Power and sample size calculations for designing rare variant sequencing association studies.

Seunggeun Lee[1], Michael C. Wu[2], Tianxi Cai[1], Yun Li[2,3], Michael Boehnke[4] and Xihong Lin[1]

1 Department of Biostatistics, Harvard School of Public Health, Boston

2 Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill

3 Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill

4 Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor

1 Introduction

Recently, Wu et al. [4] have proposed the sequence kernel machine test (SKAT) to test association between genetic variants in a gene or region and a continuous or binary trait. SKAT, which uses the kernel machine regression framework, is very flexible and computationally efficient. From extensive simulation studies and real data application, it has been shown that SKAT is more powerful than the collapsing based burden tests under many circumstances [4].

To design new sequence association study with SKAT as a testing procedure, it is important to know the required sample size to achieve a proper statistical power. Power and sample size calculation can be done by simulations, however this computer intensive approach would be time consuming. Here, we derive the analytical formula of the statistical power of SKAT. Required sample size can be computed easily by inverting the power function. In addition, We have developed user friendly R package which implements the power and sample size calculation formula (Web Resources).

2 Power and Sample Size Calculation Formula for SKAT: Statistical Derivations

2.1 Continuous Traits

For simplicity, we assume no covariates are present. However, the results presented can be easily extended to accommodate covariates. Suppose there are n individuals and $\mathbf{y} = (y_1, \ldots, y_n)'$ is a vector of continuous phenotype. We assume that p variants are observed in a particular gene or genomic region, and $\mathbf{G}$ is an $n \times p$ genotype matrix with $\mathbf{G}_i$ being the i^{th} row of the $\mathbf{G}$. To relate the SNP set to the phenotype, we consider the linear model

$$y_i = \alpha_0 + \mathbf{G}_i'\boldsymbol{\beta} + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. Without loss of generality and for the ease of presentation, we set each entry of $\mathbf{G}_i$ to be centered such that $E(\mathbf{G}_i) = 0$, and $\sigma = 1$ for continuous traits. The SKAT test statistic with a kernel $K(\cdot, \cdot)$ is

$$Q = (\mathbf{y} - \bar{y}\mathbf{1})'\mathbf{K}(\mathbf{y} - \bar{y}\mathbf{1}), \quad \text{where} \quad \bar{y} = n^{-1}\sum_{i=1}^{n} y_i.$$

In the case of weighted linear kernel with a weight function $w(\cdot)$,

$$\mathbf{K} = \mathbf{G}\mathbf{W}\mathbf{G}',$$

where $\mathbf{W} = diag\{w(\hat{m}_1), \ldots, w(\hat{m}_p)\}$, and $\hat{m}_j$ is an observed MAF of the j^{th} variant. Denote $\boldsymbol{\mu}_\beta = \mathbf{G}\boldsymbol{\beta}$ and $\mathbf{Z} = \mathbf{y} - \bar{y}\mathbf{1} - \boldsymbol{\mu}_\beta$, and then

$$Q = (\mathbf{y} - \bar{y}\mathbf{1})'\mathbf{K}(\mathbf{y} - \bar{y}\mathbf{1}) = (\mathbf{Z} + \boldsymbol{\mu}_\beta)'\mathbf{K}(\mathbf{Z} + \boldsymbol{\mu}_\beta).$$

By the spectral decomposition, $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$. Since each element of $\mathbf{Z}$ is an independent Gaussian with mean 0 and asymptotic variance 1, Q asymptotically follows $\sum_{j=1}^{m} \lambda_j \chi_1^2(\delta_j)$ with $\delta_j = \boldsymbol{\mu}_\beta'\mathbf{u}_j\mathbf{u}_j'\boldsymbol{\mu}_\beta$. Here, λ_j is the j^{th} diagonal element of $\boldsymbol{\Lambda}$, and $\mathbf{u}_j$ is the j^{th} column of $\mathbf{U}$.

