Mathématiques 1

Présentation du sujet

Cette épreuve constitue une introduction à l'analyse en composantes principales, un domaine des statistiques dans lequel l'analyse spectrale d'une matrice de covariance permet la mise en évidence de *facteurs principaux*. Ceux-ci sont les vecteurs propres de cette matrice, et leur importance est mesurée par la valeur propre (réelle positive) à laquelle ils sont associés. La fin du sujet propose, dans certains cas choisis, une méthode pour la recherche des premiers facteurs principaux d'une matrice de covariance par ordre décroissant d'importance.

La partie I reprend des résultats classiques sur l'orthodiagonalisation des matrices symétriques réelles ainsi que l'étude du rayon spectral $\rho(A)$ d'une telle matrice, établissant l'identité $\rho(A) = \max_{U \in \mathcal{M}_{n,1}(\mathbb{R}) : \|U\|=1} |U^\top A U|$.

La partie II introduit le concept de matrice de covariance associée à un vecteur aléatoire et propose l'étude de ses principales propriétés. On y démontre notamment une formule donnant la matrice de covariance d'une transformation linéaire d'un vecteur aléatoire (question 17), formule utile pour une large partie de la suite du sujet.

Enfin, s'appuyant sur les concepts et propriétés introduits en partie II, la partie III détaille une méthode d'extraction des deux premiers facteurs principaux par optimisation de la fonctionnelle $U \mapsto U^\top \Sigma_Y U$. Le résultat général pour le premier facteur principal est donné en III-B, puis la section III-C se concentre sur l'étude d'un modèle à corrélation uniparamétrée. La section III-D conclut ce sujet en revenant au cas quasi général d'une matrice de covariance présentant des valeurs propres deux à deux distinctes.

Analyse globale des résultats

Sur les 3454 copies corrigées, la moyenne constatée est de 26,2 % du barème, pour un écart-type de 16,2 %, ce qui permet de considérer le sujet comme de longueur raisonnable, et permettant un niveau de discrimination satisfaisant parmi les candidats. La meilleure copie obtient 90,3 % des points du barème total.

Comme nous le verrons plus loin, la sélection des meilleurs candidats s'est essentiellement faite sur deux points : la connaissance (parfois basique) du cours et la qualité du raisonnement, bien plus que sur le volume traité ou l'originalité des idées.

Concernant le premier point, à titre d'exemple, les questions 1 (dont la réussite s'appuie essentiellement sur la connaissance du théorème spectral) et 5 (demandant d'établir que l'application $(P,Q) \mapsto \int_0^1 P(t)Q(t)\,\mathrm{d}t$ définit un produit scalaire sur $\mathbb{R}_{n-1}[X]$) ne sont totalement réussies que par une part significativement minoritaire des candidats (un tiers pour Q1 et un quart pour Q5).

Quant au second point, il est important de vérifier la validité des hypothèses permettant l'utilisation d'un résultat précédemment établi. Par exemple, en question 29, la moitié des candidats l'ayant abordée oublient de rappeler la nécessaire positivité des coefficients diagonaux de la matrice A_2 avant d'utiliser le résultat de la question 21. Par ailleurs, le jury rappelle que la gestion des implications et équivalences dans les raisonnements doit se faire avec la plus grande rigueur : de nombreux candidats tentent de résoudre la question 1 directement par équivalences sans prêter attention à leur validité. Enfin, la manipulation d'espérances et de covariances (en questions 16 et 17) nécessite d'aborder la question de leur existence, une précaution rarement présente dans les copies.

Cette année encore, le soin apporté à la qualité des réponses est un facteur plus décisif dans les résultats finaux que la quantité de questions traitées. Par exemple, parmi les copies obtenant plus de la moitié des points du barème total, environ 85 % de la note se répartit sur seulement 25 des 38 questions du sujet.

Commentaires sur les réponses apportées et conseils aux futurs candidats

Ce sujet se caractérise par une difficulté progressive et la quasi-absence de questions nécessitant une forte prise d'initiative de la part des candidats. Malgré la présence de notions de probabilités à partir de la question Q16, les domaines mathématiques concernés par ce sujet se concentrent essentiellement autour de l'algèbre bilinéaire et de la réduction des matrices symétriques réelles. Les quelques questions probabilistes (Q21, Q27, Q28 par exemple) sont rarement abordées par les candidats (et, beaucoup plus rarement encore, réussies).

