Some old and new tests in genetic association analysis: an introduction

Wei Pan[1]

[1]Division of Biostatistics, School of Public Health

University of Minnesota

July 22, 2015

Outline

- Introduction: problem
- New method: SSU test
  Some theory, connections with others, numerical results...
- Discussion
- Main refs:
  Pan (2009, *Genet Epi*), Han and Pan (2010, *Genet Epi*), Pan (2011, *Genet Epi*), ...

Introduction

- Single Nucleotide Polymorphisms (SNP)
  DNA seq 1 – AAGC**C**TA
  DNA seq 2 – AAGC**T**TA
  two alleles, C and T; 3 genotypes: CC, TT, CT;
  SNP: a minor allele freq (MAF) $\geq 5\%$ (or 1%).
  GWAS: Genome-wide SNPs are measured as markers for each subject;
- Problem: Genome-wide *association* studies (GWAS)
  Goal: to detect assoc b/w a phenotype (e.g. disease status) and genome-wide SNPs;
  Ultimate goal: to detect *causal* genetic variants.
- The NIH Catalog of Published GWAS includes thousands of SNPs that are associated with some phenotypes, such as prostate cancer, diabetes, bipolar disorder...

- Most common study design: case-control;
  n in hundreds, then thousands, then ?
  hundreds of thousands SNPs (e.g. 500K Affy arrays);
  OR : < 1.5, typically, even only 1.1-1.2.

- Data:

Obs	Y	SNP1	...	SNP2	...	(SNP0)	...	SNPk
1	1	CT	...	AG	...	CG	...	AC
2	1	TT	...	AG	...	GG	...	AA
3	1	CT	...	AA	...	CG	...	CC
.....								
1001	0	CT	...	AG	...	CC	...	AC
1002	0	TT	...	GG	...	CC	...	AC
1003	0	CC	...	GG	...	CC	...	CC
.....								

- A binary response: $Y = 0$ or 1;
  each SNP j has up to 3 possible values; coded as $X_j = 0$, 1 or 2, though other codings are possible.
- The causal SNP0 may not be observed.
- Linkage disequilibrium (LD): SNP0 and its nearby SNPs are

correlated (and form an LD block).
$\Longrightarrow$ If SNP0 is causal, then its nearby SNPs are associated with Y!

- Statistical question: any SNP associated with Y?
  univariate or multivariate?
- Here we consider $k > 1$ SNPs inside a **given** LD block or sliding window.
  Selection of LD block or window size: relevant, not trivial.
- GxG and GxE can be similarly formulated.

Existing methods

- Single-locus (or SNP-by-SNP or univariate) analysis:
  - Model: $Y \sim SNP_j$

$$\text{Logit } \Pr(Y_i = 1) = \beta_{M,0j} + X_{ij}\beta_{M,j}, \qquad (1)$$

  - $H_{0,j}$: $\beta_{M,j} = 0$ for each $j = 1, ..., k$
    $\Longrightarrow p_j$.
  - Combining: $UminP = \min(p_1, p_2, ..., p_k)$ or ...
    Need to do multiple test adjustment!
    Time-consuming with permutation, or conservative with Bonferroni method.
    Analytical: sometimes; numerical integration.
  - Model (1): as a 2×3 table; Cochran-Armitage trend test.

- Multivariate (or global or joint) analysis:
  - Model: $Y \sim SNP_1 + ... + SNP_k$

$$\text{Logit } \Pr(Y_i = 1) = \beta_0 + \sum_{j=1}^{k} X_{ij}\beta_j, \tag{2}$$

  - H_0: $\beta_1 = ... = \beta_k = 0$
  - Use the score, Wald or LR test:
    $T_W = \hat{\beta}' V^{-1} \hat{\beta}$, $T_S = U' V_U^{-1} U \sim \chi^2_k$ under H_0;
    $V = Cov(\hat{\beta})$, $V_U = Cov(U)$;
    Possibly large $DF = k$.
  - Hotelling's T^2 test: closely related to the score test.

