An adaptive test for high-dimensional generalized linear models with application to detect gene-environment interactions

Chong Wu
Department of Statistics, Florida State University
Joint work with Gongjun Xu, Xiaotong Shen, & Wei Pan

ENAR 2019

March 25

Outline

- **Problem formulation**
- Method
- Simulation results
- Application to ADNI data

Motivations

Practical motivation: testing gene-environment interactions

- Complex diseases are often caused by the interplay of genes and the environment

Theoretical motivations:

- Testing high-dim groups of parameters with high-dim nuisance parameters is largely untouched
- Existing methods hard to control Type I error rates and maintain high power

Problem formulation

- Y_i is the phenotype (outcome) ($i = 1, \ldots, n$)
- $Z_1, \ldots, Z_q$ are the q covariates (age, gender, environmental effect, genetic effect, etc.) (high-dimensional)
- $X_1, X_2, \ldots, X_p$ are the p gene-environment interactions (high-dimensional)
- $\mu_i = E(Y_i | Z_1, \ldots, Z_q, X_1, \ldots, X_p)$

Model

$$\mu_i = g^{-1}(\alpha_0 + \alpha_1 Z_{i1} + \cdots + \alpha_q Z_{iq} + \beta_1 X_{i1} + \cdots + \beta_p X_{ip})$$

- Hypothesis of no gene-environment interaction effect

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{v.s.} \quad H_1 : \text{At least one } \beta_j \neq 0$$

New statistical challenge

- Estimating α under the H_0 is difficult
- Use a penalized regression framework:

$$\min -L(\alpha) + \lambda P(\alpha)$$

- Ridge: $P(\alpha) = \sum_{j=1}^{q} \alpha_j^2$; Lasso: $P(\alpha) = \sum_{j=1}^{q} |\alpha_j|$
- Lasso yields sparse but biased estimation

Outline

- Problem formulation
- **Method**
- Simulation results
- Application to ADNI data

Existing methods

Method	GESAT (Lin et al., Biostatistics, 2013)	Three step procedure (Zhang and Cheng, JASA, 2017)
Test statistic	SSU + Ridge penalty	$T_{\text{st}} = \max_j \frac{\sqrt{n}\lvert\hat{\beta}^{DL}\rvert}{\text{sd}(\hat{\beta}^{DL})}$
Pros	Fast; easy to use	Powerful under sparse alternative
Cons	Fail to control Type I error rates when q is large	Only for linear models; Lose power under "dense" alternatives

Note: $\hat{\beta}^{DL}$ is the de-sparsified (or de-biased) Lasso: Lasso plus a one step bias correction

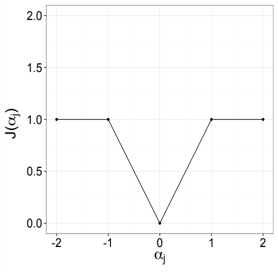
Oracle estimator

- Oracle estimator: MLE if we know which $\alpha_j = 0$
- If we know the oracle estimator, it will reduce to the low-dimensional nuisance parameter situations

Question

How to get the oracle estimator?

Our idea: using TLP to estimate nuisance parameter



$J(\alpha_j)$ with $\tau = 1$

- Truncated Lasso penalty (TLP): $J(\alpha_j) = \min(|\alpha_j|, \tau)$ (Shen et al. JASA, 2012)
- TLP consistently reconstructs the oracle estimator under some mild conditions
- TLP is a non-convex penalty. I develop an R package "glmtlp"

  Online manual: `wuchong.org/glmtlp.html`

New test: iSPU and aiSPU

- Apply the adaptive testing idea to maintain high power across different cases
- Score $U_j = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{0i})X_{ij}, \qquad 1 \leq j \leq p$

