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 $s \le p$ such that $h(\mathbf{X})$ retains "as much info$rmation as possibl" 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 $j \ge 2$,

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

    $$\text{subject to} \qquad \boldsymbol{w}_j^\top\boldsymbol{w}_j = 1 \text{ and } \mathrm{cov}(\mathbf{X}^\top\boldsymbol{w}_j, \mathbf{X}^\top\boldsymbol{w}_{j'}) = 0 \text{ for } j' = 1, \dots, j-1$$
  - (PCA Theorem) Let $\lambda_1 \ge \dots \ge \lambda_p$ be eigenvalues of $\boldsymbol{\Sigma}$. Then the above $\boldsymbol{w}_j$ is the eigenvector corresponding to λ_j.
  - 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., $\{\boldsymbol{w}_1, \dots, \boldsymbol{w}_p\}$ 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)
    * $\mathrm{cov}(Z_j, Z_{j'}) = \boldsymbol{w}_j^\top\boldsymbol{\Sigma}\boldsymbol{w}_{j'} = \lambda_j$ if $j = j'$ and 0 otherwise
    * $\sum_{j=1}^p \mathrm{var}(Z_j) = \sum_{j=1}^p \lambda_j = \mathrm{tr}(\boldsymbol{\Sigma}) = \sum_{j=1}^p \mathrm{var}(X_j)$
    * 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)

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

- 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.

Sample PCA

- Data $\mathbf{X}_{n\times p} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^\top$
  - Each row $\mathbf{X}_i \overset{\text{iid}}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Estimate the loadings $\boldsymbol{w}_j$ through the eigenvectors of sample covariance matrix $\mathbf{S}$ or sample correlation matrix $\hat{\mathbf{R}}$

$$\hat{\mathbf{R}} = \begin{bmatrix} \{\widehat{\mathrm{var}}(X_1)\}^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \{\widehat{\mathrm{var}}(X_p)\}^{-1/2} \end{bmatrix} \mathbf{S} \begin{bmatrix} \{\widehat{\mathrm{var}}(X_1)\}^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \{\widehat{\mathrm{var}}(X_p)\}^{-1/2} \end{bmatrix}$$

- Matrix of scores of the first s principal components

$$\mathbf{Z} = [Z_{ij}]_{n\times s} = \tilde{\mathbf{X}}_{n\times p} \widehat{\mathbf{W}}_{p\times s}$$

  - $\tilde{\mathbf{X}} = [\mathbf{X}_1 - \bar{\mathbf{X}}, \ldots, \mathbf{X}_n - \bar{\mathbf{X}}]^\top$: row-centered $\mathbf{X}$ (i.e. the sample mean has been subtracted from each row of $\mathbf{X}$)
  - $\widehat{\mathbf{W}} = [\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_s]$: $\hat{\boldsymbol{w}}_j$ is the estimate of $\boldsymbol{w}_j$
  - $Z_{ij} = (\mathbf{X}_i - \bar{\mathbf{X}})^\top \hat{\boldsymbol{w}}_j$: the jth PC score for the ith observation