Analyse en composantes principales d'héritabilité pour données familiales

Karim Oualkacha

Professeur au Département de Mathématiques, UQÀM

Présentation dans le cadre de la journée de

La statistique au service de la collectivité:

Hommage à Louis-Paul Rivest pour ses 60!!!

29 Août 2014

Maladies complexes

- Cancer, diabète, maladies psychiatriques, maladies cardio-vasculaires...
- Interaction entre des facteurs environnementaux, comportementaux, psychologiques, sociaux, sanitaires et **génétiques**
- Facteurs génétiques : des centaines de gènes peuvent être impliqués

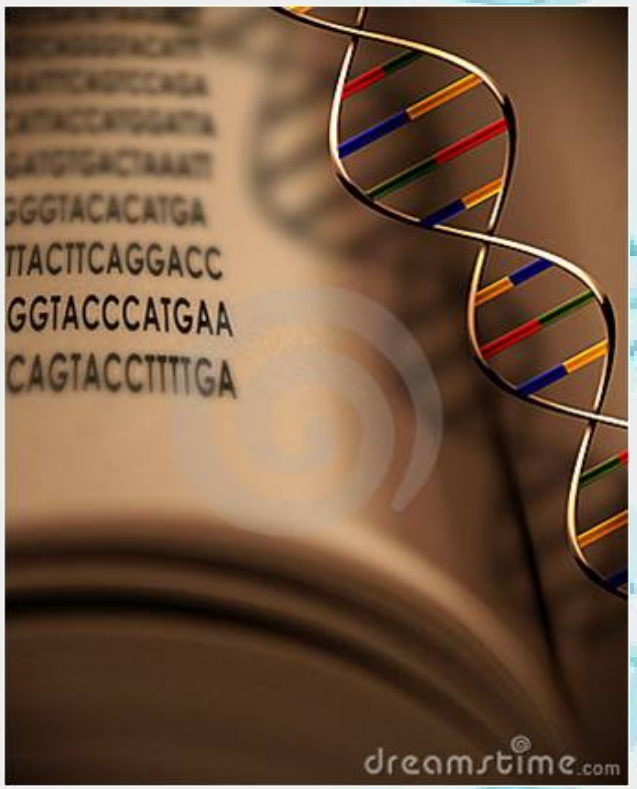
Principals défis pour les statisticiens

- Quantité et complexité des données
- **Quantité**
  - Dans le cas de maladies génétiques complexes, les chercheurs collectent le plus grand nombre possible de données
  - Matrice de données $N \times r$ où N se compte en <u>milliers</u> et r se compte en <u>millions</u>
- **Complexité** :
  - Variables mesurées souvent corrélées entre elles
  - Maladies ne sont pas monogéniques
  - Individus impliqués dans l'étude sont reliés (cas de données familiales)

QUELQUES RAPPELS

ADN, génome et gènes

- Vous pouvez voir l'ADN comme un livre qui contient toutes les informations nécessaires à la vie d'un organisme
- Ce livre (le Génome) est écrit dans un alphabet de 4 lettres : A, T, G, C
- Les gènes sont des séquences d'ADN qui portent les caractères héréditaires transmis au cours des générations
- L'endroit précis d'un segment d'ADN sur le génome est appelé locus, ou loci au pluriel



Allèles d'un gène

- On estime que le génome humain contient 3 milliards de bases (lettres : A, T, G, C)
- De plus, 99.9% d'ADN entre 2 individus quelconques est identique !
- L'idée est donc d'aller extraire seulement les loci qui varient entre les individus
- Chaque individu possède 2 copies de son ADN : une copie reçue de sa mère et une copie reçue de son père
- L'orthographe de chaque copie d'un segment d'ADN est nommée allèle

- Un marqueur génétique est une séquence d'ADN variable selon les individus, mais dont la localisation est connue sur le génome

Phénotypes et Traits quantitatifs

- Le phénotype est un caractère observable et mesurable chez les individus
  Ex : la pression artérielle, l'indice de masse corporelle, la couleur des yeux, la taille ...

- Un trait quantitatif est un phénotype pour lequel la variation chez les individus dans une population est continue

- Cette variation continue est attribuable à la variation génétique combinée à la variation environnementale

- Le diagnostic des maladies est souvent basé sur un ou plusieurs traits quantitatifs
  Ex : La pression artérielle et les niveaux de cholestérol sont des traits quantitatifs liés aux maladies cardiovasculaires

Motivation

- Dans le cas de maladies génétiques complexes (Schizophrénie, la maladie bipolaire, Diabète II), les chercheurs collectent le plus grand nombre possible de traits quantitatifs

  Ex 1 : 82 traits quantitatifs collectés chez 48 familles atteintes de la Schizophrénie et de la maladie bipolaire dans la province de Québec (CRULRG)

  Ex 2 : Vingtaine de traits quantitatifs liés à l'obésité sont collectés chez 431 familles, de la ville de Québec et la région métropolitaine, impliquées dans une étude sur les facteurs d'obésité (Dr. J. Engert, Université McGill).

