Examen terminal de Statistique SD 2 - Données manquantes

Annale examen terminal 2024-2025

Consignes

Ce sujet de 7 pages comporte deux exercices. Vous avez 3 heures pour traîter l'ensemble du sujet. Les notes de cours sont interdites. Tout *medium* informatique est interdit. Les figures et tables sont renvoyées en Annexe. Bonne chance!

Exercice 1

On note, $\forall i \in [\![1, n]\!]$, la variable aléatoire $(X_i, Y_i, \delta_i, T_i)$ telle que

$$[T_i|X_i = x_i] \sim \mathcal{N}(a + bx_i, 1), \qquad [Y_i|T_i = t_i] = \min(s_i, t_i) \qquad \text{et} \qquad [\delta_i|Y_i = y_i] = 1(y_i = s_i), \quad (1)$$

où $1(\cdot)$ est la fonction indicatrice. On note $\theta = (a, b)$ le paramètre du modèle, inconnu, et $(s_i)_{i\in[\![1,n]\!]}$ des paramètres réels connus. On observe alors un échantillon iid $\mathcal{D}_n = \{(X_i = x_i, Y_i = y_i, \delta_i = d_i)\}_{i\in[\![1,n]\!]}$ de taille n où chaque observation suit le modèle (1). De plus on suppose que les x_i ne sont pas tous égaux.

1. Comment interpréter chaque paramètre s_i ? Pourquoi est-il introduit dans le modèle ? Donner alors une interprétation de la variable δ_i.

Etude du maximum de vraisemblance

2. Soit un paramètre $\theta \in \mathbb{R}^2$, expliciter la log-vraisemblance $\ell(\theta; \mathcal{D}_n)$ de ce modèle après observation de $\mathcal{D}_n$. On note φ et Φ les fonctions de densité et de répartition de la loi normale centrée réduite.

3. En déduire que les deux équations que doivent résoudre l'estimateur du maximum de vraisemblance de θ, noté $\hat{\theta} = (\hat{a}, \hat{b})$, sont :

$$\begin{cases} \hat{a} &= \overline{y}^{(0)} - \hat{b}\overline{x}^{(0)} + \dfrac{n_1}{n_0}\overline{\gamma(\hat{\theta})}^{(1)}, \\ \hat{b} &= \dfrac{c_0(x,y)}{c_0(x,x)} + \dfrac{n_1}{n_0}\dfrac{c_1(x,\gamma(\hat{\theta}))}{c_0(x,x)}, \end{cases}$$

où et, pour $k = 0, 1$, $n_k = \sum_{i:d_i=k}^{n} 1$, $\gamma_i(\hat{\theta}) = \dfrac{\varphi\left(s_i - \hat{a} - \hat{b}x_i\right)}{1 - \Phi\left(s_i - \hat{a} - \hat{b}x_i\right)}$, $\gamma(\hat{\theta}) = (\gamma_1, \cdots, \gamma_n)(\hat{\theta})$ et, l'opérateur moyenne empirique de $\mathbb{R}^n$ dans $\mathbb{R}$: $x \mapsto \overline{x}^{(k)} = \frac{1}{n_k}\sum_{i:d_i=k} x_i$. Aussi c_k est une fonction de $\mathbb{R}^n \times \mathbb{R}^n$ dans $\mathbb{R}$ telle que $c_k : (x, y) \mapsto \overline{xy}^{(k)} - \overline{x}^{(k)}\overline{y}^{(k)}$.

La fonction $\Gamma : t \mapsto \varphi(t)(1 - \Phi(t))^{-1}$, que l'n nomera dans la suite *fonction de saturation* apparaît ici dans la définition de la variable $\gamma_i(\theta)$, son allure est représentée sur la Figure 1, visible en Annexe.