- Sum test
  - *Working* assumption: $\beta_1 = ... = \beta_k \equiv \beta_c$.
    in general, *incorrect*!
  - Model:

$$\text{Logit } \Pr(Y_i = 1) = \beta_{0,c} + \sum_{j=1}^{k} X_{ij}\beta_c = \beta_{0,c} + X_{i,c}\beta_c, \qquad (3)$$

  - $H_{0,c}$: $\beta_c = 0$
  - Apply the score, Wald or LR test:
    $T_W = \hat{\beta}_c^2 / V_c \sim \chi_1^2$ under $H_{0,c}$.
  - Feature: DF=1; no multiple test!
  - Correct test size:
    $H_0 \Longrightarrow H_{0,c}$!
  - Power: simulation results; $n = 500 + 500$

- Chapman and Whittaker (2008, *Genetic Epi*):
  The UminP and a test by Goeman et al (2006, JRSS-B) work best.

- Goeman's test:
  - Set-up: "large k, small n" as for microarray data;
  - Main idea:
    Prior for $\beta = (\beta_1, ..., \beta_k)'$: $E(\beta) = 0$, $Cov(\beta) = \tau^2 I$.
    Now test H_{0,τ^2}: $\tau^2 = 0$.
  - For logistic regression:
    $T_{Go} = \frac{1}{2}(U'U - \text{Trace}(I_F))$, where $U = X'(Y - \bar{Y})$,
    and $I_f = Cov(U) = \bar{Y}(1 - \bar{Y})(X - \bar{X})'(X - \bar{X})$.
  - Null distribution unknown; use simulation or permutation.

- Why does Goeman's test work here ("large n, small k")?

Corr	OR	Sum	LRT	T^2	UminP	Goeman
Rand	1.0	.044	.048	.051	.050	.048
	1.2	.134	.078	.079	.087	.121
	1.4	.320	.148	.153	.200	.290
	1.6	.546	.243	.246	.360	.523
	1.8	.753	.383	.391	.537	.729
	2.0	.863	.530	.540	.688	.848

HapMap CEU data for gene IL21R; #SNP=27:

n	OR	Sum	LRT	T^2	UminP	Goeman
200	1.0	.046	.098	.063	.057	.052
200	1.2	.078	.107	.078	.087	.087
200	1.4	.204	.200	.148	.256	.265
200	1.6	.351	.344	.275	.500	.474
500	1.0	.050	.054	.031	.055	.047
500	1.2	.165	.142	.107	.183	.204
500	1.4	.432	.408	.333	.652	.600
500	1.6	.607	.717	.667	.908	.831

New method

- Recall $LRT \approx Wald's \approx Score = U'V^{-1}U$,
  $U = \sum_{i=1}^{m} X_i(Y_i - \bar{Y})$,
  $V = Cov(U) = I_F = \bar{Y}(1-\bar{Y})(X-\bar{X})'(X-\bar{X})$.

- New tests:

$$SSU = U'U, \qquad SSUw = U'diag(V)^{-1}U.$$

- Null distributions for $Q = U'W^{-1}U$:
  1) $W = I$ and $W = \mathrm{Diag}(V_M)$ in the above;
  2) $Q \sim \sum_{j=1}^{k} c_j \chi_1^2$, where c_j's are the eigen values of $V_M W^{-1}$;
  3) Zhang (2005, JASA): approximate by $a\chi_d^2 + b$ with

$$a = \frac{\sum_{j=1}^{k} c_j^3}{\sum_{j=1}^{k} c_j^2}, \quad b = \sum_{j=1}^{k} c_j - \frac{\left(\sum_{j=1}^{k} c_j^2\right)^2}{\sum_{j=1}^{k} c_j^3}, \quad d = \frac{\left(\sum_{j=1}^{k} c_j^2\right)^3}{\left(\sum_{j=1}^{k} c_j^3\right)^2}.$$

4) $Pr(SSU > s|H_0) \approx Pr\left(\chi^2_d > (s-b)/a\right)$.

- Wald's versions of SSU and SSUw ...

