Testing for Association with Multiple Traits in Generalized Estimating Equations, With Application to Neuroimaging Data

Wei Pan

[1]Division of Biostatistics, School of Public Health

University of Minnesota

IG Working Group Meeting, Oct 3, 2014

Ref: Zhang Y, Xu Z, Shen X, Pan W, for ADNI (2014, NeuroImage).

Outline

- Introduction: what is the problem ...
- New method: SPU/aSPU tests
- Connection with some existing methods
- Simulation results
- Application to an ADNI dataset
- Future work

Introduction

- Problem: association testing with a multivariate (quantitative) trait and a single SNP, possibly with covariates.
- Why? to increase power; pleitropy; ...
- Existing methods: a review (Yang and Wang 2013);
  - Combining single trait analyses, e.g. by UminP (Yang et al 2010);
  - CCA/MANOVA for QTs (Ferreira and Purcell 2009);
  - GLS/LME/GLMM (Li et al 2011; ...);
  - PCA (for QTs?) (Lan et al 2003; Aschard et al 2013);
  - PCH (Klei et al 2008; Lin et al 2012);
  - A simple Average/Sum of the (Q) traits (Shen et al 2012);
  - GEE (Liang and Zeger 1986; Liu et al., 2009; Chen et al., 2011; Lange et al., 2003); may have inflated Type I errors

(Yang and Wang 2013), but mainly for the Wald test only (Pan 2001).

  - More recent ones, MDMA, KMR, TATES, MultiPhen, ...

- Why this study?
  target: possibly a medium # of traits as ROIs in ADNI data;
  adaptive?
  relationships among the methods?

New Method

- Data: for each subject $i = 1, ..., n$, k traits $Y_i = (y_{i1}$, y_{i2}, ..., $y_{ik})'$;

  genotype score $x_i = 0$, 1 or 2;

  covariates z_i;

- Marginal GLM: $\beta = (\beta_1, ..., \beta_k)'$

$$g(E(Y_i)) = x_i\beta + z_i'\kappa,$$

  link function $g() = I()$ for QTs.

- GEE: $\mu_i = E(Y_i)$,

$$U = U(\beta, \kappa) = \sum_{i=1}^{n} U_{.i} = \sum_{i} \bigtriangledown\mu_i{}' V_i^{-1}(Y_i - \mu_i) = 0,$$

$$\bigtriangledown\mu_i = \partial\mu_i/\partial\theta^{'} = \partial g^{-1}(\mu_i)/\partial\theta^{'},\ V_i = \phi A_i^{1/2} R_w(\alpha) A_i^{1/2},$$

- For simplicity of notation, assume no covariates and

$U = (U_1, ..., U_k)'$, $\Sigma = Cov(U)$
Key: $U \sim N(0, \Sigma)$ under H_0.

- Existing GEE tests:
  Wald: $T = \hat{\beta}' Cov(\hat{\beta})^{-1} \hat{\beta} \sim \chi^2_1$ under H_0;
  Score: $T = U' \Sigma^{-1} U \sim \chi^2_1$ under H_0;
  UminP: $T = \max_j U_j^2 / \Sigma_{jj}$ with $R_w = I$;
- Problem: not adaptive?
- New method: for $j = 1, 2, ..., \infty$,

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

- Special cases:
  SPU(1) = Sum;
  SPU(2) = SSU; (Pan 2009; Yang and Wang 2013)
  $SPU(\infty) = \max_j U_j^2 \approx$ UminP;

- Key idea: increasing γ puts higher weights on more significant components!
- Which γ to use?

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

  where $P_{SPU(\gamma)}$ is the p-value of $SPU(\gamma)$.
- Use simulations to calculate p-values for SPU/aSPU tests: simulate $U^{(b)} \sim N(0, \Sigma)$, calculate $SPU(\gamma)^{(b)}$ for $b = 1, 2, ..., B$, then

  $$P_{SPU(\gamma)} = \sum_{b=1}^{B} I(|SPU(\gamma)| \leq |SPU(\gamma)^{(b)}|)/B.$$

  Similarly for aSPU; just need a single loop of B simulations.
- Remarks:

  1. single trait vs multiple SNPs (RVs): done;

2. motivated by and related to polygenic testing: almost done;
3. multiple traits vs multiple SNPs: Yiwei’s thesis, in prep;