4. La solution est-elle explicite ? Interpréter la forme obtenue du système en comparant avec le cas sans saturation.

Mise en place d'un algorithme EM

5. Afin de résoudre le système précédent, on se propose de mettre en place un algorithme EM. Quel estimateur est approximé grâce à l'algorithme EM ?
6. On note que les données manquantes sont ici les T_i présentés dans le modèle statistique pour lesquels $\delta_i = 1$. Expliciter la loi de $[T_i|X_i = x_i, \delta_i = 1]$.
7. En déduire une forme de l'espérance conditionelle suivante pour un certain paramètre $\theta_k = (a_k, b_k)$:
$$\mathbb{E}_{\theta_k}[T_i|X_i = x_i, \delta_i = 1]\,,$$
que l'on notera dans la suite $\langle t_i \rangle_k$.
8. Interpréter la forme obtenue en vous aidant de l'allure da la fonction de saturation, Figure 1.
9. Pour un paramètre θ, écrire la log-vraisemblance du jeu de données $\mathcal{D}_n$ complété des $\mathbf{t}_k = (\langle t_i \rangle_k)_{\delta_i=1}$, notée $\ell_c(\theta|\mathcal{D}_n, \mathbf{t}_k)$. On pourra utiliser le changement de variable z_i qui vérifie :
$$z_{i,k} = \begin{cases} y_i & \text{si } d_i = 0, \\ \langle t_i \rangle_k & \text{sinon,} \end{cases}$$
et aussi $\overline{z_k} = 1/n \sum_{i=1}^n z_{i,k}$. Donner ensuite les équations du paramètre $\theta_{k+1} = (a_{k+1}, b_{k+1})$ qui maximise cette log-vraisemblance complétée. A-t-on une solution explicite ? Est-elle unique ?
10. Proposer un algorithme EM pour estimer le paramètre θ du modèle (1). Justifier la convergence de cet algorithme.

Fin de l'exercice 1

Exercice 2

Une étoile peut être assimilée, à grande distance, à un point lumineux émettant des photons dans toutes les directions avec la même probabilité. Dans le vide de l'espace, l'énergie à distance fixée, est constante. Ainsi la quantité de photons (variable quantitative discrète) reçue par le capteur, notée N_i, est liée à l'inverse du carré de la distance D_i de cette étoile à nous. Nous n'observons pas N_i mais un courant relevé dans le capteur (variable quantitative continue) que nous noterons T_i et qui vérifie :

$$T_i = a + \frac{b}{D_i^2} + \varepsilon_i \quad \text{où} \quad \varepsilon_i \sim \mathcal{N}(0,1), \tag{2}$$

où a et b sont des paramètres associés au capteur et ε_i est le bruit dû aux perturbations atmosphériques lors de la traversée de l'atmosphère par les photons.

Un ingénieur vient de recevoir deux capteurs photosensibles (**A** et **B**) pour observer les étoiles. Ces deux capteurs mesurent la luminosité des étoiles, mais ils sont sujets à des saturations lorsque la quantité de photons reçue est trop importante.

Les deux capteurs sont montés sur deux télescopes différents et sont pointés en direction d'étoiles de référence assez proches. Plus précisemment, on choisit les 19 étoiles les plus proches de nous sur lesquels on répette aléatoirement des visées avec l'un ou l'autre des télescopes jusqu'à obtenir 100 mesures pour chaque capteur (et donc chaque télescope). Pour chaque visée on obtient donc une valeur de courant dans le capteur, la distance théorique de l'étoile considérée, le seuil de saturation du capteur considéré et si le capteur a saturé ou pas. Certains résultats sont saturés et d'autres ne le sont pas et l'on connaît aussi les seuils de saturation. On note $\mathcal{D}_n$ l'ensemble des données recueillies.

> **Attention**
>
> Dans la plupart des figures, rapportées en Annexe, l'abscisse fait référence à la distance de l'étoile à nous mais l'échelle est en inverse de cette distance mise au carré.

1. Après avoir rappelé la définition mathématique des 3 processus de pertes de données manquantes, expliquer pourquoi le modèle (2) est un modèle de données manquantes de type MNAR.

Estimation de modèle

2. En s'insiprant du modèle (1), proposer un modèle de mesure pour les capteurs **A** et **B**. Expliciter les paramètres s_i.
3. L'ingénieur propose d'estimer les paramètres a et b par maximum de vraisemblance. Il obtient les résultats présentés sur la Figure 2. Que penser du modèle estimé ?

Il semble y avoir un décalage entre les deux capteurs, l'ingénieur pense que c'est dû à des différences dans les télescopes. Il décide donc de modéliser ce décalage.

4. La solution imaginée consiste à supposer un effet fixe sur l'ordonnée à l'origine pour chaque capteur, notés a_A et a_B pour le capteur A et le capteur B respectivement. Expliciter clairement le modèle complet des deux capteurs.

Dans la suite on notera le paramètre $\theta = (a_A, a_B, b)$.

5. On suppose construit l'algorithme EM pour estimer le paramètre θ au maximum de vraisemblance. On obtient les estimations projetées sur la Figure 3. Que penser de ce modèle contrairement au modèle estimé précédemment et visible sur la Figure 2 ?