Simulation with corr randomly b/w 0.2–0.7; #SNP=10; $n = 500 + 500$:

OR	Sum	LRT	UminP	Goeman	SSUw	SSU
1.0	.044	.048	.050	.048	.044	.046
1.2	.134	.078	.087	.121	.116	.114
1.4	.320	.148	.200	.290	.281	.284
1.6	.546	.243	.360	.523	.505	.500
1.8	.753	.383	.537	.729	.718	.721
2.0	.863	.530	.688	.848	.837	.836

HapMap CEU data for gene IL21R; #SNP=27:

OR	Sum	LRT	UminP	Goeman	SSUw	SSU
				($n = 200$)		
1.0	.046	.098	.057	.052	.047	.047
1.2	.078	.107	.087	.087	.079	.084
1.4	.204	.200	.256	.265	.265	.261
1.6	.351	.344	.500	.474	.457	.464
				($n = 500$)		
1.0	.050	.054	.055	.047	.044	.042
1.2	.165	.142	.183	.204	.208	.202
1.4	.432	.408	.652	.600	.589	.594
1.6	.607	.717	.908	.831	.836	.828

- $SSU \approx SSU_w$ if $diag(V_M) \approx v\mathbf{1}$.
- Connection b/w SSU and Goeman's test:

$$\begin{aligned} T_{Go} &= \frac{1}{2}(Y-\bar{Y})'XX'(Y-\bar{Y}) - \\ &\quad \frac{1}{2}\bar{Y}(1-\bar{Y})\text{Trace}((X-\bar{X})'(X-\bar{X})), \end{aligned}$$

  Conditional on Y the second term is fixed (i.e. non-random) and can be dropped:

$$T_{Go} = \frac{1}{2}U'U + c_0 = \frac{1}{2}U_M'U_M + c_0 \propto SSU.$$

- Why do SSU/SSUw work?
  How could they beat "optimal" score, Wald and LR tests???
- Cox and Hinkley, *Theoretical Statistics*, 1974:
  - Optimality of the score, Wald and LR tests:
    locally most powerful, but only for ...;

o/w, no uniformly most power (unbiased) (UMPU) test!

- If we knew β, then
  $T_{MP} = \beta' U$, **but** ...
- Try $\max_b b'U$ s.t. $Var(b'U) = b' I_F b = 1$?

- We estimate T_{MP} by
  $T_{EMP} = \hat{\beta}'_M U$.
- $T_{EMP} \approx SSUw = U' \mathrm{Diag}(I_F)^{-1} U$ because

$$\hat{\beta}_M = I_{M,d}^{-1} U_M + O_{\boldsymbol{p}}(m^{-1}), \qquad U = U_M. \tag{4}$$

- How about estimating β by $\hat{\beta}$?
  $T_{EMP,J} = \hat{\beta}' U \approx U' I_F^{-1} U$, which is ...

- Connection b/w SSU and kernel machine regression(KMR):
  - KMR (Kwee et al 2008, *AJHG*; Wu et al 2010, *AJHG*): use a semi-parametric regression model

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + h(X_{i1}, ..., X_{ik}), \qquad (5)$$

    $h(.)$ is an unknown function to be estimated. The form of $h(.)$ is determined by a user-specified positive and semi-definite (psd) kernel function $K(.,.)$: by the representer theorem (Kimeldorf and Wahba 1971), $h_i = h(X_i) = \sum_{j=1}^{n} \gamma_j K(X_i, X_j)$ with some γ_1,...,γ_n.
  - To test H_0: $h = (h_1(X_1), ..., h_n(X_n))' = 0$.
    let $K = (K(X_i, X_j))$, $\gamma = (\gamma_1, ..., \gamma_n)'$, then $h = K\gamma$.
    Assume h as subject-specific random effects:
    $E(h) = 0$, $Cov(h) = \tau K$.
    $H_0 = H_0'$: $\tau = 0$.

Score test for H_0':

$$Q = (Y - \bar{Y}1)'K(Y - \bar{Y}1) = SSU$$

for H_0'': $b = 0$ in

$$\text{Logit } \Pr(Y = 1) = b_0 + Zb$$

with $K = ZZ'$.

- Genomic distance based regression (GDBR) (Wessel and Schork 2006, *AJHG*), a nonparametric MANOVA:

$$\begin{aligned} F &= \frac{tr(\hat{Y}'\hat{Y})}{tr(R'R)} = \frac{tr(\hat{Y}\hat{Y}')}{tr(RR')} = \frac{tr(HYY'H)}{tr((I-H)YY'(I-H))} \\ &= \frac{tr(HGH)}{tr((I-H)G(I-H))} \propto SSU \end{aligned}$$

  for H_0'': $b = 0$ in

$$\text{Logit Pr}(Y = 1) = b_0 + Zb$$

  with $G = ZZ'$.