Le jury a relevé un certain nombre de points généraux dans la correction des copies, et en tire les recommandations suivantes.

— Le jury note des *faiblesses importantes et largement répandues sur des points de cours élémentaires.* La question 1 dont la substance repose sur la connaissance de l'énoncé du théorème spectral n'est totalement réussie que dans environ un tiers des copies (pour une question traitée par 90 % des candidats). Plus loin dans le sujet, la question Q5, demandant de vérifier que l'application $(P, Q) \mapsto \int_0^1 P(t)\,Q(t)\,\mathrm{d}t$ définit un produit scalaire sur $\mathbb{R}_n[X]$, n'est pleinement réussie que dans un quart des copies (pour une question traitée par 99 % de candidats). Il est par ailleurs faux de croire, comme vu en réponse à la question 16, qu'une variable aléatoire admet une espérance pour la seule raison qu'elle est discrète.

— *Un enchainement de calculs ou de symboles logiques ne peut constituer une réponse à part entière.* Le jury relève une proportion importante de copies présentant presque systématiquement les réponses de cette manière, avec un maniement souvent bancal des symboles logiques élémentaires (implications, équivalences en particulier), utilisés, à tort, comme des abréviations. Le jury encourage les futurs candidats à davantage rédiger, à subordonner leurs calculs et enchainements logiques à un texte constitué.

— *Les variables utilisées par les candidats sont loin d'être systématiquement déclarées.* Il n'est pas rare de voir apparaitre des indices, des polynômes ou des matrices, au milieu d'un raisonnement, sans en avoir constaté la moindre déclaration préalable, laissant au lecteur le soin de comprendre dans quel ensemble ces variables se trouvent, ou ce qu'elles désignent. En question 7, notamment, le décalage des indices entre 0 et $n-1$, par contraste avec l'habitude d'indexer les composantes d'un vecteur de 1 à n, rend nécessaire ce degré de précision avant le moindre calcul. Un manque de rigueur sur ce plan nuit à la clarté du discours et rend le raisonnement confus.

— *Le jury recommande aux candidats une posture d'humilité*, et notamment de bannir de leur vocabulaire des mots comme « clairement », « trivialement », « évidemment ». Ceux-ci n'apportent rien au contenu mathématique de la copie et ne peuvent jouer qu'en défaveur du candidat, surtout lorsqu'ils sont suivis d'erreurs manifestes ou lorsqu'ils ont pour effet d'éluder des points essentiels à la résolution de la question.

Le jury rappelle également que les *fautes d'orthographe*, malheureusement nombreuses dans les copies, nuisent au candidat et laissent au lecteur une impression négative qui peut se répercuter, consciemment ou non, sur la note finale (en plus de faire l'objet d'un malus). Citons pour exemple, malheureusement très fréquents : « théorème spectrale », « théorème de transfère », « valeure propre », « la fonction atteind ses bornes », « dévellopement », « il est admit que », etc. La validité d'un raisonnement passe aussi par la correction de la langue employée pour l'exprimer.

Voici désormais les remarques du jury, question par question.

Q1. Une question proche du cours pleinement réussie par une proportion relativement peu importante des candidats, en particulier pour l'implication consistant à montrer qu'une matrice symétrique réelle est orthodiagonalisable (terme défini dans l'énoncé), c'est-à-dire diagonalisable en base orthonormée. Les confusions entre la transposée $A^\top$ et l'inverse A^{-1} d'une matrice A sont nombreuses.

Q2. Question globalement réussie par les candidats. Toutefois, la piste donnée par l'énoncé de cette question, consistant à comprendre une opération sur les colonnes sous la forme d'une multiplication à droite (ici, par un vecteur colonne), aura été peu comprise par les candidats.

Q3. Au moment de conclure quant au spectre de A_1, on note beaucoup de calculs faisant intervenir le polynôme caractéristique de la matrice A_1, alors qu'il suffisait d'invoquer son caractère diagonalisable et l'invariance de sa trace par similitude. L'utilisation rigoureuse de la trace est relativement rare parmi les candidats.

