

December 10, 2015

Details in development of test statistic

Additional File 1 to: Analysing multiple types of molecular profiles
simultaneously: connecting the needles in the haystack

Menezes RX, Mohammadi L, Goeman JJ and Boer JM

1 Introduction

The global test statistic proposed by Goeman *et al.* [1] for testing associations between a response and a set of covariates used a test statistic proposed elsewhere [2], for testing in the context of random effects models. In [2] the authors get to their test statistic by first obtaining an approximation of the likelihood function using a truncated Taylor series expansion, and then computing the score test statistic. The advantage this yielded was partly due to the fact that the approximation was calculated around points where the random effects r_i were equal to zero. Since under H_0 it was assumed that $E(r_i) = 0$, some simplification followed.

Here we will follow on the footsteps of Le Cessie and van Houwelingen [2] to obtain a generalization of the global test statistic, for the case where two sets of covariates are involved. We will later on show that an extension to considering any fixed number of sets of covariates is straightforward.

2 Generalization: the global test for two sets of covariates

Consider the model

$$E(Y_i) = h^{-1}(\alpha + r_i), \quad i = 1, \dots, I,$$

for I samples, where the random effect r_i is here defined as a sum of random effects

$$r_i = X_i\beta + Z_i\gamma, \quad i = 1, \dots, I, \quad \mathbf{r} = (r_1, \dots, r_I)^t,$$

with $E(\beta) = E(\gamma) = 0$, $\text{Cov}(\beta) = \sigma^2 I$, $\text{Cov}(\gamma) = \tau^2 I$ and $\text{Cov}(\beta, \gamma) = 0$. These assumptions imply that $E(r_i) = 0$ and $\text{Cov}(r_i) = \sigma^2 R + \tau^2 S$, where we define $R \equiv XX^t$, $S \equiv ZZ^t$, and X, Z as being the matrices with the i th row represented by X_i, Z_i . Note that this means β, γ are $p \times 1, q \times 1$ vectors and X, Z are $n \times p, n \times q$ matrices, respectively.

The likelihood function for all $\{Y_i\}$ can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma, \tau) = \mathbb{E}_r \left[\prod_{i=1}^I f_i(Y_i | r_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma, \tau) \right], \quad (1)$$

with f_i representing the density of Y_i , given the random effect r_i . Le Cessie and van Houwelingen [2] obtained the following approximation:

$$\begin{aligned} \prod_{i=1}^I f_i(r_i) = & \prod_{i=1}^I f_i(0) \left\{ 1 + \sum_{i=1}^I [r_i u_i^1(0) + r_i^2 ((u_i^1(0))^2 + u_i^2(0)) / 2] \right. \\ & \left. + (1/2) \sum_{i,j \neq i} r_i r_j u_i^1(0) u_j^1(0) \right\} + o(\mathbf{r}^2), \end{aligned} \quad (2)$$

where $u_i^k(r_i) \equiv \partial^k \log f_i(r_i) / \partial r_i^k$, for $k = 1, 2, \dots$. The expected value of (2) in r leads to an approximation to the likelihood (1), which can be written as

$$\begin{aligned} & \left[\prod_{i=1}^I f_i(0) \right] \left\{ 1 + \frac{1}{2} \sum_{i=1}^I (\sigma^2 R_{ii} + \tau^2 S_{ii}) u_i^2(0) \right. \\ & \left. + \frac{1}{2} \sum_{i=1}^I \sum_{j=1}^I (\sigma^2 R_{ij} + \tau^2 S_{ij}) u_i^1(0) u_j^1(0) \right\} + o(\sigma^2) + o(\tau^2). \end{aligned} \quad (3)$$

For simplicity, let us represent $\boldsymbol{\theta} = (\theta_1, \theta_2) \equiv (\sigma, \tau)$. We would like to test the hypotheses

$$H_0 : \boldsymbol{\theta} = (0, 0) \equiv \boldsymbol{\theta}_0 \text{ against } H_a : \theta_j > 0, \text{ for at least one } j, j = 1, 2,$$

for at least one j , $j = 1, 2$. Such a test statistic can be derived for example from the inner product of the score vector

$$\mathbf{U}(\boldsymbol{\theta}_0)^t \mathbf{U}(\boldsymbol{\theta}_0), \quad (4)$$

where $\mathbf{U}(\boldsymbol{\theta}_0)$ represents the score vector, given by

$$\mathbf{U}(\boldsymbol{\theta}_0) = \left[\frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \Big|_{H_0}, \frac{\partial}{\partial \tau^2} \log L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \Big|_{H_0} \right].$$

