Agrégation externe de mathématiques
Épreuve de modélisation, option C : algèbre et calcul formel

(*Texte public*)

Résumé : Le texte propose un modèle pour localiser le gène responsable d'une maladie génétique. L'étude mathématique de ce modèle conduit à traiter un problème d'élimination délicat, pour lequel une solution *ad hoc* est construite.

Mots clefs : algèbre linéaire, élimination

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. Vous êtes laissé(e) libre d'organiser votre discussion comme vous l'entendez. Des suggestions de développement, largement indépendantes les unes des autres, vous sont proposées en fin de texte. Vous n'êtes pas tenu(e) de les suivre. Il vous est conseillé de mettre en lumière vos connaissances à partir du fil conducteur constitué par le texte. Le jury appréciera que la discussion soit accompagnée d'exemples traités sur ordinateur.*

1. Problème

Le programme génétique d'un individu est contenu dans un ensemble de 23 paires de chromosomes, chacun formé d'un ensemble de gènes. Chaque individu dispose donc de deux versions différentes de chaque gène ; nous supposerons [1] que l'une est héritée de sa mère (et est l'une des deux versions du gène portées par sa mère), l'autre est héritée de son père. Une maladie génétique est liée au fait qu'un gène particulier est défectueux. Cela peut signifier, selon les cas, qu'une des deux versions du gène est défectueuse (si cela suffit à déclencher la maladie, cette dernière est dite *dominante*), ou que les deux versions du gène sont défectueuses (si cela est nécessaire pour déclencher la maladie, cette dernière est dite *récessive*).

Dans la suite, nous noterons d un gène défectueux, et n un gène normal. L'étude familiale consiste, dans un large ensemble de familles dans laquelle la maladie génétique est présente, à comparer les génomes de deux frères ou soeurs atteints de la maladie pour tenter de comprendre quel est le gène responsable et son mode d'action (récessive, dominante, etc).

1. la réalité biologique est un peu plus complexe

2. Modèle

Soit p la fréquence du gène défectueux dans l'ensemble de la population. Nous supposerons que les deux versions du gène sont indépendantes pour chaque individu. Cela implique que la probabilité de l'événement nn (les deux copies du gène sont normales) est $(1-p)^2$, la probabilité des événements dn et nd est $p(1-p)$, et la probabilité de l'événement dd est p^2.

Un modèle pour la maladie est la donnée de trois réels positifs f_0, f_1, f_2, non tous trois nuls, où f_i est la probabilité d'être porteur de la maladie sachant que l'on porte i copies du gène étudié. Les modèles les plus classiques sont le modèle récessif ($f_0 = f_1 = 0$, $f_2 = f$) et le modèle dominant ($f_0 = 0$, $f_1 = f_2 = f$), mais le modèle additif ($f_i = if/2$) est aussi pertinent. Enfin, dans le cas où le gène étudié n'est pas relié à la maladie, on s'attend à ce que $f_0 = f_1 = f_2 = f$.

Étant données des familles porteuses de la maladie et un gène dont on soupçonne qu'il est en cause, on regarde le nombre de copies du gène communes entre deux frères et soeurs infectés par la maladie. Pour $i = 0, 1, 2$, notons z_i la probabilité que deux frères et soeurs aient i copies identiques (c'est-à-dire venant du même chromosome du même parent) du gène et qu'ils soient tous deux malades. En réalité, on s'intéresse plutôt à la probabilité qu'ils aient i copies identiques *sachant qu*'ils sont tous deux malades, mais ces deux questions sont équivalentes : cette probabilité diffère de z_i d'une constante.

Dans le cas où gène et maladie sont indépendants, on s'attend à ce que (z_0, z_1, z_2) soit proportionnel à $(1/4, 1/2, 1/4)$.

Notre modèle consiste alors à exprimer z_i en fonction de p et des f_i. On obtient la proposition suivante :

Proposition 1. *On a*[2]

$$\begin{pmatrix} z_0 \\ z_1 \\ z_2 \end{pmatrix} = \frac{1}{16} \begin{pmatrix} 4f_0^2 & 16f_0f_1 & 8f_0f_2 + 16f_1^2 & 16f_1f_2 & 4f_2^2 \\ 8f_0^2 & 8(f_0+f_1)^2 & 16f_1(f_0+f_1+f_2) & 8(f_1+f_2)^2 & 8f_2^2 \\ 4f_0^2 & 8(f_0^2+f_1^2) & 4f_0^2+16f_1^2+4f_2^2 & 8(f_1^2+f_2^2) & 4f_2^2 \end{pmatrix} \begin{pmatrix} (1-p)^4 \\ (1-p)^3 p \\ (1-p)^2 p^2 \\ (1-p)p^3 \\ p^4 \end{pmatrix}.$$

Démonstration. Expliquons comment obtenir le terme $(f_0, +f_1)^2/2$ en deuxième ligne et deuxième colonne de la matrice.