Connections

- All analytical
- Average=Sum = SPU(1) with $R_w = I$;
- TATES $\approx$ UminP $\approx SPU(\infty)$ with $R_w = I$;
- CCA = MANOVA =GEE Score test for any R_w;
  Score($R_{w,1}$) = Score($R_{w,2}$) for any $R_{w,1} \neq R_{w,2}$;
- MDMR (Wessel & Schork 2006; Zapala & Schork 2012; ...);
  MDMR(L2) = SPU(2) with $R_w = I$;
  MDMR: an extension of MANOVA to any distance $d_{ij} = d(Y_i, Y_j)$; but the summary measure differs from MANOVA.
  Why important? limitation of SPU(2) ...
- KMR (Maity et al 2012; Schifano et al 2012; Wang et al 2013);
  KMR = SPU(2) if $R_w = Corr(Y_i)$;

Why important? limitation of SPU(2) ...

- MultiPhen (O'Reilly et al 2012): POM $x_i \sim Y_i$; MultiPhen applies a likelihood ratio test, asymptotically equivalent to score test. The score vector for the POM is

$$U_{POM} = \frac{-n_1 - n_2}{n} \sum_{i:x_i=0} Y_i + \frac{n_0 - n_2}{n} \sum_{i:x_i=1} Y_i + \frac{n_0 + n_1}{n} \sum_{i:x_i=2} Y_i, \tag{1}$$

where $n_j = \sum_{i=1}^n I(x_i = j)$ for $j = 0$, 1 and 2. In contrast, the Score vector for the GEE with $R_w = I$ is

$$U_{GEE} = \frac{-n_1 - 2n_2}{n} \sum_{i:x_i=0} Y_i + \frac{n_0 - n_2}{n} \sum_{i:x_i=1} Y_i + \frac{2n_0 + n_1}{n} \sum_{i:x_i=2} Y_i. \tag{2}$$

$\Longrightarrow$ MultiPhen $\approx$ GEE Score = CCA=MANOVA!

- Generalized Kendall's tau (Zhang et al 2010): Generalized Kendall's tau = GEE Score test;

Simulation Results

- $n = 1000$; $k = 5$, 10, ..., 40;
- A causal SNP is associated to $k_1 = 5$ traits out of k traits;
- $\beta_j \sim U(0.2, 0.3)$ or $U(0.8, 1)$ or $\beta_j = 0$;
- The k traits have either CS(r) or AR1(r) with $r = 0.3$ or 0.5;
- Test association b/w the k traits and an SNP in LD with the causal one;
- Replicated 1000 times; $B = 1000$ for simulated-based methods;
- Empirical Type I error rates were well controlled, except for the GEE Wald test; only show (empirical) power:

Empirical power when the multiple traits were correlated with a CS structure relation coefficient r; the first five traits were associated with a causal SNP -ORs $\beta_j \sim U(0.8, 1)$, while all others had $\beta_j = 0$. An independence working on structure was used in GEE.

