**Consigne du Dr**

Concernant l'examen, vous pouvez lire et reprendre simplement les exemples du cours. Il n'aura pas de calcul. Juste des vrais ou faux puis des exercices sur le choix des méthodes statistiques appropriées. Les documents ne sont pas autorisés

1. **Qu’est-ce que l’Analyse en Composantes Principales (ACP) ?**

L’ACP est une méthode statistique utilisée pour explorer et simplifier des ensembles de données complexes et multidimensionnels. Elle permet de réduire la dimensionnalité d’un jeu de données tout en conservant le maximum d’information possible.

1. **Pourquoi utilise-t-on l’ACP dans l’analyse de données ?**

On utilise l’ACP pour identifier des structures sous-jacentes dans les données, réduire le nombre de variables tout en minimisant la perte d’information, et visualiser des tendances ou des regroupements dans un espace de plus faible dimension.

1. **Que représente une composante principale ?**

Une composante principale est une nouvelle variable obtenue comme une combinaison linéaire des variables initiales, choisie pour maximiser la variance des données. La première composante principale capture la plus grande part de la variance totale.

1. **Quelle est la différence entre la première et la deuxième composante principale ?**

La première composante principale (CP1) capture la plus grande part de la variance totale, tandis que la deuxième composante principale (CP2) est orthogonale à la première et capture la plus grande variance restante.

1. **Que signifie une valeur propre élevée en ACP ?**

Une valeur propre élevée signifie que la composante principale associée contient beaucoup d’information, c’est-à-dire qu’elle explique une grande part de la variance totale des données.

1. **Comment choisir le nombre de composantes principales à retenir ?**

On utilise souvent un scree plot (diagramme de l’éboulis) pour visualiser la décroissance de la variance expliquée et identifier le « point de coude ». On conserve généralement les composantes qui expliquent au moins 70-80% de la variance totale.

1. **Qu’est-ce que le cercle des corrélations en ACP ?**

Le cercle des corrélations est une représentation graphique qui permet d’analyser comment les variables initiales sont liées aux composantes principales. Les variables proches du bord du cercle sont bien représentées par les composantes principales.

1. **Que signifie la proximité de deux variables sur le cercle des corrélations ?**

Si deux variables sont proches et alignées sur le cercle des corrélations, elles sont corrélées positivement. Si elles sont opposées, elles sont corrélées négativement.

1. **Comment sont représentés les individus en ACP ?**

Les individus sont visualisés dans un nouvel espace réduit (souvent en 2D ou 3D). Les individus proches sur le graphique ont des profils similaires, ce qui permet d’identifier des groupements ou des structures cachées dans les données.

1. **Dans quels cas utilise-t-on la matrice de covariance ou la matrice de corrélation pour l’ACP ?**

On utilise la matrice de covariance si les variables sont sur des échelles comparables, et la matrice de corrélation si les variables sont sur des échelles différentes, car elle standardise les variables.

1. **Quelles sont les étapes principales de l’ACP ?**

Les étapes principales sont :

* Standardisation des données (si nécessaire)
* Calcul de la matrice de covariance ou de corrélation
* Détermination des valeurs propres et des vecteurs propres
* Interprétation des composantes principales
* Visualisation des résultats à l’aide de graphiques (cercle des corrélations, plan factoriel)

1. **Comment doit-on préparer les données avant d’appliquer l’ACP ?**

Il faut vérifier que les variables sont quantitatives continues, et si elles sont sur des échelles différentes, il faut centrer et réduire les données (standardisation) pour éviter qu’une variable domine l’analyse.

1. **Comment sont organisées les données dans un tableau pour l’ACP ?**

Les individus ou observations sont placés en lignes et les variables en colonnes. Chaque ligne correspond à un individu, chaque colonne à une variable quantitative.

|  |  |
| --- | --- |
| Résumé des propriétés Propriété | Explication |
| Orthogonalité | Les CPs sont perpendiculaires et indépendantes. |
| Maximisation de la variance | Chaque CP capte le maximum de variance restante. |
| Somme de la variance conservée | Pas de perte d'information, seule la distribution de variance change. |
| Combinaisons linéaires des variables initiales | Chaque CP est °définie par des poids appliqués aux variables. |
| Somme des carrés des coefficients = 1 | Normalisation des vecteurs propres pour une interprétation standardisée. |
| Interprétation géométrique | Changement de base optimisé pour maximiser l'information |

|  |  |
| --- | --- |
| Quel choix pour une ACP ?Type de données | Distance Recommandée |
| Quantitatives, normalisées | Euclidienne |
| Corrélées, avec variances très différentes | Mahalanobis |
| Qualitatives (catégoriques) | Khi-Deux |
| Données en grande dimension (ex. texte) | Cosinus |

**Interprétation des coordonnées des variables :**

* Plus une variable est proche du bord du cercle des corrélations, mieux elle est représentée par les axes factoriels.
* Une variable située près d’un axe contribue fortement à cet axe.
* Des variables proches l’une de l’autre sur le cercle sont corrélées positivement.
* Des variables placées à l’opposé (environ 180°) sont corrélées négativement.
* Des variables perpendiculaires sont non corrélées.

