Adaptive testing with high-dimensional genetic and microbiome data

Chong Wu*, Gongjun Xu+, Wei Pan*

* Division of Biostatistics, University of Minnesota
+ Department of Statistics, University of Michigan

Aug. 2nd, 2017

Outline

Introduction

- Suppose n independent samples have been collected, for which we have an n-vector response Y, an $n \times q$ matrix $\mathbb{Z}$ for q covariates, and an $n \times p$ matrix $\mathbb{X}$ for variables of interest.
- Assuming a generalized linear model:

$$E(Y|\mathbb{X}, \mathbb{Z}) = g^{-1}(\mathbb{X}\beta + \mathbb{Z}\alpha),$$

, where p-vector β and q-vector α are unknown parameters, and g is the canonical link function.

- Null hypothesis:

$$H_0 : \beta = \beta_0 \qquad \text{versus} \qquad H_1 : \beta \neq \beta_0$$

- $p >> n$.
- Motivating examples: Polygenic test, Pathway based analysis.

Problems with traditional tests

- $p \ll n$: likelihood ratio test and the Wald test.
- The power tends to diminish quite rapidly as p increases.
- Break down completely when $p > n$ since usual ordinary least squares estimator no longer exists.

Existing high dimensional tests

- Goeman et al. (2011):

$$T_{\text{Goe}} = \frac{(Y-\hat{\mu}_0)^\intercal \mathbb{X}\mathbb{X}^\intercal (Y-\hat{\mu}_0)}{(Y-\hat{\mu}_0)^\intercal \mathbb{D} (Y-\hat{\mu}_0)},$$

  where $\hat{\mu}_0$: maximum likelihood estimate of μ_0 under the null hypothesis; $\mathbb{D}$: diagonal of $\mathbb{X}\mathbb{X}^\intercal$.

- Guo and Chen (2016):

$$T_{\text{Guo}} = n^{-1}(Y-\hat{\mu}_0)^\intercal (\mathbb{X}\mathbb{X}^\intercal - \mathbb{D})(Y-\hat{\mu}_0),$$

- $n, p \to \infty$, T_{Guo} converges to a normal distribution.

Limitation of existing methods

- A large proportion of small to moderate signals: sum-of-squares of the score (existing tests) are more powerful.
- Signals are strong but highly sparse: supremum of the score (minimum p test) is more powerful.
- Signals are dense and in the same direction: Sum of the score (Sum test) is more powerful.
- Intermediate situations: neither of type of the test is powerful.

Goal

Develop an adaptive testing approach which would yield high testing power under various high-dimensional scenarios.

aSPU

- The score vector U:

$$U_j = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{0i})X_{ij}, \qquad 1 \le j \le p,$$

- SPU tests: for a $\gamma \ge 1$

$$L(\gamma, \hat{\mu}_0) = \sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{0i})X_{ij}\right)^{\gamma};$$

$$L(\infty, \hat{\mu}_0) = \max_{1\le j\le p}\frac{n\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{0i})X_{ij}\right)^2}{\sigma_{jj}};$$

$$T_{\text{aSPU}} = \min_{\gamma\in\Gamma} P_{\text{SPU}(\gamma,\hat{\mu}_0)}.$$

Theorem for SPU tests

Under H_0 and some regularity conditions, we have

- Let Γ be a set of finite positive integers,

$$[\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma)]^{\mathsf{T}}_{\gamma\in\Gamma} \xrightarrow{d} N(0, R),$$

- For any $x \in \mathbb{R}$,

$$Pr\{L(\infty, \hat{\mu}_0) - a_p \leq x\} \to \exp\{-\pi^{-1/2}\exp(-x/2)\}$$

  as $n, p \to \infty$, where $a_p = 2\log p - \log\log p$.

- $[\{L(\gamma, \hat{\mu}_0) - \mu(\gamma)\}/\sigma(\gamma)]^{\mathsf{T}}_{\gamma\in\Gamma'}$ is asymptotically independent with $L(\infty, \hat{\mu}_0)$.

