Statistics for high dimensional data

Anne Gégout-Petit
(Université de Lorraine, Faculté des Sciences, IECL, Inria Nancy)
with the Benoît Liquet help !

4 Lectures on Model Selection and high-dimensional statistical analysis

1. Multiple testing issues
   - FWER: Family Wise Error Rate
   - False Discovery Rate
2. Model selection and assessment
   - Problematic, error
   - Criteria for linear model : AIC, BIC, Cp
   - Cross Validation and bootstrap method
   - Variable selection: subset
3. Regularization Methods
   - Ridge Regression
   - Lasso method
   - Elastic-net
4. Reduction method
   - Principal Component Analysis
   - Partial Least Square regression
   - Sparse Methods
   - Discriminant Analysis version
5. Handling missing data
   - Framework, definition
   - Method using maximisation of the likelihood: EM algorithm
   - Imputation methods
   - Method for and with PCA

Introduction

Introduction

Definition (Little et Rubin [2019]) *Missing data are unobserved values that would be meaningful for analysis if observed ; in other words, a missing value hides a meaningful value.*

Some notations

$$\mathbf{X} = \begin{pmatrix} X_{11} & \dots & X_{1j} & \dots & X_{1p} \\ & & \dots & & \\ X_{i1} & \dots & X_{ij} & \dots & X_{ip} \\ & & \dots & & \\ X_{n1} & \dots & X_{nj} & \dots & X_{np} \end{pmatrix}$$

is associated with

$$\mathbf{R} = \begin{pmatrix} R_{11} & \dots & R_{1j} & \dots & R_{1p} \\ & & \dots & & \\ R_{i1} & \dots & R_{ij} & \dots & R_{ip} \\ & & \dots & & \\ R_{n1} & \dots & R_{nj} & \dots & R_{np} \end{pmatrix}$$

and $R_{ij} = 1$ if X_{ij} is observed, $R_{ij} = 0$ either.

Framework

- The whole data are $(\mathbf{X}, \mathbf{R})$
- $(\mathbf{X}^{\text{obs}}, \mathbf{R})$ are the observed data
- $\mathbf{X}^{\text{miss}}$ are the missing data
- We say that data for subject i are full if $\prod_{j=1}^{p} R_{ij} = 1$
- Number of full subjects is ..

Missing data mechanisms

$f(\mathbf{X}, \mathbf{R})$ is the notation for the distribution of $(\mathbf{X}, \mathbf{R})$

Definition We say that the mechanism leading to missing data is MCAR (Missing Completely At Random) if

$$f(\mathbf{R}|\mathbf{X}) = f(\mathbf{R})$$

In this case, inference on the complete data is not biaised

Definition We say that the mechanism leading to missing data is MAR (Missing At Random) if

$$f(\mathbf{R}|\mathbf{X}) = f(\mathbf{R}|\mathbf{X}^{\text{obs}})$$

In this case, distribution of the observed data differs from the one of the complete data

Main methods for inference in the context of missing data

- Drop subjects or variables
- In the context of a survey : ponderation of the responses
- Inference from the maximisation of the likelihood : EM algorithm
- Imputation of the missing data
- Imputation of the missing data for and with PCA
- Other possibilities
- Packages

Likelihood methods

Likelihood methods

We suppose the data are MAR and we want make inference on a parameter β. We can define

- The completed log-likelihood $L_c(\mathbf{X}, \beta) = \sum_{i=1}^{n} ln(f(\mathbf{X}_i))$.
- The observed log-likelihood $L_o(\mathbf{X}^{\text{obs}}, \beta) = \sum_{i=1}^{n} \int ln(f(\mathbf{X}_i, \beta))d(\mathbf{X}_i^{\text{miss}})$
- We want to tackle $\hat{\beta} = \text{argmax}_{\beta} L_o(\mathbf{X}^{\text{obs}}, \beta)$

Difficulties because the integral in L_o

EM algorithm, principle

Dempster, 1977. The idea is to take the conditional expectation of L_c given $\mathbf{X}^{\text{obs}}$ and based on the assumption that distribution of $\mathbf{X}$ is given by the parameter $\tilde{\beta}$:

