STAT 3690 Lecture 29

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Apr 08, 2022

CCA (Con'd)

- Sample version
  - $(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_n, \mathbf{X}_n) \overset{\text{iid}}{\sim} (\mathbf{Y}, \mathbf{X})$
    * $\mathbf{Y}_i$ and $\mathbf{X}_i$ jointly sampled
    * $p \leq q < n$
  - $n \times p$ matrix $\mathbb{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_n]^\top$ and $n \times q$ matrix $\mathbb{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^\top$
  - Sample covariance matrices
    * $\mathbf{S}_\mathbf{Y} = (n-1)^{-1} \sum_i (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^\top$
    * $\mathbf{S}_\mathbf{X} = (n-1)^{-1} \sum_i (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$
    * $\mathbf{S}_{\mathbf{YX}} = \mathbf{S}_{\mathbf{XY}}^\top = (n-1)^{-1} \sum_i (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$
  - Vocabulary
    * (The kth pair of) sample canonical directions: $(\hat{\boldsymbol{a}}_k \in \mathbb{R}^p, \hat{\boldsymbol{b}}_k \in \mathbb{R}^q)$
    * (The kth pair of) sample canonical variates: $(\mathbb{Y}_C \hat{\boldsymbol{a}}_k, \mathbb{X}_C \hat{\boldsymbol{b}}_k)$
    * (The kth) canonical correlation: $\hat{\rho}_k$
  - Goal: find $\hat{\boldsymbol{a}}_k$ and $\hat{\boldsymbol{b}}_k$, $k = 1, \ldots, r \leq p$, to maximize

$$\hat{\rho}_k = \frac{\hat{\boldsymbol{a}}_k^\top \mathbf{S}_{\mathbf{YX}} \hat{\boldsymbol{b}}_k}{\sqrt{\hat{\boldsymbol{a}}_k^\top \mathbf{S}_\mathbf{Y} \hat{\boldsymbol{a}}_k} \sqrt{\hat{\boldsymbol{b}}_k^\top \mathbf{S}_\mathbf{X} \hat{\boldsymbol{b}}_k}}$$

    subject to
    * $\hat{\boldsymbol{a}}_k^\top \mathbf{S}_\mathbf{Y} \hat{\boldsymbol{a}}_k = 1$
    * $\hat{\boldsymbol{b}}_k^\top \mathbf{S}_\mathbf{X} \hat{\boldsymbol{b}}_k = 1$
    * $\hat{\boldsymbol{a}}_k^\top \mathbf{S} \hat{\boldsymbol{a}}_\ell = 0$, $\ell = 1, \ldots, k-1$
    * $\hat{\boldsymbol{a}}_k^\top \mathbf{S}_{\mathbf{YX}} \hat{\boldsymbol{b}}_\ell = 0$, $\ell = 1, \ldots, k-1$
    * $\hat{\boldsymbol{b}}_k^\top \mathbf{S}_\mathbf{X} \hat{\boldsymbol{b}}_\ell = 0$, $\ell = 1, \ldots, k-1$
    * $\hat{\boldsymbol{b}}_k^\top \mathbf{S}_{\mathbf{XY}} \hat{\boldsymbol{a}}_\ell = 0$, $\ell = 1, \ldots, k-1$
  - Solution: Let $\widehat{\mathbf{M}} = \mathbf{S}_\mathbf{Y}^{-1/2} \mathbf{S}_{\mathbf{YX}} \mathbf{S}_\mathbf{X}^{-1/2}$
    * $\hat{\rho}_k = \sqrt{\hat{\lambda}_k}$ is the kth largest singular value of $\widehat{\mathbf{M}}$
      · $\hat{\lambda}_k$: the kth largest eigenvalue of $\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top$ (or $\widehat{\mathbf{M}}^\top\widehat{\mathbf{M}}$)
    * $\hat{\boldsymbol{a}}_k = \mathbf{S}_\mathbf{Y}^{-1/2} \hat{\boldsymbol{e}}_k$
      · $\hat{\boldsymbol{e}}_k$: the left-singular vector corresponding to the kth largest singular value of $\widehat{\mathbf{M}}$, i.e., the eigenvector corresponding to the kth largest eigenvalue of $\widehat{\mathbf{M}}\widehat{\mathbf{M}}^\top$
    * $\hat{\boldsymbol{b}}_k = \mathbf{S}_\mathbf{X}^{-1/2} \hat{\boldsymbol{f}}_k$
      · $\hat{\boldsymbol{f}}_k$: the right-singular vector corresponding to the kth largest singular value of $\widehat{\mathbf{M}}$, i.e., the eigenvector corresponding to the kth largest eigenvalue of $\widehat{\mathbf{M}}^\top\widehat{\mathbf{M}}$

