EXPLORATING A MULTIVARIATE DATASET

Prediction problem

$\underline{X} \in \mathbb{R}^p$, $Y \in \mathbb{R}$: we want to use $\underline{X}$ to predict Y.

What is the best function $f(\underline{X})$: $f: \mathbb{R}^p \to \mathbb{R}$ to predict Y?

"best" : $f(\underline{X}) = \arg\min \mathbb{E}[(Y - f(\underline{X}))^2] := \arg\min$ MSE

(mean square error)

If $f(\underline{X}) = k \implies k = \arg\min \mathbb{E}[(Y-k)^2] = \mathbb{E}[Y]$ (*)

$$\mathbb{E}[(Y - f(\underline{X}))^2] = \underbrace{\mathbb{E}[(Y - \mathbb{E}[Y|\underline{X}])^2]}_{\text{constant w.r.t. } f(\underline{X})} + \mathbb{E}[(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))^2]$$

$$\implies f(\underline{X}) = \arg\min \text{MSE} = \arg\min \mathbb{E}[(\mathbb{E}[Y|\underline{X}] - f(\underline{X}))^2] \overset{(*)}{=} \underbrace{\mathbb{E}[Y|\underline{X}]}_{\text{best guess of } Y \text{ once we know } \underline{X}}$$

$\implies$ The model is:

$$\begin{cases} Y = f(\underline{X}) + \varepsilon \\ f(\underline{X}) = \mathbb{E}[Y|\underline{X}] \\ \mathbb{E}[\varepsilon] = 0 \end{cases}$$

Suppose now that we have a model $\hat{f}$. How good is the model?

$\hat{f}$: estimation of f through $\mathbb{X}$: $\mathbb{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2p} & y_2 \\ \vdots & & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_n \end{bmatrix}$

We want to see how good $\hat{f}$ is in estimating new predictions:

$\underline{x}_0 \in \mathbb{R}^p$: new observation ($\notin \mathbb{X}$) for which we want to predict y_0

We want to know about the error:

$$\mathbb{E}[(y_0 - \hat{f}(\underline{x}_0))^2 \underbrace{|\, \mathbb{X} = x}_{\text{we have one realization}}] := \mathbb{E}_{\mathbb{X}}[(y_0 - \hat{f}(\underline{x}_0))^2]$$

Since $y_0 = f(\underline{x}_0) + \varepsilon_0$

$$\implies \underbrace{\mathbb{E}_{\mathbb{X}}[(y_0 - f(\underline{x}_0))^2]}_{\substack{\text{this estimate the error of the} \\ \text{prediction given } \mathbb{X} = x, \text{ what} \\ \text{if we want to consider all} \\ \text{the possible realizations of } \mathbb{X}?}} = (f(\underline{x}_0) - \hat{f}(\underline{x}_0))^2 + \underbrace{\text{Var}(\varepsilon_0)}_{\substack{\text{this is irreducible:} \\ \text{it's the part of } Y \text{ not} \\ \text{explainable through } \underline{X}}}$$

$$\implies \mathbb{E}[\mathbb{E}_{\mathbb{X}}[(y_0 - \hat{f}(\underline{x}_0))^2]] = \text{Var}(\varepsilon_0) + \text{Var}(\hat{f}(\underline{x}_0)) + \underbrace{(f(\underline{x}_0) - \mathbb{E}[\hat{f}(\underline{x}_0)])^2}_{:= \text{BIAS}^2}$$

(how far is the model from what we want to estimate)

How to estimate f? (How to get $\hat{f}$?)

$\implies$ LOCAL AVERAGE (fail with p large: CURSE OF DIMENSIONALITY)

$\implies$ we react with:

- reduce of dimensionality (PCA) (data driven)
- parametric models (knowledge is necessary)

Geometry of the data

We can explore the dataset by rows or by columns:

- by columns:

$$\mathbb{X} = [X_1 \; X_2 \; \cdots \; X_p]$$

Every column X_j has n realizations: $\underline{y}_j = [x_{1j}, x_{2j}, \dots, x_{nj}]^T$

- mean of X_j: $\bar{x}_j := \frac{1}{n} \sum_{i=1}^{n} x_{ij}$
- variance of X_j: $S_{jj} := \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$
- covariance of X_j, X_k: $Cov(X_j, X_k) := \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) := S_{kj}$
- covariance matrix (for $X_1, \dots, X_p$):

$$S := \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & & & \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix} \in \mathbb{R}^{p \times p}$$

- correlation of X_j, X_k: $r_{kj} = \frac{S_{kj}}{\sqrt{S_{kk} S_{jj}}} = Corr(X_j, X_k) \in [-1, 1]$
- correlation matrix (for $X_1, \dots, X_p$):

$$r := \begin{bmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & & \\ r_{p1} & \cdots & r_{pp} \end{bmatrix}$$

Whatever the distribution of X_j is, we can say (Chebyshev):

$$Fr\left(\bar{x}_j - k\sqrt{S_{jj}} \le X_j \le \bar{x}_j + k\sqrt{S_{jj}}\right) \ge 1 - \frac{1}{k^2} \qquad \forall k \neq 0$$

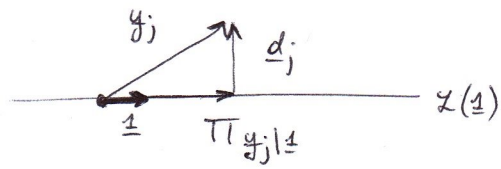
alternatively:

$$\mathbb{P}(|X - \mu| \le k\sigma) \ge 1 - \frac{1}{k^2}$$

Thanks to Chebyshev, whenever we have the mean and the standard deviation we can construct an interval.

Geometrical interpretation:

- one variable



$\underline{d}_j$ = deviation vector

$\mathcal{L}(\underline{1})$ = space with "no statistic" ($\underline{v} \in \mathcal{L}(\underline{1}) \Rightarrow \underline{v} = k \cdot \underline{1}$)

$\Pi_{\underline{y}_j|\underline{1}}$ = projection of $\underline{y}_j$ on $\mathcal{L}(\underline{1})$

Generic projection of $\underline{v}$ on $\mathcal{L}(\underline{w})$:

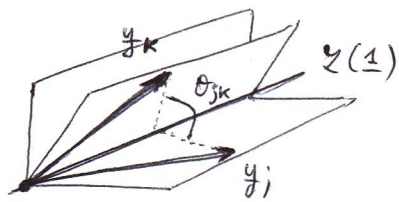
$$\Pi_{\underline{v}|\mathcal{L}(\underline{w})} = \frac{\underline{w} \cdot \underline{w}^T}{\underline{w}^T \cdot \underline{w}} \cdot \underline{v}$$

$$\Longrightarrow \Pi_{\underline{y}_j|\mathcal{L}(\underline{1})} = \frac{\underline{1} \cdot \underline{1}^T}{\underline{1}^T \underline{1}} \underline{y}_j = \bar{x}_j \underline{1}$$

$$\Longrightarrow \underline{d}_j = \underline{y}_j - \bar{x}_j \underline{1} \Longrightarrow \|\underline{d}_j\| = \sqrt{n}\sqrt{S_{jj}}$$

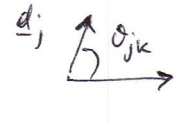
this is why S_{jj} talks about the variability (Note: $\underline{d}_j$ captures the informations that the mean doesn't)

- two variables



$$\underline{y}_j = \bar{x}_j \underline{1} + \underline{d}_j$$

$$\underline{y}_k = \bar{x}_k \underline{1} + \underline{d}_k$$



$$\cos(\vartheta_{jk}) = \frac{\underline{d}_j^T \underline{d}_k}{\|\underline{d}_j\| \cdot \|\underline{d}_k\|} = \frac{s_{jk}}{\sqrt{s_{jj}\, s_{kk}}} = r_{jk}$$

- $\vartheta_{jk} = 0 \implies r_{jk} = 1 \implies \underline{d}_j \in \mathcal{L}(\underline{d}_k)$
- $\vartheta_{jk} = \frac{\pi}{2} \implies r_{jk} = 0 \implies \underline{d}_k \perp \underline{d}_j$

- by rows:

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix}$$

where $\underline{x}_j$ = realization of $\underline{X}_j$

$\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} \underline{X} \in \mathbb{R}^p$: we manage with random vectors $\underline{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$

- mean: $\mathbb{E}[\underline{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_p] \end{bmatrix} = \underline{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$
- covariance of X_j and X_k:

$$\sigma_{jk} = \mathbb{E}[(X_j - \mu_j)(X_k - \mu_k)] = [\Sigma]_{jk}$$

- covariance matrix (for $\underline{X}$)

$$\Sigma := [\sigma_{ij}] \in \mathbb{R}^{p \times p}$$

$$\Sigma := \mathbb{E}[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T]$$

- correlation matrix (for $\underline{X}$): (ρ)

$$V = \operatorname{diag}(\sigma_1, \dots, \sigma_p) \implies V^{1/2} = \operatorname{diag}(\sqrt{\sigma_1}, \dots, \sqrt{\sigma_p})$$
$$V^{-1/2} = \operatorname{diag}\left(\frac{1}{\sqrt{\sigma_1}}, \dots, \frac{1}{\sqrt{\sigma_p}}\right)$$

$$\rho := V^{-\frac{1}{2}} \Sigma V^{\frac{1}{2}}$$

- linear combinations of the component of $\underline{X}$:

$\underline{c} \in \mathbb{R}^p$: $\mathbb{E}[\underline{c}^T \underline{X}] = \underline{c}^T \underline{\mu}$

$\operatorname{Var}(\underline{c}^T \underline{X}) = \underline{c}^T \Sigma \underline{c}$

- k linear combinations of the component of $\underline{X}$:

$C \in \mathbb{R}^{k \times p}$: $\mathbb{E}[C\underline{X}] = C\underline{\mu}$

$\operatorname{Cov}(C\underline{X}) = C \Sigma C^T$

Estimators

We consider $\mathbb{X}$ and we look at it by rows. Can we estimate $\underline{\mu}$ and Σ?
We assume every row to be a realization of a random vector:

n rows $\implies$ n random vectors $\underline{X}_j \in \mathbb{R}^p$ $\quad (\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} \underline{X} \in \mathbb{R}^p)$

- Estimator for $\underline{\mu}$:

$$\overline{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_j \xrightarrow{\text{realization}} \overline{\underline{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Assuming $\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} \underline{X}$, $\mathbb{E}[\underline{X}] = \underline{\mu}$, $\operatorname{Cov}(\underline{X}) = \Sigma$:

- $\mathbb{E}[\overline{\underline{X}}] = \underline{\mu}$ (unbiased)
- $\operatorname{Cov}(\overline{\underline{X}}) = \frac{1}{n} \Sigma$

- Estimator for Σ:

$$S = \frac{1}{n} \sum_{i=1}^{n} (\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})^T \xrightarrow{\text{realization}} S = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ & \vdots & \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}$$

$\mathbb{E}[S] = \frac{n-1}{n}\Sigma \implies \frac{n-1}{n} S$ is unbiased for Σ

$\implies$ from now on:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})^T$$

Considering that:

$$\underline{d}_j = \underline{y}_j - \frac{\underline{1}\cdot\underline{1}^T}{\underline{1}^T\cdot\underline{1}}\underline{y}_j = \left(I - \frac{\underline{1}\cdot\underline{1}^T}{\underline{1}^T\cdot\underline{1}}\right)\underline{y}_j \qquad \left(\underline{d}_j = \underline{y}_j - \Pi_{\underline{y}_j|\underline{1}}\right)$$

$$d = [\underline{d}_1, \dots, \underline{d}_p] = \left(I - \frac{\underline{1}\,\underline{1}^T}{\underline{1}^T\underline{1}}\right)X \qquad := \text{deviation matrix}$$

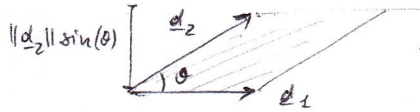
$$S = \frac{1}{n} d^T d \;\Big|\; = \frac{1}{n-1} X^T\left(I - \frac{\underline{1}\,\underline{1}^T}{\underline{1}^T\underline{1}}\right)X$$

Variability in a multivariate sense

In $\mathbb{R}$ the variability is explained by variance.
In $\mathbb{R}^p$ the variance is not enough.

- Generalized variance: **Det(S)**

$p = 2$: $S = \frac{1}{n-1} d^T d \implies \text{Det}(S) = \|\underline{d}_1\|^2 \|\underline{d}_2\|^2 \sin^2\theta \left(\frac{1}{n-1}\right)^2$



$\implies \text{Det}(S) \propto (\text{area of parall.}(\underline{d}_1, \underline{d}_2))^2$

  - $\det(S) \uparrow \implies$ area $\uparrow$: can increase because of θ (max area with $\theta = \pi/2$) or because of $\|\underline{d}_1\| \,/\, \|\underline{d}_2\|$
  - $\det(S) = 0 \implies$ either:
    - $(\|\underline{d}_1\| \vee \|\underline{d}_2\|) = 0$
    - $\theta = 0$

    In any case, it doesn't mean that $\nexists$ variability

$p = \forall$: $\text{Det}(S) \propto (\text{vol}(\text{parallelotope}(\underline{d}_1, \dots, \underline{d}_p))^2$

- Total variance: **Tr(S)**

$p = 2$: $\text{Tr}(S) = \frac{1}{n-1}(\|\underline{d}_1\|^2 + \|\underline{d}_2\|^2)$

$p = \forall$: $\text{Tr}(S) = \frac{1}{n-1}(\|\underline{d}_1\|^2 + \dots + \|\underline{d}_p\|^2)$

(capturing the sum of the marginal variabilities of the variables)

Note: $\text{Det}(S) = 0 \iff \underline{d}_1, \dots, \underline{d}_p$ are linearly dependent
$(\exists\, \underline{c} \neq \underline{0}$ s.t. $c_1\underline{d}_1 + \dots + c_p\underline{d}_p = \underline{0})$

Consequence: $\text{Det}(S) = 0 \implies \exists k$ s.t. $\underline{d}_k = -\sum_{\substack{i=1 \\ i\neq k}}^{p} \frac{c_i}{c_k}\underline{d}_i$

$$\implies \underline{y}_k = \bar{x}_k\underline{1} - \sum_{\substack{i=1 \\ i\neq k}}^{p} \frac{c_i}{c_k}(\underline{y}_i - \bar{x}_i\underline{1})$$

(there is a perfect linear relationship between the variable k and all the other variables, so the variable k is useless (is obtainable as function of the others))

Consequence: $X \in \mathbb{R}^{n\times p}$: If $p \geq n \implies \text{Det}(S) = 0$

Note that: $S \in \mathbb{R}^{p\times p}$ real and symmetric $\Rightarrow$ $S = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T$

$\Rightarrow$ $P := [\underline{e}_1, \dots, \underline{e}_p]$, $\Lambda := \text{diag}(\lambda_1, \dots, \lambda_p)$: $S = P \Lambda P^T$

$\Rightarrow$ $\begin{cases} \det(S) = \prod_{i=1}^{p} \lambda_i \\ \text{Tr}(S) = \sum_{i=1}^{p} \lambda_i \end{cases}$

S is positive semi-definite ($\lambda_i \geq 0 \ \forall i$), if $\text{Det}(S) > 0$ is pos. def. ($\lambda_i > 0 \ \forall i$)

Assuming S positive definite ($\text{Det}(S) > 0$):
(and an order $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$)