P-value calculations

- Asymptotics:

$$p_O = 1 - \int_{\substack{s=(s_\gamma:\text{odd } \gamma\in\Gamma)^\intercal \\ -T_O\le s_\gamma\le T_O}} N(0, R_O)ds,$$

$$p_E = 1 - \int_{\substack{t=(t_\gamma:\text{even } \gamma\in\Gamma)^\intercal \\ -\infty\le t_\gamma\le T_E}} N(0, R_O)dt,$$

$$p_{\min} := \min\{p_O, p_E, p_\infty\},$$

$$p_{\text{aSPU}} = 1 - (1 - p_{\min})^3.$$

Approximation for $\mu(\gamma)$

Under the null hypothesis H_0,

$$\mu(\gamma) = \begin{cases} \frac{\gamma!}{d!2^d} n^{-d} \sum_{i=1}^{p} \sigma_{ii}^d + o(pn^{-d}), & \text{if } \gamma = 2d, \\ o(pn^{-(d+1)}), & \text{if } \gamma = 2d+1, \end{cases}$$

where $\sigma_{ii} = E[(S_{1i})^2]$, $S_{ij} = (Y_i - \mu_{0i})X_{ij}$.

Approximation for $\sigma(\gamma)$

Under the null hypothesis H_0,
$\sigma^2(1) = \frac{1}{n}\sum_{1\le i,j\le p}\sigma_{ij} + o(pn^{-1})$ and for $\gamma \ge 2$,

$$\sigma^2(\gamma) = \mu(2\gamma) - \sum_{j=1}^{p}\{\mu^{(j)}(\gamma)\}^2 + o(pn^{-\gamma})$$

$$+ \frac{1}{n^\gamma}\sum_{i\neq j}\sum_{\substack{2c_1+c_3=\gamma \\ 2c_2+c_3=\gamma \\ c_3>0}} \frac{(\gamma!)^2}{c_3!c_1!c_2!2^{c_1+c_2}}\sigma_{ii}^{c_1}\sigma_{jj}^{c_2}\sigma_{ij}^{c_3},$$

where $\sigma_{ij} = E[S_{1i}S_{1j}]$.

Proof idea

- For finite γ:
  - Weak dependence among $\mathbb{X}$ (α-mixing)
  - $p \to \infty$, Lyapunov condition can be checked and central limit theorem can be applied
- For $\gamma = \infty$: similar argument as Theorem 6 in Tony Cai et al. (2014).
- With nuisance parameters: we prove $||\mu_0 - \hat{\mu}_0||$ is ignorable

Simulation Settings

Simulation settings:

- $\mathbb{X}$: from multivariate normal distributions and $X_i \sim N(\mu_i, \Sigma)$.
- $\mathbb{Z}$ from standard normal distribution $N(0, 1)$.
- $$\text{logit}[P(Y_i = 1)] = 1 + \mathbb{Z}\alpha + \mathbb{X}\beta,$$
- Under null hypothesis, $\beta = 0$.
- Under alternative, $\lfloor ps \rfloor$ elements in β were set to be non-zero, where $s \in [0, 1]$.

Simulation Results

Table: Empirical type 1 errors and powers (%) of various tests for normal samples with $n = 200$, $p = 2000$.

r	0	0.03	0.05	0.07	0.1	0.15
SPU(1)	5 (5)	33 (32)	59 (59)	73 (74)	84 (86)	92 (92)
SPU(2)	6 (5)	18 (15)	44 (39)	65 (61)	81 (78)	91 (89)
SPU(3)	4 (5)	28 (30)	58 (59)	76 (76)	89 (90)	96 (96)
SPU(4)	4 (6)	11 (14)	33 (36)	55 (58)	74 (75)	87 (87)
SPU(5)	4 (5)	15 (18)	37 (41)	59 (62)	78 (81)	88 (89)
SPU(6)	3 (6)	7 (11)	18 (24)	36 (43)	53 (59)	70 (72)
SPU(∞)	5 (5)	7 (7)	8 (9)	13 (16)	19 (22)	21 (25)
aSPU	5 (5)	22 (25)	53 (57)	75 (77)	90 (90)	96 (96)

