A powerful and adaptive association test for rare variants

Wei Pan¹, Junghi Kim¹, Yiwei Zhang¹, Xiaotong Shen², Peng Wei³,

¹ Division of Biostatistics, School of Public Health, ² School of Statistics, University of Minnesota, Minneapolis, MN 55455
³ Division of Biostatistics and Human Genetics Center, University of Texas School of Public Health, Houston, TX 77030

UNC, Sept 25, 2014
Outline

- Review: some existing methods.
- New methods: SPU and aSPU tests.
  Connections with some existing tests.
- Discussion.
Introduction

• Problem:
  - Given: a binary disease indicator $Y_i$ for subject $i$; a group of rare variants (RVs) (additively) coded as $X_i = (X_{i1}, ..., X_{ik})'$; $i = 1, ..., n >> k$.
  - Q: any association between $Y_i$ and $X_i$?
  - Approaches: global testing.

• 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

- Burden tests (Morgenthaler & Thilly 2007; Li & Leal 2008; Madsen & Browning 2009):
  Sum test (Chapman & Whittaker 2008): assuming
  $\beta_1 = \beta_2 = ... = \beta_k = \beta_c$; $H_0$: $\beta_c = 0$;

  $$\text{Logit}(Pr(Y_i = 1)) = \beta_{c,0} + \sum_{j=1}^{k} X_{ij} \beta_c.$$  

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

- Variance components tests:
  Sum of Squared Score (SSU) test (Pan 2009): assuming $\beta_1, ..., \beta_k \sim F(0, \tau^2)$, $H_0$: $\tau^2 = 0$,

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

  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:** 
  \[ T_{\text{UminP}} = \max_{j=1}^k U_j^2 / V_{jj}, \]
  close to \( T_{\text{maxU}} = \max_{j=1}^k |U_j| \)

- **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{minP}};

• 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
  $\Gamma = \{1, 2, ..., 8, \infty\}$.

- 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!
Simulations

- Using a multivariate Normal to simulated (possibly correlated) RVs (Wang and Eslston 2008);
- MAFs $\sim U(0.001, 0.01)$;
- $k = k_1 + 0, 8, 16, ..., 128; k_1 = 8$ causal ones;
- $Y_i$ from a joint logistic reg model with various values of $\beta_j$’s; case-control design, $n = 1000$.
- 1000 replicates for each set-up.
- Show power for one case: causal RVs with $\exp(\beta_j) \sim U(1, 2)$, $j = 1, ..., 8$. 
<table>
<thead>
<tr>
<th>Test</th>
<th># non-associated RVs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>UminP</td>
<td>.874</td>
</tr>
<tr>
<td>SPU(1)</td>
<td><strong>.939</strong></td>
</tr>
<tr>
<td>SPU(2)</td>
<td>.926</td>
</tr>
<tr>
<td>SPU(3)</td>
<td>.917</td>
</tr>
<tr>
<td>SPU(4)</td>
<td>.909</td>
</tr>
<tr>
<td>SPU(5)</td>
<td>.902</td>
</tr>
<tr>
<td>SPU(6)</td>
<td>.901</td>
</tr>
<tr>
<td>SPU(7)</td>
<td>.899</td>
</tr>
<tr>
<td>SPU(8)</td>
<td>.898</td>
</tr>
<tr>
<td>SPU(∞)</td>
<td>.877</td>
</tr>
<tr>
<td>aSPU</td>
<td>.923</td>
</tr>
<tr>
<td>aSum</td>
<td><strong>.948</strong></td>
</tr>
<tr>
<td>PWST</td>
<td>.823</td>
</tr>
<tr>
<td>EREC</td>
<td>.943</td>
</tr>
<tr>
<td>SKAT</td>
<td>.927</td>
</tr>
<tr>
<td>SKAT-O</td>
<td>.940</td>
</tr>
</tbody>
</table>
Example: GAW17 Data

- Mini-exome sequencing data with $n = 697$ unrelated subjects (Almasy et al 2011);

- After removing SNVs with MAFs $> 1\%$: $24,487$ RVs in $2476$ genes;

