Introduction

→ sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

→ variance

$$s_{11} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{i1} - \overline{x}_1)^2$$

→ pairwise covariance

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$$

→ sample correlation coefficient

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$$

Eigenvalues - Eigenvectors

$$q(a) = |\underbrace{\Sigma - Ia}| \overset{!}{=} 0 \to a$$

singular → $a\gamma_i = \Sigma\gamma_i$ right eigenvector of Σ to a_i

standardized if $\gamma_i^T \gamma_i = 1$

$$|\Sigma - aI| = |C||\Sigma - aC^{-1}C||C^{-1}| = |C\Sigma C^{-1} - aI|$$

$$\gamma_i \Sigma = \gamma_i a \to \gamma_i C^{-1}\Sigma C = \gamma_i a \to C\gamma_i$$

Spectral Theorem

→ every symmetric $p \times p$ matrix Σ

$$\Sigma = \Gamma A \Gamma^T = \sum_{i=1}^{p} \gamma_i a_i \gamma_i^T$$

$\Gamma^T = \Gamma^{-1}$

$\Gamma = (\gamma_1 \dots \gamma_p)$

$A = \mathrm{diag}(a_1 \dots a_n)$

Multivariate Case

$E(x) = [(E(x_{ij})]$

$E(Ax+B) = AE(x)+B$

$cov(x,y) = E(x_i - E(x_i))E(y_i - E(y_i))$

$cov(x,y) = [E(x-\mu_x)E(y-\mu_y)^T] = E(xy^T) - \mu_x \mu_y^T$

Multivariate Normal Distribution

$$f(x) = 2\pi^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

$(x-\mu)^T \Sigma^{-1}(x-\mu)$: MD^2 without $\frac{1}{2}$!

$x-\mu \sim N_p(0,\Sigma)$ $\quad$ $Ax \sim N_p(A\mu, A\Sigma A^T)$

$MD \stackrel{!}{=} c^2 \rightarrow$ ellipse $\quad$ $MD^2 \sim \chi_p^2$

Wishart Distribution

→ symmetric $p \times p$ matrix W

→ if W can be represented as:

$W = X^T X \mid x \sim N_p(0,\Sigma)$

we can use the notation:

$W \sim W_p(\Sigma, n)$

→ Properties:

→ $A^T W A \sim W_q(A^T \Sigma A, n)$

→ $\Sigma^{-\frac{1}{2}} W \Sigma^{-\frac{1}{2}} \sim W(I, n)$

→ $\frac{a^T W a}{a^T \Sigma a} \sim \chi_n^2 \mid a \neq 0$

Central Limit Theorem

→ n independent obs ⇒ for large n−p approximately:

$$\sqrt{n}(\bar{x}-\mu) \sim N(0,\Sigma)$$

$$n(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu) \sim \chi_p^2$$

⇒ approximate Normal distribution

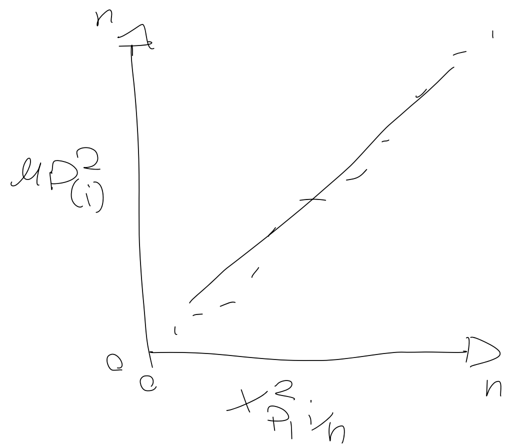
Tests for multivariate normality – χ^2 plot

(1) for every obs ⇒ $MD_i^2(x_i, \bar{x}, S)$

(2) sort MD's: $MD_{(1)}^2 \leq \dots \leq MD_{(n)}^2$

(3) compute quantiles i/n of χ_p^2: $\chi^2_{p, i/n}$

(4) Graph: if linear trend ⇒ multivar. normal



Transformation to Normality

→ "change of data scale"

→ examples:

Count y → $\sqrt{y}$ Ratios → Log Correlations → Fisher

→ Power Transformation x^λ

→ for appropriate λ: histogram or QQ-Plot

→ Adaptation: Box-Cox

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases} \quad \Big| \lambda \in \mathbb{R}, x > 0$$

→ only good approx. and not normality can be achieved

Tests, confidence regions

$H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$, $x_1, \dots, x_n \sim N_p(\mu, \Sigma)$

→ Hotelling's T^2

$$T^2 = \underbrace{(\bar{x}-\mu_0)^T \left(\frac{1}{n}S\right)^{-1} (\bar{x}-\mu_0)}_{MD^2(\bar{x}, \mu_0, \frac{1}{n}S)} \sim \frac{(n-1)p}{(n-p)} F_{p, n-p}$$

→ significance level α

$$\alpha = P\left[T^2 > \frac{(n-1)p}{(n-p)} F_{p, n-p, 1-\alpha}\right]$$

→ confidence region

→ ellipsoid determined by all μ for which:

$$T^2 \leq \frac{(n-1)p}{(n-p)} F_{p, n-p, 1-\alpha}$$

Cluster Analysis

Distance Measures

L_1: $d(i,j) = \|x_i - x_j\|_1 = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$

$D = [(d_{i,j})]$

$\begin{bmatrix} 0 & & \\ & \ddots & \\ & & 0 \end{bmatrix}$

L_2: $d(i,j) = \|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$

Hierarchical clustering measures (5)

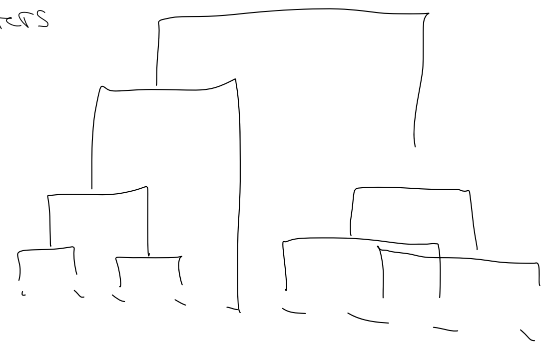
$\max_{i \in C_k, j \in C_l} d(i,j)$

$\min_{i \in C_k, j \in C_l} d(i,j)$

$\frac{1}{n_k n_l} \sum_{i \in C_k, j \in C_l} d(i,j)$

$\|\bar{x}(C_k) - \bar{x}(C_l)\|_2$

$\frac{\|\bar{x}(C_k) - \bar{x}(C_l)\|^2}{\frac{1}{n_k} + \frac{1}{n_l}}$ ... increase in variance when merging two clusters



Partitioning Measures

K-Means

$T = W(C) + B(C)$

$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j \in C_k} d(i,j)^2$ → Cluster centers (μ) to minimize

$B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in C_k, j \notin C_k} d(i,j)^2$

$W(C) = \sum_{k=1}^{K} n_k \sum_{i \in C_k} \|x_i - \bar{x}_k\|_2^2$ <— k-Means

Model-based clustering

$\pi_1 \dots \pi_k$ ... mixing coefficients $\left(\sum_{i=1}^{k} \pi_i = 1\right)$

→ estimate μ, Σ, π using EM

Σ : $p \times p$ → large samples difficult ⇒ $\Sigma_k = \sigma_k^2 I$

$k_1 \dots k_n = a \mid a \in \mathbb{R}$

$k_1 \dots k_n \neq a \mid a \in \mathbb{R}$

Fuzzy clustering – Fuzzy k-Means

$$\min \rightarrow \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^2 \|x_i - m_k\|^2$$

→ k given

→ u_{ik} to minimize

→ ≙ same algo as k-means

with: $$m_{kj} = \frac{\sum_{i=1}^{n} u_{ik}^2 x_{ij}}{\sum_{i=1}^{n} u_{ik}^2}$$

u_{ik} ... membership coeff.

m_{kj} ... weighted cluster center

→ D instead of x can be used

→ clusters spherically shaped

$$B_k = \sum_{k=1}^{K} \| \bar{x}_k - \bar{x} \|^2$$ Heterogeneity

$$W_k = \sum_{k=1}^{K} \sum_{i \in C_k} \| x_i - \bar{x}_k \|^2$$ Homogeneity

---

Validity measures (4) => ~~determine optimal k~~

$$H_k = \ln \frac{B_k}{W_k}$$

↳ identifies point when higher k does not result in better clustering "elbow point"

$$CH_k = \frac{\frac{B_k}{k-1}}{\frac{W_k}{n-k}}$$

↳ higher values ≙ better clustering

avg. silh width

$$s = \frac{1}{n} \sum_{i=1}^{n} s_i$$

-> higher s ≙ better classification

$$s_i = \frac{d_{i,C} - d_{i,C_k}}{\max(d_{i,C}, d_{i,C_k})}$$

$\in [-1, 1]$ ... −1 false clustering, 1 perfect clustering

$$d_{i,C} = \min(d_{i,C_l}) \rightarrow d_{i,C_l} = \frac{1}{n_l} \sum_{j \in C_l} d(i,j)^2$$ ... avg. dissimilarity to other clusters

$$d_{i,C_k} = \frac{1}{n_k - 1} \sum_{i,j \in C_k} d(i,j)^2$$ ... avg. dissimilarity to same cluster

$$Gap_n(K) = E_n^*\{\log(\tilde{W}_k)\} - \log(\tilde{W}_k)$$

-> smallest k that maximises distance to rndm. expectation

E^* ... exp under sample n of appropr. reference dists.