Q4. Le jury note très peu de bonnes réponses à cette question, en particulier quant à l'orthonormalisation d'une base de vecteurs propres. De nombreuses copies se contentent de proposer une base de diagonalisation de la matrice A_1 sans se préoccuper de la rendre orthonormale (par exemple en utilisant le procédé de Gram-Schmidt).

Q5. Un exemple classique cité au programme, pourtant faiblement réussi par les candidats. La justification du caractère défini positif aura été le lieu de nombreuses approximations.

Q6. Question réussie par deux tiers des copies, pleinement valorisée pour le calcul de l'intégrale $\int_0^1 t^{i+j}\,\mathrm{d}t$, un calcul qui aura posé des problèmes à une proportion non négligeable des copies. Les représentations lacunaires de la matrice H, non étayées par un calcul explicatif, ne peuvent constituer une réponse pleinement satisfaisante à cette question.

Q7. Beaucoup d'approximations de calcul dans la gestion de la somme double, pour une question réussie par un tiers des candidats.

Q8. La symétrie de H est traitée par une grande majorité de candidats, mais avec des arguments souvent surprenants (« d'après le dessin de Q6, H est symétrique » ou « H est symétrique donc $H \in S_n(\mathbb{R})$ »). Les autres aspects de la question sont rarement traités et réussis dans les copies (en particulier, les valeurs propres d'une matrice, même symétrique, ne sont pas ses éléments diagonaux).

Q9. Il est bon de préciser pour quelle raison une matrice nilpotente admet au moins une valeur propre réelle avant d'en établir la nécessaire nullité. Le jury rappelle que la notion de matrice nilpotente et, à fortiori, tout résultat théorique sur les matrices nilpotentes, est hors programme. De nombreux candidats pensent qu'une matrice nilpotente est diagonalisable, ce qui n'est pourtant vrai que pour la matrice nulle.

Q10. Question souvent traitée, où les principaux problèmes sont pour justifier la continuité de l'application $U \mapsto U^\top U$. Beaucoup de candidats pensent que le caractère borné d'une partie de $\mathbb{R}^n$ implique son caractère fermé, ce qui est faux. On rappelle également que $U^\top U$ et $UU^\top$ ne sont pas des matrices de taille identique.

Q11. Dans cette question, moins traitée que la précédente, on trouve les mêmes problèmes quant à la continuité de l'application $U \mapsto |U^\top AU|$, ainsi que la référence correcte au théorème des bornes atteintes. Une proportion significative des candidats écrivent, à tort, que l'application $U \mapsto U^\top AU$ est linéaire.

Q12 à **Q14**. Questions traitées par moins de la moitié des candidats, et réussies uniquement dans les meilleures copies.

Q15. Question souvent traitée, rarement réussie, à cause de nombreuses approximations dans la gestion des valeurs absolues. Il ne faut pas non plus oublier que la propriété d'homogénéité à établir pour une norme est une propriété d'homogénéité *positive*, un point qui aura manqué dans de nombreuses copies.

Quant à l'axiome de séparation, il nécessite d'invoquer le caractère diagonalisable d'une matrice symétrique réelle, un argument uniquement rencontré dans les meilleures copies. Enfin, il est faux de croire qu'étant donné deux matrices $A, B \in \mathcal{M}_n(\mathbb{R})$, le spectre de $A+B$ est constitué des sommes $\lambda+\mu$ pour λ parcourant le spectre de A et μ parcourant le spectre de B.

Q16 et **Q17**. Questions souvent traitées, avec une appropriation variable du concept de matrice de covariance. Il ne faut pas oublier d'établir l'existence des espérances et covariances évoquées, un point souvent négligé dans les copies. Par ailleurs, la moindre identité impliquant des vecteurs aléatoires, par exemple $\mathbb{E}(MY) = M\mathbb{E}(Y)$ nécessite de raisonner composante par composante, et ne constitue pas, comme lu dans de nombreuses copies, une extension évidente de la propriété de linéarité de l'espérance.

Q18. Question réussie par un tiers des candidats, qui voient le changement de base impliqué par la formule démontrée en Q17 lorsque M est une matrice orthogonale. De nombreux candidats pensent ou laissent entendre que la matrice orthogonale P réalisant le changement de base peut être prise quelconque, ce qui n'est pas le cas.

