05-2015-10-18-unified-kernel-statistcs-add-other-optimizations

I began by following gitboxes/b-blueb/trunk/docum/2015/grits/b-2015-05-11-tracking-down-independent-independent/19-NL-and-GBPvals051715.pdf. Now I am following 38-NL&GBPvals062415.pdf and 38-NL&GBPvals062415Tbl1.pdf

After moving to "/Users/ggong/aaa-ebony/43-gritsr2/gritsr2/gails-stuff/b-examples/02-v-9010/a-docum" I am following /Volumes/2015a-stanford/aaa/gitboxes/b-blueb/trunk/docum/2015/grits/e-2015-10-13-grits-competitor/03-kernelPvals091815.pdf

/Users/ggong/aaa-ebony/43-gritsr2/gritsr2/gails-stuff/b-examples/02-v-9010/a-docum

Created "2015-10-18 13:51:45 PDT" by copying from /Users/ggong/aaa-ebony/43-gritsr2/gritsr2/gails-stuff/b-examples/01-v-9009/f-docum/49-2015-07-08-unified-kernel-statistcs-add-lee.Rmd

saved and latexed: 2015-10-20 07:48:06

1 Notation dictionary

I changed the notation from p to π and V_G to Γ, so these no longer match my programs. However I kept J equal to the dimension of Z or $\mathcal{Z}$ because I avoid the letter L in my programs, and I kept Y_1 and Y_2 because the formulas are much easier to write down this way rather than y and p because I can use the index $k = 1, 2$. In this document, any vestiges of V without a subscript is probably meant to be Γ and p is probably meant to be π.

"2015-10-20 07:30:11 PDT" I kept Y_1 and Y_2 because y and p are not easily searchable.

R program	this document	Alice
not used	J	L
y_1	Y_1	y
y_2	Y_2	p
N_{cases}	N_{cases}	N_1
z_{standard}	Z^{standard}	$z^{(1)}$
z_{altern}	Z^{altern}	$z^{(2)}$
z_{optim}	Z^{optim}	$z^{(3)}$
p	π	π
V_G	Γ	Γ

2 davies_fn

Given the J dimensional random vector (of functions of genotypes) $Z \sim N\Big(\mu_Z, V_Z\Big)$ and the matrix A, the p-value for the statistic $Q = Z^T A Z$ can be gotten from `davies_fn(zzz, mu_z, V_z, AAA)`

This function performs the following calculations

$\tilde{Z} = V_Z^{-1/2} Z \sim N\Big(\mu_{\tilde{Z}} = V_Z^{-1/2}\mu_Z, I\Big)$.

$Q = Z^T A Z = Z^T V^{-1/2} V^{1/2} A V^{1/2} V^{-1/2} Z = \tilde{Z}^T \tilde{A} \tilde{Z}$

$\tilde{A} = V^{1/2} A V^{1/2}$

$\tilde{A} = U^T \Lambda U$. The spectral decomposition of $\tilde{A}$, so Λ is a diagonal matrix containing the eigenvalues of $\tilde{A}$, and $U^T U = U U^T = I$.

$Q = \tilde{Z}^T \tilde{A} \tilde{Z} = \tilde{Z}^T U^T \Lambda U \tilde{Z} = X^T \Lambda X$

$X = U\tilde{Z} \sim N\Big(\mu_X = U\mu_{\tilde{Z}}, UIU^T = I)$

$Q = X^T \Lambda X = \sum_{j=1}^{J} \lambda_j X_j^2$ is a sum of independent χ^2 random variables with noncentrality parameters μ_X^2.

The R package `CompQuadForm` provides the function `davies` which computes $P(Q > q)$ where $Q = \sum_{j=1}^{J} \lambda_j X_j + \sigma X_0$ where X_j are independent randomvariables having a non-central χ^2 distribution with n_j degrees of freedom and noncentrality parameter δ_j^2, and X_0 having a standard normal distribution.

In our case, the degrees of freedom is $n_j = 1$, and the noncentrality parameter is $\delta_j^2 = \mu_X^2$.

3 The genotype matrix

G is the genotype matrix.

$E\Big(G_{nm}\Big) = 2\pi_m$

$\Gamma = \mathrm{cov}(G_n)$, a covariance matrix of dimension $M \times M$. I can calculate the empirical covariance matrix by `cov(genotype)`.