- Aim : compute $\mathbb{E}_{\tilde{\beta}}[L_c(\mathbf{X}, \beta)|\mathbf{X}^{\text{obs}}]$
- EM algorithm alternates between:
  - Expectation E : Computation of $Q(\beta, \tilde{\beta}^{[r-1]}) = \mathbb{E}_{\tilde{\beta}^{[r-1]}}[L_c(\mathbf{X}, \beta)|\mathbf{X}^{\text{obs}}]$
  - Maximisation M : $\tilde{\beta}^{[r]} = \text{argmax}_{\beta}(Q(\beta, \tilde{\beta}^{[r-1]}))$
- Properties : $Q(\beta, \tilde{\beta}^{[r]}) \geq Q(\beta, \tilde{\beta}^{[r-1]})$
- Convergence towards a local maximum, and to the global maxim in some cases
- Nice if the steps E and M are explicit

Exemple : EM algorithm for a linear model

Model

$$Y_i = \mathbf{X_i}^t \beta + \varepsilon_i, \qquad \text{where } Y_i \in \mathbb{R}, \qquad \beta = (\beta_1, \ldots, \beta_p)$$

Assumption on $\mathbf{X} = \left(X_{ij}\right)_{1 \leq i \leq n, 1 \leq j \leq p}$: $(\mathbf{X_i} = (X_{i1}, \ldots, X_{ip}))_{1 \leq i \leq n} \in \mathbb{R}^p$ are i.i.d. and $\mathbf{X_i} \simeq \mathcal{N}_p(\mathbf{0}_p, \mathbf{Id}_p)$,

Assumption on $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t \in \mathbb{R}^n$ and $\varepsilon \simeq \mathcal{N}_n(\mathbf{0}_n, \mathbf{Id}_n)$.

We want to infer β despite of missing data on some X_{ij} with a ignorable mechanism trough an EM algorithm

Exemple : EM algorithm for a linear model

Model

$$Y_i = \mathbf{X_i}^t \beta + \varepsilon_i, \qquad \text{where } Y_i \in \mathbb{R}, \qquad \beta = (\beta_1, \ldots, \beta_p)$$

Assumption on $\mathbf{X} = \left(X_{ij}\right)_{1 \le i \le n, 1 \le j \le p}$: $(\mathbf{X_i} = (X_{i1}, \ldots, X_{ip}))_{1 \le i \le n} \in \mathbb{R}^p$ are i.i.d. and $\mathbf{X_i} \simeq \mathcal{N}_p(\mathbf{0}_p, \mathbf{Id}_p)$,

Assumption on $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t \in \mathbb{R}^n$ and $\varepsilon \simeq \mathcal{N}_n(\mathbf{0}_n, \mathbf{Id}_n)$.

We want to infer β despite of missing data on some X_{ij} with a ignorable mechanism trough an EM algorithm

First task : write the log likelihood in the full observed case

$$L(\beta) = -\frac{1}{2} \sum_{i=1}^{n} \left(\beta^t \mathbf{X_i} \mathbf{X_i}^t \beta + \beta^t Y_i \mathbf{X_i} + Y_i^2\right)$$

whose solution is

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y}$$

$$\hat{\beta} = \left(\sum_{i=1}^{n} \mathbf{X_i} \mathbf{X_i}^t\right)^{-1} * \left(\sum_{i=1}^{n} Y_i \mathbf{X_i}\right)$$

Conditionnal expectation of the log-likelihood

For a current β^c we have to take the conditional expectation of the complete log-likelihood

$$\mathbb{E}_{\beta^c}[L(\beta)|\mathbf{X}^{\text{obs}}, Y] = -\frac{1}{2}\sum_{i=1}^{n}\left(\beta^t\mathbb{E}_{\beta^c}[\mathbf{X_i}\mathbf{X_i}^t|\mathbf{X_i}^{\text{obs}}, Y_i]\beta + \beta^t Y_i\mathbb{E}_{\beta^c}[\mathbf{X_i}|\mathbf{X}_i^{\text{obs}}, Y_i] + Y_i^2\right)$$

And to compute the β that maximise it :: *

$$\hat{\beta}^{c+1} = \left(\sum_{i=1}^{n}\mathbb{E}_{\beta^c}[\mathbf{X_i}\mathbf{X_i}^t|\mathbf{X_i}^{\text{obs}}, Y_i]\right)^{-1} * \left(\sum_{i=1}^{n} Y_i\mathbb{E}_{\beta^c}[\mathbf{X_i}|\mathbf{X_i}^{\text{obs}}, Y_i]\right)$$