Les événements possibles dépendent des types des gènes des deux parents et de la façon dont ils ont été transmis. Si les gènes des parents sont numérotés de 1 à 4, 1 et 2 représentant les gènes du père et 3 et 4 ceux de la mère, le type des gènes des parents est un élément de $\{n,d\}^4$ et les gènes transmis aux deux enfants un élément de $(\{1,2\} \times \{3,4\})^2$. Chaque élément du produit de ces deux ensembles définit un événement différent.

Par exemple, considérons l'événement $(n,n,d,n),(1,3,1,4)$. Cet événement a probabilité $p(1-p)^3/16$ (3 des 4 copies du gène considérées sont normales, une défectueuse), et signifie que le premier enfant sera porteur de gènes de type (n,d) et le second de type (n,n). En

2. Dans la perspective de l'illustration informatique, on vérifiera attentivement qu'on n'a pas commis d'erreur dans la saisie de la matrice ci-dessous.

particulier, un seul gène (le 1) est partagé entre les deux enfants. Si l'on prend en compte le fait que les deux enfants doivent être malades (cf. la définition de z_i), l'événement contribue au terme correspondant à z_1 pour une valeur $f_0 f_1 p(1-p)^3/16$. On dénombre 16 événements contribuant à ce terme pour cette même valeur : (n,n,d,n), $(1,4,1,3)$, (n,n,d,n), $(2,3,2,4)$, ..., (d,n,n,n), $(1,3,2,3)$, etc... Par la même méthode on dénombrera 8 événements contribuant à ce terme correspondant à z_1 pour une valeur de $f_0^2 p(1-p)^3/16$, par exemple : (n,d,n,n), $(1,3,1,4)$, (n,d,n,n), $(1,4,1,3)$, ... ,(n,n,n,d), $(1,3,2,3)$, etc.. et 8 événements contribuant pour une valeur de $f_1^2 p(1-p)^3/16$. D'où le coefficient $(f_0+f_1)^2/2$ dans la matrice. □

Dans la suite, nous noterons la matrice du membre droit $M(f_0,f_1,f_2)$ et le vecteur $v(p)$.

3. Étude du modèle

Nous allons maintenant entreprendre l'étude du modèle. L'objectif final est, étant données les données (z_0,z_1,z_2), de comprendre le mode d'action de la maladie (f_0,f_1,f_2), et en particulier de comprendre si la maladie est liée au gène étudié. Dans le cas contraire, on s'attend à $f_0=f_1=f_2$. Dans la suite, nous étudierons le triplet $(\hat{z}_0,\hat{z}_1,\hat{z}_2)\approx(0,12,0,49,0,39)$.

3.1. *Quelques cas simples*

Commençons par étudier le cas simple où $f_0=f_1=f_2=f$.

Lemme 1. *La maladie est indépendante du gène étudié (autrement dit $f_0=f_1=f_2=f$) si et seulement si (z_0,z_1,z_2) est colinéaire à $(1/4,1/2,1/4)$*

Démonstration. Quand $f_0=f_1=f_2=f$, les équations de la proposition 1 se simplifient et montrent que l'on doit avoir $(z_0,z_1,z_2)=f^2(1/4,1/2,1/4)$.

Réciproquement, étudions ce qui se passe si l'on prend $(z_0,z_1,z_2)=f^2(1/4,1/2,1/4)$ pour un certain $f>0$. La première des trois équations de la proposition 1 se factorise sous la forme :

$$f^2=\left(f_0(1-p)^2+2f_1p(1-p)+f_2p^2\right)^2.$$

L'hypothèse sur la positivité de p,f_0,f_1,f_2,f entraîne

$$f_0=\frac{1}{(1-p)^2}\left(f-2f_1p(1-p)-f_2p^2\right).$$

En injectant cette valeur dans la seconde équation puis en factorisant cette dernière, on exprime f_1 en fonction de f_2,f et p. Puis en injectant ces valeurs dans la troisième équation on trouve $f_2=f$ (indépendamment de p) et, en remontant, on en déduit $f_0=f_1=f_2=f$. □

Ce cas correspond à la situation où $\operatorname{rang} M(f_0,f_1,f_2)\leqslant 1$. On peut traiter par des arguments similaires le cas où $\operatorname{rang} M(f_0,f_1,f_2)=2$; remarquons d'abord qu'il correspond, en particulier, au cas additif.