Simulation Results

Chong Wu,
Gongjun Xu,
Wei Pan

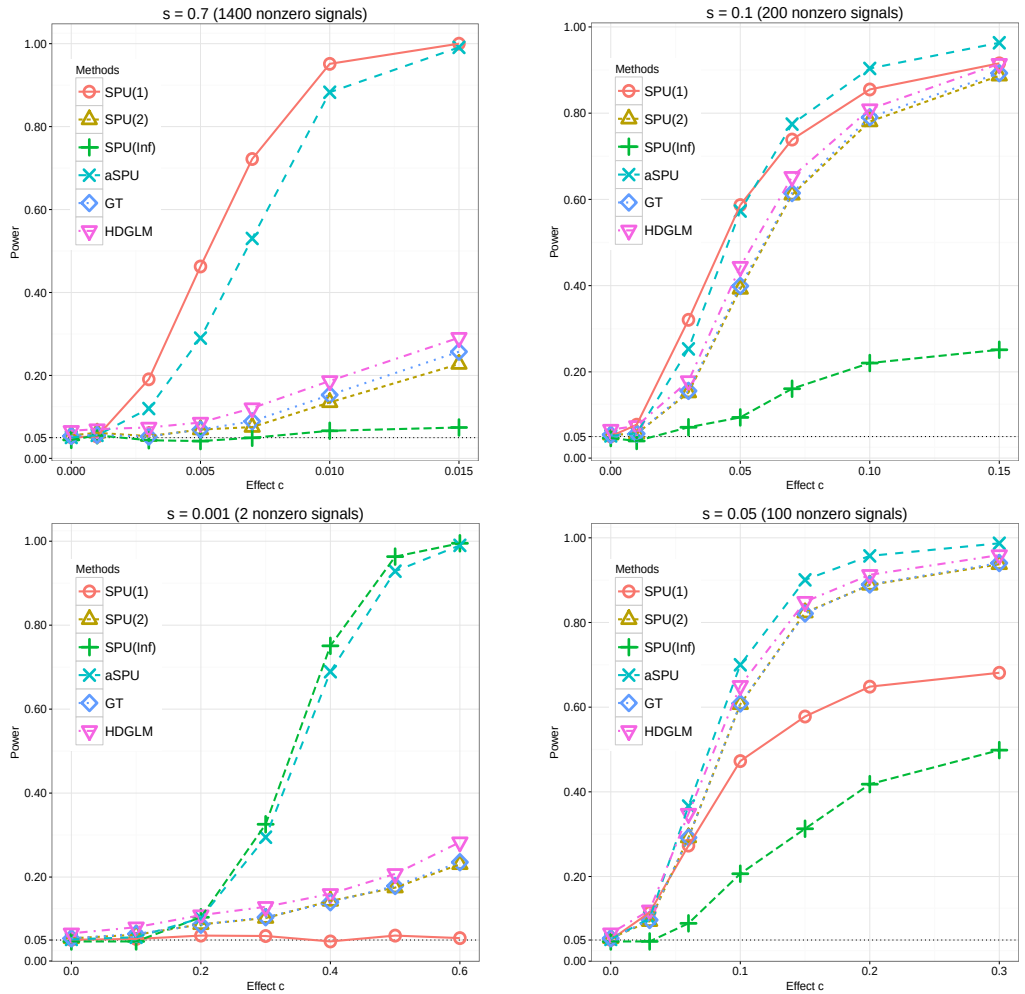

Figure: Power comparison for different methods.

Application to ADNI data

- Alzheimer's disease (AD) is the most common form of dementia.
- ADNI is a longitudinal multisite observational study of healthy elders, mild cognitive impairment, and AD. ADNI has recruited more than 1,500 subjects.
- We retrieved a total of 214 human biological pathways from the KEGG database (Only analyze the pathway with 10 to 200 genes, $\#\text{SNPs} > 1000$).