**Règles d’interprétation :**

* Un cos² élevé (proche de 1) indique que la variable est bien représentée par l’axe.
* Un cos² faible (proche de 0) signifie que la variable est peu expliquée par l’axe et pourrait être mieux représentée par un autre.

Des valeurs propres élevées signifient que l’axe explique une part importante de l’information.

**Applications et cas d’utilisation :**

* En agriculture et biologie : étude des caractéristiques des variétés végétales
* En marketing : segmentation des clients selon leurs comportements d’achat
* Médecine : réduction de la dimensionnalité dans des bases de données médicales

**Qualités et défauts de l’ACP normée :**

|  |  |
| --- | --- |
| Avantages | Inconvénients |
| Permet de comparer des variables de nature différente | Perte d’information sur les variances initiales |
| Évite qu’une variable domine l’analyse | Résultats parfois moins interprétables |
| Applicable à des variables fortement corrélées | Ne fonctionne pas bien avec des données discrètes ou qualitatives |
| Facilite l’interprétation graphique (cercle des corrélations, biplots…) | Hypothèse de linéarité entre les variables (peu adapté aux relations complexes) |

**Partie A**

ACP – Analyse en Composantes Principales

1. L’ACP est applicable uniquement aux variables qualitatives. Faux. L’ACP s’applique uniquement aux variables quantitatives, nécessite des variables quantitatives continues pour calculer les distances et corrélations.
2. L’ACP permet de réduire la dimension tout en conservant une grande partie de l’inertie. Vrai.
3. Les axes de l’ACP sont orthogonaux. Vrai.
4. L’ACP nécessite toujours un centrage et une réduction des données. Faux. Le centrage et la réduction sont nécessaires si les variables n’ont pas la même unité ou échelle pour rendre les variables comparables.
5. L’inertie totale correspond à la somme des valeurs propres. Vrai.
6. En ACP, un individu proche de l’origine a un comportement moyen. Vrai.
7. Les variables sont projetées dans le cercle des corrélations. Vrai.
8. Les variables fortement corrélées sont représentées par des vecteurs proches. Vrai.
9. Une valeur propre proche de zéro signifie que la composante correspondante explique beaucoup d'inertie. Faux. Une valeur propre proche de zéro indique que l’axe n’explique presque aucune variance.
10. L’ACP est insensible aux valeurs aberrantes. Faux. L’ACP est sensible aux outliers qui peuvent influencer les axes principaux.

1**. Qu’est-ce que l’Analyse Factorielle des Correspondances (AFC) et à quel type de données s’applique-t-elle ?**

L’AFC est une méthode d’analyse descriptive multidimensionnelle adaptée à l’étude des tableaux de contingence issus de variables qualitatives. Elle permet d’étudier les relations entre les catégories de ces variables.

**2. Que mesure la distance du khi-deux dans l’AFC ?**

La distance du khi-deux mesure l’écart entre la distribution observée des modalités et la distribution attendue si les variables étaient indépendantes. Elle sert à quantifier la proximité ou la dissimilarité entre modalités ou individus.

**3. Comment interpréter la projection des modalités et des individus sur les axes factoriels en AFC ?**

Les modalités ou individus proches sur le plan factoriel ont des profils similaires. Les axes factoriels permettent de visualiser les principales oppositions ou regroupements entre modalités.

**4. Qu’est-ce que l’inertie en AFC et à quoi sert-elle ?**

L’inertie représente la part de l’information (ou de la variance) expliquée par chaque axe factoriel. Elle permet de choisir le nombre d’axes à retenir pour une bonne interprétation.

**5. Quelles sont les principales applications de l’AFC ?**

L’AFC est utilisée dans l’analyse d’enquêtes, le marketing, la sociologie, la biologie, etc., pour explorer les liens entre variables qualitatives (par exemple, entre professions et loisirs, ou entre symptômes et diagnostics).