For computational efficiency, we approximate the mixture of chi-square distributions of Q us-

ing the non-central chi-square approximation with ν degrees of freedom and non-centrality parameter δ [2] under the null and alternative. Results using the Davies method [1] for power calculations are similar. Specifically, we compute $c_k = \sum_{j=1}^{p} \lambda_j^k$ for the null distribution, and $c_k = \sum_{j=1}^{p} \lambda_j^k + k \sum_{j=1}^{p} \lambda_j^k \delta_j$ for the alternative distribution up to $k = 4$. These values can be obtained from

$$\sum_{j=1}^{p} \lambda_j^k = trace(\mathbf{K}^k) = trace\{(\mathbf{G}'\mathbf{G}\mathbf{W})^k\} \tag{1}$$

and

$$\sum_{j=1}^{p} \lambda_j^k \delta_j = \boldsymbol{\mu}_\beta' \mathbf{K}^k \boldsymbol{\mu}_\beta = trace(\boldsymbol{\mu}_\beta' \mathbf{K}^k \boldsymbol{\mu}_\beta) = trace\{(\mathbf{G}'\mathbf{G}\mathbf{W})^{k-1} \mathbf{G}' \boldsymbol{\mu}_\beta \boldsymbol{\mu}_\beta' \mathbf{G}\mathbf{W}\}. \tag{2}$$

Suppose $\mathbf{A} = E(\mathbf{G}'\mathbf{G}\mathbf{W}/n)$ and $\mathbf{B} = E(\mathbf{G}'\boldsymbol{\mu}_\beta\boldsymbol{\mu}_\beta'\mathbf{G}\mathbf{W}/n^2)$. Since the distribution of $\mathbf{G}$ can be inferred from population genetic simulations or existing data (e.g. 1000 genome project data), we can obtain both $\mathbf{A}$ and $\mathbf{B}$. By the continuity of trace and matrix multiplication, $trace(\mathbf{K}^k) = n^k trace(\mathbf{A}^k)$ and $trace(\boldsymbol{\mu}_\beta' \mathbf{K}^k \boldsymbol{\mu}_\beta) = n^{k+1} trace(\mathbf{A}^{k-1}\mathbf{B})$. After computing $c_1, \ldots, c_4$, we obtain following values.

$$\begin{aligned}
&\mu_Q = c_1, \quad \sigma_Q = \sqrt{2c_2}, \quad s_1 = c_3/c_2^{3/2}, s_2 = c_4/c_2^2, \\
&a = \begin{cases} 1/\left(s_1 - \sqrt{s_1^2 - s_2}\right) & \text{if } s_1^2 > s_2 \\ 1/\sqrt{s_2} & \text{if } s_1^2 \le s_2 \end{cases}, \\
&\delta = \begin{cases} s_1 a^3 - a^2 & \text{if } s_1^2 > s_2 \\ 0 & \text{if } s_1^2 \le s_2 \end{cases}, \\
&l = a^2 - 2\delta, \quad \mu_X = l + \delta, \quad \text{and } \sigma_X = \sqrt{2}\sqrt{l + 2\delta}.
\end{aligned}$$

Note that we modified the approximation of Liu *et al.*(2009) [2] when $s_1^2 \le s_2$ by matching kurtosis, instead of skewness, to improve the estimation of tail probability. To estimate the power, we first compute μ_Q, μ_X, σ_Q, and σ_X under the null. A critical value with level α is

$$q_c = (q(1-\alpha; \chi_l^2(\delta)) - \mu_X)\frac{\sigma_Q}{\sigma_X} + \mu_Q,$$

where $q(\cdot; \chi_l^2(\delta))$ is a quantile function of $\chi_l^2(\delta)$. Then, we recompute μ_Q, μ_X, σ_Q, and σ_X under

the alternative and estimate the power

$$P\left(\chi^2_l(\delta) > \frac{\sigma_X}{\sigma_Q}(q_c - \mu_Q) + \mu_X\right).$$

2.2 Dichotomous Traits in Cross-Sectional and Prospective Studies

In the absence of covariates, the logistic model we consider is

$$\text{logit}\,(\pi_i) = \alpha_0 + \mathbf{G}_i'\boldsymbol{\beta}, \tag{3}$$

where y_i is a disease status (1 = disease, 0 = non-disease). We assume that the prevalence/incidence of disease is known. Our SKAT test statistic with a weighted linear kernel $\mathbf{K}$ is

$$Q = (\mathbf{y} - \hat{\pi}_0\mathbf{1})'\mathbf{K}(\mathbf{y} - \hat{\pi}_0\mathbf{1}),$$

where $\hat{\pi}_0 = n^{-1}\sum_{i=1}^n y_i$, the estimated disease probability under H_0. Denote $\boldsymbol{\mu}_\beta = (\pi_1 - \hat{\pi}_0, \ldots, \pi_n - \hat{\pi}_0)'$, where π satisfies (3), and $\mathbf{V} = diag[v_1, \ldots, v_n]$, and $v_i = \pi_i(1-\pi_i)$ is $var(y_i)$. Then Q can be written as

