Polygenic testing and two-sample testing with high-dimensional data

Wei Pan

Division of Biostatistics, School of Public Health, University of Minnesota,
Minneapolis, MN 55455

Shanghai, Nov 2014

Outline

- Introduction: problem.
- Part 1: Polygenic testing
  ISC-Poly vs aSPU
- Part 2: 2-sample tests for high-dim data
  Review: some existing tests;
  SPU/aSPU
  Comparison, theory
- Application in neuroimaging?
- Discussion.

Introduction

- Problem:
  - Given: a binary disease indicator Y_i for subject i; a group of of (genome-wide) genetic variants (SNPs) (additively) coded as $X_i = (X_{i1}, ..., X_{ik})'$ with $X_{ij} = 0$, 1 or 2; $i = 1, ..., n << k$.
  - Q: any association between Y_i and X_i?
  - Approaches: global testing.
- Polygenic testing: X_i genome-wide; 100s–1000s genes.
  Why? missing heritability from genome-wide association studies (GWAS);
  Any association?
- Example: the International Schizophrenia Consortium (ISC) (2009, *Nature*)

- Goal: to maximize the power of a test
- Logistic reg model:

$$\text{Logit}[Pr(Y_i = 1)] = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j.$$

or, for $j = 1, ..., k$,

$$\text{Logit}[Pr(Y_i = 1)] = \beta_{M,j0} + X_{ij}\beta_{M,j}.$$

- H_0: $\beta = (\beta_1, ..., \beta_k)' = 0$, or $\beta_M = (\beta_{M,1}, ..., \beta_{M,k})' = 0$.
- Remark: other phenotypes or covariates can be accommodated.
- The score vector $U = (U_1, ..., U_k)'$ and its covariance:

$$U = \sum_{i=1}^{n}(Y_i - \bar{Y})X_i,$$

$$V = Cov(U|H_0) = \bar{Y}(1 - \bar{Y})\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'.$$

Some existing tests

- Five global tests (Pan 2009, *Genetic Epi*) for $k < n$:

$$T_{Score} = U'V^{-1}U,$$

$$T_{SSU} = U'U = \sum_{j=1}^{k} U_j^2,$$

$$T_{SSUw} = U'\mathrm{diag}(V)^{-1}U = \sum_{j=1}^{k} U_j^2/V_{jj},$$

$$T_{UminP} = \max_{j=1}^{k} U_j^2/V_{jj},$$

$$T_{Sum} = 1'U/\sqrt{1'V1} = \sum_{j=1}^{k} U_j/\sqrt{1'V1},$$

where $V_{jj} = \mathrm{Var}(U_j)$.

- Variance components tests:
  Sum of Squared Score (SSU) test (Pan 2009): assuming $\beta_1, \ldots, \beta_k \sim F(0, \tau^2)$, H_0: $\tau^2 = 0$,
  $T_{SSU} = U'U = \sum_{j=1}^{k} U_j^2$.
  SSU test: equivalent to KMR (Liu et al 2008) with $K = XX'$ (Pan 2011), i.e. SKAT with no weighting and a linear kernel (Wu et al 2011); C-alpha (Neal et al 2011), an EB test (Goeman et al 2006), GDBR/MDMR (Schork et al), ...
- UminP test: $T_{UminP} = \max_{j=1}^{k} U_j^2 / V_{jj}$,
  close to $T_{maxU} = \max_{j=1}^{k} |U_j|$
- A challenge: no uniformly most powerful test!

- Adaptive tests: with weights $\zeta = (\zeta_1, ..., \zeta_k)'$,

$$T_G = \zeta' U = \sum_{j=1}^{k} \zeta_j U_j,$$

  - aSum (Han and Pan 2010): $\zeta_j = -1$ (or 1) if $\hat{\beta}_{M,j} < 0$ (or > 0) and p-value $p_j < 0.1$;
  - PWST (Zhang et al 2011): $\zeta_j = 2(p_j - 0.5)$;
  - EREC (Lin and Tang 2011): $\zeta_j = \hat{\beta}_{M,j} \pm d$.