Chong Wu,
Gongjun Xu,
Wei Pan

Application to ADNI data

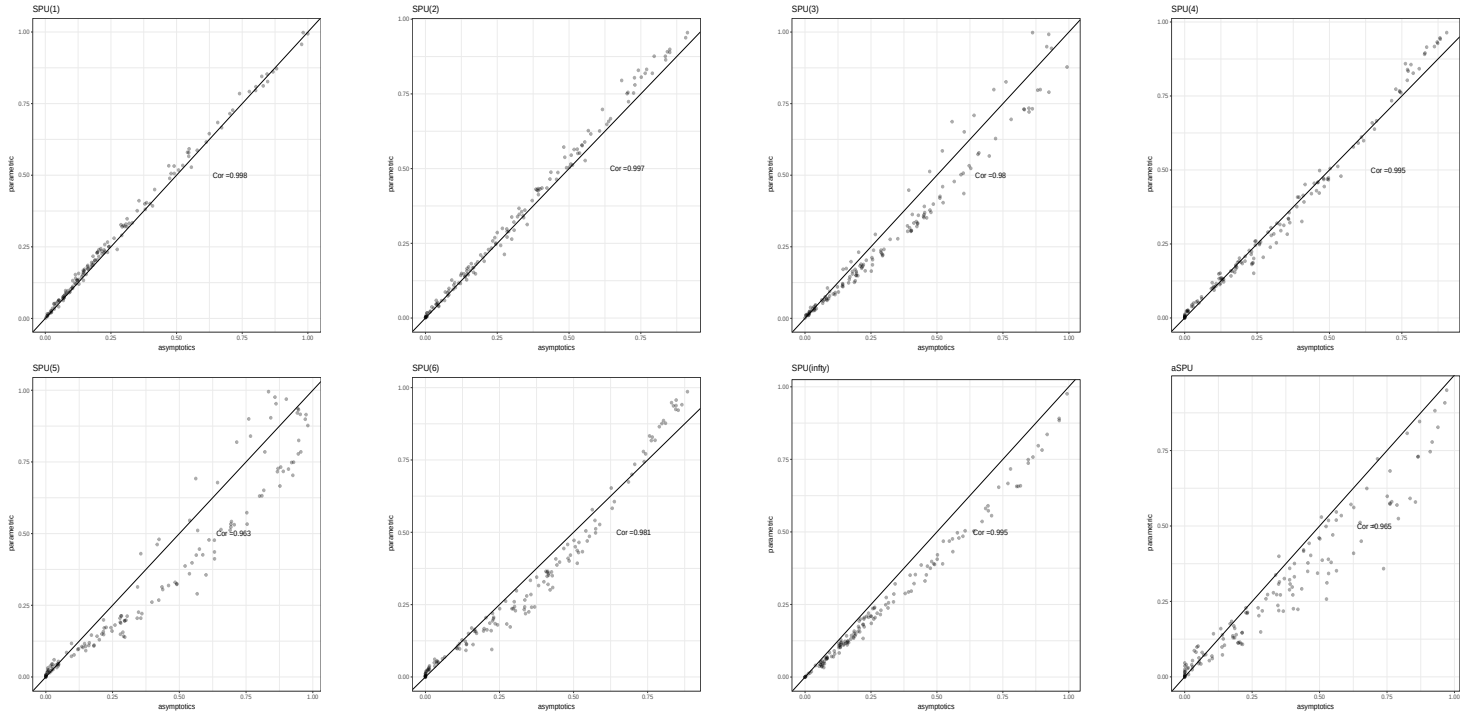

Figure: Comparison between the asymptotic-based and the parametric bootstrap-based *p*-values of SPU(γ) and aSPU.

Application to ADNI data

Table: Results of the ADNI Data Application: KEGG Pathways with p Values $< 3 \times 10^{-4}$ by Any of aSPU, GT, and HDGLM

		p values		
Pathway Name	# G	aSPU	GT	HDGLM
Alzheimer's disease	151	0.0E+00	3.8E-03	1.4E-03
Amyotrophic lateral sclerosis	52	0.0E+00	2.3E-03	3.2E-04
Acute myeloid leukemia	55	0.0E+00	2.6E-03	7.6E-04
Adherens junction	72	9.0E-09	4.4E-01	4.7E-01
Fatty acid degradation	40	5.3E-08	1.6E-02	8.0E-03
Retinol metabolism	61	2.1E-07	4.1E-03	7.9E-04
Tyrosine metabolism	38	4.0E-07	7.7E-03	2.4E-03
Drug metabolism	70	2.2E-05	3.6E-02	2.6E-02
Heparin	26	6.4E-05	6.2E-04	1.1E-05
Metabolism of xenobiotics	68	1.6E-04	9.5E-02	9.1E-02

Acknowledgement

Thank you!

