Next steps

Antoine Ogier

19 juin 2018

Table des matières

Plan

Choses à faire :

- Simuler des jeux de données pour la prédiction.
- Insérer des données manquantes
- Choisir une méthode d'imputation.

- Appliquer cette méthode en séparant les données en un jeu d'entraînement et un d'apprentissage, de différentes manières
- Evaluer l'imputation, d'une part sur la distance aux données de test et d'autre part sur la performance de prédiction
- Répéter l'opération de nombreuses fois pour avoir une bonne estimation de chaque méthode

1 Simulation

1.1 Simulation des covariables

En premier lieu, on simule un tableau de données X qui servira de variable explicative. On utilise un nombre p à définir de colonnes et n de lignes.

1.1.1 Gaussienne multivariée

Option la plus simple, simuler X par tirage d'une loi normale $\mathcal{N}(\mu, \Sigma)$ pour des paramètres choisis. Sans perte de généralité on prend μ nul. Pour Σ, plusieurs possibilités :

- $\Sigma = I_p$ gaussiennes indépendantes
- $\Sigma = (1-\rho)I_p + \rho\mathbb{1}$ où $\mathbb{1}$ est la matrice remplie de 1 : Corrélation identique entre toutes les variables.
- $\Sigma = \begin{pmatrix} 1 & \rho & \dots & \rho & \rho & & & \\ \rho & 1 & \dots & \rho & \rho & & & \\ \vdots & \vdots & \ddots & \vdots & \vdots & & 0 & \\ \rho & \rho & \dots & 1 & \rho & & & \\ \rho & \rho & \dots & \rho & 1 & & & \\ & & & & & 1 & \dots & \rho \\ & & 0 & & & \vdots & \ddots & \vdots \\ & & & & & \rho & \dots & 1 \end{pmatrix}$: deux groupes de variables corrélées entre elles mais indépendantes d'un groupe à l'autre.

1.1.2 Noised low rank matrix with LRsim

Use *LRsim* from package *denoiseR* : a matrix of size $n \times p$ is drawn from a multivariate standard normal then projected on its first k columns and gaussian noise is added :

- $X_i \sim \mathcal{N}(0, 1)$
- $X_i = UDV^T$
- $X = U_k D_k V_k^T + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$

1.2 Simulation de la réponse

On choisit y la vaiable de réponse

1.2.1 Choix d'une colonne

Une colonne de X est choisie comme variable y.

1.2.2 Modèle de régression

On choisit β un coefficient de régression et $y = X\beta$

1.2.3 •

1.2.4 Modèle plus complexe que celui de l'imputation

Choisir un modèle qui ne peut pas complètement être capturé par le modèle d'imputation, comme une régression qui prend en compte le carré d'une des variables. Intéressant pour voir si l'imputation par la moyenne n'est pas meilleure quand les prérequis des modèles d'imputation sont violés.

2 Ajout de données manquantes

2.1 MCAR

Enlever des observations complètement au hasard, sur une variable ou sur un ensemble de variables.

2.2 MAR

Enlever des observations sur une variable en fonction des autres variables.

Mécanismes de données manquantes MAR

3 Méthodes d'imputation

4 Validation/Apprentissage

5 Résultats