								GEE									
					MDMR					SPU(γ)							
SNP	#trait	Average	MultiPhen	TATES	L_1	L_2	MANOVA	Score	UminP	$\gamma = 1$	2	3	4	5	6	∞	aSPU
2	5	0.888	0.683	0.823	0.878	0.883	0.683	0.682	0.815	0.889	0.880	0.868	0.862	0.851	0.843	0.812	0.865
	10	0.567	0.685	0.727	0.786	0.826	0.686	0.684	0.708	0.567	0.830	0.777	0.819	0.796	0.807	0.772	0.795
	20	0.218	0.613	0.665	0.616	0.729	0.615	0.611	0.657	0.223	0.724	0.667	0.787	0.756	0.792	0.762	0.757
	30	0.116	0.528	0.607	0.435	0.574	0.531	0.528	0.591	0.117	0.577	0.541	0.725	0.695	0.742	0.738	0.703
	40	0.084	0.424	0.536	0.262	0.442	0.435	0.424	0.534	0.084	0.432	0.445	0.644	0.609	0.680	0.678	0.639
3	5	0.334	0.178	0.292	0.328	0.330	0.178	0.177	0.281	0.331	0.327	0.321	0.315	0.312	0.305	0.289	0.320
	10	0.184	0.167	0.203	0.240	0.269	0.167	0.167	0.197	0.182	0.273	0.260	0.282	0.267	0.273	0.244	0.249
	20	0.092	0.138	0.179	0.149	0.189	0.141	0.137	0.173	0.090	0.188	0.191	0.242	0.246	0.257	0.242	0.229
	30	0.074	0.121	0.143	0.107	0.120	0.127	0.119	0.146	0.079	0.128	0.141	0.185	0.184	0.203	0.204	0.181
	40	0.058	0.105	0.132	0.088	0.113	0.109	0.104	0.134	0.062	0.112	0.118	0.161	0.168	0.179	0.188	0.168
2	5	0.829	0.602	0.784	0.821	0.822	0.604	0.601	0.763	0.832	0.821	0.811	0.806	0.800	0.793	0.769	0.806
	10	0.424	0.725	0.694	0.629	0.729	0.728	0.725	0.685	0.430	0.734	0.714	0.766	0.750	0.765	0.737	0.723
	20	0.163	0.665	0.624	0.344	0.524	0.666	0.662	0.634	0.161	0.534	0.567	0.697	0.695	0.725	0.722	0.694
	30	0.093	0.570	0.549	0.186	0.318	0.577	0.570	0.570	0.093	0.319	0.440	0.593	0.609	0.666	0.707	0.653
	40	0.067	0.484	0.487	0.119	0.203	0.496	0.483	0.508	0.072	0.202	0.333	0.518	0.544	0.612	0.654	0.613
3	5	0.291	0.149	0.270	0.288	0.290	0.150	0.148	0.259	0.290	0.294	0.293	0.285	0.284	0.279	0.263	0.287
	10	0.129	0.181	0.209	0.171	0.207	0.182	0.180	0.201	0.126	0.203	0.223	0.245	0.249	0.255	0.245	0.223
	20	0.077	0.138	0.160	0.098	0.130	0.139	0.136	0.168	0.075	0.131	0.158	0.196	0.212	0.220	0.228	0.205
	30	0.067	0.117	0.129	0.077	0.092	0.121	0.117	0.139	0.065	0.091	0.118	0.144	0.156	0.166	0.195	0.181
	40	0.055	0.110	0.113	0.071	0.078	0.116	0.109	0.129	0.054	0.079	0.105	0.134	0.148	0.163	0.190	0.166

correlation structure was used in GEE.

								GEE									
					MDMR					SPU(γ)							
NP	#traits	Average	MultiPhen	TATES	L_1	L_2	MANOVA	Score	UminP	$\gamma = 1$	2	3	4	5	6	∞	aSPU
1	5	0.664	0.468	0.551	0.653	0.652	0.469	0.468	0.531	0.660	0.658	0.636	0.609	0.597	0.569	0.533	0.632
	10	0.263	0.574	0.452	0.441	0.506	0.576	0.573	0.437	0.267	0.501	0.460	0.493	0.472	0.481	0.444	0.456
	20	0.114	0.535	0.335	0.202	0.245	0.536	0.535	0.330	0.114	0.249	0.261	0.330	0.321	0.348	0.345	0.305
	30	0.084	0.458	0.283	0.126	0.158	0.462	0.456	0.282	0.085	0.162	0.188	0.254	0.262	0.288	0.293	0.257
	40	0.058	0.412	0.252	0.089	0.103	0.421	0.409	0.250	0.058	0.100	0.128	0.180	0.192	0.236	0.263	0.211
2	5	0.226	0.110	0.165	0.209	0.213	0.110	0.108	0.160	0.221	0.214	0.206	0.188	0.188	0.181	0.166	0.211
	10	0.087	0.142	0.120	0.098	0.117	0.143	0.142	0.115	0.088	0.117	0.116	0.129	0.129	0.128	0.122	0.118
	20	0.064	0.132	0.085	0.074	0.083	0.135	0.131	0.091	0.064	0.086	0.089	0.099	0.097	0.098	0.095	0.089
	30	0.058	0.130	0.098	0.063	0.069	0.131	0.129	0.097	0.060	0.071	0.076	0.086	0.092	0.098	0.102	0.087
	40	0.050	0.091	0.067	0.057	0.056	0.098	0.091	0.066	0.049	0.055	0.057	0.064	0.066	0.070	0.071	0.063

Empirical power when the multiple traits were correlated with an AR1 structure
relation coefficient r; the non-zero $\beta_j \sim U(0.2, 0.3)$. An independence working
on structure was used in GEE.

