STAT 475/575 Midterm Review

Outline

Data

- Vector-format: $\boldsymbol{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$
- Matrix-format (n observations): $X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$

Multivariate normal distribution (MVN)

- $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$
- PDF: $f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$
- Contour is ellipse/ellipsoid: $(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c$
- Check normality:
  - Plot: histogram, QQ plot, ...
  - Test: Shapiro-Wilk test, ...

Compare center (hypothesis test)

- $H_0 : \cdots = \boldsymbol{c}$ vs $H_a : \ldots \neq \boldsymbol{c}$
- Assumption/condition
- Test statistics & distribution
- Case
  - One-sample (Hotelling's T2): $\boldsymbol{\mu} = \boldsymbol{c}$
  - Two-sample (Hotelling's T2): $\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)} = \boldsymbol{c}$
  - More samples (MANOVA): $\boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(g)}$
- Confidence region
  - Region/ellipsoid (p-dim)
  - Interval (1-dim)
    * One-at-a-time
    * Bonferroni (simultaneous)
    * T^2 (simultaneous)

Principal Component Analysis (PCA)

$PC_j = \sum_{i=1}^{p} a_{ji} X_i$

Factor Analysis (FA)

$X_i - \mu_i = \sum_{j=1}^{m} l_{ij} F_j + \epsilon_i.$

- Principal component method
- Principal factor method
- Maximum likelihood estimation

One-sample Hotelling's T2 Test

Univariate case: one-sample t-test

$H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$.

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

Multivariate case: one-sample Hotelling's T^2 test

$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs $H_a : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ (at least one variable not equal).

$$T^2 = (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \left(\frac{S}{n}\right)^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

where sample covariance $S = \frac{1}{n-1} \sum_i (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$.

In R: `DescTools::HotellingsT2Test`

Conditions

1. Independency (most important)
2. Sample from a population with mean $\boldsymbol{\mu}$, covariance Σ
3. Multivariate-normality (in R: `mvShapiroTest::mvShapiro.Test`)

Repeated measure

$H_0 : C\boldsymbol{\mu} = \boldsymbol{c}$ vs $H_a : C\boldsymbol{\mu} \neq \boldsymbol{c}$.

$$T^2 = (C\bar{\boldsymbol{X}} - \boldsymbol{c})' \left(\frac{CSC'}{n}\right)^{-1} (C\bar{\boldsymbol{X}} - \boldsymbol{c}) \sim \frac{(n-1)q}{n-q} F_{q,n-q}$$

where $q = \text{rank}(C)$.

Two-sample Hotelling's T2 Test

Univariate case: two-sample t-test

$H_0 : \mu_1 - \mu_2 = c$ vs $H_a : \mu_1 - \mu_2 \neq c$.

$$t = \frac{\bar{X}_1 - \bar{X}_2 - c}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim t_{n_1+n_2-2}$$

where s_1^2, s_2^2 can be replaced with pooled variance $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$.

Multivariate case: two-sample Hotelling's T^2 test

$H_0 : \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)} = \boldsymbol{c}$ vs $H_a : \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)} \neq \boldsymbol{c}$ (at least one variable not equal).

$$T^2 = (\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)} - \boldsymbol{c})' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} (\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)} - \boldsymbol{c}) \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}$$

where S_1, S_2 have to be replaced with pooled variance $S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{(n_1-1)+(n_2-1)}$.

In R: `DescTools::HotellingsT2Test`

Conditions

1. Independency (within group & between group) (most important)
2. Sample 1 from a population with mean $\boldsymbol{\mu}_1$, covariance Σ_1
   Sample 2 from a population with mean $\boldsymbol{\mu}_2$, covariance Σ_2
3. Homogeneity of variance ($\Sigma_1 = \Sigma_2$) (in R: `biotools::boxM`)
4. Multivariate-normality (for each group) (in R: `mvShapiroTest::mvShapiro.Test`)

MANOVA ($g > 2$)

Univariate case: ANOVA

$H_0 : \mu_1 = \cdots = \mu_g$ vs $H_a : \mu_k \neq \mu_l$ for some k, l.

$$F = \frac{\text{MS}_{trt}}{\text{MS}_{err}} = \frac{\text{SS}_{trt}/(g-1)}{\text{SS}_{err}/(n-g)} \sim F_{g-1,n-g}$$

Multivariate case: MANOVA

$H_0 : \boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(g)}$ vs $H_a : \boldsymbol{\mu}^{(k)} \neq \boldsymbol{\mu}^{(l)}$ for some k, l (at least two groups, at least one variable not equal).