Proposition 2. *La matrice $M(f_0,f_1,f_2)$ est de rang $\leqslant 2$ si et seulement si $2f_1=f_0+f_2$ ou $f_0=f_2=0$.*

Démonstration. Étudier les mineurs d'ordre 3 de $M(f_0, f_1, f_2)$. □

Si la matrice M est de rang 2, on obtient une condition nécessaire sur (z_0, z_1, z_2) en écrivant que le vecteur ${}^t(z_0, z_1, z_2)$ doit être dans le plan engendré par deux colonnes indépendantes de M. À un facteur près, on obtient la condition $E_A := z_0 - z_1 + z_2 = 0$ dans le cas du modèle additif, et $E_0 := z_1 - z_2$ dans le cas du (peu naturel) modèle $f_0 = f_2 = 0, f_1 = f$.

3.2. *Calcul direct des f_i*

Une solution pour comprendre l'influence du gène étudié consiste simplement à injecter les valeurs de (z_0, z_1, z_2) dans le modèle pour obtenir un système de 3 équations polynomiales en 4 inconnues f_0, f_1, f_2, p. Par analogie avec la situation linéaire, on s'attend à obtenir une infinité de solutions, à savoir un ensemble qui soit géométriquement une courbe.

Injectons dans le système d'équations les valeurs $(z_0, z_1, z_2) = (12/100, 49/100, 39/100)$ (noter que pour éviter l'accumulation d'erreurs numériques dans les calculs subséquents, on prend des valeurs rationnelles et non flottantes). Il est alors possible d'éliminer p entre ces 3 équations par des calculs de résultants ; on obtient sans peine trois équations en f_0, f_1, f_2, une par paire d'équations initiales, à savoir, en notant E_i l'équation correspondant à la i-ème ligne du produit matrice-vecteur,

$$R_{ij} = \mathrm{Res}_p(E_i, E_j), 1 \leqslant i < j \leqslant 3.$$

On parvient ensuite à éliminer, par exemple f_0, en considérant $r_2 = \mathrm{Res}_{f_0}(R_{12}, R_{23})$, $r_3 = \mathrm{Res}_{f_0}(R_{13}, R_{23})$ mais les calculs peuvent commencer à devenir lourds pour le logiciel. À ce stade, toute tentative d'éliminer f_1 ou f_2 échoue en donnant un résultant nul (si du moins le calcul aboutit).

Cependant, avec un logiciel de calcul formel, on peut observer que $(f_1 - f_2)^{24}$ est un facteur commun de r_2 et r_3. On ne s'attendrait pas à trouver une solution exacte comme $f_1 = f_2$ en ayant injecté dans l'équation des valeurs estimées. Aussi, si l'on considère ce facteur $(f_1 - f_2)^{24}$ comme parasite, et que l'on substitue dans $r_2' := r_2/(f_1 - f_2)^{24}$ et $r_3' := r_3/(f_1 - f_2)^{24}$ par exemple $f_2 = 1$, on trouve une seule racine commune admissible à savoir $f_1 \approx 0.6376$. En réinjectant ces valeurs dans les R_{ij}, on peut trouver une racine commune approchée mais elle est négative (≈ -1.48). On ne peut donc pas choisir $f_2 = 1$: il faudrait décider pour quelles valeurs de f_2 on obtiendrait des valeurs f_1 puis f_0 réelles et appartenant à $[0, 1]$, ce qui est difficile.

3.3. *Une méthode alternative d'élimination*

Nous allons maintenant développer une méthode alternative d'élimination, avec pour objectif de construire une condition nécessaire sur les z_i du même type que celles de la partie 3.1, mais dans le cas où le rang de la matrice $M(f_0, f_1, f_2)$ est 3.

En regardant, parmi les modèles classiques, quelle condition est la plus proche d'être satisfaite au point $(\hat{z}_0, \hat{z}_1, \hat{z}_2)$, on pourra ainsi inférer le mécanisme d'action du gène défectueux, ou, si les z_i sont proches du point $(1/4, 1/2, 1/4)$, supposer que le gène étudié n'a pas d'action sur la maladie.

Ces conditions pourraient être obtenues, à nouveau, en utilisant des résultants. Nous allons cependant développer ici une méthode d'élimination ad hoc.

Notons $H(z_0, z_1, z_2)$ la matrice $\begin{pmatrix} z_1 & -z_0 & 0 \\ z_2 & 0 & -z_0 \end{pmatrix}$. On a alors la proposition suivante :

Proposition 3. *Soient Φ et Ψ les deux coordonnées du vecteur $H(z_0, z_1, z_2)M(f_0, f_1, f_2)v(p)$. S'il existe p tel que ${}^t(z_0, z_1, z_2) = M(f_0, f_1, f_2)v(p)$, alors* $\mathrm{Res}_p(\Phi, \Psi) = 0$.

