+1/1/60+

EPFL

Professeur : Gioele La Manno
Examen blanc BIOENG-210: Biological data science
I - MA
22 Mai 2025
1h30

AC-345

Student 1

SCIPER : **999000** Salle : **R-A** Signature :

Ne pas tourner la page avant le début de l'examen. Ce document est imprimé recto-verso, a 10 pages, la dernière éventuellement vierge. Ne pas retirer l'agrafe.

- Placer votre carte étudiante sur la table en évidence.
- **Aucun document papier** n'est autorisé pendant la durée de l'examen.
- **Vous pouvez vous servir d'une calculatrice.**
- Cet examen contient des questions à choix multiples. Chaque question admet une unique réponse correct. Le barème est le suivant:

  1 points si la réponse indiquée est juste,
  0 points s'il n'y pas de réponse ou si plus d'une unique réponse est indiquée,
  0 points si la réponse indiquée est fausse.
- Utiliser un **stylo bleu ou noir** et effacer clairement avec **du fluide correcteur** si nécessaire.
- Si une question est mal formulée ou fausse, le professeur pourra décider de l'annuler.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien

choisir une réponse | select an answer
Antwort auswählen

ne PAS choisir une réponse | NOT select an answer
NICHT Antwort auswählen

Corriger une réponse | Correct an answer
Antwort korrigieren

ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte

Formules

Fonction de masse de probabilité pour une variable aléatoire X suivant une loi de Poisson où λ est à la fois la moyenne et la variance.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Estimateur de James-Stein : si l'on estime simultanément $p \geq 3$ moyennes notées Y_i alors l'estimation suivante est plus précise que l'estimation par maximum de vraisemblance:

$$\hat{\mu}_i^{JS} = \left(1 - \frac{(p-2)}{\sum_{j=1}^{p} Y_j^2}\right) \cdot Y_i$$

Lien complet ("complete linkage"): distance entre les points les plus éloignés.

$$d(C_i, C_j) = \max_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y})$$

Matrice de covariance sur les observations ("sample covariance matrix") : avec $\mathbf{X}_c$ la matrice de données centrées et n le nombre d'observations

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1} \mathbf{X}_c^T \mathbf{X}_c$$

Questionnaire à choix multiples

Pour chaque question, cocher la case correspondant à la bonne réponse.

Question 1 En statistiques multivariées, les matrices de covariance jouent un rôle fondamental pour capturer les relations entre variables aléatoires. Pour 2 variables, la matrice de covariance sera une matrice 2×2. Cependant, toutes les matrices 2×2 ne sont pas des matrices de covariance valides. Laquelle de ces matrices 2×2 n'est PAS une matrice de covariance valide ?

- ☐ $\begin{pmatrix} 2 & 0.5 \\ 0.5 & 3 \end{pmatrix}$
- ☐ $\begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}$
- ☐ $\begin{pmatrix} 4 & 3 \\ 3 & 9 \end{pmatrix}$
- ☐ $\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$

Question 2 Soit Σ une matrice de covariance et Σ_{ij} son entrée à la colonne i, ligne j. Étant donné $\Sigma_{12} = 4$, $\Sigma_{11} = 16$ et $\Sigma_{22} = 9$, la corrélation ρ_{12} est :

- ☐ $\frac{4}{25}$
- ☐ $\frac{4}{144}$
- ☐ $\frac{1}{3}$
- ☐ $\frac{1}{9}$

Question 3 En statistiques multivariées, nous travaillons avec la matrice de covariance des observations Σ calculée à partir d'un ensemble de données avec n observations pour p variables. Que se passe-t-il avec la matrice de covariance des observations lorsque p (nombre de variables) dépasse n (nombre d'observations) ?

- ☐ Elle se factorise en un produit de covariances de dimensions inférieures
- ☐ Elle devient singulière (non inversible) car $\text{rang}(\Sigma) \leq n - 1 < n < p$)
- ☐ Elle s'inverse pour donner une matrice de précision bien définie
- ☐ Elle conserve son rang complet mais n'est plus positive définie (ce qui signifie que certaines valeurs propres deviennent négatives)