Dans la suite l'ingénieur cherche à générer des paramètres vraisemblables en se basant sur sa propre expériernce. Pour celà il utilise un formalisme bayésien.

6. On suppose dans la suite un *a priori* indépendant sur le paramètre θ sous la forme $p(\theta) = p(a_A)p(a_B)p(b)$, tel que
$$a_A \sim \mathcal{N}(\alpha_A, \sigma_A^2), \quad a_B \sim \mathcal{N}(\alpha_B, \sigma_B^2) \quad \text{et} \quad b \sim \mathcal{N}(\alpha, \sigma^2),$$
où $(\alpha_A, \sigma_A^2, \alpha_B, \sigma_B^2, \alpha, \sigma^2)$ est un hyper-paramètre connu basé sur l'expérience de l'ingénieur. La vraisemblance du modèle considéré, notée ici $\mathcal{L}(\theta)$ est proportionelle à $p(\mathcal{D}_n|\theta)$. Donner, à une constante multiplicative près, la loi *a posteriori* du paramètre θ : $p(\theta|\mathcal{D}_n)$.

Imputations

7. Il est possible, grâce aux résultats de l'algorithme EM, de construire des valeurs vraisemblables pour les données manquantes. Rappeler de quelle étape de l'algorithme EM il s'agit et de quelle grandeur. On obtient un jeu de données complété et tracé sur la Figure 4. Que penser de ce jeu de données ? Vous décrirez avec soin la position des données imputées.

On introduit une nouvelle variable z_i qui vient compléter le jeu de données. L'ensemble accessible aux données manquantes $Z = (z_i)_{i:\delta_i=1}$ est noté $\mathcal{Z}$. On a donc accès à un jeu de données complété.

8. La densité $p(\mathcal{D}_n, Z|\theta)$ est proportionelle à la vraisemblance complétée, notée $\mathcal{L}_c(\theta)$. Donner son expression.

9. En s'inspirant du travail effectué sur la loi *a posteriori*, donner les lois conditionelles du paramètre θ sachant le jeu de données complété $(\mathcal{D}_n, Z)$: $[a_A|a_B, b, \mathcal{D}_n, Z]$, $[a_A|a_B, b, \mathcal{D}_n, Z]$ et $[b|a_A, a_A, \mathcal{D}_n, Z]$, pour l'*a priori* déjà détaillé plus haut :

10. Donner la forme de la loi conditionelle de z_i pour les valeurs saturées sachant les données $\mathcal{D}_n$ et un paramètre courant θ.

11. En supposant que les conditions de convergence d'un tel algorithme soient réunies, proposer un algorithme de Gibbs qui permette de simuler la loi jointe $[Z, \theta|\mathcal{D}_n] = [Z, a_A, a_B, b|\mathcal{D}_n]$.

Grâce à l'algorithme précédent, l'ingénieur a un échantillon Monte-Carlo de taille $H > 0$ de (Z, θ) qui suit la loi $[Z, \theta|\mathcal{D}_n]$. On note chaque observation de cet échantillon $(Z^{(h)}, a_A^{(h)}, a_B^{(h)}, b^{(h)}) = (Z^{(h)}, \theta^{(h)})$ pour $h = 1, 2, \dots, H$.

12. On définit la loi postérieure des données manquantes sachant les données observées $\mathcal{D}_n$ par la relation de marginalisation suivante :
$$p(Z|\mathcal{D}_n) = \int_{\mathbb{R}^3} p(Z, \theta|\mathcal{D}_n)d\theta = \int_{\mathbb{R}^3} p(Z|\theta, \mathcal{D}_n)p(\theta|\mathcal{D}_n)d\theta, .$$
Soit i tel que $\delta_i = 1$, proposer une approximation Monte-Carlo de la loi marginale de z_i sachant les données, on notera cette fonction $z \mapsto \hat{p}_i(z|\mathcal{D}_n)$.

13. En reprenant la Figure 4, l'ingénieur se rend compte que seulement 5 astres font saturer les capteurs (5 pour le capteur A et 4 pour le capteur B). Dans l'ordre de proximité avec nous, il s'agit des astres répertoriés dans la Table 1. Sur la Figure 5, sont représentées les 9 densités prédictives postérieures pour $z \in [30, 55]$. Aidez l'ingénieur à associer chacune des courbes à chacun des couples (astre,capteur) en saturation.

14. La Figure 6 représente 6 jeux de données imputés. S'agit-il d'imputation simple ou multiple ? Propre ou impropre ? Justifier.

