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

- 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?

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 \( j = 1, \ldots, k, \)

\[
\text{Logit}[Pr(Y_i = 1)] = \beta_{M,j_0} + X_{ij} \beta_{M,j}.
\]

\( H_0: \beta = (\beta_1, \ldots, \beta_k)' = 0, \) or \( \beta_M = (\beta_{M,1}, \ldots, \beta_{M,k})' = 0. \)

Remark: other phenotypes or covariates can be accommodated.

The score vector \( U = (U_1, \ldots, U_k)' \) and its covariance:

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

\[
V = \text{Cov}(U|H_0) = \bar{Y}(1 - \bar{Y}) \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'.
\]
Some existing tests

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

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

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

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

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

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

where $V_{jj} = \text{Var}(U_j)$. 
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 $\zeta = (\zeta_1, \ldots, \zeta_k)'$,

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

- 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: $\hat{\beta}_M = \text{Diag}(V)^{-1} U + O_p(1/n)$,
1) If $|\hat{\beta}_M|$ is large, $\zeta \approx \hat{\beta}_M \propto U \implies \text{EREC} \approx \text{SSU}$;
2) If $|\hat{\beta}_M|$ is small, $\zeta \approx \pm d \implies \text{EREC} \approx \text{Sum}$;

... 

Key: how to choose $\zeta$? Is any given choice of $\zeta$ 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(\infty) = \max U \approx U_{\text{min}} P;

- Intuition in the choice of \( \gamma \):
  1) the more sparse the signals, the larger \( \gamma \);
  2) if (most) associations in one direction, then use an odd \( \gamma \).
Our experience: often $SPU(8) \approx SPU(16) \approx SPU(\infty)$; If $SPU(\gamma) \approx SPU(\infty)$, then no need to increase $\gamma$.

In practice, how to choose $\gamma$? 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 $\Gamma = \{1, 2, \ldots, 8, \infty\}$.

Computing: one loop of permutations or parametric 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

\[ T_{Poly}(P_T) \propto \frac{\sum_j U_j(D_1)U_j(D_2)I(p_j(D_1) < P_T)}{\text{Var}(U_j(D_1))}, \]

Better to use

\[ T_{tSSUw}(P_T) = \frac{\sum_j U_j(D)U_j(D)I(p_j(D) < P_T)}{\text{Var}(U_j(D))}, \]

Thresholding and inverse-variance weighting are not really effective \( \implies \)

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

or even better, SPU(\( \gamma \)), and aSPU!