From (3) we have that, under H_0 , the likelihood is reduced to

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})|_{H_0} = \prod_{i=1}^I f_i(0).$$

Also using (3) we can write

$$\left. \frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \right|_{H_0} = \sum_{i=1}^I R_{ii} u_i^2(0) + \sum_{i=1}^I \sum_{j=1}^I R_{ij} u_i^1(0) u_j^1(0). \quad (5)$$

If we assume that $Y_i \sim \mathcal{G}$ where $\mathcal{G}$ is a distribution in the exponential family with canonical link, we can write its density function as

$$f(y; \delta, \phi) = \exp \{ [y\delta - b(\delta)] / a(\phi) + c(y, \phi) \},$$

where δ represents the parameter of interest and ϕ is a known nuisance parameter, with $a(\cdot)$ known. Then

$$\begin{aligned} \frac{\partial}{\partial r_i} \log f(y; \delta, \phi) &= \frac{y_i - \mu_{1i}(r_i)}{a(\phi)} \equiv u_i^1(r_i) \\ \frac{\partial^2}{\partial r_i^2} \log f(y; \delta, \phi) &= \frac{-\mu_{2i}(r_i)}{a(\phi)^2} \equiv u_i^2(r_i), \end{aligned}$$

where $\mu_{1i}(r_i) = \partial b(\delta) / \partial r_i$ and $\mu_{2i}(r_i) = a(\phi) \partial^2 b(\delta) / \partial r_i^2$. Replacing these expressions into (5), we can write

$$\begin{aligned} \left. \frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \right|_{H_0} &= \frac{1}{2a(\phi)^2} \left\{ (Y - \mu_1)^t R (Y - \mu_1) - \text{tr}(RV) \right\} \\ &\equiv \frac{1}{2a(\phi)^2} \{Q(R) - \text{tr}(RV)\}, \end{aligned} \quad (6)$$

where we define $Q(A) \equiv (Y - \mu_1)^t A (Y - \mu_1)$ for a square matrix A with as many rows as the number I of entries in Y , $V = \text{diag}\{\mu_{2i}(r_i)\}$ and μ_1 is a $I \times 1$ vector with i th entry given by μ_{1i} . Following the same steps, we obtain a similar expression for the partial derivative with respect to τ^2 :

$$\begin{aligned} \left. \frac{\partial}{\partial \tau^2} \log L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \right|_{H_0} &= \frac{1}{2a(\phi)^2} \left\{ (Y - \mu_1)^t S (Y - \mu_1) - \text{tr}(SV) \right\} \\ &\equiv \frac{1}{2a(\phi)^2} \{Q(S) - \text{tr}(SV)\}, \end{aligned} \quad (7)$$

where here $Q(S) \equiv (Y - \mu_1)^t S (Y - \mu_1)$.

Note that the expressions (6,7) are the entries in $\mathbf{U}(\boldsymbol{\theta})$. We have $E[Q(R)] = \text{tr}(RV)$ and, equivalently, $E[Q(S)] = \text{tr}(SV)$. Replacing this in (6) and (7), we can express the inner product of the score vector as

$$\{Q(R) - E[Q(R)]\}^2 + \{Q(S) - E[Q(S)]\}^2,$$

up to a constant. This is the test statistic sought.

3 Considering any K number of sets of covariates

The extension to considering an arbitrary number K of sets of covariates is relatively straightforward, as we will show below¹. Consider the model

$$E(Y_i) = h^{-1}(\alpha + r_i), \quad i = 1, \dots, I,$$

where the random effect r_i is here defined as a sum of random effects

$$r_i = \sum_{k=1}^K X_{ki} \beta_k, \quad i = 1, \dots, I, \quad \mathbf{r} = (r_1, \dots, r_I)^t,$$

with $E(\beta_k) = 0$, $\text{Cov}(\beta_k) = \sigma_k^2 I$ and $\text{Cov}(\beta_{k1}, \beta_{k2}) = 0$ for $k = 1, \dots, K$, $k_1 \neq k_2$. These assumptions have similar implications to the 2-gene set case, namely that $E(r_i) = 0$ and $\text{Cov}(r_i) = \sum_{k=1}^K \sigma_k^2 R_k$, where we now define $R_k \equiv X_k X_k^t$, with X_k representing the matrix which i th row is represented by X_{ki} .