$$\begin{aligned} Q &= (\mathbf{y} - \hat{\pi}_0\mathbf{1})'\mathbf{K}(\mathbf{y} - \hat{\pi}_0\mathbf{1}) \\ &= (\mathbf{y} - \hat{\pi}_0\mathbf{1} - \boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta)'\mathbf{V}^{-1/2}\mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}\mathbf{V}^{-1/2}(\mathbf{y} - \hat{\pi}_0\mathbf{1} - \boldsymbol{\mu}_\beta + \boldsymbol{\mu}_\beta) \\ &= (\mathbf{Z} + \mathbf{V}^{-1/2}\boldsymbol{\mu}_\beta)'\widetilde{\mathbf{K}}(\mathbf{Z} + \mathbf{V}^{-1/2}\boldsymbol{\mu}_\beta), \end{aligned}$$

where $\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{y} - \hat{\pi}_0\mathbf{1} - \boldsymbol{\mu}_\beta)$, and $\widetilde{\mathbf{K}} = \mathbf{V}^{1/2}\mathbf{K}\mathbf{V}^{1/2}$. Since each element of $\mathbf{Z}$ has mean 0 and variance 1, $(\mathbf{u}_j'\mathbf{Z})^2$ asymptotically follows independent χ^2_1 distribution. Now we apply the same argument shown in the Section 2.1 using $\widetilde{\mathbf{K}}$ instead of $\mathbf{K}$, and estimate the power.

2.3 Modifications of Power Calculations for Rare Variants

With finite sample size n, causal variants that are rare may not be observed. Our power and sample size calculations can account for this uncertainty. Suppose the population MAF for the j^{th}

variant is m_j. Let θ_j be the chance the variant j is observed (polymorphic) in sample size n. Then

$$\theta_j = 1 - (1 - m_j)^{2n},$$

and the sample size required to observe this variant with at least π_j chance is

$$n > \frac{ln(1-\theta_j)}{2ln(1-m_j)}.$$

For example, to have $\theta_j = 99.9\%$ chance to observe a variant with a given MAF, the required minimum sample size is

MAF	0.1	0.01	0.001	0.0001
Minimum n	33	344	3453	34537

One can see larger sample size is needed to observe a rarer variant.

Suppose there are p variants in a region in the population, with rare variants, the model we fit is actually $y_i = \alpha_0 + \widetilde{\mathbf{G}}_i\boldsymbol{\beta} + \epsilon_i$ for continuous traits, and logit $(\pi_i) = \alpha_0 + \widetilde{\mathbf{G}}_i\boldsymbol{\beta}$ for dichotomous traits, where $\widetilde{\mathbf{G}}_i = (G_{i1}\Delta_1, \ldots, G_{ip}\Delta_p)' = \boldsymbol{\Delta}\mathbf{G}_i$, and $\boldsymbol{\Delta} = diag[\Delta_1, \ldots, \Delta_p]$. Here, Δ_j is an indicator that variant j is observed in sample size n. Under this model, the weighted linear kernel $\mathbf{K} = \widetilde{\mathbf{G}}W\widetilde{\mathbf{G}} = \mathbf{G}\boldsymbol{\Delta}\mathbf{W}\boldsymbol{\Delta}\mathbf{G}'$, and $trace(\mathbf{K}^k)$ and $trace(\boldsymbol{\mu}'_\beta\mathbf{K}^k\boldsymbol{\mu}_\beta)$ can be approximated as $trace(\mathbf{K}^k) \approx n^k trace((\mathbf{A}\boldsymbol{\Pi})^k)$ and $trace(\boldsymbol{\mu}'_\beta\mathbf{K}^k\boldsymbol{\mu}_\beta) \approx n^{k+1}trace((\mathbf{A}\boldsymbol{\Pi})^{k-1}\mathbf{B}\boldsymbol{\Pi})$, where $\boldsymbol{\Pi} = diag[\theta_1, \ldots, \theta_p]$. We further improve the approximation by incorporating the fact that $E(\boldsymbol{\Delta}_i) = E(\boldsymbol{\Delta}_i^2)$. Let $\mathbf{A}_2$ be a $p \times p$ matrix with the (i,j)th element being $a_{ij}\theta_i\theta_j^{I(i\neq j)}$, where a_{ij} is the $(i,j)^{th}$ element of $\mathbf{A}$, and then $E(\boldsymbol{\Delta}\mathbf{G}'\mathbf{G}\mathbf{W}\boldsymbol{\Delta}/n) \approx \mathbf{A}_2$. Let us denote $\mathbf{A}_1 = \boldsymbol{\Pi}$, $\mathbf{A}_3 = \boldsymbol{\Pi}\mathbf{A}\mathbf{A}_2$, $\mathbf{A}_4 = \mathbf{A}_2\mathbf{A}\mathbf{A}_2$, and then $trace(\mathbf{K}^k) \approx n^k trace(\mathbf{A}\mathbf{A}_k)$, and $trace(\boldsymbol{\mu}'_\beta\mathbf{K}^k\boldsymbol{\mu}_\beta) \approx n^{k+1}trace(\mathbf{B}\mathbf{A}_k)$. Now we compute the power from the χ^2 approximation method which is described in the previous section.

