Discussion

Wei Pan[1]
[1]Division of Biostatistics, School of Public Health
University of Minnesota

ENAR Meeting on March 19, 2014

Outline

- General: why/how does KMR work?
  its connections to other methods.
- Specifics: choice of the kernel
- Main refs:
  - Pan (2009, *Genetic Epi*): SSU, SSU = an EB test of Goeman et al (2006, *JRSS-B*);
  - Han and Pan (2011, *Genetic Epi*): SSU = GDBR (Wessel and Schork 2006, *AJHG*; McArdle and Anderson 2001, *Ecology*);
  - Pan (2011, *Genetic Epi*): KMR = SSU = GDBR

KMR, SSU, Goeman's EB test, GDBR, ...

- My experiences mainly with SNP/seq data:
  1) SNP data: Goeman's test (Chapman and Whittaker 2008);
  SSU=Goeman's test (Pan 2009);
  2) SNP data: GDBR (Lin and Schaid 2009);
  3) Seq data (RVs): SSU=KMR (Basu and Pan 2011); SKAT (Wu et al 2011, 2012, ...)
  Recently, neuroimaging data.
- KMR: a semi-parametric model

$$\text{Logit } \Pr(Y_i = 1) = \beta_0 + h(X_i), \tag{1}$$

  $h()$ is unspecified, but determined by a kernel K.
- $h = (h_1(X_1), ..., h_n(X_n))' \sim F(0, \tau^2 K)$,
  $K = K(\rho) = (K_{ij})$ with $K_{ij} = K(X_i, X_j)$.
- H_0': $h = 0$ becomes H_0: $\tau = 0$.

- Score test statistic for H_0 is (proportional to)

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

- Since K is symmetric and p.s.d, $K = ZZ'$.
  A linear kernel $K = XX'$, $Z = X$.
- Fit a parametric logistic reg model:

$$\text{Logit } \Pr(Y = 1) = \beta_0 + Z\beta, \quad (2)$$

- Score vector $U = Z'(Y - \bar{Y}1)$
- SSU test: $T_{SSU} = U'U = Q \Longrightarrow$ SSU=KMR if $K = ZZ'$.
  $T_{Sco} = U'\text{Cov}(U)^{-1}U$.
- GDBR: nonparametric MANOVA

$$F = \frac{tr(HGH)}{tr[(I - H)G(I - H)]},$$

$G = (I - 11'/n)A(I - 11'/n)$, $A = (-D_{ij}^2/2)$, $D = (D_{ij})$ with

$D_{ij} = d(X_i, X_j)$.
$H = y(y'y)^{-1}y'$.

- If $G = ZZ'$, then $F = T_{SSU}$.
  More, if $K = ZZ'$, then $F = T_{SSU} = Q$, GDBR=SSU=KMR!
- SSU = Goeman's test (Pan 2009).
  Ballard et al (2009): Goeman's test = a variance component-based score test (Tzeng and Zhang 2007).
- Why these relevant?
- 1) Choice of the kernel: not easy,
  K has to be p.s.d., why? if not, then ...
  SSU=KMR: use transformed Z, not X, in logistic reg;
  BUT ......
- 2) Can use multiple kernels, even transformed Z, then combine, or use other tests (e.g. Score test) (Han and Pan 2011);

- 3) Can generalize KMR, through SSU, to more complex data (Wang et al 2013);
- 4) Some optimality property:
  Goeman's test: highest **average** local power (Goeman et al 2006).
  No (local) uniformly most powerful test for multiple parameters (Cox and Hinkley 1974).
- Extensions to multivariate phenotypes: Hua and Ghosh (2014).

Specific choice of the kernel

- Metabolomic data:
  Two types: missing (0) or not; if not then abundance.
  Missing: truncation and more ?
- A distance kernel:
$$K_d(X_i, X_j) = \exp\left\{\frac{-d^2(X_i, X_j)}{\rho}\right\}.$$
- $$d(X_i, X_j) = \sqrt{\sum_k I(\delta_{X_{ik}} = \delta_{X_{jk}}) + \sum_k (X_{ik} - X_{jk})^2}.$$
- +: use the two types of data;
  challenge: trade-off b/w the two components;
- A stratified kernel:
  1) if the same missing pattern ($\delta_{X_i} = \delta_{X_j}$), then
  $K_s(X_i, X_j) = K_d(X_i, X_j)$;

2) o/w, $K_s(X_i, X_j) = 0$;

- +: more general, but maybe too extreme.
- Other features: testing a group of metabolites;
  An interesting grouping method: connected subgraphs based on marginal $\text{Corr}(X_i, X_j)$'s.

Acknowledgement: This research was supported by NIH.

You can download our papers from

http://www.biostat.umn.edu/rrs.php

Thank you!