Ψ is 2 times the kinship matrix. $\Psi_{n_1\, n_2} = \mathrm{Cor}\Big(G_{n_1\, m}, G_{n_2\, m}\Big)$ while holding m equal to any number in $1, \cdots, M$.

$\mathrm{Cov}\Big(G_{n_1 m_1}, G_{n_2 m_2}\Big) = \Gamma_{m_1 m_2} \Psi_{n_1 n_2}$

$W = W_{M\times M}$ is the diagonal matrix of weights.

4 Y_1, Y_2, $N_{\textbf{cases}}$ and e

Y_1 is a vector of dimension N, the n-th element being the indicator for disease of the n-th person.

N_{cases} is the number of cases.

Y_2 is a vector of outcome predictors. By design, each element of this vector is equal to N_{cases}/N.

$e = Y_1 - Y_2$ is the vector of residuals.

5 Lemma 1 Various Z and their vitals

5.1 Lemma 1a: $Z^{\textbf{standard}} = WG^Te$

$Z = Z^{\text{standard}}$ is a random vector of length M.

$E(Z) = E\Big(Z_1 - Z_2\Big) = 0_M$

$z_m = \sum_n w_m G_{nm} e_n$

$\text{Cov}\Big(z_{m_1}, z_{m_2}\Big) = \text{Cov}\Big(\sum_{n_1} w_{m_1} G_{n_1 m_1} e_{n_1}, \sum_{n_2} w_{m_2} G_{n_2 m_2} e_{n_2}\Big) = \sum_{n_1}\sum_{n_2} e_{n_1} e_{n_2} \text{Cov}\Big(G_{n_1 m_1}, G_{n_2 m_2}\Big) w_{m_1} w_{m_2}$
$= \sum_{n_1}\sum_{n_2} e_{n_1} e_{n_2} \Gamma_{m_1 m_2} \Psi_{n_1 n_2} w_{m_1} w_{m_2} = e^T \Psi e \;\; w_{m_1} \Gamma_{m_1 m_2} w_{m_2}$

$\text{Cov}(Z) = e^T \Psi e \; W \Gamma W$

5.2 Lemma 1b: $Z^{\textbf{altern}} = \textbf{rbind}\Big(WG^TY_1, WG^TY_2\Big)$

For $k = 1, 2,$ $Z_k = WG^TY_k$ is a random vector of length M with $E\Big(Z_k\Big) = 2N_{cases} W\pi$.

$Z = Z^{\text{altern}} = \text{rbind}\Big(Z_1, Z_2\Big)$ is a random vector of length $2M$.

$E(Z) = \text{rbind}\Big(2N_{cases} W\pi, 2N_{cases} W\pi\Big)$

$z_{km} = \sum_n w_m G_{nm} y_{kn}$

$\text{Cov}\Big(z_{k_1\, m_1}, z_{k_2\, m_2}\Big) = \text{Cov}\Big(w_{m_1} \sum_{n_1} y_{k_1\, n_1} G_{n_1\, m_1}, w_{m_2} \sum_{n_2} y_{k_2\, n_2} G_{n_2\, m_2}\Big)$

$= \quad w_{m_1} w_{m_2} \sum_{n_1}\sum_{n_2} y_{k_1\, n_1} y_{k_2\, n_2} \text{Cov}\Big(G_{n_1\, m_1}, G_{n_2\, m_2}\Big) \quad = \quad w_{m_1} w_{m_2} \sum_{n_1}\sum_{n_2} y_{k_1\, n_1} y_{k_2\, n_2} \Gamma_{m_1 m_2} \Psi_{n_1 n_2}$
$= w_{m_1} \Gamma_{m_1 m_2} w_{m_2} Y_{k_1}^T \Psi Y_{k_2}$

$\text{Cov}\Big(Z_{k_1}, Z_{k_2}\Big) = Y_{k_1}^T \Psi Y_{k_2} \; W \Gamma W$

$\text{Cov}(Z) = \text{kronecker}\Big(\Phi, W\Gamma W\Big)$ with $\Phi_{k_1 k_2} = Y_{k_1}^T \Psi Y_{k_2}$

5.3 Lemma 1c: $Z^{\textbf{optim}} = \textbf{rbind}\Big(Z^{\textbf{standard}}, Z^{\textbf{altern}}\Big)$

$U = \text{rbind}\Big(U_1, U_2, U_3\Big) = \text{rbind}\Big(e, Y_1, Y_2\Big)$

$Z = Z^{\text{optim}} = \text{rbind}\Big(Z_1, Z_2, Z_3\Big) = WG^TU$ is a random vector of length $3M$.

