Modélisation avancée

(C. Preda).

Sommaire

1. Données manquantes

- type de d.m.
- gestion de d.m.
- méthods d'imputation
- imputation multiple
- NIPALS.

2. Régression et dimension

- rappel sur la régression linéaire
- $n \geq p$ et $n < p$
- régularisation de la régression linéaire
  PCR, PLS, Ridge, Lasso
- la régression logistique et Poisson

3. Analyse de la variance

- facteurs fixes et facteurs aléatoires
- estimation des effets et régression
- interraction des facteurs.

Références :

- The Elements of Statistical Learning
  Data Mining, inférence and Prediction
  (T. Hastie, R. Tibshirani, J. Friedman)

  http://web.stanford.edu/~hastie/local.ftp/Springer

- Probabilités, analyse des données et statistique. (Ed. 3ème)
  Disponible à la Bibliothèque de Polytech'Lille.
  Voir aussi : http://cedric.cnam.fr/~saporta

- Advanced Statistical Modelling.
  (Course notes for University of Auckland)
  Alan Lee, Ross Ihaka, Chris Triggs

  https://www.stat.auckland.ac.nz/~stats330/coursebook.pdf

- Design and Analysis of Experiments
  (Douglas Montgomery)

Volume :

- 16 h cours / TD
- 12h TP (R et SAS).

Examen :

- DS écrit (2h) : coeff = 1.25
- TP - projet : coeff = 1.25

(UE 8.1 : Modélisation aléatoire : 7 ECTS)

Gestion des données manquantes

- une problématique fréquente en analyse des données (économie, santé, etc).

- differentes causes responsables de l'information manquante
  - non réponse à une question
  - perte des données
  - impossible d'observer la réponse
  - ⋮

- La question principale est :

Dans quelles conditions, les analyses réalisées en présence de valeurs manquantes restent valides ?

valide = estimateurs et leur précision (écart-type) restent consistents

Il faut comprendre le mécanisme étant à la base de la génération des données manquantes.

On se place dans un cadre multivarié

où

p variables aléatoires ($p \geq 2$) sont

observées sur

n unités statistiques :

	$X_1, \dots$	X_j	$\dots X_p$
1			
2			
$\vdots$			
i		x_{ij}	
$\vdots$			
n			

$$= \mathbb{X} = X_{obs} \cup X_m$$

$X_{obs} \downarrow$ données observées

$X_m \downarrow$ donnés manquant.

indicateur de données manquantes :

$$R_{ij} = \begin{cases} 1 & \text{si } x_{ij} \text{ est observé} \\ 0 & \text{si } x_{ij} \text{ est manquant} \end{cases}$$

$R(ij)$	$1 \dots$	j	$\dots p$
1			
2			
$\vdots$			
i		R_{ij}	
$\vdots$			
n			

$$= R$$

Mecanismes de génération des données manquants

M1 : Données manquantes complètement aléatoires
(MCAR = missing completely at Random)

$\Leftrightarrow$ la donnée manquante ne dépend pas des données observées ni des données manquantes :

$$\mathbb{P}(R|X) = \underbrace{\mathbb{P}(R|X_{obs}, X_m) = \mathbb{P}(R)}$$

Ce type de mécanisme est parfois appelé uniforme.

Dans ce type de scénario, l'analyse des cas complets conduit à des résultats valides (mais échantillon réduit).

M2 : Données manquantes aléatoires :
(MAR = missing at random)

$\Leftrightarrow$ la donnée manquante ne dépend pas des autres donnés manquantes mais que des donnés observées :

$$\mathbb{P}(R|X_{obs}, X_m) = \mathbb{P}(R|X_{obs}).$$

Exemple :

matière / étudiant	épreuve 1	épreuve 2	...
1	12	13	
2	8	?	
3	14	17	
4	6	?	
⋮			

Mécanisme générant les valeurs manquantes :

⚠ l'épreuve 2 est réalisée si épreuve 1 ≥ 10

Cas particulier : non-réponses par niveaux de la variable :

- X = poids : le patient refuse de donner son poids car obèse
- X = revenu : refus car trop élevé/petit.

Biais dans l'estimation des moments (moyennes, ...)

(pas de biais si analyses par classes (niveaux) !

M3 : Données manquantes non-aléatoires
(MNAR = Missing Not At Random)

$\Longleftrightarrow$ $\mathbb{P}(R \mid X_{obs}, X_m)$ ne peut pas être simplifié !

La présence d'une donnée manquante n'est plus explicable qu'par les données observées, mais dépend aussi des observations manquantes.

inférences valides $\Rightarrow$ préciser un modèle pour $\mathbb{P}(R \mid X_{obs}, X_m)$

⚠ Remarques :

- il est difficile de statuer sur le mécanisme générateur des données manquantes MCAR, MAR, MNAR.
- il est rare de savoir le bon modèle pour MNAR
- il est important de savoir comment les analyses sont influencées par le mécanisme choisi : MAR/MCAR. (analyse de sensitivité).

Quelle méthodologie adopter en présence de données manquantes ?

Méthode 1 : Cas complets

- élimine les sujets/unités statistiques ayant des données manquantes
- très simple !
- méthode par défaut des analyses multivariées (ACP, ...)

⚠ Problèmes :

- perte de puissance (tests statistiques, ...)
- biais (si MAR)
- on élimine des informations disponibles (observations partielles).

Recommandations :

- si % de données manquantes $\leq 5\%$ OK
- sinon : utiliser une méthode de gestion des données manquantes et utiliser toute l'information.

<u>Méthode 2</u> (point de vue très utilisateur !
Ce n'est pas le cas d'un étudiant(e) GIS !).

Données manquantes gérées par la procédure (technique) statistique employée

<u>Exemple</u> : la procédure FASTCLUS de SAS
– classification non-supervisée

$$d(i,j) = \sqrt{\frac{n}{m}\left(\sum_{k=1}^{m}(x_{ik}-x_{jk}\right)^2}$$

n – nombre de variables
m – nombre de variables <u>sans</u> valeurs manquantes pour les <u>individus i et j</u>

<u>Problèmes</u> – dépend fortement du logiciel utilisé

– certaines procédures sont naïves (remplacement avec la moyenne, mode, etc)

Méthode 3 : Imputation des données manquantes

imputation = remplacer une donnée manquante par une valeur plausible

Le fichier ainsi complété peut être traité ensuite par des techniques standard.

Conjecture : L'imputation est meilleure que l'analyse des cas complets.

⚠

il y a en pratique deux types d'imputation

- imputation
  - simple
  - multiple