

Contrôle de l'erreur de première espèce. Approche de Müller et al. V2

Théorie (de Müller et al)

1) Cas d'un locus biallélique, pas d'effet de nuisance

On suppose un locus (q) avec les allèles A_q et B_q , d'effets additifs. Les effets des génotypes seront $+\alpha_q, 0$ et $-\alpha_q$ pour A_qA_q, A_qB_q et B_qB_q . Ici donc, α_q est un effet allélique.

Le caractère est décrit par le modèle

$$y = W_q \alpha_q + e$$

Où W_q la matrice d'incidence décrivant le locus q . Elle est composée de +1, 0 et -1

e est la résiduelle, supposée multinormale de variance $V = A\sigma_a^2 + I\sigma_e^2$.

On suppose les variances σ_a^2 et σ_e^2 connues. La matrice A est établie à partir du pedigree. Elle peut être remplacée par la matrice génomique G .

En chaque locus, on estime l'effet α_q par $\hat{\alpha}_q = (W_q^T V^{-1} W_q)^{-1} W_q^T V^{-1} y$ et on fait un test de Wald qui dans ce cas s'écrit simplement : $T_q = \hat{\alpha}_q^2 / \text{var}(\hat{\alpha}_q)$, avec $\text{var}(\hat{\alpha}_q) = (W_q^T V^{-1} W_q)^{-1}$. Cette statistique teste l'hypothèse que $\alpha_q = 0$.

Les tests aux différents points $q = 1 \dots Z$ sont corrélés, ce qui fait que l'erreur ne peut pas être valablement contrôlée par une correction de Bonferroni, trop conservatrice. Il faut considérer le processus statistique $T = (T_1 \dots T_q \dots T_Z)$, en chercher la loi et en déduire le seuil de rejet de H_0 : aucun des $\alpha_q = 0$.

Ce processus est engendré par les variables aléatoires $\hat{\alpha} = (\hat{\alpha}_1 \dots \hat{\alpha}_q \dots \hat{\alpha}_Z)$. Sous H_0 ces variables sont distribuées dans une gaussienne d'espérance nulle et de variance :

$$\text{var}(\hat{\alpha}) = \begin{pmatrix} M_{11} & M_{12} & \dots & M_{1Z} \\ M_{21} & M_{22} & \dots & M_{2Z} \\ \vdots & \vdots & \ddots & \vdots \\ M_{Z1} & M_{Z2} & \dots & M_{ZZ} \end{pmatrix}$$

Avec $M_{qq'} = (W_q^T V^{-1} W_q)^{-1} (W_q^T V^{-1} W_{q'}) (W_{q'}^T V^{-1} W_{q'})^{-1}$

On fait une triangularisation de la matrice de covariances : $\text{var}(\hat{\alpha}) = U F U^T$

La distribution du processus est obtenue par simulations :

Pour $sim = 1 \dots S$,

on génère $u_{sim} = (u_{1sim} \dots u_{qsim} \dots u_{Zsim})$ dans une multinormale centrée réduite

on en déduit un vecteur d'estimation d'effet par transformation $\hat{\alpha}_{sim} = U \sqrt{F} u_{sim}$

On calcule le vecteur $T_{sim} = (T_{1sim} \dots T_{qsim} \dots T_{Zsim})$ dont les éléments sont les statistiques de test $T_{qsim} = \hat{\alpha}_{qsim}^2 / \text{var}(\hat{\alpha}_q)$

$$T_q = \hat{\beta}_q^T H_q^T \left(H_q (X_q^T V^{-1} X_q)^{-1} H_q^T \right)^{-1} H_q \hat{\beta}_q$$

Avec $\hat{\beta}_q = (X_q^T V^{-1} X_q)^{-1} X_q^T V^{-1} y$.

Ici, la statistique de test suit une loi de $\chi_{r_q}^2$

La matrice $M_{qq'}$ devient

$$M_{qq'} = H_q (X_q^T V^{-1} X_q)^{-1} (X_q^T V^{-1} X_{q'}) (X_{q'}^T V^{-1} X_{q'})^{-1} H_{q'}^T$$

Comme r_q varie entre locus, on considère en chaque point la P-value $p_{qsim} = \text{prob}(\chi_{r_q}^2 > T_{qsim})$, et on retient $p_{infsim} = \inf(p_{1sim} \cdots p_{qsim} \cdots p_{Zsim})$

4) Cas d'un locus multiallélique, avec des effets de nuisance et des données manquantes