$$\text{(Wilk's) } \Lambda = \frac{|W|}{|B+W|} = \frac{\left|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{X}_j^{(l)} - \bar{\boldsymbol{X}}^{(l)})(\boldsymbol{X}_j^{(l)} - \bar{\boldsymbol{X}}^{(l)})'\right|}{\left|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\boldsymbol{X}_j^{(l)} - \bar{\boldsymbol{X}})(\boldsymbol{X}_j^{(l)} - \bar{\boldsymbol{X}})'\right|} \qquad i.e. \frac{|\text{Within-group SS}|}{|\text{Total SS}|}$$

In R: `car::Manova`

Conditions

1. Independency (within group & between group) (most important)
2. Each sample from a population with mean $\boldsymbol{\mu}_l$, covariance Σ_l
3. Homogeneity of variance ($\Sigma_1 = \cdots = \Sigma_g$) (in R: `biotools::boxM`)
4. Multivariate-normality (for each group) (in R: `mvShapiroTest::mvShapiro.Test`)

Confidence Region/Interval

(p-dim) Region for $\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$

One-sample:

$$T^2 = (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \left(\frac{S}{n}\right)^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$
$$\implies \boldsymbol{\mu} : (\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \left(\frac{S}{n}\right)^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{(p,n-p),1-\alpha}$$

Two-sample:

$$T^2 = (\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)} - \boldsymbol{c})' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} (\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)} - \boldsymbol{c}) \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}$$
$$\implies \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)} : (\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)} - (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}))' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} (\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)} - (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}))$$
$$\leq \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{(p,n_1+n_2-p-1),1-\alpha}$$

where S_1, S_2 have to be replaced with pooled variance S_p.

(1-dim) Interval for $\mu_j, \mu_j^{(1)} - \mu_j^{(2)}, \mu_j^{(k)} - \mu_j^{(l)}$

Formula:

$$\text{center} \pm \text{coef} \times \text{std.err}$$

If $g \geq 2$:

- individual variances: $S_{jj}^{(l)}$
- pooled variance: $S_{p,jj} = \frac{1}{n-g} \sum_{l=1}^{g} (n_l - 1) S_{jj}^{(l)}$

(1). One-at-a-time CI

$$\mu_j : \left[\bar{X}_j \pm t_{n-1,1-\alpha/2} \cdot \sqrt{\frac{S_{jj}}{n}} \right]$$
$$\mu_j^{(1)} - \mu_j^{(2)} : \left[\bar{X}_j^{(1)} - \bar{X}_j^{(2)} \pm t_{n_1+n_2-2,1-\alpha/2} \cdot \sqrt{\frac{S_{jj}^{(1)}}{n_1} + \frac{S_{jj}^{(2)}}{n_2}} \right]$$
$$\mu_j^{(k)} - \mu_j^{(l)} : \left[\bar{X}_j^{(k)} - \bar{X}_j^{(l)} \pm t_{n-g,1-\alpha/2} \cdot \sqrt{\frac{S_{jj}^{(k)}}{n_k} + \frac{S_{jj}^{(l)}}{n_l}} \right]$$

(2). Bonferroni simultaneous CI

Trick: $\cdot/m$ in one-at-a-time CI, where m is the no. of comparisons (CIs) you want. E.g.

- (One sample) $\mu_1, \cdots, \mu_p \implies m = p$
- (Two samples) $\mu_1^{(1)} - \mu_1^{(2)}, \cdots, \mu_p^{(1)} - \mu_p^{(2)} \implies m = p$
- (More samples) $\begin{cases} \mu_1^{(1)} - \mu_1^{(2)}, \cdots, \mu_p^{(1)} - \mu_p^{(2)} \\ \mu_1^{(1)} - \mu_1^{(3)}, \cdots, \mu_p^{(1)} - \mu_p^{(3)} \\ \mu_1^{(2)} - \mu_1^{(3)}, \cdots, \mu_p^{(2)} - \mu_p^{(3)} \end{cases} \implies m = p \cdot \binom{g}{2}$

$$\mu_j : \left[\bar{X}_j \pm t_{n-1,1-\alpha/2m} \cdot \sqrt{\frac{S_{jj}}{n}}\right]$$

$$\mu_j^{(1)} - \mu_j^{(2)} : \left[\bar{X}_j^{(1)} - \bar{X}_j^{(2)} \pm t_{n_1+n_2-2,1-\alpha/2m} \cdot \sqrt{\frac{S_{jj}^{(1)}}{n_1} + \frac{S_{jj}^{(2)}}{n_2}}\right]$$

