Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test (Michael C. Wu et al., 2011)

Xu (Rita) Shi

Presentation 2: Update

Outline

Review

- Rare variants: "rare" defined by MAF
- Association between a group of rare variants and disease

  Burden and nonburden test
  Sequence Kernel Association Test (SKAT)

- $g(E(Y)) = f(G) + X\beta$

  $E(Y) = \sum_{j=1}^{p} G_j b_j + X\beta$

- Test for $H_0 : f(G) = 0$

  $Q = (y - \hat{y}_0)^T GWG^T (y - \hat{y}_0)$, $K = GWG^T$

  $\hat{y}_0$: fitted value under H_0

  $\sim$ mixture of χ_1^2

Kernel K vs Model $f(G)$

- Linear Mixed Model (Liu 2005, 2008):
  $\boldsymbol{Y} = \boldsymbol{f(G)} + \boldsymbol{X}\beta + \boldsymbol{\epsilon}$

  Unknown $\boldsymbol{f(G)} \sim N(0, \lambda^{-1}\sigma^2\boldsymbol{K})$

  $\boldsymbol{\epsilon} \sim N(0, R = \sigma^2 I)$

- $\boldsymbol{Y} = \sum_{j=1}^{p} \boldsymbol{G_j} b_j + \boldsymbol{X}\beta + \boldsymbol{\epsilon}$

  $b_j \sim N(0, \lambda^{-1}\sigma^2 w_j)$

  $\Rightarrow \boldsymbol{f(G)} \sim N(0, \lambda^{-1}\sigma^2\boldsymbol{GWG^T})$ where $W = diag(w_j)$

Kernel K vs Model $f(G)$

- BLUP Estimator $\hat{\boldsymbol{f}}(\boldsymbol{G})$ (from $E[f(G)|y]$)

$$\begin{aligned}\hat{\boldsymbol{f}}(G) &= \lambda^{-1}\hat{\sigma}^2\boldsymbol{K}\boldsymbol{\gamma} \\ &= \text{ linear combination of } Var(f(G)) = \lambda^{-1}\hat{\sigma}^2\boldsymbol{K} \\ &= \lambda^{-1}\hat{\sigma}^2 K \ (\hat{\sigma}^2 I + \lambda^{-1}\hat{\sigma}^2 K)^{-1} \ (y - X\hat{\beta}) \\ &= \lambda^{-1} K \ (I + \lambda^{-1} K)^{-1} \ (y - X\hat{\beta})\end{aligned}$$

Notes from Biost 571:

Thus, the conditional mean (BLUP/empirical Bayes) values of γ_i are

$$\begin{aligned}\tilde{\gamma}_i &= E(\gamma_i|\,\boldsymbol{Y}_i;\boldsymbol{\beta},\boldsymbol{\alpha}) \\ &= \boldsymbol{G}\boldsymbol{Z}_i^\top\boldsymbol{\Sigma}_i(\boldsymbol{\alpha})^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\end{aligned}$$

where we plug in $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, generally from ML or REML

Chapter 3.69; Last updated February 21, 2014

Score Test

- $\boldsymbol{f(G)} \sim N(0, \lambda^{-1}\sigma^2 \boldsymbol{K})$
- Score Test (Lin 1997): $H_0 : f(G) = 0 \Leftrightarrow \theta = \lambda^{-1}\hat{\sigma}^2 = 0$

$$\boldsymbol{Y} = \boldsymbol{f(G)} + \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

$$L(\beta, \theta) = exp\{\sum_{i=1}^{n} l_i(\beta, \theta)\}$$

$$= \int_{-\infty}^{\infty} exp\{\sum_{i=1}^{n} l_i(\beta, f(G))\}\, dF(f(G)|\theta)$$

$$exp\{\sum_{i=1}^{n} l_i(\beta, f(G))\} = \text{ Taylor expansion at 0, 1st and 2nd order}$$

$$= \mathcal{L}(E(f(G)), Var(f(G)), \beta) = \mathcal{L}(\theta K, \beta)$$

Score Test

Under the null

$$\frac{\partial l(\beta,\theta)}{\partial \theta} = \frac{1}{2\sigma^4}(y - X\hat{\beta})^T K(y - X\hat{\beta}) - tr\{\frac{1}{\sigma^2}P_0 K\}$$
$$\text{where } P_0 = I - X(X^T X)^{-1}X$$