- Note: $\hat{\beta}_M = Diag(V)^{-1}U + O_p(1/n)$,
  1) If $|\hat{\beta}_M|$ is large, $\zeta \approx \hat{\beta}_M \propto U \Longrightarrow$ EREC $\approx$ SSU;
  2) If $|\hat{\beta}_M|$ is small, $\zeta \approx \pm d \Longrightarrow$ EREC $\approx$ Sum;
- ...
- Key: how to choose ζ? Is any given choice of ζ sufficiently adaptive?
  Our answers:

New Tests: SPU and aSPU

- $\zeta_j = f(U_j) = U_j^{\gamma-1}$ for $\gamma \geq 1$;
- SPU tests: for a $\gamma \geq 1$,

$$T_{SPU(\gamma)} = \sum_{j=1}^{k} U_j^{\gamma}.$$

$$T_{SPU(\infty)} \propto \lim_{\gamma \to \infty} \left(\sum_{j=1}^{k} |U_j|^{\gamma} \right)^{1/\gamma} = \max_{j=1}^{k} |U_j|.$$

- Special cases:
  SPU(1) = Sum;
  SPU(2) = SSU;
  SPU(∞) = maxU $\approx$ UminP;
- Intuition in the choice of γ:
  1) the more sparse the signals, the larger γ;
  2) if (most) associations in one direction, then use an odd γ.

- Our experience: often $SPU(8) \approx SPU(16) \approx SPU(\infty)$;
  If $SPU(\gamma) \approx SPU(\infty)$, then no need to increase γ.
- In parctice, how to choose γ?
  choose the one giving the most significant p-value?
- Use an adaptive SPU (aSPU) test:

  $$T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)},$$

  where $P_{SPU(\gamma)}$ is the p-value of $SPU(\gamma)$, and
  $\Gamma = \{1, 2, ..., 8, \infty\}$.
- Computing: one loop of permutations or parameteric bootstrap is sufficient to calculate the p-values of $SPU(\gamma)$ for $\gamma \in \Gamma$ and aSPU tests!
- Ref: Pan et al (2014, *Genetics*)

Connections

- The ISC-Poly test:
  1) Divide data $D = D_1 \cup D_2$;
  2) $w_j = w_j(D_1) = \hat{\beta}_{M,j} I(p_j < P_T)$ from the marginal model;
  3) $s_i = \sum_j w_j(D_1) X_{ij}(D_2)$;
  4) t-test on s_i's with $i \in D_2$;
- The ISC-Poly is the same as the Sum (Poly-Sum) test on H_0': $\alpha_1 = 0$ in

$$\text{Logit}[Pr(Y_i = 1)] = \alpha_0 + \alpha_1 \sum_{j=1} w_j X_{ij},$$

  with the new genotype score $w_j X_{ij}$ and $i \in D_2$.
- Can construct Poly-SSU, Poly-UminP, ...
- Key: use a half of the sample to construct weights w_j's; use the other half for hypothesis testing.
  sample splitting is **not** efficient!

- Some algebra (and asymptotics) shows

$$T_{Poly(P_T)} \propto \frac{\sum_j U_j(D_1)U_j(D_2)I(p_j(D_1) < P_T)}{\mathrm{Var}(U_j(D_1))},$$

- Better to use

$$T_{tSSUw(P_T)} = \frac{\sum_j U_j(D)U_j(D)I(p_j(D) < P_T)}{\mathrm{Var}(U_j(D))},$$

- Thresholding and inverse-variance weighting are not really effective $\Longrightarrow$

$$T_{SSU} = \sum_j U_j(D)U_j(D),$$

  or even better, SPU(γ), and aSPU!

- aSSU (Pan and Shen 2011, *Genetic Epi*; Fan 1997, *JASA*) vs aSPU (Pan et al 2014, *Genetics*)...

Simulations

Empirical Type I error rate (for $OR = 1$) and power (for $a > 1$) for polygenic tests (with sample splitting) and SPU/aSPU tests (without sample splitting) for 1000 independent SNPs, including k_1 causal SNPs with OR_j's $\sim U(1, a)$.