  $\hat{\mu}_{0i} = g^{-1}(\hat{\alpha}_0^{\mathrm{TLP}} + Z_{1i}\hat{\alpha}_1^{\mathrm{TLP}} + \cdots + Z_{1q}\hat{\alpha}_q^{\mathrm{TLP}})$
- iSPU(γ): $iSPU(\gamma) = \sum_{j=1}^{p} U_j^{\gamma}$
- iSPU(∞): $iSPU(\infty) = \max_{1\leq j\leq p} nU_j^2/\sigma_{jj}$
- aiSPU: $T_{\mathrm{aiSPU}} = \min_{\gamma\in\Gamma} P_{\mathrm{iSPU}(\gamma)}$
  - $\Gamma = \{1, 2, \ldots, 6, \infty\}$

Asymptotic distribution under the null

Theorem
Under some mild assumptions and the null hypothesis H_0:

- Let Γ be a set of finite positive integers, $[\{\text{iSPU}(\gamma) - \mu(\gamma)\}/\sigma(\gamma)]'_{\gamma\in\Gamma}$ converges weakly to a normal distribution $N(0, R)$ as $n, p \to \infty$
- When $\gamma = \infty$, let $a_p = 2\log p - \log\log p$, for any $x \in \mathbb{R}$, $Pr\{\text{iSPU}(\infty) - a_p \leq x\} \to \exp\{-\pi^{-1/2}\exp(-x/2)\}$ as $n, p \to \infty$
- $[\{\text{iSPU}(\gamma) - \mu(\gamma)\}/\sigma(\gamma)]'_{\gamma\in\Gamma}$ is asymptotically independent with $\text{iSPU}(\infty)$

Outline

- Problem formulation
- Method
- **Simulation results**
- Application to ADNI data

Simulation results: validation of theorem

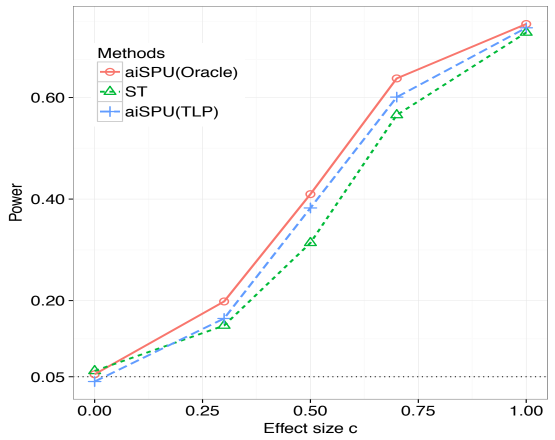
Empirical Type I errors and powers (%) for a linear model with $n = 200$, $p = 1000$, $q = 1000$, and $\eta = 0.99$

Asymptotics (parametric bootstrap)

c	0	0.3	0.5	0.7
iSPU(1)	5.6 (5.4)	6.7 (6.1)	6.6 (6.3)	7.5 (7.2)
iSPU(2)	3.6 (3.3)	4.2 (5.7)	6.6 (8.2)	15.3 (18.9)
iSPU(3)	5.0 (4.8)	6.4 (5.6)	14.6 (13.5)	41.7 (40.1)
iSPU(4)	3.8 (1.8)	9.1 (7.5)	29.5 (26.4)	54.6 (52.1)
iSPU(6)	4.9 (2.2)	18.2 (13.3)	38.8 (33.8)	61.9 (58.2)
iSPU(∞)	3.5 (4.6)	16.1 (18.3)	36.5 (38.7)	61.4 (61.9)
aiSPU	5.3 (4.1)	16.6 (16.5)	38.5 (38.3)	61.4 (60.1)

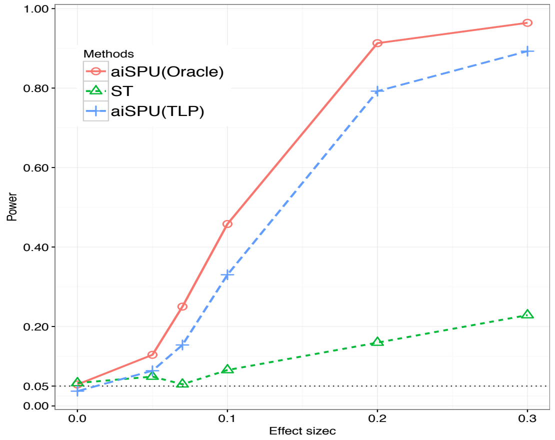
Power comparison under a linear model

Sparse alternative ($\eta = 0.99$)