Question :

Quels traits ou combinaisons de traits seront choisis pour identifier les gènes responsables de ces maladies ?

Modèle statistique pour des traits quantitatifs

Modèle génétique simple

- Supposons que Y dénote le trait quantitatif d'intérêt
- Modèle génétique simple avec un seul gène

$$Y = \mu + G + E, \quad G \perp E$$

- Les composantes de la variation phénotypique :

$$Var(Y) = \sigma_Y^2 = \sigma_g^2 + \sigma_e^2$$

- Le rapport σ_g^2/σ_Y^2 donne une mesure d'héritabilité (h^2)
- Comment des gènes partagés entre deux individus de la même famille contribuent à leur covariance ?

$$Cov(Y_j, Y_k) = Cov(\mu + G_j + E_j, \mu + G_k + E_k) = Cov(G_j, G_k)$$

$$Cov(Y_j, Y_k) = 2\Phi_{jk}\sigma_g^2$$

Φ_{jk} est la probabilité que les individus j et k partagent 1 allèle par descendance

Le modèle dans le cas multivarié

- Supposons que Y est le vecteur de r traits d'intérêt
- Le modèle est

$$Y = \mu + G + E, \quad G \perp E$$

$$Var(G) = \Sigma_g, \quad Var(E) = \Sigma_e$$

- La covariance de l'effet génétique entre deux individus dans la même famille

$$Cov(G_j, G_k) = 2\Phi_{jk}\Sigma_g$$

- Φ_{jk} est le coefficient de kinship entre les individus j et k
- La matrice de variance-covariance de Y

$$Var(Y) = \Sigma_g + \Sigma_e, \quad Cov(Y_j, Y_k) = 2\Phi_{jk}\Sigma_g$$

Motivation

- Dans le cas de maladies génétiques complexes (Schizophrénie, la maladie bipolaire, Diabète II), les chercheurs collectent le plus grand nombre possible de traits quantitatifs

Question :

Quels traits ou combinaisons de traits seront choisis pour identifier les gènes responsables de ces maladies ?

Techniques de réduction de la dimension des données : Données familiales

ACP et ACPH

- L'analyse en composantes principales (ACP standard) maximise

$$ACP = \arg\max_{||\beta||=1} \left[\beta^T Var(Y)\beta\right]$$

$$Var(Y) = \Sigma_g + \Sigma_e$$

- Quand les données sont familiales, l'ACP ne tient pas compte de la structure familiale
- L'approche de Ott et Rabinowitz (1999) maximise l'héritabilité h^2

$$CPH = \arg\max_{||\beta||=1} h^2(\beta^T Y) = \arg\max_{||\beta||=1} \frac{\beta^T \Sigma_g \beta}{\beta^T(\Sigma_g + \Sigma_e)\beta}$$

ACP et ACPH

- Cette maximisation est équivalente à

$$CPH = \arg\max_{||\beta||=1} \frac{\beta^T \Sigma_g \beta}{\beta^T \Sigma_e \beta}$$

- Solution : La première composante est le premier vecteur propre de $\Sigma_e^{-1}\Sigma_g$ (Mardia et al., 1979)

- En présence d'un plus grand nombre de traits que de familles, l'approche PCH rencontre le problème connu sous le nom "la petite taille d'échantillon"
  Ex : Analyse génétique dans Morley et al. (2004), r = 3554 phénotypes mesurés sur 14 familles

Régularisation des CPH

- Conséquences :
  - Plusieurs valeurs propres de Σ_e peuvent être estimées par zéro
  - Tout β qui satisfait $\Sigma_e \beta = 0_r$ sera une solution de la maximisation (maximum non identifiable)
- Wang et al. (2007) ont proposé une régularisation pour l'approche CPH

$$CPH_\lambda = \arg\max_{||\beta||=1} \frac{\beta^T \Sigma_g \beta}{\beta^T (\Sigma_e + \lambda I_r) \beta}$$

- Le paramètre optimal de régularisation est donné par une technique standard de validation croisée

Estimation des CPH

- L'estimation des CPH revient à estimer Σ_g et Σ_e
- Lorsque les données sont collectées à partir des familles de frères et soeurs seulement
  - La corrélation entre les individus de la même famille est la même pour toutes les paires d'individus
- Wang et al. (2007) ont estimé Σ_g par la variance échantillonnale inter familles et Σ_e par la variance échantillonnale intra familles

Estimation des CPH

- **Défis statistiques :**
  - Lorsque les données proviennent de familles quelconques
  - Les familles n'ont pas nécessairement la même taille
  - La structure de corrélation change entre les individus de la même famille
  - Nous sommes face aux données multivariées non balancées