$\underline{x}, \underline{y} \in \mathbb{R}^p$: $d^2_{S^{-1}}(\underline{x}, \underline{y}) = (\underline{x}-\underline{y})^T S^{-1} (\underline{x}-\underline{y})$:= Mahalanobis distance (squared)

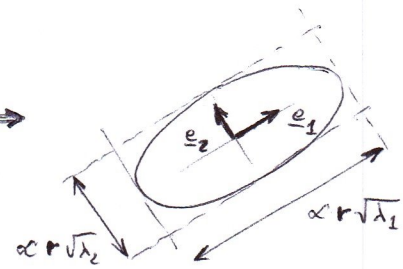
Note: $S^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i^T$

"Neighborhood" in Mahalanobis' distance:

$$\mathcal{E}_{r^2, S^{-1}}(\underline{\bar{x}}) = \{ \underline{x} \in \mathbb{R}^p : (\underline{x}-\underline{\bar{x}})^T S^{-1} (\underline{x}-\underline{\bar{x}}) \leq r^2 \}$$

Graphically ($p=2$):

$S^{-1} = \sum_{i=1}^{2} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i^T$ $\Rightarrow$

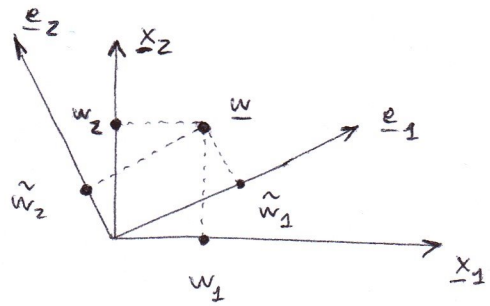


The Mahalanobis' distance is the right one for capturing the distance between data.
(The euclidean distance is not good, in fact the Mahalanobis' distance is the Euclidean distance after we standardized the data)

Note that: it would be convenient to find a system s.t. S is diagonal w.r.t. that system. Graphically ($p=2$):

in this way the axes of the ellipse are // to the axes of the system

$\Rightarrow$ We introduce a system identified by the eigenvectors of S:



old system	new system
$\underline{w} = [x_1 \ \dots \ x_p]^T$	$\tilde{\underline{w}} = [\underline{e}_1^T \underline{w} \ \dots \ \underline{e}_p^T \underline{w}] = P^T \underline{w}$
$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix}$	$\tilde{\mathbb{X}} = \mathbb{X} P$
$S = \frac{1}{n-1} \mathbb{X} \left(\mathbb{I} - \frac{\underline{1}\underline{1}^T}{\underline{1}^T\underline{1}} \right) \mathbb{X}$	$\tilde{S} = P^T S P = \Lambda$

Observation: $\|\underline{w}\| = \|\tilde{\underline{w}}\|$

Do we lose informations? No:

- $\text{Det}(S) = \prod_{i=1}^{p} \lambda_i = \text{Det}(\tilde{S})$
- $\text{Tr}(S) = \sum_{i=1}^{p} \lambda_i = \text{Tr}(\tilde{S})$

There is always a reference system for which the coordinates are uncorrelated: correlation is NOT a property of the data

PRINCIPAL COMPONENT ANALYSIS (PCA)

Let's refer to the model (not the data):

$\underline{X}$ random vector in $\mathbb{R}^p$:
- $\mathbb{E}[\underline{X}] = \underline{\mu}$
- $\mathrm{Cov}(\underline{X}) = \Sigma$

Problem: Find $\underline{a} \in \mathbb{R}^p$ s.t. $\|\underline{a}\| = 1$ and $\mathrm{Var}(\underline{a}^T \underline{X})$ is maximized
(i.e. find the direction $\underline{a}$ s.t. the variability of the projection of $\underline{X}$ on $\underline{a}$ is maximum)

$$\max_{\underline{a} \in \mathbb{R}^p : \|\underline{a}\| = 1} \mathrm{Var}(\underline{a}^T \underline{X}) = \max_{\underline{a} \in \mathbb{R}^p : \|\underline{a}\| = 1} \underline{a}^T \Sigma \underline{a} = \max_{\underline{a} \in \mathbb{R}^p} \frac{\underline{a}^T \Sigma \underline{a}}{\underline{a}^T \underline{a}}$$

Geometrical Lemma:

$B \in \mathbb{R}^{p \times p}$ positive semidefinite, $B = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T$: $(\lambda_1 \geq \lambda_2 \geq \dots)$

$\Longrightarrow$

1. $\max_{\underline{x} \in \mathbb{R}^p} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_1$, $\arg\max(\dots) = \underline{e}_1$
2. $\max_{\underline{x} \in \mathbb{R}^p, \ \underline{x} \perp \underline{e}_1} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_2$, $\arg\max(\dots) = \underline{e}_2$
3. $\vdots$
4. p. $\max_{\underline{x} \in \mathbb{R}^p, \ \underline{x} \perp \underline{e}_1, \dots, \underline{x} \perp \underline{e}_{p-1}} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}} = \lambda_p = \min_{\underline{x} \in \mathbb{R}^p} \frac{\underline{x}^T B \underline{x}}{\underline{x}^T \underline{x}}$

Back to PCA:

$$\max_{\underline{a} \in \mathbb{R}^p :} \mathrm{Var}(\underline{a}^T \underline{X}) = \max_{\underline{a} \in \mathbb{R}^p} \frac{\underline{a}^T \Sigma \underline{a}}{\underline{a}^T \underline{a}} = \lambda_1, \quad \arg\max(\dots) = \underline{e}_1$$

$\Longrightarrow$ First Principal Component (PC1): $Y_1 = \underline{e}_1^T \underline{X}$ $\left(/ \ Y_1 = \underline{e}_1^T (\underline{X} - \underline{\mu})\right)$

Problem: Find $\underline{a} \in \mathbb{R}^p$ s.t. $\|\underline{a}\| = 1$, $\mathrm{Var}(\underline{a}^T \underline{X})$ is max and $\underline{a} \perp \underline{e}_1$
(i.e. we want to find an other direction of max variability but we want the projection on this direction to be uncorrelated with the previous projection)

$$\max_{\substack{\underline{a} \in \mathbb{R}^p : \|\underline{a}\| = 1 \\ \mathrm{Cov}(\underline{a}^T \underline{X}, \underline{e}_1^T \underline{X}) = 0}} \mathrm{Var}(\underline{a}^T \underline{X}) = \max_{\substack{\underline{a} \in \mathbb{R}^p \\ \underline{a} \perp \underline{e}_1}} \frac{\underline{a}^T \Sigma \underline{a}}{\underline{a}^T \underline{a}}$$

$\Longrightarrow$

$$\max_{\underline{a} \in \mathbb{R}^p : \ \underline{a} \perp \underline{e}_1} \frac{\underline{a}^T \Sigma \underline{a}}{\underline{a}^T \underline{a}} = \lambda_2, \quad \arg\max(\dots) = \underline{e}_2$$

$\Longrightarrow$ Second Principal Component (PC2): $Y_2 = \underline{e}_2^T \underline{X}$ $\left(/ \ Y_2 = \underline{e}_2^T (\underline{X} - \underline{\mu})\right)$

Generalized problem:

$$\max_{\substack{\underline{a} \in \mathbb{R}^p : \|\underline{a}\| = 1 \\ \mathrm{Cov}(\underline{a}^T \underline{X}, \underline{e}_j^T \underline{X}) = 0 \ \ j = 1, \dots, k-1}} \mathrm{Var}(\underline{a}^T \underline{X}) = \lambda_k, \quad \arg\max(\dots) = \underline{e}_k$$

$\Longrightarrow$ k^{th} Principal Component (PCk): $Y_k = \underline{e}_k^T \underline{X}$ $\left(/ \ Y_k = \underline{e}_k^T (\underline{X} - \underline{\mu})\right)$

$\underline{Y} := [Y_1 \dots Y_p]^T = P^T \underline{X}$:= vector of Principal Components

- $\mathbb{E}[\underline{Y}] = \mathbb{E}[P^T \underline{X}] = P^T \underline{\mu}$
- $Cov(\underline{Y}) = Cov(P^T \underline{X}) = P^T \Sigma P = \Lambda$
  $Cov(Y_i, Y_j) = 0 \quad \forall i \neq j \longrightarrow$ no correlation between the coordinates of $\underline{Y}$
  $Cov(Y_i, Y_i) = Var(Y_i) = \lambda_i \quad \forall i$
- Ordering: we ordered the components $\Rightarrow$ the first component is the one with longer variability, the second is the one with the second longer var. and so on $\Rightarrow$ we can capture the most of the variability just with the first components (the last express small variability)
- We're not losing variability:
  - generalized variance with $\underline{Y}$: $Det(\Lambda) = \prod_{i=1}^{p} \lambda_i = Det(\Sigma)$
  - total variance with $\underline{Y}$: $Tr(\Lambda) = \sum_{i=1}^{p} \lambda_i = Tr(\Sigma)$
- $Y_i = e_{1i} X_1 + \dots + e_{pi} X_p$
  $\{e_{1i}, \dots, e_{pi}\}$ are the loading (weights):
  $$Corr(Y_i, X_k) = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

PCA on standardized variables

(we're still working with the model, not data)

$V := diag(\sigma_{11}, \dots, \sigma_{pp})$

$$\underline{X} \longrightarrow \underline{Z} = V^{-1/2}(\underline{X} - \underline{\mu}) = \left[\frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \dots, \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}\right]^T$$

$\underline{Z}$:
- $\mathbb{E}[\underline{Z}] = \underline{0}$
- $Cov(\underline{Z}) = V^{-1/2} \Sigma V^{-1/2} := \rho$:= correlation matrix (covariance matrix of the standardized variables)

$\rho = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T$ (eigenvalues/vectors of ρ)

$$\Longrightarrow Y_i = \underline{e}_i^T \underline{Z} = \underline{e}_i^T V^{-1/2}(\underline{X} - \underline{\mu})$$

(Note: $Tr(\rho) = \sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$)

- $PCA(\Sigma) \neq PCA(\rho)$
- If we think that there are variables with very different variabilities (maybe because of the units of measure or maybe because of the phenomena) then it would be better $PCA(\rho)$:
  if X_1 is in km and X_2 in mm $\Rightarrow$ $Var(X_1)$ will mask $Var(X_2) \Rightarrow PCA(\rho)$

PCA on data

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} : \underline{x}_i \text{ realization of } \underline{X}_i, \quad \underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} \underline{X} \in \mathbb{R}^p$$

Usually $\underline{\mu}$ and Σ are unknown, we use data:

$\left.\begin{array}{l} S \text{ estimates } \Sigma \\ \overline{\underline{X}} \text{ estimates } \underline{\mu} \end{array}\right\} \Longrightarrow$ PCA on S

$S = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^T \implies$ PCA: Y_i projection on $\underline{e}_i$

$\underline{x}_i$ (i-th stat. unit) $\xrightarrow{\text{PCA}}$ $y_i = \begin{bmatrix} \underline{e}_1^T \underline{x}_i \\ \vdots \\ \underline{e}_p^T \underline{x}_i \end{bmatrix}$ ← "score" of $\underline{x}_i$ on the 1st PC ← "score" of $\underline{x}_i$ on the p^{th} PC

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} \xrightarrow{\text{PCA}} \begin{bmatrix} y_1^T \\ \vdots \\ y_n^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & & y_{2p} \\ \vdots & & & \vdots \\ y_{n1} & y_{n2} & & y_{np} \end{bmatrix} \quad \begin{matrix} \text{scores on the} \\ p\text{-th principal} \\ \text{component} \end{matrix}$$

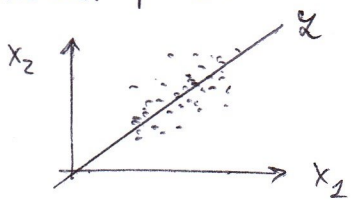
$$Y_1 \quad Y_2 \quad \cdots \quad Y_p$$

- PCA for categorical data is called "Correspondence analysis" and it's performed on the table of joint frequencies among variables
- we have also to observe the smaller λ_i: if $\lambda_p \sim 0$ it means that $\exists$ linear relationship between $\underline{X}_1, \dots, \underline{X}_k$

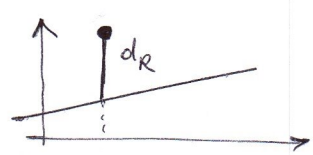
Geometrical meaning of PCA

Problem: given $\underline{x}_1, \dots, \underline{x}_n \in \mathbb{R}^p$ find a linear space $\mathcal{L}$ s.t. $\dim(\mathcal{L}) = k \leq p$ which best approximate the data (i.e. such that $\mathcal{L}$ is "closest" to data).

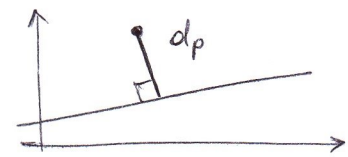
Consider $p = 2$:



Is $\mathcal{L}$ the regression line? **NO!**



(Regression)



(Projection)

Here we're trying to minimize d_p, not d_R ($\neq d_p$)!
(Note that the regression depends on the system, the projection no)

Problem:

Find $\eta_1, \dots, \eta_k$ (we identify $\mathcal{L}$ with an orthonormal basis: $\mathcal{L} = \text{span}(\eta_1, \dots, \eta_k)$)

such that:

$$\underbrace{\sum_{i=1}^{n} \left\| (\underline{x}_i - \bar{\underline{x}}) - \sum_{j=1}^{k} \eta_j \eta_j^T (\underline{x}_i - \bar{\underline{x}}) \right\|^2}_{(*)} \quad \text{is minimum.}$$

$(*)$ = (distance between every single datum (after centering) and the projection on $\mathcal{L}$ of the datum (after centering))^2

$$\min (*) \iff \max \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\eta_j^T (\underline{x}_i - \bar{\underline{x}}) \right)^2$$

$$\iff \max (n-1) \sum_{j=1}^{k} \eta_j^T S \eta_j \quad : \quad k = 1 \implies \max(\cdot) = \lambda_1, \ \operatorname{argmax}(\cdot) = \underline{e}_1$$

by induction on k:

$\eta_1 = \underline{e}_1, \dots, \eta_k = \underline{e}_k \implies$ The directions of the first k principal components identify the closest linear space (of dim k) to the data

Error of approximation: $(n-1) \sum_{j=k+1}^{p} \lambda_i$

MULTIVARIATE GAUSSIAN DISTRIBUTION

$\underline{X} \sim N_p(\underline{\mu}, \Sigma)$

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

- $\underline{X} \sim N_p(\underline{\mu}, \Sigma) \Longrightarrow \mathbb{E}[\underline{X}] = \underline{\mu}, \quad Cov(\underline{X}) = \Sigma$
- $\underline{X} \sim N_p(\underline{\mu}, \Sigma) \iff \underline{a}^T \underline{X} \sim N_1(\underline{a}^T \underline{\mu}, \underline{a}^T \Sigma \underline{a}) \qquad \forall \underline{a} \in \mathbb{R}^p$
  (∀ linear combination of the components is gaussian)
- $\underline{X} = [X_1, \dots, X_p]^T \sim N_p(\underline{\mu}, \Sigma), \; \Sigma = [\sigma_{ij}] \Longrightarrow X_j \sim N_1(\mu_j, \sigma_{jj})$
- $\underline{X} \sim N_p(\underline{\mu}, \Sigma), \; A \in \mathbb{R}^{q \times p} \Longrightarrow A\underline{X} \sim N_q(A\underline{\mu}, A\Sigma A^T)$
- $\underline{X} \sim N_p(\underline{\mu}, \Sigma), \; \underline{d} \in \mathbb{R}^p \Longrightarrow \underline{X} + \underline{d} \sim N_p(\underline{\mu} + \underline{d}, \Sigma)$
- $\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix} \sim N_p\left(\begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \Longrightarrow \underline{X}_1 \sim N_q(\underline{\mu}_1, \Sigma_{11})$