- A simulated binary phenotype; $200$ sets

- Two analyses:
  1) Applying gene-based testing on the set 1 of the binary phenotype, adjusting for age, gender and smoking status; Show p-values;
  2) Applying causal gene-based testing on each of the $200$ sets of the binary phenotype at the nominal level of $0.05$; Show empirical power

- Resampling: start with $B = 10^3$; if a p-value $< 5/B$, then
increase $B$ to $10^4$, ...., up to $10^6$

- Computing time for a genome-scan for one phenotype: in R; with 100 cores,
  it took about 0.2 hours to test 2476 genes based on $B = 10^3$
  permutations;
  0.05 hours to test 50 genes with $B = 10^4$;
  0.12 hours to test 5 with $B = 10^5$.

- Analysis I: give p-values;
<table>
<thead>
<tr>
<th>gene</th>
<th>#RVs</th>
<th>SPU(1)</th>
<th>SPU(2)</th>
<th>SPU(∞)</th>
<th>aSPU</th>
<th>SKATa</th>
<th>SKAT</th>
<th>SKAT-Oa</th>
<th>SKAT-O</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEGFA*</td>
<td>5, 1</td>
<td>0.00090</td>
<td>0.00060</td>
<td>0.02020</td>
<td>0.00120</td>
<td>0.00082</td>
<td>0.00150</td>
<td>0.00091</td>
<td>0.00110</td>
</tr>
<tr>
<td>PTK2*</td>
<td>9, 2</td>
<td>0.00080</td>
<td>0.00300</td>
<td>0.00240</td>
<td>0.00230</td>
<td>0.00178</td>
<td>0.00350</td>
<td>0.00105</td>
<td>0.00120</td>
</tr>
<tr>
<td>SOS2*</td>
<td>7, 2</td>
<td>0.18600</td>
<td>0.03500</td>
<td>0.01400</td>
<td>0.03200</td>
<td>0.03838</td>
<td>0.04496</td>
<td>0.06176</td>
<td>0.05195</td>
</tr>
<tr>
<td>RARB*</td>
<td>9, 2</td>
<td>0.07600</td>
<td>0.04900</td>
<td>0.04900</td>
<td>0.07100</td>
<td>0.06313</td>
<td>0.05794</td>
<td>0.09546</td>
<td>0.09690</td>
</tr>
<tr>
<td>SIRT1*</td>
<td>23, 9</td>
<td>0.35500</td>
<td>0.09500</td>
<td>0.13600</td>
<td>0.07200</td>
<td>0.08091</td>
<td>0.10589</td>
<td>0.14543</td>
<td>0.17383</td>
</tr>
<tr>
<td>GAB4</td>
<td>5, 0</td>
<td>0.05240</td>
<td>0.00940</td>
<td>0.00160</td>
<td>0.00370</td>
<td>0.01358</td>
<td>0.01230</td>
<td>0.02204</td>
<td>0.02040</td>
</tr>
</tbody>
</table>
• Analysis II: empirical power based on the significance level of 0.05 and the 200 replicates
• Show for causal genes; many with power < 0.05
<table>
<thead>
<tr>
<th>Chr</th>
<th>Gene</th>
<th>#RVs</th>
<th>SPU(1)</th>
<th>SPU(2)</th>
<th>SPU(3)</th>
<th>SPU(4)</th>
<th>SPU(5)</th>
<th>SPU(∞)</th>
<th>aSPU</th>
<th>SKAT</th>
<th>SKAT-O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PIK3C2B</td>
<td>60, 23</td>
<td>0.565</td>
<td>0.445</td>
<td><strong>0.650</strong></td>
<td>0.400</td>
<td>0.395</td>
<td>0.340</td>
<td><strong>0.600</strong></td>
<td>0.435</td>
<td>0.560</td>
</tr>
<tr>
<td>6</td>
<td>VNN1</td>
<td>6, 1</td>
<td>0.185</td>
<td>0.230</td>
<td>0.315</td>
<td>0.235</td>
<td><strong>0.380</strong></td>
<td>0.140</td>
<td><strong>0.270</strong></td>
<td>0.215</td>
<td>0.185</td>
</tr>
<tr>
<td>3</td>
<td>BCHE</td>
<td>28, 13</td>
<td>0.110</td>