Fin de l'exercice 2

Annexes

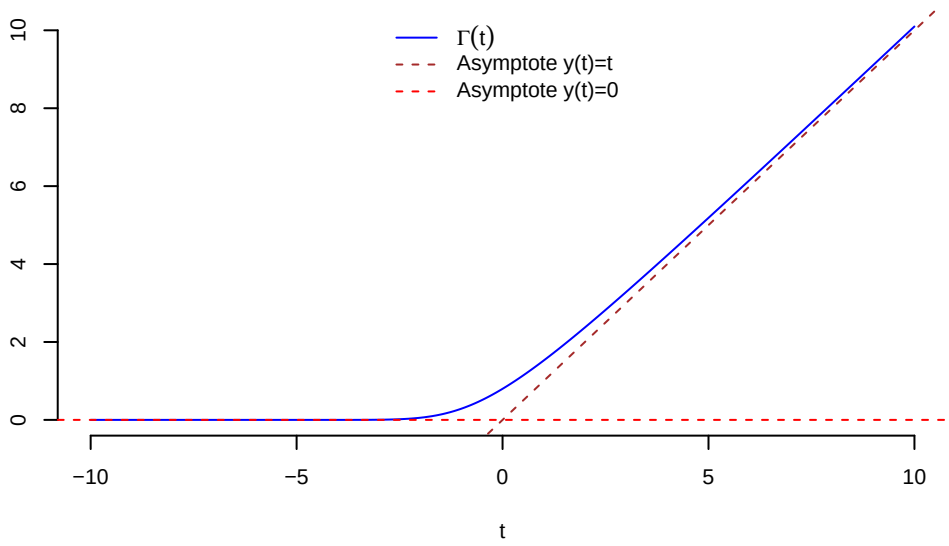

Figure 1: Allure de la fonction de saturation

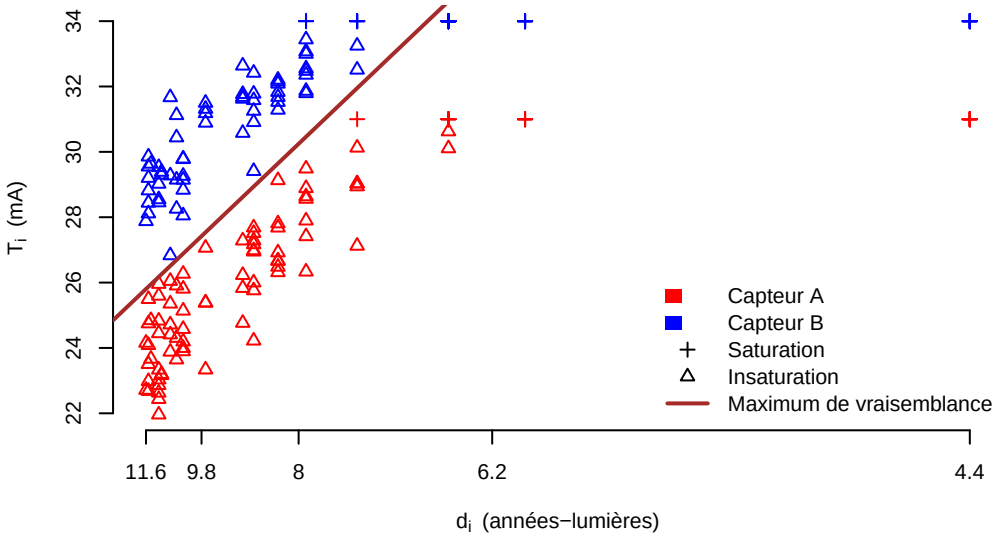

Figure 2: Estimation du modèle initial

Table 1: Description des astres dont la mesure est saturée

Nom	Distance	Capteur(s) saturé(s)
Alpha Centauri A	4.4	A et B
Barnard star	6.0	A et B
Luhman 16 A	6.5	A et B
WISE 0855-0714	7.3	A et B
Wolf 359	7.9	B