  So for example:
  $\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N \Longrightarrow \begin{bmatrix} X_1 \\ X_3 \end{bmatrix} \sim N$

  $\underline{X}_1 \perp\!\!\!\perp \underline{X}_2 \iff \Sigma_{12} = \Sigma_{21} = 0$

- $\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix} \sim N_p\left(\begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right), \; \det(\Sigma) \neq 0:$

  $$\Longrightarrow \underline{X}_1 \mid \underline{X}_2 = \underline{x}_2 \sim N_q\left(\underline{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2 - \underline{\mu}_2), \; \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right)$$

  $$Cov(\underline{X}_1 \mid \underline{X}_2 = \underline{x}_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad \perp\!\!\!\perp \underline{x}_2$$

  := Partial covariances
  (even without knowing $\underline{x}_2$ we already know how informations will be modified)
  (This generates the Regression Effect)

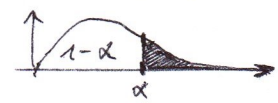
- $\underline{X} \sim N_p(\underline{\mu}, \Sigma), \; \det(\Sigma) > 0 \Longrightarrow (\underline{X}-\underline{\mu})^T \Sigma^{-1} (\underline{X}-\underline{\mu}) \sim \chi^2(p)$
- $\underline{X} \sim N_p(\underline{\mu}, \Sigma), \; \det(\Sigma) = 0 \Longrightarrow (\underline{X}-\underline{\mu})^T \Sigma^{-} (\underline{X}-\underline{\mu}) \sim \chi^2(k)$

  $k = rank(\Sigma)$

  $$\Sigma^{-} = \sum_{i=1}^{k} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i^T$$

  $(\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > 0 = \lambda_{k+1} = \dots = \lambda_p)$

- $\alpha \in (0,1), \; \det(\Sigma) > 0 \Longrightarrow \mathbb{P}\left((\underline{X}-\underline{\mu})^T \Sigma^{-1} (\underline{X}-\underline{\mu}) \leq \chi^2_\alpha(p) \right) = 1 - \alpha$



Estimators for μ and Σ

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} : \; \underline{x}_i \text{ realization of } \underline{X}_i \Longrightarrow \underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma)$$

- $\overline{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i$
- $S = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})^T$ (unbiased)
- $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})^T$ (biased, but MLE)

Distributions: assuming $\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma)$

- $\bar{\underline{X}} \sim N_p(\underline{\mu}, \frac{1}{n}\Sigma)$
- $S \sim \text{Wish}(\frac{1}{n-1}\Sigma, n-1)$ $\qquad (n-1)S \sim \text{Wish}(\Sigma, n-1)$
- $\hat{\Sigma} \sim \text{Wish}(\frac{1}{n}\Sigma, n-1)$ $\qquad n\hat{\Sigma} \sim \text{Wish}(\Sigma, n-1)$

About Wishart distribution (matrix distribution!):

(Def.) $\underline{z}_1, \dots, \underline{z}_m$ iid $\sim N_p(\underline{0}, \Sigma)$, $\det(\Sigma) > 0$

$\Longrightarrow \sum_{i=1}^{n} \underline{z}_i \underline{z}_i^T \sim \text{Wishart}(\Sigma, m)$ (the param p is in Σ)

Properties:

- $A_1 \sim \text{Wish}(\Sigma, m_1)$, $A_2 \sim \text{Wish}(\Sigma, m_2)$, $A_1 \perp\!\!\!\perp A_2$
  $\Longrightarrow A_1 + A_2 \sim \text{Wish}(\Sigma, m_1 + m_2)$
- $A \sim \text{Wish}(\Sigma, m)$, $C \in \mathbb{R}^{k \times p}$ const
  $\Longrightarrow CAC^T \sim \text{Wish}(C\Sigma C^T, m)$
- $A \sim \text{Wish}(\Sigma, m)$, $\sigma^2 > 0$
  $\Longrightarrow \sigma^2 A \sim \text{Wish}(\sigma^2\Sigma, m)$
- $A \sim \text{Wish}(\Sigma, m)$, Σ 1×1 $\Longrightarrow A \sim \Sigma\, \chi^2(m)$

Note that: $\bar{\underline{X}} \perp\!\!\!\perp S$ and are sufficient statistics (i.e. if the data is generated by a gaussian distribution (no matter n) then all we need to know is $\bar{\underline{X}}, S$)

LLN

$\underline{X}_1, \dots, \underline{X}_n$ random vectors iid such that $\mathbb{E}[\underline{X}_i] = \underline{\mu}$, $\text{Cov}(\underline{X}) = \Sigma$ exists:

- $\bar{\underline{X}} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i \xrightarrow{\mathbb{P}} \underline{\mu}$ as $n \to \infty$
- $S = \frac{1}{n-1}\sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})^T \xrightarrow{\mathbb{P}} \Sigma$ as $n \to \infty$

CLT

$\underline{X}_1, \dots, \underline{X}_n$ random vectors iid such that $\mathbb{E}[\underline{X}_i] = \underline{\mu}$, $\text{Cov}(\underline{X}) = \Sigma$ exists:

$\Longrightarrow \sqrt{n}(\bar{\underline{X}} - \underline{\mu}) \sim \underbrace{AN_p(\underline{0}, \Sigma)}$

asintotically normal

meaning: for large n one can approximate the distribution of $\sqrt{n}(\bar{\underline{X}} - \underline{\mu})$ with a $N_p(\underline{0}, \Sigma)$

In practice, for large n:

$$\bar{\underline{X}} \sim N_p(\underline{\mu}, \frac{1}{n}\Sigma)$$

It doesn't mean that if the sample is large then it is gaussian. It means that if a sample is large enough then the sample mean is gaussian

INFERENCE FOR THE MEAN $\underline{\mu}$

$\underline{X}_1, \dots, \underline{X}_n \in \mathbb{R}^p$ random vectors s.t. $(\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} \underline{X})$

- $\mathbb{E}[\underline{X}] = \underline{\mu}$
- $\mathrm{Cov}(\underline{X}) = \Sigma$
- $\mathrm{Det}(\Sigma) > 0$

$\overline{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i$ pointwise estimator for $\underline{\mu}$

Two cases:

- **n** large $(n >> p)$:

  CLT: $\sqrt{n}\,(\overline{\underline{X}} - \underline{\mu}) \sim N_p(\underline{0}, \Sigma)$

  Pivotal statistic distr.:

  $$n(\overline{\underline{X}} - \underline{\mu})^T \Sigma^{-1} (\overline{\underline{X}} - \underline{\mu}) \sim \chi^2(p)$$

  If Σ is not known:

  LLN: $S \xrightarrow{n \to \infty} \Sigma$

  Pivotal statistic distr.:

  $$n(\overline{\underline{X}} - \underline{\mu})^T S^{-1} (\overline{\underline{X}} - \underline{\mu}) \sim \chi^2(p)$$

  - Confidence regions:

    $$\mathbb{P}\left(d^2_{S^{-1}}(\overline{\underline{X}}, \underline{\mu}) \le \chi^2_\alpha(p)\right) = \mathbb{P}\left(\underline{\mu} \in \mathcal{E}^{\alpha}_{S^{-1}}(\overline{\underline{X}})\right) = 1 - \alpha$$

    (i.e. the random ellipse centered in $\overline{\underline{X}}$ and shaped with S will be covering the true (fixed) value of $\underline{\mu}$, $(1-\alpha)\cdot 100\%$ of the times that we generate it)

    $$CR_{1-\alpha}(\underline{\mu}) = \left\{ \underline{\eta} \in \mathbb{R}^p : n(\underline{\eta} - \overline{\underline{X}})^T S^{-1} (\underline{\eta} - \overline{\underline{X}}) \le \chi^2_\alpha(p) \right\}$$

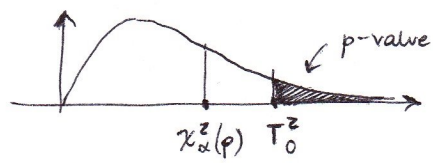
  - Testing:

    $$\begin{cases} H_0: \underline{\mu} = \underline{\mu}_0 \\ H_1: \underline{\mu} \neq \underline{\mu}_0 \end{cases}$$

    Test statistic: $T_0^2 = n(\overline{\underline{X}} - \underline{\mu}_0)^T S^{-1} (\overline{\underline{X}} - \underline{\mu}_0)$

    If H_0 is true $\Rightarrow T_0^2 \sim \chi^2(p)$

    Rejection Region$_\alpha$ $= \{ T_0^2 > \chi^2_\alpha(p) \}$

    $T_0^2 > \chi^2_\alpha(p) \iff$ p-value $\le \alpha$

- **n** small

  Here we have to add: $\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma)$, $\det(\Sigma) > 0$

  Pivotal statistic distr.:

  $$n(\overline{\underline{X}} - \underline{\mu})^T S^{-1} (\overline{\underline{X}} - \underline{\mu}) \sim \frac{(n-1)p}{n-p} F(p, n-p)$$

  Hotelling's theorem: $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$ $(\det(\Sigma) > 0)$ $\perp\!\!\!\perp$ $W \sim \mathrm{Wish}\left(\frac{1}{m}\Sigma, m\right)$

  $$\Rightarrow \frac{m-p+1}{m} (\underline{X} - \underline{\mu})^T W^{-1} (\underline{X} - \underline{\mu}) \sim F(p, m-p+1)$$

- Confidence regions:

$$\mathbb{P}\left(n(\bar{\underline{x}}-\underline{\mu})^T S^{-1}(\bar{\underline{x}}-\underline{\mu}) \le \frac{(n-1)p}{n-p} F_\alpha(p, n-p) \right) = 1-\alpha$$

$$CR_{1-\alpha}(\underline{\mu}) = \left\{ \underline{\eta} \in \mathbb{R}^p : n(\bar{\underline{x}}-\underline{\mu})^T S^{-1}(\bar{\underline{x}}-\underline{\mu}) \le \frac{(n-1)p}{n-p} F_\alpha(p, n-p) \right\}$$

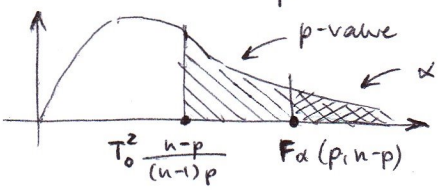
- Testing:

$$\begin{cases} H_0 : \underline{\mu} = \underline{\mu}_0 \\ H_1 : \underline{\mu} \neq \underline{\mu}_0 \end{cases}$$

$$T_0^2 = n(\bar{\underline{x}}-\underline{\mu}_0)^T S^{-1}(\bar{\underline{x}}-\underline{\mu}_0)$$

If H_0 is true $\Rightarrow$ $T_0^2 \sim \frac{(n-1)p}{n-p} F_\alpha(p, n-p)$

Reject if $T_0^2 > \frac{(n-1)p}{n-p} F_\alpha(p, n-p)$



CI for linear combinations of the mean

$\underline{x}_1, \dots, \underline{x}_n \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma)$, small n case

$\bar{\underline{x}} = \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i$ estimator for $\underline{\mu}$

- One-at-the-time $CI(\underline{\mu})$ (1! $\underline{a}$)

Let $\underline{a} \in \mathbb{R}^p$, $CI_{1-\alpha}(\underline{a}^T\underline{\mu})$?

$$\frac{\underline{a}^T\bar{\underline{x}} - \underline{a}^T\underline{\mu}}{\sqrt{\underline{a}^T S \underline{a}}} \sqrt{n} \sim t(n-1)$$

$$\Rightarrow \mathbb{P}\left(\underline{a}^T\underline{\mu} \in \left[\underline{a}^T\bar{\underline{x}} \pm t_{\frac{\alpha}{2}}(n-1) \sqrt{\frac{\underline{a}^T S \underline{a}}{n}} \right] \right) = 1-\alpha \qquad \forall \underline{a} \in \mathbb{R}^p$$

$$\Rightarrow CI_{1-\alpha}(\underline{a}^T\underline{\mu}) = \left[\underline{a}^T\bar{\underline{x}} \pm t_{\frac{\alpha}{2}}(n-1) \sqrt{\frac{\underline{a}^T S \underline{a}}{n}} \right]$$

(Testing:

$$\begin{cases} H_0 : \underline{a}^T\underline{\mu} = \delta_0 \\ H_1 : \underline{a}^T\underline{\mu} \neq \delta_0 \end{cases}$$

Reject if:

$$\frac{|\underline{a}^T\bar{\underline{x}} - \delta_0|}{\sqrt{\underline{a}^T S \underline{a}}} \sqrt{n} > t_{\frac{\alpha}{2}}(n-1)$$

$$\begin{cases} H_0 : \underline{a}^T\underline{\mu} = \delta_0 \\ H_1 : \underline{a}^T\underline{\mu} > \delta_0 \end{cases}$$

Reject if:

$$\frac{\underline{a}^T\bar{\underline{x}} - \delta_0}{\sqrt{\underline{a}^T S \underline{a}}} \sqrt{n} > t_\alpha(n-1) \;)$$

Examples:

$\underline{a} = [0 \dots 0\,1\,0 \dots 0]^T \Rightarrow \underline{a}^T\underline{\mu} = \mu_i$

$\underline{a} = [0 \dots 0\,1\,0 \dots 0\,{-1}\,0 \dots 0]^T \Rightarrow \underline{a}^T\underline{\mu} = \mu_i - \mu_j$

- Simultaneous $CI(\underline{a}^T\underline{\mu})$ ($\forall \underline{a}$)

$$\mathbb{P}\left(\frac{|\underline{a}^T(\bar{\underline{x}}-\underline{\mu})|}{\sqrt{\underline{a}^T S \underline{a}}} \sqrt{n} \le \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)}, \ \forall \underline{a} \in \mathbb{R}^p \right) = 1-\alpha$$

$$\text{Sim } CI_{1-\alpha}(\underline{a}^T\underline{\mu}) = \left[\underline{a}^T\bar{\underline{x}} \pm \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{\underline{a}^T S \underline{a}}{n}} \right]$$

- Bonferroni's correction (k $\underline{a}$, small k)

  Given $\underline{a}_1, \dots, \underline{a}_k \in \mathbb{R}^p$ find $CI(\underline{a}_1^T \underline{\mu}), \dots, CI(\underline{a}_k^T \underline{\mu})$ with simultaneous confidence of $1-\alpha$, $\alpha \in (0,1)$:

$$\text{sim } CI_{1-\alpha}(\underline{a}_i^T \underline{\mu}) = \left[\underline{a}_i^T \bar{\underline{x}} \pm t_{\frac{\alpha}{2k}}(n-1) \sqrt{\frac{\underline{a}^T S \underline{a}}{n}} \right]$$

Testing k assumptions simultaneously

$\underline{X}_1, \dots, \underline{X}_n \overset{iid}{\sim} N_p(\underline{\mu}, \Sigma)$ $\quad \det(\Sigma) > 0$

