Refreshment Matrix Algebra

Spectral decomposition (A symmetric $p \times p$):

$$A = \Gamma \cdot \Lambda \cdot \Gamma^\top$$
$$\Gamma = (\gamma_1, ..., \gamma_p) \quad \text{Eigenvectors}$$
$$\Lambda = \text{diag}(\lambda_1, ..., \lambda_p) \quad \text{Eigenvalues}$$

Singular value decomposition (A $n \times p$, $\text{rk}(A) = r$)

$$A = \Gamma \cdot \Lambda \cdot \Delta^\top$$
$$\Lambda = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r})$$
$$\Gamma = \text{Eigenvector}(AA^\top)$$
$$\Delta = \text{Eigenvector}(A^\top A)$$

Decomposition of Data Matrices

n observations in variables space $\mathbb{R}^p$

- projection of x on vector with direction u: $p_x = \|p_x\| \cdot u$
- $\|p_x\| = \langle x, u\rangle$ (dot product between x and u)
- $\min\limits_{\|u\|=1} \sum \|x_i - px_i\|^2 = \max\limits_{\|u\|=1} \sum \|p_{x_i}\|^2 = \max\limits_{\|u\|=1} u^\top X^\top X u$

$\Rightarrow$ $u = \text{Eigenvector}(X^\top X)$

p variables in observation space $\mathbb{R}^n$: $X \to X^\top$

- $\max\limits_{\|v\|=1} v^\top X X^\top v$

$\Rightarrow$ $v = \text{Eigenvector}(XX^\top)$

Relations between subspaces:

$$u = \lambda^{-1/2} \cdot X^\top \cdot v$$
$$v = \lambda^{-1/2} \cdot X \cdot u$$

Representation:

$$z_i = X u_i \quad \text{observations}$$
$$w_i = \sqrt{\lambda_i} u_i \quad \text{parameters}$$

Principal Component Analysis

Theory:

- $\max\limits_{\|\delta\|=1} \text{Var}(\delta^\top X) \Rightarrow \delta = \text{Eigenvector}(\text{Var}(X)) = \text{Eigenvector}(X^\top X)$
- $Y = \Gamma^\top (X - \mu)$ principal components ($\text{Var}(Y_i) = \lambda_i$)
- $\text{Cov}(X, Y) = \Gamma \cdot \Lambda$
- $\text{Cor}(X_i, Y_i) = \gamma_{ij} \sqrt{\frac{\lambda_j}{\text{SD}(X_i X_i)}}$

Practise:

- scale should be roughly the same
- plot $\text{Cor}(X_1, Y_1)$ vs $\text{Cor}(X_2, Y_2)$ shows which of the original variables are most correlated with the PCs, namely those which are near the periphery of the circle of radius 1.

Asymptotic properties:

- 95% CI for explained variance ψ:

$$\widehat{\psi} \pm 1.96 \sqrt{\frac{\widehat{\omega^2}}{n-1}}$$
$$\widehat{\omega^2} = \frac{2\,\text{tr}(S^2)}{\text{tr}(S)^2} \left(\widehat{\psi}^2 - 2\widehat{\beta}\widehat{\psi} + \widehat{\beta}\right)$$
$$\widehat{\beta} = \frac{l_1^2}{l_1^2 + \cdots + l_p^2}$$

- $\sqrt{n-1}(l - \lambda) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, 2\Lambda^2)$
- $\sqrt{\frac{n-1}{2}}(\log(l_j) - \log(\lambda_j)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$
- $\sqrt{n-1}\left(\widehat{\psi_q} - \psi_q\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \omega^2)$

Factor Analysis

Factor analysis model:

- $X = \mu + QF + U$
- $\text{Var}(X) = QQ^\top + \text{Var}(U) = \Gamma\Lambda\Gamma^\top$
- $\text{Var}(X_i) = \sum_{l=1}^k q_{il}^2 + \psi_{ii}$ = communality h_j^2 + specic variance
- $Q = \Gamma\Lambda^{1/2}$ (principal component method, assuming $\text{Var}(U) = 0$)
- other methods: maximum likelihood method, method of principal factors

Factor model for correlation matrix:

- Choice of k:

$$d = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)$$

- $d < 0$: ∞ exact solutions, $d = 0$: 1 exact solution, $d > 0$: approximation
- example: $p = 2$, $k = 1 \Rightarrow d = -1$

$$R = \begin{pmatrix} 1 & \\ \rho & 1 \end{pmatrix} = \begin{pmatrix} q_1^2 + \psi_1 & \\ q_1 q_2 & q_2^2 + \psi_2 \end{pmatrix}$$

4 parameters and 3 equations

Rotation: $X = \mu + (QG)(G^\top F) + U$

- orthogonal: $\perp$ between factors
- oblique: no $\perp$ between factors

$\Rightarrow$ i.e. varimax: $\sum_{j=1}^k \text{Var}(q_j^2) \to \max$ (each factor has small or large loadings on variable)