$E(Z) = \text{rbind}\Big(0_M, 2N_{cases} W\pi, 2N_{cases} W\pi\Big)$

$\text{Cov}\Big(Z_{k_1}, Z_{k_2}\Big) = U_{k_1}^T \Psi U_{k_2} W \Gamma W$

$\text{Cov}(Z) = \text{kronecker}\Big(\Omega, W\Gamma W\Big)$ with $\Omega_{k_1 k_2} = U_{k_1}^T \Psi U_{k_2}$

6 The burden statistics

$Z = Z^{\text{standard}}$

The burden statistic is $Z^T 1$ if we want a two-tailed test or $Z^T 1 1^T Z$ if we are interested in a one-tailed test.

Davies useful things about the squared burden statistic (burd)

$A = A^{\text{standard}}_{\text{burden}} = 1_M 1_M^T$

Using Lemma 1a . . .

$J = M$

$\mu_Z = E(Z) = 0$

$V_Z = \text{cov}(Z) = e^T \Psi e \, W \Gamma W$

The burden normal statistic (bnorm)

$Z \sim \mathcal{N}\Big(\mu_Z = 0, V_Z = e^T \Psi e W \Gamma W\Big)$

`burd_normal` $= 1^T Z \sim \mathcal{N}\Big(1^T \mu_Z = 0, 1^T V_Z 1 = e^T \Psi e 1^T W V W 1 = e^T \Psi e w^T \Gamma w\Big)$

7 The skat (linear kernel) statistic

$H = G W^2 G^T$

$T = \Big(Y_1 - Y_2\Big)\Big(Y_1 - Y_2\Big)^T$

$Q = \Big(Y_1 - Y_2\Big)^T H \Big(Y_1 - Y_2\Big)$ the linear kernel in its familiar form

$Q = \sum_{n_1} \sum_{n_2} \Big(Y_{1\,n_1} - Y_{2\,n_1}\Big) H_{n_1 n_2} \Big(Y_{1\,n_2} - Y_{2\,n_2}\Big) = \sum_{n_1} \sum_{n_2} H_{n_1 n_2} T_{n_1, n_2} = 1^T \Big(H * T\Big) 1.$ This shows the two expressions coincide.

Also we can write

$Q_{\text{lin}} = \Big(Y_1 - Y_2\Big)^T H \Big(Y_1 - Y_2\Big) = \Big(Y_1 - Y_2\Big)^T G W^2 G^T \Big(Y_1 - Y_2\Big) = \bigg(W G^T \Big(Y_1 - Y_2\Big)\bigg)^T \bigg(W G^T \Big(Y_1 - Y_2\Big)\bigg)$
$= \Big(Z_1 - Z_2\Big)^T \Big(Z_1 - Z_2\Big) = Z^T A Z$

Davies useful things about skat (lin)

$Z = Z^{\text{standard}}$

$A = A^{\text{standard}}_{\text{skat}} = I_M$

J, μ_Z, V_Z are the same as in the squared burden statistic.

8 The alternative burden statistic

$Z = Z^{\text{altern}}$

$A = A^{\text{altern}}_{\text{burden}} = \text{rbind}\Big\{\text{cbind}\Big\{A^{\text{standard}}_{\text{burden}}, 0_{M\times M}\Big\}, \text{cbind}\Big\{0_{M\times M}, -A^{\text{standard}}_{\text{burden}}\Big\}\Big\}$

Using Lemma 1b ...

$J = 2M$

$\mu_Z = E(Z) = \text{rbind}\Big(2N_{cases}W\pi, 2N_{cases}W\pi\Big)$

$V_Z = \text{cov}(Z) = \text{kronecker}\Big(\Phi, W\Gamma W\Big)$

9 The alternative skat statistic (newl)

$H = GW^2G^T$

$T = Y_1Y_1^T - Y_2Y_2^T$

$Q = 1^TK1 = 1^T(H * T)1 = 1^T(H * Y_1Y_1^T)1 - 1^T(H * Y_2Y_2^T)1 = \sum_{n_1n_2} H_{n_1n_2}Y_{1n_1}Y_{1n_2} - \sum_{n_1n_2} H_{n_1n_2}Y_{2n_1}Y_{2n_2} = Y_1^THY_1 - Y_2^THY_2 = Y_1^TGW^2G^TY_1 - Y_2^TGW^2G^TY_2 = Z_1^TZ_1 - Z_2^TZ_2$