Given $\underline{a}_1, \dots, \underline{a}_k \in \mathbb{R}^p$:

$$\begin{cases} H_0 : \begin{cases} \underline{a}_1^T \underline{\mu} = \delta_1 \\ \vdots \\ \underline{a}_k^T \underline{\mu} = \delta_k \end{cases} \\ H_1 : \exists i : \underline{a}_i^T \underline{\mu} \neq \delta_i \end{cases}$$

(Bonferroni method:
Reject at level α if for at least one i: $\dfrac{|\underline{a}_i^T \bar{\underline{x}} - \delta_i|}{\sqrt{\underline{a}_i^T S \underline{a}_i}} \sqrt{n} > t_{\frac{\alpha}{2k}}(n-1)$)

Truth:	not-rejecting H_0	rejecting H_0
H_0	U	V (false positive)
H_1	T (false negative)	S
	$k-R$ # not rejected	R # rejected

k_0 = # true hypothesis

$k - k_0$ = # false hypothesis

- Family-wise error rate : $FWER := \mathbb{P}(V \geq 1)$ = probability that we'll reject at least one of the true null hypothesis
- False discovery rate : $FDR := \mathbb{E}\left[\frac{V}{R}\right]$

Property :

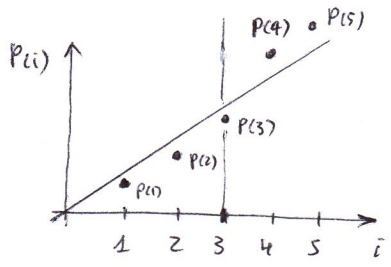
$FDR \leq FWER$

Methods :

- Bonferroni guarantees : $FWER \leq \alpha$ but with large k it becomes a problem
- Benjamin and Hockberg strategy for controlling FDR:
  1. For each of the k tests compute the p-value p_i
  2. Order the p-values : $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$
  3. $m := \max\{ i \in \{1, \dots, k\} : p_{(i)} \leq \frac{i}{k} \alpha \}$
  4. Reject $H_{0(1)}, \dots, H_{0(m)}$

Ex :

$m = 3$:



Remember :

$H_{0_1} \neq H_{0(1)}$

In this way, if the p-values are $\perp\!\!\!\perp$ $\implies$ $FDR \leq \alpha$

Comparing means of Gaussian distributions

- **Paired data**

Each unit is observed 2 times:

$$\underline{X}_{1i} = \begin{bmatrix} X_{1i1} \\ X_{1i2} \\ \vdots \\ X_{1ip} \end{bmatrix} \qquad \underline{X}_{2i} = \begin{bmatrix} X_{2i1} \\ X_{2i2} \\ \vdots \\ X_{2ip} \end{bmatrix} \qquad i = 1, \dots, n$$

$\underline{X}_{1i}$ iid with mean $\underline{\mu}_1$

$\underline{X}_{2i}$ iid with mean $\underline{\mu}_2$

Goal: inference on $\underline{\mu}_1 - \underline{\mu}_2$

$\Longrightarrow$ $\underline{D}_i := \underline{X}_{1i} - \underline{X}_{2i}$ $i = 1, \dots, n$, $\mathbb{E}[\underline{D}_i] = \underline{\mu}_1 - \underline{\mu}_2$

Goal: inference on $\mathbb{E}[\underline{D}_i]$

Assumptions: $\underline{D}_1, \dots, \underline{D}_n \overset{iid}{\sim} N_p(\underline{\delta}, \Sigma_D)$ (we're not assuming on distr. of $\underline{X}_{1i}$ / $\underline{X}_{2i}$)

Pivotal statistic:

$$n(\bar{\underline{D}} - \underline{\delta})^T S_D^{-1} (\bar{\underline{D}} - \underline{\delta}) \sim \frac{(n-1)p}{n-p} F(p, n-p)$$

- $CR_{1-\alpha}(\underline{\mu}_1 - \underline{\mu}_2) = \left\{ \underline{\delta} \in \mathbb{R}^p : n(\bar{\underline{D}} - \underline{\delta})^T S_D^{-1} (\bar{\underline{D}} - \underline{\delta}) \le \frac{(n-1)p}{n-p} F_\alpha(p, n-p) \right\}$
- $H_0 : \underline{\mu}_1 - \underline{\mu}_2 = \underline{\delta}_0$ vs. $H_1 : \underline{\mu}_1 - \underline{\mu}_2 \neq \underline{\delta}_0$
  Reject at level α if: $n(\bar{\underline{D}} - \underline{\delta}_0)^T S_D^{-1} (\bar{\underline{D}} - \underline{\delta}_0) > \frac{(n-1)p}{n-p} F_\alpha(p, n-p)$
- Sim $CI_{1-\alpha}(\mu_{1i} - \mu_{2i}) = \left[\bar{D}_i \pm \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{1}{n}} \sqrt{S_{D\,ii}} \right]$
- Bonf sim $CI_{1-\alpha} = \left[\bar{D}_i \pm t_{\frac{\alpha}{2p}}(n-1) \sqrt{\frac{1}{n}} \sqrt{S_{D\,ii}} \right]$

- **Repeated univariate measures**

Each unit is observed q times

$\underline{X}_i = [X_{i1}, \dots, X_{iq}]^T$ $i = 1, \dots, n$

Ex. n patients, X_{ij} = blood pressure of patient i at time j

$$\mathbb{E}[\underline{X}_i] = \underline{\mu} = [\mu_1, \dots, \mu_q]^T$$

Goal: $H_0 : \mu_1 = \mu_2 = \dots = \mu_q$ vs. $H_1 : \exists\, i, j$ s.t. $\mu_i \neq \mu_j$

We introduce the contrast matrix C:

examples: $C = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 & \cdots & -1 \\ 0 & 1 & \cdots & -1 \end{bmatrix}$ (*)

$\Longrightarrow$ Test: $H_0 : C\underline{\mu} = \underline{0}$ vs. $H_1 : C\underline{\mu} \neq \underline{0}$

Pivotal statistic:

$$n(C\bar{\underline{X}} - C\underline{\mu})^T (CSC)^{-1} (C\bar{\underline{X}} - C\underline{\mu}) \sim \frac{(n-1)(q-1)}{n-q+1} F(q-1, n-q+1)$$

Reject at level α if:

$$n(C\bar{\underline{x}} - C\underline{\mu})^T (CSC)^{-1} (C\bar{\underline{x}} - C\underline{\mu}) > F_\alpha(q-1, n-q+1)$$

(*) the first is testing $\mu_1 - \mu_2$, $\mu_2 - \mu_3$, ...
the second is testing $\mu_1 - \mu_q$, $\mu_2 - \mu_q$, ...

M(ANOVA) ONE-WAY

Cases:

	p # features	g # groups	Description
1	≥ 1	2	We have n patients and one treatment. We apply the treatment to n_2 patients and not to n_1 patients. For each patient we have p features
2	1	≥ 2	We have n patients and one treatment. We apply the treatment at different levels: for n_1 patients we don't apply; for n_2 pat.s we apply level 1; ⋮; for n_g pat.s we apply level g-1. For each patients we have 1 feature
3	≥ 1	≥ 2	Like in the second case but for each patient we have p features.

Case 1: $p \geq 1,\ g = 2$

$$\left.\begin{array}{l} \underline{X}_{11}, \dots, \underline{X}_{1n_1} \overset{iid}{\sim} N_p(\underline{\mu}_1, \Sigma) \\ \underline{X}_{21}, \dots, \underline{X}_{2n_2} \overset{iid}{\sim} N_p(\underline{\mu}_2, \Sigma) \end{array}\right\} \perp\!\!\!\perp$$

Goal: inference on $\underline{\mu}_1 - \underline{\mu}_2$

$$\left.\begin{array}{l} \underline{\mu}_1 \leftarrow \bar{\underline{X}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \underline{X}_{1i} \sim N_p(\underline{\mu}_1, \frac{1}{n_1}\Sigma) \\ \underline{\mu}_2 \leftarrow \bar{\underline{X}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \underline{X}_{2i} \sim N_p(\underline{\mu}_2, \frac{1}{n_2}\Sigma) \end{array}\right\} \perp\!\!\!\perp$$

$$S_{pooled} := \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

with

$$\begin{cases} S_1 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\underline{X}_{1i} - \bar{\underline{X}}_1)(\underline{X}_{1i} - \bar{\underline{X}}_1)^T & (\Sigma \text{ in group 1}) \\ S_2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(\underline{X}_{2i} - \bar{\underline{X}}_2)(\underline{X}_{2i} - \bar{\underline{X}}_2)^T & (\Sigma \text{ in group 2}) \end{cases}$$

Pivotal statistic:

$$\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-1}\left[(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)\right]^T S_{pooled}^{-1}\left[(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)\right] \sim$$

$$\sim \frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F(p,\ n_1+n_2-1-p)$$

Testing: $\begin{cases} H_0: \mu_1-\mu_2=\delta_0 \\ H_1: \mu_1-\mu_2\neq\delta_0 \end{cases}$

Reject at level α if:

$$\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-1}\left((\bar{X}_1-\bar{X}_2)-\delta_0\right)^T S_{pool}^{-1}\left((\bar{X}_1-\bar{X}_2)-\delta_0\right) > \underbrace{\frac{(n_1+n_2-2)p}{n_1+n_2-1-p} F_\alpha(p, n_1+n_2-1-p)}_{(*)}$$

Confidence region:

$$CR_{1-\alpha}(\mu_1-\mu_2)=\left\{\delta\in\mathbb{R}^p:\ \left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-1}\left((\bar{X}_1-\bar{X}_2)-\delta\right)^T S_p^{-1}\left((\bar{X}_1-\bar{X}_2)-\delta\right)\le (*)\right\}$$

Note: for n_1 and n_2 very large:

$$\left((\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)\right)^T\left(\frac{S_1}{n_1}+\frac{S_2}{n_2}\right)^{-1}\left((\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)\right) \sim \chi^2(p)$$

Case 2: $p=1$, $g\geq 2$

$$\left.\begin{array}{l} X_{11},\dots,X_{1n_1} \overset{iid}{\sim} N_1(\mu_1,\sigma^2) \\ \vdots \\ X_{g1},\dots,X_{gn_g} \overset{iid}{\sim} N_1(\mu_g,\sigma^2) \end{array}\right\} \quad \text{II} \qquad n=n_1+\dots+n_g$$

Goal: $\begin{cases} H_0: \mu_1=\mu_2=\dots=\mu_g \\ H_1: \exists\, i\neq j:\ \mu_i\neq\mu_j \end{cases}$

Parametrization: $\mu_i = \mu + \tau_i \qquad i=1,\dots,g$

μ: overall mean

$$\Longrightarrow X_{ij}=\mu+\tau_i+\varepsilon_{ij} \qquad \varepsilon_{ij}\sim N_1(0,\sigma^2)$$

$i=1,\dots,g \qquad j=1,\dots,n_i$

Constraint: $\sum_{i=1}^{g} n_i\tau_i = 0$

$$\mu \leftarrow \bar{X}=\frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i} X_{ij}$$

$$\tau_i \leftarrow \bar{X}_i-\bar{X} = \left[\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}-\bar{X}\right]$$

Variance decomposition:

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2}_{SS_{centered}} = \underbrace{\sum_{i=1}^{g}(\bar{x}_i-\bar{x})^2 n_i}_{SS_{treatment}} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}_{SS_{residuals}}$$

Pivotal statistic:

$$F_0 = \frac{SS_{treat}/(g-1)}{SS_{res}/(n-g)} \sim F(g-1, n-g)$$

Reject H_0 if $F_0 > F_\alpha(g-1, n-g)$

Case 3: $p \geq 1$, $g \geq 2$

$$\left.\begin{array}{l} \underline{X}_{11}, \dots, \underline{X}_{1n_1} \overset{iid}{\sim} N_p(\underline{\mu}_1, \Sigma) \\ \vdots \\ \underline{X}_{g1}, \dots, \underline{X}_{gn_g} \overset{iid}{\sim} N_p(\underline{\mu}_g, \Sigma) \end{array}\right\} \perp\!\!\!\perp$$

modello:

$$\underline{X}_{ij} = \underline{\mu} + \underline{\tau}_i + \underline{\varepsilon}_{ij} \qquad i = 1, \dots, g \quad j = 1, \dots, n_i$$

$$\sum_{i=1}^{g} n_i \underline{\tau}_i = \underline{0}$$

$$\underline{\varepsilon}_{ij} \sim N(\underline{0}, \Sigma)$$

$$\underline{\mu} \leftarrow \bar{\underline{X}}$$

$$\underline{\tau}_i \leftarrow \bar{\underline{X}}_i - \bar{\underline{X}}$$

Goal:
$H_0: \underline{\tau}_1 = \underline{\tau}_2 = \dots = \underline{\tau}_g = \underline{0}$
$H_1: \exists\, \underline{\tau}_i \neq \underline{0}$

Covariance decomposition:

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underline{X}_{ij}-\bar{\underline{X}})(\underline{X}_{ij}-\bar{\underline{X}})^T}_{\text{covariability that we would compute under } H_0} = \underbrace{\sum_{i=1}^{g}(\bar{\underline{X}}_i-\bar{\underline{X}})(\bar{\underline{X}}_i-\bar{\underline{X}})^T n_i}_{:= B \text{ between}} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underline{X}_{ij}-\bar{\underline{X}}_i)(\underline{X}_{ij}-\bar{\underline{X}}_i)^T}_{:= W \text{ within}}$$

We want again see "$\frac{B}{W}$" to decide wether to reject or not

Proposals:

Λ	reject H_0 if Λ is ..
WILKS: $\Lambda_W = \frac{Det(W)}{Det(W+B)}$	small (*)
LAWLEY-HOTELLING $\Lambda_{LH} = Tr(BW^{-1})$	large
PILLAI: $\Lambda_P = Tr(B\cdot(B+W)^{-1})$	large

(interpretations on notes)

(*) when all the n_i are big enough:
Barlett's approximation:

$$-\left(n-1-\frac{p+g}{2}\right)\log\Lambda_W \sim \chi^2(p(g-1))$$

Reject at level α if:

$$-\left(n-1-\frac{p+g}{2}\right)\log\Lambda_W > \chi^2(p(g-1))$$

If we reject H_0: at what level the effect made effect?
(i.e. we want to compare τ_i and τ_k component-wise so we discover if level i produced a different effect than level k and on what component)

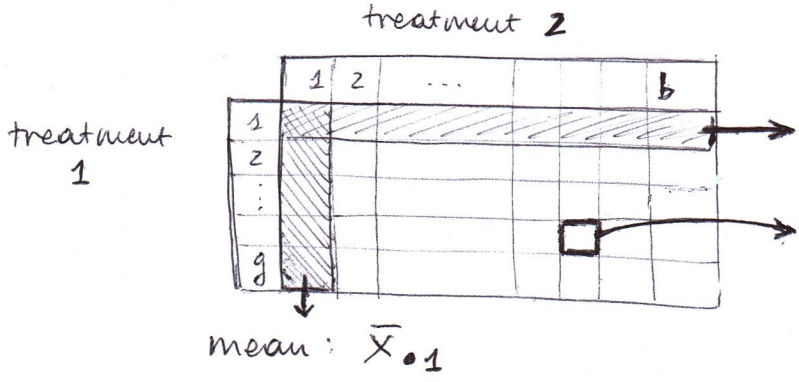
⟹ Bonferroni $CI(\tau_{il}-\tau_{ik})$ $\quad k,i = 1,\dots,g$ $\quad l = 1,\dots,p$

