**Université d'Aix-Marseille** **Année 2024-2025**
**Master 1 Mathématiques Appliquées, Statistique**
**Cours de classification, fiche 1**

---

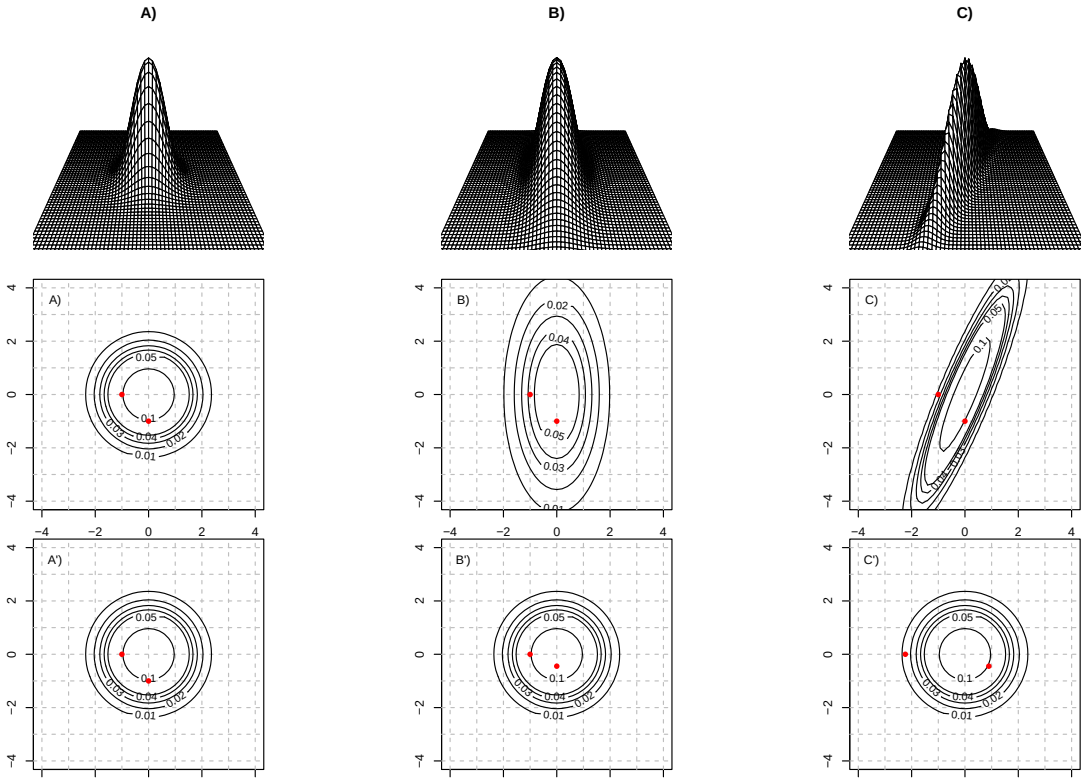
Objectif : comprendre la signification des distances usuelles.

---

Pour les programmes `R` que vous générerez, créer un dossier au nom de `TP-CLASSIF`. Le dossier `TP-CLASSIF` contiendra les dossiers des différents TP de l'UE Classification. Créer dans `TP-CLASSIF` un nouveau dossier au nom de `TPdistances`.

**Exercice 1** La figure 1 présente pour trois distributions bivariées deux points indiqués en rouge dans les sous-figures A), B) et C). Chaque distribution est une loi normale bivariée d'espérance le point $(0, 0)$ et de matrice de variance-covariance valant respectivement :

A) $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ B) $\begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}$ C) $\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$




1. Que valent les coordonnées des points indiqués en rouge pour chaque cas ? En déduire la distance euclidienne entre les points indiqués en rouge puis entre chaque point indiqué en rouge et le point de coordonnées (0,0).

2. Calculer maintenant, pour chaque cas, la distance de Mahalanobis entre les entre les points indiqués en rouge puis entre chaque point indiqué en rouge et le point de coordonnées (0,0).
   On donne $\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}^{-1} = \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}$

3. Que représentent les sous-figures A'), B') et C') ? En quoi ces figures sont-elles cohérentes avec les résultats trouvés à la question précédente ?

**Exercice 2** 1. On considère deux points de $\mathbb{R}^2$. Laquelle des distances entre ces deux points est la plus élevée : la distance euclidienne ou la distance de Manhattan ?

2. On considère le point A de coordonnée $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$. Calculer les distances euclidienne et de Manhattan entre le point A et le point de coordonnées $(0, 0)$.