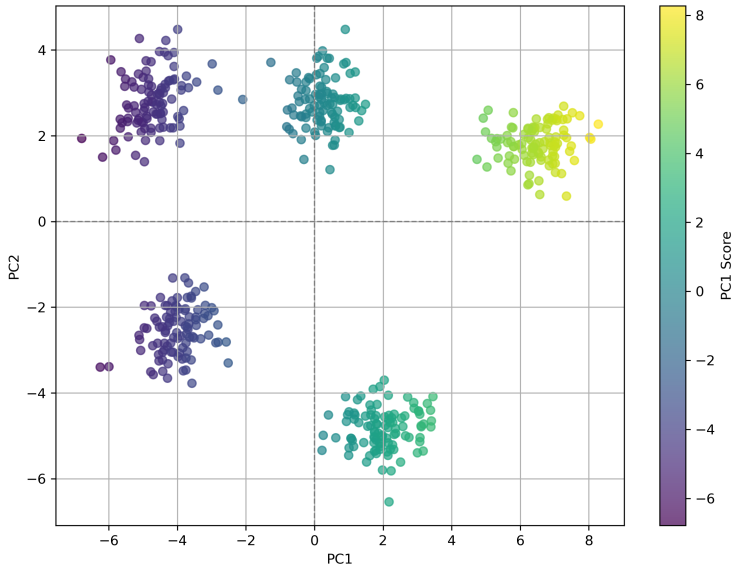
Question 4 $\mathbf{I_p}$ est la matrice identité en p dimensions. Si $\mathbf{\Sigma} = \mathbf{I}_p$ pour une distribution normale p-dimensionnelle, les lignes de niveau de la distribution font apparaitre :

- ☐ Une sphère p-dimensionnelle (hypersphère)
- ☐ Une "boîte" bloc-diagonale
- ☐ Des ellipsoïdes orientés arbitrairement
- ☐ Un point unique à la moyenne

Question 5 Étant donné la Décomposition en Valeurs Singulières d'une matrice de données $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$, quelle affirmation est vraie concernant les vecteurs singuliers droits $\mathbf{V}$?

- ☐ Ce sont des vecteurs propres de $\mathbf{X}\mathbf{X}^T$
- ☐ Ce sont des vecteurs propres de $\mathbf{X}^T\mathbf{X}$ (la matrice de covariance des variables)
- ☐ Ce sont des vecteurs propres à la fois de $\mathbf{X}\mathbf{X}^T$ et de $\mathbf{X}^T\mathbf{X}$
- ☐ Ils sont égaux aux vecteurs singuliers gauches $\mathbf{U}$
- ☐ Ils diagonalisent $\mathbf{X}\mathbf{X}^T$

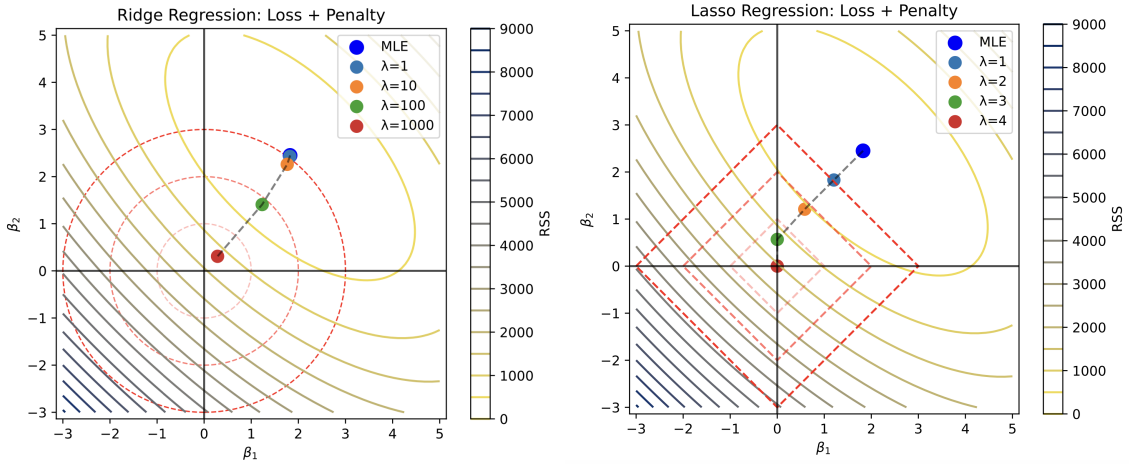
Question 6 Un chercheur effectue une Analyse en Composantes Principales sur un ensemble de données contenant l'expression de plusieurs gènes à travers de nombreuses cellules. Le graphique des scores CP représente les cellules comme des points.