AFC – Analyse Factorielle des Correspondances

1. L’AFC s’applique à un tableau croisé de fréquences. Vrai.
2. L’AFC suppose des données quantitatives. Faux. Elle s’applique à des données qualitatives résumées sous forme de fréquences.
3. L’AFC maximise la variance totale. Faux. Elle maximise l’inertie (variance pondérée par les effectifs).
4. En AFC, les lignes et colonnes sont projetées dans le même plan factoriel. Vrai.
5. Les profils ligne et colonne doivent être centrés autour de leur moyenne pour l’analyse. Vrai.
6. Une association forte entre deux modalités se traduit par leur proximité dans le plan factoriel. Vrai.
7. En AFC, les distances entre les modalités sont mesurées avec la distance euclidienne simple. Faux. On utilise la distance du Khi-2.
8. Un point proche de l’origine indique qu’il a un comportement moyen. Vrai.
9. L’AFC est sensible aux effectifs faibles dans les cases du tableau. Vrai. Les faibles effectifs peuvent fausser l’analyse.
10. Les points les plus extrêmes dans le plan sont toujours les plus représentatifs. Faux. Ils peuvent être atypiques ou peu représentatifs.

AFCM – Analyse Factorielle des Correspondances Multiples

1. L’AFCM est utilisée pour l’analyse de plusieurs variables qualitatives. Vrai.
2. L’AFCM suppose un encodage en variables numériques centrées réduites. Faux. Elle utilise un codage binaire (tableau disjonctif complet).
3. L’AFCM est une généralisation de l’ACP. Faux. C’est une généralisation de l’AFC.
4. Le total de l’inertie est égal au nombre de variables actives. Faux. Il dépend du nombre de modalités moins le nombre de variables.
5. L’AFCM est insensible aux modalités rares. Faux. Les modalités rares peuvent influencer l’analyse.
6. L’AFCM est utile pour les enquêtes avec des questions à réponses fermées. Vrai.
7. Les modalités très proches dans le plan ont un profil similaire. Vrai.
8. Les coordonnées des modalités sont pondérées par leur fréquence. Vrai.
9. L’AFCM permet d'étudier la structure des individus à partir des variables continues. Faux. Elle s’applique aux variables qualitatives.
10. Le plan factoriel en AFCM est interprété comme celui de l’ACP. Vrai.

CAH – Classification Ascendante Hiérarchique

1. La CAH construit des groupes d’individus à partir d’une matrice de distances. Vrai.
2. Le dendrogramme permet de visualiser l’ordre des fusions. Vrai.
3. En CAH, l’algorithme commence par un seul groupe contenant tous les individus. Faux. Il commence par autant de groupes que d’individus.
4. Le critère de Ward maximise l'inertie intra-classe. Faux. Il minimise l’inertie intra-classe (et maximise l’inertie inter-classe).
5. Le dendrogramme peut être coupé pour définir un nombre de classes. Vrai.
6. La CAH est sensible à l’échelle des variables. Vrai.
7. Les groupes obtenus par CAH sont toujours optimaux. Faux. Ils dépendent de la méthode et peuvent ne pas être optimaux.
8. La CAH peut être combinée à l’ACP pour la visualisation. Vrai.
9. La distance de Manhattan est une des distances possibles pour la CAH. Vrai.
10. L’algorithme de CAH est itératif et non récursif. Vrai.

K-Means – Classification des Moyennes

1. K-Means est une méthode de classification non supervisée. Vrai.
2. Le nombre de groupes doit être défini à l’avance. Vrai.
3. K-Means fonctionne mieux avec des variables qualitatives. Faux. Elle est adaptée aux variables quantitatives.
4. L’algorithme minimise la variance intra-groupe. Vrai.
5. K-Means est sensible aux valeurs extrêmes. Vrai.
6. Le résultat de K-Means dépend des points initiaux choisis. Vrai.
7. K-Means peut produire des clusters non sphériques. Faux. Elle tend à former des clusters sphériques.
8. L’ACP peut être utilisée pour visualiser les résultats de K-Means. Vrai.

ICA – Analyse en Composantes Indépendantes

* L’ICA cherche des composantes orthogonales. Faux. Elle cherche des composantes indépendantes, pas nécessairement orthogonales.
* L’ICA est souvent utilisée pour le traitement du signal. Vrai.

Méthodes multidimensionnelles : concepts avancés

1. En ACP, un axe principal est un vecteur propre de la matrice des corrélations. Vrai.
2. En ACP, les valeurs propres reflètent la corrélation entre variables. Faux. Les valeurs propres représentent la variance expliquée par chaque axe ?
3. En AFC, les masses sont les effectifs totaux des lignes ou des colonnes. Vrai.
4. L’AFC peut être vue comme une ACP sur un tableau transformé. Vrai.
5. En CAH, deux classes peuvent être fusionnées même si leur distance est grande. Faux. La CAH fusionne toujours les classes les plus proches.
6. En K-Means, chaque individu est affecté à plusieurs clusters avec une probabilité. Faux. K-Means est une classification dure : chaque individu appartient à un seul cluster.
7. La CAH produit des classes imbriquées (hiérarchiques). Vrai.
8. L’ACP permet de détecter des groupes d’individus sans faire de classification. Vrai.
9. L’AFCM permet d’inclure des variables quantitatives dans l’analyse. Faux. L’AFCM s’applique aux variables qualitatives uniquement.
10. L’ICA suppose une indépendance des composantes sources. Vrai.