$$\hat{W}_k = \sum_{k=1}^{K} \frac{1}{2n_k} \sum_{i,j \in C_k} d(i,j)^2$$

=> Smallest k, so that $Gap_n(k) \leq 1$ std. error away from 1. local max

$$Gap_n(K) = E_n^*\{\log(\hat{W}_k)\} - \log(\hat{W}_k)$$

Gap indicates how good cluster is compared to rndm. one

$$y = X\beta + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

$\downarrow$

$$y_1 = \beta_0 + \beta_1 x_{11} + \dots \beta_p x_{1p} + \varepsilon_1$$

$\vdots$

$$y_n = \beta_0 + \beta_1 x_{n1} + \dots \beta_p x_{np} + \varepsilon_n$$

$\longrightarrow y = X\beta + \varepsilon$

$y \dots n \times 1$

$X \dots n \times (p+1)$

$\beta = (p+1) \times 1$

$\varepsilon \dots n \times 1$

$E(\varepsilon) = 0 \quad cov(\varepsilon) = \sigma^2 I_n$

$LS = (y - X\beta)^T (y - X\beta) \longrightarrow min \Rightarrow$

$\hat{\beta} = (X^T X)^{-1} X^T y \longrightarrow \hat{y} = X\hat{\beta}$

X : $p+1 \leq n$ ... not full rank $\Rightarrow H = X(X^T X)^{-1} X^T$

$\hat{y} = Hy \longrightarrow y - \hat{y} = \hat{\varepsilon} = (I - H)y$

$E(\hat{\varepsilon}) = 0 \quad cov(\hat{\varepsilon}) = \sigma^2 (I - H)$

$E(\hat{\beta}) = \beta \quad cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

$$s^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n-p-1}$$

$$\left.\begin{array}{l} y_{1j} = \beta_{0j} + \beta_{1j} x_{11} + \dots \beta_{pj} x_{p1} + \varepsilon_{1j} \\ \vdots \\ y_{nj} = \beta_{0j} + \beta_{1j} x_{1n} + \dots + \beta_{pj} x_{pn} + \varepsilon_{nj} \end{array}\right\} Y = XB + \varepsilon$$

$Y \dots n \times m$

$X \dots n \times (p+1)$

$B \dots (p+1) \times m$

$\varepsilon \dots n \times m$

$$LS = (Y - XB)^T (Y - XB) \longrightarrow \hat{B} = (X^T X)^{-1} X^T Y$$

$$\hat{Y} = X\hat{B}$$

$$Cov(\varepsilon_j, \varepsilon_k) = \sigma_{jk}^2 I_n \qquad E(\varepsilon_j) = 0$$

$y \sim N_n(XB, \sigma^2 I) \longrightarrow$ a for β_j and σ^2

$$\Lambda = \left(\frac{|S_R|}{|S_{R_0}|} \right)^{\frac{n}{2}} \quad \text{Wilk's Lambda}$$

$$S_R = \frac{R R^T}{n-p-1} \qquad R = Y - \hat{Y}$$

$$S_{R_0} = \frac{R_0 R_0^T}{n-q-1} \qquad R_0 = Y^* - \hat{Y}_0$$

$$q < p$$

$\longrightarrow$ approximate for large n's to $-2\log\Lambda \sim \chi^2_{m(p-q)}$

→ $IF(x, T, G) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)G + \delta_x \varepsilon) - T(G)}{\varepsilon}$ ← rate of change

→ $maxbias(m, T, x) = \sup_{\tilde{x}} \| T(\tilde{x}) - T(x) \|$

→ $\varepsilon_n^*(T, x) = \min\{ \frac{m}{n} ; maxbias(m, T, x) = \infty \}$

→ efficiency $\approx \frac{1}{Var}$ (+ Fisher Information)

→ M-estimator $n \to x \dots n+(p+1)$

$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} r_i(\beta)^2 \longrightarrow \hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(r_i(\beta))$

→ $\rho(r) = r^2 \dots LS$

$\rho(r) = |r| \dots L1$

→ scale invariance:

$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\beta)}{\hat{\sigma}}\right)$

$\hat{\sigma}$ … robust scale estimator of residuals

→ enhances robustness by getting rid of dependence on data scale

→ M-estimating equations

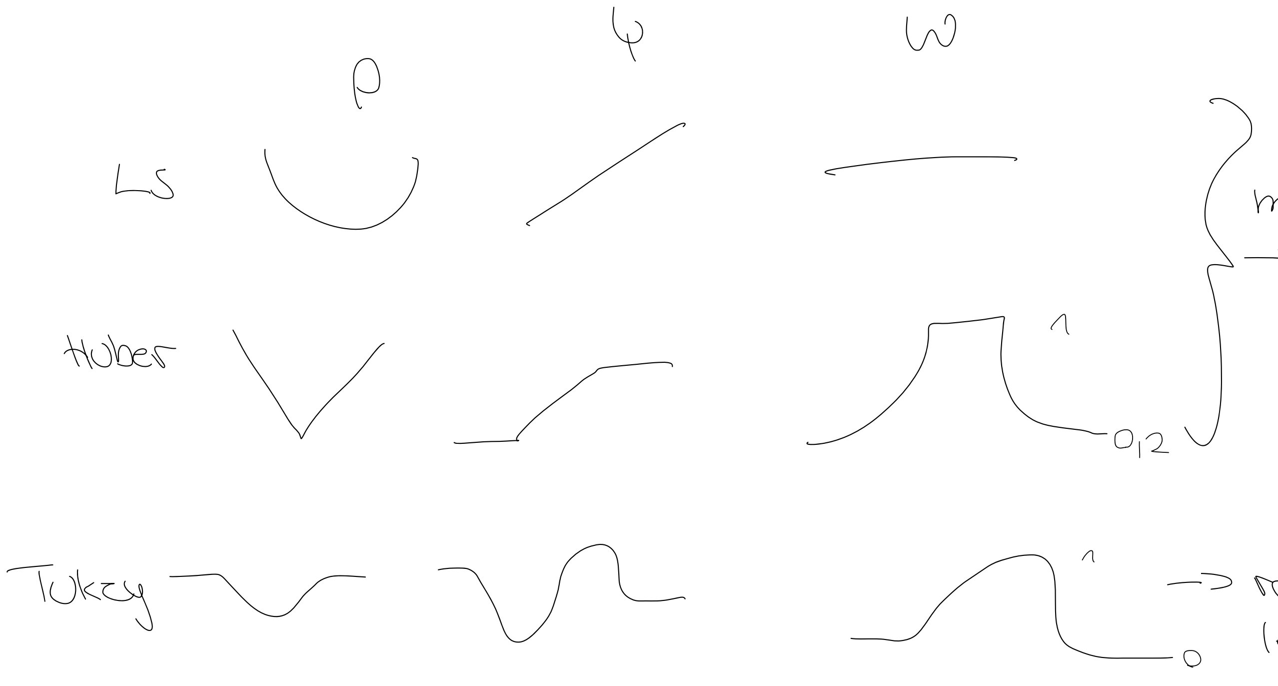
$\sum_{i=1}^{n} \psi\left(\frac{r_i(\beta)}{\hat{\sigma}}\right) x_i = 0 \longrightarrow \psi(r) \dots$ normal equations

$\psi = \rho'$

$w(r) = \psi(r)/r \longrightarrow \sum_{i=1}^{n} r\, w\left(\frac{r_i(\beta)}{\hat{\sigma}}\right) x_i$

$w_i = w(r_i(\beta)/\hat{\sigma}) \longrightarrow \sum_{i=1}^{n} w_i (y_i - x_i^T \beta) x_i = 0$

$r_i = r_i(\hat{\beta}_m)$



monotonic f's
→ ψ increasing

→ robust against leverage points
→ bounded ρ

$$\psi_{H,b}(r) = \begin{cases} r & : |r| \le b \\ b\,\text{sign}(r) & : |r| > b \end{cases}$$

$b = 0 \to L_1$

$b = \infty \to LS$

$$\rho_{B,k}(r) = \begin{cases} \left(\frac{r}{k}\right)^2 \left(3 - 3\left(\frac{r}{k}\right)^2 + \left(\frac{r}{k}\right)^4\right) & : |r| \le k \\ 1 & : |r| > k \end{cases}$$

k... tuning parameter
$k \Rightarrow \infty$ ... LS
→ efficiency vs. robust

Computation

IRWLS: (1) $r_i = r_i(\hat{\beta}_m)$ in iter=0: $\hat{\beta}_m = \hat{\beta}_0$ → m... approx. at iteration m

(2) $w_i = W(r_i/\hat{\sigma})$ ... update weights

(3) $\hat{\beta}_{m+1}$ ... update β, reiterate → $\hat{\beta}_0$ must be robust

→ no leverage points: L_1 as initial $\hat{\beta}$ and $\hat{\sigma}$ computed as robust scale of r_i's (e.g. MAD)

→ leverage points: $\hat{\beta}_0$ cannot use previous res. scale ⇒

Scale Estimators

$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, \hat{\sigma}(r(\beta))$ Based on $\hat{\sigma}$ we differentiate:

→ Estimators using Scales based on ordered values

→ LMS: $\hat{\sigma}(r) = \underset{i}{\operatorname{med}} |r_i|$

→ LTS: $\hat{\sigma}(r) = \left(\frac{1}{h} \sum_{i=1}^{h} |r_i|^2\right)^{\frac{1}{2}}$ $\mid h \in [n/2, n]$

→ S-Estimator

Regression estimators β with σ given by:

→ Scale M-estimators

M-estimator of scale: Solution σ of

$$\sigma \to \frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{r_i}{\sigma}\right) = \delta \quad \mid \delta \in (0, \rho(\infty))$$

$\delta = 1, \rho(z) = z^2 \Rightarrow$ RMSE

$\delta = 0{,}5, \rho(z) = 1(|z| > 1) \Rightarrow \sigma = \operatorname{med}(|r|)$

bp: $\min(\delta, 1-\delta)$

with $\omega_\sigma(z) = \frac{\rho(z)}{z^2} \to \hat{\sigma}^2 = \frac{1}{n\delta} \sum_{i=1}^{n} w_i r_i^2$

... can be solved iteratively

→ computed using IRWLS using initial approximation

→ achieve max breakdown but low efficiency

MM-estimators

$\hat{\beta}_0$ from S-estimator → IRWLS with σ from M-scale ($\rho = \rho_{B,1}$)

→ BP of $\hat{\beta}_0$ and efficiency γ (recommended 0,95)

Affine equivariance

$t(Ax_1+b, \dots, Ax_n+b) = A\,t(x_1, \dots x_n)+B$

$C(Ax_1+b, \dots, Ax_n+b) = A\,C(x_1, \dots x_n)A^T$

d(r)

MCD

h data points for which determinant of empirical covariance matrix minimal $\quad h \approx \frac{n}{2}$ → max Breakdown point
→ low efficiency

t… mean of h obs.

C… given by cov matrix with smallest det, scaled to obtain consistency for normal distribution

Multivariate S-estimator

residuals ⇒ small MD's : $MD^2 := (x-t)^T C^{-1}(x-t)$

→ using M-scale: min: $\hat{\sigma}(MD^2(x_1,t,C), \dots MD^2(x_n,t,C)) \quad |C|=1$

→ low asymptotic efficiency

Multivariate MM-estimator

high ε_n^*, high efficiency ⇒ Affine equivariant, bounded IF

→ using IRWLS

multivariate outlier detection

1) MCD for t and C

2) $MD^2 := (x-t)^T C^{-1} (x-t)$

3) $MD^2 > \chi^2_{p,0.975}$

$h_{ii} = \frac{1}{n-1} MD^2 + \frac{1}{n}$ → MD^2 non robust

$h_{ii} > 2 \cdot \frac{p+1}{n}$ identification of lvg. points

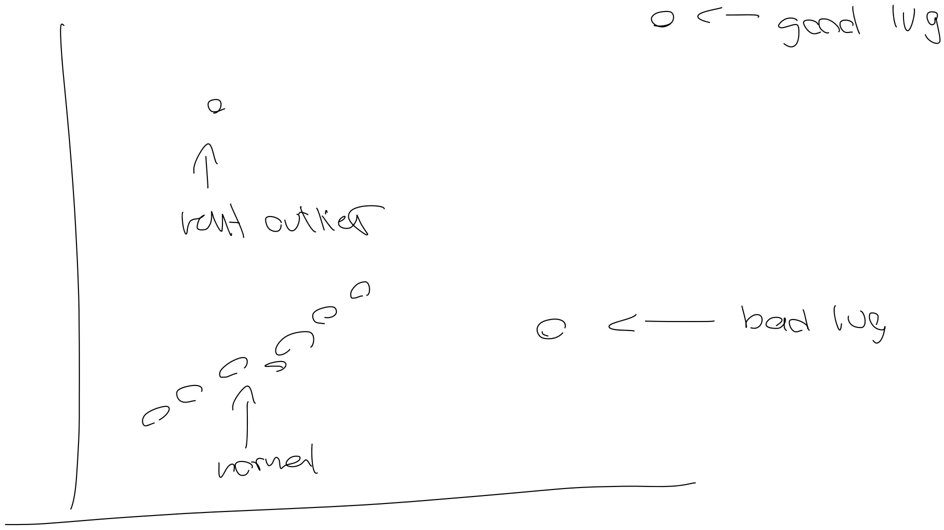
$tr(H) = p+1$ → $avg(h_{ii}) = \frac{p+1}{n}$

→ Use robust $MD^2 := (x-t)^T C^{-1} (x-t)$

$MD^2 > \chi^2_{p,0.975}$

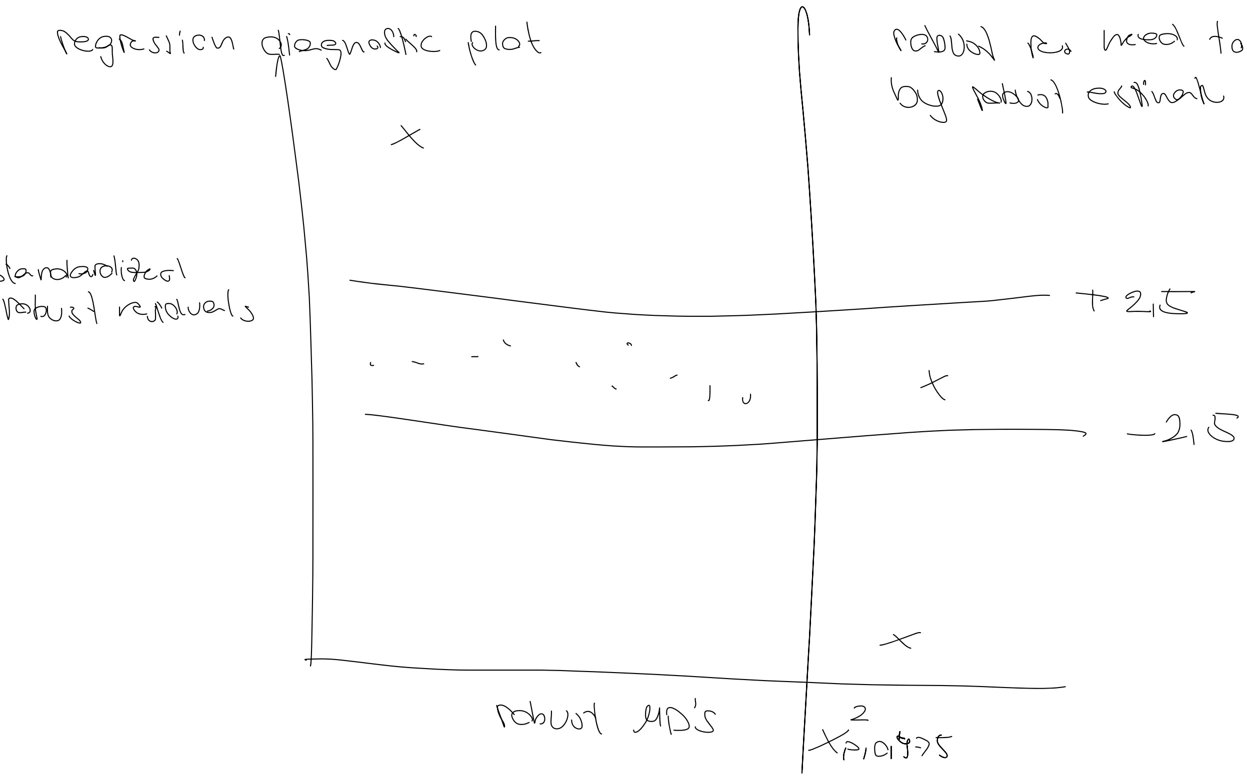
→ not possible to decide between good and bad lvg. using h_{ii}

→ only x-values considered in h_{ii}



regression diagnostic plot

robust res need to be scaled by robust estimate of reg. scale



Robust multivariate regression

$$Y = XB + \mathcal{E}$$

→ robustly estimate B and Σ

→ multivar. S-estimator: MD's based on r_i and C

→ minimize $\hat{\sigma}(r_1^T C^{-1} r_1, \ldots)$ using M-scale $|C| = 1$

PCA

$z = \Gamma^T(x-\mu)$ $\quad x \dots p \times 1$

$\Gamma \dots p \times p$ with $(\gamma_1, \dots \gamma_p)$

$z \dots p \times 1$ … Principal components

$Var(z_i) = \gamma_i^T \Sigma \gamma_i \qquad \gamma_i \gamma_i^T = 1 \,,\, \gamma_i \gamma_j^T = 0 \qquad \Gamma^{-1} = \Gamma^T$

-> Calculation of Principal Components

-> Maximise Var under restrictions => Lagrange optimization

-> $\phi_1 = \gamma_1^T \Sigma \gamma_1 - a_1(\gamma_1^T \gamma_1 - 1)$

-> $\Sigma \gamma_1 = a_1 \gamma_1$ … γ_1 eigenvector of Σ to eigenvalue a_1

$Var(z_1) = \gamma_1^T \Sigma \gamma_1 = \gamma_1^T a_1 \gamma_1 = a_1$

of Σ's p eigenvalues, the largest one is chosen for PC_1 since we want to maximize $Var(z_1)$

-> $\phi_2 = \gamma_2^T \Sigma \gamma_2 - a_2(\gamma_2^T \gamma_2 - 1) - b \gamma_2^T \gamma_1$ … $\gamma_2^T \gamma_1 \overset{!}{=} 0$ for uncorrelatedness