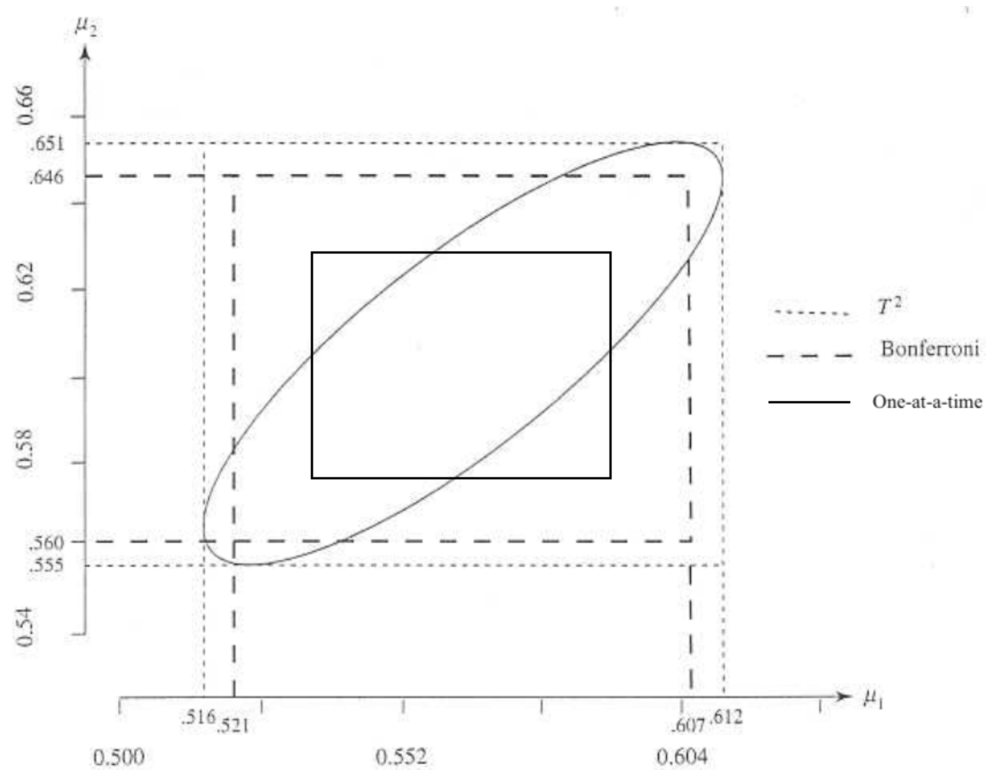
$$\mu_j^{(k)} - \mu_j^{(l)} : \left[\bar{X}_j^{(k)} - \bar{X}_j^{(l)} \pm t_{n-g,1-\alpha/2m} \cdot \sqrt{\frac{S_{jj}^{(k)}}{n_k} + \frac{S_{jj}^{(l)}}{n_l}}\right]$$

(3). T^2 simultaneous CI

Trick: $\sqrt{\cdot}$ in CR.

$$\mu_j : \left[\bar{X}_j \pm \sqrt{\frac{(n-1)p}{n-p}F_{(p,n-p),1-\alpha}} \cdot \sqrt{\frac{S_{jj}}{n}}\right]$$

$$\mu_j^{(1)} - \mu_j^{(2)} : \left[\bar{X}_j^{(1)} - \bar{X}_j^{(2)} \pm \sqrt{\frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{(p,n_1+n_2-p-1),1-\alpha}} \cdot \sqrt{\frac{S_{jj}^{(1)}}{n_1} + \frac{S_{jj}^{(2)}}{n_2}}\right]$$



Principal Component Analysis (PCA)

Model

Idea: uncorrelated linear combinations of original covariates $\boldsymbol{X}$.

$$\boldsymbol{Y}_{p\times 1} = A'_{p\times p}\boldsymbol{X}_{p\times 1} \quad \begin{cases} Y_1 = \boldsymbol{a}'_1\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 = \boldsymbol{a}'_2\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ \vdots \\ Y_p = \boldsymbol{a}'_p\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{cases}$$

$Var(\boldsymbol{X}) = \Sigma$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. $\boldsymbol{a}'_k\boldsymbol{a}_k = 1, Cov(Y_k, Y_l) = \boldsymbol{a}'_k\Sigma\boldsymbol{a}_l = 0$.

Properties

- Spectral(eigen) decomposition: $\Sigma = E\Lambda E' = \lambda_1\boldsymbol{e}_1\boldsymbol{e}'_1 + \cdots + \lambda_p\boldsymbol{e}_p\boldsymbol{e}'_p$, with eigenvalue-eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$.
- Loading matrix: $A = E = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_p]$, with coefficients $\boldsymbol{a}_j = \boldsymbol{e}_j$.
- $Var(\boldsymbol{Y}) = \Lambda = \text{diag}(\lambda_i)$ and $Cov(\boldsymbol{Y}, \boldsymbol{X}) = \Lambda E'$.
  $Var(Y_j) = \lambda_j$ or $\frac{\lambda_j}{\lambda_1 + \cdots + \lambda_p}$ (proportion).
- PC score: $Y_{n\times p} = X_{n\times p}A$ or $\boldsymbol{Y}_{p\times 1} = A'\boldsymbol{X}$.
- Choose number of PCs: (1) scree plot + elbow method (2) % variance (3) context (4) meaningful interpretation (5) desire for simplicity (6)...
- Interpretation: size (absolute value) and sign (positive or negative) of coefficients $\boldsymbol{a}_j$.