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 $\sim U(1, a)$. 
<table>
<thead>
<tr>
<th>Test</th>
<th>$P_T$</th>
<th>Null</th>
<th>$k_1 = 20$</th>
<th>$k_1 = 50$</th>
<th>$k_1 = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$a = 1$</td>
<td>$a = 1.2$</td>
<td>$a = 1.3$</td>
<td>$a = 1.4$</td>
</tr>
<tr>
<td>Poly-ISC</td>
<td>0.05</td>
<td>0.044</td>
<td>0.109</td>
<td>0.344</td>
<td>0.728</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.053</td>
<td>0.115</td>
<td>0.299</td>
<td>0.676</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.041</td>
<td>0.101</td>
<td>0.258</td>
<td>0.488</td>
</tr>
<tr>
<td>Poly-Sum</td>
<td>0.05</td>
<td>0.044</td>
<td>0.111</td>
<td>0.344</td>
<td>0.730</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.053</td>
<td>0.114</td>
<td>0.299</td>
<td>0.676</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.042</td>
<td>0.103</td>
<td>0.258</td>
<td>0.489</td>
</tr>
<tr>
<td>Poly-SSU</td>
<td>0.05</td>
<td>0.046</td>
<td>0.163</td>
<td>0.610</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.041</td>
<td>0.143</td>
<td>0.593</td>
<td>0.917</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.030</td>
<td>0.124</td>
<td>0.584</td>
<td>0.907</td>
</tr>
<tr>
<td>Poly-SSUw</td>
<td>0.05</td>
<td>0.043</td>
<td>0.144</td>
<td>0.494</td>
<td>0.845</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.038</td>
<td>0.113</td>
<td>0.418</td>
<td>0.781</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.023</td>
<td>0.053</td>
<td>0.198</td>
<td>0.398</td>
</tr>
<tr>
<td>Poly-UminP</td>
<td>0.05</td>
<td>0.050</td>
<td>0.134</td>
<td>0.458</td>
<td>0.787</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.039</td>
<td>0.123</td>
<td>0.415</td>
<td>0.751</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.039</td>
<td>0.097</td>
<td>0.287</td>
<td>0.590</td>
</tr>
<tr>
<td>SPU(1)</td>
<td></td>
<td>.053</td>
<td>.139</td>
<td>.182</td>
<td>.296</td>
</tr>
<tr>
<td>SPU(2)</td>
<td></td>
<td>.062</td>
<td>.234</td>
<td>.565</td>
<td>.819</td>
</tr>
<tr>
<td>SPU(4)</td>
<td></td>
<td>.058</td>
<td>.364</td>
<td>.817</td>
<td>.984</td>
</tr>
<tr>
<td>SPU(8)</td>
<td></td>
<td>.049</td>
<td>.348</td>
<td>.830</td>
<td>.982</td>
</tr>
<tr>
<td>SPU(16)</td>
<td></td>
<td>.056</td>
<td>.308</td>
<td>.769</td>
<td>.961</td>
</tr>
<tr>
<td>SPU(32)</td>
<td></td>
<td>.056</td>
<td>.293</td>
<td>.741</td>
<td>.950</td>
</tr>
<tr>
<td>SPU(∞)</td>
<td></td>
<td>.058</td>
<td>.297</td>
<td>.737</td>
<td>.949</td>
</tr>
<tr>
<td>aSPU</td>
<td></td>
<td>.055</td>
<td>.348</td>
<td>.806</td>
<td>.971</td>
</tr>
</tbody>
</table>
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 researched 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!
<table>
<thead>
<tr>
<th>Test</th>
<th>$P_T$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly-ISC</td>
<td>0.01</td>
<td>0.0042</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>$7.29 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>$5.04 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>$1.61 \times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>$5.85 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>$1.37 \times 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>$1.23 \times 10^{-6}$</td>
</tr>
<tr>
<td>Bonferroni-adjusted p-value</td>
<td></td>
<td>$8.64 \times 10^{-6}$</td>
</tr>
<tr>
<td>SPU(1)</td>
<td></td>
<td>$5.12 \times 10^{-4}$</td>
</tr>
<tr>
<td>SPU(2)</td>
<td></td>
<td>$&lt; 1 \times 10^{-7}$</td>
</tr>
<tr>
<td>SPU(3)</td>
<td></td>
<td>0.0433</td>
</tr>
<tr>
<td>SPU(4)</td>
<td></td>
<td>$&lt; 1 \times 10^{-7}$</td>
</tr>
<tr>
<td>SPU(5)</td>
<td></td>
<td>0.1925</td>
</tr>
<tr>
<td>SPU(6)</td>
<td></td>
<td>$6.54 \times 10^{-5}$</td>
</tr>
<tr>
<td>SPU(7)</td>
<td></td>
<td>0.3111</td>
</tr>
<tr>
<td>SPU(8)</td>
<td></td>
<td>0.0235</td>
</tr>
<tr>
<td>SPU($\infty$)</td>
<td></td>
<td>0.3383</td>
</tr>
<tr>
<td>aSPU</td>
<td></td>
<td>$9.00 \times 10^{-7}$</td>
</tr>
</tbody>
</table>
Part 2: two-sample tests

- Set-up: two samples, \( \{x_{1i}, i = 1, 2, \ldots, n_1\} \) and \( \{x_{2j}, j = 1, 2, \ldots, n_2\} \) with \( p > \max\{n_1, n_2\} \).
  