		Null	$k_1 = 20$			$k_1 = 50$			$k_1 = 100$		
Test	P_T	$a = 1$	$a = 1.2$	1.3	1.4	1.1	1.2	1.3	1.1	1.15	1.2
Poly-ISC	0.05	.044	.109	.344	.728	.056	.298	.769	.093	.240	.674
	0.1	.053	.115	.299	.676	.057	.311	.767	.106	.284	.738
	0.5	.041	.101	.258	.488	.078	.298	.731	.121	.377	.769
Poly-Sum	0.05	.044	.111	.344	.730	.056	.299	.769	.093	.240	.674
	0.1	.053	.114	.299	.676	.057	.311	.768	.106	.284	.738
	0.5	.042	.103	.258	.489	.078	.299	.731	.121	.377	.768
Poly-SSU	0.05	.046	.163	.610	.925	.066	.350	.887	.086	.228	.645
	0.1	.041	.143	.593	.917	.072	.379	.896	.094	.253	.693
	0.5	.030	.124	.584	.907	.062	.363	.906	.093	.284	.760
Poly-SSUw	0.05	.043	.144	.494	.845	.065	.306	.838	.074	.220	.595
	0.1	.038	.113	.418	.781	.060	.319	.827	.078	.233	.631
	0.5	.023	.053	.198	.398	.041	.179	.553	.091	.184	.525
Poly-UminP	0.05	.050	.134	.458	.787	.072	.191	.642	.066	.131	.364
	0.1	.039	.123	.415	.751	.063	.202	.592	.064	.136	.326
	0.5	.039	.097	.287	.590	.063	.166	.442	.066	.111	.241
SPU(1)		.053	.139	.182	.296	.162	.439	.733	.490	.781	.946
SPU(2)		.062	.234	.565	.819	.158	.657	.966	.327	.756	.981
SPU(4)		.058	.364	.817	.984	.159	.763	.994	.292	.782	.986
SPU(8)		.049	.348	.830	.982	.122	.630	.978	.166	.495	.918
SPU(16)		.056	.308	.769	.961	.105	.465	.924	.114	.339	.744
SPU(32)		.056	.293	.741	.950	.103	.413	.903	.110	.307	.682
SPU(∞)		.058	.297	.737	.949	.109	.408	.887	.115	.307	.674
aSPU		.055	.348	.806	.971	.203	.747	.992	.464	.877	.995

Example

- SAGE GWAS on alcohol dependence (Bierut et al 2010); $n = 1165$ cases $+1379$ controls; a total of 948,658 SNPs; 607,033 SNPs after QC; **None** reseached the genome-wide significance by univariate testing!
- Previous twin/familial studies showed heritability of alcohol dependence!
- Any here?
- Use Plink to trim to 62,801 nearly uncorrelated SNPs ($r^2 \leq 0.1$ with a sliding window of 200 SNPs and a step size of 20 SNPs).
- Results: based on 10 million permutations!

Test	P_T	p-value
Poly-ISC	0.01	0.0042
	0.05	7.29×10^{-5}
	0.10	5.04×10^{-5}
	0.20	1.61×10^{-5}
	0.30	5.85×10^{-6}
	0.40	1.37×10^{-6}
	0.50	1.23×10^{-6}
Bonferroni-adjusted p-value		8.64×10^{-6}
SPU(1)		5.12×10^{-4}
SPU(2)		$< 1 \times 10^{-7}$
SPU(3)		0.0433
SPU(4)		$< 1 \times 10^{-7}$
SPU(5)		0.1925
SPU(6)		6.54×10^{-5}
SPU(7)		0.3111
SPU(8)		0.0235
SPU(∞)		0.3383
aSPU		9.00×10^{-7}

Part 2: two-sample tests

- Set-up: two samples, $\{\mathbf{x}_{1i}, i = 1, 2, \ldots, n_1\}$ and $\{\mathbf{x}_{2j}, j = 1, 2, \ldots, n_2\}$ with $p > \max\{n_1, n_2\}$.
  H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. (Or more generally, H_0: $F_1 = F_2$.)
- Sample means and covariance matrices: $n = n_1 + n_2$,
  $\bar{\mathbf{x}}_k = \sum_{i=1}^{n_k} \mathbf{x}_{ki} / n_k$.
  $S_n = \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k\right)\left(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k\right)^T / n$.
- Bai and Saranadasa (1996):

$$Z = \frac{\frac{n_1 n_2}{n_1+n_2}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^T\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) - \mathrm{tr} S_n}{\sqrt{\frac{2(n+1)}{n}} B_n}, \qquad (1)$$