3. En déduire un point B qui serait situé à une distance de Manhattan égale à 1 du point de coordonnées $(0, 0)$.

4. Représenter sur un plan les points situés à une distance de 1 du point $(0, 0)$ pour la distance de Manhattan et pour la distance euclidienne.

**Exercice 3** 10 étudiants ont été interrogés sur les moyens de transport qu'ils utilisent pour leurs déplacements quotidiens. On cherche à mesurer la proximité entre les réponses obtenues. Pour cela on note 1 lorsque le moyen de transport est cité dans la réponse et 0 s'il n'est pas cité. Les résultats sont présentés dans le tableau 1[1].

| | Voiture | Moto | Bus | Train | Vélo | A pied |
|---|---|---|---|---|---|---|
| Lucas | 1 | 1 | 1 | 0 | 0 | 0 |
| Clara | 1 | 1 | 1 | 0 | 0 | 1 |
| Lola | 1 | 1 | 0 | 1 | 0 | 0 |
| Matéo | 0 | 0 | 1 | 1 | 1 | 1 |
| Leni | 0 | 1 | 1 | 1 | 0 | 0 |
| Raphaël | 0 | 1 | 1 | 0 | 0 | 1 |
| Anna | 1 | 1 | 1 | 1 | 0 | 0 |
| Mathilde | 1 | 1 | 0 | 0 | 1 | 1 |
| Jean | 1 | 1 | 0 | 0 | 1 | 1 |
| Claire | 1 | 0 | 0 | 0 | 1 | 0 |

Table 1: Réponses de dix étudiants sur leur utilisation des moyens de transports.

Pour un couple $(i, j)$ d'étudiants on note :

$a$ le nombre de caractéristiques communes possédées par $i$ et par $j$
$b$ le nombre de caractéristiques possédées par $i$ mais pas par $j$
$c$ le nombre de caractéristiques possédées par $j$ mais pas par $i$
$d$ le nombre de caractéristique possédées ni par $i$ ni par $j$

et les indices de similarité classiques :

- Jaccard : $J(i, j) = a/(a + b + c)$
- Dice (ou Sorensen) : $D(i, j) = 2a/(2a + b + c)$
- Simple matching : $S(i, j) = (a + d)/(a + b + c + d)$
- Russel et Rao : $R(i, j) = a/(a + b + c + d)$

1. Calculer les indices de similarité précédents entre les réponses de Lucas et Clara et entre celles de Lucas et Lola. En déduire les indices de dissimilarités correspondants puis les distances associées.

2. En utilisant la fonction `dist.binary()` du package `ade4` de `R`, calculer les matrices de similarité au sens de Jaccard, Dice, Russel et Rao, de simple matching entre les dix réponses.

[1] Sous `R` vous pouvez charger l'objet `transports` en exécutant

```
> load("transports.rds")
```

le fichier `transports.rds` étant disponible sur AMeTICE.

NB : On pourra également utiliser la fonction `dist()` avec l'option `method="binary"` pour l'indice de Jaccard et avec l'option `method="manhattan"` renormalisée par la longueur du vecteur de réponse pour l'indice du simple matching, ou la fonction `vegdist()` du package `vegan` pour les indices de dissimilarité de Jaccard, Dice et du simple matching.

3. Exprimer dans le cas de données binaires la distance de Manhattan en fonction de $a, b, c, d$. Quel est le lien entre la similarité simple matching et le distance de Manhattan ? Dans le cas de données binaires, quel est le lien entre la distance euclidienne et la distance de Manhattan ?

**Exercice 4** La table 2[2] est issue du site[3] education.gouv.fr du ministère de l'éducation nationale, de la jeunesse et des sports. Elle présente le nombre d'étudiant.e.s inscrits à l'université en France pour l'année 2020-2021 suivant leurs origines sociales et les filières suivies.

| | AgrArtCommChEnt | CadPrIntel | Pinter | Employés | Ouvriers | Retraités-inactifs |
|---|---|---|---|---|---|---|
| Droit sciences politiques | 19041 | 68475 | 25592 | 32148 | 17771 | 27960 |
| Economie, AES | 21279 | 54497 | 27658 | 37731 | 25970 | 28790 |
| Arts, lettres, langues, SHS | 35988 | 122705 | 72565 | 91446 | 53805 | 85196 |
| Sciences | 26757 | 103009 | 49844 | 50914 | 33089 | 36909 |
| Staps | 5252 | 18114 | 11388 | 12674 | 7126 | 5328 |
| Santé | 18081 | 96241 | 26260 | 21964 | 12470 | 24439 |

Table 2: Nombre d'occurrences des filières universitaires suivies suivant l'origine sociale des étudiants. L'origine `AgArtCommChEnt` signifie Agriculteurs, artisans, commerçants et chefs d'entreprise, l'origine `CadPrIntel` signifie Cadres et professions intellectuelles supérieures, l'origine `Pinter` signifie Professions Intermédiaires.

