Lecture 8: Supervised Burden Tests and Variance Component Test for Rare Variants

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2015

Lecture Overview

Recall: Region Based Analysis of Rare Variants

- Single variant test is not powerful to identify rare variant associations
- Strategy: Region based analysis
  - Test the joint effect of rare/common variants in a gene/region while adjusting for covariates.

Major Classes of Tests

- Burden/Collapsing tests
- Supervised/Adaptive Burden/Collapsing tests
- Variance component (similarity) based tests
- Omnibus tests: hedge against difference scenarios

Burden Tests So Far

- Tests
  - Binary Collapsing: CAST
  - CMC
  - Count Collapsing: MZ (GRANVIL)
  - Weighted Sum Test
- Power of burden tests depends on
  - Number of associated variants
  - Number of non-associated variants
  - Direction of the effects.
- Powerful if most variants are causal and have effects in the same direction.

Burden vs. Single Variant Test

	Single Variant Test	Combined Test
10 variants / all have risk 2 / All have frequency .005	.05	.86
10 variants / all have risk 2 / Unequal Frequencies	.20	.85
10 variants / average risk is 2, but varies / frequency .005	.11	.97

[Li and Leal (2008) AJHG]

- Power from simulated data
- Combining variants can greatly increase the power.

Burden vs. Single Variant Test

	Single Variant Test	Combined Test
10 disease associated variants	.05	.86
10 disease associated variants + 5 null variants	.04	.70
10 disease associated variants + 10 null variants	.03	.55
10 disease associated variants + 20 null variants	.03	.33

[Li and Leal (2008) AJHG]

- Null variants reduce the power.
- Existence of variants whose effects are in different directions can reduce power more substantially (Next Topic).

Burden Test: Mixed effect directions

- Lose power if variants have positive and negative effects.

Y	G_1	G_2	G_3	G_4
1	1	0	0	0
1	0	1	0	0
1	0	0	0	0
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
0	0	0	0	0
0	0	0	1	0
0	0	0	0	1

Burden Test: Mixed effect directions

- Lose power if variants have positive and negative effects.

Y	G_1	G_2	G_3	G_4		C
1	1	0	0	0		1
1	0	1	0	0		1
1	0	0	0	0	→	0
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
0	0	0	0	0		0
0	0	0	1	0		1
0	0	0	0	1		1

Burden Test: Mixed effect directions

- Several methods have been developed to estimate association directions and incorporate them in the burden test framework.
  - Adaptive Sum Test
  - Estimated regression coefficient (EREC) test

Adaptive sum test

Han F and Pan W. (2010) *Hum Hered*

- Model:

$$C_i = \sum_{j=1}^{p} w_j g_{ij}$$

$$logit(Pr(Y = 1)) = \alpha_0 + C_i \beta$$

- Fit individual SNP models

$$logit(Pr(Y = 1)) = \alpha_0 + g_j \beta_j$$

  - Assign $w_j = -1$ if $\widehat{\beta}_j < 0$ and the p-value is small
  - $w_j = 1$ otherwise.

Adaptive sum test

- Compute p-values using permutation.
- Step-up procedure assign $w_j = 0$ if g_j is unlikely associated with the trait (Hoffmann *et al.*Plos One, 2010)

Estimated regression coefficient (EREC) test

Lin DY. and Tang Z. (2011) *AJHG*

- Estimate regression coefficient β and use it as a weight.

$$C_i = \sum_{j=1}^{p} w_j g_{ij}, \quad w_j = \widehat{\beta}_j$$

  - Motivation: True β_j is the optimal weight
  - Estimate $\widehat{\beta}_j$ by fitting individual SNP regression models
  - Use $w_j = \widehat{\beta}_j + \delta$ when the sample size is small (n$<$ 2000)

Estimated regression coefficient (EREC) test

- Calculate

$$C_i = \sum_{j=1}^{p} w_j g_{ij}, \quad w_j = \widehat{\beta}_j$$

- Test statistic:

$$T_{EREC} = \sum_{i=1}^{n} C_i (y_i - \widehat{\mu}_{0,i}).$$

  - Use score test statistics
  - P-values from the parametric bootstrap.

Estimated regression coefficient (EREC) test

- Cons:
  - Individual SNP regression models are difficult to fit for very rare variants.
  - The constant δ is arbitrary.

Adaptive burden test

- Adaptive burden tests have robust power.
- Compute p-values through permutation or bootstrap
  - Computationally intensive

Variance component test

- Burden tests are not powerful, if there exist variants with different association directions or many non-causal variants
- Variance component tests have been proposed to address it.
- "Similarity" based test

C-alpha test

Neale BM, *et al.*(2011). *Plos Genet.*

- Case-control studies without covariates.
- Assume the jth variant is observed n_{j1} times, with r_{j1} times in cases.

	a	A	Total
Case	r_{j1}	r_{j2}	r
Control	s_{j1}	s_{j2}	s
Total	n_{j1}	n_{j2}	n

- Under H_0

$$r_{j1} \sim Binomial(n_{j1}, q) \quad (q = r/n)$$

C-alpha test

- Risk increasing variant:

$$r_{j1} - qn_{j1} > 0$$

- Risk decreasing variant:

$$r_{j1} - qn_{j1} < 0$$

- Test statistic:

$$T_\alpha = \sum_{j=1}^{p} (r_{j1} - qn_{j1})^2 - \sum_{j=1}^{p} n_{j1} q(1-q)$$

- This test is robust in the presence of the opposite association directions.