  Under H_0, $Z \xrightarrow{D} N(0, 1)$.
- Key:

$$M_n = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^T\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) - \frac{n_1 + n_2}{n_1 n_2} \mathrm{tr} S_n. \qquad (2)$$

- Chen et al (2010, *Ann Statist*):

$$T_n = \frac{\sum_{i\neq j}^{n_1} \mathbf{x}_{1i}^T \mathbf{x}_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} \mathbf{x}_{2i}^T \mathbf{x}_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \mathbf{x}_{1i}^T \mathbf{x}_{2j}}{n_1 n_2}, \quad (3)$$

  which is the terms after removing $\sum_{i=1}^{n_k} \mathbf{x}_{ki}^T \mathbf{x}_{ki}$ for $k = 1, 2$ from $\|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\|^2$. Hence

$$\frac{T_n - \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{\sqrt{\mathrm{Var}(T_n)}} \xrightarrow{D} N(0,1) \quad (4)$$

  as $n \longrightarrow \infty$ and $p \longrightarrow \infty$.

- Cai et al (2014, *JRSS-B*): $\boldsymbol{\delta}^{\mathbf{A}} = \mathbf{A}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$,

$$M_{\mathbf{A}} = \frac{n_1 n_2}{n_1 + n_2} \max_{1\leq i\leq p} \frac{(\delta_i^{\mathbf{A}})^2}{b_{ii}}, \quad (5)$$

  an asymptotic extreme value distribution.

- Chen et al (2014):

$$L_n(s) = \sum_{j=1}^{p} \left\{ n\left(\bar{\mathbf{x}}_{1,j} - \bar{\mathbf{x}}_{2,j}\right)^2 - 1 \right\} I\left\{ n\left(\bar{\mathbf{x}}_{1,j} - \bar{\mathbf{x}}_{2,j}\right)^2 > \lambda_n(s) \right\}, \tag{6}$$

with $\lambda_n(s) = 2s\log p$ as the thresholding level. Then

$$M_{L_n} = \max_{s \in (0,1-\eta)} \frac{L_n(s) - \hat{\mu}_{L_n(s),0}}{\hat{\sigma}_{L_n(s),0}}, \tag{7}$$

with an asymptotic extreme value distribution.

- Our SPU tests:

$$\mathbf{U} = \frac{n_1 + n_2}{n_1 n_2} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right).$$

  Then for a positive integer γ

$$\mathrm{SPU}(\gamma) = \sum_{j=1}^{p} \left(\bar{\mathbf{x}}_{1,j} - \bar{\mathbf{x}}_{2,j}\right)^{\gamma},$$

$$\mathrm{SPU}(\infty) = \max_{j=1}^{p} \left(\frac{\bar{\mathbf{x}}_{1,j}}{\sigma_{1,j}} - \frac{\bar{\mathbf{x}}_{2,j}}{\sigma_{2,j}}\right)^{2}.$$

- Remarks:
  Chen et al (2010): $\sim$ SPU(2)=SSU;
  Chen et al (2014): $\sim$ tSPU(2)=aSPU(2)=aSSU;
  Cai et al (2014): $\sim$ SPU(∞).

Theorem for SPU tests

Let Γ be a set of finite positive integers. Under H_0, we have

$$\{\sigma(\gamma)^{-1}(\mathrm{SPU}(\gamma)-\mu(\gamma)):\gamma\in\Gamma\}' \xrightarrow{d} N(\mathbf{0},\boldsymbol{\xi}),$$

and for $x \in \mathbb{R}$,

$$P(n\mathrm{SPU}(\infty)-a_p\le x)\to\exp\left\{-\frac{1}{\sqrt{\pi}}\exp\left(-\frac{x}{2}\right)\right\}$$

as $n, p \to \infty$, where $a_p = 2\log p - \log\log p$ and
$n = n_1 n_2/(n_1+n_2)$.
Moreover, $\{\sigma(\gamma)^{-1}(\mathrm{SPU}(\gamma)-\mu(\gamma)):\gamma\in\Gamma\}$ and $n\mathrm{SPU}(\infty)-a_p$ are asymptotically independent.