1. La table 2 est une table de contingence. Donner une définition générale d'une table de contingence.
2. Ordonner les filières suivant le nombre d'étudiants qu'elles accueillent.
3. A l'aide de la fonction `dist()` calculer les distances euclidiennes entre chaque paire de filières. Donner tous les couples pour lesquelles la distance est supérieure à 100000. Quelle est la valeur de la distance entre les deux filières accueillant le plus d'étudiants ? Quelle conclusion peut-on en tirer ?
4. Exécuter le code suivant :

```
>(profils<-etudiants/rowSums(etudiants))
> distchi2<-function(ni1,ni2){
ni.<-ni1+ni2
d<-sum(((ni1/sum(ni1))-(ni2/sum(ni2)))^2/(ni./sum(ni.)))
return(d)
}
>distchi2(etudiants["Arts, lettres, langues, SHS",],etudiants["Staps",])
>distchi2(etudiants["Arts, lettres, langues, SHS",],etudiants["Santé",])
```

   puis décrire les sorties obtenues.

5. Pourquoi la distance du chi2 est plus appropriée que la distance euclidienne pour l'étude d'individus décrits à l'aide d'un tableau de contingence ?

[2]Sous `R` vous pouvez charger l'objet `etudiants` à partir du fichier `transports.rds` disponible sur AMeTICE.
[3]`https://www.education.gouv.fr/reperes-et-references-statistiques-2021-308228`

**Exercice 5** La table 3[4] présente pour cinq sites la température à la surface d'un champ et la température du sol mesurée à 2 cm sous la surface.

| site | température de surface | température du sol |
|---|---|---|
| A | -4 | -1 |
| B | 0 | 1 |
| C | 1 | 0 |
| D | -5 | -2 |
| E | 2 | 2 |

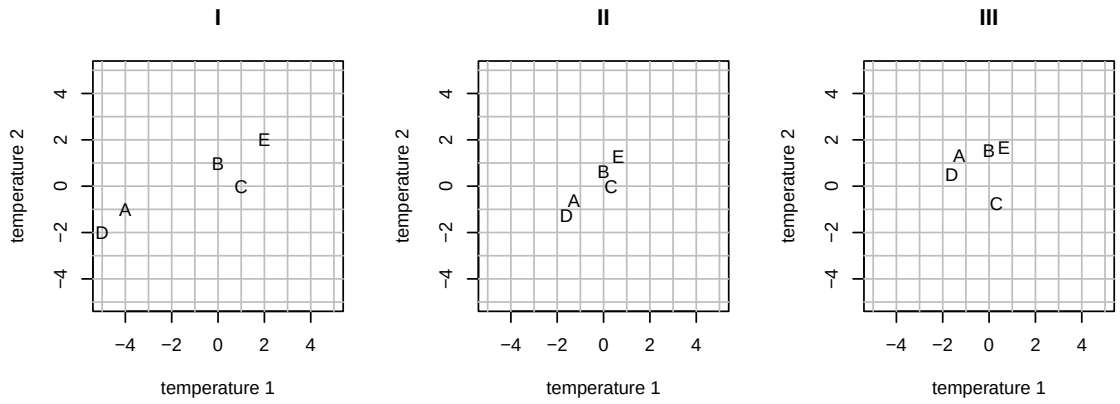
Table 3: Température de surface et du sol pour six sites notés A à E.

1. Représenter graphiquement la température du sol en fonction de la température de surface. En déduire quel est le point le plus proche de du point B suivant la distance euclidienne.
   Que vaut d(B,C)/d(B,E) ?

2. Les variances empiriques des deux températures valent respectivement 9.7 et 2.5. En déduire les distances pondérées par l'inverse de la variance entre les points B, C et E. Quelle remarque peut-on faire ?

3. Calculer maintenant la distance de Mahalanobis entre les points B et C.
   On donne $V = \begin{pmatrix} 9.7 & 4.5 \\ 4.5 & 2.5 \end{pmatrix}$ et $V^{-1} = \begin{pmatrix} 0.625 & -1.125 \\ -1.125 & 2.425 \end{pmatrix}$.

4. Que représente chacune des sous-figures de la figure ci-dessous ? Proposer une interprétation de ces résultats.




[4]Sous `R` vous pouvez charger l'objet `temperatures` à partir du fichier `temperatures.rds` disponible sur AMeTICE.