Power comparison under a linear model

Dense alternative ($\eta = 0.23$)



Type I error rates under a logistic model

Empirical Type I error rates of various tests under $G \times E$ interaction simulations with $n = 2000$ and various q
* Inflated Type I error rates

q	25	50	100	300	500
GESAT	0.061	0.055	0.103*	0.636*	1.000*
aiSPU(Oracle)	0.067	0.049	0.052	0.057	0.047
aiSPU(TLP)	0.061	0.054	0.053	0.042	0.047

Outline

- Problem formulation
- Method
- Simulation results
- **Application to ADNI data**

ADNI data analysis: pathway-gender interactions

- Brain development and adult brain structure differ by gender (Cosgrove et al. 2007)
- 214 healthy controls ($Y = 1$); 364 MCI subjects ($Y = 0$)
- Main effects: years of education, age, intracranial volume measured at baseline, gender, and genetic variants
- Bonferroni correction; 96 KEGG pathways ($0.05/100 = 5 \times 10^{-4}$)
- aiSPU identified one significant pathway *Fructose and mannose metabolism* (hsa00051, p-value $= 3 \times 10^{-4}$);

  GESAT failed to do so (p-value $= 0.016$)

ADNI data analysis: gene-gender interactions

- Candidate gene study (Gene *APOE*)
- aiSPU identified APOE and gender interaction effects (p-value $= 0.039$)

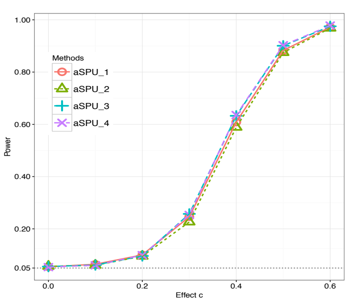
  GESAT failed to identify (p-value $= 0.56$)
- Women who are positive for the APOE $\epsilon 4$ are at greater risk of developing AD than men with this allele (Altmann et al. 2014)

Acknowledgement

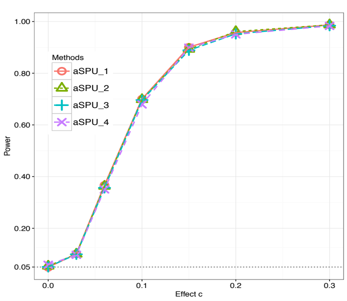
Thank you!

Robustness of choice of Γ

Sparse alternative: 2 "causal"



Dense alternative: 100 "causal



Empirical powers of aSPU with different Γ set. Γ set aSPU_1, aSPU_2, aSPU_3, aSPU_4 represent aSPU with $\Gamma_1 = \{1, 2, \ldots, 4, \infty\}$, $\Gamma_2 = \{1, 2, \ldots, 6, \infty\}$, $\Gamma_3 = \{1, 2, \ldots, 8, \infty\}$, and $\Gamma_4 = \{1, 2, \ldots, 10, \infty\}$, respectively. We set $n = 200$ and $p = 2000$.

Asymptotics-based method

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

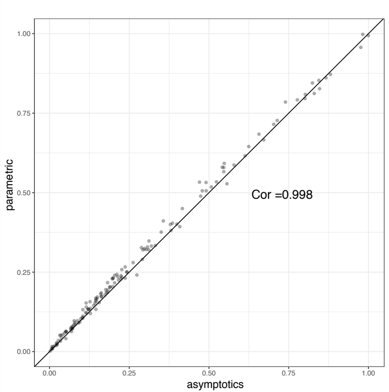
$$p_E = 1 - \int_{\substack{t=(t_\gamma:\text{even } \gamma\in\Gamma)' \\ -\infty \le t_\gamma \le T_E}} N(0, R_E)dt$$