Simulations

- Simulation set-ups follow Chen et al (2014).
- $n_1 = 30$, $n_2 = 40$, $p = 200$.
- Under H_0, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$; under H_1, $\boldsymbol{\mu}_1 = \mathbf{0}$, and $\boldsymbol{\mu}_2$ has $\lfloor p^{1-\beta} \rfloor$ non-zero entries of equal value, which are uniformly allocated among $\{1, 2, \ldots, p\}$. $\beta = 0, 0.1, 0.2, \ldots, 0.9$.
- The values of the non-zero entries are $\sqrt{2r \log p(1/n_1 + 1/n_2)}$. $r = 0.1, 0.2, 0.3, 0.4$.
- $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = (\sigma_{ij})$, where $\sigma_{ij} = \rho^{|i-j|}$. $\rho = 0.6$.
- Results:
- Based on 1000 replicates; all used permutations $B = 1000$
- Used true $\Omega = \Sigma^{-1}$ if needed.

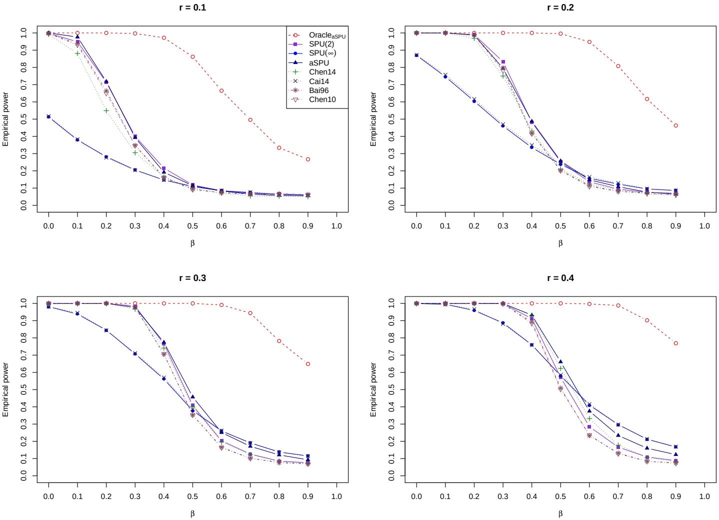

Figure: No data transformation

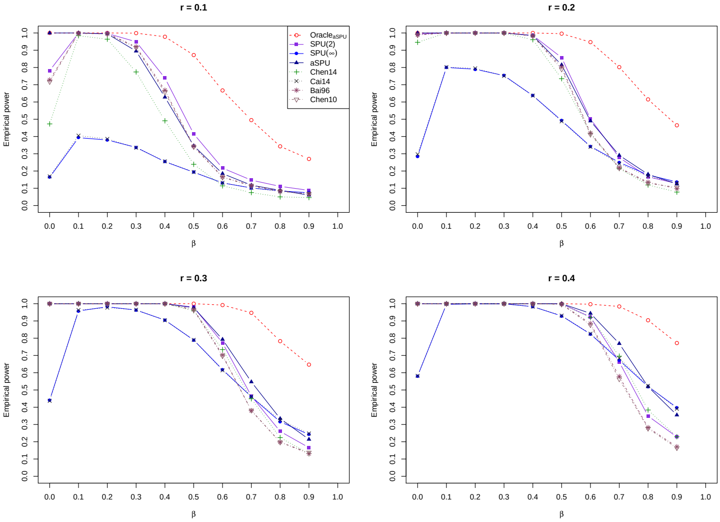

Figure: Data transformation with $\boldsymbol{\Omega}^{1/2}$

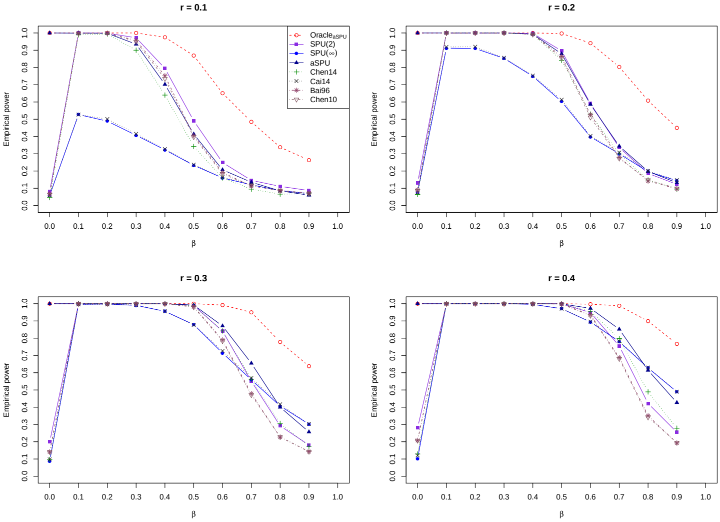

Figure: Data transformation with $\boldsymbol{\Omega}$

Discussion

- Conclusion: aSPU test is promising (and general/flexible)
- Current work:
  applied to real data;
  develop an R package;
- Extensions:
  Pathway analysis; ongoing ...
  Multivariate (neuroimaging) traits-single SNP (Zhang et al 2014);
  Multivariate traits-multiple SNPs; ongoing ...
  To familial and/or longitudinal data; ongoing ...

Another Application

- To brain connectivity data: $k >> n$; Kim et al (2014).