-> $\Sigma \gamma_2 = a_2 \gamma_2$ -> γ_2 eigvector to a_2 which is PC_2 (second largest PC)

-> PC solution

$\Gamma \Sigma = \Gamma A$ -> $\Sigma = \Gamma^T A \Gamma$ … based on decomp of Σ

… $\Gamma = (\gamma_1, \dots \gamma_p) \qquad A = Diag(a_1, \dots a_p)$

↳ $Cor(z_i, z_j) = 0$

->Properties

not scale invariant

$E(z) = \Gamma^T[E(x-\mu)]$ -> if centered: $= 0$

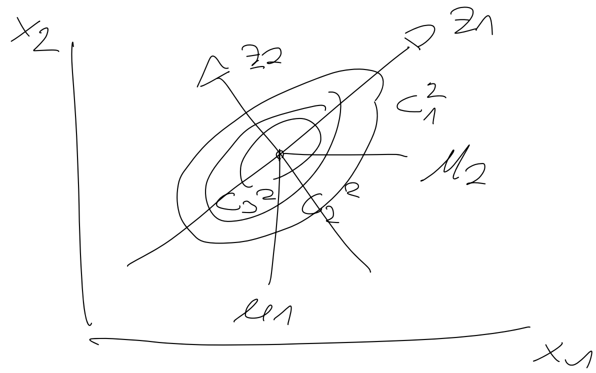
$Cov(z) = \Gamma^T cov(x-\mu)\Gamma = \Gamma^T \Sigma \Gamma = A$ -> different & uncorrelated

$Cov(x, z) = \Gamma A$

$Cor(x, z) = \Lambda = (Diag(\Sigma))^{-\frac{1}{2}} \Gamma A^{\frac{1}{2}}$ ↳ $= \sigma_{ii}^{-\frac{1}{2}}$

Graphical representation

$$(x-\mu)^T \Sigma^{-1}(x-\mu) := c^2$$



$$c^2 = (x-\mu)^T \Sigma^{-1}(x-\mu) = (x-\mu)^T \Gamma A^{-1} \Gamma^T (x-\mu)$$

$$= [\Gamma^T(x-\mu)]^T A^{-1} \Gamma^T (x-\mu) = z^T A^{-1} z = \sum_{i=1}^{p} \frac{z_i^2}{a_i}$$

due to scaling → expansion = $\sqrt{a_i}$

→ components of z represent main axes of ellipsoids
↳ eigenvectors give direction
→ PC1 along largest expansion of ellipsoid and so on
↳ eigenvalues give strength

---

estimation in real world conditions

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\Sigma} = S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

$$Z = (X - 1\bar{x}^T)\hat{\Gamma} \longrightarrow \hat{\Gamma}^T S \hat{\Gamma} = \hat{A}$$

n×p ↵ ↳ p×p

$1 \ldots (1_1 \ldots 1_n)^T \ldots$ needed for dimensionality

$$\hat{\Gamma} = (\hat{\gamma}_1, \ldots \hat{\gamma}_p)$$

$$\mathrm{Cor}(x, z) = \hat{\Lambda} = (\mathrm{Diag}(S))^{-\frac{1}{2}} \hat{\Gamma} \hat{A}^{\frac{1}{2}}$$

Number of relevant PC's

-> total variance: trace($\hat{A}$) ... sum of eigenvalues

(1) Statistical tests

H_0: last p-k PC's contain same variance

-> start with k = 0 until H_0 can't be rejected

$$\left(n - \frac{2p+11}{6}\right)(p-k)\ln\left(\frac{m_a}{m_g}\right) \sim \chi^2_{(p-k+2)(p-k-1)/2}$$

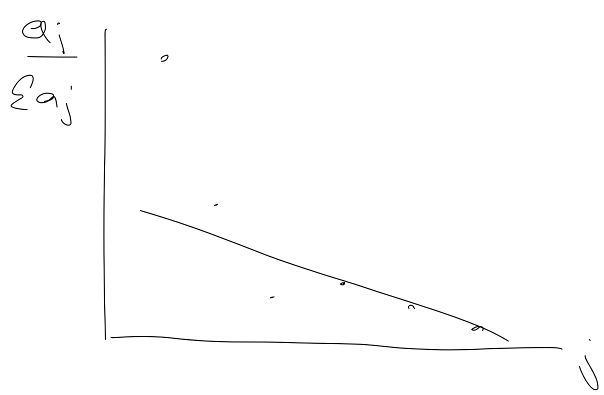
$$m_a = \frac{\hat{a}_{k+1} + \dots \hat{a}_p}{p-k} \qquad m_g = \sqrt[p-k]{\hat{a}_{k+1}, \dots \hat{a}_p}$$

(2) Rules of thumb

-> $\frac{\sum_{j=1}^{K} \hat{a}_j}{\sum_{i=1}^{p} \hat{a}_i} \geq \alpha_i$ $\qquad \alpha_i$ ... 80-90 %

-> exclude PC's with var lower average (1 if standardized)

(3) Scree Graph



(4) Squared correlation coefficient between x and z

$$\lambda_{ij}^2 = \frac{r_{ij}^2 a_j}{s_{ii}}$$ if low for last few PC's => exclude

(5) resampling procedures (e.g.: Jackknife)

-> estimate uncertainty (e.g. of explained variance by k-PC's)

SVD

$X = UDV^T$

$X \ldots n \times p$

$D \ldots n \times p$ $\quad d_{ii} \geq 0$ singular values $\quad i = \text{rank}(X)$

$U \ldots n \times n$, n orthogonal eigenvectors of XX^T in cols

$V^T \ldots p \times p$, p orthogonal eigenvectors of X^TX in cols

$Z = (X - 1\bar{x}^T)\Gamma \longrightarrow$ mean centered: $\hat{Z} = X\Gamma \longrightarrow X = Z\Gamma^T$

$\longrightarrow X = UDV^T = Z\Gamma^T$

$(X-\mu)^T(X-\mu) = X^TX$ since centered

$S = \frac{1}{n-1} X^TX = \hat{\Gamma}\hat{\Lambda}\hat{\Gamma}^T \longrightarrow \Gamma$ matrix with normed eigenvectors of S which in this case is V (of X^TX) $\longrightarrow \hat{\Gamma} \equiv V$

$\hat{=} \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-n}$

$\Longrightarrow X = UDV^T = ZV^T \longrightarrow Z = UD$

$\|X\|_F = \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2}$ ← measure of total power of matrix

$XV = UD = Z$ transformation matrix

$X \ldots n \times p$

$V_m \ldots n \times m$

$\text{rank}(B) \leq m < \text{rank}(X) \quad B^TB = 1 \quad B \ldots p \times m$

$\longrightarrow$ "maximising variance of projected data"

$V_m = \underset{B}{\text{argmax}} \|XB\|_F^2$

mapping of X onto lower dim space $\quad n \times p \Rightarrow n \times m$

↳ maximising power of projection $\hat{=}$ maximising variance

$\longrightarrow$ "minimising RSS between og. and projection"

$X = XVV^T = XV_mV_m^T + \varepsilon$

$V_m = \underset{B}{\text{argmin}} \|X - \tilde{X}\|_F^2 \qquad \tilde{X} = XBB^T$

↳ minimizing F of rs $\hat{=}$ minimizing power difference

Biplots

$$X \approx X_{(2)} = UDV^T = (u_1\ u_2)\begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}\begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix}$$

$$X_{(2)} = GH^T$$

$$G = (u_1\ u_2)\begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{c-1}$$ → gives coord's of data points in reduced space

$0 \leq c \leq 1$ ... determines distribution of singular values to G and H

$$H = (v_1\ v_2)\begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{c}$$ → determines direction and magnitude of arrows indicating og. var's

obtained for $c = 1$ (+ rescaling):

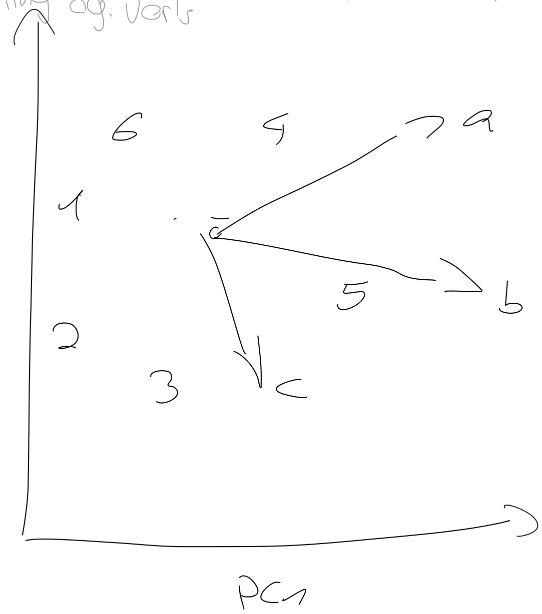
→ $g_i^T h_j \approx x_{ij}$

→ $HH^T \approx S = \frac{1}{n-1} X^T X$

→ $\|h_j\|^2 \approx Var(x_j)$

→ $\cos(h_i, h_j) \approx r_{ij}$

→ $\|g_i - g_j\|^2 \approx MD^2 = (x_i - x_j)^T S^{-1} (x_i - x_j)$



Diagnostics

$$SD_i = \left(\sum_{j=1}^{k} \frac{z_{ij}^2}{\hat{a}_j}\right)^{\frac{1}{2}}$$ ... only k relevant PC's considered