Davies useful things about contrast skat (newl)

$Z = Z^{\text{altern}}$

$A = A^{\text{altern}}_{\text{skat}} = \text{rbind}\Big\{\text{cbind}\Big\{A^{\text{standard}}_{\text{skat}}, 0_{M\times M}\Big\}, \text{cbind}\Big\{0_{M\times M}, -A^{\text{standard}}_{\text{skat}}\Big\}\Big\}$

J, μ_Z, V_Z are the same as in the alternative burden.

10 The lee statistic

$Z = Z^{\text{standard}}$

$Q_\rho = Z^TA_\rho Z$

$A_\rho = (1-\rho)A^{\text{standard}}_{\text{burden}} + \rho A^{\text{standard}}_{\text{skat}}$

For ρ_l in the sequence $0 = \rho_1 < \cdots < \rho_L = 1$, compute Q_{ρ_l} and its p-value p_l. Use `davies_fn`.

Define the lee statistic $Q = \min_{l=1}^{L} p_l$. What is the distribution of Q. Use a simulation and then if feasible, use a gritsr to get the answer.

11 The alternative lee statistic

$Z = Z^{\text{altern}}$

$Q_\rho = Z^TA_\rho Z$

$A_\rho = (1-\rho)A^{\text{altern}}_{\text{burden}} + \rho A^{\text{altern}}_{\text{skat}}$

For ρ_l in the sequence $0 = \rho_1 < \cdots < \rho_L = 1$, compute Q_{ρ_l} and its p-value p_l. Use `davies_fn`.

Define the lee statistic $Q = \min_{l=1}^{L} p_l$. What is the distribution of Q. Use a simulation and then if feasible, use a gritsr to get the answer.

12 Optimized burden statistic

$Z = Z^{\text{optim}} = WG^T U$ with $U = \text{rbind}\Big(U_1, U_2, U_3\Big) = \text{rbind}\Big(e, Y_1, Y_2\Big)$

$A = \text{rbind}\Big(\text{cbind}\Big((1-\tau)A^{\text{standard}}_{\text{burden}}, 0_{M\times 2M}\Big), \text{cbind}\Big(0_{2M\times M}, \tau A^{\text{alternative}}_{\text{burden}}\Big)\Big)$

Davies useful things

$J = 3M$

$\mu_Z = \text{rbind}\Big(0_M, 2N_{cases}W\pi, 2N_{cases}W\pi\Big)$

$V_Z = \text{Cov}(Z) = \text{kronecker}\Big(\Omega, W\Gamma W\Big)$ with $\Omega_{k_1 k_2} = U^T_{k_1}\Psi U_{k_2}$

13 Optimized skat statistic

$Z = Z^{\text{optim}} = WG^T U$ with $U = \text{rbind}\Big(U_1, U_2, U_3\Big) = \text{rbind}\Big(e, Y_1, Y_2\Big)$

$A = \text{rbind}\Big(\text{cbind}\Big((1-\tau)A^{\text{standard}}_{\text{skat}}, 0_{M\times 2M}\Big), \text{cbind}\Big(0_{2M\times M}, \tau A^{\text{alternative}}_{\text{skat}}\Big)\Big)$

Davies useful things Same as in optimizied burden statistic

14 Optimized statistic

$Z = Z^{\text{optim}} = WG^T U$ with $U = \text{rbind}\Big(U_1, U_2, U_3\Big) = \text{rbind}\Big(e, Y_1, Y_2\Big)$

$A = \text{rbind}\Big\{\text{cbind}\Big\{$

$(1-\tau)\Big\{\Big(1-\rho_{\text{standard}}\Big)A^{\text{standard}}_{\text{burden}} + \rho_{\text{standard}}A^{\text{standard}}_{\text{skat}}\Big\}, 0_{M\times 2M}$

$\Big\}, \Big\{$

$0_{2M\times M}, \tau\Big\{\Big(1-\rho_{\text{altern}}\Big)A^{\text{altern}}_{\text{burden}} + \rho_{\text{altern}}A^{\text{altern}}_{\text{skat}}\Big\}$

$\Big\}\Big\}$

Davies useful things Same as in optimizied burden statistic

THE END