

Données manquantes, modèle gaussien, estimation au maximum de vraisemblance, imputation multiple

Instructions

Ce TP est volontairement moins guidé que les précédents. Il s'agit ici de mettre ensemble tout ce que nous avons vu sur la gestion de données manquantes et l'algorithme EM.

Bonne chance !

Considérons le jeu de données suivant :

x	0.04	1.4	0.15	NA	NA	NA	0.49	-0.55
y	-1.15	3.7	0.73	1.02	-2.05	-1.74	NA	NA

On suppose qu'il peut être modélisé par deux variables aléatoires X et Y qui suivent une loi normale bivariée $(X, Y)^\top \sim \mathcal{N}(0, \Sigma)$.

On souhaite répondre à deux problématiques :

- approximer l'estimateur du maximum de vraisemblance de Σ ,
- générer des valeurs plausibles pour :
 - les données manquantes, en respectant le principe d'imputation propre,
 - le paramètre Σ .

On pourra noter $\mathcal{D}_n = ((x_i)_{1,2,3,7,8}, (y_i)_{1,2,3,4,5,6})$ le jeu de donnée observé.