Estimation des CV Σ_g et Σ_e

- En statistique génétique, les estimateurs des composantes de variance (CV) les plus recommandés sont les EMV de Lange (2002, chap 8)
- Ces estimateurs sont utilisés en pratique dans tous les logiciels courants en génétique
- Les EMV de Lange ont de belles propriétés asymptotiques
- Non explicites et très exigeants au niveau du calcul
- Pour un r grand ($r > 30$), la méthode de Lange ne produit pas les estimateurs des CV

Estimation des CV Σ_g et Σ_e

- Nous avons proposé une nouvelle méthode d'ANOVA qui estime des CV pour un nombre quelconque de traits et pour des familles de taille large (Oualkacha et al. 2012)
- Les estimateurs d'ANOVA adaptés sont :
  - Explicites
  - Faciles à calculer pour r quelconque et pour des pedigrees quelconques

Estimation des CV Σ_g et Σ_e

- Par une étude de simulation, nous avons montré que nos estimateurs ont des propriétés asymptotiques similaires à ceux de Lange
  - Les critères de comparaison :

$$R(\hat{\Sigma}, \Sigma) = tr\left\{\boldsymbol{E}\left[(\hat{\Sigma} - \Sigma)^2\right]\right\} \qquad cb(\hat{\Sigma}, \Sigma) = tr\left\{\left[\boldsymbol{E}(\hat{\Sigma}) - \Sigma\right]^2\right\}$$

- Dans le cas de grand nombre de traits, les estimateurs de Lange ne sont pas calculables
- Nos estimateurs
  - ont une forme explicite simple (toujours calculables) et
  - Les estimateurs d'ANOVA performent bien, à l'exception de sérieux cas de données non balancées

APPLICATIONS

Schizophrénie et maladie bipolaire (en collaboration avec M. Maziade et M. A. Roy, CRULRG)

- 82 traits quantitatifs (symptômes) collectés chez 48 familles atteintes de la Schizophrénie et de la maladie bipolaire dans la province de Québec (CRULRG)
- BP est principalement caractérisée par des changements d'humeur extrêmes
- SZ est caractérisée plus particulièrement par les symptômes de délire et d'hallucination
- Les 82 symptômes peuvent êtres regroupés en 11 Dimensions : Délire (15 traits), Hallucination (6), Comportement bizarre (5), Anhédonie (5), Apathie (4), Catatonie (6), Retrait affectif (8), Pensée désorganisée (7), Alogie (4), Manie (8), Dépression (14)

Schizophrénie et maladie bipolaire (en collaboration avec M. Maziade et M. A. Roy, CRULRG)

TABLE : Estimation de la première CPH et de son héritabilité

	$CPH1$
Délire	0.160
Halluc	0.276
C. bizarre	0.209
Anhédonie	0.531
Apathie	0.419
Catatonie	0.077
R. affectif	0.388
P. désorganisée	0.141
Alogie	0.221
Manie	-0.348
Dépression	-0.223
héritabilité	0.925

- La première CPH discrimine les sujets en deux groupes : manie-dépression versus autres symptômes
- Ceci est intéressant puisque la manie et la dépression caractérisent la maladie BP
- La première CPH a une large héritabilité
- CPH1 peut être considérée comme un nouveau phénotype

Future application : Traits quantitatifs liés à l'obésité (en collaboration avec J. C. Engert, McGill)

- Vingtaine de traits quantitatifs liés à l'obésité sont collectés chez 431 familles, de la ville de Québec et la région métropolitaine, impliquées dans une étude sur les facteurs d'obésité
- L'indice de masse corporelle, le poids, la masse grasse, la masse sans gras, et les plis de la peau mesurés dans différents endroits du corps, ont tous montré une association (corrélation) significative avec des marqueurs génétiques situés sur le gène "FTO"
- Les analyses génétiques qui ont montré de telles associations ont traité chaque phénotype séparément
- Les phénotypes sont très corrélés entre eux

Objectif :

Nous voulons représenter les traits liés à l'obésité par un nouveau phénotype optimal (au sens de l'héritabilité)

Merci infiniment Louis-Paul pour tout ! ! ! !

Références

- Mardia KV, Kent JT, Bibby JM (1979). *Multivariate Analysis*. Academic Press, London.
- Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung BG (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430 (7001), 743-747.
- Ott J, Rabinowitz D (1999). A principal-components approach based on heritability for combining phenotype information. *Human Heredity*, 49, 106-111.
- Oualkacha K, Labbe A, Ciampi A, Roy MA and Maziade M (2012). Principal components of heritability for high dimension quantitative traits and general pedigrees. *Journal of Statistical Applications in Genetics and Molecular Biology*, Volume 11. Issue 2, Article 4.
- Wang Y, Fang Y, Jin M (2007). A ridge penalized principal-components approach based on heritability for high-dimensional data. *Human Heredity*, 64, 182-191.