.. $\hat{=}$ MD score vector to PCA center with respect to covariance Matrix A

$$OD_i = \|x_i - \hat{\Gamma}_k z_i\|_2$$

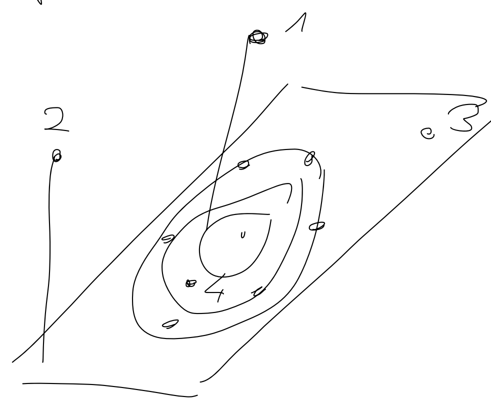
$\hat{=}$ $\|\ \|_2$ of observation to projection into space of first k PC's

1: big OD, low SD → vertical outlier

2: big OD, big SD → bad leverage point

3: low OD, big SD → good leverage point

4: low OD, low SD → good point



→ outlier detection:

$$SD_i > \sqrt{\chi^2_{k,0.975}}$$

$$OD_i > \left(\text{median}_i\left(OD_i^{\frac{2}{3}}\right) + MAD\left(OD_i^{\frac{2}{3}}\right) z_{0.975}\right)^{\frac{3}{2}}$$ … $OD^{\frac{2}{3}}$ closer to normality

$z_{0.975}$ … 0.975 quantile of $N(0,1)$

$$MAD = 1.483 \cdot \text{median}_i\left(\left| y_i - \text{median}_j(y_j)\right|\right)$$

Factor Analysis

$y_i = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}} \mid i = 1, \dots, p$ ... data standardized

$y \dots (y_1, \dots, y_p)^T$ $p \times 1$

$f \dots (f_1, \dots, f_k)^T$ $k < p$ factors

$y = \Lambda f + e$

$\Lambda = [(\lambda_{ij})] \dots p \times k$ loadings matrix

$e \dots (e_1, \dots, e_p)^T$ $p \times 1$ unique factor

$E(f) = 0 = E(e)$

$cov(e_i, e_j) = 0 = cov(f, e)$ $i \neq j$ $\longrightarrow$ $cov(e) = \Psi = Diag(\psi_{11}, \dots, \psi_{pp})$

$Var(f_i) = 1$

data standardized

$P = cor(x) = cov(y) = cov(\Lambda f + e) =$

$= \Lambda\, cov(f) \Lambda^T + \Lambda \underbrace{cov(f, e)}_{0} + \underbrace{cov(e, f)}_{0} \Lambda^T + cov(e) = \Lambda \Phi \Lambda^T + \Psi$

factors uncorrelated $\Rightarrow \Phi = I \Rightarrow P = \Lambda\Lambda^T + \Psi$

$P_{red} = P - \Psi = \Lambda\Lambda^T = \begin{bmatrix} k_1^2 & \cdots & \rho_{1p} \\ \rho_{21} & \ddots & \vdots \\ \vdots & & \vdots \\ \rho_{p1} & \cdots & k_p^2 \end{bmatrix} \longrightarrow k_i^2 = 1 - \psi_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 \dots$ communalities

$\lambda_{ij} \hat{=}$ % variance explained by factors

Non-uniqueness

$\Rightarrow$ factors loadings not unique $\Rightarrow$ additional constraints for uniqueness:

$\Lambda^T \Psi^{-1} \Lambda$ or $\Lambda\Lambda^T$ ... diagonal

---

\# parameters

upper bound for k: $s > 0 \rightarrow$ fewer parameters than corr. matrix

$s =$ corr. matrix $-$ (factor model $-$ restrictions) $= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)$

Parameter estimation using PFA (principal factor analysis)

→ communalities

(1) highest correlation coefficient

$$\max_{i \neq j} |\hat{\rho}_{ij}|$$ → may indicate strongest lin. relationship

(2) Squared multiple correlation coefficient

$$\hat{\rho}^2_{i|1,2,\dots,p} = 1 - \frac{1}{\hat{\rho}^{ii}}$$ → var. proportion of i-th var explained by other var's

(3) Iterative estimation

1) fix k (criterion from PCA on # components)

2) initialize $\hat{k}_i^2$ using (1) or (2)

3) replace $\hat{\rho}_{ii}$ with $\hat{k}_i^2$

4) estimate $\hat{\Lambda}$ (see below)

5) re-estimate $\hat{k}_i^2 = \sum_{j=1}^{k} \hat{l}_{ij}^2 \quad | \; i = 1, \dots, p$

6) repeat until stable

→ loadings

$$\rho_{red} = \hat{\Lambda}\hat{\Lambda}^T = \hat{\rho} - \hat{\Psi} = \hat{\Lambda}\hat{\Lambda}^T = \underbrace{\hat{\Gamma}_{1:k}\hat{A}_{1:k}\hat{\Gamma}_{1:k}^T}_{\text{model}} + \sum_{i=k+1}^{p} a_i \gamma_i \gamma_i^T \Rightarrow \hat{\Lambda} = \hat{\Gamma}_{1:k}\hat{A}_{1:k}^{\frac{1}{2}}$$

→ update uniqueness

$$\hat{\Psi}_{ii} = 1 - \sum_{j=1}^{k} \hat{l}_{ij}^2$$ → valid if $\hat{\Psi}_{ii} > 0$

Rotation

$$\sum_{s<j=1}^{k} \sum_{i=1}^{p} (\lambda_{is} \lambda_{ij})^2 \rightarrow min$$ ⟹ loadings matrix simple: many small absolute values, only a few large ones

→ orthogonal rotation

orthogonal $k \times k$ matrix T → $\hat{\Lambda} = \Lambda T$

→ factors uncorrelated (90° to each other)

→ do not change k

→ $(k_i^2)^2 = const$

} ≙ maintains overall structure of factor solution in terms of variance

→ sum over all variables ≠ const. ⟹ maximise one term ≙ minimise the other

(1) $QMAX = \sum_{i=1}^{p} \sum_{j=1}^{k} \hat{\lambda}_{ij}^4$ ⟹ one dominant factor

(2) $VMAX = p \sum_{j=1}^{k} \sum_{i=1}^{p} \left(\frac{\hat{\lambda}_{ij}}{k_i} \right)^4 - \sum_{j=1}^{k} \left[\sum_{i=1}^{p} \left(\frac{\hat{\lambda}_{ij}}{k_i} \right)^2 \right]^2$ ⟹ max ≙ absolute large and small loadings

(p: #var's)

→ normalize with communalities, that dominate

→ maximises variance of squared loadings of each factor

→ oblique rotation

→ factors no longer uncorrelated ⟹ Φ needs to be considered

→ $\hat{\Lambda} = \Lambda T$ ... T no longer orthogonal

→ $P_{red} = \hat{\Lambda} Cov(\hat{f}) \hat{\Lambda}^T$ → $\hat{f} = T^{-1} f$ ⟹ T must be invertible

(1) $QMIN = \sum_{s<j=1}^{k} \sum_{i=1}^{p} \hat{\lambda}_{is}^2 \hat{\lambda}_{ij}^2$ ⟹ discourages multiple large loadings

(2) $OBMIN = \sum_{s<j=1}^{k} \left(\sum_{i=1}^{p} \hat{\lambda}_{is}^2 \hat{\lambda}_{ij}^2 - \frac{\gamma}{p} \sum_{i=1}^{p} \hat{\lambda}_{is}^2 \sum_{i=1}^{p} \hat{\lambda}_{ij}^2 \right)$ ⟹ high or low loadings but not medium

→ $\gamma = 0$ → QMIN

→ $\gamma = 1$ → COVARMIN

Estimation of factor scores

→ Bartlett

$y = \Lambda f + e$ ... heteroscedastic ($Cov(e)$ not equal) ⇒ · $\Psi^{-\frac{1}{2}}$ (weights)

⇒ $\Psi^{-\frac{1}{2}} y = \Psi^{-\frac{1}{2}} \Lambda f + \Psi^{-\frac{1}{2}} e$ → $\Psi^{-\frac{1}{2}} Cov(e) \Psi^{-\frac{1}{2}} = \Psi^{-\frac{1}{2}} \Psi \Psi^{-\frac{1}{2}} = I$ ⇒ homoscedastic

⇒ $\hat{f} = (\Lambda^T \Psi^{-1} \Lambda)^{-1} \Lambda^T \Psi^{-1} y$ ... used when aim is to reproduce factors as close to original as possible

($\hat{f} \sim \beta$, $\Lambda^T \Psi^{-1} \sim X^T$, $\Lambda \sim X$, $\Lambda^T \Psi^{-1} \sim X^T$, $y \sim y$)

⇒ $\hat{F} = Y \Psi^{-1} \Lambda (\Lambda^T \Psi^{-1} \Lambda)^{-1}$

→ Regression

$f = By + \delta$ regress f on y

→ $\hat{B} = f y^T (y y^T)^{-1}$ ⇒ $\hat{f} = \hat{B} y = f y^T (y y^T)^{-1} y$

→ $\hat{f} = f (\Lambda f + e)^T (y y^T)^{-1} y = (\Lambda f f^T + e f^T)(y y^T)^{-1} y = \Lambda f f^T (y^T y)^{-1} y$

