File S1

Supplementary materials for "On Assessing Genome-Wide Statistical Significance for Large p Small n Problems"

Guoqing Diao and Anand N. Vidyashankar

Department of Statistics, George Mason University, Fairfax, VA 22030

Additional Simulation Studies

GWAS study with SNP data

We conducted additional simulation studies to evaluate the performance of the proposed method for genome-wide association studies with single nucleotide polymorphisms (SNPs) data. We considered a study that scans 10 independent genome regions with 200 biallelic SNPs in each region. For each SNP, we assume Hardy-Weinberg equilibrium and set the minor allele frequency to be 0.4. Within each genome region, the linkage disequilibrium (LD) between successive two loci varied from 0 to 0.18. Under the null hypothesis, we generated the quantitative traits from a standard normal distribution; under the alternative hypothesis, we assume that the 35th SNP in genome region 1 has an additive effect. The effect sizes were set to be 1.2 and 0.8 for sample sizes of 50 and 100, respectively. The number of resamples B was set to be 2,000.

Figure 1 presents the sizes and powers of the two resampling approaches at genome-wide significance (GWS) level of 0.05 and 0.01 based on 10,000 replicates. Under all scenarios, the method of our paper has type I error rate close to the nominal level while the approach of Zou *et al*. (2004) tends to be conservative especially for $n = 50$ and significance level of 0.01. The proposed method substantially improves the power of the test over that of Zou *et al*. (2004). For example, with

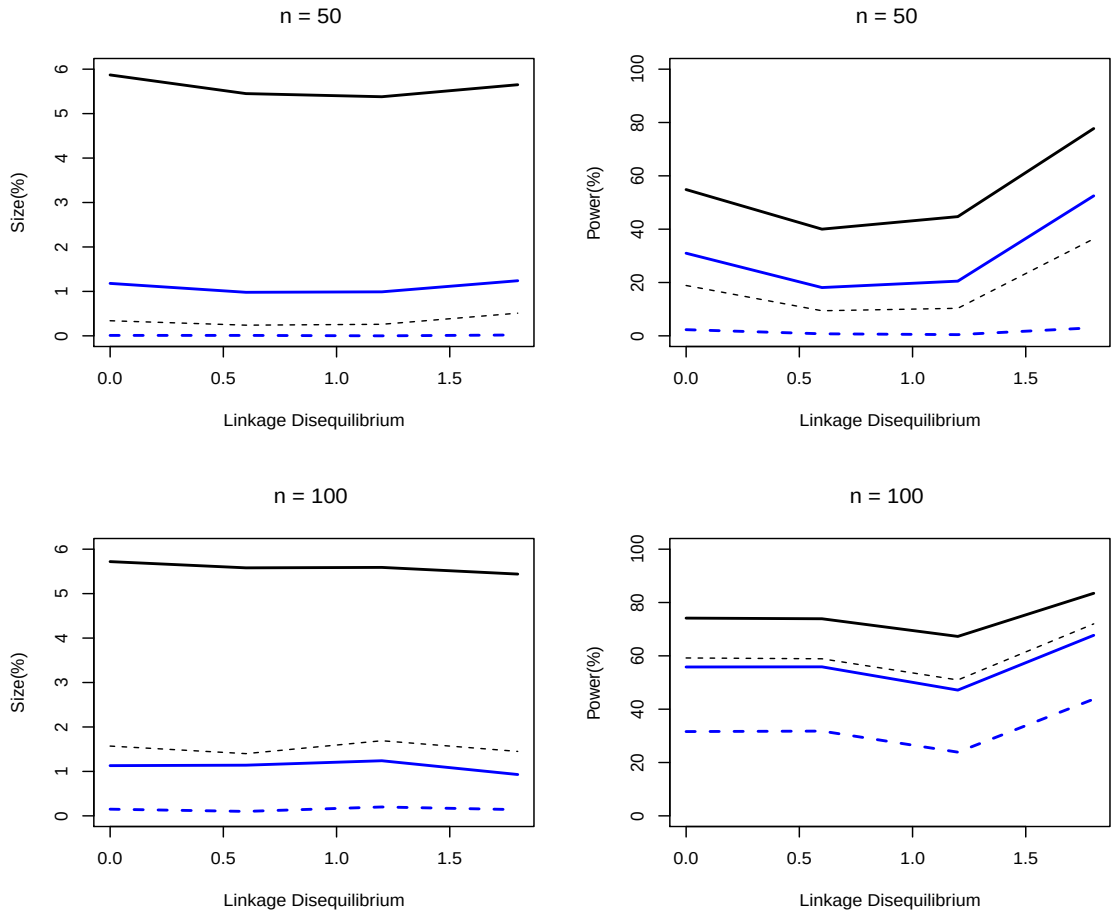

Figure 1: Sizes/powers(%) at nominal genome-wide significance level of 0.05 and 0.01. The black solid and black dashed curves correspond to the sizes/powers of the proposed method at significance levels of 0.05 and 0.01, respectively. The blue solid and blue dashed curves correspond to the sizes/powers of the method of ZOU *et al.* (2004) at significance levels of 0.05 and 0.01, respectively.

$n = 50$ and a LD of 0.18, the powers were 77.73% and 52.49% at significance levels of 0.05 and 0.01, respectively, compared to 36.37% and 3.04% of ZOU *et al.* (2004).

Simulation studies for $n = 500$

We have conducted additional simulation studies for the case of $n = 500$ and $p = 100, 1000$, and 2000. Table 1 in the Supplementary Materials presents the sizes and powers of the proposed method and the resampling approach of ZOU *et al.* (2004). These two methods yielded similar results. The reason for this is that when n is large and p is not much larger than n, the asymptotic theory takes effect. It would be desirable to conduct simulation studies to compare the two methods under the scenario of $p \gg n$ for large n. While it is feasible to analyze a real data set with both large n and large p, it is computationally prohibitive to conduct simulation studies given the current computing technology.

TABLE 1

Sizes/powers(%) at nominal genome-wide significance level of α with $n = 500$

Setup		Proposed[c]		ZOU *et al.* (2004)[d]	
p[a]	μ[b]	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
100	0.0	4.80	0.96	4.78	0.90
	0.2	73.99	50.88	73.66	50.04
1000	0.0	4.74	1.08	4.43	0.95
	0.2	44.95	25.73	43.88	24.13
2000	0.0	5.06	1.08	4.59	0.88
	0.2	37.21	20.32	35.76	18.64

[a] Total number of markers.

[b] Additive effect.

[c] Sizes/powers based on the proposed resampling method.

[d] Sizes/powers based on the resampling method of ZOU *et al.* (2004).

Computation of the standard error estimates

For the simulations described in the main manuscript, we also computed the standard error estimates on the thresholds by using the function *quantileSE* in the R package *broman*, which implements the method described in COX and HINKLEY (1974). The average of the standard error estimates based on the proposed method agree well with the standard error estimates of the empirical thresholds, obtained from 10,000 replicates under the null hypothesis. For example, for $n = 50$, $p = 100$, and $\alpha = 0.05$, the empirical threshold was 7.33 (SE=0.071) and the average of the proposed threshold was 7.25 and the average of the standard error estimates was 0.071.

References

COX, D. R., and D. V. HINKLEY, 1974 *Theoretical Statistics*. Chapman and Hall, London.

ZOU, F., J. FINE, J. HU, and D. Y. LIN, 2004 An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. Genetics **168**: 2307–2316.