								GEE									
					MDMR					SPU(γ)							
SNP	#traits	Average	MultiPhen	TATES	L_1	L_2	MANOVA	Score	UminP	$\gamma = 1$	2	3	4	5	6	∞	aSPU
1	5	0.661	0.458	0.554	0.629	0.634	0.459	0.458	0.522	0.651	0.630	0.624	0.594	0.582	0.564	0.525	0.624
	10	0.390	0.371	0.426	0.496	0.527	0.373	0.388	0.447	0.388	0.555	0.534	0.533	0.513	0.513	0.471	0.516
	20	0.217	0.262	0.332	0.362	0.365	0.263	0.286	0.334	0.214	0.414	0.390	0.427	0.397	0.402	0.343	0.400
2	5	0.223	0.113	0.165	0.202	0.201	0.113	0.113	0.153	0.220	0.206	0.193	0.182	0.178	0.173	0.162	0.208
	10	0.124	0.107	0.122	0.131	0.129	0.107	0.100	0.112	0.124	0.150	0.137	0.139	0.127	0.129	0.114	0.139
	20	0.084	0.080	0.105	0.121	0.118	0.081	0.069	0.104	0.090	0.113	0.106	0.122	0.109	0.116	0.103	0.111
1	5	0.780	0.547	0.571	0.698	0.706	0.571	0.546	0.551	0.774	0.706	0.709	0.647	0.637	0.602	0.551	0.737
	10	0.487	0.442	0.469	0.568	0.596	0.444	0.442	0.443	0.482	0.592	0.569	0.546	0.530	0.511	0.454	0.572
	20	0.274	0.309	0.366	0.448	0.490	0.312	0.307	0.349	0.277	0.478	0.456	0.467	0.438	0.434	0.368	0.459
2	5	0.245	0.129	0.154	0.177	0.179	0.129	0.127	0.146	0.244	0.180	0.185	0.165	0.164	0.153	0.149	0.190
	10	0.147	0.120	0.129	0.157	0.156	0.121	0.119	0.122	0.146	0.161	0.146	0.143	0.139	0.136	0.122	0.156
	20	0.077	0.085	0.093	0.121	0.126	0.087	0.085	0.087	0.078	0.131	0.113	0.115	0.098	0.103	0.091	0.113

- GEE-Score (or MANOVA) and GEE-aSPU are complementary; A strange behavior of GEE-Score or MANOVA: adding non-associated traits may **increase** power!
- GEE-aSPU.Sco: combines GEE-aSPU and GEE-Score tests. always close to the winner!
- Remark: compared to $R_w = I$, using a non-diag R_w in GEE may or may **not** improve power of a test; it also depends on the test being used.
  A correct model/assumption may not help, depending on how to use it!

Application

- To an ADNI dataset;
- $n = 680$ non-Hispanic Caucasians: 192 HCs, 327 MCIs, 161 ADs;
- $k = 26$ cortical thickness in 26 ROIs; FreeSurfers;
- Covariates: gender, age, education, brain volume;
- followed Shen et al (2012); downloaded from the ADNI website;
- Started with $B = 10^4$, then increase B to 10^5, 10^6 up to 10^7 if p-value $< 5/B$.

Table 4: P-values of testing on a pooled set of 26 univariate traits.

				GEE				
SNPs	Average	MultiPhen	TATES	UminP	Score	SPU(1)	SPU(2)	aSPU
rs7526034	1.40e-04	5.82e-04	1.72e-05	2.10e-05	5.86e-04	7.30e-05	7.00e-06	7.00e-06
rs429358	1.42e-04	1.68e-05	1.23e-04	1.50e-04	2.32e-05	1.10e-04	7.00e-05	1.60e-04

Future Work

- Compare with PCH, Projection regression of Liu et al (2012). For large k, dimension reduction may be necessary and beneficial, but interpretation/motivation?

- An example: vGWAS-like; multiple voxels replace multiple ROIs.
  Data-adaptive parcellation (to form ROIs);
  Hongtu's Multi-scale modeling? Spatially varying coefficient modeling?
  Regularized matrix/tensor regression (Zhou et al 2013; Zhou and Li 2014)?
  ...

- More extreme: to high-dim data (with even larger $k >> n$: compare with some new tests (Chen et al 2014, 2010; Cai et al 2014; ...); theory.

Acknowledgment: This research was supported by NIH.

Thank you!