectors # loadings
variation1 = data.frame(
  idx = 1:length(pca1$values),
  var = pca1$values
)
plot(variation1, type='b') # scree plot
cumsum(pca1$values)/sum(pca1$values) # cummulative contribution of PCs
```

- Population PCA (based upon correlation matrix $\mathbf{R}$)
  - (Pearson) correlation matrix

$$\mathbf{R} = [\mathrm{corr}(X_i, X_j)]_{p\times p} = \begin{bmatrix} \{\mathrm{var}(X_1)\}^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \{\mathrm{var}(X_p)\}^{-1/2} \end{bmatrix} \boldsymbol{\Sigma} \begin{bmatrix} \{\mathrm{var}(X_1)\}^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \{\mathrm{var}(X_p)\}^{-1/2} \end{bmatrix}$$

  - Loadings and PCs from $\mathbf{R}$ are not identical to those obtained from $\boldsymbol{\Sigma}$
  - General advice: use $\mathbf{S}$ when entries of $\mathbf{X}$ are of the same units and comparable; use $\mathbf{R}$ otherwise.
    * Using $\mathbf{R}$ rather than $\boldsymbol{\Sigma} \Leftrightarrow$ normalizing entries of $\mathbf{X}$ (i.e., $\{X_i - \mathrm{E}(X_i)\}/\sqrt{\mathrm{var}(X_i)}$) before carrying on PCA
    * Without normalizing, the component with the "smallest" units (e.g., centimeter vs. meter) could be driving most of overall variance.

```
# pca based upon correlation matrix
pca2 = eigen(cov2cor(Sigma), symmetric = T)
pca2$vectors # loadings
variation2 = data.frame(
  idx = 1:length(pca2$values),
  var = pca2$values
); plot(variation2, type='b') # scree plot
cumsum(pca2$values)/sum(pca2$values) # cummulative contribution of PCs
```