The score statistic $Q = (y - X\hat{\beta})^T K(y - X\hat{\beta}) \sim$ mixture of χ^2_1

Asymptotic distribution approximated by the Satterthwaite method (Satterthwaite 1946)

Simulation and Application

- Simulation for type I error
  1. Generate genotype data using COSI, p = 10,000
  2. Generate phenotype data using model $y = 0.5N(0,1) + 0.5Bernoulli(0.5) + N(0,1)$
  3. Fit null model to get score statistic
  4. Get asymptotic distribution: need to compute eigen value!
  5. Get p-value
  6. Empirical type I error estimated as proportion of p-values less than $\alpha = 10^{-4},\ 10^{-5},\ 10^{-6}$

Simulation and Application

- Simulation for power
  1. Generate genotype data using COSI, p = 10,000
  2. Generate phenotype data using model
     $y = 0.5N(0,1) + 0.5Bernoulli(0.5) + N(0,1) + \sum_{j=1}^{p} G_j b_j$,
     G_j are 5% of rare variantes randomly selected with MAF $< \%3$
     $|b_j| = c|log_{10} MAF_j|$, percentage of negative values varies to reflect different directions of effects
  3. Fit null model to get score statistic
  4. Get asymptotic distribution: need to compute eigen value!
  5. Get p-value
  6. Estimate empirical power
  7. Compare to (1) counting-based burden test (2) CAST (3) Weighted-sum burden test

Simulation and Application

- Application to Dallas Heart Study Data
  1. 3476 individuals, 93 variants, covariates: sex and ethnicity
  2. Get p-values from (0) SKAT (1) counting-based burden test (2) CAST (3) Weighted-sum burden test

Kernel K vs Model $f(G)$

- Least Square Kernel Machine Estimator (Liu 2005, 2008; Cristianini 2000)
- $g(E(Y)) = f(G) + X\beta$

  Unknown $f(G) \in$ function space $\mathcal{H}$

  $\mathcal{H} = Span\{\phi_1(G), ..., \phi_J(G)\}$

  $f(G) = \sum_{i=1}^{J} w_j \phi_j(G) = \boldsymbol{w^T \phi(G)}$

  $\phi_i(G)$: orthogonal basis functions

Kernel K vs Model $f(G)$

- $max_{w,\beta}\ J(w,\beta) = -\frac{1}{2}\sum_{i=1}^{n}\{y_i - x_i^T\beta - w^T\phi(G_i)\}^2 - \frac{1}{2}\lambda w^T w$
- $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$

  $V := (I + \lambda^{-1} K)$
  $K_{n\times n} = \left(\phi(G_i)^T \phi(G_j)\right)_{i,j}$

- $\hat{w} = \lambda^{-1}\{\phi(G_1), ..., \phi(G_n)\}\gamma$

  $\gamma = V^{-1}(y - X\hat{\beta})$

- $\hat{\boldsymbol{f}}(G) = \hat{w}^T\{\phi(G_1), ..., \phi(G_n)\} = \lambda^{-1}\boldsymbol{K}\gamma$

Kernel K vs Model $f(G)$

- Reproducing Kernel Hilbert Space $\mathcal{H}_K$
- Primal representation: $\hat{\boldsymbol{f}} = \hat{w}^T\{\phi(G_1), ..., \phi(G_n)\}$

  $$||f||^2_{\mathcal{H}_K} = w^T w$$

- Dual representation: $\hat{\boldsymbol{f}} = \lambda^{-1}\boldsymbol{K}\boldsymbol{\gamma} = \sum_{i=1}^{n} \lambda^{-1}\gamma_i K(G, G_i)$
- Instead of $\phi_j(G)$ representation, can use $K(\cdot, \cdot)$ representation

Thank you!

References

- Wu, M. et. al. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 92: 841-853, 2013.
- Liu D. et al. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, 2007 Dec;63(4):1079-1088.
- Lin X. Variance Component Testing in Generalised Linear Models with Random Effects. *Biometrika*, 1997 Jun;84(2):309-326.
- Liu D. et al. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 2008 Jun;9, 292.