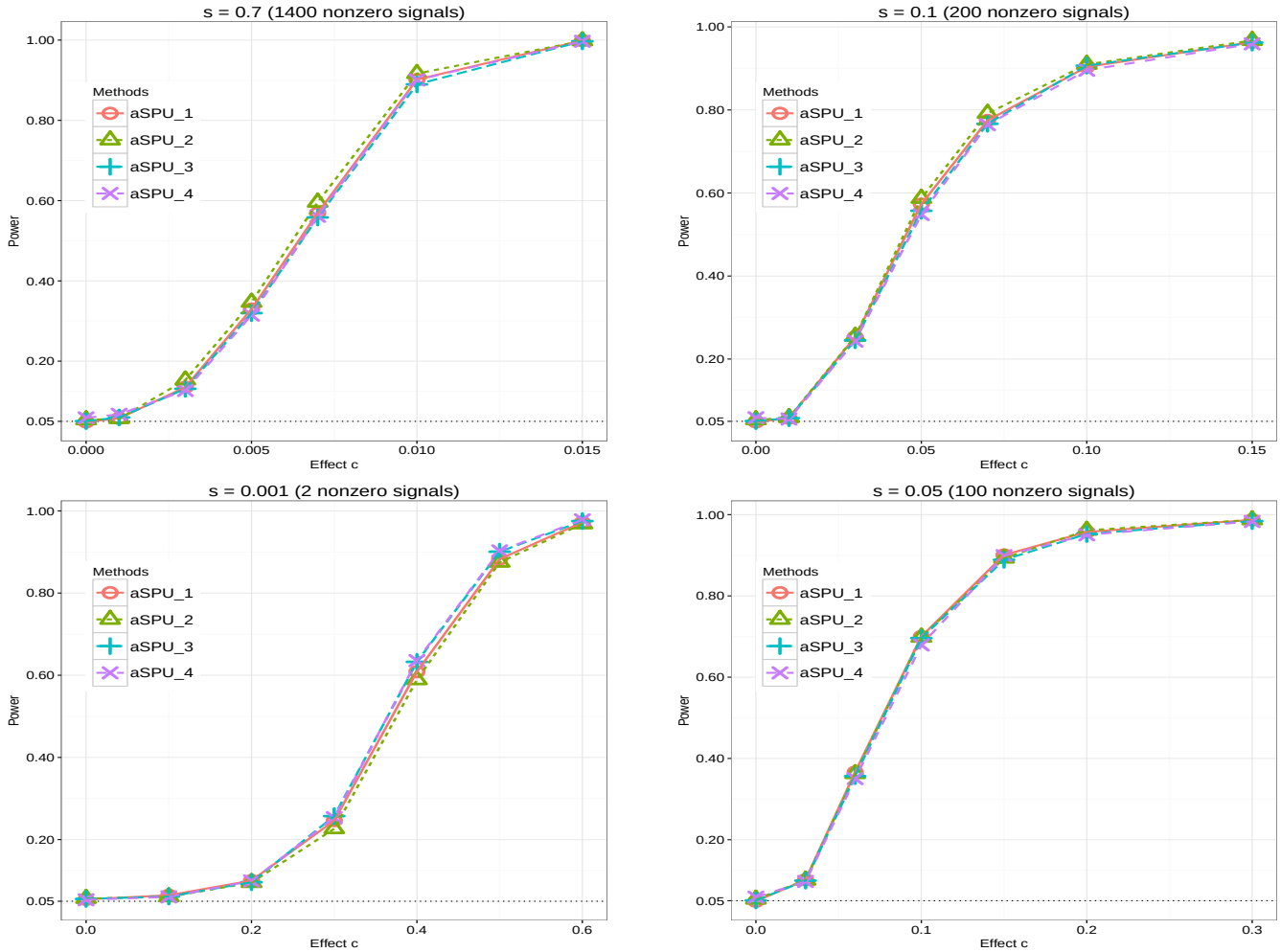

Figure: Empirical powers of aSPU with different Γ set.