Correspondence Analysis

- expectation of an element $\mathbb{E}(x_{ij}) = E_{ij} = \frac{x_{i\cdot} \cdot x_{\cdot j}}{x_{\cdot\cdot}}$
- χ^2-Test:

$$t = \sum_i \sum_j \frac{(x_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(l-1)(k-1)}$$

- find r_k (row factor) and s_k (column factor)

$$C = (c_{ij}) = \left(\frac{x_{ij} - E_{ij}}{\sqrt{E_{ij}}}\right) = \Gamma\Lambda\Delta^\top$$
$$c_{ij} = \sum_k \sqrt{\lambda_k} \cdot \gamma_{ik} \cdot \delta_{jk}$$
$$\approx \sum_k \sqrt{\lambda_1} \cdot \gamma_{ik} \cdot \delta_{jk}$$

(one λ explains much χ^2). Then

$$r_k = A^{-1/2} \cdot C \cdot \delta_k$$
$$s_k = B^{-1/2} \cdot C \cdot \gamma_k$$

$A = \text{diag}(\text{row sums})$, $B = \text{diag}(\text{column sums})$

- plot r_1 vs r_2 and s_1 vs s_2

Canonical Correlation Analysis

- $\text{Cor}(aX, bY) \to \max$ under constrains $a^\top \Sigma_{XX} a = b^\top \Sigma_{YY} b = 1$
- Define

$$K = \Sigma_{XX}^{-1/2} \cdot \Sigma_{XY} \cdot \Sigma_{YY}^{-1/2} = \Gamma\Lambda\Delta^\top$$

- Set

$$a_r = \Sigma_{XX}^{-1/2} \gamma_r$$
$$b_r = \Sigma_{YY}^{-1/2} \delta_r$$

then $\text{Cor}(a_r X, b_r Y) = \sqrt{\lambda_r}$

Multidimensional Scaling

metric MDS: euclidian matrix D

- D euclidian $\Leftrightarrow B = HAH$ is positive semidefinite ($\Leftrightarrow$ all eigenvalues ≥ 0)
- $A = (a_{ij}) = -\frac{1}{2} d_{ij}^2$

$\Rightarrow$ $B = \Gamma\Lambda\Gamma^\top \Rightarrow$ coordinates $\Gamma\Lambda^{1/2}$

- similarity $C \to$ distance D: $d_{ij} = \sqrt{c_{ii} - 2c_{ij} + c_{jj}}$

nonmetric MDS: ranks instead of distances, Shepard-Kruskal-Algorithm

- metric MDS $\to$ coordinates $\to$ distances δ_{ij}
- $d_{ij} = \text{rk}(\delta_{ij})$
- compare D and d_{ij} $\to$ monotone? $\to$ mean of non-monotone points
- calc $STRESS1 = \sqrt{\frac{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}}$ (small is good)

$$x_{il}^{new} = x_{il} + \frac{\alpha}{n-1} \sum_{j=1, j\neq i} \left(1 - \frac{\hat{d}_{ij}}{d_{ij}}\right)(x_{jl} - x_{il})$$

Discriminant Analysis

- ML Discriminant Rule: $R_1 = \{x \mid L_1(x) > L_{\neq 1}(x)\}$
- ECM Rule: $R_1 = \left\{x \mid \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)} \cdot \frac{\pi_2}{\pi_1}\right\}$
- LDA: Group $i \sim \mathcal{N}_p(\mu_i, \Sigma)$ (squared Mahalanobis distance): $x \in R_1 \Leftrightarrow w^\top (x - \mu) > 0$, $w = \Sigma^{-1}(\mu_1 - \mu_2)$
- QDA: Group $i \sim \mathcal{N}_p(\mu_i, \Sigma_i)$, $x \in R_1 \Leftrightarrow -\frac{1}{2} x^\top (\Sigma_1^{-1} - \Sigma_2^{-1}) x + (\mu_1 \Sigma_1^{-1} - \mu_2 \Sigma_2^{-1}) x - k \geq \log\left(\frac{c(1|2)}{c(2|1)} \cdot \frac{\pi_2}{\pi_1}\right)$ where $k = \frac{1}{2} \log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_2)}\right) + \frac{1}{2}(\mu_1^\top \Sigma_1^{-1} \mu_1 - \mu_2^\top \Sigma_2^{-1} \mu_2)$
- Bayes Rule: $\max \pi_i \cdot f_i(x)$ (all Bayes rules are admissible, no rule exists where $p'_{ii} > p_{ii}$)
- apparent error rate (APER) $= \frac{\#\text{misclassified}}{\#\text{all}}$ $\to$ too optimistic
- actual error rate (AER) with CV $= \frac{\#\text{misclassified}}{\#\text{all}}$
- Fisher: $\max\limits_w \frac{w^\top B w}{w^\top W w} = \frac{w^\top \sum n_j (\bar{y}_j - \bar{y})^2 w}{\sum (w^\top X_i) H_i (X_i w)}$, $w = \text{Eigenvector}(W^{-1}B)$, $x \in R_j \Leftrightarrow j = \arg\min\limits_i |w^\top (x - \bar{x_i})|$