($e f^T$: $Cov(e, f) = 0$)

→ sample version: $\hat{F} = \Lambda \frac{\hat{F} \hat{F}^T}{n-1} \frac{n-1}{(Y^T Y)} Y =$

$= Y R^{-1} \Lambda \hat{\Phi} \rightarrow Y R^{-1} \Lambda$ ... used for orthogonal factors

(R: Correlation matrix of observed variables (y); $\hat{\Phi} = I$ for orthogonal)

Correlation analysis

multiple correlation analysis

→ dependency of feature x on p-dim. feature y

→ minimize $MSE = E(x - a_0 - a^T y)^2$

$x \ldots 1 \times 1$
$y \ldots 1 \times p$

Linear prediction function

→ $a_0 + a^T y$

→ $a_0 = \mu_x - a^T \mu_y$ $\quad a = \Sigma_{yy}^{-1} \sigma_{yx}$

→ $\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$ $\quad \Sigma = \begin{pmatrix} \sigma_{xx} & \sigma_{yx}^T \\ \sigma_{yx} & \Sigma_{yy} \end{pmatrix}$

→ $MSE = \sigma_{xx} - \sigma_{yx}^T \Sigma_{yy}^{-1} \sigma_{yx}$

→ linear prediction function of x with minimal MSE and max corr with x

$$corr(x, a_0 + a^T y) = \sqrt{\frac{\sigma_{yx}^T \Sigma_{yy}^{-1} \sigma_{yx}}{\sigma_{xx}}}$$

multiple correlation coefficient:

→ correlation between x and best linear predictor function

$$\rho_{x,y} = \sqrt{\frac{\sigma_{yx}^T \Sigma_{yy}^{-1} \sigma_{yx}}{\sigma_{xx}}}$$

→ $\rho_{x,y}^2$ ... multiple coefficient of determination

→ indicates how well feature x is explained by properties $y = (y_1, \ldots y_p)^T$

Test for hypothesis multiple correlation is zero $\hat{=}$ bivariate correlations zero

$H_0: \rho_{x,y} = 0$

assume normal distribution, significance lvl α, n... # observations in sample

$$F = \frac{(n-1-p)\, r_{x,y}^2}{p(1 - r_{x,y}^2)} > F_{p, n-1-p, 1-\alpha}$$

Canonical correlation

$x \dots 1 \times p$ $\quad y \dots 1 \times q$ $\quad p \leq q$

→ linear dependence between two group of variables ⇒ subspace (one corr. coeff does no longer suffice)

→ k-th canonical pair of variables

$$\varphi_k = a_k^T x = e_k^T \Sigma_{11}^{-\frac{1}{2}} x \qquad \eta_k = b_k^T y = f_k^T \Sigma_{22}^{-\frac{1}{2}} y$$

… linear combinations

with:

$$\Sigma_{11} = E[(x-\mu_1)(x-\mu_1)^T]$$

$$\Sigma_{22} = E[(y-\mu_2)(y-\mu_2)^T]$$

$$\Sigma_{12} = \Sigma_{21}^T = E[(x-\mu_1)(y-\mu_2)^T]$$

→ maximise $\mathrm{Corr}(\varphi_k, \eta_k) = \rho_k$ over all linear comb. uncorrelated with last $1, 2, \dots, k-1$ canonical variables

→ eigenvalues and eigenvectors

$\rho_1^2 \geq \dots \rho_p^2$ a_i of $\Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}}$ → respective γ_i: $e_1, \dots e_p$

→ ≥ … eigenvalues sorted in descending corr. magnitude

$\rho_1, \dots \rho_p$: p largest a_i of $\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$ → respective γ_i: $f_1, \dots f_p$

→ Every f_i proportional to $\Sigma_{22}^{-\frac{1}{2}} \Sigma_{21} \Sigma_{11}^{-\frac{1}{2}} e_i$

→ standard: same order for a_i's as above

→ solution is derived from eigenvalues of combined covariance matrices, thus an analysis of covariance structures between the sets

→ Proof: using Cauchy-Schwarz Inequality and reformulation as Eigenvalue Problem

⇒ eigenvalues of combined matrix represent directions in which corr. is maximised

→ Properties

$\mathrm{Var}(\varphi_k) = \mathrm{Var}(\eta_k) = 1$

$\mathrm{cov}(\varphi_k, \varphi_l) = \mathrm{cov}(\eta_k, \eta_l) = \mathrm{cov}(\varphi_k, \eta_l) = 0 \qquad k \neq l$

$$\Rightarrow \mathrm{Cov}\begin{pmatrix} \varphi \\ \eta \end{pmatrix} = \begin{pmatrix} I & P \\ P & I \end{pmatrix}$$

$\varphi \dots (\varphi_1, \dots \varphi_p)^T$

$\eta \dots (\eta_1, \dots \eta_p)^T$

$P \dots \mathrm{Diag}(\rho_1, \dots \rho_p)$

$\rho_1 = 1$?

-> at least one $x_i = y_i$

-> indicates perfect linear relationship between linear comb. of sets x and y

-> implications: linear-comb. of these variables completely correlated -> lin. dependency between x and y, no conclusions on relationships of other variables can be made if
$x = (1, 0, \dots, 0)$ $y = (1, 0, \dots, 0)$

---

canonical cor. coeff. invariant to transformations

$x^* = U^T x + u$ $y^* = V^T y + v$ $U \dots p \times p$ $V \dots q \times q$

$u \dots 1 \times p$ $v \dots 1 \times q$

-> can. corr. between of x, y = between x^*, y^*

-> $a_i^* = U^{-1} a_i$ $b_i^* = V^{-1} b_i$

Tests

-> Likelyhood ratio test $H_0: \Sigma_{12} = 0$

$\Lambda^{2/n} = |I - S_{22}^{-1} S_{21} S_{11}^{-1} S_{12}| = \prod_{i=1}^{p} (1 - r_i^2) \longrightarrow \Lambda(q, n-1-p, p)$ Wilks' distr.

↑ sample canonical correlation coefficient

-> Bartlett's approximation

-> for large n instead of Λ

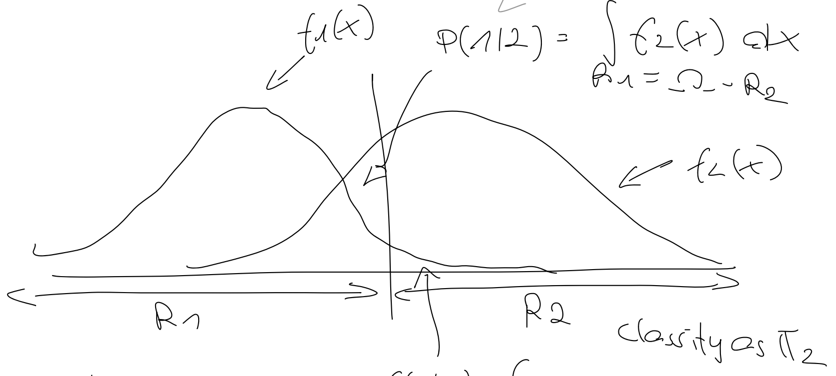
$$-\left[n - \frac{1}{2}(p+q+3)\right] \ln \prod_{i=1}^{p} (1 - r_i^2) \sim \chi^2_{pq}$$

-> Test only s can. cor. are $\neq 0$

$$-\left[n - \frac{1}{2}(p+q+3)\right] \ln \prod_{i=s+1}^{p} (1 - r_i^2) \sim \chi^2_{(p-s)(q-s)}$$

Discriminant Analysis

Satz von Bayes: prob's updated based on new evidence → $P(A|B) = P(B|A) \cdot P(A)$



$P(1|2) = \int_{R_1 = \Omega - R_2} f_2(x)\,dx$

$f(2|1) = \int_{R_2 = \Omega \cdot R_1} f_1(x)\,dx$

R_x ... space of observations to which we assign objects π_x

π_x ... "real" population of x

$f(2|1) = P(x \in R_2 | \pi_1)$ ... x is falsely classified as π_2

→ Application of Bayes Theorem

$$P(\pi_i|x) = \frac{P(x|\pi_i)P(\pi_i)}{\underbrace{\sum_j P(x|\pi_j)P(\pi_j)}_{=1}} \rightarrow P(x \in \pi_i) = P(x \in R_i|\pi_i)P(\pi_i)$$

→ Prior propability: p_i ... prop that obj. comes from π_i ... $P(\pi_i)$

→ ECM: $c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$

→ class. rule to minimize ECM

R_1, obs which: $\frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)\,p_1}{c(2|1)\,p_2}$ R_2: $<$

→ TPM: $p_1 P(2|1) + p_2 P(1|2)$

g = 2

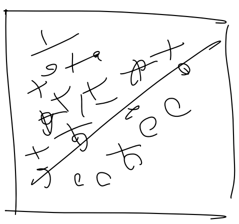
$P(x \in \pi_1) = P(x \in R_1|\pi_1)P(\pi_1)$

$\Sigma_1 = \Sigma_2 = \Sigma$

→ minimizing ECM for $\pi_1, \pi_2 \sim N(\mu_i, \Sigma)$

(1) x_0 to π_1: $(\mu_1 - \mu_2)^T \Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 + \mu_2) \geq \ln \frac{c(1|2)\,p_1}{c(2|1)\,p_2}$