Dans le cas où certains individus n'ont pas de génotypes pour toutes les positions testées, on réduit en chaque position l'analyse aux seuls individus possédant un génotype. Les auteurs introduisent une matrice D_q qui extrait de y les informations utiles : $y_q = D_q y$. La matrice de variance restreinte $V_{qq} = D_q V D_q^T$ remplace V dans la procédure. On notera notamment que

$$\hat{\beta}_q = (X_q^T V_{qq}^{-1} X_q)^{-1} X_q^T V_{qq}^{-1} y_q$$

$$T_q = \hat{\beta}_q^T H_q^T \left(H_q (X_q^T V_{qq}^{-1} X_q)^{-1} H_q^T \right)^{-1} H_q \hat{\beta}_q$$

Où, dans le cas biallélique

$$T_q = \hat{\alpha}_q^2 / \left(H_q (X_q^T V_{qq}^{-1} X_q)^{-1} H_q^T \right)$$

$$M_{qq'} = H_q (X_q^T V_{qq}^{-1} X_q)^{-1} (X_q^T V_{qq}^{-1} V_{qq'} V_{q'q'}^{-1} X_{q'}) (X_{q'}^T V_{q'q'}^{-1} X_{q'})^{-1} H_{q'}^T$$

7. On calcule et stocke son inverse $G_{qq} = F_{qq}^{-1}$ (i.e. $(X_q^T V_{qq}^{-1} X_q)^{-1}$ de dimension $n_q \times n_q$). On peut être amené à utiliser une inverse généralisée $G_{qq} = F_{qq}^{-}$
8. On construit $M_{qq} = H_q G_{qq} H_q^T$ où $H_q = (\mathbf{0}, I_{n_{\alpha}(q)}, -\mathbf{1}_{n_{\alpha}(q)})$ est la matrice des contraintes pour réduire le modèle complet au modèle sans effet QTL.
9. Calcul de la statistique de test
 - 9.1. On calcule $RHS_q = K_{qq} y_q$
 - 9.2. On en déduit $\hat{\beta}_q = G_{qq} RHS_q$
 - 9.3. On calcule $\hat{\gamma}_q = H_q \hat{\beta}_q$
 - 9.4. On inverse M_{qq}
 - 9.5. On en déduit la statistique $T_q = \hat{\gamma}_q^T M_{qq}^{-1} \hat{\gamma}_q$

Etape 3 : calcul des éléments $M_{qq'}$

Pour $q = 1 \cdots Z - 1$, et $q' = q \cdots Z$

1. On construit $V_{qq'} = D_q V D_{q'}^T$ (les éléments de V correspondant aux listes L_q en ligne et $L_{q'}$ en colonne)
2. On calcule $G_{qq'} = K_{qq} V_{qq'} K_{q'q'} (= X_q^T V_{qq}^{-1} V_{qq'} V_{q'q'}^{-1} X_{q'})$
3. On calcule et stocke $M_{qq'}$, qui est l'élément $H_q G_{qq'} F_{q'q'} G_{q'q'}^T H_{q'}^T$

Etape 4 : Construction de la matrice de transformation

1. On construit la matrice des covariances des estimateurs qui compile les éléments $M_{qq'}$:

$$var(\hat{\gamma}) = \begin{pmatrix} M_{11} & \cdots & M_{1Z} \\ \vdots & \ddots & \vdots \\ M_{Z1} & \cdots & M_{ZZ} \end{pmatrix} = M$$
2. On estime le rang de M , soit r_M
3. On fait une décomposition spectrale de la matrice des covariances : $M = U F U^T$, F étant une matrice diagonale dont les r_M premiers éléments sont non nuls
4. On forme la matrice P à partir des r_M premières colonnes de $U \sqrt{F}$. On a $M = P P^T$

Etape 5 : Simulation des seuils

1. Pour $s = 1 \cdots S$,
 - 1.1. on génère $u_s = (u_{1s} \cdots u_{qs} \cdots u_{Zs})$ dans une multinormale centrée réduite
 - 1.2. on calcule $\hat{\gamma}_s = P u_s$
 - 1.3. pour $q = 1 \cdots Z$, on calcule $T_{sq} = \hat{\gamma}_{sq}^T M_{qq}^{-1} \hat{\gamma}_{sq}$
 - 1.4. on convertit chaque T_{sq} en P-value p_{sq} associée à un $\chi_{n_{\alpha}(q)}^2$ ($p_{sq} = prob(\chi_{n_{\alpha}(q)}^2 > T_{sq})$)
 - 1.5. On retient p_{smin} le minimum du vecteur des p_{sq}
2. On trie les p_{smin}
3. On affiche les quantiles de la distribution des p_{smin}