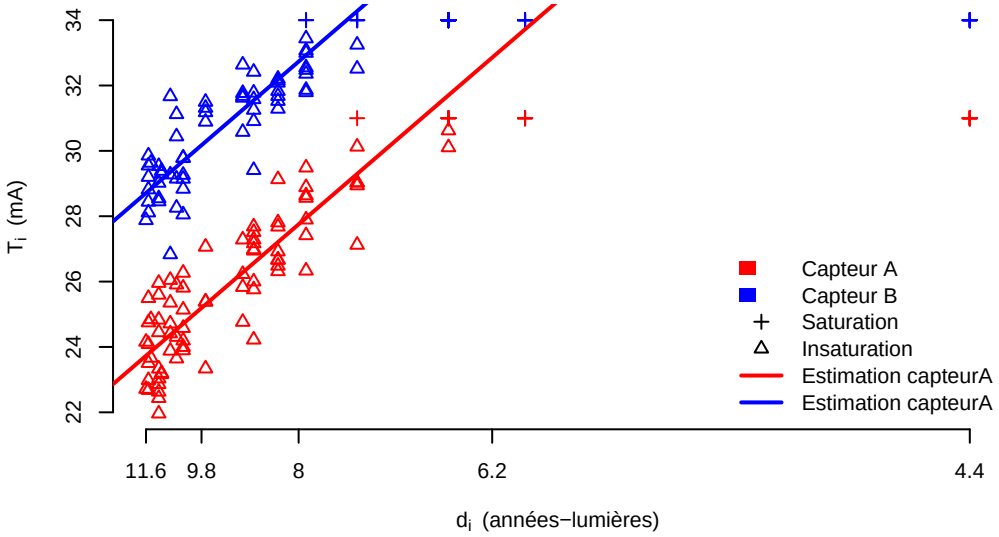

Figure 3: Estimation du modèle à effets fixes

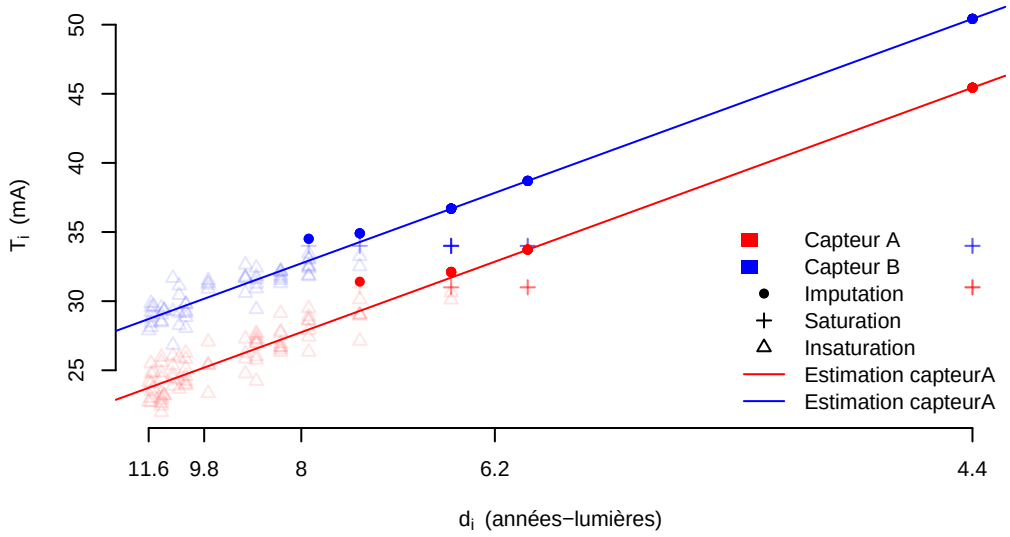

Figure 4: Premier jeu de données imputé