Conditionnal expectation of the log-likelihood

For a current β^c we have to take the conditional expectation of the complete log-likelihood

$$\mathbb{E}_{\beta^c}[L(\beta)|\mathbf{X}^{\text{obs}}, Y] = -\frac{1}{2}\sum_{i=1}^{n}\left(\beta^t\mathbb{E}_{\beta^c}[\mathbf{X_i}\mathbf{X_i}^t|\mathbf{X_i}^{\text{obs}}, Y_i]\beta + \beta^t Y_i \mathbb{E}_{\beta^c}[\mathbf{X_i}|\mathbf{X}_i^{\text{obs}}, Y_i] + Y_i^2\right)$$

And to compute the β that maximise it :: *

$$\hat{\beta}^{c+1} = \left(\sum_{i=1}^{n}\mathbb{E}_{\beta^c}[\mathbf{X_i}\mathbf{X_i}^t|\mathbf{X_i}^{\text{obs}}, Y_i]\right)^{-1} * \left(\sum_{i=1}^{n} Y_i\mathbb{E}_{\beta^c}[\mathbf{X_i}|\mathbf{X_i}^{\text{obs}}, Y_i]\right)$$

There are different kind of terms

- $$\mathbb{E}_{\beta^c}[X_{ij}|\mathbf{X_i}^{\text{obs}}, Y_i] = \begin{cases} X_{ij} & \text{if } R_{ij} = 1 \\ \frac{(Y_i - \sum_k \beta_k^c R_{ik} X_{ik})\beta_j^c}{\sum_{k=1}^{p}(1-R_{ik})(\beta_k^c)^2+\sigma^2} & \text{if } R_{ij} = 0 \end{cases}$$

- $$\mathbb{E}_{\beta^c}[X_{ij}X_{ij'}|\mathbf{X_i}^{\text{obs}}, Y_i] = \begin{cases} X_{ij}X_{ij'} & \text{if } R_{ij}R_{ij'} = 1 \\ X_{ij}\mathbb{E}_{\beta^c}[X_{ij'}|\mathbf{X_i}^{\text{obs}}, Y_i] & \text{if } (R_{ij}, R_{ij'}) = (1,0) \\ -\frac{\beta_j^c\beta_{j'}^c}{\sum_{k=1}^{p}(1-R_{ik})(\beta_k^c)^2+\sigma^2} & \text{if } (R_{ij}, R_{ij'}) = (0,0) \text{ and } j \neq j' \\ 1 - \frac{(\beta_j^c)^2}{\sum_{k=1}^{p}(1-R_{ik})(\beta_k^c)^2+\sigma^2} & \text{if } (R_{ij}, R_{ij'}) = (0,0) \text{ and } j = j' \end{cases}$$

EM algorithm

Algorithm

- Step 0
  - Initialise the missing values by their expectation (given by the model : 0) $\mathbf{X_0}$
  - Compute $\beta^0 = \left(\mathbf{X_0}'\mathbf{X_0}\right)^{-1}\mathbf{X_0}'\mathbf{Y}$
- Step k+1
  - Step E, calculus of terms of $Q(\beta, \beta^k) = \mathbb{E}_{\beta^k}[L(\beta)|\mathbf{X}^{\text{obs}}, Y]$
  - Step M

$$\hat{\beta}^{k+1} = \left(\sum_{i=1}^{n} \mathbb{E}_{\beta^k}[\mathbf{X_i}\mathbf{X_i}^t|\mathbf{X_i}^{\text{obs}}, Y_i]\right)^{-1} * \left(\sum_{i=1}^{n} Y_i \mathbb{E}_{\beta^k}[\mathbf{X_i}|\mathbf{X_i}^{\text{obs}}, Y_i]\right)$$

$\hat{\beta}^{k+1}$ converges towards $\hat{\beta}$

Imputation methods

First Ideas

Imputation by mean

- X^j quantitative : missing X_{ij} imputed by mean on observed values $\frac{1}{\sum_{i=1}^n R_{ij}} \sum_{i=1}^n R_{ij} X_{ij}$
- X^j qualitative : missing X_{ij} imputed by the mode of observed $(X_{ij})_{1 \leq i \leq n}$
- X^j qualitative, you can also consider the NA as a level of X^j

Drawback
Does not take other variables into account

Exemple

Model $\mathbf{X_i} \simeq \mathcal{N}_2(0, \Gamma)$ with variances 1 and correlation ρ.

Observation : X^1 is observed but $P(R_{i2} = 0 | X_{i1}) = (1 + e^{1 - X_{i1}})^{-1}$

Exemple

Model $\mathbf{X_i} \simeq \mathcal{N}_2(0, \Gamma)$ with variances 1 and correlation ρ.