Dans ce graphique, le fait que les groupes soient serrés (les cellules semblent être regroupées étroitement au sein de chaque sous-groupe, et les différents groupes semblent bien séparés les uns des autres) indique :

- ☐ Que les gènes ont une expression identique dans toutes les cellules
- ☐ Que les groupes biologiques varient le long des axes principaux
- ☐ Que le bruit technique domine les premières composantes principales
- ☐ Que l'Analyse en Composantes Principales n'a pas réussi à réduire la dimensionnalité

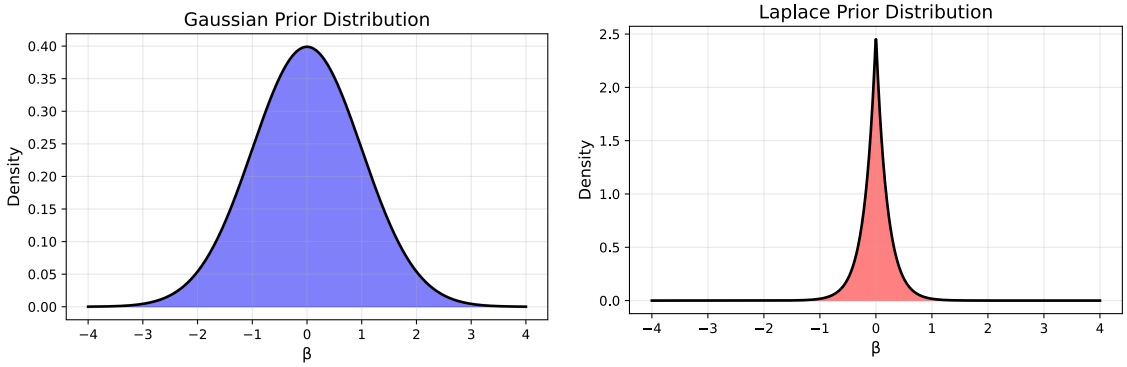
Question 7 Les graphiques suivants montrent (en basses dimensions) l'une des propriétés qui caractérise la régularisation Lasso par rapport à la régularisation Ridge.



Quelle est cette propriété ?

- ☐ Lasso ne pénalise pas β_1 et β_2 via la valeur absolue, contrairement à ridge.
- ☐ Lasso tend à renvoyer plus de coefficients égaux à zéro.
- ☐ Lasso utilise une prior gaussienne, contrairement à Ridge.
- ☐ Lasso produit toujours une Erreur Quadratique Moyenne plus faible que la pénalité Ridge.

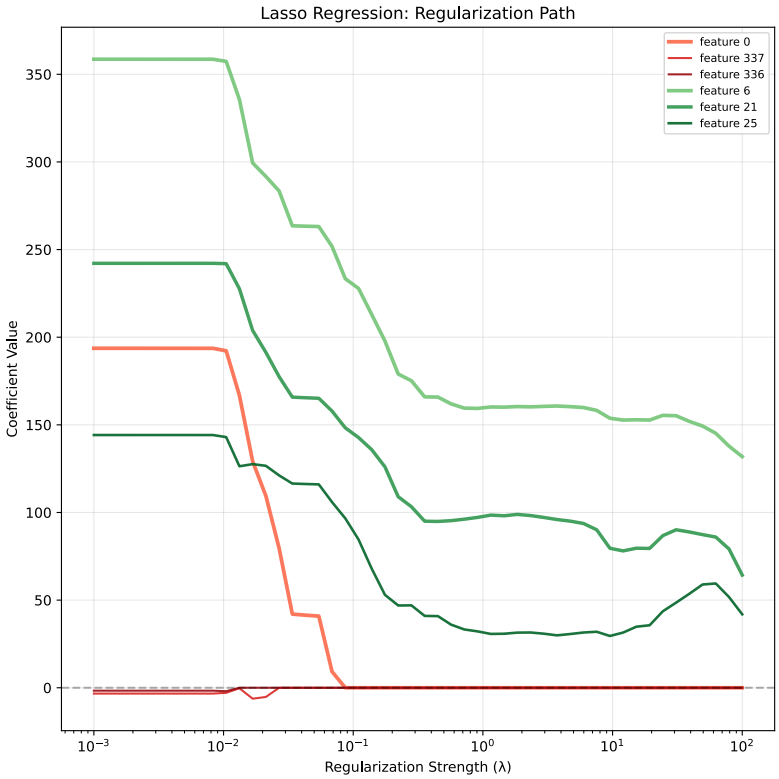
Question 8 Dans le contexte de l'Estimation du Maximum A Posteriori, la densité de l'a priori de Laplace a un pic plus prononcé à zéro que l'a priori Gaussien.