- $\bar{x}_{il}-\bar{x}_{ik} \sim N\left(\tau_{il}-\tau_{kl}, \frac{1}{n_i}\sigma_{ll}+\frac{1}{n_k}\sigma_{ll}\right)$
- $\frac{1}{n-g}W$ estimate Σ ⟹ $\frac{W_{ll}}{n-g}$ estimate σ_{ll}

$$\text{Bouf Sim } CI_{1-\alpha}(\tau_{il}-\tau_{kl}) = \left[\bar{x}_{il}-\bar{x}_{kl} \pm t_{\frac{\alpha}{pg(g-1)}}(n-g)\sqrt{\frac{W_{ll}}{n-g}\left(\frac{1}{n_k}+\frac{1}{n_i}\right)}\right]$$

(M)ANOVA TWO-WAYS

We have two treatments



i = level treat. 1
j = level treat. 2
$k = 1,\dots,n$
(we have n patients in this cell)

Model: $X_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$ $\quad i = 1,..,g \quad j = 1,..,b \quad k = 1,..,n$

(τ_i: treat. 1, β_j: treat. 2, γ_{ij}: interactions)

Constraints:

$$\sum_{i=1}^{g} \tau_i = 0 \qquad \sum_{j=1}^{b} \beta_j = 0 \qquad \sum_{i=1}^{g} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0$$

Decomposition of variance:

$$\sum_{i=1}^{g}\sum_{j=1}^{b}\sum_{k=1}^{n} (x_{ijk} - \bar{x})^2 = \sum_{i=1}^{g} bn (\bar{x}_{i\cdot} - \bar{x})^2 + \quad SS_{treat\ 1}$$

$$+ \sum_{j=1}^{b} gn (\bar{x}_{\cdot j} - \bar{x})^2 + \quad SS_{treat\ 2}$$

(this turn up in the additive model (with no interact))

$$+ \sum_{i=1}^{g}\sum_{j=1}^{b} (\bar{x}_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2 n + \quad SS_{interactions}$$

$$+ \sum_{i=1}^{g}\sum_{j=1}^{b}\sum_{k=1}^{n} (x_{ijk} - \bar{x}_{ij})^2 \quad SS_{residuals}$$

$\begin{cases} H_0 : \gamma_{ij} = 0 \\ H_1 : \exists\, \gamma_{ij} \neq 0 \end{cases} \longrightarrow$ Reject at level α if:

$$\frac{\frac{SS_{interactions}}{(g-1)(b-1)}}{\frac{SS_{res}}{gb(n-1)}} > F_\alpha\left((g-1)(b-1),\ gb(n-1)\right)$$

If we accept $H_0 \Longrightarrow$ additive model, and so:

$\begin{cases} H_0 : \tau_1 = \tau_2 = \dots = \tau_g = 0 \\ H_1 : \exists\, \tau_i \neq 0 \end{cases} \longrightarrow$ Reject at level α if:

$$\frac{\frac{SS_{treat\ 1}}{g-1}}{\frac{SS_{residuals}}{gbn-g}} > F_\alpha(g-1,\ gbn-g)$$

CLASSIFICATION

Supervised

$$\mathbb{X} = \begin{bmatrix} X_1 & X_2 & \dots & X_p & L \\ x_{11} & x_{12} & \dots & x_{1p} & \ell_1 \\ \vdots & & & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} & \ell_n \end{bmatrix}$$

Goal: Find $\delta : X \to \{1,..,g\}$

($\underline{X} = [X_1,..,X_p] \in X$)

(DISCRIMINANT ANALYSIS)

Unsupervised

$$\mathbb{X} = \begin{bmatrix} X_1 & X_2 & & X_p \\ x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

we believe there are hidden labels

Goal:
- estimate lables (CLUSTER ANALYSIS)
- Find $\delta : X \to \{1,..,g\}$ (DISCRIMINANT ANALYSIS)

Ingredients: **(SUPERVISED CLASSIFICATION)**

- $X|L=i \sim f_i(\underline{x})$ = distribution of the features in group 1
- $\mathbb{P}(L=i) = p_i$ = prior probabilities
- $c(i|j)$ = cost of misclassification ($c(i|j)$ = what we pay if we attribute to **i** a unit belonging to **j**)

Optimality criterium:

(Note that instead of defining δ we define a partition R_i)

Goal: min $ECM(\delta)$ (Expected Cost of Misclassification of δ)

- $g = 2$

$$R_1 = \{ \underline{x} \in \mathbb{R}^p : c(1|2) f_2(\underline{x}) p_2 \leq c(2|1) f_1(\underline{x}) p_1 \}$$

$$R_2 = \{ \underline{x} \in \mathbb{R}^p : c(2|1) f_1(\underline{x}) p_1 \leq c(1|2) f_2(\underline{x}) p_2 \}$$

Explaination:

we put $\underline{x}$ in R_1 if:

$$\text{cost}\left(\begin{matrix} \underline{x} \text{ belonging to } 2 \\ \underline{x} \text{ attributed to } 1 \end{matrix} \right) \leq \text{cost}\left(\begin{matrix} \underline{x} \text{ belonging to } 1 \\ \underline{x} \text{ attributed to } 2 \end{matrix} \right)$$

we're choosing where to put $\underline{x}$ basing on the minimum cost that we have to pay if we're wrong

- $g \geq 2$

$$R_1 = \{ \underline{x} \in \mathbb{R}^p : \sum_{k=2}^{g} c(1|k) f_k(\underline{x}) p_k \leq \sum_{k \neq j} c(j|k) f_k(\underline{x}) p_k, \; j = 2, \dots, g \}$$

$$R_2 = \{ \underline{x} \in \mathbb{R}^p : \sum_{k \neq 2} c(2|k) f_k(\underline{x}) p_k \leq \sum_{k \neq j} c(j|k) f_k(\underline{x}) p_k, \; j = 1, 3, \dots, g \}$$

$$\vdots$$

$$R_i = \{ \underline{x} \in \mathbb{R}^p : \sum_{k \neq i} c(i|k) f_k(\underline{x}) p_k \leq \sum_{k \neq j} c(j|k) f_k(\underline{x}) p_k, \; j \neq i \}$$

Explaination;

we put $\underline{x}$ in R_1 if:

$$\text{cost}\left(\begin{matrix} \underline{x} \text{ belonging to } k \; (\in \{2, \dots, g\}) \\ \underline{x} \text{ attributed to } 1 \end{matrix} \right) \leq \text{cost}\left(\begin{matrix} \underline{x} \text{ belonging to } j \; (j \neq k) \\ \underline{x} \text{ attributed to } j \end{matrix} \right) \quad \forall j$$

we chose to put $\underline{x}$ in R_1 if the cost of misclassification is minimum

$$\delta(\underline{x}) = i \iff \underline{x} \in R_i$$

Optimal classifier:

$$\delta(\underline{x}) = i \iff \left[\sum_{k \neq i} c(i|k) f_k(\underline{x}) p_k \leq \sum_{k \neq j} c(j|k) f_k(\underline{x}) p_k \quad \forall j \neq i \right]$$

$(i \in \{1, \dots, g\})$

$$\iff \left[\sum_{k \neq i} c(i|k) \mathbb{P}(L=k \mid \underline{X} = \underline{x}) \leq \sum_{k \neq j} c(j|k) \mathbb{P}(L=k \mid \underline{X} = \underline{x}) \ \forall j \neq i \right]$$

(i.e. the expected posterior cost $\leq$ all the other expected costs for all the other groups)

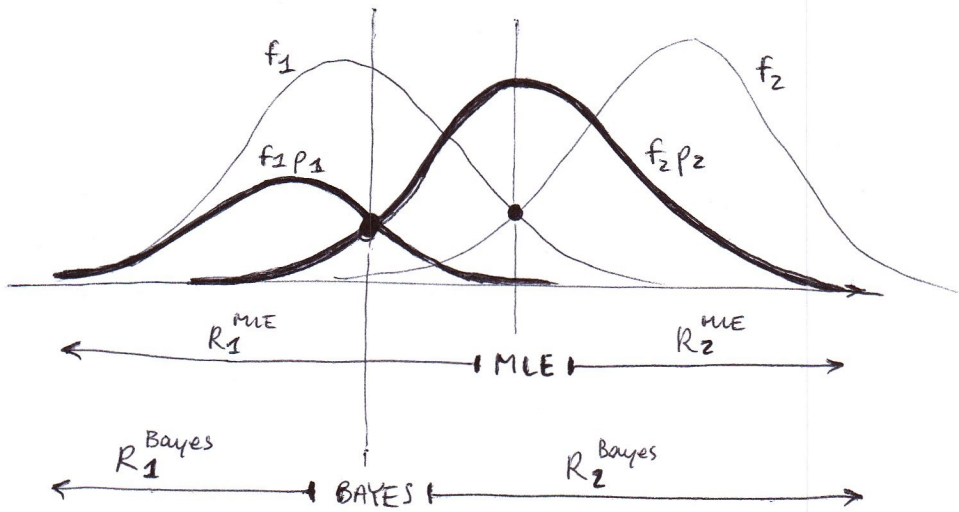
- BAYES CLASSIFIER: all costs are constant (and equal)

$$\delta(\underline{x}) = i \iff \mathbb{P}(L=j \mid \underline{X} = \underline{x}) \leq \mathbb{P}(L=i \mid \underline{X} = \underline{x}) \quad \forall j \neq i$$

(i.e. we attribute $\underline{x}$ to i if the posterior probability of belonging to group i is maximum)

- MLE CLASSIFIER: $c(i|j) = \text{const} \ \forall i \neq j$, $p_1 = p_2 = \dots = p_g = \frac{1}{g}$

$$\delta(\underline{x}) = i \iff f_j(\underline{x}) \leq f_i(\underline{x}) \quad \forall j \neq i$$



Bayes classifier

Bayes classifier is more flexible than it seems.
If we have costs (so Bayes assumptions seem to be violated) we can modify the priors in order to take into account the costs and then we go back to Bayes classifier.
(Costs and priors play a similar role)

Ex. $c(i|k) = c_k \geq 0$ (we pay c_k for every unit belonging to k not attributed to k (wherever it goes))

$$\pi_k := \frac{c_k p_k}{\sum c_j p_j}$$ are acting like priors

BAYES CLASSIFIER

$$\Rightarrow \quad R_i = \left\{ \underline{x} \in \mathbb{R}^p : \sum_{k \neq i} f_k(\underline{x}) \pi_k \leq \sum_{k \neq j} f_k(\underline{x}) \pi_k \quad j = 1, \dots, g \right\}$$

Special Bayes classifiers

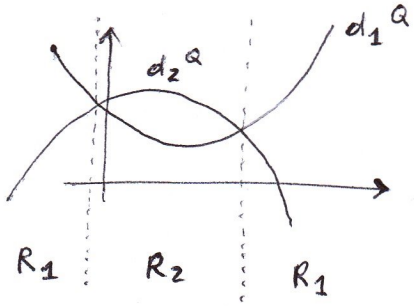
- **QDA** (Quadratic Discriminant Analysis)

$\underline{X} \mid L=i \sim N_p(\underline{\mu}_i, \Sigma_i)$

$\Rightarrow$ we can substitute the generic $\mathbb{P}(L=j \mid \underline{X}=\underline{x}) \leq \mathbb{P}(L=i \mid \underline{X}=\underline{x})$:

$$R_i = \{ \underline{x} \in \mathbb{R}^p : d_i^Q(\underline{x}) \geq d_j^Q(\underline{x}),\ j=1,\dots,g \}$$

$$d_i^Q(\underline{x}) = \log(p_i) - \frac{1}{2}\log(\det(\Sigma_i)) - \frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \Sigma_i^{-1} (\underline{x}-\underline{\mu}_i)$$



$(g=2)$

- **LDA** (Linear Discriminant Analysis)

$\underline{X} \mid L=i \sim N_p(\underline{\mu}_i, \Sigma)$ $\quad (\Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma)$

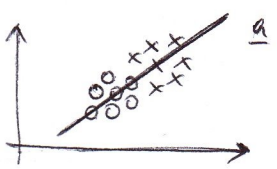
$$R_i = \{ \underline{x} \in \mathbb{R}^p : d_i^L(\underline{x}) \geq d_j^L(\underline{x}),\ j=1,\dots,g \}$$

$$d_i^L(\underline{x}) = \log(p_i) + \underline{\mu}_i^T \Sigma^{-1} \underline{x} - \frac{1}{2}\underline{\mu}_i \Sigma \underline{\mu}_i$$

Fisher's argument for LDA

Suppose $\underline{X} \mid L=i \sim (\underline{\mu}_i, \Sigma)$ (no gaussianity)

Goal: we want to find the direction $\underline{a}$ which maximizes the variability BETWEEN groups w.r.t. the variability WITHIN groups



$$\Rightarrow \begin{cases} \underline{a}_1 = \Sigma^{-1/2} \underline{e}_1 \\ \vdots \\ \underline{a}_s = \Sigma^{-1/2} \underline{e}_s \end{cases}$$

$$\bar{\underline{\mu}} = \frac{1}{g}\sum_{i=1}^{g} \underline{\mu}_i$$

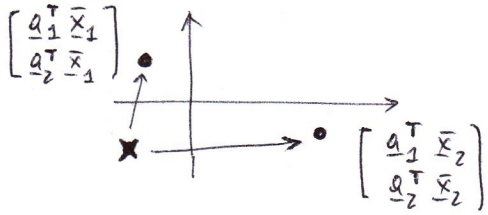
where $\underline{e}_j$ are the eigenvectors of $\Sigma^{-\frac{1}{2}} B \Sigma^{-\frac{1}{2}}$ $\quad \left(B = \frac{1}{g-1}\sum_{i=1}^{g} (\underline{\mu}_i - \bar{\underline{\mu}})(\underline{\mu}_i - \bar{\underline{\mu}})^T\right)$

Estimates: $\hat{\underline{\mu}}_i = \bar{\underline{X}}_i$, $\hat{\Sigma} = \frac{1}{n-g}\sum_{i=1}^{g} (n_i - 1) S_i$

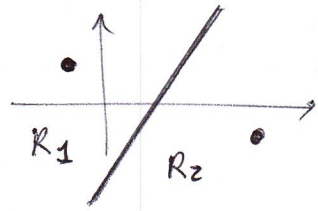
Classifier:

1. $\bar{\underline{X}}_i \longrightarrow [\, \underline{a}_1^T \bar{\underline{X}}_i, \dots, \underline{a}_k^T \bar{\underline{X}}_i \,|\, \dots, \underline{a}_s^T \bar{\underline{X}}_i \,]^T$ — we consider only the first k projections
2. observation $\underline{x} \longrightarrow [\, \underline{a}_1^T \underline{x}, \dots, \underline{a}_k^T \underline{x} \,]^T$
3. $R_i = \{ \underline{x} \in \mathbb{R}^p : \sum_{j=1}^{k} (\underline{a}_j^T \underline{x} - \underline{a}_j^T \bar{\underline{X}}_i)^2 \leq \sum_{j=1}^{k} (\underline{a}_j^T \underline{x} - \underline{a}_j^T \bar{\underline{X}}_h)^2,\ h=1,\dots,g \}$

We are attributing $\underline{x}$ to the closest mean $\bar{\underline{x}}_i$:



$\underline{x}$ closest to the mean of group 1 $\Rightarrow \delta(\underline{x}) = 1$



Estimate of the parameters

- QDA : $\hat{\underline{\mu}}_k = \frac{1}{n_k} \sum_{\{i : l_i = k\}} \underline{x}_i = \bar{\underline{x}}_k$

$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{\{i : l_i = k\}} (\underline{x}_i - \bar{\underline{x}}_k)(\underline{x}_i - \bar{\underline{x}}_k)^T = S_k$$

- LDA : $\hat{\underline{\mu}}_k = \bar{\underline{x}}_k$

$$\hat{\Sigma} = \frac{1}{n-g} \sum_{k=1}^{g} S_k (n_k - 1)$$

Evaluating a classifier

We have to estimate the Actual Error Rate of δ ($AER(\delta)$).