→ linear in x



→ sample version

$$S_{pooled} = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$$

→ unbiased estimate of Σ, $\hat{=}$ cov. estimate of group data

(1b) x_0 to Π_1: $(\bar{x}_1 - \bar{x}_2) S_{pooled}^{-1} x_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T \Sigma^{-1}(\bar{x}_1 + \bar{x}_2) \geq \ln \frac{c(1|2)p_1}{c(2|1)p_2}$

→ Simplification

→ summands scalars; if $\ln\left(\frac{c(1|2)p_1}{c(2|1)p_2}\right) = 1$ → $\ln(1) = 0$, then:

$\hat{y} := \underbrace{(\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1}}_{a^T} x = a^T x$

$\hat{m} := \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$

→ $\bar{y}_1 = (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1} \bar{x}_1 = a^T \bar{x}_1$

→ $\bar{y}_2 = (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1} \bar{x}_2 = a^T \bar{x}_2$

} y results from linear comb. of observations

⇒ x_0 to Π_1: $\hat{y}_i \geq \hat{m}$ ⇒ linear decision boundary m

⇒ for p-dim features → 1.dim variable y

$\Sigma_1 \neq \Sigma_2$

→ minimizing ECM for $\Pi_1, \Pi_2 \sim N(\mu_i, \Sigma_i)$

(2) x_0 to Π_1: $-\frac{1}{2}x_0^T \overbrace{(\Sigma_1^{-1} - \Sigma_2^{-1})}^{\Sigma \text{ alone}} x_0 + \overbrace{(\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1})}^{\mu^T \Sigma} x_0 - k \geq \ln \frac{c(2|1)p_1}{c(1|2)p_2}$

with $k = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}\underbrace{(\mu_1^T \Sigma_1^{-1}\mu_1 - \frac{1}{2}\mu_2^T \Sigma_2^{-1}\mu_2)}_{\mu^T \Sigma \mu \text{ together}}$

→ quadratic in x_0

Evaluation

→ training set

(1) $AER = p_1 \int_{\tilde{R}_2} f_1(x)\,dx + p_2 \int_{\tilde{R}_1} f_2(x)\,dx$

(2) $\frac{\#\text{misclassified}}{\text{total}}$

→ no training set

(3) jackknife

→ classify Π_1, Π_2 multiple times, omitting an observation as test case

estimated error rate: $\frac{\overline{n_1} + \overline{n_2}}{n_1 + n_2}$, $\hat{P}(1|2) = \frac{\overline{n_2}}{n_2}$ $\hat{P}(2|1) = \frac{\overline{n_1}}{n_1}$

(4) k-fold cross validation

→ k parts, k−1 parts computed, last part for evaluation

Fisher

→ projected onto straight line → $\hat{a}$ varied until max separation

$\frac{|\bar{y}_1 - \bar{y}_2|}{s_y} \to \max$ → pooled variance: $S_y^2 = \frac{1}{n_1 + n_2 - 2} \sum_{g=1}^{2} \sum_{i \in n_g} (x_{ig} - \bar{x}_g)^2$

$\hat{y} = \hat{a}^T x = (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1} x$ maximises $\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{a}^T \bar{x}_1 - \hat{a}^T \bar{x}_2)^2}{\hat{a}^T S_{pooled} \hat{a}}$ over all $\hat{a}$

(s_y^2: pooled variance of y-values)

→ maximum ratio: $D^2 := (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2)$ … $\in D^2$

→ transform multivariate ob's to univariate, maximising separation

→ classification rule

x_0 to Π_1: $\hat{y}_0 \geq \hat{m} := (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1} (\bar{x}_1 + \bar{x}_2)$ … $\hat{y}_0 = \hat{y}$ with x_0

$g > 2$

areas minimizing ECM: x to π_k if $\sum_{l=1, k \neq i}^{g} p_i f_i(x) c(k|i)$ minimal

→ c equal for all ⇒ x to π_1 if $p_k f_k(x) > p_i f_i(x) \; \forall i \neq k$

→ p_i, c, f must be known

$N(\mu_i, \Sigma_i)$ assumed

→ $\Sigma_1 = \Sigma_2 = \Sigma_n$

similar to (1) with μ_k instead of $\mu_1 \pm \mu_2$

(3) $d_k(x) = \mu_k^T \Sigma^{-1} x - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln p_k$ largest $d_1(x) \ldots d_g(x)$

→ pooled version (no training set)

$$S_{pooled} = \frac{1}{\sum_{i=1}^{g} n_i - g} \sum_{l=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T$$

(3b) x to π_k if: $\hat{d}_k(x) = \bar{x}_k^T S_{pooled}^{-1} x - \frac{1}{2} \bar{x}_k S_{pooled}^{-1} \bar{x}_k + \ln p_k$ largest $\hat{d}_1(x) \ldots \hat{d}_g(x)$

→ $\Sigma_1 \neq \Sigma_2 \neq \Sigma_n$

(4) x to π_k if: $d_k^Q(x) = -\frac{1}{2} \ln |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \ln p_k$ largest $d_1^Q \ldots d_g^Q$

first term of k — MD^2

→ estimated from training set

(4b) $\hat{d}_k^Q(x) = -\frac{1}{2} \ln |S_i| - \frac{1}{2} |x - \bar{x}_i| S_i^{-1} (x - \bar{x}_i) + \ln p_i$ largest $\hat{d}_1^Q \ldots \hat{d}_g^Q$

S_i … sample covariance matrix

Jackknife

$$\frac{\sum_{i=1}^{g} \overline{n_i}}{\sum_{i=1}^{g} n_i}$$

Fisher

→ $\Sigma_1, \ldots, \Sigma_g = \Sigma$

$$\frac{\sum_{i=1}^{g} p_i (\mu_{i,y} - \bar{\mu}_y)^2}{\sigma_y^2} \longrightarrow \max$$

← weighted sum of squared distances of group- to the total mean

σ_y^2 ↳ variance

$\bar{\mu}_y = a^T \bar{\mu} = a^T \sum_{i=1}^{g} p_i \mu_i$

$\mu_{i,y} = E(y | x \in \Pi_i) = a^T E(x | x \in \Pi_i) = a^T \mu_i$

$\sigma_{y,i}^2 = Var(y | x \in \Pi_i) = a^T Cov(x | x \in \Pi_i) a = a^T \Sigma_i a \Rightarrow \sigma_y^2 = a^T \Sigma a$

↑ $\Sigma_1 = \ldots \Sigma_h$

→ $\Sigma_1, \ldots, \Sigma_g \neq \Sigma$

→ resulting classification rule non-optimal ⇒ replace Σ by pooled version

→ solution of maximization problem given by $a_1, \ldots, a_l$ of $W^{-1}B$ ← here eigenvectors!

→ must be scaled: $a_i^T W a_i = 1 \mid i = 1, \ldots, l$

→ $l \leq \min(g-1, p)$ … l = # strictly positive eigenvalues … # significant directions

max rank $W = g-1$ ↲ ↳ max rank B: # var's p

$$\Rightarrow \frac{a^T B a}{a^T W a} \mid a \in \mathbb{R}^p, a \neq 0$$

→ variation within groups: $W = \sum_{i=1}^{g} p_i \Sigma_i$

→ variation between groups: $B = \sum_{i=1}^{g} p_i (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T$

→ discriminant function: $y = a_j^T x$ … projection of rndm var x in direction a_j

→ classification rule

x to Π_k: if $d_i^F(x) = \sum_{j=1}^{l} (y_j - \mu_{i,y_j})^2 - 2\log p_i$ smallest $d_1^F(x), \ldots, d_g^F(x)$

(Euclidean distance in discriminant space)

$y_j = a_j^T x$

$\mu_{i,y_j} \hat{=} a_j^T \mu_i$

→ also: $\sum_{j=1}^{l} (a_j^T (x - \mu_j))^2 - 2\log p_i = (x - \mu_i)^T A A^T (x - \mu_i) - 2\log p_i$

$\hat{=} MD^2$ in original space

Symbolic Data Analysis

$Y_j : S \to B_j$

$S = \{s_1, \dots, s_n\}$ ... set of n units to be analysed

$Y_1, \dots, Y_p$ ... variables

O_j ... underlying domain of Y_j
B_j ... observation space of Y_j
$\Big\} j = 1, \dots, p$

→ Y_j single valued variable $B_j = O_j$

→ Y_j multi valued variable $B_j = P(O_j)$

→ Y_j interval variable (Numeric) $Y_j : S \to B : Y_j(s_j) = [l_{ij}, u_{ij}], \ l_{ij} \leq u_{ij}$

→ Y_j categorical modal or histogram variable: $Y_j : B_j$ set of distributions on O_j

→ interval variable special case of histogram with $H_{Y(s_i)} = ([l_{ij}, u_{ij}], 1)$

Parametric Models

→ probabilistic models for interval variables

$Y_j : S \to B : Y_j(s_i) = [l_{ij}, u_{ij}], \ l_{ij} \leq u_{ij} \mid \forall s_i$

	Y_1	...	Y_p
s_1	$[l_{11}, u_{11}]$	...	$[l_{1p}, u_{1p}]$
⋮	⋮	⋮	⋮
s_n	$[l_{n1}, u_{n1}]$	...	$[l_{np}, u_{np}]$

→ Represent $Y_j(s_i)$ using:

- midpoint $c_{ij} = \frac{l_{ij} + u_{ij}}{2}$
- range $r_{ij} = u_{ij} - l_{ij}$