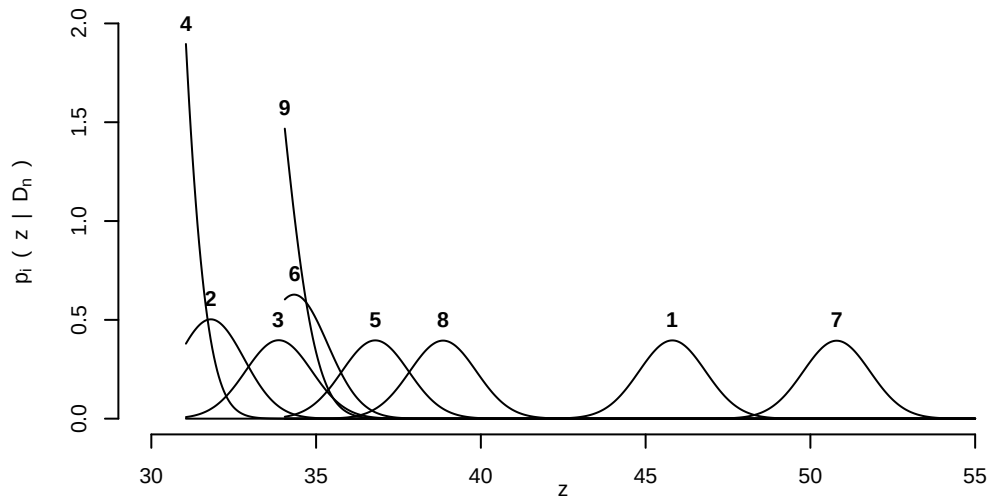

Figure 5: Densités prédictives postérieures pour les astres saturés

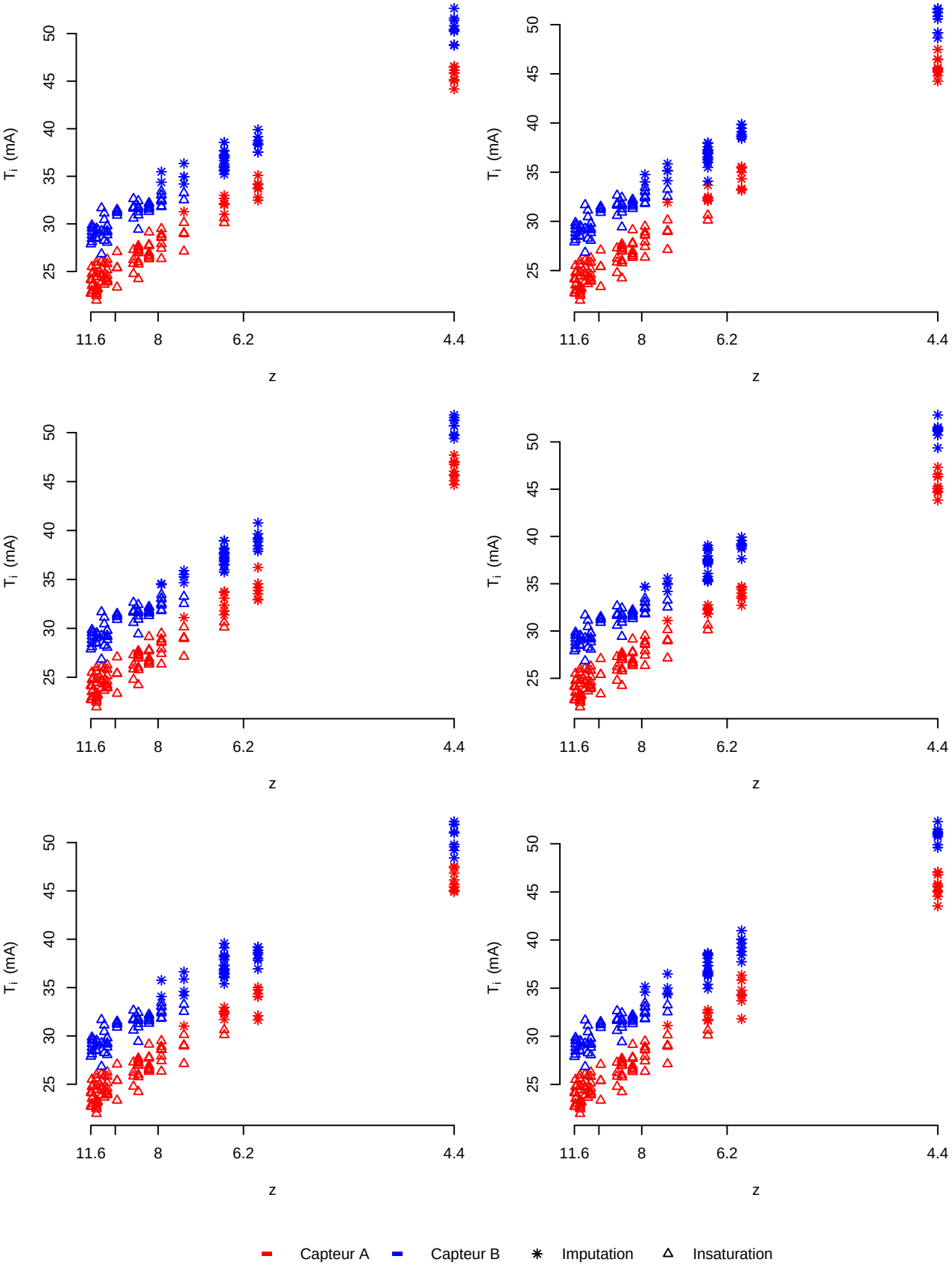

Figure 6: Exemple de jeux de données imputés