Cette caractéristique géométrique explique pourquoi :

- ☐ Les deux produisent des coefficients d'une densité identique
- ☐ Une priori Laplacienne tend à produire moins de coefficients non nuls.
- ☐ Ridge produit des zéros exacts, pas Lasso
- ☐ Les prédictions faites avec Lasso surpassent toujours celles faites avec Ridge

Question 9 Le chemin de régularisation Lasso montre que de nombreux coefficients atteignent zéro à différentes valeurs de λ.



Un coefficient qui atteint zéro très tôt (petit λ) est probablement :

- ☐ Constant à travers les échantillons
- ☐ Hautement prédictif, protégé par la pénalité
- ☐ Moins prédictif, facilement pénalisé jusqu'à disparaître
- ☐ Parfaitement corrélé avec un autre coefficient

Question 10 Étant donné des données d-dimensionnelles $\{\mathbf{x}_i\}_{i=1}^N$, vous exécutez une analyse en composantes principales et choisissez P composantes principales. Vous souhaitez essayer de reconstruire les points de données originaux en utilisant uniquement les informations capturées par les P composantes principales. Pour effectuer cette reconstruction, vous exprimez chaque point de données comme une combinaison linéaire des P composantes principales. Pouvez-vous toujours reconstruire n'importe quel point de données $\mathbf{x}_i$ pour $i \in \{1, \ldots, N\}$ à partir des P composantes principales avec une erreur de reconstruction nulle ?

- ☐ Non, il est impossible de reconstruire avec une erreur nulle.
- ☐ Oui, si $P < N$
- ☐ Oui, si $P = d$
- ☐ Oui, si $P < d$

Question 11 Vous prenez un échantillon de taille n dans une population de variance inconnue, et calculer la moyenne de cette échantillon $\bar{x}$. Quelle statistique de test et quelle distribution nulle devriez-vous utiliser pour tester $H_0 : \mu = \mu_0$?

- ☐ $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ avec $t(n-1)$
- ☐ $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ avec $N(0,1)$
- ☐ $F = \frac{s^2}{\sigma^2}$ avec la distribution F
- ☐ $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ avec $t(n)$

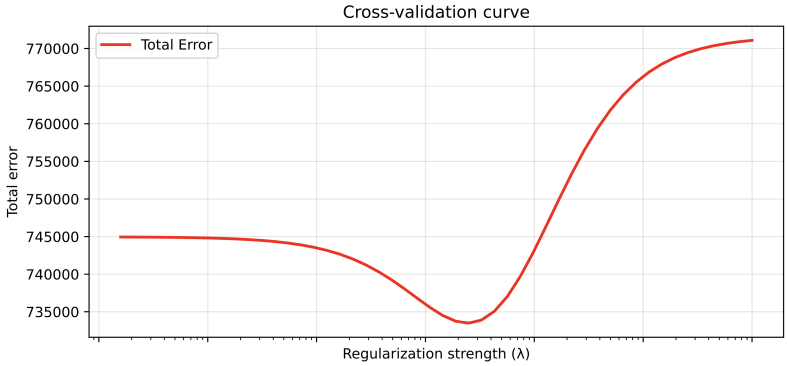
Question 12 Dans une régression linéaire où nous visons à prédire les valeurs de Y basées sur le prédicteur X, la largeur d'un intervalle de confiance de la réponse moyenne pris à une valeur spécifique du prédicteur $X = x_0$, notée $\hat{y}_0$, dépend de plusieurs facteurs. Lequel des changements suivants ne conduirait PAS à un intervalle de confiance à 95% plus étroit pour la réponse moyenne à $X = x_0$?

- ☐ Réduire la variance d'erreur estimée $\hat{\sigma}^2$
- ☐ Augmenter la taille de l'échantillon n
- ☐ Rapprocher x_0 de la moyenne de l'échantillon $\bar{x}$
- ☐ Diminuer la variabilité du prédicteur $\sum(x_i - \bar{x})^2$

Question 13 Une biologiste recueille des données d'expression génique de 50 cellules et veut déterminer si l'expression d'un gène particulier suit une distribution de Poisson. Après avoir ajusté un modèle de Poisson en utilisant l'Estimation du Maximum de Vraisemblance (EMV), elle trouve que l'EMV pour le paramètre de taux λ est $10,5$. Laquelle des affirmations suivantes est correcte concernant cette estimation ?