- non-parametric estimate

  We apply δ to the training set and we compute the confusion matrix (suppose we have $g = 2$):

actual L \ attributed L	1	2
1	n_{11}	n_{12}
2	n_{21}	n_{22}

$$\hat{AER}(\delta) = \frac{n_{12} + n_{21}}{n}$$

- leave-one-out cross validation

  For $i = 1, \dots, n$

  1. we take out "i" from the training set
  2. we train δ on $\mathcal{X}_{-i} \Rightarrow \delta_{-i}$
  3. $\delta_{-i}(\underline{x}_i) = \hat{l}_i$
  4. $\varepsilon_i = \begin{cases} 1 & \text{if } \hat{l}_i \neq l_i \\ 0 & \text{if } \hat{l}_i = l_i \end{cases}$

$$\Rightarrow \hat{AER}(\delta) = \frac{\sum_{i=1}^{n} \varepsilon_i}{n}$$

(the final considered classifier is δ, not δ_{-i})

- K-Fold cross validation

  Equal to leave-one-out but we take out a set of k elements and not only one element. This is to reduce the variance of $\hat{AER}(\delta)$ estimated with leave-one-out

K-fold cross-validation algorithm:

0. Set $k < n$ and randomly split the units of the training set in k parts (randomly = permute the rows (randomly) and then split in k parts)

For $j = 1, .., k$

1. Take out part j from training set
2. Train δ on $\mathcal{X}_{-\text{part}\,j} \Longrightarrow \delta_{-\text{part}\,j}$
3. Apply $\delta_{-\text{part}\,j}$ to the part j:

$$Err_j = \frac{1}{n_j} \sum_{\{i \in \text{part}\,j\}} \varepsilon_i$$

$$\varepsilon_i = \begin{cases} 1 & \delta_{-\text{part}\,j}(\underline{x}_i) \neq \ell_i \\ 0 & \delta_{-\text{part}\,j}(\underline{x}_i) = \ell_i \end{cases}$$

$$\Longrightarrow \hat{AER}(\delta) = \frac{1}{n} \sum_{j=1}^{k} n_j \, Err_j$$

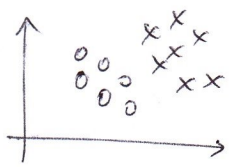
Note: by initializing B times we get: $\hat{AER}_1(\delta), .., \hat{AER}_B(\delta)$

$$\Longrightarrow \begin{cases} \mathbb{E}\left[\hat{AER}(\delta)\right] = \frac{1}{B} \sum_{j=1}^{B} \hat{AER}_j(\delta) \\ Var\left(\hat{AER}(\delta)\right) = \frac{1}{B-1} \sum_{j=1}^{B} \left(\hat{AER}_j(\delta) - \mathbb{E}\left[\hat{AER}(\delta)\right]\right)^2 \end{cases}$$

$$CI_{1-\alpha}\left(\mathbb{E}\left[\hat{AER}(\delta)\right]\right) = \left[\mathbb{E}\left[\hat{AER}(\delta)\right] \pm \sqrt{\frac{Var(\hat{AER}(\delta))}{B}}\, z_{\frac{\alpha}{2}}\right] \quad \text{(CLT)}$$

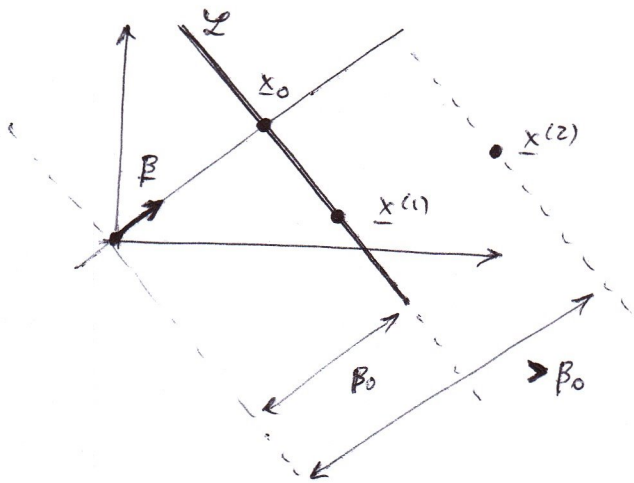
Support Vector Machines (SVM)

If we have a set (suppose $g = 2$):



$\Longrightarrow$ either there exists a separating hyperplane or not

- $\exists$ SEPARATING HYPERPLANE



To identify the hyperplane we introduce a vector $\underline{\beta}$ s.t. $\underline{\beta} \in \mathbb{R}^p$,

$\underline{\beta} \perp \mathcal{L}$, $\|\underline{\beta}\| = 1$

$\underline{x}_0 \in span(\underline{\beta}) \cap \mathcal{L}$

$\beta_0 = \|\underline{x}_0\|$

A generic point $\underline{x}$ can (or not) be on $\mathcal{L}$:

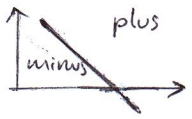
$$\underline{x}^{(1)} \in \mathcal{L} \iff \Pi_{\underline{x}^{(1)}|\mathcal{L}(\underline{\beta})} = \underline{x}_0$$

$$\iff \underline{\beta}^T \underline{x}^{(1)} = \beta_0$$

For example, $\underline{x}^{(2)} \notin \mathcal{L} \Rightarrow \underline{\beta}^T \underline{x}^{(2)} > \beta_0$.

$\underline{\beta}^T \underline{x} - \beta_0$ measures (with sign) the distance between $\underline{x}$ and $\mathcal{L}$

Let say:



$$\Rightarrow \begin{cases} \underline{\beta}^T \underline{x} - \beta_0 > 0 \Rightarrow \underline{x} \in \text{Plus} \\ \underline{\beta}^T \underline{x} - \beta_0 < 0 \Rightarrow \underline{x} \in \text{Minus} \end{cases}$$

Suppose we have 2 groups:

$$\begin{cases} y_i = 1 & \text{if } l_i = 1 \\ y_i = -1 & \text{if } l_i = 2 \end{cases} \quad (= \text{"plus"} = 1 \quad \text{"minus"} = 2)$$

$\Rightarrow$ Distance between $\underline{x}_i$ and $\mathcal{L}$: $y_i(\underline{\beta}^T \underline{x}_i - \beta_0)$:

Let $M := \min\{ y_i(\underline{\beta}^T \underline{x}_i - \beta_0) : i = 1, \dots \}$

Optimal separating plane:

$\mathcal{L}$ (to $\underline{\beta}$ and β_0) s.t. $\max_{\beta_0, \underline{\beta}} M$

under the constraints:

$$\begin{cases} \|\underline{\beta}\| = 1 \\ y_i(\underline{\beta}^T \underline{x}_i - \beta_0) \geq M \end{cases} \qquad i = 1, \dots, n$$

- ~~$\exists$~~ SEPARATING HYPERPLANE

  We can either:

  - Allow some overlapping:

$$\max_{\underline{\beta}, \beta_0} M \quad \text{s.t.} \quad \begin{cases} \|\underline{\beta}\| = 1 \\ y_i(\underline{\beta}^T \underline{x}_i - \beta_0) \geq M(1 - \varepsilon_i) \\ \varepsilon_i \geq 0 \\ \sum \varepsilon_i \leq c \end{cases} \leftarrow \text{Budget constraint}$$

  - Use non-linear boundaries

UNSUPERVISED CLASSIFICATION

Idea: units belonging to the same group are more similar than units belonging to other groups.

Dissimilarity functions:

1. $d(\underline{x}, \underline{y}) = \sqrt{(\underline{x} - \underline{y})^T (\underline{x} - \underline{y})} = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$ — Euclidean
2. $d_{\Sigma^{-1}}(\underline{x}, \underline{y}) = \sqrt{(\underline{x} - \underline{y})^T \Sigma^{-1} (\underline{x} - \underline{y})}$ — Mahalanobis
3. $d(\underline{x}, \underline{y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^m \right)^{1/m}$ — ℓ^m distance (Minkowski)

4. $d(\underline{x}, \underline{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{x_i + y_i}$ Canberra

Dissimilarity matrix:

$$X = \begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} \longrightarrow d_{ij} = d(\underline{x}_i, \underline{x}_j) \longrightarrow D = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots \\ d_{21} & 0 & d_{23} & \cdots \\ \vdots & & & \end{bmatrix}$$

Dissimilarity between clusters:

U, V two sets of points: $d(U,V) = ?$

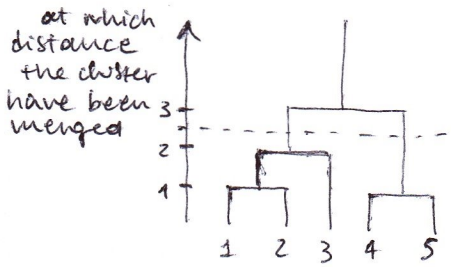
- Single linkage: $d(U,V) = \min \{ d(\underline{x}, \underline{y}) : \underline{x} \in U, \underline{y} \in V \}$
- Complete linkage: $d(U,V) = \max \{ d(\underline{x}, \underline{y}) : \underline{x} \in U, \underline{y} \in V \}$
- Average linkage: $d(U,V) = \frac{1}{\#U \, \#V} \sum_{\substack{\underline{x} \in U \\ \underline{y} \in V}} d(\underline{x}, \underline{y})$

Hierarchical Agglomerative Clustering Algorithm

- Set a d and a linkage
- Initialization: every unit is a cluster
- Until convergence repeat:
  1. merge the two cluster which are less similar
  2. compute the new distance matrix

Here we use the linkage (and in the linkage we use d)

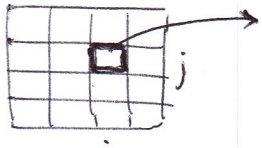
Graphical representation:
Dendrogram



← if we look at the data at this distance we see two clusters

Another representation of the Dendrogram:
Cophenetic distance

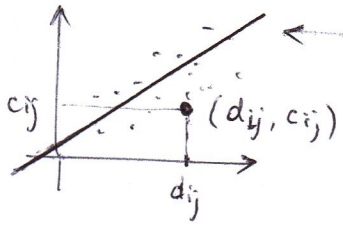
Dendrogram ⟶



level at which i and j have been merged

For every couple of elements: $(\underline{x}_i, \underline{x}_j) \longrightarrow (d_{ij}, c_{ij})$ $\quad i,j = 1, \dots, n$

⟹



$|CPCC| = |Corr(C,D)|$

Cophenetic correlation coefficient

$= |Corr \{ (d_{ij}, c_{ij}) : i,j = 1, \dots, n \}|$

(the closest to 1 the better the coph. matrix is repr. the dendrogr.)

Ward's method for Hierarchical clustering

Suppose we split data into k groups.
Let $C_1, .., C_k$ be clusters.

$$ESS_j := \sum_{\underline{x}_i \in C_j} (\underline{x}_i - \bar{\underline{x}}_j)^T (\underline{x}_i - \bar{\underline{x}}_j) = \sum_{\underline{x}_i \in C_j} \| \underline{x}_i - \bar{\underline{x}}_j \|^2$$

↑ barycenter of the cluster

$$ESS := ESS_1 + .. + ESS_k.$$

At the next iteration we merge the two cluster which generate the minimum increase of ESS.
(Method focus on the minimization of the loss of informations due to merging clusters)

K-means : non-hierarchical method of clustering

Def. Given a cluster $C_j \subseteq$ training set, we call centroid of C_j:

$$\bar{\underline{x}}_j = \underset{\underline{x} \in \mathbb{R}^p}{\arg\min} \sum_{\underline{x}_i \in C_j} d^2(\underline{x}_i, \underline{x})$$

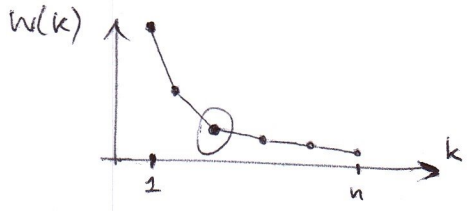
Optimal clustering: find $C_1, .., C_k$ s.t. $\sum_{j=1}^{k} \left[\sum_{\underline{x}_i \in C_j} d^2(\underline{x}_i - \bar{\underline{x}}) \right]$ is min.

k-means algorithm:

- Initialization:
  - randomly create $C_1, .. C_k$ in the training set ($\Rightarrow$ step. 1)
  - randomly assign k centroids: $\bar{\underline{x}}_1, .., \bar{\underline{x}}_k$ ($\Rightarrow$ step. 2)
- Iterate until convergence:
  1. for $j = 1, .., k$ compute the centroids of $C_j \Rightarrow \bar{\underline{x}}_1, .., \bar{\underline{x}}_k$
  2. for each stat. unit $\underline{x}_i$:
     assign $\underline{x}_i$ to cluster C_j if:
     $$d^2(\underline{x}_i, \bar{\underline{x}}_j) = \min \{ d^2(\underline{x}_i, \bar{\underline{x}}_1), .., d^2(\underline{x}_i, \bar{\underline{x}}_k) \}$$

How to choose k?

$$w(k) = \sum_{j=1}^{k} \left[\sum_{\underline{x}_i \in C_j} d^2(\underline{x}_i, \bar{\underline{x}}_j) \right] \quad .:$$



Multidimensional scaling

Given the distances* among n statistical units, look for the k-dimensional representation of the n statistical units s.t. the distances* among the representations of the n units are as close as possible to the original distances* among the n units.