On a donc ramené notre problème d'élimination initial à un calcul de résultant entre deux polynômes de degré 4 en p, à coefficients dans $\mathbb{C}(z_0, z_1, z_2, f_0, f_1, f_2)$.

Il reste à calculer ce résultant. On peut encore aider quelque peu le système de calcul formel pour le calcul du résultant, qui reste assez lourd en raison du nombre élevé de variables des polynômes mis en jeu. Pour ce faire, étant donnés deux polynômes de degré 4 notés $U(X) = \sum_{i=0}^{4} u_i X^i$ et $V(X) = \sum_{i=0}^{4} v_i X^i$, posons

$$R(X, Y) = (U(X)V(Y) - U(Y)V(X))/(X - Y) = \sum_{0 \leqslant i,j \leqslant 3} r_{ij} X^i Y^j.$$

Proposition 4. *Le déterminant de la matrice (r_{ij}) est le résultant de U et V.*

Démonstration. Par exemple, en vérifiant dans un système de calcul formel. □

Tous calculs faits (à l'aide d'un système de calcul formel), on trouve une condition du type

$$z_0^4 f_2^2 f_0^2 (2f_1 - f_0 - f_2)^4 E(f_0, f_1, f_2, z_0, z_1, z_2) = 0. \tag{1}$$

Nous traiterons les trois premiers facteurs comme des facteurs parasites que l'on ignorera. Le quatrième est lié au fait que quand $2f_1 = f_0 + f_2$, on a déjà observé que le rang de la matrice $M(f_0, f_1, f_2)$ est $\leqslant 2$; on l'ignorera également.

Dans le cas du modèle dominant, on applique la stratégie en spécialisant les valeurs $f_0 = 0$, $f_1 = f_2 = 1$ dans l'expression E. On trouve

$$(z_2 + z_0 - z_1)(-4z_0^2 z_2 + 8z_0 z_2 z_1 - 4z_0 z_2^2 - z_0 z_1^2 + 4z_2^3 - 4z_2^2 z_1) = 0.$$

On notera dans la suite E_D le second facteur de cette condition.

Dans le modèle récessif, la condition E s'annule identiquement, bien que $M(f_0, f_1, f_2)$ soit de rang 3. En reprenant les calculs dans ce cas précis, on constate que Ψ et Φ ont toujours 0 comme racine commune. En divisant ces deux polynômes par la puissance de p adaptée, puis en calculant le résultant, on trouve (à des facteurs parasites près) la condition $z_0(z_1^2 - 4z_0 z_2) = 0$, dont on notera E_R le second facteur.

Dans le cas de notre exemple, en utilisant les conditions ci-dessus et la condition E_A trouvée dans 3.1 pour le cas additif, on trouve

$$E_D(\hat{z}_0, \hat{z}_1, \hat{z}_2) \approx 1.7 \cdot 10^{-3}, E_R(\hat{z}_0, \hat{z}_1, \hat{z}_2) \approx 5.3 \cdot 10^{-2}, E_A(\hat{z}_0, \hat{z}_1, \hat{z}_2) = 2 \cdot 10^{-2}.$$

Cela suggère que le gène étudié est impliqué dans la maladie, avec un mécanisme de type dominant.

Suggestions pour le développement

► *Soulignons qu'il s'agit d'un menu à la carte et que vous pouvez choisir d'étudier certains points, pas tous, pas nécessairement dans l'ordre, et de façon plus ou moins fouillée. Vous pouvez aussi vous poser d'autres questions que celles indiquées plus bas. Il est très vivement souhaité que vos investigations comportent une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats.*

— Prouver les assertions non démontrées du texte ;
— Reproduire et détailler les différents calculs conduits dans le texte ;
— Expliquer l'occurrence des différents facteurs parasites rencontrés dans les parties 3.2 et 3.3 ;
— Tracer les courbes $E_D = 0$, $E_A = 0$, $E_R = 0$ dans le plan $z_0 + z_1 + z_2 = 1$ et placer le point $(\hat{z}_0, \hat{z}_1, \hat{z}_2)$ sur le dessin ;
— Étudier la généralisation de la proposition 4 en degré quelconque ;
— Interpréter la matrice $H(z_0, z_1, z_2)$;
— La condition dans le cas additif est-elle nécessaire et suffisante ?
— La méthode utilisée dans le paragraphe 3.3 fournit-elle une condition suffisante ? Discuter ;
— Comparer les conditions R_{ij} de la partie 3.2 et la condition E de la partie 3.3, et discuter ;
— Dans le cas du modèle dominant, retrouver la condition E_D au moyen de calculs de résultants.