- ☐ L'EMV de $\lambda = 10,5$ indique que la variance des données d'expression doit être $10,5$
- ☐ L'EMV de $\lambda = 10,5$ signifie qu'exactement $10,5$ molécules sont exprimées dans chaque cellule
- ☐ L'EMV de $\lambda = 10,5$ signifie que la fonction de log-vraisemblance égale $10,5$ à son maximum
- ☐ L'EMV de $\lambda = 10,5$ garantit que le gène suit une distribution de Poisson
- ☐ L'EMV de $\lambda = 10,5$ représente la valeur qui maximise la probabilité d'observer les données données sous un modèle de Poisson

Question 14 Une courbe de validation croisée montre l'erreur en fonction de la force de régularisation λ dans l'ensemble de validation, un ensemble différent de points de données que ceux utilisés pour entraîner le modèle. Une courbe typique de validation croisée a en forme de U :



Le côté gauche (petit λ) correspond à :

- ☐ Surapprentissage—biais faible mais variance élevée
- ☐ Le plateau d'erreur irréductible
- ☐ Sous-apprentissage—biais élevé mais variance faible
- ☐ La règle d'un écart-type

Question 15 Considérez un ensemble de données bivariées provenant d'une expérience biologique où les mesures de deux variables X et Y semblent suivre une distribution normale conjointe. Si l'information mutuelle $I(X;Y)$ est calculée comme étant proche de zéro, laquelle des interprétations suivantes est correcte ?

- ☐ La relation entre X et Y doit être non linéaire et ne peut pas être détectée avec une analyse de corrélation
- ☐ La distribution conjointe peut être bien approximée par le produit des marginales
- ☐ La variance de la variable X doit être approximativement égale à la variance de la variable Y
- ☐ La distribution conditionnelle de Y sachant $X = x$ a exactement la même variance quelle que soit la valeur de x

Question 16 Un chercheur menant une étude génomique teste 5000 gènes pour voir si leur expression diffère significativement du contrôle et utilise la procédure de Benjamini-Hochberg pour contrôler le Taux de Fausse Découverte (FDR) à $0,10$. Après avoir appliqué cette correction, 800 gènes sont déclarés significatifs. Laquelle des affirmations suivantes interprète correctement ces résultats ?

- ☐ Si la même expérience était répétée, exactement 800 gènes seraient à nouveau trouvés significatifs
- ☐ 10% de tous les 5000 gènes testés devraient être des faux positifs
- ☐ Environ 80 des 800 résultats significatifs devraient être des faux positifs
- ☐ La probabilité qu'un résultat significatif individuel soit un faux positif est exactement $0,10$

Question 17 Un biologiste étudiant la relation entre la température (X, en °C) et le taux de réaction enzymatique (Y, en μmol/min) obtient l'équation de régression suivante : $\hat{Y} = 2,5 + 0,8X$ avec $R^2 = 0,65$. L'erreur standard du coefficient de pente est $0,2$. Laquelle des affirmations suivantes est correcte concernant l'interprétation de ce modèle ?

- ☐ Si la température est de 0°C, le taux de réaction enzymatique sera exactement de $2,5$ μmol/min.
- ☐ À partir des données collectées, la moyenne d'échantillon du taux de réaction enzymatique est $2,5$.
- ☐ La relation entre la température et le taux de réaction n'est pas statistiquement significative à $\alpha = 0,05$.
- ☐ Le modèle explique 80% de la variabilité des taux de réaction enzymatique.
- ☐ Pour chaque augmentation de 1°C de température, nous nous attendrions à une augmentation de $0,8$ μmol/min du taux de réaction enzymatique, en moyenne.

Question 18 Lors de la comparaison de deux modèles de régression imbriqués à l'aide du test du rapport de vraisemblance (likelihood ratio test), quelle est l'interprétation correcte si le test donne une valeur p de $0,03$?

- ☐ Il y a 3% de chances que le modèle plus simple soit correct
- ☐ Le modèle plus simple modèle mieux aux données que le modèle plus complexe
- ☐ Le modèle plus complexe explique 3% de variance de plus que le modèle plus simple
- ☐ Il y a suffisamment de preuves pour rejeter l'affirmation comme quoi que le modèle le plus complexe fonctionne mieux que le modèle plus simple.
- ☐ Il y a suffisamment de preuves pour rejeter l'hypothèse nulle selon laquelle les prédicteurs supplémentaires n'améliorent pas le modèle
- ☐ Le modèle plus simple est préféré car il a moins de paramètres et est plus parcimonieux

Question 19 Dans le contexte de la régression Ridge, nous notons la force de régularisation λ. Si $\lambda \to \infty$, les coefficients ajustés $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ tendent vers :

- ☐ Le vecteur propre associé à la plus grande valeur propre
- ☐ La solution de la méthode des Moindres Carrés
- ☐ Zéro pour tous les prédicteurs
- ☐ L'infini

Question 20 Un chercheur applique la régression logistique et la régression de Poisson pour analyser différents ensembles de données biologiques. Laquelle des affirmations suivantes décrit correctement ces modèles dans le cadre des Modèles Linéaires Généralisés (GLM) ?

