STAT 3690 Lecture 22

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Mar 23, 2022

Sample PCA

- Data $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^\top_{n\times p}$
  - 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 $\hat{\mathbf{R}}$

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

- Matrix of scores of the first s principal components

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

  - $\tilde{\mathbf{X}} = [\mathbf{X}_1 - \bar{\mathbf{X}}, \ldots, \mathbf{X}_n - \bar{\mathbf{X}}]^\top_{n\times p}$: row-centered $\mathbf{X}$ (i.e. the sample mean has been subtracted from each row of $\mathbf{X}$)
    - $\bar{\mathbf{X}} = n^{-1}\sum_{i=1}^n \mathbf{X}_i$
  - $\widehat{\mathbf{W}} = [\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_s]_{p\times s}$: $\hat{\boldsymbol{w}}_j$ 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

---

```r
Mu <- c(1, 2, 2)
Sigma <- matrix(
  c(10, 5, 1,
    5, 6, 5,
    1, 5, 8),
  ncol = 3)
n = 100
set.seed(1)
X <- mvtnorm::rmvnorm(n, mean = Mu, sigma = Sigma)

# pca based upon sample covariance matrix
pca3 = eigen(cov(X), symmetric = T)
pca3$vectors # loadings
variation3 = data.frame(
  idx = 1:length(pca3$values),
  var = pca3$values
); plot(variation3, type='b') # scree plot
```

```r
cumsum(pca3$values)/sum(pca3$values) # cummulative contribution of PCs
Z3 = scale(X, center = T, scale = F) %*% pca3$vectors # PC scores

pca4 = prcomp(X)
pca4$rotation
screeplot(pca4, type = 'l') # scree plot
cumsum((pca4$sdev)^2)/sum((pca4$sdev)^2) # cummulative contribution of PCs
Z4 = pca4$x # PC scores

# pca based upon sample correlation matrix
pca5 = eigen(cor(X), symmetric = T)
pca5$vectors # loadings
cumsum(pca5$values)/sum(pca5$values) # cummulative contribution of PCs
Z5 = scale(X, center = T, scale = F) %*% pca5$vectors # PC scores

pca6 = prcomp(X, scale. = T)
pca6$rotation
cumsum((pca6$sdev)^2)/sum((pca6$sdev)^2) # cummulative contribution of PCs
Z6 = pca6$x # PC scores

pca7 = prcomp(scale(X))
pca7$rotation
cumsum((pca7$sdev)^2)/sum((pca7$sdev)^2) # cummulative contribution of PCs
Z7 = pca7$x # PC scores

pca8 = prcomp(scale(X), scale. = T)
pca8$rotation
cumsum((pca8$sdev)^2)/sum((pca8$sdev)^2) # cummulative contribution of PCs
Z7 = pca7$x # PC scores
```

Geometric interpretation of (sample) PCA

- The definition of PCA as a linear combination that maximises variance is due to H. Hotelling (1933, Journal of Educational Psychology, 24, 417–441).
- PCA was introduced earlier by K. Pearson (1901, Philosophical Magazine, Series 6, 2(11), 559–572) to minimize the overall error in reconstructing data points

$$(\bar{\mathbf{X}}, \widehat{\mathbf{W}}, \mathbf{Z}_i) = \arg\min_{\boldsymbol{\theta}, \mathbf{A}, \mathbf{B}_i} \sum_{i=1}^{n} \|\mathbf{X}_i - \boldsymbol{\theta} - \mathbf{A}\mathbf{B}_i\|^2$$

  - $\mathbf{Z}_i$: the ith row of score matrix $\mathbf{Z}$