The likelihood function for all $\{Y_i\}$ can be written as

$$L(\beta, \gamma, \sigma, \tau) = E_r \left[\prod_{i=1}^I f_i(Y_i | r_i, \{\beta_k\}, \{\sigma_k\}) \right], \quad (8)$$

with f_i representing the density of Y_i , given the random effect r_i . Here $\prod_i f_i(r_i)$ can be approximated by (2) as before, and taking the expected value in r now leads to

$$\begin{aligned} & \left[\prod_{i=1}^I f_i(0) \right] \left\{ 1 + \frac{1}{2} \sum_{i=1}^I u_i^2(0) \sum_{k=1}^K \sigma_k^2 R_{k,ii} \right. \\ & \left. + \frac{1}{2} \sum_{i=1}^I \sum_{j=1}^I u_i^1(0) u_j^1(0) \sum_{k=1}^K \sigma_k^2 R_{k,ij} \right\} + \sum_{k=1}^K o(\sigma_k^2). \end{aligned} \quad (9)$$

Here we shall define $\boldsymbol{\theta} = (\sigma_1, \sigma_2, \dots, \sigma_K)$. The k th entry of the score vector is now

$$\frac{\partial}{\partial \sigma_k^2} \log L(\{\beta_k\}, \boldsymbol{\theta}) \Big|_{H_0}$$

Under H_0 , the likelihood is again reduced to $\prod_i f_i(0)$. In addition, similarly to (5) we now have

$$\frac{\partial}{\partial \sigma_k^2} \log L(\{\beta_k\}, \boldsymbol{\theta}) \Big|_{H_0} = \sum_{i=1}^I R_{k,ii} u_i^2(0) + \sum_{i=1}^I \sum_{j=1}^I R_{k,ij} u_i^1(0) u_j^1(0). \quad (10)$$

¹Note that in the main text we have represented the number of covariate sets by M .

By using again the fact that Y_i has a distribution $\mathcal{G}$ in the exponential family with canonical link, we get that the k th entry of the score vector $\mathbf{U}(\boldsymbol{\theta})$ is

$$\begin{aligned} \left. \frac{\partial}{\partial \sigma_k^2} \log L(\{\beta_k\}, \boldsymbol{\theta}) \right|_{H_0} &= \frac{1}{2a(\phi)^2} \left\{ (Y - \mu_1)^t R_k (Y - \mu_1) - \text{tr}(R_k V) \right\} \\ &\equiv \frac{1}{2a(\phi)^2} \{Q(R_k) - \text{tr}(R_k V)\}, \quad k = 1, \dots, K \end{aligned} \quad (11)$$

So, the inner product of the score vector is given by, up to a constant,

$$\sum_{k=1}^K \{Q(R_k) - \text{tr}(R_k V)\}^2 \equiv \sum_{k=1}^K \{Q(R_k) - E(R_k)\}^2, \quad (12)$$

which can then be used to test

$H_0 : \boldsymbol{\theta} = (0, \dots, 0) \equiv \boldsymbol{\theta}_0$ against $H_a : \theta_k > 0$, for at least one k , $k = 1, \dots, K$.

4 Alternative scaling of the test statistics

In the Methods section of the main article we suggest using the scaled individual test statistics, in order to combine effects of individual covariate sets onto the same scale on the joint test statistic. Here we give an alternative to using the scaled test statistics, by scaling and centering the covariates prior to computing the individual test statistics $Q(X)$, $Q(Z)$ and subsequently computing $Q(X, Z) \equiv Q(X) + Q(Z)$. Indeed, we could compute directly $Q(X_S)$ and $Q(Z_S)$, where $X_S \equiv Xw_X = X(I - H_X) / \text{trace}(XX^t)$ and $Z_S \equiv Zw_Z = (I - H_Z) / \text{trace}(XX^t)$. That is, we suggest using

$$Q(X, Z) = Q(Xw_X) + Q(Zw_Z) = \frac{(Y - \mu)^t [Xw_X^2 X^t + Zw_Z^2 Z^t] (Y - \mu)}{(Y - \mu)^t (Y - \mu)}, \quad (13)$$

where $w_X = [\text{trace}(XX^t)(I - H_X)]^{-1}$ and $I - H_X$ is the matrix that centers X around its column-specific means, with a similar definition for w_Z . Since $E[Q(X)] = \text{trace}(XX^t)/J$ [1], the proposed weights correspond to bringing mean and variance of $Q(X)$, $Q(Z)$ onto the same scale.

References

- [1] Goeman J, van de Geer S, de Kort F, van Houwelingen H: **A global test for groups of genes: testing association with a clinical outcome.** *Bioinformatics* 2004, **20**:93–99.
- [2] Le Cessie S, van Houwelingen H: **Testing the Fit of A Regression-Model Via Score Tests in Random Effects Models.** *Biometrics* 1995, **51**:600–614.