  \( H_0: \mu_1 = \mu_2. \) (Or more generally, \( H_0: F_1 = F_2. \))

- Sample means and covariance matrices: \( n = n_1 + n_2, \)
  \[ \bar{x}_k = \frac{\sum_{i=1}^{n_k} x_{ki}}{n_k}. \]
  \[ S_n = \sum_{k=1}^{2} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k) (x_{ki} - \bar{x}_k)^T / n. \]

- Bai and Saranadasa (1996):
  \[
  Z = \frac{n_1 n_2}{n_1 + n_2} \left( \bar{x}_1 - \bar{x}_2 \right)^T \left( \bar{x}_1 - \bar{x}_2 \right) - \frac{n_1 + n_2}{n_1 n_2} \text{tr} S_n \\
  \cdot \sqrt{\frac{2(n+1)}{n} B_n}, \tag{1}
  \]

Under \( H_0, Z \overset{D}{\to} N(0, 1). \)

- Key:
  \[
  M_n = \left( \bar{x}_1 - \bar{x}_2 \right)^T \left( \bar{x}_1 - \bar{x}_2 \right) - \frac{n_1 + n_2}{n_1 n_2} \text{tr} S_n. \tag{2}
  \]

\[ T_n = \frac{\sum_{i \neq j}^{n_1} x_{1i}^T x_{1j}}{n_1(n_1 - 1)} + \frac{\sum_{i \neq j}^{n_2} x_{2i}^T x_{2j}}{n_2(n_2 - 1)} - 2 \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} x_{1i}^T x_{2j}}{n_1 n_2}, \]  

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

\[ \frac{T_n - \|\mu_1 - \mu_2\|^2}{\sqrt{\text{Var}(T_n)}} \xrightarrow{D} N(0, 1) \]

as \( n \to \infty \) and \( p \to \infty \).

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

\[ M_A = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \leq i \leq p} \frac{(\delta^A_i)^2}{b_{ii}}, \]

an asymptotic extreme value distribution.
Chen et al (2014):

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

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

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

with an asymptotic extreme value distribution.
Our SPU tests:

\[ U = \frac{n_1 + n_2}{n_1 n_2} (\bar{x}_1 - \bar{x}_2). \]

Then for a positive integer \( \gamma \)

\[ SPU(\gamma) = \sum_{j=1}^{p} (\bar{x}_{1,j} - \bar{x}_{2,j})^{\gamma}, \]

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

Remarks:
Chen et al (2010): \( \sim \) SPU(2)=SSU;
Chen et al (2014): \( \sim \) tSPU(2)=aSPU(2)=aSSU;
Theorem for SPU tests

Let $\Gamma$ be a set of finite positive integers. Under $H_0$, we have

$$\{\sigma(\gamma)^{-1}(\text{SPU}(\gamma) - \mu(\gamma)) : \gamma \in \Gamma\}' \overset{d}{\to} N(0, \xi),$$

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

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

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, $\{\sigma(\gamma)^{-1}(\text{SPU}(\gamma) - \mu(\gamma)) : \gamma \in \Gamma\}$ and $n\text{SPU}(\infty) - a_p$ are asymptotically independent.
Simulations

- \( n_1 = 30, n_2 = 40, p = 200. \)
- Under \( H_0, \mu_1 = \mu_2 = 0; \) under \( H_1, \mu_1 = 0, \) and \( \mu_2 \) 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 \( \sqrt{2r \log p(1/n_1 + 1/n_2)} \). \( r = 0.1, 0.2, 0.3, 0.4. \)
- \( \Sigma_1 = \Sigma_2 = \Sigma = (\sigma_{ij}), \) 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 $\Omega^{1/2}$
Figure: Data transformation with $\Omega$
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: $k \gg n$; 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;
Figure: Structural networks (from DTI); taken from Moo Chung's website at UW-Madison.
### Table: P-values after adjusting for age and gender for the FASD data.

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