* distances = dissimilarities

REGRESSION

General goal: explain Y in terms of $\underline{X}$.
More specifically the regression is concentrated in estimating $\mathbb{E}[Y|\underline{X}=\underline{x}] = f(\underline{x})$
(:= Regression function)

Two basically approach:

1. Totally data-driven (non-parametric) $\longrightarrow$ CART classification and regression trees
2. Parametric approach (model based)

Linear Models

$$\mathbb{X} \longrightarrow \underset{\text{design matrix}}{\mathbb{Z}} = \begin{bmatrix} & z_1 & \cdots & z_r \\ 1 & z_{11} & \cdots & z_{1r} \\ \vdots & \vdots & & \vdots \\ 1 & z_{n1} & \cdots & z_{nr} \end{bmatrix} \in \mathbb{R}^{n \times r}$$

$z_1, \dots, z_r$ known functions of $X_1, \dots, X_p$

Linear model: $\mathbb{E}[Y|z_1, \dots, z_r] = \beta_0 + \beta_1 z_1 + \dots + \beta_r z_r$

Model for Y:

$$\underline{Y} = \mathbb{Z}\underline{\beta} + \underline{\varepsilon}$$

$$\underline{\beta} = [\beta_0, \beta_1, \dots, \beta_r]^T \in \mathbb{R}^{r+1}$$

$$\underline{\varepsilon} \in \mathbb{R}^n \text{ s.t. } \begin{cases} \mathbb{E}[\underline{\varepsilon}] = \underline{0} \\ \text{Cov}(\underline{\varepsilon}) = \sigma^2 I \end{cases}$$

i.e. $Y_i = \beta_0 + \beta_1 z_{i1} + \dots + \beta_r z_{ir} + \varepsilon_i$

$\varepsilon_1, \dots, \varepsilon_n$:
- uncorrelated
- same variance
- mean 0
- independent of the z's

Estimating $\underline{\beta}$ and σ^2

(i.e. fitting the model)

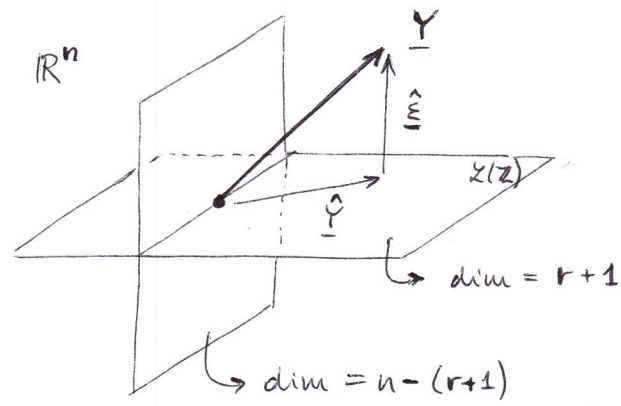
OLS : Ordinary Least Squares

$$\hat{\underline{\beta}} = \arg\min_{\underline{\beta}} \|\underline{Y} - \mathbb{Z}\underline{\beta}\|^2$$

$\mathbb{Z}\hat{\underline{\beta}} = \hat{\underline{Y}}$ fitted values

$\underline{Y} - \hat{\underline{Y}} = \hat{\underline{\varepsilon}}$ residuals

Goal: find $\hat{\underline{Y}} \in \mathcal{L}(\mathbb{Z})$ closest to $\underline{Y}$
(closest in the sense of Euclidean distance in $\mathbb{R}^n$)



Note: we can't take $r+1$ too large because we need degrees of freedom for the estimation of variance

Cases:

- $\operatorname{rank}(\mathbb{Z}) = r+1 < n$ ($\mathbb{Z}$ full rank)

$$\hat{\underline{Y}} = \mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{Y} := H\underline{Y} \qquad H = \Pi_{\cdot|\mathcal{Z}(\mathbb{Z})}$$

$$\hat{\underline{\beta}} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\underline{Y}$$

$$\hat{\underline{\varepsilon}} = (\underline{I}-H)\underline{Y}$$

- $\operatorname{rank}(\mathbb{Z}) = k < r+1 < n$ ($\dim(\mathcal{Z}(\mathbb{Z})) = k$)

$$\hat{\underline{Y}} = \mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-}\mathbb{Z}^T\underline{Y} \qquad (\exists!\ \hat{\underline{Y}})$$

$$\hat{\underline{\beta}} = (\mathbb{Z}^T\mathbb{Z})^{-}\mathbb{Z}^T\underline{Y} \qquad (\nexists!\ \hat{\underline{\beta}})$$

$$(\mathbb{Z}^T\mathbb{Z})^{-} = \sum_{i=1}^{k} \frac{1}{\lambda_i}\underline{e}_i\underline{e}_i^T \qquad \text{where} \begin{cases} \mathbb{Z}^T\mathbb{Z} = \sum_{i=1}^{r+1}\lambda_i\underline{e}_i\underline{e}_i^T \\ \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k = 0 = \lambda_{k+1} = \dots = \lambda_{r+1} \end{cases}$$

- $\operatorname{rank}(\mathbb{Z}) = r+1 = n$ ($\mathcal{Z}(\mathbb{Z}) = \mathbb{R}^n$)

$$\hat{\underline{Y}} = \underline{Y}, \qquad \hat{\underline{\varepsilon}} = \underline{0}$$

Coefficients of determination

- $$\underbrace{\sum_{i=1}^{n} Y_i^2}_{SS_{TOT}} = \underbrace{\sum_{i=1}^{n} \hat{Y}_i^2}_{SS_{REG}} + \underbrace{\sum_{i=1}^{n} \varepsilon_i^2}_{SS_{RES}}$$

- Decomposition of variance:

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{CSS_{TOT}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{CSS_{Reg}} + \underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}_{SS_{Res}}$$

(Centered sum of squares)

$$\Longrightarrow \quad R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = 1 - \frac{SS_{Res}}{CSS_{TOT}}$$

Remember that it doesn't hold if $\underline{1} \notin \mathbb{Z}$ (in that case: $\tilde{R}^2 = 1 - \frac{\|\hat{\underline{\varepsilon}}\|^2}{\|\underline{Y}\|^2}$)

$$\Longrightarrow \quad R^2_{adj} = 1 - \frac{\frac{\sum_{i=1}^{n}\varepsilon_i^2}{n-(r+1)}}{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}$$

Properties of $\hat{\underline{\beta}}$ and $\hat{\underline{\varepsilon}}$:

- $\mathbb{E}[\hat{\underline{\beta}}] = \underline{\beta}$
- $\mathrm{Cov}(\hat{\underline{\beta}}) = \sigma^2 (\mathbb{Z}^T \mathbb{Z})^{-1}$
- $\mathbb{E}[\hat{\underline{\varepsilon}}] = \underline{0}$
- $\mathrm{Cov}(\hat{\underline{\varepsilon}}) = \sigma^2 (I - H)$
- $\mathbb{E}[\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}] = \mathbb{E}\left[\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right] = \sigma^2 (n-(r+1))$

  $\Longrightarrow S^2 := \dfrac{\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}}{n-(r+1)}$ is s.t. $\mathbb{E}[S^2] = \sigma^2$

From now on: $\underline{\varepsilon} \sim N_n(\underline{0}, \sigma^2 I)$:

- $\hat{\underline{\beta}}$ and $\hat{\sigma}^2 = \dfrac{\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}}{n}$ are MLE
- $\hat{\underline{\beta}} \sim N_{r+1}(\underline{\beta}, \sigma^2 (\mathbb{Z}^T \mathbb{Z})^{-1})$
- $\hat{\underline{\varepsilon}} \sim N_n(\underline{0}, \sigma^2 (I - H))$ } $\perp\!\!\!\perp$
- $\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \sim \sigma^2 \chi^2(n-(r+1))$

Inference:

$$\begin{cases} \dfrac{1}{S^2} (\hat{\underline{\beta}} - \underline{\beta})^T (\mathbb{Z}^T \mathbb{Z}) (\hat{\underline{\beta}} - \underline{\beta}) \sim (r+1) F(r+1, n-(r+1)) & \left(S^2 = \dfrac{\hat{\underline{\varepsilon}}^T \hat{\underline{\varepsilon}}}{n-(r+1)}\right) \\ CR_{1-\alpha}(\underline{\beta}) = \left\{ \underline{\beta} \in \mathbb{R}^p : \dfrac{1}{S^2} (\hat{\underline{\beta}} - \underline{\beta})^T (\mathbb{Z}^T \mathbb{Z}) (\hat{\underline{\beta}} - \underline{\beta}) \le F_\alpha(r+1, n-(r+1)) \cdot (r+1) \right\} \end{cases}$$

σ^2

$$\begin{cases} \dfrac{(n-(r+1)) S^2}{\sigma^2} \sim \chi^2(n-(r+1)) \\ CI_{1-\alpha}(\sigma^2) = \left\{ \sigma^2 \in \mathbb{R} : \dfrac{(n-(r+1)) S^2}{\chi^2_{\frac{\alpha}{2}}(n-(r+1))} \le \sigma^2 \le \dfrac{(n-(r+1)) S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-(r+1))} \right\} \end{cases}$$

Simultaneous **CI for** $\underline{a}^T \underline{\beta}$:

$$\underbrace{Sim\, CI_{1-\alpha}(\underline{a}^T \underline{\beta})}_{\text{Simultaneous } CI_{1-\alpha} \text{ for any linear combination of } \beta\text{'s}} = \left[\underline{a}^T \hat{\underline{\beta}} \pm \sqrt{\underline{a}^T (\mathbb{Z}^T \mathbb{Z})^{-1} \underline{a}} \sqrt{S^2 (r+1) F_{1-\alpha}(r+1, n-(r+1))} \right]$$

Special case

$$Sim\, CI_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i \pm \sqrt{diag_i (\mathbb{Z}^T \mathbb{Z})^{-1}} \sqrt{S^2 (r+1) F_\alpha(r+1, n-(r+1))} \right]$$

Testing the β's

$$\begin{cases} H_0: C\underline{\beta} = \underline{0} \\ H_1: C\underline{\beta} \neq \underline{0} \end{cases} : \quad \frac{1}{S^2}\left(C\hat{\underline{\beta}}\right)^T\left(C(\mathbb{Z}^T\mathbb{Z})^{-1}C^T\right)^{-1}\left(C\hat{\underline{\beta}}\right) \sim p\,F(p, n-(r+1))$$

Reject H_0 : at level α if : $\frac{1}{S^2}\left(C\hat{\underline{\beta}}\right)^T\left(C(\mathbb{Z}^T\mathbb{Z})^{-1}C^T\right)^{-1}\left(C\hat{\underline{\beta}}\right) > p\,F_{1-\alpha}(p, n-(r+1))$

Special case I :

$$\begin{cases} H_0: \beta_r = \beta_{r-1} = \dots = \beta_{r-(p-1)} = 0 \\ H_1: \exists \beta_j \neq 0 \end{cases} \quad \text{(testing } p \text{ parameters)}$$

we're comparing $\underline{Y} = \mathbb{Z}\underline{\beta} + \underline{\varepsilon}$ vs. $\underline{Y} = \mathbb{Z}_1\underline{\beta}_1 + \underline{\varepsilon}_1$

$$\Longrightarrow \quad \frac{SS_{res}(\mathbb{Z}_1) - SS_{res}(\mathbb{Z})}{S^2 p} \sim F(p, n-(r+1)) \qquad \left(S^2 = \frac{\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}}{n-(r+1)}\right)$$

where $SS_{res}(\mathbb{Z}_1) - SS_{res}(\mathbb{Z}) = \hat{\underline{\varepsilon}}_1^T\hat{\underline{\varepsilon}}_1 - \hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}$

Special case II :

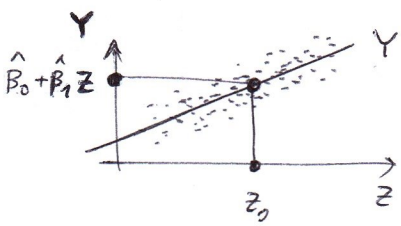
$$\begin{cases} H_0: \beta_1 = \beta_2 = \dots = \beta_r = 0 \\ H_1: \exists \beta_j \neq 0 \end{cases}$$

$$\Longrightarrow \quad \frac{SS_{res}(\mathbb{Z}_1) - SS_{res}(\mathbb{Z})}{S^2 r} = \frac{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n \hat{\varepsilon}_i^2}{r}}{\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-(r+1)}} \sim F(r, n-(r+1))$$

Prediction

$$\underline{Y} = \mathbb{Z}\underline{\beta} + \underline{\varepsilon}$$

$$\underline{z}_0 = [1 \;\; z_{0_1} \;\; \dots \;\; z_{0_r}]^T \longrightarrow Y_0 ? \qquad \begin{cases} Y_0 = \underline{z}_0^T\underline{\beta} + \varepsilon_0 & \varepsilon_0 \sim N(0, \sigma^2) \\ \mathbb{E}[Y_0] = \underline{z}_0^T\underline{\beta} & (\varepsilon_0 \perp\!\!\!\perp \underline{\varepsilon}) \end{cases}$$

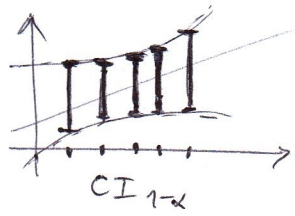


$$\frac{\underline{z}_0^T\hat{\underline{\beta}} - \underline{z}_0\underline{\beta}}{S\sqrt{\underline{z}_0(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{z}_0}} \sim t(n-(r+1))$$

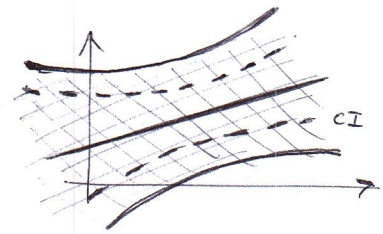
$$CI_{1-\alpha}(\underline{z}_0^T\underline{\beta}) = \left[\underline{z}_0^T\hat{\underline{\beta}} \pm S\sqrt{\underline{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{z}_0}\; t_{1-\frac{\alpha}{2}}(n-(r+1))\right]$$

$$\text{Sim } CI_{1-\alpha}(\underline{z}_0^T\underline{\beta}) = \left[\underline{z}_0^T\hat{\underline{\beta}} \pm S\sqrt{\underline{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\underline{z}_0}\;\sqrt{(r+1)F_{1-\alpha}(r+1, n-(r+1))}\right]$$

With $CI_{1-\alpha}(\underline{z}_0^T \underline{\beta})$ we generate a CI for any $\underline{z}_0$ but fixed.
With $Sim\,CI_{1-\alpha}(\underline{z}_0^T \underline{\beta})$ we generate a band of CI (so for any $\underline{z}_0$):



only the vertical line has a level of $1-\alpha$



all the band has level $1-\alpha$

Note: this is the prediction of $\mathbb{E}[Y_0 | \underline{z}_0]$, not Y_0!

$$\frac{Y_0 - \underline{z}_0^T \hat{\underline{\beta}}}{S\sqrt{1 + \underline{z}_0^T (\underline{Z}^T \underline{Z})^{-1} \underline{z}_0}} \sim t(n-(r+1))$$

$$\underbrace{PI_{1-\alpha}(Y_0)}_{\text{prediction interval of probability } 1-\alpha} = \left[\underline{z}_0^T \hat{\underline{\beta}} \pm S\sqrt{1 + \underline{z}_0^T (\underline{Z}^T \underline{Z})^{-1} \underline{z}_0}\; t_{1-\frac{\alpha}{2}}(n-(r+1)) \right]$$

$$\mathbb{P}(Y_0 \in PI_{1-\alpha}(Y_0)) = 1-\alpha$$

(Remember: they're one at the time!)