Observation : X^1 is observed but $P(R_{i2} = 0 | X_{i1}) = (1 + e^{1 - X_{i1}})^{-1}$

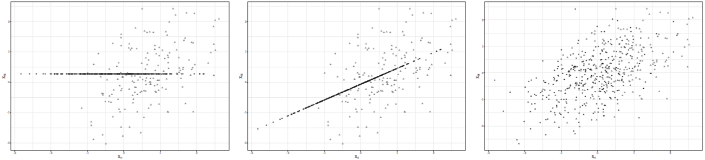


Figure 2.2 : *Jeu de données avec trois imputations différentes : imputation par l'espérance (gauche), imputation par l'espérance conditionnelle (centre), imputation stochastique par la loi conditionnelle (droite). Les données complètes sont en gris et les données ayant été imputées en noir.*

What about the mean ? the correlation between X^1 and X^2 according to the imputation method ?

Imputation methods by neighbours method

k-N-N imputation

Algorithm

- Suppose that X^1 is partially observed but all the other $(p-1)$ variables are observed
- choose a integer k
- If $R_{i1} = 0$, compute the distance between the vector $\mathbf{X}_i^{-1} = (X_{i2}, \ldots, X_{ip})$ and all the $\mathbf{X}_{i'}^{-1}$ with $R_{i1} = 1$
- Select the k-nearest neighbours of $\mathbf{X}_{i_1}^{(-1)}, \ldots, \mathbf{X}_{i_k}^{(-1)}$ of $\mathbf{X}_i^{(-1)}$
- Imput X_{i1} by the mean $\frac{1}{k}(\mathbf{X}_{i_1}^{(-1)} + \ldots + \mathbf{X}_{i_k}^{(-1)})$

regression imputation

Algorithm

- Suppose that X^1 is partially observed but all the other $p-1$ variables are observed
- Regress $(X^1)^{\text{obs}}$ on the $p-1$ other corresponding variables by the model $(X^1)^{\text{obs}} = \mathbf{X}^{(-1),\text{obs}}\beta + \varepsilon$, compute $\hat{\beta}$
- Imput X_{i1} by the $\hat{X}_{i1} = \mathbf{X}_i^{(-1)}\hat{\beta}$

Some times, it is only a "local regression" : evaluation of β is done on the k-NN of $\mathbf{X}_i^{-1}$

Random forest Imputation

Algorithm

- Suppose that X^1 is partially observed but all the other $p-1$ variables are observed
- Fit $(X^1)^{\text{obs}}$ on the $p-1$ other corresponding variables to have $(x^1)^{\text{obs}} \simeq RF(\mathbf{x}^{(-1),\text{obs}})$
- Imput x_{i1} by $\hat{x}_{i1} \simeq RF(\mathbf{x}_i^{(-1)})$

Random forest Imputation

Algorithm

- Suppose that X^1 is partially observed but all the other $p-1$ variables are observed
- Fit $(X^1)^{\text{obs}}$ on the $p-1$ other corresponding variables to have $(x^1)^{\text{obs}} \simeq RF(\mathbf{x}^{(-1),\text{obs}})$
- Imput x_{i1} by $\hat{x}_{i1} \simeq RF(\mathbf{x}_i^{(-1)})$

Miss Forest Algorithm

- Step 0: Complete X^1 by a simple method
- (while criterion), do
  - New matrix X replaced by imputed matrix
  - Fit $(X^1)^{imput}$ with $\mathbf{X}^{(-1)}$ by RF
  - If $R_{i1} = 1$, imput X_{i1} by $\hat{X}_{i1} \simeq RF(\mathbf{X}_i^{(-1)})$
- evaluate criterion

package Miss Forest

Imputation methods by PCA

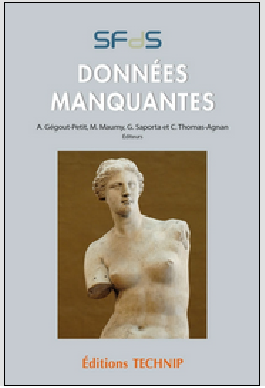
Multiple imputation

- For instance, in regression add the noise to reduce the over-fitting
- Do it several times and infer parameters
- Give an idea of the variability

R packages

- VIM to visualize the missing data
- mice
- Miss Forest
- missMDA based on FactoMineR

References



Wiki stat Toulouse
https://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-m-app-idm.pdf