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

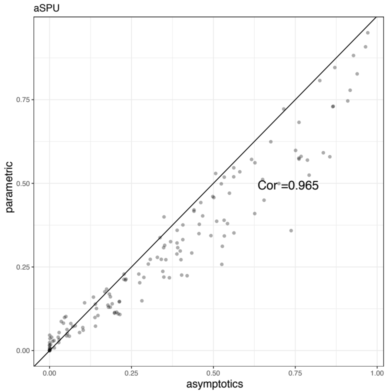
$$p_{\text{aSPU}} = 1 - (1 - p_{\min})^3$$

Application to ADNI data: validation of theorem

SPU(1)



aSPU



Comparison between the asymptotics- and the parametric bootstrap-based p-values for KEGG pathways

More details on proof outline

- For finite γ: if all SNPs are independent, we can apply CLT directly; use Bernstein's block to make the leading term almost independent
- For asymptotically independent: the distribution of SPU(γ) conditional on SPU(∞) is the same as the unconditional version

Difference of convex (DC) algorithm

- Estimate α by minimizing $\min S(\alpha) = -L(\alpha) + \lambda P(\alpha)$
- DC decomposition of $S(\alpha)$:

$$S(\alpha) = S_1(\alpha) - S_2(\alpha)$$

$$S_1(\alpha) = -L(\alpha) + \lambda \sum_{j=1}^{q} |\alpha_j|$$

$$S_2(\alpha) = \lambda \sum_{j=1}^{q} \max(|\alpha_j| - \tau, 0)$$

- Approximate the $S_2(\alpha)$, then we have

$$S^{(m)}(\alpha) = -L(\alpha) + \lambda \sum_{j=1}^{q} |\alpha_j| I(|\hat{\alpha}_j^{(m-1)}| \leq \tau)$$

Details on GESAT

- $Q = (Y - \mu(\hat{\alpha}^R))' X X' (Y - \mu(\hat{\alpha}^R))$
- Follow a mixture of χ^2 distribution under the null
- $\sqrt{n}$-consistent (Knight and Fu 2000): $\sqrt{n}(\hat{\alpha}^R - \alpha) = O_p(1)$
  Only valid when the cov(Z) **is non-negative** (small q)
- Cannot control **Type I error rate when** q **is large**

Details on three-step procedure

- Desparsifying the Lasso: Lasso plus a one step bias correction
- Three-step procedure (Zhang and Cheng, 2017)
  - Random sampling splitting: $\mathcal{D}_1$ & $\mathcal{D}_2$
  - Marginal screening based on $\mathcal{D}_1$
  - Testing after screening based on $\mathcal{D}_2$:
    $T_{\text{nst}} = \max_j \sqrt{n}|\hat{\beta}^{DL}|$; $T_{\text{st}} = \max_j \sqrt{n}|\hat{\beta}^{DL}|/\text{sd}(\hat{\beta}^{DL})$
  - Error term will be **out of control** for other type statistics (Sum, SSU)
  - Only apply to a **linear model**

Asymptotic power analysis

$$Pr(T_{\text{aiSPU}} = \min_{\gamma \in \Gamma} P_{\text{iSPU}(\gamma)} < p^*_\alpha) \geq Pr(P_{\text{iSPU}(\gamma)} < p^*_\alpha)$$

- p^*_α: critical threshold under H_0 with significance level α
- The asymptotic power of aiSPU is 1 if there exists $\gamma \in \Gamma$ such that $Pr(P_{\text{iSPU}(\gamma)} < p^*_\alpha) \to 1$

Asymptotic power analysis

- Unknown truth: size of $P_0 = \{j : \beta_j \neq 0\}$ is $k = p^{1-\eta}$
- "Dense" alternatives ($\eta < 1/2$)
  - All variables are associated and with the same effect size: iSPU(1) is asymptotically most powerful among iSPU(γ)'s
  - Half variables are positively associated; the other half are negatively associated: iSPU(2) is asymptotically most powerful
- "Sparse" alternatives ($\eta > 1/2$):
  - The asymptotic power of iSPU with finite γ is strictly less than 1
  - iSPU(∞) is more powerful