- Problem: based on fMRI data, estimate a functional connectivity (FC) network for each subject using marginal correlations (i.e. sample covariance) or partial correlations (i.e. precision matrix).
- Key Q: group comparisons; not many studies ...
- Example: a rs-fMRI dataset (Wozniak et al 2013);
  Group 1: patients with fatal alcohol spectrum disorder (FASD), $n_1 = 24$;
  Group 2: controls, $n_2 = 31$;
  $N = 62 + 12 = 74$ cortical and sub-cortical ROIs; $k = 2701$ possible edges;
  Each subject measured at 180 time points;

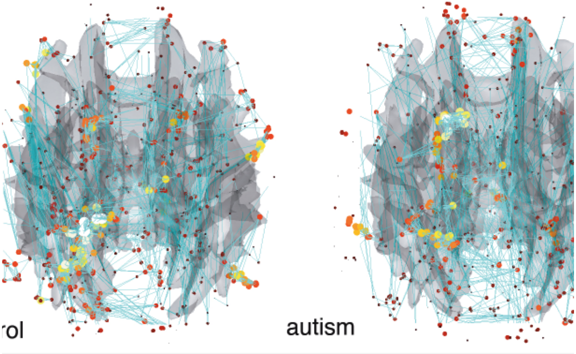
rol
autism

Table: P-values after adjusting for age and gender for the FASD data.

Test	SPU(1)	SPU(2)	SPU(3)	SPU(4)	SPU(5)	SPU(6)	SPU(7)	SPU(8)	SPU(∞)	aSPU
P-value	0.009	0.312	0.085	0.348	0.236	0.391	0.366	0.437	0.759	0.031
Test	MDMR	DiProPerm	nbs(0.1)	nbs(0.25)	nbs(0.5)	nbs(0.75)	CharPath	Eclust	Eglob	Eloc
P-value	0.468	-	0.009	0.017	0.064	0.081	0.673	0.862	0.919	0.925

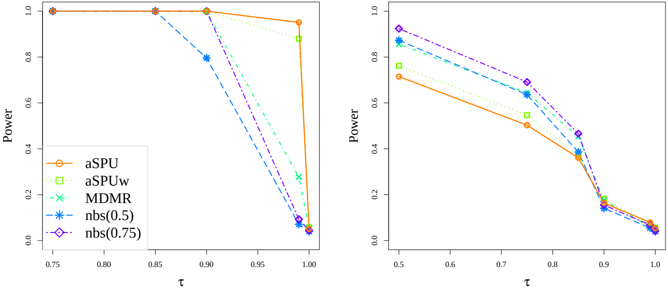

Figure: Sparse networks: empirical Type I error (for $\tau = 1$) and power (for $\tau < 1$) based on 1000 simulations.

Acknowledgement

- This research was supported by NIH.
- Polygenic testing: Peng Wei, Yue-Ming Chen;
- SPU/aSPU for RVs: Peng Wei, Junghi Kim, Yiwei Zhang, Xiaotong Shen;
- 2-sample tests: Lifeng Lin, Gongjun Xu.
- You can download our papers from http://www.biostat.umn.edu/rrs.php

- **Thank you!**