GLS Generalized least squares:

$$\hat{\underline{\beta}} = \arg\min_{\underline{\beta}} (\underline{Y} - \underline{Z}\underline{\beta})^T W^{-1} (\underline{Y} - \underline{Z}\underline{\beta})$$

$$\hat{\underline{\beta}} = (\underline{Z}^T W^{-1} \underline{Z})^{-1} \underline{Z}^T W^{-1} \underline{Y}$$

Collinearity

$\hat{\underline{\beta}} = (\underline{Z}^T \underline{Z})^{-1} \underline{Z}\,\underline{Y}$: if $(\underline{Z}^T \underline{Z})$ is close to be singular we have a problem due to collinearity (one regressor can be expressed as a linear combination of the others)

For every column:

$$\underline{z}_j := \gamma_0 + \gamma_1 \underline{z}_1 + \dots + \gamma_{j-1} \underline{z}_{j-1} + \gamma_{j+1} \underline{z}_{j+1} + \dots + \gamma_r \underline{z}_r$$

We substitute the j^{th} column of $\underline{Z}$ with $\underline{z}_j$ (↑) and we perform the linear model analysis again, getting R^2_j.

$$\Longrightarrow \quad VIF_j = \frac{1}{1 - R^2_j}$$

Variance Inflation Factor

If $VIF(\beta_j) > 5/10 \Rightarrow$ probably β_j can be expressed as a linear combination of the other regressors

Collineanity and variable selection

Let's work with the problem "centered":

$$\mathbb{Z} \longrightarrow \mathbb{Z}^* = \begin{bmatrix} z_{11} - \bar{z}_1 & \cdots & z_{1r} - \bar{z}_r \\ z_{21} - \bar{z}_1 & & z_{2r} - \bar{z}_r \\ \vdots & & \vdots \\ z_{n1} - \bar{z}_1 & & z_{nr} - \bar{z}_r \end{bmatrix} := \mathbb{Z}$$

$$\underline{Y} \longrightarrow \underline{Y}^* = [\, Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y} \,]^T := \underline{Y}$$

PCA Regression

- $\mathbb{Z} = [\, \underline{z}_1, \ldots, \underline{z}_r \,]$
- PCA on $\mathbb{Z} \Longrightarrow PC_1, \ldots, PC_r$
- Reduce dimensionality: $PC_1, \ldots, PC_K \quad K \leq r$

$$\mathbb{Z}^* := [\, \underline{PC}_1, \ldots, \underline{PC}_K \,]$$

- Fit $\underline{Y} = \mathbb{Z}^* \underline{\beta} + \underline{\varepsilon}$

$$\Longrightarrow \quad Y_0 = \underline{z}_0^T \hat{\underline{\beta}}$$

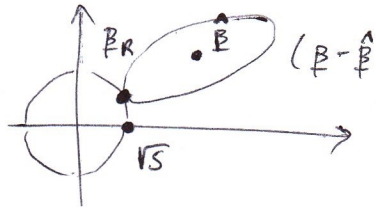
$$= \hat{\beta}_1 PC_1 + \ldots + \hat{\beta}_K PC_K$$

$$= \hat{\beta}_1 (e_{11} z_1 + \ldots + e_{r1} z_r) + \ldots + \hat{\beta}_K (e_{1K} z_1 + \ldots + e_{rK} z_r)$$

$$= z_1 (e_{11} \hat{\beta}_1 + \ldots + e_{1K} \hat{\beta}_K) + \ldots + z_r (e_{r1} \hat{\beta}_1 + \ldots + e_{rK} \hat{\beta}_K)$$

$$= \hat{\gamma}_0 z_1 + \ldots + \hat{\gamma}_r z_r$$

RIDGE Regression



$$(\underline{\beta} - \hat{\underline{\beta}})^T (\mathbb{Z}^T \mathbb{Z}) (\underline{\beta} - \hat{\underline{\beta}}) = t$$

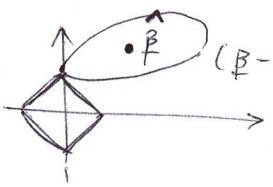
(not scale invariant!)

Problem:

$$\begin{cases} \arg\min_{\underline{\beta}} \| \mathbb{Z} (\underline{\beta} - \hat{\underline{\beta}}) \|^2 \\ \| \underline{\beta} \|^2 \leq s \\ \hat{\underline{\beta}} = (\mathbb{Z}^T \mathbb{Z})^{-1} \mathbb{Z}^T \underline{Y} \end{cases}$$

$$\Longrightarrow \quad \hat{\underline{\beta}}_R = (\mathbb{Z}^T \mathbb{Z} + \lambda I)^{-1} \mathbb{Z}^T \underline{Y}$$

LASSO Regression



$$(\underline{\beta} - \hat{\underline{\beta}})^T (\mathbb{Z}^T \mathbb{Z}) (\underline{\beta} - \hat{\underline{\beta}}) = t$$

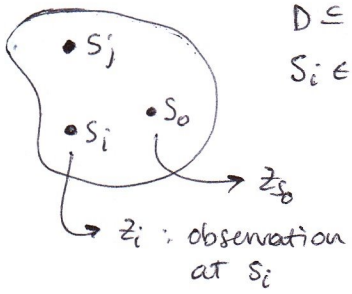
(λ with cross validation and prediction error for Y)

Problem:

$$\begin{cases} \arg\min (\underline{\beta} - \hat{\underline{\beta}})^T (\mathbb{Z}^T \mathbb{Z}) (\underline{\beta} - \hat{\underline{\beta}}) \\ \| \underline{\beta} \|_1 \leq s \\ \hat{\underline{\beta}} \ \text{OLS} \end{cases}$$

$$\Longrightarrow \arg\min \| \underline{Y} - \mathbb{Z} \underline{\beta} \| + \lambda \| \underline{\beta} \|_1$$

SPATIAL DATA



$D \subseteq \mathbb{R}^d$

$s_i \in D$

$s_1, \dots, s_n \in D \subseteq \mathbb{R}^d$ fixed

$z_{s_1}, \dots, z_{s_n}$ r.v. observations

Goals:
- Spatial dependence estimation
- Prediction (spatial) : kriging

$s_1, \dots, s_n \in D$ sites

$z_{s_1}, \dots, z_{s_n}$ observations : $\{ z_s, s \in D \}$

Assumptions:
- $\mathbb{E}[z_s] < \infty \quad \forall s \in D$
- $\mathrm{Var}(z_s) < \infty \quad \forall s \in D$

Definitions:
- $m_s := \mathbb{E}[z_s]$ spatial mean of $\{ z_s, s \in D \}$
- $C(s_1, s_2) := \mathrm{Cov}(z_{s_1}, z_{s_2})$ covariance function
- $\{ z_s, s \in D \}$ second order stationary if:
  1. $\mathbb{E}[z_s] = m \quad \forall s \in D$
  2. $\mathrm{Cov}(z_{s_1}, z_{s_2}) = \underbrace{C(s_1 - s_2)}_{:= \text{COVARIOGRAM}} \quad s_1, s_2 \in D$

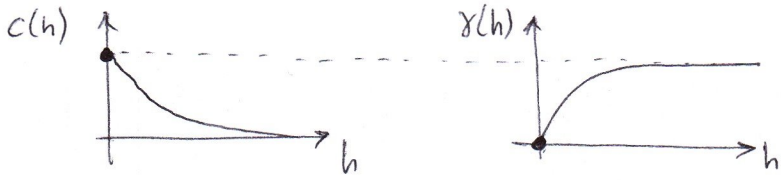
Properties:
- Simmetry: $C(-h) = C(h) \quad \forall h \in \mathbb{R}^d$
- Boundness: $|C(h)| \leq C(\underline{0}) \quad \forall h \in \mathbb{R}^d$
- Pos. def.: $\sum_{i,j} \lambda_i \lambda_j C(s_i - s_j) \geq 0 \quad \forall \lambda_i, \lambda_j \in \mathbb{R}, \ \forall s_i, s_j \in D$

Def. VARIOGRAM (under 2nd order of stationarity): $2\gamma(s_1 - s_2) := \mathrm{Var}(z_{s_1} - z_{s_2}) = \mathbb{E}\left[(z_{s_1} - z_{s_2})^2\right]$

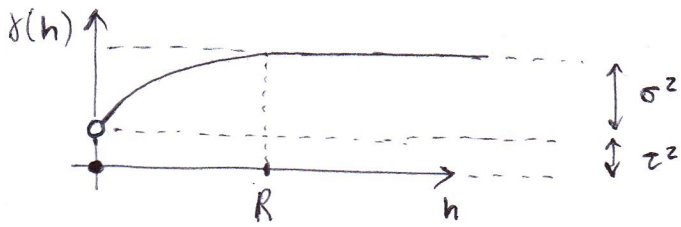
SEMI-VARIOGRAM: $\gamma(s_1 - s_2) := C(0) - C(s_1 - s_2)$

Properties:
- Simmetry: $\gamma(-h) = \gamma(h)$
- Null at 0: $\gamma(0) = 0$
- Conditional negative definite: $\sum_{i,j} \lambda_i \lambda_j \gamma(s_i - s_j) \leq 0$ $\quad \lambda_i, \lambda_j \in \mathbb{R}$, $\sum \lambda_i = 0$, $\forall s_i, s_j \in D$

Examples:



Structural properties:



partial still

nugget

$\sigma^2 + \tau^2$ still (variance of the process)

range (value of h s.t. the variogram reaches the still) it quantify the amount of dependence that we have in our data

Def. ISOTROPY: (under 2nd order of stationarity)

A field is isotropic if: $Cov(Z_{s_i}, Z_{s_j}) = C(\|s_i - s_j\|)$

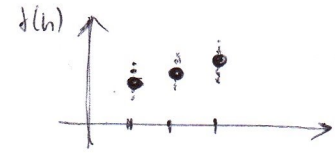
$s_i, s_j \in D$

Estimate Spatial Dependence

(= estimate the variogram)

- Empirical estimate:

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} (Z_{s_i} - Z_{s_j})^2$$

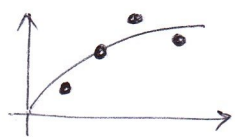


$$N(h) = \{(i,j) : \|s_i - s_j\| = h\}$$

$$N_2(h) = \{(i,j) : \|s_i - s_j\| = \Delta h\}$$ ← since we don't have so much data

- Parametric model:

$\gamma(h;\theta)$: we have to estimate θ (← best parameter vector to fit the empirical estimate)



• = empirical estimate

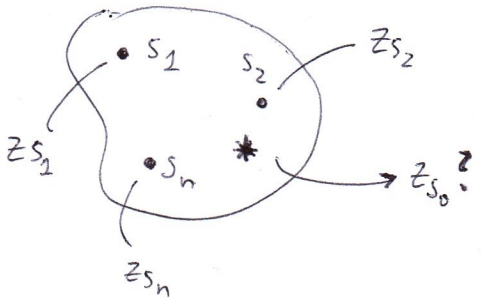
— = model

For example: $\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} \sum_{k=1}^{K} (\gamma(h_k) - \gamma(h_k;\theta))^2$ (OLS)

Properties:

- γ_1, γ_2 valid variograms $\Rightarrow$ $\gamma_1 + \gamma_2$ valid variogram
- $\alpha > 0$, γ_1 valid variogram $\Rightarrow$ $\alpha\gamma_1$ valid variogram

(Spatial) Predictor



$s_1, \dots, s_n$ sites in $D \subseteq \mathbb{R}^d$

$z_{s_1}, \dots, z_{s_n}$ obs. (real random variables)

How to make prediction at a location not observed?

Problem:

$$z^*_{s_0} = f(z_{s_1}, \dots, z_{s_n})$$

$$f = \text{argmin}\ \mathbb{E}\left[(z_{s_0} - f(z_{s_1}, \dots, z_{s_n}))^2\right]$$

Solution:

$$f(z_{s_1}, \dots, z_{s_n}) = \mathbb{E}\left[z_{s_0} \mid z_{s_1}, \dots, z_{s_n}\right]$$

Problem: if data are not gaussian f can be highly non-linear.
We need an alternative way:

KRIGING: it looks for the best
- linear
- unbiased

predictor

Problem (kriging):

$$z^*_{s_0} = \lambda_0^* + \sum_i \lambda_i^* z_{s_i}$$

where $\lambda_0^*, \dots, \lambda_n^*$ solve:

$$\begin{cases} \min\limits_{\lambda_0, \lambda_1, \dots \lambda_n \in \mathbb{R}} \mathbb{E}\left[(z_{s_0} - (\lambda_0 + \sum_i \lambda_i z_{s_i}))^2\right] \\ \text{such that}: \ \mathbb{E}\left[\lambda_0 + \sum_i \lambda_i z_{s_i}\right] = \mathbb{E}\left[z_{s_0}\right] \end{cases}$$

• Ordinary kriging (• unknown mean • stationarity 2nd order)

Assumptions:
- unknown mean $\mathbb{E}[z_s] := m \quad \forall s \in D$
- stationarity 2nd order
- $\text{Cov}(z_{s_1}, z_{s_2}) = C(s_1 - s_2) \quad \forall s_1, s_2 \in D$ **known**

(we actually estimate it: $\hat{\delta} \to \hat{C} \to \hat{\Sigma}$)

Solution:

Solve: $$\begin{bmatrix} \Sigma & \underline{1} \\ \underline{1}^T & 0 \end{bmatrix} \begin{bmatrix} \underline{\lambda} \\ \mu \end{bmatrix} = \begin{bmatrix} \underline{\sigma}_0 \\ 1 \end{bmatrix} \Longrightarrow \underline{\lambda}^* \Longrightarrow z^*_{s_0} = \underline{\lambda}^{*T} \underline{z}$$

where:

$$\Sigma = \text{Cov}(\underline{z}): \quad \Sigma_{ij} = \text{cov}(z_{s_i}, z_{s_j}) = C(s_i - s_j)$$

$$\sigma_{0i} = \text{Cov}(z_{s_0}, z_{s_i}) = C(s_0 - s_i)$$

μ = lagrangian multiplier

Note: we can also get the variance (of Ordinary kriging)

$$\sigma^2_{OK}(s_0) = C(0) - \sum_i \lambda_i^* C(s_0 - s_i) - \mu^* = \mathbb{E}\left[(z_{s_0} - z^*_{s_0})^2\right]$$

- **Universal Kriging** (• unknown mean • non stationarity)

Model : $\begin{cases} Z_s = m_s + \delta_s & s \in D \\ m_s = \mathbb{E}[Z_s] & \text{drift} \\ \delta_s = Z_s - m_s & \text{residual} \end{cases}$

Assumptions :

- $m_s = \sum_{j=1}^{L} a_j f_j(s)$: • a_j unknown coeff. • $f_j(s)$ known regressors
- δ_s random field ($\{\delta_s, s \in D\}$)

  2nd order stationarity random field for the residuals
  - $\mathbb{E}[\delta_s] = 0$
  - $Cov(\delta_{s_1}, \delta_{s_2}) = Cov(Z_{s_1}, Z_{s_2}) = C(s_1 - s_2)$

Solution :

Solve : $\begin{bmatrix} \Sigma & F \\ F^T & 0 \end{bmatrix} \begin{bmatrix} \underline{\lambda} \\ \underline{\mu} \end{bmatrix} = \begin{bmatrix} \underline{\sigma}_0 \\ \underline{f}_0 \end{bmatrix} \Rightarrow \underline{\lambda}^* \Rightarrow Z_{s_0}^* = \underline{\lambda}^{*T} \underline{Z}$

where :

$\Sigma = Cov(\underline{Z})$

F = design matrix

0 = matrix of zeros

$\sigma_{0i} = Cov(Z_{s_0}, Z_{s_i})$

$\underline{f}_0$ = design vector : $f_{0j} = f_j(s_0)$: value of the j-th regressor at the new location

Estimations :

If Σ is unknown we should estimate $\hat{\gamma} \to \hat{C} \to \hat{\Sigma}$.
We know how to estimate $\hat{\gamma}$ from stationary data, here we have to use the residuals (since the data are not stationary)

- $\delta_{s_1}, \dots, \delta_{s_n}$ given $\Rightarrow \hat{\gamma} \to \hat{C} \to \hat{\Sigma}$
- $\delta_{s_1}, \dots, \delta_{s_n}$ unknown $\Rightarrow \hat{\delta}_{s_i} \to \hat{\gamma} \to \hat{C} \to \hat{\Sigma}$

$\hat{\delta}_{s_i}$? : $\begin{cases} \hat{\delta}_{s_i} = Z_{s_i} - \hat{m}_{s_i} \\ \hat{m}_{s_i} = \sum_j \hat{a}_j f_j(s_i) \end{cases}$

where $\hat{\underline{a}}$ is estimated with an iterative procedure

Note: Here we can't get a real (not underestimate) kriging variance (since it's based on Σ and we only have $\hat{\Sigma}$)