- Example: olive oil data
  - 572 olive oils
  - 10 features
    - `region` indicates the general region (in Italy) of origin.
    - `area` details the area of Italy.
    - Remaining variables are continuous valued and measure the percentage composition of 8 different fatty acids
  - Interested in the correlations between the region of origin and the fatty acid measurements
    - $\mathbb{Y} \in \mathbb{R}^{572\times 3}$ an indicator matrix, i.e., each row of Y indicates the region with a 1 and otherwise has 0
    - $\mathbb{X} \in \mathbb{R}^{572\times 8}$ contains the 8 fatty acid measurements

---

- Proportion of explained correlation
  - Determining r, the number of pairs of canonical variates to retain
  - $p \times r$ matrix $\mathrm{corr}(\mathbf{Y}, \mathbf{A}_r^\top \mathbf{Y})$ and $q \times r$ matrix $\mathrm{corr}(\mathbf{X}, \mathbf{B}_r^\top \mathbf{X})$
    - The correlation matrix between $\mathbf{Y}$ (or $\mathbf{X}$) and canonical variates
    - $\mathbf{A}_r = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_r]$ and $\mathbf{B}_r = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_r]$
  - $\|\mathrm{corr}(\mathbf{Y}, \mathbf{A}_r^\top \mathbf{Y})\|_F^2/p$ and $\|\mathrm{corr}(\mathbf{X}, \mathbf{B}_r^\top \mathbf{X})\|_F^2/q$
    - Proportion of explained correlation of $\mathbf{Y}$ (or $\mathbf{X}$)
    - $\|\cdot\|_F^2$: squared Frobenius norm, i.e., sum of squared entries

---

- Interpreting canonical variates
  - $\mathrm{corr}(\mathbf{Y}, \mathbf{A}_r^\top \mathbf{Y})$
  - $\mathrm{corr}(\mathbf{X}, \mathbf{B}_r^\top \mathbf{X})$
  - $\mathrm{corr}(\mathbf{Y}, \mathbf{B}_r^\top \mathbf{X})$
  - $\mathrm{corr}(\mathbf{X}, \mathbf{A}_r^\top \mathbf{Y})$

Testing the uncorrelatedness of canonical variates

- LRT for $H_0 : \mathbf{\Sigma}_{\mathbf{YX}} = 0$ vs. H_1 : otherwise
  - LRT statistic $\lambda = \prod_{k=1}^p (1 - \hat{\rho}_k^2)^{n/2}$
    - $\hat{\rho}_k$: the kth sample canonical correlation
    - Under H_0, $-2\ln\lambda = -n\sum_{k=1}^p \ln(1-\hat{\rho}_k^2) \approx \chi^2(pq)$
- Sequential inference
  - Determining r, the number of pairs of canonical variates to retain
  - Note that $\mathbf{\Sigma}_{\mathbf{YX}} = 0 \Leftrightarrow \rho_1 = \cdots = \rho_p = 0 \Leftrightarrow \rho_1 = 0$
    - Since $\rho_1 \geq \cdots \geq \rho_p$
  - Consider a sequence of p pairs of hypotheses: $H_{0,k} : \rho_{k-1} > 0, \rho_k = 0$ vs. $H_{1,k} : \rho_k > 0$
    - LRT statistic $\lambda_k = \prod_{\ell=k}^p (1-\hat{\rho}_\ell^2)^{n/2}$
      - Under $H_{0,k}$, $-2\ln\lambda_k = -n\sum_{\ell=k}^p \ln(1-\hat{\rho}_\ell^2) \approx \chi^2((p-k+1)(q-k+1))$
  - Stopping rules
    - p_k: the p-value associated with the testing on $H_{0,k}$ vs. $H_{1,k}$
    - $p_{(k)}$: the kth smallest value among $\{p_1, \dots, p_p\}$
    - Holm-Bonferroni procedure (Holm (1979), Scandinavian Journal of Statistics, 6, 65–70): if $p_{(k)} < \alpha/(p+1-k)$, reject $H_{0,(k)}$ and proceed to larger p-values; otherwise EXIT.
    - B-H procedure (Benjamini & Hochberg (1995), Journal of the Royal Statistical Society, Series B., 57, 289–300):
      1. For a given level α, find $k^* = \max\{k \in \{1, \dots, p\} \mid p_{(k)} \leq k\alpha/p\}$
      2. Reject $H_{0,(k)}$ for $k = 1, \dots, k^*$
