Web-based Supplementary Materials for "Sequence Robust Association Test for Familial Data"

Wei Dai, Ming Yang, Chaolong Wang, and Tianxi Cai

November 30, 2016

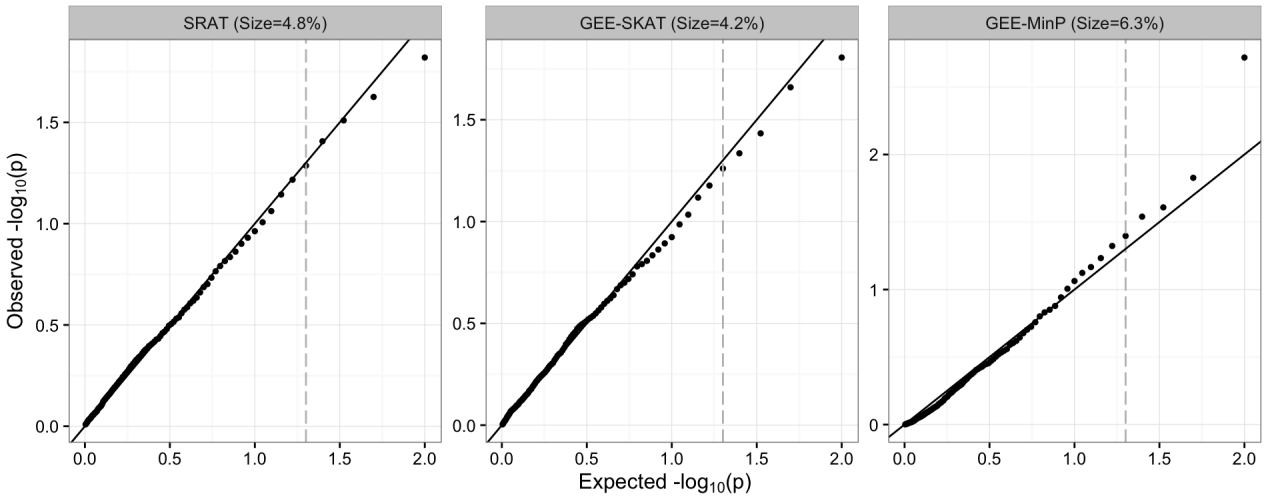
Web Appendix A: Setting without Covariates

We also considered simplified settings where there are no covariates in the true model. The SNP data were generated from the same model as described in the simulation section. Two types of alternatives are considered: (i) multiple causal variants with $\eta_{ij} = \boldsymbol{\beta}^\top \mathbf{G}_{ij} + a_{ij}$, $\beta_k = \tau |\log_{10}(\text{MAF}_k)|$ and $\tau = 0.01, 0.02, \ldots, 0.05$, and (ii) single causal variant where for $k = 1, ..., p$, we let the k^{th} SNP being the casual variant with $\eta_{ij} = 0.1G_{ijk} + a_{ij}$. For both settings, we generate continuous phenotype from exponential with mean $e^{\eta_{ij}}$ and binary phenotype from Bernoulli with success probability $\pi_{ij} = 1 - e^{-e^{\eta_{ij}}}$. Figure 2 (a) shows the empirical size of different tests under the null model of no association between Y_{ij} and $\mathbf{G}_{ij}$. The empirical power under alternative (i) and (ii) are shown in Figure 2 (b) and (c), respectively.