2.4 Power and Sample Size Calculations for Retrospective Case-Control Studies

It is well known that logistic regression can be used to analyze case-control data [3]. However, it is necessary to incorporate the retrospective nature of case-control studies to properly estimate the power. Let S be a selection indicator such that $S = 1$ denotes a subject is selected in the case-control sample. Then the conditional distribution of G and y given $S = 1$, instead of the unconditional distribution of G and y, should be used to compute power. Denote by $\tilde{\pi}_i = Pr(y_i =$

$1|\mathbf{G}_i, S_i = 1)$ the case-control probability , if the the population disease probability follows the logistic model (3), then the case-control probability $\tilde{\pi}_i$ also follows the same logistic model except for a different intercept [3] as

$$\text{logit}\,(\tilde{\pi}_i) = \tilde{\alpha}_0 + \mathbf{G}_i'\boldsymbol{\beta}, \tag{4}$$

where

$$\tilde{\alpha}_0 = \alpha_0 + log\left\{\frac{P(S=1|y=1)}{P(S=1|y=0)}\right\} = \alpha_0 + log\left\{\frac{\hat{\pi}_0 P(y=0)}{(1-\hat{\pi}_0)P(y=1)}\right\}, \tag{5}$$

where $P(S=1|y=1)$ is the probability that a case is sampled and $P(S=1|y=0)$ is the probability that a control is sampled, and $P(y=1)$ is the population disease prevalence/incidence. Further one can show that

$$\begin{aligned} P(G|S=1) &= P(G|y=1,S=1)P(y=1|S=1) + P(G|y=0,S=1)P(y=0|S=1) \\ &= \frac{\hat{\pi}_0}{P(y=1)}P(y=1|G)P(G) + \frac{1-\hat{\pi}_0}{P(y=0)}P(y=0|G)P(G). \end{aligned} \tag{6}$$

We compute $\mathbf{A}$, $\mathbf{B}$, $\mathbf{W}$, $\mathbf{A}_2$ and $\mathbf{\Pi}$ by estimating μ_β and $\mathbf{V}$ using (5) and by using conditional distribution (6), and subsequently estimate the power.

3 Computing the average power over different regions

The statistical power of rare variants analysis depends on the LD structure of genomic regions to be investigated, and MAFs of causal variants. If one is interested in only one known region and knows in advance which variants are causal, they can directly estimate power using the power formula given above. In practice, however, one is often interested in more than one region and only hypothesizes a disease model of causal variants. For example, one may hypothesize that a certain percentage of rare variants are causal, instead of selecting priori causal variants. In this case, we propose to use the average power of regions given a disease model. This average power can be easily computed from taking mean of obtained powers from the power formula using randomly selected regions/causal variants. Our experience shows that 50 $\sim$ 100 sets of different regions/causal variants are enough to compute the average power stably.

4 Web Resources

An implementation of SKAT and power/sample size calculations in the R language can be found at http://www.hsph.harvard.edu/∼xlin/software.html.

References

[1] Davies, R. (1980). Algorithm AS 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics 29*(3), 323–333.

[2] Liu, H., Y. Tang, and H. Zhang (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis 53*(4), 853–856.

[3] Prentice, R. and R. Pyke (1979). Logistic disease incidence models and case-control studies. *Biometrika 66*(3), 403–411.

[4] Wu, M., S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin (2011). Rare Variant Association Testing for Sequencing Data Using the Sequence Kernel Association Test (SKAT). *Manuscript*.