Regression

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top y$$
$$\mathbb{E}(\hat{\beta}) = \beta$$
$$\text{Var}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$$
$$\mathbb{E}((y - x^\top \hat{\beta})^2) = \sigma^2 + (x^\top \beta - \mathbb{E}(x^\top \beta))^2 + x^\top \text{Var}(\hat{\beta}) x$$

if $\log(y_i) = z_i = \beta_0 + \beta_1 x_1 + \ldots$ then forecast y by $\exp(z)$ is wrong (biased), better $y = \exp(z + \frac{1}{2}\text{Var}(z))$

RESET-test tests for non-linearity

Detect multicollinearity:

- $\det(\text{Cor}(X, X)) \approx 0 \Rightarrow$ MC
- condition number $\sqrt{\frac{\lambda_1}{\lambda_p}} \geq 30 \Rightarrow$ MC
- variance inflation factor $VIF_j = \frac{1}{1 - R_j} \geq 5 \Rightarrow$ x_j contributes to MC

$\Rightarrow$ more orthogonal data, remove variables, PCR, Ridge Regression

Model building:

- general to simple, otherwise estimators are biased and have lower variance
- Goodness-of-fit-measures:

$$R_{adj}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \to \max$$
$$AIC = \log(\hat{\sigma}^2) + \frac{2p}{n} \to \min$$
$$BIC = \log(\hat{\sigma}^2) + \frac{p}{n}\log(n) \to \min$$

comparing non-nested models:

- $y = Z\gamma + X_2\beta_2 + \varepsilon$, $\beta_2 = 0$?
- R^2, AIC, BIC
- J-Test: $y = Z\gamma + \delta \cdot$ prediction from $y = X\beta$, $\delta = 0$?

leverage: $h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$,
$\mathrm{diag}(h_{11}, ..., h_{nn}) = X(X^\top X)^{-1}X^\top$

non-parametric regression: use kernel density estimation

PCR:
- find first k eigenvectors G_k
- $Z_k = X \cdot G_k$
- $y = Z_k \cdot \alpha_k \Rightarrow \hat{\alpha}_k = (Z_k^\top Z_k)^{-1} Z_k^\top y$
- $\hat{\beta}_k = G_k \cdot \hat{\alpha}_k$
- number of parameters: use MSE

Ridge-Regression:
- $\|y - X\beta\|_2^2 + \lambda \|\beta\|_2 \to \min$

$\Rightarrow$ $\hat{\beta}_{RR} = (X^\top X + \lambda I)^{-1} X^\top y$
- $\mathrm{Var}(\hat{\beta}_{OLS}) = \sigma^2 \cdot DL^{-2}D^\top$
- $\mathrm{Var}(\hat{\beta}_{RR}) = \sigma^2 \cdot DL_\lambda^{-2}D^\top$ where $L_\lambda^{-1} = \mathrm{diag}\left(\frac{\sqrt{l_i}}{l_i + \lambda}\right)$

Ridge-Regression via ML approach:

$$p(\theta \mid X) = \frac{p(X \mid \theta) \cdot p(\theta)}{p(X)} \to \max \quad \text{MAP}$$

(LS maximizes likelihood)

$$\hat{\beta}_{MAP} = (X^\top X + \sigma^2 \lambda I)^{-1} X^\top y$$

LASSO-Regression:
- $\|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \to \min$
- no closed-form solution
- shrinks parameters to zero
- adaptive LASSO: $\|y - X\beta\|_2^2 + \lambda \sum w_i |\beta_i| \to \min$

Logistic Regression

$$\mathbb{P}(Y = 1 \mid X) = \frac{1}{1 + \exp(-X\beta)} \quad \text{logit}$$
$$= \Phi(X\beta) \quad \text{probit}$$

Interpretation:
- $\beta_j > 0$:$X\beta$ raises by $\beta_j \to \mathbb{P}$ raises by $\exp(\beta_j)$
- $\beta_j < 0$:$X\beta$ falls by $\beta_j \to \mathbb{P}$ falls by $\exp(\beta_j)$

Goodness of Model: R^2 can't be used
- pseudo R^2: L_0 log likelihood where $b_1 = b_2 = ... = 0$, L_v log likelihood for full model
  - Deviance: $D = -2L_v$
  - Mc-Faddens-R^2: $1 - \frac{L_v}{L_0}$
- accuracy: $\frac{TP+TN}{P+N}$ (same as APER)
- ROC: sensitivity $= \frac{TP}{P}$, specificity $= \frac{TN}{N}$
  - ROC-curve: sensitivity values as a function of 1 - specificity
  - AUC: area under curce $\to$ max
  - ROC-curve diagonal $\Rightarrow$ random guessing