In R: `prcomp` or `princomp`
`prcomp(x, center = T, scale. = F)`

- `scale. = F`: raw/centered data (covariance)
- `scale. = T`: standardized data (correlation)

`princomp(x, cor = F)` or `princomp(covmat, cor = F)`

- `cor = F`: centered data (covariance)
- `cor = T`: standardized data (correlation)

Factor Analysis (FA)

Model

Idea: latent variables behind original covariates $\boldsymbol{X}$.

$$\boldsymbol{X}_{p\times 1} - \boldsymbol{\mu}_{p\times 1} = L_{p\times m}\boldsymbol{F}_{m\times 1} + \boldsymbol{\epsilon}_{p\times 1} \qquad \begin{cases} X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \epsilon_1 \\ X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \epsilon_2 \\ \vdots \\ X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \epsilon_p \end{cases}$$

$E(\boldsymbol{F}) = \boldsymbol{0}, Var(\boldsymbol{F}) = I_m, E(\boldsymbol{\epsilon}) = \boldsymbol{0}, Var(\boldsymbol{\epsilon}) = \Psi = \text{diag}(\psi_i),$ and $\boldsymbol{F} \perp \boldsymbol{\epsilon}$.

Properties

- $Var(\boldsymbol{X}) = \Sigma = LL' + \Psi$ and $Cov(\boldsymbol{X}, \boldsymbol{F}) = L$.
  $Var(X_i) = \sigma_{ii} = l_{i1}^2 + \cdots + l_{im}^2 + \psi_i$ and $Cov(X_i, X_k) = \sigma_{ik} = l_{i1}l_{k1} + \cdots + l_{im}l_{km}$.
- $\sigma_{ii} = h_i^2 + \psi_i$. Communality: $h_i^2 = \sum_{j=1}^m l_{ij}^2$. Uniqueness: $\psi_i = \sigma_{ii} - h_i^2$.
- Infinite rotations: $\boldsymbol{X} - \boldsymbol{\mu} = L\boldsymbol{F} + \boldsymbol{\epsilon} = (LT)(T'\boldsymbol{F}) + \boldsymbol{\epsilon}$ for any orthogonal matrix T.
  E.g. varimax, quartimax, promax, ...
- Contribution of j-th factor: $\sum_{i=1}^p l_{ij}^2$ or $\frac{\sum_{i=1}^p l_{ij}^2}{\lambda_1 + \cdots + \lambda_p}$ (proportion).
- Factor score: $F_{n\times m} = (X_{n\times p} - \boldsymbol{1}_n\bar{\boldsymbol{X}}')(LL' + \Psi)^{-1}L$ or $\boldsymbol{F}_{m\times 1} = L'(LL' + \Psi)^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})$.
- Interpretation: size (absolute value) + sign (positive or negative) of coefficients $\boldsymbol{l}_j$.

Three common methods of estimating factor loadings:

1. Principal component method:
   - Loading matrix: $L = [\boldsymbol{l}_1, \cdots, \boldsymbol{l}_m] = [\sqrt{\lambda_1}\boldsymbol{e}_1, \cdots, \sqrt{\lambda_m}\boldsymbol{e}_m]$, with coefficients $\boldsymbol{l}_j = \sqrt{\lambda_j}\boldsymbol{e}_j$.
   - Choose number of Factors: (1) no. of PCs in PCA (2) % variance (3) context (4) meaningful interpretation (5) desire for simplicity (6)...
2. Principal factor method:
   - Iterative method: $\Psi \to L \to \Psi \to L \to \cdots$.
3. Maximum likelihood estimation:
   - Normality assumption: $\boldsymbol{X}, \boldsymbol{F}, \boldsymbol{\epsilon}$ are normal.
   - Maximize the likelihood function with respect to $(\boldsymbol{\mu}, L, \Psi)$.
   - Likelihood ratio test: H_0: m-factors are sufficient.
   - Choose number of Factors: (1) likelihood ratio test (2) % variance (3) context (4) meaningful interpretation (5) desire for simplicity (6)...

In R: `prcomp` or `factanal`
1. PC method:

(1) PCA: `prcomp` or `princomp`
(2) Factor loading: `rotation %*% diag(sdev)`
(3) Rotation (optional): `varimax` on first m columns of factor loading

3. MLE:
`factanal(x, factors, scores = "regression", rotation = "varimax")` or
`factanal(factors, covmat, n.obs, scores = "regression", rotation = "varimax")`

- `rotation`: `"varimax"` (default) or `"none"`
- `scores`: `"none"` (default) or `"regression"`