C-alpha test

- Weighting scheme

$$T_\alpha = \sum_{j=1}^{p} w_j (r_{j1} - q n_{j1})^2 - \sum_{j=1}^{p} w_j n_{j1} q(1-q)$$

- Test for the over-dispersion due to genetic effects
  - Neyman's $C(\alpha)$ test.

C-alpha test, P-value calculation

- Using normal approximation, since the test statistic is the sum of random variables.

$$T_\alpha = \sum_{j=1}^{p} (r_{j1} - q n_{j1})^2 - \sum_{j=1}^{p} n_{j1} q (1-q)$$

- Doesn't work well when p is small (or moderate).
  - P-value is computed using permutation.

C-alpha test

- C-alpha test is robust in the presence of the different association directions
- Disadvantages:
  - Permutation is computationally expensive.
  - Cannot adjust for covariates.

Sequence Kernal Association Test (SKAT)

Wu *et al.*(2010, 2011). *AJHG*

- Recall the original regression models:

$$\mu_i / logit(\mu_i) = \alpha_0 + \mathbf{X}_i^T \boldsymbol{\alpha} + \mathbf{G}_i^T \boldsymbol{\beta}$$

- Variance component test:
  - Assume $\beta_j \sim dist.(0, w_j^2 \tau)$.
  - $H_0 : \beta_1 = \cdots = \beta_p = 0 <=> H_0 : \tau = 0$.

Sequence Kernel Association Test (SKAT)

- $\beta_j \sim dist.(0, w_j^2\tau)$: $\tau = 0$ is on the boundary of the hypothesis.
- Score test statistic for $\tau = 0$:

$$Q_{SKAT} = (\mathbf{y} - \widehat{\boldsymbol{\mu}}_0)'\mathbf{K}(\mathbf{y} - \widehat{\boldsymbol{\mu}}_0),$$

- $\mathbf{K} = \mathbf{GWWG}'$: weighted linear kernel ($\mathbf{W} = diag[w_1, \ldots, w_p]$).

Sequence Kernel Association Test (SKAT)

- The C-alpha test is a special case of SKAT
  - With no covariates and flat weights:

$$Q_{SKAT} = \sum_{j=1}^{p} (r_{j1} - q n_{j1})^2$$

SKAT

- Q_{SKAT} is a weighted sum of single variant score statistics

$$Q_{SKAT} = (\mathbf{y} - \widehat{\boldsymbol{\mu}}_0)' \mathbf{GWWG}' (\mathbf{y} - \widehat{\boldsymbol{\mu}}_0)$$
$$= \sum_{j=1}^{p} w_j^2 [\boldsymbol{g}_j' (\mathbf{y} - \widehat{\boldsymbol{\mu}}_0)] = \sum_{j=1}^{p} w_j^2 U_j^2$$

  where $U_j = \sum_{i=1}^{n} g_{ij}(y_i - \widehat{\mu}_{0i})$.

- U_j is a score of individual SNP j only model:

$$\mu_i / logit(\mu_i) = \alpha_0 + \mathbf{X}_i^T \boldsymbol{\alpha} + g_{ij}\beta_j$$

SKAT

- Q_{SKAT} (asymptotically) follows a mixture of χ^2 distribution under the NULL.

$$
\begin{aligned}
Q &= (\mathbf{y} - \widehat{\mu}_0)' \mathbf{K} (\mathbf{y} - \widehat{\mu}_0) \\
&= (\mathbf{y} - \widehat{\mu}_0)' \widehat{\mathbf{V}}^{-1/2} \widehat{\mathbf{V}}^{1/2} \mathbf{K} \widehat{\mathbf{V}}^{1/2} \widehat{\mathbf{V}}^{-1/2} (\mathbf{y} - \widehat{\mu}_0) \\
&= \sum_{j=1}^{p} \lambda_j [\mathbf{u}_j' \widehat{\mathbf{V}}^{-1/2} (\mathbf{y} - \widehat{\mu}_0)]^2 \\
&\approx \sum_{j=1}^{p} \lambda_j \chi^2_{1,j}
\end{aligned}
$$

SKAT

- λ_j and $\mathbf{u}_j$ are eigenvalues and eigenvectors of $\mathbf{P}^{1/2}\mathbf{K}\mathbf{P}^{1/2}$. where $\mathbf{P} = \widehat{\mathbf{V}}^{-1} - \widehat{\mathbf{V}}^{-1}\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widehat{\mathbf{V}}^{-1}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widehat{\mathbf{V}}^{-1}$ is the project matrix to account that $\boldsymbol{\alpha}$ is estimated.