- ☐ Dans la régression logistique, un coefficient $\beta_1 = 0,5$ signifie que la probabilité du résultat augmente de $0,5$ pour chaque augmentation d'une unité de X_1, tandis que dans la régression de Poisson, le même coefficient signifie que le comptage augmente de $0,5$.
- ☐ Le modèle de régression logistique utilise la fonction de lien logit et suppose que la variance est égale à la moyenne, tandis que la régression de Poisson utilise le lien logarithmique et suppose une variance constante.
- ☐ La régression logistique et la régression de Poisson utilisent toutes deux la fonction de lien logarithmique, mais diffèrent par leurs distributions de composantes aléatoires (Bernoulli vs. Poisson).
- ☐ Les deux modèles permettent de passer d'un espace de réponse borné à un espace linéaire non borné, mais la régression logistique est choisie pour les résultats binaires bornés entre 0 et 1, tandis que la régression de Poisson est choisie pour les comptages bornés à 0.

Question 21 Trois chercheurs modélisent le nombre de mutations par cellule comme un processus de Poisson avec le paramètre de taux λ. À partir de leurs données pilotes, chacun propose un taux fixe différent :

(a) Dr. Adams : $\lambda = 2$

(b) Dr. Baxter : $\lambda = 4$

(c) Dr. Collins : $\lambda = 6$

Ils observent ensuite une nouvelle cellule et comptent 3 mutations. Sur la base de ce nouveau point de données, quel modèle modèle le mieux les données ?

☐ Dr. Collins
☐ Dr. Adams
☐ Dr. Baxter

Question 22 Un bioinformaticien estime simultanément les niveaux d'expression moyens ($\log_2$) de 5 gènes. Chaque mesure a une variance connue $\sigma^2 = 1,0$. Les changements de $\log_2$-fold observés (par rapport au contrôle) sont :

(a) Gène A : 3,10

(b) Gène B : -1,75

(c) Gène C : 2,42

(d) Gène D : 0,87

(e) Gène E : -0,39

Ces estimations ont été obtenues en utilisant l'approche standard de calcul de la moyenne d'échantillon. Cependant, le chercheur se souvient d'un article statistique des années 1960 montrant que cette méthode d'estimation est inadmissible lors de l'estimation simultanée de trois moyennes ou plus. L'article décrivait une approche alternative garantissant une erreur quadratique moyenne (Mean Squarred Error - MSE) globale plus faible, mais le chercheur ne se souvient pas de la formule exacte. Lequel des ensembles d'estimations suivants améliorerait la MSE des estimations du chercheur ?

☐ $[2,62, -1,48, 2,05, 0,74, -0,33]$
☐ $[2,30, -1,30, 1,80, 0,65, -0,29]$
☐ $[2,75, -1,35, 2,18, 0,87, -0,20]$
☐ $[2,72, -1,53, 2,12, 0,76, -0,34]$

Question 23 Dans un espace de haute dimension, considérez 4 points dont on mesure les distances deux à deux. La matrice de distance entre ces points est :

	a	b	c	d
a	$0,0$	$1,0$	$4,0$	$6,1$
b	$1,0$	$0,0$	$3,0$	$5,1$
c	$4,0$	$3,0$	$0,0$	$2,2$
d	$6,1$	$5,1$	$2,0$	$0,0$

En utilisant le regroupement hiérarchique agglomératif avec la méthode du lien complet, quelle est la séquence correcte de regroupements successifs ?

- ☐ D'abord fusionner (c,d), puis ((c,d),a), puis (((c,d),a),b)
- ☐ D'abord fusionner (c,d), puis ((c,d),b), puis (((c,d),b),a)
- ☐ D'abord fusionner (a,b), puis ((a,b),c), puis (((a,b),c),d)
- ☐ D'abord fusionner (a,b), puis fusionner (c,d), puis fusionner ((a,b),(c,d))