

**Question 1.** Quelle sont les coordonnées de la première composante principale des données décrites sur la figure 1 ?

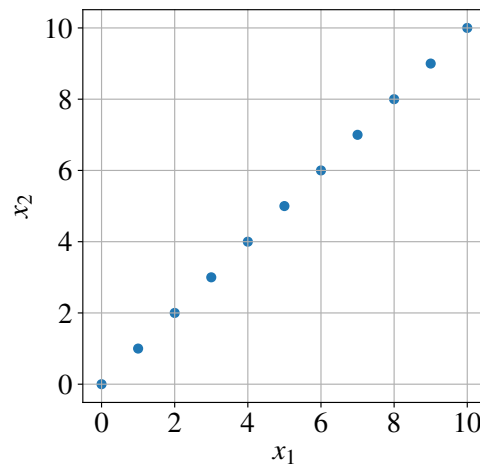


FIGURE 1 – 11 individus représentés par 2 variables  $x_1$  et  $x_2$ .

- ☐ (1,1)
- ☐  $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$
- ☐ (1,0)
- ☐  $(\sqrt{2}, 0)$

**Question 2.** Parmi les affirmations ci-dessous, lesquelles sont vraies ? On considère un jeu de données  $X \in \mathbb{R}^{n \times p}$  de  $n$  individus en  $p$  dimensions.

- ☐ La réduction de dimension relève de l'apprentissage supervisé.
- ☐ La réduction de dimension relève de l'apprentissage non-supervisé.
- ☐ La réduction de dimension facilite la visualisation des données.
- ☐ L'analyse en composantes principales de  $X$  permet de créer jusqu'à  $n$  nouvelles dimensions.
- ☐ Les nouvelles variables créées par une analyse en composantes principales sont des combinaisons linéaires des  $p$  variables.
- ☐ L'analyse en composantes principales de  $X$  s'obtient par une décomposition spectrale de  $X$ .
- ☐ La sélection de variables consiste à conserver uniquement les variables dont la variance est la plus faible.

## Solution

**Question 1.** La direction de plus grande variation des données est la diagonale d'équation  $x_1 = x_2$ . Ainsi, la première composante principale est le vecteur directeur de la diagonale, de norme 1, soit donc  $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$ .

### Question 2.

- La réduction de dimension peut relever de l'apprentissage supervisé (par exemple, l'élimination des variables indépendantes de l'étiquette requièrent évidemment une étiquette) ou de l'apprentissage non-supervisé (par exemple, l'ACP). Elle est cependant souvent plutôt classée dans l'apprentissage non-supervisé car il s'agit d'analyse exploratoire des données et non pas d'analyse prédictive, ce qui peut prêter à confusion.
- ☒ La réduction de dimension facilite la visualisation des données.
- ☐ L'analyse en composantes principales de  $X$  permet de créer jusqu'à  $n$  nouvelles dimensions.  
FAUX, elle permet de créer jusqu'à  $p$  nouvelles dimensions.
- ☒ Les nouvelles variables créées par une analyse en composantes principales sont des combinaisons linéaires des  $p$  variables.
- ☐ L'analyse en composantes principales de  $X$  s'obtient par une décomposition spectrale de  $X$ .  
FAUX, il s'agit de la décomposition spectrale de  $X^\top X$ .
- ☐ La sélection de variables consiste à conserver uniquement les variables dont la variance est la plus faible.  
FAUX, *une des techniques* de sélection de variables consiste à *éliminer* les variables dont la variance est la plus faible.