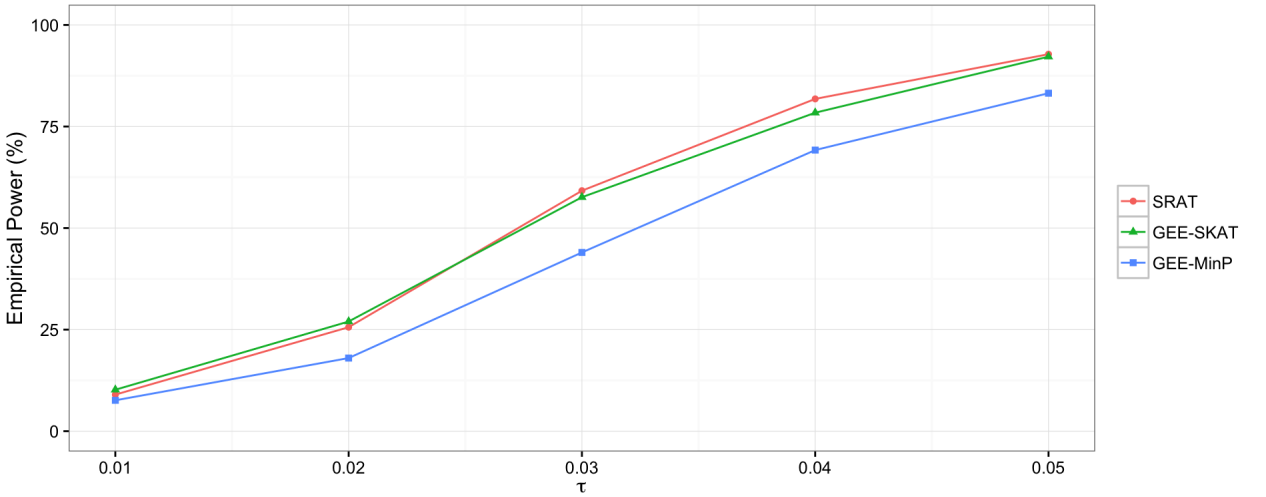
Web Appendix B: Sensitivity Analysis

Our objective here is to investigate whether the results obtained from SRAT are sensitive to the choice of bandwidth h. We follow the approach that is used in the simulation section to generate the covariates $\mathbf{X}$. For illustration, we only consider one setting where $\mathbf{X}$ is independent of $\mathbf{G}$ and perform unweighted analysis for association testing. Since it is a common practice to under-smooth the data, we consider a few smaller bandwidths other than the optimal one h_{opt}. The other three candidate bandwidths are $h_{opt} \times n^{-0.05}$, $h_{opt} \times n^{-0.10}$, and $h_{opt} \times n^{-0.15}$, where n is the number of families in the data. Figure 3 shows that the empirical size and power of SRAT are both robust to difference choices of bandwidth. Based on our experience and some theoretical considerations, we recommend using $h = h_{opt} \times n^{-0.1}$ as the default bandwidth in SRAT.

(a)



(b)



(c)

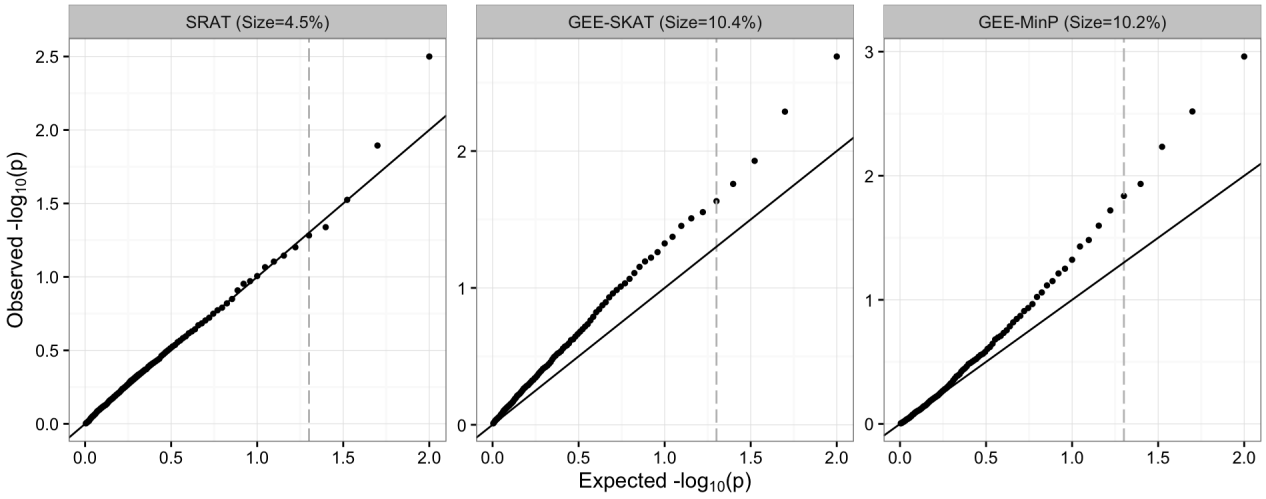


Figure 1: Empirical results under the setting with binary phenotype: (a) size with uncorrelated $\mathbf{X}$ and $\mathbf{G}$; (b) power with uncorrelated $\mathbf{X}$ and $\mathbf{G}$; (c) size with correlated $\mathbf{X}$ and $\mathbf{G}$.