Logiciel R

1. En R, la fonction prcomp() effectue une ACP. Vrai. prcomp() est une fonction standard pour réaliser une ACP en R.
2. En R, la fonction princomp() est recommandée pour les grands jeux de données. Faux. prcomp() est généralement préférée pour les grands jeux de données.
3. La fonction factoextra::fviz\_pca\_ind() permet de visualiser les individus d'une ACP. Vrai.
4. Le package FactoMineR permet de faire de l’ICA en R. Faux. FactoMineR est dédié aux analyses factorielles, pas à l’ICA.
5. En R, cluster::agnes() permet de faire une CAH. Vrai. agnes() réalise une classification ascendante hiérarchique.
6. En R, l’AFC se fait avec CA() de FactoMineR. Vrai.
7. En R, ICASSO est utilisé pour stabiliser les résultats d’ICA. Vrai.

**Partie B : exercices sur le choix des méthodes statistiques appropriées**

Applications en biologie/agronomie

1. L’ACP peut être utilisée pour explorer des données morphométriques d’organismes. Vrai.
2. L’AFC peut être utilisée pour explorer la relation entre espèces et milieux. Vrai.
3. L’AFCM est adaptée pour étudier les préférences de consommateurs sur des variétés de fruits. Vrai.
4. La CAH permet de classifier des parcelles agricoles selon des variables environnementales. Vrai.
5. L’ICA est souvent utilisée pour la séparation de signaux EEG ou EMG. Vrai.
6. K-Means permet de classer des génotypes selon leur performance dans différents environnements. Vrai.
7. Les analyses factorielles sont peu pertinentes dans les essais agronomiques. Faux. Elles sont très utiles pour réduire la dimension et interpréter les données complexes.
8. En écologie, l’AFC aide à identifier des gradients environnementaux. Vrai. L’AFC met en évidence des associations entre modalités et gradients.
9. En biologie, les axes factoriels sont souvent interprétés comme des gradients de sélection. Vrai.
10. L’ACP est utilisée pour détecter des outliers génétiques. Vrai.

**Partie C**

Erreurs fréquentes et précautions

1. Il faut toujours garder le maximum d’axes dans une ACP. Faux. Il faut sélectionner un nombre limité d’axes expliquant une part significative de la variance.
2. L’interprétation graphique repose uniquement sur la proximité entre points. Faux. L’interprétation prend aussi en compte les contributions et cos² des points.
3. Le premier axe en ACP explique toujours plus de 50 % de la variance. Faux. Cela dépend des données, le premier axe peut expliquer moins.
4. En AFC, les modalités rarement présentes doivent être ignorées systématiquement. Faux. Elles peuvent contenir des informations importantes.
5. En K-Means, les clusters sont toujours bien séparés visuellement. Faux. Les clusters peuvent se chevaucher ou être proches.
6. L’ordre des colonnes d’un tableau de données change les résultats d’ACP. Faux. L’ACP est insensible à l’ordre des variables.
7. En AFCM, il est inutile de recoder les modalités absentes. Faux. Les modalités absentes doivent être codées pour éviter des biais.
8. Une modalité fortement corrélée avec un axe est bien représentée sur ce dernier. Vrai. Une forte corrélation signifie une bonne qualité de représentation.
9. L’ACP peut être utilisée avant une CAH pour réduire le bruit. Vrai.
10. L’AFC est équivalente à une ACP sur un tableau de fréquences. Vrai.

Statistiques et interprétations

1. En ACP, un cos² élevé signifie une bonne qualité de représentation. Vrai.
2. L’angle entre vecteurs de variables en ACP reflète leur corrélation. Vrai.
3. Un individu éloigné de l’origine dans le plan ACP est atypique. Vrai.
4. En AFC, l’inertie totale est égale au chi² du tableau rapporté à l’effectif total. Vrai.
5. En AFCM, les contributions permettent d’identifier les modalités qui structurent l’axe. Vrai.
6. En CAH, une coupe basse du dendrogramme donne peu de classes. Faux. Une coupe basse donne beaucoup de classes (plus fine).
7. Le K optimal en K-Means peut être estimé avec la méthode du coude. Vrai.
8. L’ICA est robuste aux rotations de l’espace initial. Faux. L’ICA est sensible à la rotation, contrairement à l’ACP.
9. Les axes en AFCM sont orthogonaux. Vrai.
10. L’AFCM impose que toutes les modalités aient les mêmes poids. Faux. Les modalités sont pondérées selon leur fréquence.