- A side-product (Pan 2011, *Genet Epi*): KMR=GDBR=SSU if $K = G = ZZ'$.

Application to Rare Variants

- RV: X is sparse with most ($> 95\%$ or 99%) elements as 0's.
- Some dim reduction is necessary, e.g. variable selection;
  Most popular: pooling/collapsing SNP/SNV together, as done in the Sum test.
- Problems:
  Pooled assoc tests: bad with 1) opposite assoc directions; 2) large # neutral RVs.
- How about the SSU/SSUw and related tests?
- Some simulation results:

8 causal RVs with a common $OR = 2$; and a number of non-functional RVs. no LD.

Test/#nfRVs	0	4	8	16	32	64
UminP	.441	.336	.296	.222	.175	.117
Score	.746	.632	.595	.471	.332	.245
SSU	.756	.702	.694	.626	**.499**	**.423**
SSUw	.743	.638	.593	.477	.339	.268
Sum	**.951**	**.875**	**.808**	**.673**	.484	.313
KMR(Linear)	.762	.711	.699	.631	**.509**	**.438**
KMR(Quad)	.755	.707	.699	.629	.501	**.410**
CMC	**.938**	**.853**	.777	.616	.399	.211
wSum	**.940**	.846	**.782**	.618	.424	.267
aSum-P	.933	**.858**	**.780**	**.669**	**.499**	.313
C-alpha-P	.771	.712	.688	.627	.484	.378
Step-up	.859	.801	.769	**.679**	**.521**	.335

$OR = (3, 3, 2, 2, 2, 1/2, 1/2, 1/2)$; no LD.

Test/#nfRVs	0	4	8	16	32
UminP	.607	.532	.481	.417	.346
Score	.869	.772	.721	.632	.483
SSU	**.895**	**.835**	**.815**	**.774**	**.696**
SSUw	.867	.773	.732	.633	.501
Sum	.682	.566	.465	.365	.258
KMR(Linear)	**.897**	**.842**	**.824**	**.783**	**.707**
KMR(Quad)	.893	.835	.815	.781	.698
CMC	.661	.544	.456	.336	.204
wSum	.659	.548	.459	.335	.228
aSum-P	.854	.745	.684	.574	.430
C-alpha-P	**.906**	**.844**	**.823**	**.775**	**.674**
Step-up	.839	.767	.724	.640	.527

$OR = (3, 1/3, 2, 2, 2, 1/2, 1/2, 1/2)$; with LD.

Test/#nfRVs	0	4	8	16	32
UminP	.489	.479	.452	.365	.318
Score	.599	.538	.491	.380	.276
SSU	.603	**.624**	**.635**	**.581**	**.574**
SSUw	.532	.561	.574	.506	.493
Sum	.342	.312	.315	.258	.239
KMR(Linear)	.611	**.630**	**.644**	**.597**	**.590**
KMR(Quad)	.545	.563	.565	.493	.474
CMC	.296	.283	.189	.182	.365
wSum	.369	.297	.287	.191	.200
aSum-P	.350	.323	.325	.258	.243
C-alpha-P	**.629**	**.650**	**.668**	**.607**	**.598**
Step-up	.524	.516	.532	.429	.409

Discussion

- No UMPU test!
  Test selection? selecting the most powerful one (Pan et al 2009, *Hum Hered*).
  Highly adaptive tests, e.g. aSPU (Pan et al 2014 , *Genetics*).
- SSU: Applied to detect gene-gene and gene-environment interactions (Pan 2010 *Hum Hered*).
  aSPU?
- Main results applicable to other GLMs or regressions in general!
  Why do we always use the score/Wald/LR test in regression?
  They are **not** UMPU (though they are UMPI).
  Ignore correlations, as in the SSU test?
  Reduce # parameters, as in the Sum test? Tukey's 1-DF test!

Acknowledgement: I'd like to thank my collaborators and especially my current and former students. This research was supported by NIH.

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

Thank you!