Q19 et **Q20**. Questions peu traitées, et peu réussies pour ceux qui s'y attellent. Peu reconnaissent que $\mathrm{cov}(X_i, X_i) = \mathbb{V}(X_i) \geqslant 0$ pour Q19 ou remarquent l'argument d'invariance de la trace par similitude en Q20. On rappelle également qu'une covariance entre deux variables aléatoires n'est pas toujours positive.

Q21. Question difficile qui demande d'utiliser la richesse de l'espace probabilisé mentionnée en introduction. Moins de 1 % des copies proposent une solution complète.

Q22. Question traitée par moins d'un tiers des candidats, mais davantage réussie que la précédente, certains voyant la déduction qui s'opère à partir de Q20 et de la formule démontrée en Q17.

Q23. Peu de réussite (10 % des copies) pour une question se référant presque directement à Q17 et demandant de remarquer la taille $(1,1)$ de la matrice obtenue.

Q24. Question réussie par la moitié des copies s'y étant consacrées (une moitié des candidats).

Q25 à **Q28**. Questions peu traitées par les candidats, et réussies uniquement dans les meilleures copies. Le jury note de nombreuses confusions entre les notions de supplémentaire et de complémentaire en Q25. La question Q28 est probablement la plus difficile du sujet, demandant la mise en place d'une intersection d'événements et une utilisation judicieuse des questions Q25 à Q27, un raisonnement très rarement rencontré dans les copies.

Q29. La référence presque directe à la question Q21 aura été remarquée par une proportion significative de copies.

Q30 et **Q31**. Questions rarement traitées, avec des références à la partie I, souvent remarquées par les candidats, rarement bien mises en place avec rappel des hypothèses.

Q32 et **Q33**. L'étude spectrale de la matrice J est abordée dans de nombreuses copies, avec de la réussite et aussi quelques approximations dans le calcul des dimensions des sous-espaces propres.

Q34 à **Q38**. Questions très rarement abordées par les candidats, avec des solutions proposées uniquement dans les toutes meilleures copies.

Conclusion

Il est absolument primordial de se présenter à une épreuve de ce niveau avec une connaissance précise des éléments de cours et une capacité à les manier avec précision et rigueur. Il est également important d'apporter une attention particulière à ce qui semble être considéré par de nombreux candidats – à tort – comme des détails : déclaration des variables, utilisation pertinente des liens logiques (implications, équivalences) et des mots de liaison. Il importe également que les candidats sélectionnent et mentionnent

explicitement la totalité des arguments nécessaires pour répondre à chaque question. En effet, les correcteurs, à l'écrit (contrairement aux examinateurs, à l'oral), ne peuvent interroger les candidats afin de leur demander d'étayer leurs affirmations ou de les compléter ; il faut donc que tout soit exprimé sur la copie. Ce manque de rigueur explique que de nombreux candidats risquent de se retrouver déçus par leur note, ayant eu l'impression de traiter de nombreuses questions du sujet, alors que la plupart des réponses sont incomplètes ou insuffisamment précises.

Le jury tient également à rappeler l'impact significatif d'une copie bien présentée, rédigée dans un français correct. Il en aura été tenu compte dans la notation. Les désagréments impliqués par un manquement à ces règles d'usage sont doubles :

— sur le fond, un certain manque de soin ou une rédaction précipitée fait manquer des points importants de la question ou certaines étapes cruciales d'un raisonnement ;

— sur la forme, l'impression laissée au correcteur par une copie négligée est forcément négative.

Pour éviter tout désagrément, le jury recommande aux candidats de soigner leur écriture, de limiter les ratures, d'éviter de multiplier les insertions plus ou moins lisibles ou les renvois vers une autre page, et d'écrire dans un français correct.

Enfin, il n'est pas nécessaire de se précipiter et de traiter un nombre impressionnant de questions pour obtenir un très bon total : il suffit de procéder avec soin, dans un esprit scientifique empreint de rigueur et de précision. Le jury encourage les futurs candidats à prendre ces bonnes habitudes dans leur préparation. Les bonnes et très bonnes copies sont, presque sans exception, de cette sorte.