Gaussian model

→ assume: joint distribution of c's and logs of r's is multivariate normal

$(C, R^*) \sim N_{2p}(\mu, \Sigma)$

$\Sigma_{xy} \dots p \times p \quad \mu_C, \mu_R \dots 1 \times p$

with $\mu = [\mu_C^T, \mu_R^T]$ $\quad \Sigma = \begin{bmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{bmatrix}$ $\quad R^* := \ln(R)$

→ advantage: simple application of inference methods

Skew-Normal model

→ same assumption

→ more general model, less limitations

$(C, R^T) \sim SN_{2p}(\xi, \Omega, \alpha)$

ξ … location $1 \times p$

Ω … scale, $p \times p$, symmetric positive definite

α … shape (additional), $1 \times p$

↳ indicates skewness (e.g. $\alpha > 0$ → longer right tail)

$f(y; \alpha, \xi, \Omega) = 2\,\phi_p(x - \xi; \Omega)\,\Phi(\alpha^T \omega^{-1}(x - \xi)) \quad | \; x \in \mathbb{R}^p$

ω … $\mathrm{Diag}(\sqrt{\Omega_{ii}})$

ϕ_p … dens. of p-dim std. Gaussian vector

Φ … Distr. func of std. norm var

↳ becomes 0,5 for $\alpha = 0$ ⇒ then $SN_{2p} = N$

Covariance structures

→ c_{ij}, r_{ij}: two quantities related to one variable → dependencies?

→ Parametrization of covariance matrix which takes this into account

1) Non-restricted

$$\begin{bmatrix} \Sigma_{CC} & C_{CR^*} \\ \Sigma_{R^*C^*} & \Sigma_{R^*R^*} \end{bmatrix}$$

no-restrictions

2) i_j's non correlated

$$\begin{bmatrix} \ddots & 0 & \ddots & 0 \\ 0 & \ddots & 0 & \ddots \\ \ddots & 0 & \ddots & 0 \\ 0 & \ddots & 0 & \ddots \end{bmatrix}$$

p 2×2 blocks

Diag(all)

3) C's non correlated with r's

$$\begin{bmatrix} m & 0 \\ 0 & m \end{bmatrix}$$

2 p×p blocks

$\Sigma_{CR^*} = 0 = \Sigma_{R^*C}$

4) All C's and r's non correlated (incl. with self)

$$\begin{bmatrix} \ddots & 0 & 0 \\ 0 & 0 & \ddots \end{bmatrix}$$

2p blocks … single real elements

(3) & $\mathrm{Diag}(\Sigma_{CC})$, $\mathrm{Diag}(\Sigma_{R^*R^*})$

ML-estimation for interval data

→ Gaussian

empirical covariance matrix

$$\ln L(\mu, \Sigma) = -np\,(\ln(2\pi)) - \frac{n}{2}(\bar{x}-\mu)^T \Sigma^{-1}(\bar{x}-\mu) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\text{tr}\,\tilde{\epsilon}\,\Sigma^{-1}$$

holds always due to $N(\mu, \Sigma)$; MD^2

→ Σ^{-1} positive definite ⇒ ML estimate of μ is $\bar{x}$, reduces ML to:

$$\ln L(\mu, \Sigma) = \text{const.} - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\text{tr}(S_i^{-1})$$

→ Σ structures ⇒ max L = separately maximising with respect to each block Σ

→ Skew-Normal

≙ Gaussian version; additional term

$$l = \text{constant} - \frac{n}{2}\ln|\Omega| - \frac{n}{2}\text{tr}(\Omega^{-1}V) + \sum_i \left(\zeta_0\left(\alpha^T\omega^{-1}(x_i - \xi)\right)\right)$$

$$V = \frac{1}{n}\sum_i\left[(x_i-\xi)(x_i-\xi)^T\right] \approx S \qquad \zeta_0(x) = \ln(2\Phi(x))$$

↑ Population covariance matrix since $\frac{1}{n}$ instead of $\frac{1}{n-1}$ (sample version)

Robust Parameter estimation

→ replace $\sum_{i=1}^{n}$ by a trimmed sum using TLE (↳ Trimmed Likelyhood Estimators)

→ remove obs. whose values would highely unlikely occur if fitted model true

→ Gaussian data: MCD and Weighted Trimmed Likelyhood ⇒ same cov. estimators (↳ form of TLE)

→ TLE applied to all config's of Σ

Outlier Detection

1) Represent interval data as C and R
2) Assume $(C, \ln(R)) \sim N(\mu, \Sigma)$, possibly restrict Σ
3) Robust parameter estimation → $\hat{\mu}, \hat{\Sigma}$ (TLE)
4) robust MD's
5) interpret using EDA-graphics ($MD_i^2 > \chi^2_{d, 0.975}$)

Analysis of Variance

→ Comparison of means of ≥ 1 numerical variables in ≥ 2 populations

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$ $\quad H_1: \exists j, l: \mu_j \neq \mu_l$

→ for interval valued variables

Each Y_j modelled by pair (C_j, R_j^a) ⇒ analysis of Var: 2D-MANOVA of (C_j, R_j^a)

→ MANOVA

→ > 1 variable considered (difference to ANOVA)

→ Compares | | (determinant) of within-group matrix W to | | of global matrix T

$$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)^T$$

$$T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})(x_{ij} - \bar{x})^T$$

x ... $n \times k$

$x_{ij}, \bar{x}_j, \bar{x}$ ... $1 \times p$

Similar to covariance formula

→ we consider p variables

→ Gaussian homoscedastic case:

→ for one Variable ⇒ Wilk's Lamda

$$\Lambda = \frac{|W|}{|T|} = \frac{|E_{alt}|}{|E_{null}|}$$

$E_{null/alt}$ ... 2×2

→ General case

similar to end of chapter (3)

→ Likely ratio approach: maximise log-likelyhood for H_0 and H_1

$$\lambda = \frac{L_{null}}{L_{alt}} = \left(\frac{|E_{alt}|}{|E_{null}|}\right)^{-\frac{n}{2}}$$

→ $-2\ln\lambda$ follows asymptotically chi-square distribution

Discriminant Analysis

→ Gaussian

→ ∀ config, estimate of optimum classification rule can be obtained with $\hat{=} \Sigma$

→ direct generalisation of classical linear and quadratic discriminant classification rules

→ Linear

$$Y = \text{argmax}_g \left(\hat{\mu}_g^T \hat{\Sigma}^{-1} x - \frac{1}{2} \hat{\mu}_g^T \Sigma^{-1} \hat{\mu}_g + \log(\hat{\pi}_g) \right)$$

$g > 2, \Sigma_1 = \Sigma_n$

$\hat{=}$

→ Quadratic

$$Y = \text{argmax}_g \left(\hat{\mu}_g \hat{\Sigma}_g^{-1} x - \frac{1}{2} \hat{\mu}_g^T \hat{\Sigma}_g^{-1} \hat{\mu}_g + \log(\hat{\pi}_g) - \frac{1}{2}\left(\log(|\hat{\Sigma}_g| + x^T \hat{\Sigma}_g^{-1} x \right) \right)$$

≈ ML-term

→ Skew-Normal

→ more complex due to skewness-parameter

→ alternative formulas based on configs

(1) groups differ only on μ → location

(2) groups differ on μ and Σ → location and scatter

(3) groups differ on μ and Σ and μ_1 → general

→ Location:

$$Y = \text{argmax}_g \left(\hat{\xi}_g^T \hat{\Omega}^{-1} x - \frac{1}{2} \xi_g^T \hat{\Omega}^{-1} \xi_g + \log(\hat{\pi}_g) + \zeta_0 \left(\hat{\delta}^T \omega^{-1} (x - \hat{\xi}_g) \right) \right)$$

linear term — additional term

→ General

$$Y = \text{argmax}_g \left(\hat{\xi}_g^T \hat{\Omega}^{-1} x - \frac{1}{2} \hat{\xi}_g^T \Omega_g^{-1} \hat{\xi}_g + \log(\hat{\pi}_g) + \zeta_0 \left(\hat{\delta}^T \omega^{-1} (x - \hat{\xi}_g) \right) - \frac{1}{2} \left(\log|\hat{\Omega}_g| + x^T \hat{\Omega}^{-1} x \right) \right)$$

ML-term

→ same between $\mathcal{N}_p$ and SN_p – only difference: term ζ_0

Model based Clustering

(1) finite mixture model

$$f(x_i;\rho) = \sum_{l=1}^{k} \pi_l f_l(x_i;\Theta_l)$$

k → # clusters

data point x_i ; ↳ set of all parameters

π_l ... mixing proportions

$f_l(\ldots)$ component density function

(2) ML estimation

$$l(\rho, x) = \sum_{i=1}^{n} \ln f(x_i;\rho)$$

n → # data points

→ entire dataset constructed from mixture of several probability distributions, each corresponding to different cluster of model

(3) EM - Algorithm

→ avoid local optima, find max. likelyhood estimate

(4) Model selection

→ Select model and # components K

→ Bayesian Information Criterion

$$BIC: -2 l(\hat{\rho}, x) + d_\rho \ln(n)$$

→ min BIC ≙ optimal model

n → sample size

$-2l(\hat{\rho}, x)$: penalises worse model fit

d_ρ → # free parameters; $d_\rho \ln(n)$: penalises more parameters

(5) Model based Clustering using best Model according to BIC