SKAT: P-value calculation

- P-values can be computed by inverting the characteristic function using Davies' method (1973, 1980)
  - Characteristic function

$$\varphi_x(t) = E(e^{itx}).$$

  - Characteristic function of $\sum_{j=1}^{p} \lambda_j \chi^2_{1,j}$

$$\varphi_x(t) = \prod_{i=j}^{p} (1 - 2\lambda_j it)^{-1/2}.$$

  - Inversion Formula

$$P(X < u) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{Im[e^{-itu}\varphi_x(t)]}{t} \, \mathrm{d}t.$$

Small sample adjustment

Lee *et al.*(2012). *AJHG*

- When the sample size is small and the trait is binary, asymptotics does not work well.
- SKAT test statistic:

$$\begin{aligned} Q_{SKAT} &= (\mathbf{y} - \widehat{\boldsymbol{\mu}}_0)' \mathbf{K} (\mathbf{y} - \widehat{\boldsymbol{\mu}}_0) \\ &= \sum_{v=1}^{p} \lambda_v \eta_v^2, \end{aligned}$$

- η_vs are asymptotically independent and follow N(0,1).

Small sample adjustment

- When the trait is binary and the sample size is small:
  - $Var(\eta_v) < 1$.
  - η_vs are negatively correlated.

Small sample adjustment

- Mean and variance of the Q_{SKAT}

	Mean	Variance
Large Sample	$\sum \lambda_j$	$\sum \lambda_j^2$
Small Sample	$\sum \lambda_j$	$\sum \lambda_j \lambda_k c_{jk}$

- Adjust null distribution of Q_{SKAT} using the estimated small sample variance.

Small sample adjustment

- Variance adjustment is not enough to accurately approximate far tail areas.
- Kurtosis adjustment:
  - Estimate the kurtosis of Q_{SKAT} using parametric bootstrapping:
  - $\widehat{\gamma}$ (estimated kurtosis)
  - D.F. estimator: $\widehat{df} = 12/\hat{\gamma}$
  - Null distribution

$$(Q_{SKAT} - \sum \lambda_j^2) \frac{\sqrt{2\widehat{df}}}{\sqrt{\sum \lambda_j \lambda_k c_{jk}}} + \widehat{df} \sim \chi^2_{\widehat{df}}$$

Small sample adjustment

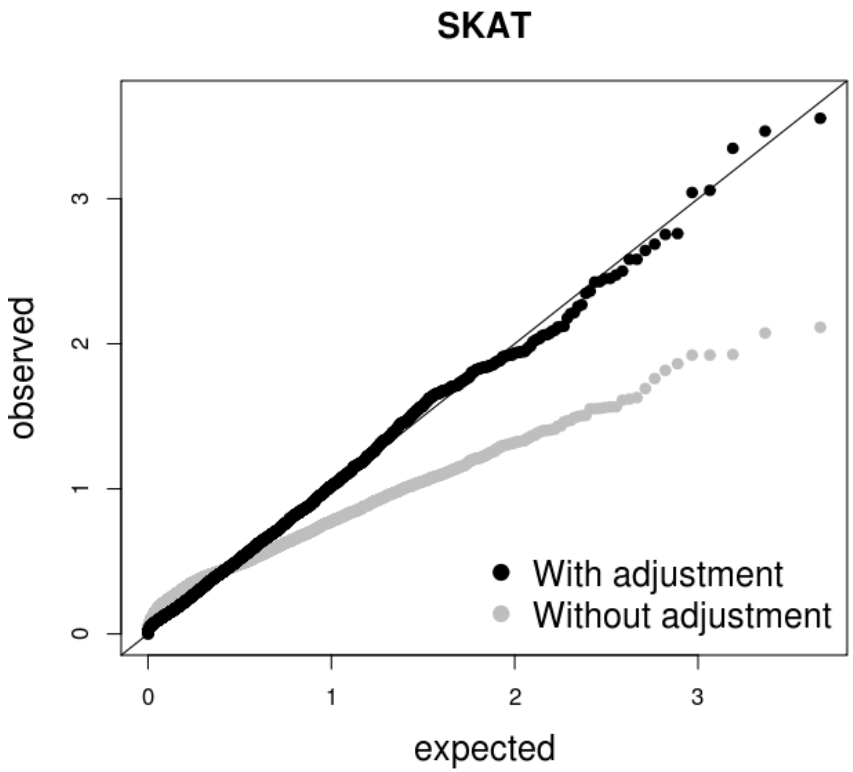

Figure: ARDS data (89 samples)

General SKAT

- General SKAT Model:

$$\mu_i / logit(\mu_i) = \alpha_0 + X_i \boldsymbol{\alpha} + h_i$$

  where $h_i \sim GP(0, \tau K)$.
- Kernel $K(\mathbf{G}_i, \mathbf{G}_{i'})$ measures genetic similarity between two subjects.

General SKAT

- Examples:
  - Linear kernel=linear effect

$$K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = w_1^2 Z_{i1} Z_{i'1} + \cdots w_p^2 Z_{ip} Z_{i'p}$$

  - IBS Kernel (Epistatic Effect: SNP-SNP interactions)

$$K(\mathbf{Z}_i, \mathbf{Z}_j) = \frac{\sum_{k=1}^{p} w_k^2 IBS(Z_{ik}, Z_{jk})}{2p}$$