<td>0.190</td>
<td><strong>0.215</strong></td>
<td>0.185</td>
<td>0.175</td>
<td>0.160</td>
<td><strong>0.210</strong></td>
<td>0.195</td>
<td>0.175</td>
</tr>
<tr>
<td>8</td>
<td>LPL</td>
<td>15, 2</td>
<td>0.090</td>
<td>0.130</td>
<td><strong>0.135</strong></td>
<td>0.110</td>
<td>0.115</td>
<td>0.110</td>
<td><strong>0.135</strong></td>
<td>0.125</td>
<td>0.125</td>
</tr>
<tr>
<td>10</td>
<td>SIRT1</td>
<td>23, 9</td>
<td>0.095</td>
<td>0.105</td>
<td><strong>0.110</strong></td>
<td>0.065</td>
<td>0.070</td>
<td>0.015</td>
<td><strong>0.105</strong></td>
<td>0.090</td>
<td>0.100</td>
</tr>
<tr>
<td>14</td>
<td>SOS2</td>
<td>7, 2</td>
<td>0.100</td>
<td>0.270</td>
<td><strong>0.285</strong></td>
<td>0.265</td>
<td>0.275</td>
<td>0.245</td>
<td>0.220</td>
<td><strong>0.255</strong></td>
<td>0.200</td>
</tr>
<tr>
<td>19</td>
<td>RRAS</td>
<td>5, 2</td>
<td><strong>0.235</strong></td>
<td>0.140</td>
<td>0.155</td>
<td>0.145</td>
<td>0.155</td>
<td>0.100</td>
<td>0.180</td>
<td>0.135</td>
<td><strong>0.200</strong></td>
</tr>
<tr>
<td>8</td>
<td>PLAT</td>
<td>25, 8</td>
<td><strong>0.225</strong></td>
<td>0.135</td>
<td>0.145</td>
<td>0.110</td>
<td>0.105</td>
<td>0.070</td>
<td>0.155</td>
<td>0.130</td>
<td><strong>0.195</strong></td>
</tr>
<tr>
<td>9</td>
<td>VLDLR</td>
<td>23, 8</td>
<td>0.080</td>
<td>0.120</td>
<td><strong>0.125</strong></td>
<td>0.110</td>
<td>0.120</td>
<td>0.075</td>
<td>0.090</td>
<td><strong>0.125</strong></td>
<td>0.090</td>
</tr>
<tr>
<td>17</td>
<td>SREBF1</td>
<td>21, 10</td>
<td>0.050</td>
<td>0.085</td>
<td>0.090</td>
<td><strong>0.105</strong></td>
<td>0.100</td>
<td>0.100</td>
<td>0.085</td>
<td><strong>0.090</strong></td>
<td>0.070</td>
</tr>
<tr>
<td>4</td>
<td>KDR</td>
<td>14, 8</td>
<td><strong>0.365</strong></td>
<td>0.350</td>
<td>0.160</td>
<td>0.105</td>
<td>0.105</td>
<td>0.020</td>
<td>0.280</td>
<td>0.365</td>
<td><strong>0.390</strong></td>
</tr>
<tr>
<td>13</td>
<td>FLT1</td>
<td>25, 8</td>
<td>0.125</td>
<td>0.160</td>
<td><strong>0.170</strong></td>
<td>0.150</td>
<td>0.160</td>
<td>0.065</td>
<td>0.125</td>
<td>0.150</td>
<td><strong>0.165</strong></td>
</tr>
<tr>
<td>14</td>
<td>HSP90AA1</td>
<td>20,3</td>
<td>0.050</td>
<td>0.275</td>
<td>0.180</td>
<td>0.195</td>
<td>0.170</td>
<td>0.030</td>
<td>0.155</td>
<td><strong>0.335</strong></td>
<td>0.250</td>
</tr>
</tbody>
</table>
Discussion

• Conclusion: aSPU test is promising (and general/flexible)

• Current work:
  applied to real data;
  develop an R package;
  analytical null distribution?

• 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;
Figure 1: Structural networks (from DTI); taken from Moo Chung’s website at UW-Madison.
Table 1: 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 2: 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.

You can download our papers from http://www.biostat.umn.edu/rrs.php

Thank you!