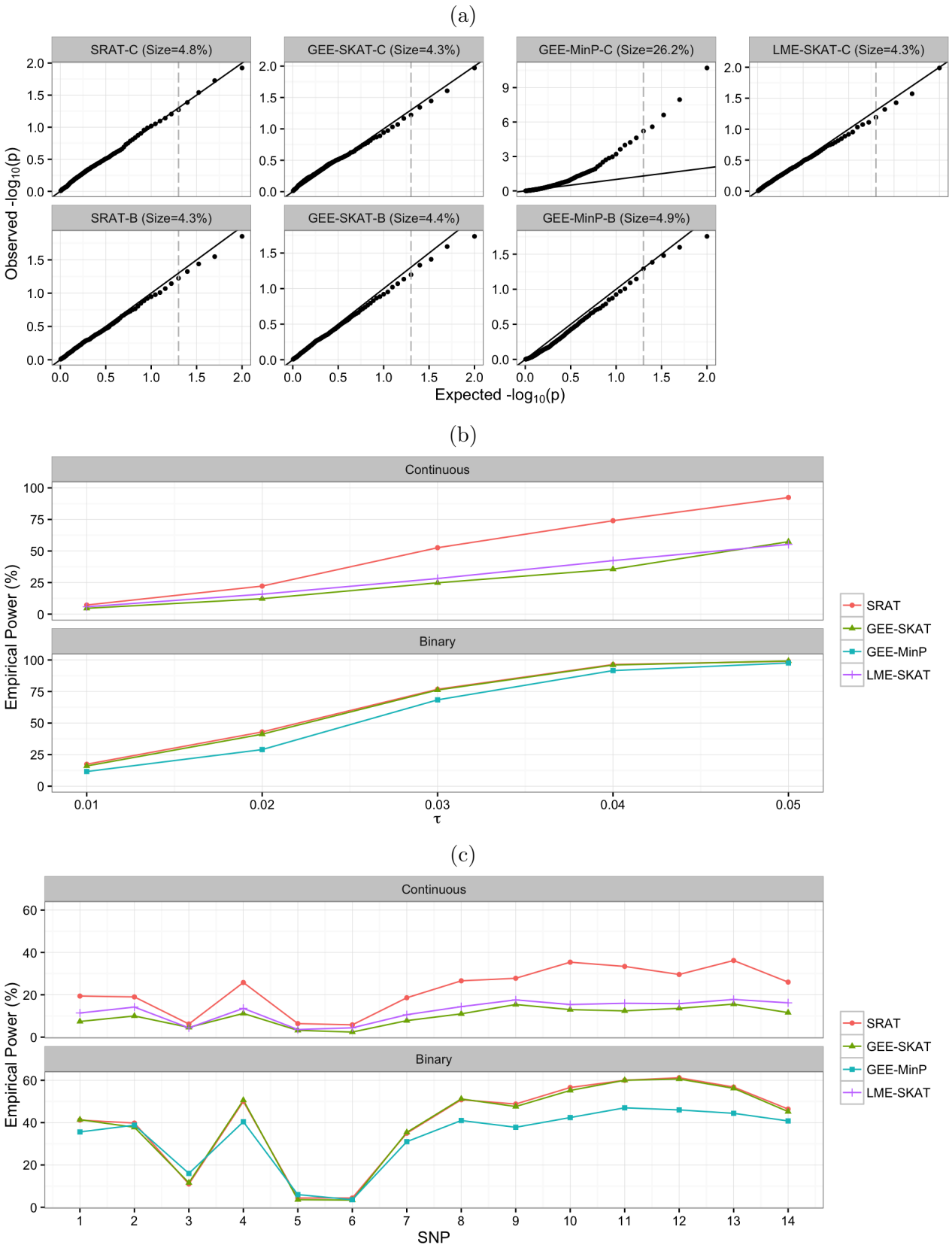
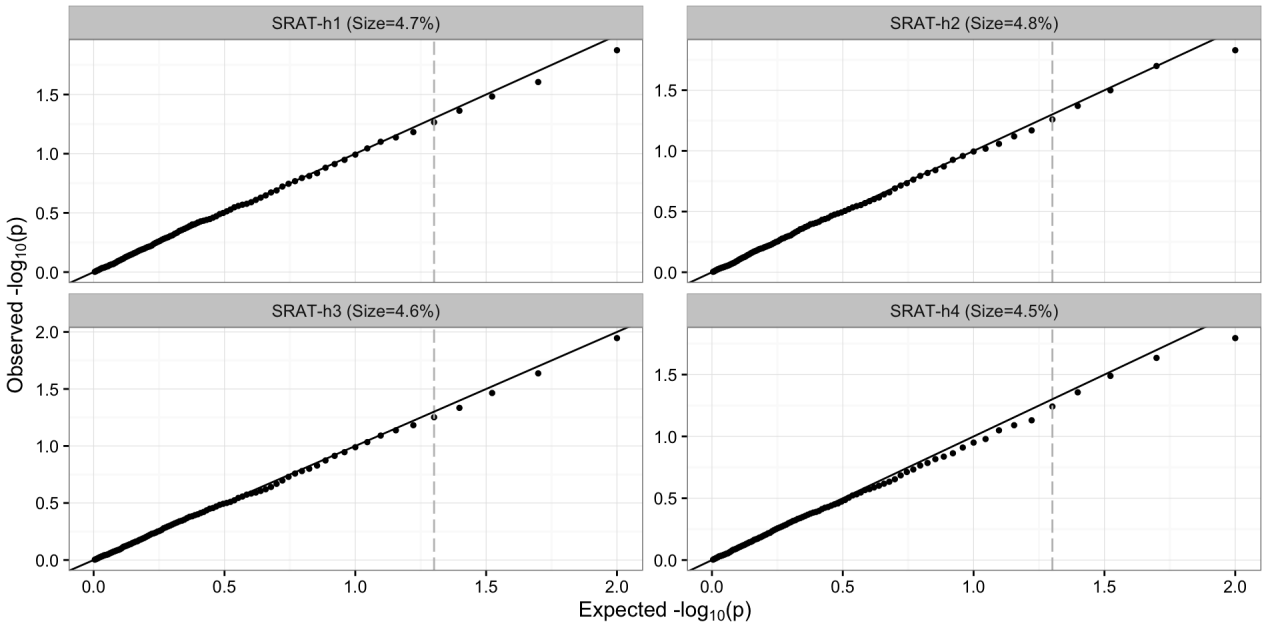
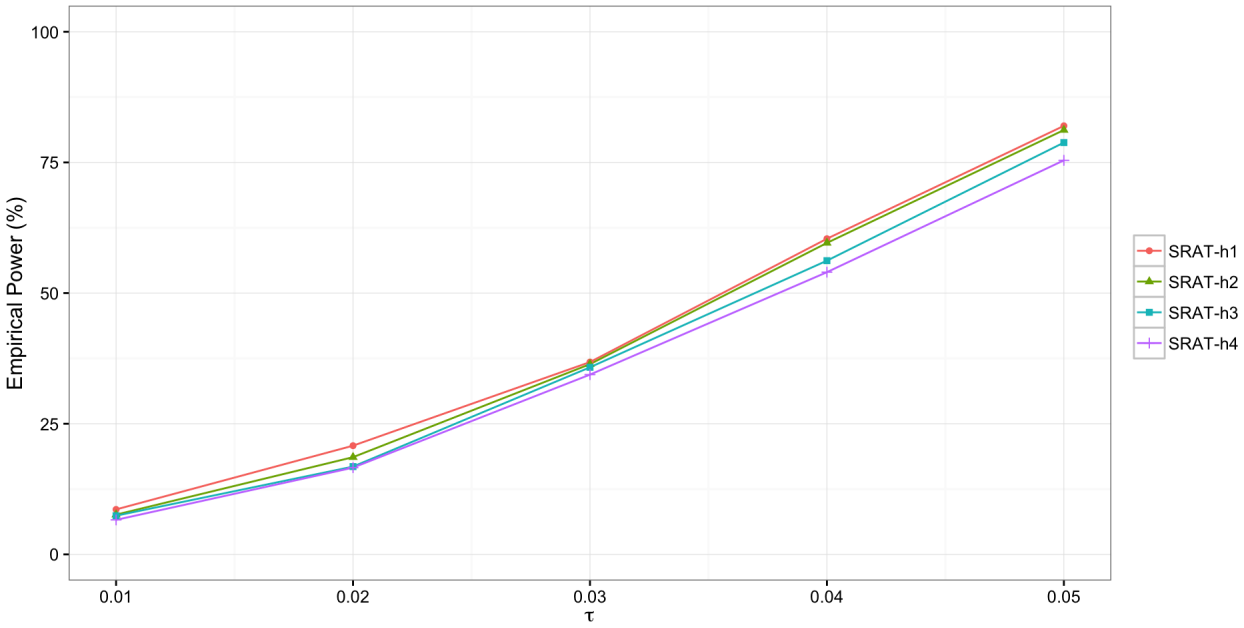


Figure 2: Empirical results under the setting without covariates for both continuous (C) and binary (B) phenotypes: (a) QQ plot and type I error rate; (b) power under the scenario with multiple causal SNPs; (c) power under the scenario with a single causal SNP.



(a)



(b)

Figure 3: Empirical size (a) and power (b) of SRAT under different bandwidths: $h_1 = h_{opt}$, $h_2 = h_{opt} \times n^{-0.05}$, $h_3 = h_{opt} \times n^{-0.10}$, and $h_4 = h_{opt} \times n^{-0.15}$, where h_{opt} is the optimal bandwidth and n is the number of families.